

Fast Detection of Community Structures using Graph Traversal in Social Networks

Partha Basuchowdhuri^{*1}, Satyaki Sikdar^{†2}, Varsha Nagarajan¹, Khusbu Mishra¹,
Surabhi Gupta¹, and Subhashis Majumder¹

¹Department of Computer Science and Engineering,
Heritage Institute of Technology, Kolkata, WB, India

²Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame, IN, USA

Abstract

Finding community structures in social networks is considered to be a challenging task as many of the proposed algorithms are computationally expensive and does not scale well for large graphs. Most of the community detection algorithms proposed till date are unsuitable for applications that would require detection of communities in real-time, especially for massive networks. The Louvain method, which uses *modularity maximization* to detect clusters, is usually considered to be one of the fastest community detection algorithms even without any provable bound on its running time. We propose a novel graph traversal-based community detection framework, which not only runs faster than the Louvain method but also generates clusters of better quality for most of the benchmark datasets. We show that our algorithms run in $O(|V| + |E|)$ time to create an initial cover before using modularity maximization to get the final cover.

Keywords — community detection; Influenced Neighbor Score; brokers; community nodes; communities

1 Introduction

Networks can be realized by a graph data structure defined by $G = (V, E)$, where V denotes the vertex set, E denotes the edge set and $n = |V|$, $m = |E|$. These networks exhibit certain implicit characteristics like community structures. *Communities* in a network represent groups of vertices having dense intra-connections but sparse inter-connections. In this paper, we use the terms *cluster* and *community* interchangeably. In a social network, a basic assumption is that the shortest paths are used to propagate information. Communities are connected via nodes, termed as *brokers*, that are present on a large number of shortest paths among pairs of nodes in the network and hence, control the spread of information between communities. These nodes mark the boundary of communities. In our algorithm, we find such broker nodes that help to reach a new community and then spread the information among its members. In the process, we identify the nodes that are influenced by the brokers and also exhibit a high probability of belonging to the same community.

Most of the existing community detection algorithms involve lots of computations and hence are time-consuming. In 2002, in one of the early attempts to find communities from social networks, Girvan and Newman (Girvan and Newman, 2002) proposed an algorithm to detect hierarchical communities using repeated occurrence of an edge in all pair shortest paths as a metric. It was termed as the *brokerage* value of

^{*}Corresponding Author: parthabasu.chowdhuri@heritageit.edu

[†]The work was done when the author was at Heritage Institute of Technology

an edge. This is one of the first papers that points out the significance of finding broker nodes or edges for detecting communities in social networks. This algorithm has a very high computational demand even for the sparse networks. Likewise, most of the other community detection algorithms are also computationally expensive and some of them suffer from the major disadvantage of detecting only disjoint communities. However, real world networks are generally found to have overlapping communities. Some algorithms take the number of communities required as an input, but, it may not always be possible to estimate the number of clusters without prior analysis of the network, if someone wants to find the best set of clusters.

In our paper, we try to overcome most of these existing drawbacks found in the known algorithms (Fortunato and Hric, 2016). Our algorithm uses popular traversal techniques like depth first traversal and breadth first traversal and proves to perform considerably better and faster than other existing algorithms.

2 Prior Works

In this section, we mainly focus on the community detection algorithms known for having lesser running time. OPTICS (Ankerst et al., 1999), a density based clustering works well with the benchmark datasets. It chooses an outlier point to start the algorithm and then traverses through the points to draw a plot to mark the denser and sparser regions and thereby detecting the clusters. We have guided our algorithm in a similar manner using a graph structure but have achieved a linear running time in the process. An overlapping community detection algorithm, ONDOCS (Chen et al., 2009), takes help of visualization like OPTICS. It orders the nodes on the basis of their *reachability scores*, which helps the user to understand the emerging network structure. After initial visualization, selected parameters are used for extracting communities, hubs, outliers from the network. It finds overlapping communities in a network with a worst case running time of $O(n \log n)$.

A modified version of overlapping Girvan-Newman (GN) algorithm (Gregory, 2008) was proposed to detect overlapping communities on the basis of a local form of betweenness. It discovers small-diameter communities in large networks and has a time complexity of $O(n \log n)$ for sparse networks. In one of his seminal papers, Mark Newman presented the notion of modularity (Q) (Newman, 2006) of a clustering or a cover in a network, using a concept of minimization of inter-cluster edges and maximization of intra-cluster edges. Modularity, equipped with mathematical versatility, was very popularly used by the researchers in measuring goodness of covers and related topics. It opened up a new problem, popularly known as the modularity maximization problem. The decision problem corresponding to this optimization problem was later proved to be NP-complete (Brandes et al., 2006) and it is well accepted that heuristics can provide reasonably good solutions to the problem (Good et al., 2010). Newman himself presented a solution (Newman, 2004) to the problem but it was computationally expensive and practically infeasible for massive graphs. Later, as an extension of that work, a disjoint community detection technique was proposed by Clauset, Newman and Moore, now popularly known as CNM (Clauset et al., 2004). It is essentially a greedy algorithm that uses efficient data structures to store and find the maximum gain in modularity incrementally and eventually finds communities in sub-quadratic running time. Another greedy modularity maximization algorithm by Blondel et. al. (Blondel et al., 2008), used a simplistic approach of looking into the neighbors of a node to look for increase in modularity. After deciding on a tentative split, it shrinks the network, thereby drastically reducing future computation. This method is popularly known as the Louvain method. Local notion of modularity has also been used for detection of communities by modifying the equation of modularity and including a parameter to address the resolution limit (Xiang et al., 2016).

Raghavan et. al. proposed a fast community detection algorithm (Raghavan et al., 2007) popularly known as the label propagation algorithm (LPA). It runs multiple breadth-first searches in successive iterations in a random manner such that labels propagate locally and after a few iterations converge to provide a stable final cover. This algorithm has a running time $O(m + n)$. The algorithm has several disadvantages - for example, it searches for the similar nodes locally by spreading labels to adjacent nodes and it needs multiple iterations of breadth-first traversals. The convergence of node labels can be mathematically guaranteed but it is not known if the number of iterations needed for the convergence of the node labels is dependent on n , m or some other network parameter. Another community detection algorithm that claims to work fast in practice

is Infomap (Rosvall and Bergstrom, 2008). In this algorithm, the problem of community detection has been transformed into compression of information during its flow in the network. The algorithm uses random walks to move within the network and uses entropy-based information compression policies to find out the final cover. More recently, another linear time community detection technique (CGA) was presented by Yu Wang et. al. (Wang et al., 2010) that detects communities in social networks taking into account information diffusion. It detects community structure in social networks using an approach of label propagation with a worst case running time of $O(m)$. Although it has a provable bound, there was no empirical proof to justify that it runs faster than the Louvain method. Also, their results could not be reproduced due to unavailability of any public release from the authors of the papers. The theoretical background of our method is different from these algorithms or methods but it has a similar bound for its running time. Another community detection technique that uses depth first traversal is LexDFS (Creusefond et al., 2017). The worst case time complexity of LexDFS algorithm has been reported to be $O(n \log n)$. We pit the performances of our method against the popular fast community detection techniques with publicly available releases. The Louvain method is widely accepted as the present state-of-the art in terms of finding disjoint communities from a network. Therefore, any improvement on the results obtained from the Louvain method can be considered as an improvement of the state-of-the-art. If Louvain method is started from a bias (i.e., a cover is fed as an input) instead of starting from the original network, the method can be faster and the structure of the final cover will largely depend on the initial bias. Therefore, a cover generated from a given bias might be quite different from the final cover generated by the Louvain method when it is run on the original network. This motivates us to generate a bias such that we can generate final covers that are better in quality.

Recently, local searches have been used to look into the neighborhood of a vertex to find the best possible community for that vertex (Cui et al., 2014). A new area of classifying nodes and thereby predicting communities is node and community embedding (Zheng et al., 2016; Wang et al., 2017; Lin et al., 2017). Usually node embedding outputs a vector representation for each node in the graph, such that two nodes being “close” on the graph have similar vector representations. Another recent method has used Jaccard co-efficient to find communities (Meghanathan, 2016). In this paper, Jaccard co-efficient has been used as a measure to detect locally dense group of nodes as the metric finds similarity between two adjacent nodes by detecting common neighbors between them. Similar to GN method, it detects communities by successive removal of edges ordered on the basis of non-decreasing values of Jaccard co-efficient.

3 Problem Formulation

A vertex v is said to be influenced if a piece of information spreads to it from its neighbors (referred to as $\Gamma(v)$). Evidently, this is a temporal feature of the nodes and we assume that if a node u gets influenced at time $t = i$, it remains influenced thereon, and all its uninfluenced neighbors get influenced at time $t = i + 1$. This is similar to applying the breadth-first traversal algorithm in a graph.

Definition 1 Influence. *A node $v \in V$ is said to be influenced, once it has been discovered by using a graph traversal method.*

Definition 2 Influenced Neighbors Score. *The influenced neighbors score of a vertex v at time $t = i$, $INS(v)_{t=i}$, is calculated as the fraction of the number of neighbors of vertex v that have been influenced up to the previous time-stamp, i.e., $t = i - 1$.*

$$INS(v) = \frac{\text{Number of influenced neighbors of } v}{\text{Degree}(v)} \quad (1)$$

Clearly, $INS(v)$ lies between 0 and 1. For the starting node, it is zero. If all the neighbors of v have been influenced before v is processed, then $INS(v)$ will be 1. Initially INS value for all the nodes are unassigned. During the influence propagation, each node is accessed and its INS value is calculated. Clearly, INS value of a node is calculated only once, i.e., when it is discovered for the first time from the neighborhood of the

node that is being processed. Therefore, calculation of INS value of a node is dependent on the order of information spread (i.e., the traversal order), which, in turn, is dependent on the choice of the starting node.

If t is not explicitly mentioned for $INS(v)$, then it means INS value has to be calculated for the current time-stamp. We explain how to calculate INS value of a node with an example shown in Figure 1. In this figure, we see that the INS value for node B is being calculated. At the time of calculating $INS(B)$, a few nodes have already been influenced. In Figure 1, we can see that C, D, I, K, L, M and N have been influenced. They have been shown in grey color in order to categorize them differently from the other nodes. If INS value of B is being calculated at $t=i$, then the nodes in grey have been influenced at any time-stamp between time-stamps $t = 1$ to $i - 1$. At the time of calculating $INS(B)$, we see how many nodes are in the neighborhood of B and how many of them are already influenced. C is the only influenced neighbor of B out of its four neighbors $\{A, C, E, F\}$, therefore $INS(B)$ is calculated to be $\frac{1}{4}$.

The intuition behind the definition of INS comes from the fact that if two neighboring nodes are in same community and they are highly likely to have lots of common neighbors. During the influence propagation, if one of them is reached first, the influence will reach the other node and the common neighbors in the next time-stamp. If the other node is processed in the next time-stamp then it will find that many of its neighbors have already been influenced. It gives us an idea that the node is in a closely knit community and the information it wanted to propagate is already with many of its neighbors. The idea is intuitive and was first mentioned by Granovetter (Granovetter, 1973). Here, we present a mathematical form to use it in our proposed method.

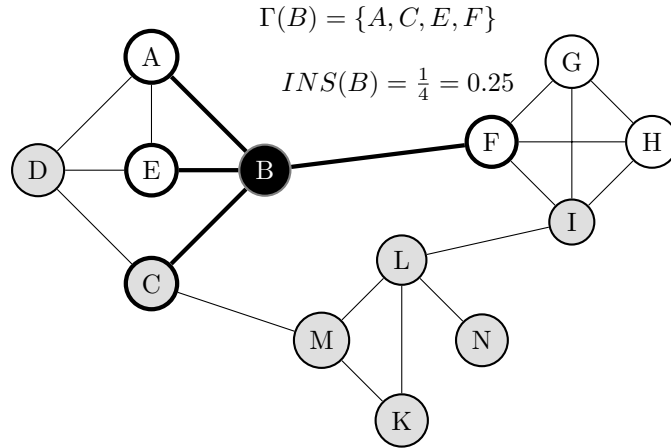


Figure 1: Calculating INS value of a node. The figure shows how $INS(B)$ is calculated.

Definition 3 Cut. In a graph $G(V, E)$, cut of a cluster s is defined as the sum of the weight of edges from cluster s to its complement $V \setminus V_s$ (Whang et al., 2013).

Definition 4 Conductance. Let V_s be the set of vertices in a cluster s . The conductance of s can be defined as the cut dividing the least number of edges incident on either the cluster or the rest of nodes in the network ($V \setminus V_s$). In other words, it is the probability of leaving the cluster by a one-hop walk starting from the smaller set between V_s and $V \setminus V_s$ (Whang et al., 2013).

The formulation of *conductance* can be given by,

$$COND(s) = \frac{C(V_s, V \setminus V_s)}{\min(C(V_s, V), C(V \setminus V_s, V))}, \quad (2)$$

where, $C(A, B)$ is defined as the sum of the weights of edges between subgraphs with node sets A and B . Note that $A \cap B$ may not always be a null set. Fig. 2 shows an example how conductance of a cut set $s = \{A, B, C, D, E\}$ is computed for the given graph.

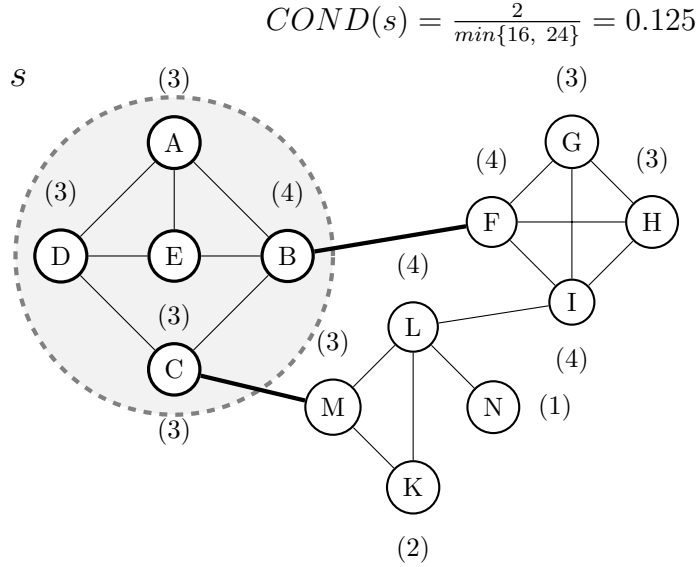


Figure 2: Calculating conductance of a cut set $s = \{A, B, C, D, E\}$. Degrees of the nodes are given in parentheses.

Definition 5 Broker Nodes. A vertex v is said to be a broker node, if at present time-stamp, $INS(v)$ is less than a predefined threshold r , where $0 \leq r \leq 1$.

The fraction r is termed as *community threshold* and may differ from one network to another for finding the best available community structure present in the network. Empirically, it was found (as shown in Table 8) that the variation of the quality of the communities obtained is marginal when r lies in the range of 0.6 to 0.8. In a sparse graph, the community threshold for a network may have a smaller value than that of in a dense graph. The nodes that are not broker nodes are declared as *community nodes*.

Definition 6 Community Nodes. A vertex v is said to be a community node, if at the present time-stamp or at $t = i$, $INS(v)$ is greater than or equal to a threshold r , where $0 \leq r \leq 1$.

When using conductance as the objective function, broker nodes are defined as those nodes that when added to a cluster, cause an increase or no change in the conductance value. Nodes that are not brokers are declared as community nodes. These definitions have been repeatedly used in the methods we have proposed in this paper.

Problem Statement

Given a graph $G = (V, E)$, where $n = |V|$ and $m = |E|$, find a cover (set) of k partitions (communities) from the hypothesis space consisting of all possible covers, represented by $C = \{C_1, C_2, C_3, \dots\}$, where C_i is the i -th cover from the hypothesis space, $V_i = \cup_{j=1}^k V_{ij}$ and $E_i = \cup_{j=1}^k E_{ij}$, such that the value of the clustering goodness measure is maximized over all such possible covers where $1 \leq k \leq n$.

Note that the goodness measure can be performed by any goodness function such as modularity (Newman, 2006) or overlapping modularity (Nepusz et al., 2007; Shen et al., 2009; Nicosia et al., 2009). In other words, for a graph G , if the value of the goodness function for cover C_i is given by $Q(G, C_i)$, then our aim is to find

$$\arg \max_{c \in C} Q(G, c) = \{Q(G, C_i) \mid \forall C_i, C_j \in C, Q(G, C_i) \geq Q(G, C_j)\}.$$

The target is to find the c , for which $Q(G, c)$ is maximum out of all possible covers from the hypothesis space C . Cover c is considered to be an overlapping partition if intersection of the vertex sets of any two clusters produces a set that is not null. Otherwise, the cover c is considered to be a disjoint partition. The problem presented here is essentially a goodness maximization problem. The decision version of goodness maximization problem, with modularity being the function for measuring goodness, has been proved to be NP-complete (Brandes et al., 2006).

For clarification, it should be noted that in our proposed method we try to achieve final clusters with good cluster quality. It is independent of any particular clustering goodness measure. We try to achieve high goodness measure without using any particular goodness measure as the objective function. Instead we use $INS(v)$ and $COND(s)$ as objective functions to maximize intra-cluster edges and minimize inter-cluster edges.

4 Traversal-based Community Detection

In this section, we propose a framework for two community detection methods which aim to find communities from a social network in linear running time (linear in terms of the size of the network) using traversal techniques. During the traversal, we first label nodes as brokers or community nodes based on an objective function. Next, we place the broker nodes in the community where most of its neighbors lie. We then reduce the graph on the basis of this initial split, where every cluster is merged into one super-vertex, with a weighted self-loop depicting the total number of intra-cluster edges. Finally, modularity maximization is run to generate the final cover. The method works as if a seed node is chosen to spread some information to the whole network. So the seed node spreads the information to all its neighbors and they in turn spread the information to their neighbors, who do not have the information yet. This process goes on iteratively and in the process we analyze the path of traversal to find out the communities. In this community detection algorithm, we use traversal techniques in sequential manner and thereby achieve a linear running time. We use the same framework to generate different clusters by devising two different methods on the basis of the objective functions mentioned in Section 3.

4.1 Part 1: Finding the Broker Nodes to Outline the Communities

The main idea of our method is to traverse the whole graph and partition it into a set of communities by observing how much of influence has reached a node’s neighborhood or how connected its neighborhood is. As mentioned earlier in this paper, communities are subgraphs with dense intra-connections and sparse inter-connections. Broker nodes reside in the bordering areas of a community, i.e., the areas where communities overlap. As a result, a piece of information that is exclusive to one community can spread to an adjacent community only via the broker nodes. These broker nodes behave as transition points between two or more communities. When a piece of information reaches to a community, it first reaches the broker nodes and then spreads among its members. Then, through some other broker nodes the information is passed on to another adjacent community. In our method, we follow the pattern of information flow via the broker nodes to explore and thereby detect the community structure of a graph in the process.

INS-based Algorithm:

In this method, a node in a graph is identified as either a broker node or a community node on the basis of its INS value. When a node is encountered during the traversal, we find the number of its neighbors that are carrying the information during that time stamp. If the fraction of neighbors of a node v carrying the information is less than r (say 0.75) then we assume that the information has not yet reached the community v belongs to, and v is one of the first nodes from its community to receive the information. Therefore, v is termed as a *broker* node and it is marked with a community label, which is same as its node label. Otherwise, the node is categorized as a *community* node.

Conductance-based Algorithm:

In this method, a node is identified as a broker when addition of the node in the presently growing cluster s increases $COND(s)$. Initially, when the process starts from a single node, the cluster consists of a single node and $COND(s)$ is 1. As new nodes are picked up during the traversal, we try to include the newly reached node in the cluster and check how the conductance changes. If it increases then it is considered to be a broker node, else the process to grow the cluster continues. Intuitively, it should stop at the borders of the prospective clusters. Nodes, other than the broker nodes, are marked as community nodes.

The community nodes identified during the traversal from one broker node to the next identified broker node are said to belong to the same community as the brokers mark the border of two communities. The community label for a group of community nodes found subsequently during traversal is same as the label of the broker node that led to the traversal of those community nodes.

4.2 Part 2: Allocating Broker Nodes to the Communities

In this part, we put the broker nodes in the communities they are most likely to belong. This decision is made based on a metric called belonging probability that calculates the ratio between the number of edges that exists from a broker node to all the nodes in a community out of the maximum number of edges possible between the broker node and that community. The idea behind defining such a metric is to find out the community to which a broker node is connected to with most number of edges. The absolute value of the number of connections to a community may not correctly represent the association of a broker node to that community. Increase in size of communities leads to increase in the possible number of connections with the broker node. Hence, if the number of actual connections are normalized by the possible number of connections that can be made with a community, the probability of a broker node belonging to that community remains independent of the size of the community. Eventually, broker nodes are placed in the community that leads to the highest belonging probability.

Some of the broker nodes may still fall into communities they are not supposed to be a part of as identified by the ground truth. The sequential nature of the traversal-based detection of communities may lead to such a situation. We solve this problem by identifying the community in which most of the neighbors of a broker node lie. In case of a tie (i.e., if the cardinality of the set of neighbors in a particular community maximizes for more than one community), we do not assign a community label to the broker node but leave it to the modularity maximization step to merge it with one of the clusters on the basis of the topological structure. A similar policy is maintained for the broker nodes with no community node in its neighbor. In the first part of modularity maximization, we reduce the present clusters into super-vertices and thereby transform the network into an undirected network of super-vertices. Every super-vertex consists of a self-loop, which has a weight equivalent to twice the number of intra-cluster edges in that cluster. The edges between a pair of super-vertices have weights equivalent to the number of inter-cluster edges between those two clusters. This step is termed as the reduction of the network. After the reduction, modularity changes for potential merges are calculated and increase in modularity is greedily maximized. In this step, the super-vertices are merged iteratively to produce the final cover. Due to the nature of the algorithms, the initial clusters are expected to be fragmented, however, with high precision. Therefore, a modularity maximization process may merge those high precision community fragments to achieve communities with both high precision and high recall. Such communities are likely to be more similar to the ground truth than the initial fragments.

4.3 Mathematical Definition of Community Nodes

The following definition of conductance is equivalent to the one defined in equation (2). In this section, we also abbreviate $COND$ to C for convenience.

$$C(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{k_S, k_{\bar{S}}\}}$$

where the numerator is the cut size, i.e., the number of edges from S to \bar{S} , k_S is the sum of degree of nodes in cluster S .

We are interested in the change in conductance of cluster S triggered by the addition of a single neighboring node (called the target node) to the cluster. If the conductance decreases after the addition, the target node is classified as a **community** node and added to the community S , otherwise it is classified as a **broker** node.

Given a graph and a partial cluster S , we can classify the nodes in the graph into three types,

1. the nodes currently in the cluster S ,
2. the target node t , which potentially could be added to S ,
3. the other nodes o which belongs to O (i.e., $V - S - \{t\}$).

Say, k_t is the degree of the target node, k_S and k_O are the degrees of the clusters S and O respectively, $k_{t,S}$ is the number of the edges incident from the target t to the cluster S , α is the number of edges incident from the cluster S to the set O . In other words, α represents the number of edges in the cut set **not** incident on the target node t .

We investigate the numerator of the definition, i.e., the cut size. Notice that the cut size initially (before t is added to S) is $\alpha + k_{t,S}$. After t is added to S , the cut size becomes $\alpha + (k_t - k_{t,S})$. This is because when t becomes a part of S , it brings k_t many edges with itself. Out of which only $k_t - k_{t,S}$ contribute to the cut size, since $k_{t,S}$ edges become part of the cluster S after the merge.

For the denominator, initially it is $\min\{k_S, (k_t + k_O)\}$. After t 's addition, it becomes $\min\{(k_S + k_t), k_O\}$. Since we pick the minimum of the two quantities, three cases arise:

1. $k_S < k_t + k_O$ and $k_S + k_t < k_O$,
2. $k_S < k_t + k_O$ and $k_S + k_t \geq k_O$,
3. $k_S \geq k_t + k_O$ (this guarantees that $k_S + k_t > k_O$, so there is no 4th case).

The initial conductance before t is added to S is represented as C_{old} and after t 's addition it is C_{new} . For each case, we find the condition on $k_{t,S}$.

Case I: $k_S < k_t + k_O$ and $k_S + k_t < k_O$

$$\begin{aligned} C_{old} &= \frac{\alpha + k_{t,S}}{k_S}, C_{new} = \frac{\alpha + k_t - k_{t,S}}{k_S + k_t} \\ C_{old} - C_{new} &= \frac{\alpha + k_{t,S}}{k_S} - \frac{\alpha + k_t - k_{t,S}}{k_S + k_t} \\ k_S \cdot (k_S + k_t) \cdot (C_{old} - C_{new}) &= (\alpha + k_{t,S}) \cdot (k_S + k_t) - k_S \cdot (\alpha + k_t - k_{t,S}) \\ &= \alpha \cdot k_S + \alpha \cdot k_t + k_S \cdot k_{t,S} + k_t \cdot k_{t,S} - \alpha \cdot k_S - k_S \cdot k_t + k_S \cdot k_{t,S} \\ &= k_{t,S} \cdot (2k_S + k_t) + k_t \cdot (\alpha - k_S) \end{aligned}$$

For t to be a community node, $C_{old} - C_{new} > 0$. Note that $k_S(k_S + k_t)$ is always positive. Therefore,

$$\begin{aligned} k_{t,S} \cdot (2k_S + k_t) + k_t \cdot (\alpha - k_S) &> 0 \\ k_{t,S} \cdot (2k_S + k_t) &> k_t \cdot (k_S - \alpha) \\ k_{t,S} &> \frac{k_t \cdot (k_S - \alpha)}{2k_S + k_t} \end{aligned} \tag{3}$$

Case II: $k_S < k_t + k_O$ and $k_S + k_t \geq k_O$

$$\begin{aligned}
C_{old} &= \frac{\alpha + k_{t,S}}{k_S}, C_{new} = \frac{\alpha + k_t - k_{t,S}}{k_O} \\
C_{old} - C_{new} &= \frac{\alpha + k_{t,S}}{k_S} - \frac{\alpha + k_t - k_{t,S}}{k_O} \\
k_S \cdot k_O (C_{old} - C_{new}) &= (\alpha + k_{t,S}) \cdot k_O - k_S \cdot (\alpha + k_t - k_{t,S}) \\
&= \alpha \cdot k_O + k_O \cdot k_{t,S} - \alpha \cdot k_S - k_S \cdot k_t + k_S \cdot k_{t,S} \\
&= k_{t,S} \cdot (k_O + k_S) + \alpha \cdot (k_O - k_S) - k_S \cdot k_t
\end{aligned}$$

For t to be a community node, $C_{old} - C_{new} > 0$. Note that $k_S k_O$ is always positive. Therefore,

$$\begin{aligned}
k_{t,S} \cdot (k_O + k_S) + \alpha \cdot (k_O - k_S) - k_S \cdot k_t &> 0 \\
k_{t,S} \cdot (k_O + k_S) &> k_S \cdot k_t + \alpha \cdot (k_S - k_O) \\
k_{t,S} &> \frac{k_S \cdot k_t + \alpha \cdot (k_S - k_O)}{(k_S + k_O)}
\end{aligned} \tag{4}$$

Case III: $k_S \geq k_t + k_O$

$$\begin{aligned}
C_{old} &= \frac{\alpha + k_{t,S}}{k_t + k_O}, C_{new} = \frac{\alpha + k_t - k_{t,S}}{k_O} \\
C_{old} - C_{new} &= \frac{\alpha + k_{t,S}}{k_t + k_O} - \frac{\alpha + k_t - k_{t,S}}{k_O} \\
k_O \cdot (k_O + k_t) \cdot (C_{old} - C_{new}) &= (\alpha + k_{t,S}) \cdot k_O - (k_t + k_O) \cdot (\alpha + k_t - k_{t,S}) \\
&= \alpha \cdot k_O + k_O \cdot k_{t,S} - \alpha \cdot k_t - k_t^2 + k_t \cdot k_{t,S} - \alpha \cdot k_O - k_O \cdot k_t + k_O \cdot k_{t,S} \\
&= k_{t,S} \cdot (2k_O + k_t) - k_t \cdot (\alpha + k_t + k_O)
\end{aligned}$$

For t to be a community node, $C_{old} - C_{new} > 0$. Note that $k_O(k_O + k_t)$ is always positive. Therefore,

$$\begin{aligned}
k_{t,S} \cdot (2k_O + k_t) - k_t \cdot (\alpha + k_t + k_O) &> 0 \\
k_{t,S} \cdot (2k_O + k_t) &> k_t \cdot (\alpha + k_t + k_O) \\
k_{t,S} &> \frac{k_t \cdot (\alpha + k_t + k_O)}{2k_O + k_t}
\end{aligned} \tag{5}$$

Equations 3, 4 and 5 give the conditions for classifying the target node to be a community node. Combining them together we can say,

$$k_{t,S} > \begin{cases} \frac{k_t \cdot (k_S - \alpha)}{2k_S + k_t} & \text{if } k_S < k_t + k_O \text{ and } k_S + k_t < k_O \\ \frac{k_S \cdot k_t + \alpha \cdot (k_S - k_O)}{(k_S + k_O)} & \text{if } k_S < k_t + k_O \text{ and } k_S + k_t \geq k_O \\ \frac{k_t \cdot (\alpha + k_t + k_O)}{2k_O + k_t} & \text{if } k_S > k_t + k_O \end{cases} \tag{6}$$

4.4 Illustrating Examples - Classifying Community and Broker Nodes

4.4.1 Example 1

Figure 3 and 4 have been used to illustrate the first example. From Figure 3, we can calculate the parameters needed to classify the target node M as a community node or a broker node. The parameters, in this example, are as follows

$$S = \{L, N\}, k_S = 5, t = M, k_t = 3, k_{t,S} = 1, \alpha = 2, k_O = 32.$$

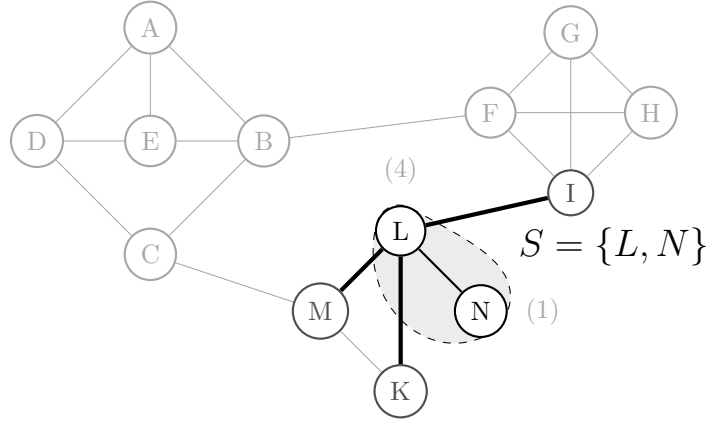


Figure 3: The graph with the cluster $S = \{L, N\}$. The cut edges are drawn with thick lines. The number in parentheses near the nodes represent the degree.

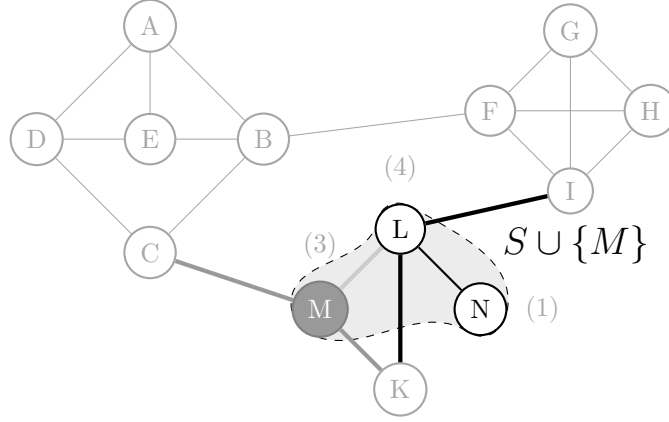


Figure 4: M is the target node. The number in parentheses above the nodes represent the degree of the nodes. The edges contributing to $k_{t,S}$, $(k_t - k_{t,S})$ and α are shown in light gray, gray and black respectively.

We have $k_S < k_t + k_O$ and $k_S + k_t < k_O$. Thus, we check the condition for case I to check if M is a community node.

$$k_{t,S} > \frac{k_t \cdot (k_S - \alpha)}{2k_S + k_t}$$

Here, $LHS = k_{t,S} = 1$. Plugging in the values in the RHS, we obtain,

$$\begin{aligned} RHS &= \frac{k_t \cdot (k_S - \alpha)}{2k_S + k_t} \\ &= \frac{3 \cdot (5 - 2)}{2 \cdot 5 + 3} \\ RHS &= \frac{9}{13} \end{aligned}$$

Thus, we have LHS equal to RHS. From Eq. 6, we can say, I is **not** a community node, and is therefore added to S (as shown in Figure 4).

4.4.2 Example 2

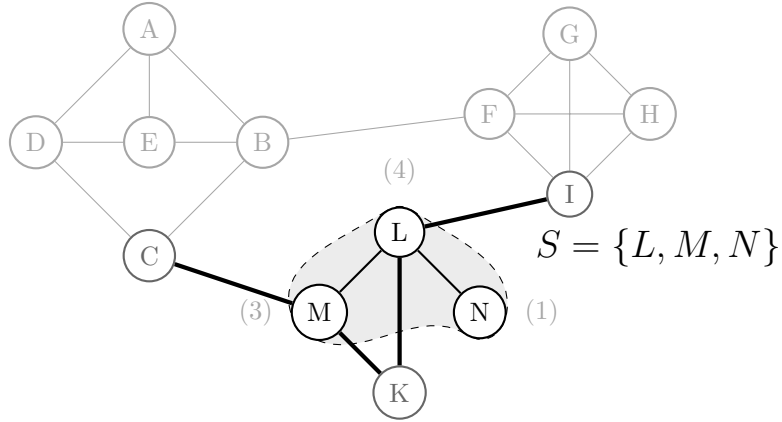


Figure 5: The graph with the cluster $S = \{L, M, N\}$. The cut edges are drawn with thick lines. The number in parentheses near the nodes represent the degree.

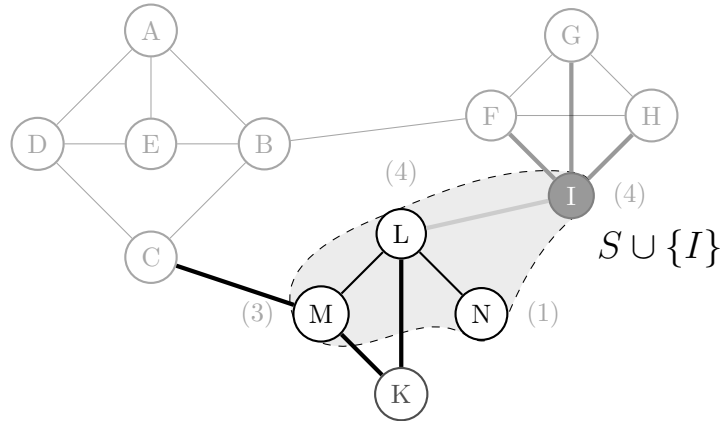


Figure 6: I is the target node. The number in parentheses above the nodes represent the degree of the nodes. The edges contributing to $k_{t,S}$, $(k_t - k_{t,S})$ and α are shown in light gray, gray and black respectively.

Figures 5 and 6 illustrate the second example. From Fig. 3, we can calculate the parameters needed to classify the target node I . The parameters, in this example, are as follows:

$$S = \{L, M, N\}, k_S = 8, t = I, k_t = 4, k_{t,S} = 1, \alpha = 3, k_O = 28.$$

We have $k_S < k_t + k_O$ and $k_S + k_t < k_O$. Thus, we check the condition for case I to check if I is a community node.

$$k_{t,S} > \frac{k_t \cdot (k_S - \alpha)}{2k_S + k_t}$$

Here, $LHS = k_{t,S} = 1$. Plugging in the values in the RHS, we obtain,

$$\begin{aligned} RHS &= \frac{k_t \cdot (k_S - \alpha)}{2k_S + k_t} \\ &= \frac{4 \cdot (8 - 3)}{2 \cdot 8 + 4} \\ RHS &= \frac{4 \cdot 5}{16 + 4} = 1 \end{aligned}$$

Thus, we have LHS equal to RHS . From Equation 6 we can say, M is **not** a community node, but a **broker** and is therefore not added to S .

4.5 Proposed Algorithms

We have described our proposed method in algorithms 1, 2, 3, 4, 5, 6, 7, 8 and 9, incrementally. Algorithm 1 describes the central LINCOS method algorithm replaced by method, where a broker stack S and a community queue Q have been used as primary data structures to facilitate the traversal. We call other procedures from Algorithm 1 to update S and Q and detect communities in the process. The phenomenon termed as ‘‘spread of information’’ has been described by Algorithms 3 and 2. The process of spread imitates breadth-first traversal. As a part of the data structures used, we have also used three node attributes to store flags for all the nodes, regarding whether it is a broker node or a community node (using `nodeType`), whether it has been covered during the traversal or not (using `covered`) and which community it presently belongs to (using `community`).

Algorithm 2 describes a procedure NODE-CAT-INS that is called by LINCOS (when using INS), until all the nodes in the graph have been categorized. In Algorithm 1, we call a generic version of the procedure referred to as NODE-CAT. But while presenting the pseudocodes, we have named the INS-based and COND-based versions of NODE-CAT as NODE-CAT-INS (presented in Algorithm 2) and NODE-CAT-COND (presented in Algorithm 9), respectively.

The purpose of NODE-CAT-INS is to traverse the untraversed nodes, mark them as traversed, categorize them as broker nodes or community nodes and finally, place the broker nodes in `brokerStack` and community nodes in `communityQueue`. Categorization and placement of nodes is performed in a few steps. In line 2 of the procedure NODE-CAT-INS, we use a procedure called SPREAD, which spreads the influence from a node v to all its uncovered neighbors. NODE-CAT very often makes use of a procedure that calculates the INS values of the neighbors of a node v , as described in Algorithm 4. Similarly, NODE-CAT-COND, is also used for categorization of nodes. Note that we do not use SPREAD in the COND-based method. In NODE-CAT-COND procedure, $COND(u)$ represents the subroutine to find out the value of conductance, as described in Algorithm 5. Please note that for a network with multiple connected components, just like any other traversal method, LINCOS can also be applied sequentially to all the components.

Algorithm 6 describes the procedure POST-PROCESS, which places the broker nodes in the clusters of the present cover G_s . Community label of a broker node v is assigned to the community that contributes to the largest fraction of community nodes in its neighborhood. If there is a tie or all the neighbors of a broker node consist only of brokers then the community label of the broker remains unassigned. Such brokers are assigned a community label during one of the latter procedure calls, referred to as MOD-MAXIMIZE. MOD-MAXIMIZE merges such brokers with the existing clusters such that the merged clusters would generate a cover with improved modularity. In MOD-MAXIMIZE, first we reduce the network by converting each of the initial clusters into a super-vertex. This is done by the procedure REDUCE. After reduction of the network, MOD-MAXIMIZE is applied in a way similar to the Louvain method (Blondel et al., 2008).

Complexity Analysis: Our algorithm uses breadth first and depth first traversal techniques, both known to have worst case time complexities of $O(|V| + |E|)$, where $|E|$ denotes the number of edges in the graph. The categorization of nodes takes $O(|E|)$ time as it involves traversal through all the edges. The time taken to place the broker depends upon the number of brokers obtained and their respective degrees. Number of brokers can never exceed the number of nodes and as the sum of degrees of all nodes is bounded by $2|E|$,

Algorithm 1: Traversal-based Linear Time Community Detection (LINCOM(G, r))**Input** : Undirected, unwt. graph $G(V, E)$, threshold (r)**Output:** Cover of k communities, $G_s = \{G_{s_1}, G_{s_2}, \dots, G_{s_k}\}$

```

1 begin
2   find  $v_{start} \in V, \exists v_{start}$  is the node with lowest degree
3   forall  $v \in V$  do
4     |  $v.nodeType \leftarrow v.covered \leftarrow 0$ 
5     |  $v.community \leftarrow v$ 
6    $S \leftarrow Q \leftarrow \emptyset$                                  $\triangleright$ brokerStack( $S$ ), communityQueue( $Q$ )
7   coverCount  $\leftarrow 1$ 
8    $v \leftarrow v_{start}$ 
9   NODE-CAT( $G, v, Q, S$ )                                     $\triangleright$ NODE-CAT-INS or NODE-CAT-COND
10  while coverCount  $< n$  do
11    | if  $Q$  is non-empty then
12    | |  $v \leftarrow dequeue(Q)$ 
13    | else
14    | |  $v \leftarrow pop(S)$ 
15    | | NODE-CAT( $G, v, Q, S$ )                                 $\triangleright$ INS or COND-based NODE-CAT
16  forall  $v \in V$  do
17    | check  $v.community$  to form  $G_s$ 
18  POST-PROCESS( $G_s$ )
19  REDUCE( $G_s$ )
20  MOD-MAXIMIZE( $G_s$ )
21  return  $G_s$ 

```

Algorithm 2: Categorizing uncovered nodes in $\Gamma(v)$ (NODE-CAT-INS(G, v, Q, S))**Input** : Undirected, unwt. graph $G(V, E)$, threshold (r),brokerStack (S), communityQueue (Q)**Output:** Update Q, S

```

1 begin
2   SPREAD( $v$ )
3   for  $u \in \Gamma(v)$  do
4     | if  $u.nodeType \neq 0$  then continue
5     | if  $INS(u) < r$  then
6     | |  $u.nodeType \leftarrow 1$                                  $\triangleright$ marking the broker nodes
7     | | push( $S, u$ )
8     | else
9     | |  $u.nodeType \leftarrow 2$                                  $\triangleright$ marking the community nodes
10    | | enqueue( $Q, u$ )
11    | |  $u.community \leftarrow v.community$ 

```

Algorithm 3: Spreading Influence to the Neighbors (SPREAD(v))**Input** : Undirected, unwt. graph $G(V, E)$, root node (v)**Output:** Update coverCount, covered list

```

1 begin
2   for  $u \in \Gamma(v)$  do
3     if  $u.covered = 0$  then
4        $u.covered \leftarrow 1$ 
5        $coverCount \leftarrow coverCount + 1$ 

```

Algorithm 4: Influenced Neighbors Score (INS(v))**Input** : Undirected, unwt. graph $G(V, E)$, root node (v)**Output:** INS(v)

```

1 begin
2    $coveredCount \leftarrow 0$  ▷keeps a count of the covered neighbors of  $v$ 
3   for  $u \in \Gamma(v)$  do
4     if  $u.covered = 1$  then
5        $coveredCount \leftarrow coveredCount + 1$ 
6    $INS(v) \leftarrow coveredCount / deg(v)$ 
7   return INS( $v$ )

```

Algorithm 5: Conductance (COND(s))**Input** : Undirected, unwt. graph $G(V, E)$, set of nodes V_s in cluster s **Output:** Conductance score of the set of nodes V_s in cluster s

```

1 begin
2    $cutSize \leftarrow 0$ 
3   forall  $u \in V_s$  do
4     for  $v \in \Gamma(u)$  do
5       if  $v \in V \setminus V_s$  then
6          $cutSize \leftarrow cutSize + 1$ 
7   return  $cutSize / \min\{\sum_{w \in V_s} deg(w), \sum_{w \in V \setminus V_s} deg(w)\}$ 

```

Algorithm 6: Post-processing of the Broker Nodes (POST-PROCESS(G_s))**Input :** Undirected, unwtcd. graph $G(V, E)$, root node (v)**Output:** Update v .community value for broker nodes

```

1 begin
2   forall  $v \in V$  do
3      $max(v) \leftarrow 0$ 
4   forall  $v \in V$  do
5     if  $v.nodeType = 1$  then
6       forall  $G_{s_i} \in G_s$  do
7         if  $\frac{|Neighbors\ of\ v\ in\ G_{s_i}|}{|G_{s_i}|} > max(v)$  then
8            $max(v) \leftarrow \frac{|Neighbors\ of\ v\ in\ G_{s_i}|}{|G_{s_i}|}$ 
9            $v.community \leftarrow$  community label that leads to  $max(v)$ 
10        else
11          if  $\frac{|Neighbors\ of\ v\ in\ G_{s_i}|}{|G_{s_i}|} = max(v)$  then
12             $v.community \leftarrow$  append community label to community list

```

Algorithm 7: Reduction of the Network (REDUCE(G_s))**Input :** G_s **Output:** Updated G_s (G'_s)

```

1 begin
2   forall  $G_{s_i} \in G_s$  do
3      $v'_i$  represents all the nodes in  $V_{s_i}$ 
4      $w(v'_i, v'_i) \leftarrow \sum_{u, v \in V_{s_i}} w(u, v)$ 
5      $w(v'_i, v'_j) \leftarrow \sum_{u \in V_{s_i}, v \in V_{s_j}} w(u, v) | u \neq v, i \neq j$ 
6   return  $G'_s$ 

```

Algorithm 8: Modularity Maximization (MOD-MAXIMIZE(G_s))**Input :** G_s **Output:** Updated G_s (G'_s)

```

1 begin
2   repeat
3     forall  $v \in G_s$  do
4       forall  $u \in \Gamma(v)$  do
5         find  $\Delta Q$  if  $v$  moved to  $u$ .community
6         if  $\Delta Q_{max} > 0$  then
7           move  $v$  to  $u_{max}$ .community with  $\Delta Q_{max}$ 
8   until there is no  $\Delta Q_{max} > 0$ 
9   return  $G'_s$ 

```

Algorithm 9: Categorizing uncovered nodes in $\Gamma(v)$ (NODE-CAT-COND(G, v, Q, S))

Input : Undirected, unwtcd. graph $G(V, E)$, brokerStack (S), communityQueue (Q)
Output: Update Q, S

```

1 begin
2   for  $u \in \Gamma(v)$  do
3     if  $u.nodeType \neq 0$  then continue
4     if  $u$  satisfies condition from Equation 6 then
5        $u.nodeType \leftarrow 2$                                 ▷marking the community nodes
6       enqueue( $Q, u$ )
7        $u.community \leftarrow v.community$ 
8     else
9        $u.nodeType \leftarrow 1$                                 ▷marking the broker nodes
10      push( $S, u$ )

```

the time complexity for this step is $O(|E|)$.

Lastly, the reduction procedure merges all the nodes in a community into one super-vertex and adds up all the inter-cluster and intra-cluster edges separately. Given, that community membership of a node can be accessed in constant running time, it has to traverse all the edges once. Therefore, the network reduction runs in $O(|E|)$. In modularity maximization part, we already work on a reduced graph and so its running time will not exceed the order of the reduced graph, because every node will need to traverse its neighbors trying to find the maximum modularity gain. Therefore, the overall worst case running time of our algorithm to create the initial bias for the community detection will be $O(|E|)$. The time complexity of the modularity maximization part is unknown (Blondel et al., 2008) but as it works on a reduced graph it works very fast in practice. We proceed to verify the performance of our method by showing some of the experimental results that we performed on some benchmark datasets.

5 An Example to Illustrate the Split Produced by INS-based Method

We illustrate our proposed method with a graph consisting of 13 nodes and 20 edges. The nodes are labeled by uppercase letters, as shown in Figure 7a. For this example, we take the threshold value of r to be 0.66. As in OPTICS, our algorithm also starts from an outlier-like point, i.e., a pendant node, which does not play a significant role in a community. The starting node will always be considered as a broker node with an INS value of 0. Initially, from the only pendant node N, influence is propagated to L, as seen in Figures 7a and 7b. $INS(L)$ evaluates to 0.25 because only one node (N) out of all the four of its neighbors (I, K, M, N) has become influenced up to this stage. According to our algorithm, L is identified as a broker and pushed into the broker stack. Community queue is still empty at this stage. So we pop the top element from the broker stack, i.e., L, and process it. By processing a node v we mean, propagating the influence to all the neighbors of v , then calculating INS value for all its neighbors, thereby categorizing them as broker nodes and community nodes and finally storing them in broker stack and community queue, respectively. For L, we spread the influence to I, K and M. INS values of I, K and M turn out to be 0.25, 1.0 and 0.67, respectively. Since $INS(I)$ being less than the value of r (0.66), I is placed in the broker stack and K and M are placed in community queue (we store them in lexicographic order, but it is not necessary). It should be noted that all the broker nodes discovered till this point have been assigned a community label that is same as their node label, as shown in Table 1.

But when a community node is discovered, the node is given a community label same as the label of the last processed broker node, which is essentially the broker node that leads to its discovery. For example, here $community(K) = community(M) = L$. Now, after insertion of K and M, community queue is non-empty

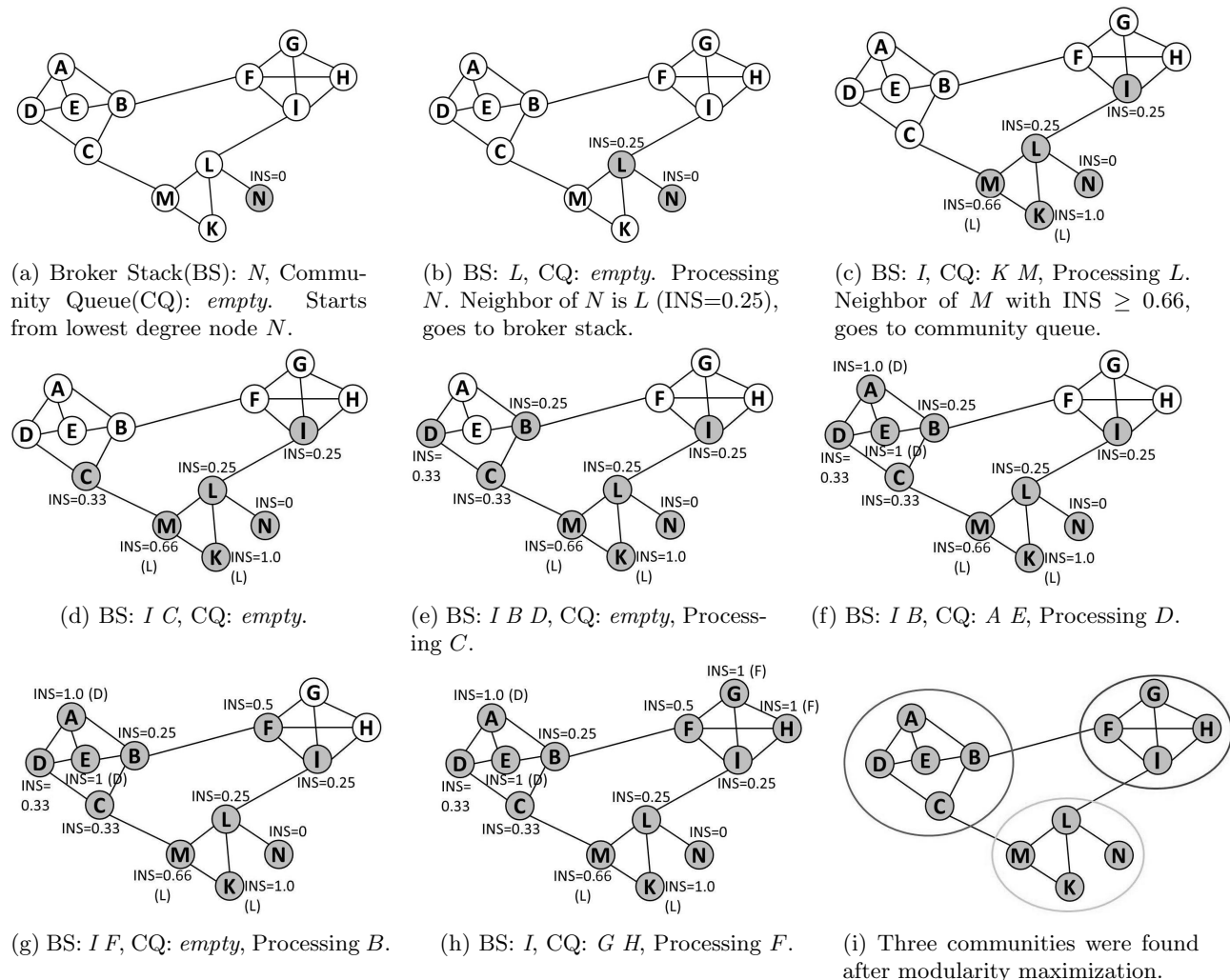


Figure 7: Steps to detect communities during the traversal-based algorithm LINCOM.

and hence the elements in the queue will be dequeued and processed until the queue is empty. The stack is not processed until the queue is empty. Therefore, K is processed first without spreading further influence to any other node in the graph. Then M reaches out to a single uncovered node C . $INS(C)$ is 0.33 and it is subsequently placed in the stack. At this stage, the stack consists of I and C . Figures 7c and 7d show these steps. In Figure 7e, we see the broker node at the top of the stack is getting popped and thereby getting processed. C spreads influence to all its uncovered neighbors (B, D). Due to low INS values, B and D are both identified as broker nodes and get pushed into stack. Now, after popping D and processing it, we reach nodes A and E . Both of them have the same INS value 1 and are identified as community nodes. The nodes in the graph's adjacency list are stored and processed in lexicographic order prompting B to be pushed to the stack before D . Therefore, D first comes to the top while popping and is processed before B . That is why A and E will also have a community label of D . This is shown in Figure 7f. Next node to be processed is B . It spreads to a new node F . As $INS(F) = 0.5$, F is stored at the top of the stack. Community queue being empty, F is processed in the next step. F spreads to the remaining two uncovered nodes (G and H) of the network. For both of them, INS value evaluates to 1.0 and hence they are categorized as community nodes with community label F , as shown in Figures 7g and 7h. After post processing, the communities turn out

Table 1: Sequence of discovering the nodes in the example

Node	INS	Category	Initial Label	Final Label
N	0	Broker	N	L
L	0.25	Broker	L	L
I	0.25	Broker	I	F
K	1	Community Node	L	L
M	0.67	Community Node	L	L
C	0.33	Broker	C	D
D	0.33	Broker	D	D
B	0.25	Broker	B	D
A	1	Community Node	D	D
E	1	Community Node	D	D
F	0.50	Broker	F	F
G	1	Community Node	F	F
H	1	Community Node	F	F

to be as shown in Figure 7i. Please note that this is an intermediate cover and not the final cover. The final cover is obtained by applying the procedure MOD-MAXIMIZE on the intermediate cover.

It is important to understand that the initial clustering generated by LINCOS is done by running two graph traversal methods, i.e., depth first traversal and breadth first traversal, in parallel. We start from a node assuming it is getting the spread started. Since it is getting the spread started, irrespective of the node actually being a broker node or a node inside a community, it will be considered as a broker node (once the traversals are over, during the post-processing, it will be placed into the correct community eventually). The traversal methods and the corresponding trees have been shown in Figure 8. In Figure 8a, we see how the nodes have been categorized into two different types - namely the broker nodes (white) and the community nodes (black). Note that the community nodes store the label of the broker node from which is was discovered. In Figure 8b, such broker nodes, that are used to label a community, is placed inside that community. In Figure 8c, the other broker nodes are placed inside the community, which has most number of edges connected to it. If there is a tie, then the broker node is considered as a singleton cluster and is passed on to modularity maximization process as a part of the initial cover. In Figure 8d, we observe that node C has equal number edges to both C_D and C_L . Modularity maximization agglomerates such a node to a cluster such that the overall modularity of the final cover is maximized.

From the illustrations, it can be observed that initially a lot of broker nodes are obtained. This is because, at the beginning of the spread propagation inside a community, most of the nodes within that community are unvisited. Therefore, in this method, some small-sized (possibly singleton) communities may be generated at the initial stage. But as the brokers are placed and the community labels of the nodes are stabilized, the communities grow larger and we get the final labels for all the nodes (shown as the final labels in Table 1). Similarly, the initial split can also be generated by using conductance. After applying modularity maximization on the obtained cover, the average size of the clusters grows further.

5.1 Effect of different starting points on the clusters

Due to the nature of the algorithm, we can pick any node to be the starting point for the traversal. Depending on the starting point, the order of traversal may change and the broker node that leads to the discovery of a community may change but discovery of the communities remain more or less similar. We have already

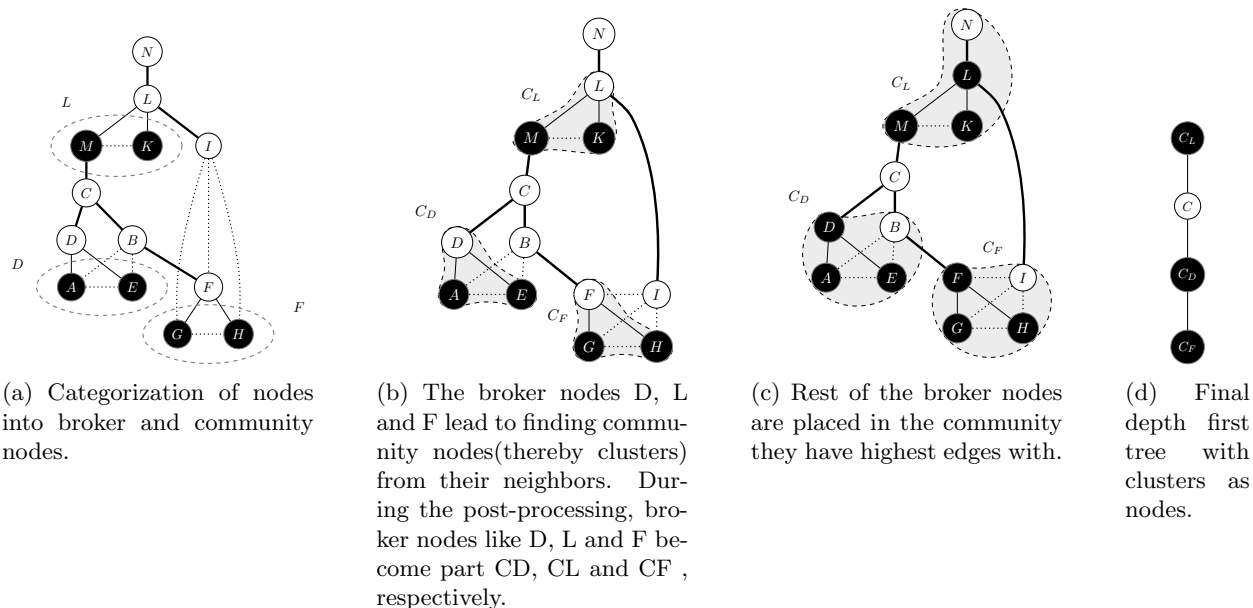


Figure 8: Transforming the network into a depth first tree of clusters. Breadth first traversal is used when traversing within the communities. The white nodes are the broker nodes and the black nodes are the community nodes. The thick lines mark the discovery of a broker node.

provided an example in Figure 7. If other starting points are used on the same network, same cover is produced every time. Figure 9a shows the traversal tree for starting node E, with back edges and cross edges in addition to the tree edges. In Figure 9b, only the tree edges have been shown along with the clusters after the first phase of the algorithm. In this figure, node B is actually connected to cluster C_E with a back edge. After modularity maximization, node B moves into C_E , C_M and C_L get merged and C_I remains as it is. This leads to the same final cover as in Figure 7. In experimental results part, we have provided sufficient empirical evidence that our method is independent of starting point.

6 Experimental Results

We have performed our experiments, including running the public releases of the Louvain method and CNM, on an Intel Xeon 2.4 GHz quad-core CPU desktop with 32GB RAM, 500 GB hard disk and Fedora LINUX version 3.3.4 OS. The source code has been written in C and it is publicly available¹.

We evaluate the performance of our algorithm on different well-known benchmark datasets (Zachary, 1977; Lusseau and Newman, 2004; Girvan and Newman, 2002; Yang and Leskovec, 2012; Klimt and Yang, 2004; Leskovec et al., 2007) by assessing the accuracy of its covers obtained, using a Newman modularity for the disjoint case and Nicosia modularity for the overlapping case. We further test our algorithm on large datasets to test its efficiency. We evaluate the effect of parameters used in our algorithm, i.e., threshold values for INS and choice of the starting nodes. The threshold value r for the INS value has a great impact on the size of the communities obtained. If r is assigned a low value, many nodes move to the same community during traversal, leading to mostly large sized communities and only a few smaller ones. On the other hand, taking high value for r leads to fragmented and small communities along with many broker nodes, specially in the initial part of the traversal. A large number of broker nodes may converge to a single community in the latter part of our method, thereby forming large-sized communities. We have observed that setting $r =$

¹The code can be downloaded from <https://github.com/sna-lincom/LINCOM>

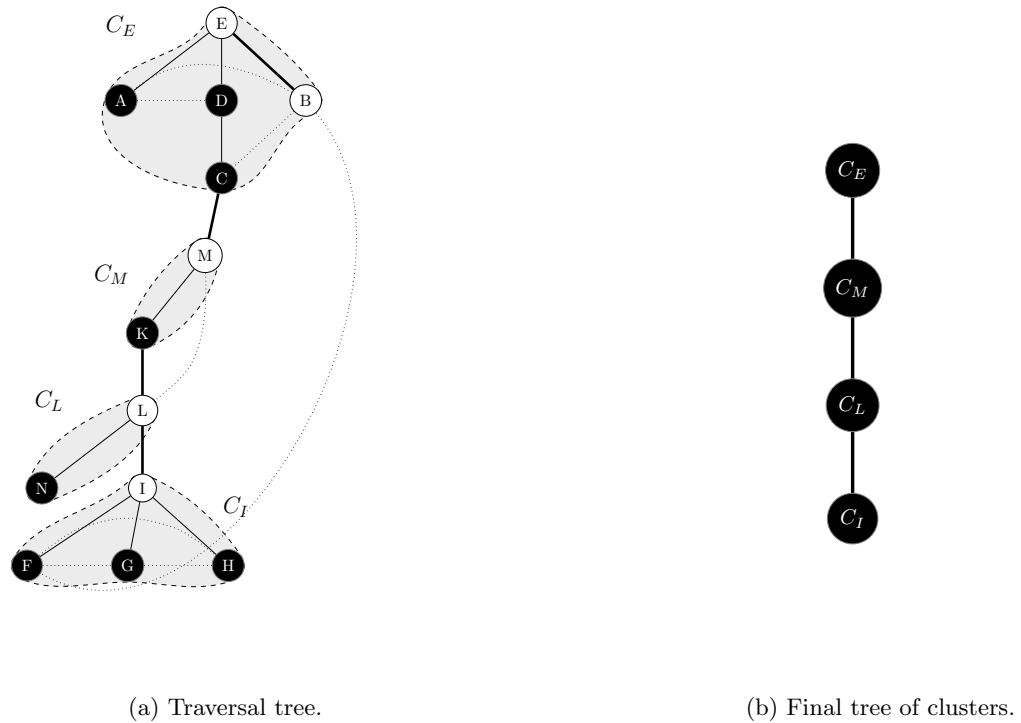


Figure 9: Discovery of clusters with E as the starting node.

0.75 as threshold value gives us appropriately sized (with significant relevance to ground truth) communities and therefore in other experiments, INS-based LINCOM was run with $r = 0.75$.

6.1 Testing efficiency of LINCOM

We run INS-based LINCOM and COND-based LINCOM along with the implementation of the Louvain method and CNM, released by their respective authors for a comparative analysis, using the same set-up declared above. The time taken to run those algorithms have been summarized in Table 2. Results show that both our methods run faster than the Louvain method. Clearly, CNM proves to be much slower, which is in coherence to the provable bounds for the running time of the algorithm. CNM cannot even find communities for massive datasets such as Orkut even in several hours. So, from the results, it can be empirically established that the concept of traversing the graph to detect communities turned out to be faster than the present state-of-the-art. The variation of the objective function did not seem to generate much perturbation in terms of running times. Hence, other objective functions suitable to maximize intra-cluster edges and minimize inter-cluster edges, can also be tested with similar efficiency unless the objective function itself is computationally expensive.

We have used CODACOM platform Creusefond et al. (2017) to run all the state-of-the-art community detection algorithms and have tested their performance on the real-world benchmark datasets and the synthetic graph data originally used by Newman and Girvan Girvan and Newman (2002). In Table 3, we have compared LINCOM with other methods in terms of efficiency using real-world benchmark datasets. In Table 6, we have shown the results of running LINCOM and a few other state-of-the-art community detection algorithms on synthetic datasets. We created these synthetic datasets using the LFR benchmark generator Lancichinetti et al. (2008). These graphs have 128 nodes with each node having degree of 16. All the nodes in the network is divided into 4 communities with each community consisting of 32 nodes. In these

Table 2: Comparison of running time for different types of community detection algorithms for different benchmark datasets.

Network Dataset	Nodes (n)	Edges (m)	LINC <small>OM</small>		Louvain (sec)	CNM (sec)
			INS (sec)	COND (sec)		
Karate	34	78	0	0	0	0
Dolphin	62	159	0	0	0	0
Lesmis	77	254	0	0	0	0
Football	115	613	0	0	0	0
GrQc	4,158	13,422	0	0	0	4
Enron	33,696	180,811	0.21	0.22	0.38	362
Epinions	75,877	405,739	0.71	0.76	0.97	1953
Amazon	334,863	925,872	4.58	4	6	3578
DBLP	317,080	1,049,866	5.141	4	11	10,440
Orkut	3,072,441	117,185,083	246	377	456	-

Table 3: Comparing running times of the state-of-the-art community detection algorithms with LINCOM for different benchmark datasets. Boldfaced numbers highlight the best time obtained for a given dataset.

Method	Amazon	DBLP	Dolphins	Enron	Epinion	Football	GrQc	Karate	LesMis
Louvain	3.46	4.36	0	0.28	0.73	0	02	0	0
LexDFS	13.02	14.18	0	3.40	11.24	0	0.10	0	0
Label Prop	27.72	43.87	0	0.69	0.93	0	03	0	0
Infomap	432.71	507.85	0	22.08	110.30	0	0.54	0	0
LINC <small>OM</small>	4.00	4.00	0	0.22	0.76	0	0	0	0

graphs, mixing parameter (μ) defines the fraction of edges with nodes outside its own community. We have changed the value of μ from 0.1 to 0.3 to test the robustness of our algorithm as the communities become less recognizable with increase in μ .

In order to reinforce our claim that the first phase of LINCOM (i.e., LINCOM without modularity maximization) runs in linear time, we have tested LINCOM on sample subgraphs of large datasets by randomly sampling 25%, 50% and 75% of the total edges of the networks. When we use LINCOM with modularity maximization, it shows non-linear growth in running time. If we run LINCOM without the modularity maximization part, results show that the running times scale up linearly. The results have been shown in Figure 10. This can be considered as an empirical confirmation of the fact that the running time of LINCOM is indeed bounded by a linear function of the number of edges in the network.

6.2 Testing Quality of the Clusters

Here, we have used overlapping modularity (Nepusz et al., 2007; Shen et al., 2009; Nicosia et al., 2009) for evaluating overlapping communities, whereas for disjoint communities, we use modularity defined by Newman (Newman, 2006). The variants of LINCOM are naturally overlapping, however, in order to compare them with disjoint communities on the basis of one goodness measure, we convert the overlapping commu-

Table 4: Comparison of the modularity values of the final covers generated by the Louvain method, CNM and the variants of LINCOS, for different benchmark datasets.

Network Dataset	Overlapping		Disjoint			
	LINCOS	LINCOS	LINCOS	LINCOS	Louvain	CNM
	INS	COND	INS	COND		
Karate	0.729	0.445	0.402	0.3793	0.415	0.38
Dolphin	0.75	0.756	0.518	0.526799	0.518	0.492
LesMis	0.579	0.579	0.544	0.448	0.55	0.5
Football	0.673	0.694	0.582	0.543	0.604	0.57
GrQc	0.879	0.87	0.847	0.841375	0.847	0.79
Enron	0.73	0.7	0.587	0.598	0.596	0.49
Epinions	0.529	0.51	0.44	0.44895	0.45	0.385
Amazon	0.931	0.963	0.961	0.92054	0.926	0.87
DBLP	0.831	0.819	0.818	0.809	0.819	0.73
Orkut	0.59	0.801375	0.548	0.679	0.679	-

Table 5: Modularity values of the final cover of the state-of-the-art community detection algorithms for different benchmark datasets. Boldfaced numbers highlight the highest values of modularity obtained for a given dataset.

Method	Amazon	DBLP	GrQc	Enron	Epinion	Football	Dolphins	Karate	LesMis
Louvain	0.926	0.821	0.847	0.613	0.452	0.6	0.518	0.392	0.554
LexDFS	0.536	0.414	0.552	0.171	0.094	0.576	0.338	0.23	0.434
Label Prop	0.785	0.7	0.769	0.317	0.044	0.584	0.410	0.352	0.523
Infomap	0.825	0.722	0.771	0.511	0.347	0.6	0.517	0.402	0.546
LINCOS	0.961	0.818	0.847	0.596	0.44	0.582	0.518	0.402	0.544

nities into disjoint communities. We place each overlapping node in one of its neighboring communities such that the modularity is maximized. A summary of the results have been presented in Table 4. Overlapping modularity of the covers generated by the variants of LINCOS are consistently greater than the modularity values of the covers produced by the Louvain method and CNM. The comparisons between the modularity values of the disjoint covers generated by the variants of LINCOS and the modularity maximization methods are close. Variants of LINCOS always seem to produce better clusters than that of CNM, particularly in large networks. The Louvain method and variants of LINCOS often produce covers with the same modularity value with a very low tolerance level.

In Table 5, we have compared performance of LINCOS with the performance of some of the state-of-the-art community detection algorithms. LINCOS consistently performs better than the others except the Louvain method. Louvain method outputs covers with similar modularity values in some cases. In Table 6, we have evaluated LINCOS’s performance on synthetic networks with 4 existing communities where the community structure becomes more obscure as the mixing factor increases. The ground truth is available and the modularity of the ground truth cover decreases as the mixing factor increases. From the results, LINCOS seems to be more robust than some of the other methods as it can recognize the ground truth structure more readily than the others.

Table 6: Performance of the state-of-the-art community detection algorithms on LFR generated synthetic benchmark datasets.

Method	μ		
	0.1	0.2	0.3
Ground truth	0.648	0.548	0.452
Louvain	0.648	0.548	0.452
LexDFS	0.648	0.548	0.328
Label Prop	0.648	0.548	0.230
Infomap	0.648	0.548	0.452
LINCOM	0.648	0.545	0.452
CNM	0.648	0.548	0.452

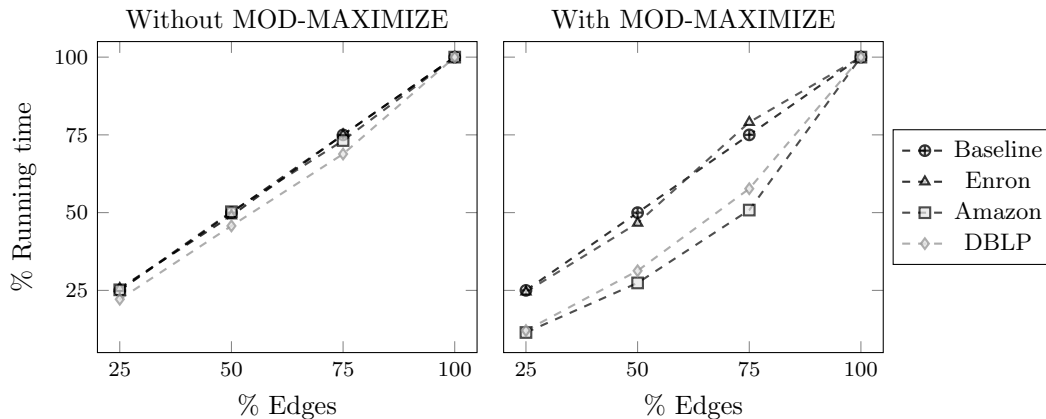
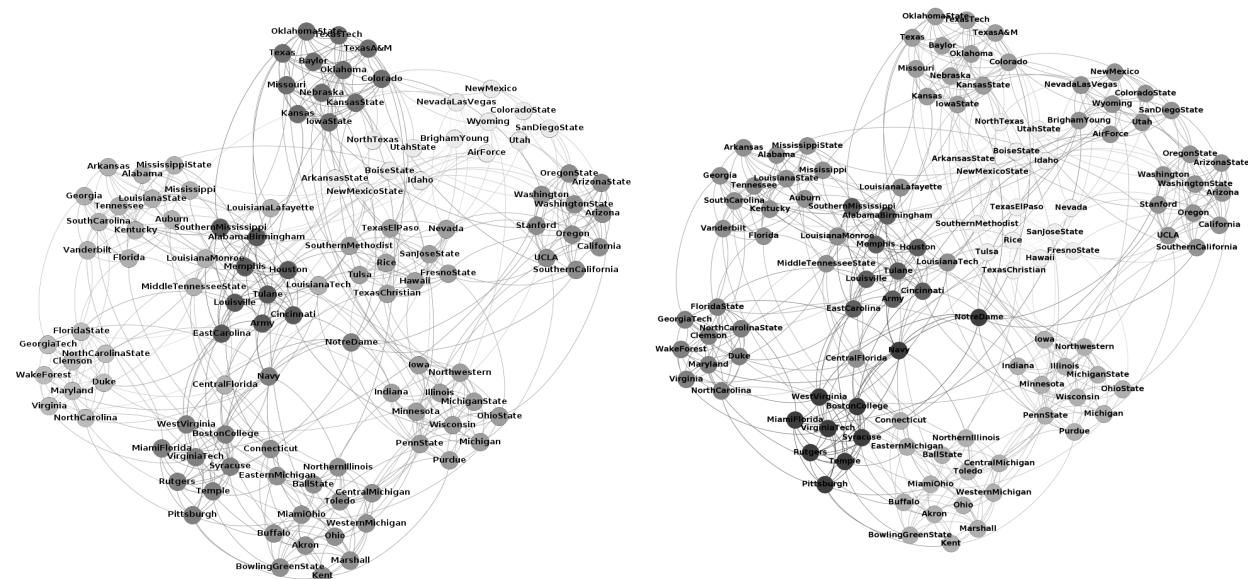


Figure 10: Growth of running time for LINCOM when tested on 25%, 50% and 75% of the edges of Enron, Amazon and DBLP datasets.

In every community detection algorithm, it is important to understand the practical significance of the method and what it can be used for. To understand that we compare the final cover obtained by INS-based LINCOM method (Figure 12b) with the ground-truth of the US College football network as described by Newman (Girvan and Newman, 2002) (Figure 11) and the final cover obtained by Louvain method (Figure 12a). This is one of the very few datasets where ground-truth from real-world is available and therefore can be used for benchmarking communities. We observe that Louvain generates a cover with ten clusters whereas LINCOM ends up generating a cover with eight clusters. Interestingly, the significance of the ground-truth communities is maintained both in Louvain as well as LINCOM. In the final cover generated by Louvain method we can observe that there are two cases where a couple of conferences have merged. Similarly, a further degree of coarsening has taken place in LINCOM, where two pairs of communities have merged to reduce the number of clusters further. Understandably, due to LINCOM’s use of modularity maximization technique, inability to detect small clusters has become an inherent problem. But the clustering found in LINCOM is defined by dense connection, a higher modularity value and evidently has a high precision when compared to the ground truth.



(a) Final cover obtained from Louvain method.

(b) Final cover obtained from LINCOM.

Figure 12: Visualization of the final covers as obtained from Louvain method and LINCOM.

Louvain method. Therefore, LINCOM guarantees that the communities produced are not too small to be considered insignificant. Smaller cover size ensures average size of clusters are larger. A cover with larger sized clusters produce higher modularity only if the community structures are more meaningful than a cover with smaller clusters.

7 Conclusion

We have convincingly shown that efficient detection of *high quality* communities can be achieved in near linear running time (in terms of the size of the graph) by combining simple traversal methods such as breadth first and depth first traversals. Our methods run faster than the state-of-the-art community detection techniques. Also, the quality of the communities generated by the proposed methods are at least as good as the state-of-the-art community detection techniques.

Acknowledgement

We would like to thank the reviewers for their effort for thoroughly reading our paper and for suggesting valuable changes. We would also like to thank Upasana Dutta for pointing out and correcting errors in some of the figures and algorithms that appear in this work.

References

M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60, 1999. ISSN 0163-5808. doi: <http://doi.acm.org/10.1145/304181.304187>. URL <http://portal.acm.org/citation.cfm?id=304187>.

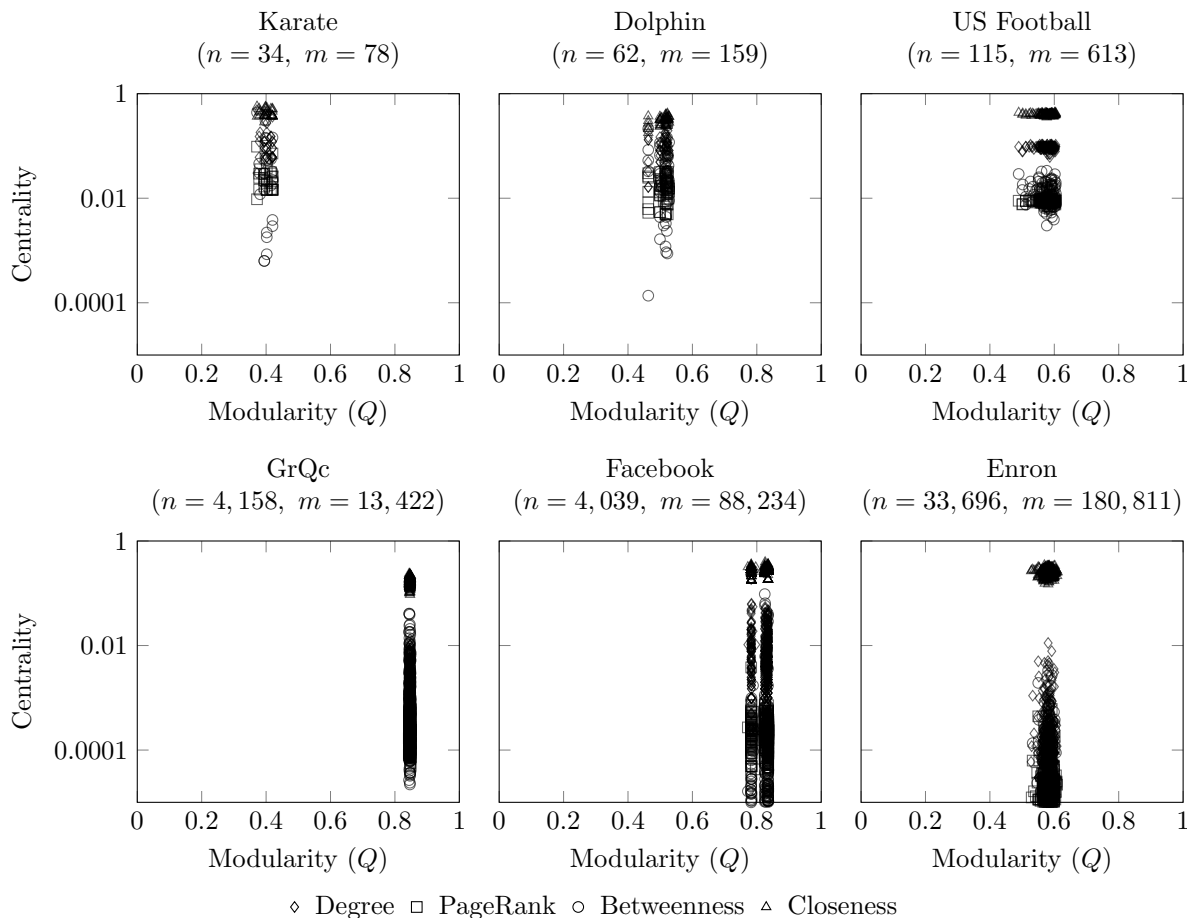


Figure 13: For the networks in the first row, every node in the network was sequentially tested as the starting node. For the networks in the bottom row, 500 nodes were randomly sampled and then their centrality data is plotted. The modularity of the final cover found using that starting node was plotted against its degree centrality, PageRank, betweenness centrality and closeness centrality.

V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008.

U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner. Maximizing modularity is hard, 2006. URL <http://arxiv.org/abs/physics/0608255>.

J. Chen, O. R. Zaiane, and R. Goebel. A visual data mining approach to find overlapping communities in networks. In Nasrullah Memon and Reda Alhajj, editors, *ASONAM*, pages 338–343. IEEE Computer Society, 2009. ISBN 978-0-7695-3689-7. URL <http://dblp.uni-trier.de/db/conf/asunam/asunam2009.html#ChenZG09a>.

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

Jean Creusefond, Thomas Largillier, and Sylvain Peyronnet. A lexdfs-based approach on finding compact

Table 7: Comparison of the reported modularity values of the final covers generated by LINCOM, and the mean modularity value of all the covers generated by using every node in the network as starting point.

Network Dataset	Mean Modularity	Standard Deviation	Reported Modularity
Karate	0.402	0.015	0.402
Dolphin	0.518	0.014	0.518
Lesmis	0.544	0.011	0.544
Football	0.581	0.017	0.582
GrQc	0.847	0.001	0.847
Enron	0.578	0.013	0.587
Facebook	0.818	0.023	0.835

communities. In *From Social Data Mining and Analysis to Prediction and Community Detection*, pages 141–177. Springer, 2017.

Wanyun Cui, Yanghua Xiao, Haixun Wang, and Wei Wang. Local search of communities in large graphs. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 991–1002. ACM, 2014.

Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659: 1–44, 2016.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

B.H. Good, Y.A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

M.S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(78):1360–1380, 1973.

S. Gregory. A fast algorithm to find overlapping communities in networks. In W. Daelemans, B. Goethals, and K. Morik, editors, *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science*, pages 408–423. Springer, 2008. ISBN 978-3-540-87478-2.

B. Klimt and Y. Yang. Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS) Proceedings*, 2004.

Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, 2008.

J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1(1), 2007. doi: 10.1145/1217299.1217301.

C. Lin, P. Ishwar, and W. Ding. Node embedding for network community discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4129–4133, March 2017. doi: 10.1109/ICASSP.2017.7952933.

D. Lusseau and M.E.J. Newman. Identifying the role that animals play in their social networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271:S477–S481, 2004.

Table 8: Comparison of the modularity values of the final covers generated by the INS-based LINCOM, by varying the value of INS threshold (r), for different benchmark datasets. Boldfaced numbers highlight the highest values of modularity obtained for a given dataset.

Threshold (r)	Karate	Dolphin	Football	LesMis	Facebook	GrQc	Enron	Amazon	DBLP
0.4	0	0.475	0	0.331	0.651	0.765	0.574	0.959	0.70
0.45	0	0.483	0	0.408	0.700	0.822	0.574	0.961	0.74
0.5	0	0.480	0	0.282	0.776	0.832	0.582	0.961	0.78
0.55	0.372	0.526	0.428	0.504	0.826	0.840	0.598	0.961	0.80
0.6	0.372	0.518	0.435	0.529	0.821	0.843	0.604	0.961	0.81
0.65	0.372	0.526	0.541	0.529	0.832	0.842	0.581	0.961	0.82
0.7	0.372	0.526	0.541	0.529	0.835	0.843	0.602	0.961	0.82
0.75	0.372	0.526	0.6	0.521	0.835	0.847	0.575	0.961	0.82
0.8	0.42	0.525	0.586	0.529	0.835	0.846	0.6	0.960	0.82
0.85	0.39	0.524	0.574	0.560	0.834	0.846	0.604	0.959	0.82

Table 9: Comparison of cover size for different types of community detection algorithms for different benchmark datasets.

Network Dataset	LINCOM		Louvain	CNM
	INS	COND		
	$ G_s $	$ G_s $	$ G_s $	$ G_s $
Karate	2	3	4	3
Dolphin	3	2	5	4
Lesmis	2	2	6	5
Football	6	5	9	7
GrQc	28	13	42	61
Enron	61	23	170	567
Epinions	671	541	733	2,983
Amazon	121	59	246	1,409
DBLP	101	228	244	3,113
Orkut	136	6	11	-

Natarajan Meghanathan. A greedy algorithm for neighborhood overlap-based community detection. *Algorithms*, 9(1):8, 2016.

T. Nepusz, A. Petroczi, L. Negyessy, and F. Bazso. Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E*, 77:016107, 2007.

M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, Jun 2004. doi: 10.1103/PhysRevE.69.066133. URL <http://link.aps.org/doi/10.1103/PhysRevE.69.066133>.

- MEJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03024, 2009.
- Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007. doi: 10.1103/PhysRevE.76.036106. URL <http://link.aps.org/doi/10.1103/PhysRevE.76.036106>.
- M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, page 1118, 2008.
- H. W. Shen, X. Q. Cheng, and J. F. Guo. Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):P07042, 2009. URL <http://stacks.iop.org/1742-5468/2009/i=07/a=P07042>.
- Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *AAAI*, pages 203–209, 2017.
- Y. Wang, G. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1039–1048, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0055-1. doi: <http://doi.acm.org/10.1145/1835804.1835935>. URL <http://doi.acm.org/10.1145/1835804.1835935>.
- J. J. Whang, D. F. Gleich, and I. S. Dhillon. Overlapping community detection using seed set expansion. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 2099–2108, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. doi: 10.1145/2505515.2505535. URL <http://doi.acm.org/10.1145/2505515.2505535>.
- Ju Xiang, Tao Hu, Yan Zhang, Ke Hu, Jian-Ming Li, Xiao-Ke Xu, Cui-Cui Liu, and Shi Chen. Local modularity for community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 443:451–459, 2016.
- J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, editors, *ICDM*, pages 745–754. IEEE Computer Society, 2012. ISBN 978-1-4673-4649-8. URL <http://dblp.uni-trier.de/db/conf/icdm/icdm2012.html#YangL12>.
- W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.
- Vincent W Zheng, Sandro Cavallari, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. From node embedding to community embedding. *arXiv preprint arXiv:1610.09950*, 2016.