

On the Analysis of a Label Propagation Algorithm for Community Detection^{*}

Kishore Kothapalli¹, Sriram V. Pemmaraju², and Vivek Sardeshmukh²

¹ International Institute of Information Technology, Hyderabad, India 500 032.
kkishore@iiit.ac.in

² Department of Computer Science, The University of Iowa, Iowa City, IA 52242-1419, USA. firstname-lastname@uiowa.edu

Abstract. This paper initiates formal analysis of a simple, distributed algorithm for community detection on networks. We analyze an algorithm that we call MAX-LPA, both in terms of its convergence time and in terms of the “quality” of the communities detected. MAX-LPA is an instance of a class of community detection algorithms called *label propagation* algorithms. As far as we know, most analysis of label propagation algorithms thus far has been empirical in nature and in this paper we seek a theoretical understanding of label propagation algorithms. In our main result, we define a clustered version of Erdős-Rényi random graphs with clusters V_1, V_2, \dots, V_k where the probability p , of an edge connecting nodes within a cluster V_i is higher than p' , the probability of an edge connecting nodes in distinct clusters. We show that even with fairly general restrictions on p and p' ($p = \Omega\left(\frac{1}{n^{1/4-\epsilon}}\right)$ for any $\epsilon > 0$, $p' = O(p^2)$, where n is the number of nodes), MAX-LPA detects the clusters V_1, V_2, \dots, V_n in just two rounds. Based on this and on empirical results, we conjecture that MAX-LPA can correctly and quickly identify communities on clustered Erdős-Rényi graphs even when the clusters are much sparser, i.e., with $p = \frac{c \log n}{n}$ for some $c > 1$.

1 Introduction

The problem of efficiently analyzing large social networks spans several areas in computer science. One of the key properties of social networks is their *community structure*. A *community* in a network is a group of nodes that are “similar” to each other and “dissimilar” from the rest of the network. There has been a lot of work recently on defining, detecting, and identifying communities in real-world networks [9, 7, 24]. It is usually, but not always, the tendency for vertices to be gathered into distinct groups, or communities, such that edges between vertices in the same community are dense but inter-community edges are sparse [20, 9]. A

^{*} This work was done when the first author (KK) was visiting The University of Iowa on an Indo-US Science and Technology Forum Fellowship. The work of the second author (SP) was partially supported by National Science Foundation grant CCF 0915543.

community detection algorithm takes as input a network and outputs a partition of the vertex set into “communities”. Detecting communities can allow us to understand attributes of vertices from the network topology alone.

There are many metrics to measure the “quality” of the communities detected by a community detection algorithm. A popular and widely adopted metric is *graph modularity* defined by Newman [21]. This measure is obtained by summing up, over all communities of a given partition, the difference between the observed fraction of links inside the community and the expected value of this quantity for a null model, that is, a random network having the same size and same degree sequence. Other popular measures include *graph conductance* [12] and *edge betweenness* [9].

The community detection problem has connections to the *graph partitioning* problem which has been well studied since 1970s [5, 13, 14, 25]. Graph partitioning problems are usually modeled as combinatorial optimization problems and this approach requires a precise sense of the objective function being optimized. Sometimes additional criteria such as the number of parts or the sizes of parts also need to be specified. In contrast, the notion of communities is relatively “fuzzy” [8] and changes from application to application. Furthermore, researchers in social network analysis are reluctant to over-specify properties of communities and would rather let algorithms “discover” communities in the given network. For a survey of the different approaches that have been proposed to find community structure in networks, see Fortunato’s work [8].

The focus of this paper is a class of seemingly simple community detection algorithms called *label propagation* algorithms (LPA). Raghavan et al. [24] seem to be the first to study label propagation algorithms for detecting network communities. The advantage of a LPA is, in addition to its simplicity, the fact that it can be easily parallelized or distributed. The generic LPA works as follows: initially each node in the network is assigned a unique label. In each iteration every node updates its label to the label which is the most frequent in its neighborhood; ties are broken randomly. One obtains variants of LPA by varying how the initial label assignment is made, how ties are broken, and whether a node includes itself in computing the most frequent label in its neighborhood. In this paper, we analyze a specific instance of LPA called MAX-LPA in which nodes are assigned initial labels uniformly at random from some large enough space. Also, if there is a tie, it is broken in favor of the larger label. Finally, a node includes its own label in determining the most frequent label in its neighborhood.

At any point during the execution of a LPA, a community is simply all nodes with the same label. The intuition behind using a LPA for community detection is that a single label (the maximum label in the case of MAX-LPA) can quickly become the most frequent label in neighborhoods within a dense cluster whereas labels have trouble “traveling” across a sparse set of edges that might connect two dense clusters. A LPA is said to have *converged* if it starts cycling through a collection of states. Ideally, we would like LPA to converge to a cycle of period one, i.e., to a state in which any further execution of LPA yields the same state. However, this is not always possible. In fact, part of the

difficulty of analyzing LPA stems from the randomized tie-breaking rule. This way of breaking ties makes it difficult to estimate the period of the cycle that the algorithm eventually converges to. The version of LPA that we analyze, namely MAX-LPA, does not suffer from this problem because Poljak and Sůra [23] have shown in a different context that MAX-LPA converges to a cycle of period 1 or 2.

Despite the simplicity of LPA, there has been very little formal analysis of either the convergence time of LPA or the quality of communities produced by it. There have been papers [24, 15, 4] that provide some empirical results about LPAs. For example, the number of iterations of label updates required for the correct convergence of LPA is around 5 [24], but it is hard to derive any fundamental conclusions about LPA’s behavior, even on specific families of networks, from these empirical results. One reason for this state of affairs is that despite its simplicity, even on simple networks, LPA can have complicated behavior, not unlike epidemic processes that model the spread of disease in a networked population [19]. Our goal in this paper is to initiate a systematic analysis of the behavior of MAX-LPA, both in terms of its convergence time and in terms of the “quality” of communities produced.

Watts and Strogatz [27] have pointed out that the classical Erdős-Rényi model of random graphs differs from real-world social, technological, and biological networks in several critical ways. Following this, a variety of other random graph models have been considered as models of real-world networks. These include the *configuration* model [17, 2], the *Watts-Strogatz* model [27], *preferential attachment* models [1], etc. (for definitions and more examples, see [18]). There is no empirical study or formal analysis of LPAs on these classes of networks. As our first step towards developing analysis techniques for LPAs we define a clustered version of Erdős-Rényi random graphs and present a formal proof of the running times of LPAs on these networks. We realize that Erdős-Rényi networks and even clustered Erdős-Rényi networks are inadequate models of real world networks, but believe that our analysis techniques could be useful in general.

The variants of LPA can naturally be viewed as *distributed* algorithms, meaning each node only has *local* knowledge, i.e., knowledge of its label and the labels of its neighbors obtained by means of message passing along edges of the networks. Distributed algorithms are generally classified as *synchronous* or *asynchronous* algorithms. (The reader is referred to standard books (e.g., [22]) for a full exposition of these terms). Here we analyze a synchronous version of MAX-LPA. The algorithm proceeds in rounds and in each round each node sends its label to all neighbors and then updates its label based on the labels received from neighbors and its own label.

1.1 Preliminaries

We use $G = (V, E)$ to denote an undirected connected graph (network) of size $n = |V|$. For $v \in V$, we denote by $N(v) = \{u : u \in V, (u, v) \in E\}$ the neighborhood of v in graph G , by $\deg(v) = |N(v)|$ the degree of v , and by $\Delta(G) = \max_{v \in V} \deg(v)$ the maximum degree over all the vertices in G . A k -hop

Algorithm 1 MAX-LPA on a node v

```
 $i = 0$   
 $l_v[i] \leftarrow \text{random}(0,1)$   
while true do  
   $i++$ ;  
  send  $l_v[i-1]$  to  $\forall u \in N(v)$   
  receive  $l_u[i-1]$  from  $\forall u \in N(v)$   
   $l_v[i] \leftarrow \max \left\{ \ell \mid \sum_{u \in N'(v)} [\ell_u[i-1] == \ell] \geq \sum_{u \in N'(v)} [\ell_u[i-1] == \ell'] \text{ for all } \ell' \right\}$   
end while
```

neighborhood ($k \geq 1$) of v is defined as $N_k(v) = \{w : \text{dist}_G(w, v) \leq k\} \setminus \{v\}$. We denote the *closed neighborhood* (respectively, *closed k -hop neighborhood*) of v as $N'(v) = N(v) \cup \{v\}$ (respectively, $N'_k(v) = N_k(v) \cup \{v\}$).

Denote by $\ell_u(t)$ the label of node u just before round t . When the round number is clear from the context, we use ℓ_u to denote the current label of u . Since the number of labels in the network is finite, LPA will behave periodically starting in some round t^* , i.e., for some $p \geq 1$, $0 \leq i < p$, and $j = 0, 1, 2, \dots$,

$$\ell_u(t^* + i) = \ell_u(t^* + i + j \cdot p)$$

for all $u \in V$. Then we say that MAX-LPA has *converged* in t^* rounds.

We now describe MAX-LPA precisely (see **Algorithm 1**). Every node $v \in V$ is assigned a unique label uniformly and independently at random. For concreteness, we assume that these labels come from the range $[0, 1]$. At the start of a round, each node sends its label to all neighboring nodes. After receiving labels from all neighbors, a node v updates its label as:

$$l_v \leftarrow \max \left\{ \ell \mid \sum_{u \in N'(v)} [\ell_u == \ell] \geq \sum_{u \in N'(v)} [\ell_u == \ell'] \text{ for all } \ell' \right\}, \quad (1)$$

where $[\ell_u == \ell]$ evaluates to 1 if $\ell_u = \ell$, otherwise evaluates to 0. Note that there is no randomness in the algorithm after the initial assignments of labels.

By “w.h.p.” (with high probability) we mean with probability at least $1 - \frac{1}{n^c}$ for some constant $c \geq 1$. In this paper we repeatedly use the following versions of a tail bound on the probability distribution of a random variable, due to Chernoff and Hoeffding [3, 11]. Let X_1, X_2, \dots, X_m be independent and identically distributed binary random variables. Let $X = \sum_{i=1}^m X_i$. Then, for any $0 \leq \epsilon \leq 1$ and $c \geq 1$,

$$\Pr[X > (1 + \epsilon) \cdot E[X]] \leq \exp\left(-\frac{\epsilon^2 E[X]}{3}\right) \quad (2)$$

$$\Pr[X < (1 - \epsilon) \cdot E[X]] \leq \exp\left(-\frac{\epsilon^2 E[X]}{2}\right) \quad (3)$$

$$\Pr[|X - E[X]| > \sqrt{3c \cdot E[X] \cdot \log n}] \leq \frac{1}{n^c} \quad (4)$$

1.2 Results

As mentioned earlier, the purpose of this paper is to counterbalance the predominantly empirical line of research on LPA and initiate a systematic analysis of MAX-LPA. Our main results can be summarized as follows:

- As a “warm-up” we prove (Section 2) that when executed on an n -node path MAX-LPA converges to a cycle of period one in $\Theta(\log n)$ rounds w.h.p. Moreover, we show that w.h.p. the state that MAX-LPA converges to has $\Omega(n)$ communities.
- In our main result (Section 3), we define a class of random graphs that we call *clustered Erdős-Rényi graphs*. A clustered Erdős-Rényi graph $G = (V, E)$ comes with a node partition $\Pi = (V_1, V_2, \dots, V_k)$ and pairs of nodes in each V_i are connected with probability p_i and pairs of nodes in distinct parts in Π are connected with probability $p' < \min_i \{p_i\}$. Since p' is small relative to any of the p_i 's, one might view a clustered Erdős-Rényi graph as having a natural community structure given by Π . We prove that even with fairly general restrictions on the p_i 's and p' and on the sizes of the V_i 's, MAX-LPA converges to a period-1 cycle in just 2 rounds, w.h.p. and “correctly” identifies Π as the community structure of G .
- Roughly speaking, the above result requires each p_i to be $\Omega\left(\left(\frac{\log n}{n}\right)^{1/4}\right)$.

We believe that MAX-LPA would correctly and quickly identify Π as the community structure of a given clustered Erdős-Rényi graph even when the p_i 's are much smaller, e.g. even when $p_i = \frac{c \log n}{n}$ for $c > 1$. However, at this point our analysis techniques do not seem adequate for situations with smaller p_i values and so we provide empirical evidence (Section 4) for our conjecture that MAX-LPA correctly converges to Π in $O(\text{polylog}(n))$ rounds even when $p_i = \frac{c \log n}{n}$ for some $c > 1$ and p' is just a logarithmic factor smaller than p_i .

1.3 Related Work

There are several variants of LPA presented in the literature [4, 10, 26, 16]. Most of these are concerned about “quality” of the output and present empirical studies of output produced by LPA.

Raghavan et al. [24], based on the experiments, claimed that at least 95% of the nodes are classified correctly by the end of 5 rounds of label updates. But the experiments that they carried out were on the small networks.

Cordasco and Gargano [4] proposed a semi-synchronous approach which is guaranteed to converge without oscillations and can be parallelized. They provided a formal proof of convergence but did bound the running time of the algorithm. Lui and Murata [16] presented a variation of LPA for bipartite networks which converges but no formal proof is provided, neither for the convergence nor for the running time.

Leung et al. [15] presented empirical analysis of quality of output produced by LPA on larger data sets. From experimental results on a special structured network they claimed that running time of LPA is $O(\log n)$.

2 Analysis of MAX-LPA on a Path

Consider a path \mathcal{P}_n consisting of vertices $V = [n]$ and edge set $E = \{(i, i+1) \mid 1 \leq i < n\}$. In this section, we analyze the execution of MAX-LPA on a path network \mathcal{P}_n and prove that in $O(\log n)$ rounds MAX-LPA converges to a state from which no further label updates occur and furthermore in such a state the number of communities is $\Omega(n)$ w.h.p..

Lemma 1. *When MAX-LPA is executed on path network \mathcal{P}_n , independent of the initial label assignment, it will converge to a state from which no further label updates occur.*

Proof. First we show that at any point in the execution of MAX-LPA, the subgraph of \mathcal{P}_n induced by all nodes with the same label, is a single connected component. This is true before the first round since the initial label assignment assigns distinct labels to the nodes. Suppose the claim is true just before round t . Let $S = (i, i+1, \dots, j)$ be the subgraph of \mathcal{P}_n consisting of nodes with label ℓ , just before round t of MAX-LPA.

- If S contains two or more nodes then none of the nodes in S will ever change their label. Moreover, the only other nodes that can acquire label ℓ in round t are nodes $i-1$ and $j+1$. Hence, after round t , the set of nodes with label ℓ still induces a single connected component.
- If S contains a single node, say i , then the only way in which label ℓ might induce multiple connected components after round t would be if in round t : (i) node $i-1$ acquires label ℓ , (ii) node $i+1$ acquires label ℓ , and (iii) node i changes its label to some $\ell' \neq \ell$. (i) and (ii) above can only happen if ℓ is larger than the labels of nodes $i-1$ and $i+1$ just before round t . But, if this were true, then node i would not change its label in round t .

Hence, in either case the nodes with label ℓ would induce a connected component.

According to Poljak and Sůra [23], MAX-LPA has a period of 1 or 2 on any network with any initial label assignment. To obtain a contradiction we suppose that MAX-LPA has a period of 2 when executed on \mathcal{P}_n for some n and some initial label assignment. Therefore for some $v \in V$ and some time t , $\ell_v(t+2i) = \ell$ and $\ell_v(t+2i+1) = \ell'$ for $\ell \neq \ell'$ and all $i = 0, 1, 2, \dots$. For v to change its label from ℓ to ℓ' in a round it must be the case that $\ell < \ell'$. This is because v cannot have two neighbors with label ℓ' since ℓ' can only induce one connected component. Hence, v acquires the new label ℓ' by tie breaking. By a symmetric argument, for v to change its label from ℓ' to ℓ in the next round, it must be the case that $\ell' < \ell$. Both conditions cannot be met and we have a contradiction. \square

Definition 1. *A node v is said to be k -hop maxima if its label ℓ_v is (strictly) greater than the labels of all nodes in its k -neighborhood. As a short form, we will use local maxima to refer to any node that is a 1-hop maxima.*

Let $M = \{i_1, i_2, \dots, i_r\}$, $i_1 < i_2 < \dots < i_r$ be the set of nodes which are 2-hop maxima in \mathcal{P}_n for the given initial label assignment. For any $1 \leq j < r$, nodes i_j and i_{j+1} are said to be *consecutive* 2-hop maxima.

Lemma 2. *When MAX-LPA converges, the number of communities it identifies is bounded below by the number of 2-hop maxima in the initial label assignment.*

Proof. Since all initial node labels are assumed to be distinct, in the first round of MAX-LPA every node $u \in V$ acquires a label by breaking ties. Since ties are broken in favor of larger labels, all neighbors of each $i_j \in M$ will acquire the corresponding 2-hop maxima label ℓ_{i_j} . Thus after one round of MAX-LPA, for each $i_j \in M$, there are three consecutive nodes in \mathcal{P}_n with label ℓ_{i_j} . None of these nodes will change their label in future rounds and hence there will be a community induced by label ℓ_{i_j} when MAX-LPA converges. \square

Lemma 3. *Let D be the maximum distance in \mathcal{P}_n between a pair of consecutive nodes in M . Then the number of rounds that MAX-LPA takes to converge is bounded above by $D + 2$.*

Proof. Call a node v *isolated* if its label is distinct from the labels of its neighbors. After the first round of MAX-LPA each node $i_j \in M$ and its neighbors acquire label ℓ_{i_j} . Therefore, after the first round, every connected component of the graph induced by isolated nodes has size bounded above by D . We now show that in each subsequent round, the size of every connected component of size two or more will reduce by at least one. Let S be a component in the graph induced by isolated nodes, just before round t . Let i be the node with maximum label in S . Since S contains at least two nodes, without loss of generality suppose that $i + 1$ is also in S . In round t , node i could acquire the label of a node outside S . If this happens i would cease to be isolated after round t . Similarly, in round t , node $i + 1$ could acquire the label of a node outside S and would therefore cease to be isolated after round t . If neither of these happens in round t , then node $i + 1$ will acquire the label of node i in round t and node i will not change its label. In this case, both i and $i + 1$ will cease to be isolated nodes after round t . In any case, we see that the size of the component S has shrunk by at least one in round t . Thus in $D + 1$ rounds \mathcal{P}_n we will reach a state in which all components in the graph induced by isolated nodes have size one. Isolated nodes whose labels are larger than the labels of neighbors will make no further label updates. The remaining isolated nodes will disappear in one more round. \square

Theorem 1. *When MAX-LPA is executed on a path \mathcal{P}_n , it converges to a state from which no further label updates occur in $O(\log n)$ rounds w.h.p. Furthermore, in such a state there are $\Omega(n)$ communities.*

Proof. Partition \mathcal{P}_n into “segments” of 5 nodes each. Let S denote the set of center nodes of these segments. The probability that a node in \mathcal{P}_n is a 2-hop maxima is $\frac{1}{5}$. Therefore the expected number of nodes in S that end up being 2-hop maxima is $n/25$. Now note that for any two nodes $i, j \in S$, node i being a 2-hop maxima is independent of node j being a 2-hop maxima due to the fact that there are at least 4 nodes between i and j . Therefore, we can apply the lower tail Chernoff bound (3) to conclude that w.h.p. at least $n/50$ nodes in \mathcal{P}_n are 2-hop maxima. Combining this with Lemma 2 tell us that when MAX-LPA

converges, it does so to a state in which there are at least $n/50$ communities with high probability.

Now consider a contiguous sequence of k 5-node segments. The probability that none of the centers of the k segments are 2-hop maxima is $(4/5)^k$. Note that here we use the independence of different segment centers becoming 2-hop maxima. Hence, for a large enough constant c , the probability that none of the centers of $k = c \log n$ consecutive segments are 2-hop maxima is at most $1/n^2$. Using the union bound and observing that there at most n consecutive segment sequences of length k , we see that the probability that there is a sequence of $k = c \log n$ consecutive segments, none of whose centers are 2-hop maxima, is at most $1/n$. Therefore, with probability at least $1 - 1/n$ every sequence of $k = c \log n$ consecutive segments contains a segment whose center is a 2-hop maxima. It follows that the distance between consecutive 2-hop maxima is at most $5c \log n$ with probability at least $1 - 1/n$. The result follows by combining this with Lemma 3. \square

The argument given here establishing a linear lower bound on the number of communities can be easily extended to graphs with maximum degree bounded by a constant. The argument bounding the convergence time depended crucially on two properties of the underlying graph: (i) degrees being bounded and (ii) number of paths of length $O(\log n)$ being polynomial in number. Thus the convergence bound can be extended to other graph classes satisfying these two properties (e.g., trees with bounded degree).

3 Analysis of MAX-LPA on Clustered Erdős-Rényi Graphs

We start this section by introducing a family of “clustered” random graphs that come equipped with a simple and natural notion of a community structure. We then show that on these graphs MAX-LPA detects this natural community structure in only 2 rounds, w.h.p. provided certain fairly general sparsity conditions are satisfied.

3.1 Clustered Erdős-Rényi graphs

Recall that for an integer $n \geq 1$ and $0 \leq p \leq 1$, the Erdős-Rényi graph $G(n, p)$ is the random graph obtained by starting with vertex set $V = \{1, 2, \dots, n\}$ and connecting each pair of vertices $u, v \in V$, independently with probability p . Let Π denote a partition (V_1, V_2, \dots, V_k) of V , let π denote the real number sequence (p_1, p_2, \dots, p_k) , where $0 \leq p_i \leq 1$ for all i and let $0 \leq p' < \min_i \{p_i\}$. The *clustered Erdős-Rényi* graph $G(\Pi, \pi, p')$ has vertex set V and edges obtained by independently connecting each pair of vertices $u, v \in V$ with probability p_i if $u, v \in V_i$ for some i and with probability p' , otherwise (see Figure 1). Thus each induced subgraph $G[V_i]$ is the standard Erdős-Rényi graph $G(n_i, p_i)$, where $n_i = |V_i|$.

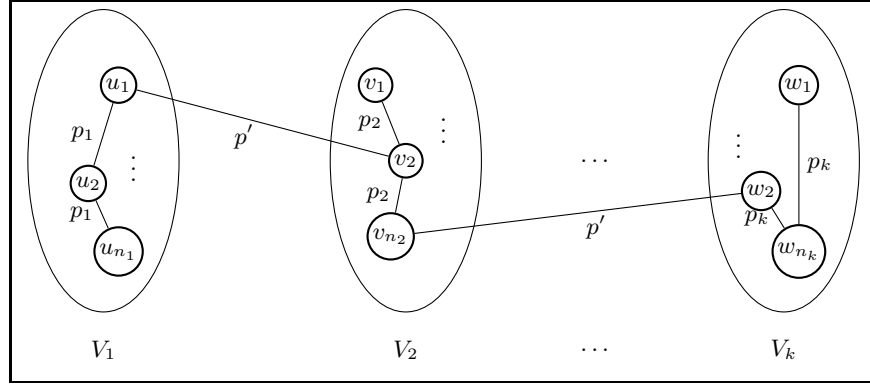


Fig. 1: The clustered Erdős-Rényi graph. We connect two nodes in the i -th ellipse (i.e., V_i) with probability p_i and nodes from different ellipses are connected with probability $p' < \min_i \{p_i\}$.

Given that $p' < p_i$ for all i , one might view $G(\Pi, \pi, p')$ as having a natural community structure given by the vertex partition Π . Specifically, when p' is much smaller than $\min_i \{p_i\}$, the inter-community edge density is much less than the intra-community edge density and it may be easier to detect the community structure Π . On the other hand as the intra-community probabilities p_i get closer to p' , it may be hard for an algorithm such as MAX-LPA to identify Π as the community structure. Similarly, if an intra-community probability p_i becomes very small, then the subgraph $G[V_i]$ can itself be quite sparse and independent of how small p' is relative to p_i , any community detection algorithm may end up viewing each V_i as being composed of several communities.

In the rest of the section, we explore values of the p_i 's and p' for which MAX-LPA “correctly” and quickly identifies Π as the community structure of $G(\Pi, \pi, p')$.

3.2 Analysis

In the following theorem we establish fairly general conditions on the probabilities $\{p_i\}$ and p' and on the node subset sizes $\{n_i\}$ and n under which MAX-LPA converges correctly, i.e., to the node partition Π , w.h.p. Furthermore, we show that under these circumstances just 2 rounds suffice for MAX-LPA to reach convergence!

Lemma 4. *Let $G(\Pi, \pi, p')$ be a clustered Erdős-Rényi graph such that $p' < \min_i \{\frac{p_i}{n}\}$. Let ℓ_i be the maximum label of a node in V_i . Then for any node $v \in V_i$ the probability that v is not adjacent to a node outside V_i with label higher than ℓ_i is at least $1/2e$.*

Proof. Let v' be a node in $V \setminus V_i$. Given that $|V_i| = n_i$ and ℓ_i is the maximum label among these n_i nodes, the probability that the label assigned uniformly at

random to v' is larger than ℓ_i is $1/(n_i + 1)$. The probability that v has an edge to v' and v' has a higher label than ℓ_i is $p'/(n_i + 1)$. Therefore the probability that v' has no edge to a node outside V_i with label larger than ℓ_i is

$$\left(1 - \frac{p'}{n_i + 1}\right)^{n - n_i}.$$

We bound this expression below as follows:

$$\left(1 - \frac{p'}{n_i + 1}\right)^{n - n_i} > \left(1 - \frac{p'}{n_i}\right)^n > \left(1 - \frac{1}{n}\right)^n > \frac{1}{2e}.$$

□

Theorem 2. Let $G(\Pi, \pi, p')$ be a clustered Erdős-Rényi graph. Suppose that the probabilities $\{p_i\}$ and p' and the node subset sizes $\{n_i\}$ and n satisfy the inequalities:

$$(i) \ n_i p_i^2 > 8np' \quad \text{and} \quad (ii) \ n_i p_i^4 > 1800c \log n,$$

for some constant c . Then, given input $G(\Pi, \pi, p')$, MAX-LPA converges correctly to node partition Π in two rounds w.h.p. (Note that condition (ii) implies for each i , $p_i > \frac{\log n_i}{n_i}$ and hence each $G[V_i]$ is connected.)

Proof. Let $V_i = \{u_1, u_2, \dots, u_{n_i}\}$ and without loss of generality assume that $\ell_{u_1} > \ell_{u_2} > \dots > \ell_{u_{n_i}}$. Since all initial node labels are assumed to be distinct, in the first round of MAX-LPA every node $u \in V$ acquires a label by breaking ties. Since ties are broken in favor of larger labels, all neighbors of u_1 in V_i that have no neighbor outside V_i with a label larger than ℓ_{u_1} will acquire the label ℓ_{u_1} . Consider a node $v \in V_i$. Let β denote the probability that v has no neighbor outside V_i with label larger than ℓ_{u_1} . Note that inequality (i) in the theorem statement implies the hypothesis of Lemma 4 and therefore $\beta > 1/2e$. The probability that v is a neighbor of u_1 and does not have a neighbor outside V_i is $\beta \cdot p_i$. Hence, after the first round of MAX-LPA, in expectation, $n_i \cdot \beta \cdot p_i$ nodes in V_i would have acquired the label ℓ_{u_1} . In the rest of the proof we will use

$$X := n_i \cdot \beta \cdot p_i.$$

Now consider node u_j for $j > 1$. For a node $v \in V_i$ to acquire the label ℓ_{u_j} it must be the case that v is adjacent to u_j , not adjacent to any node in $\{u_1, u_2, \dots, u_{j-1}\}$, and not adjacent to any node outside V_i with a label higher than ℓ_{u_j} . Since ℓ_{u_j} is smaller than ℓ_{u_1} , the probability that v is not adjacent to a node outside V_i with label higher than ℓ_{u_j} is less than β . Thus the probability that a node in V_i acquires the label ℓ_{u_j} is at most $p_i(1 - p_i)^{j-1} \cdot \beta < p_i(1 - p_i) \cdot \beta$. Furthermore, the probability that a node outside V_i will acquire the label ℓ_{u_j} at the end of the first round is at most p' . Therefore, the expected number of nodes in V that acquire the label u_j , at the end of the first round, is in expectation at most $n_i \cdot p_i(1 - p_i) \cdot \beta + (n - n_i)p'$. We now use inequality (i) and the fact that $2\beta e > 1$ to upper bound this expression as follows:

$$n_i \cdot p_i(1 - p_i) \cdot \beta + (n - n_i)p' < n_i \cdot p_i(1 - p_i) \cdot \beta + \frac{2\beta e \cdot n_i p_i^2}{8} < n_i \cdot p_i \left(1 - \frac{3p_i}{4}\right) \cdot \beta.$$

Therefore, the expected number of nodes in V that acquire the label u_j , at the end of the first round, is in expectation at most

$$Y := n_i \cdot p_i \left(1 - \frac{3p_i}{4}\right) \cdot \beta.$$

It is worth mentioning at this point that $X - Y = n_i p_i^2 \beta / 4$.

Note that all the random variables we have utilized thus far, e.g., the number of nodes adjacent to u_1 and not adjacent to any node outside V_i with label higher than ℓ_{u_1} , can be expressed as sums of independent, identically distributed indicator random variables. Hence we can bound the deviation of such random variables using the tail bound in (4). In particular, let Y' denote $Y + \sqrt{3cY \log n}$ and X' denote $X - \sqrt{3cX \log n}$. With high probability, at the end of the first round of MAX-LPA, the number of nodes in V_i that acquire the label u_1 is at least X' and the number of nodes in V that acquire the label ℓ_{u_j} , $j > 1$, is at most Y' . Next we bound the “gap” between X' and Y' as follows:

$$\begin{aligned} X' - Y' &= X - Y - \sqrt{3cX \log n} - \sqrt{3cY \log n} \\ &> \frac{3n_i p_i^2 \beta}{4} - 2\sqrt{3cX \log n} \\ &> \frac{3n_i p_i^2 \beta}{4} - 2\sqrt{3cn_i p_i \beta \log n} \\ &> \frac{3n_i p_i^2 \beta}{4} - \frac{3n_i p_i^2 \beta}{5} \\ &= \frac{3n_i p_i^2 \beta}{20} \end{aligned}$$

The second inequality follows from $X - Y = 3n_i p_i^2 \beta / 4$ and $Y < X$, the third from the fact that $X = n_i p_i \beta$, and the fourth by using inequality (ii) from the theorem statement.

We now condition the execution of the second round of MAX-LPA on the occurrence of the two high probability events: (i) number of nodes in V_i that acquire the label u_1 is at least X' and (ii) the number of nodes in V that acquire the label ℓ_{u_j} , $j > 1$, is at most Y' . Consider a node $v \in V_i$ just before the execution of the second round of MAX-LPA. Node v has in expectation at least $p_i X'$ neighbors labeled ℓ_{u_1} in V_i . Also, node v has in expectation at most $p_i Y'$ neighbors labeled ℓ_{u_j} , for each $j > 1$, in V . Let us now use X'' to denote the quantity $p_i X' - \sqrt{3cp_i X' \log n}$ and Y'' to denote the quantity $p_i Y' + \sqrt{3cp_i Y' \log n}$. By using (4) again, we know that w.h.p. v has at least X'' neighbors with label ℓ_{u_1} and at most Y'' neighbors with a label ℓ_{u_j} , $j > 1$. We will now show that $X'' > Y''$ and this will guarantee that in the second round of MAX-LPA v will acquire the label ℓ_{u_1} , with high probability. Since v is an arbitrary node in V_i , this implies that all nodes in V_i will acquire the label ℓ_{u_1} .

in the second round of MAX-LPA w.h.p.

$$\begin{aligned}
X'' - Y'' &= p_i(X' - Y') - \sqrt{3cp_iX'\log n} - \sqrt{3cp_iY'\log n} \\
&> \frac{3n_i p_i^3}{20} - 2\sqrt{3cp_iX'\log n} \\
&> \frac{3n_i p_i^3}{20} - 2\sqrt{3cn_i p_i^2 \beta \log n} \\
&> \frac{3n_i p_i^3}{20} - \frac{n_i p_i^3 \beta}{10} \\
&= \frac{3n_i p_i^2}{20} \\
&> 0
\end{aligned}$$

The second inequality follows from the bound on $X' - Y'$ derived earlier and $Y' < X'$, the third from the fact that $X' < n_i p_i \beta$, and the fourth by using inequality (ii) from the theorem statement.

Thus at the end of the second round of MAX-LPA, w.h.p., every node in V_i has label ℓ_{u_1} . This is of course true, w.h.p., for all of the V_i 's. Now note that every node $v \in V_i$ has, in expectation $n_i p_i$ neighbors in V_i and fewer than np' neighbors outside V_i . Inequality (i) implies that $np' < n_i p_i / 8$ and inequality (ii) implies that $n_i p_i = \Omega(\log n)$. Pick a constant $\epsilon > 0$ such that $n_i p_i (1 + \epsilon) / 8 < n_i p_i (1 - \epsilon)$. By applying tail bound (2), we see that w.h.p. v has more than $n_i p_i (1 - \epsilon)$ neighbors in V_i and fewer than $n_i p_i (1 + \epsilon) / 8$ neighbors outside V_i . Hence, w.h.p. v has no reason to change its label. Since v is an arbitrary node in an arbitrary V_i , w.h.p. there are no further changes to the labels assigned by MAX-LPA. \square

To understand the implications of Theorem 2 consider the following example. Suppose that the clustered Erdős-Rényi graph has $O(1)$ clusters and each cluster had size $\Theta(n)$. In such a setting, inequality (ii) from the theorem simplifies to requiring that each $p_i = \Omega((\log n / n)^{1/4})$ and inequality (i) simplifies to $p' < p_i^2 / c$ for all i . This tells us, for instance, that MAX-LPA converges in just two rounds on a clustered Erdős-Rényi graph in which each cluster has $\Theta(n)$ vertices and an intra-community probability of $\Theta(1/n^{1/3})$ and the inter-community probability is $\Theta(1/n^{2/3})$.

This example raises several questions. If we were willing to allow more time for MAX-LPA to converge, say $O(\log n)$ rounds, could we significantly weaken the requirements on the p_i 's and p' . Specifically, could we permit an intra-community probability p_i to become as small as $c \log n / n$ for some constant $c > 1$? Similarly, could we permit the inter-community probability p' to come much closer to the smallest p_i , say within a constant factor.

We believe that it may be possible to obtain such results, but only via substantively different analysis techniques.

4 Empirical Results on Sparse Erdős-Rényi Graphs

In the previous section we proved that if the clusters (each V_i) in a clustered Erdős-Rényi graphs were dense enough and the inter-cluster edge density (fraction of edges between nodes in different V_i) was relatively low, then MAX-LPA would correctly converge in just 2 rounds. Specifically, our result requires each cluster to be Erdős-Rényi random graph $G(n, p)$ with $p = O\left(\left(\frac{\log n}{n}\right)^{1/4}\right)$. In this section we ask: how does MAX-LPA behave if individual clusters are much sparser? For example, how does MAX-LPA behave on $G(n, p)$ with much smaller p , say $p = \frac{c \cdot \log n}{n}$ for some $c > 1$. The proof technique used in the previous section does not extend to such small values of p . However, we believe that MAX-LPA converges quickly and correctly even on clustered Erdős-Rényi graphs whose clusters are of the type $G(n, p)$ for $p = \frac{c \cdot \log n}{n}$ for $c > 1$. In this section, we ask (and empirically answer) two questions:

1. Can one expect there to be a constant c such that MAX-LPA, when run on $G(n, p)$ with $p \geq \frac{c \cdot \log n}{n}$ will, with high probability, terminate with one community. If the answer to Question 1 is “yes” what might the running time of MAX-LPA, as a function of n be for appropriate values of p .
2. Consider a clustered Erdős-Rényi graph with two parts V_1 and V_2 of equal size, and each $p_i = \frac{c \cdot \log n}{n}$ for some $c > 1$. Let $p' = \frac{c'}{n}$ for some c' . Are there constants c, c' for which MAX-LPA will quickly converge and correctly identify (V_1, V_2) as the community structure?

We are interested in values of p of the form $\frac{c \cdot \log n}{n}$ because $\frac{\log n}{n}$ is the threshold for connectivity in Erdős-Rényi graphs [6].

4.1 Simulation Setup

We implemented MAX-LPA in a C program and executed on a Linux machine (with 2.4 GHz Intel(R) Core(TM)2 processor). We examined the number of rounds it takes and also number of communities it declares at the end of the execution. We executed MAX-LPA on $G(n, p)$ and on $G(\Pi, \pi, p')$ with $\Pi = (V_1, V_2)$, $|V_1| = |V_2| = n/2$, $\pi = (p, p)$, $p' = 0.6/n$ for various values of n and p . For each n, p combination we ran MAX-LPA 50 times. We used p values of the form $\frac{c \cdot \log n}{n}$ for various values of $c \geq 1$.

4.2 Results

We executed MAX-LPA using the setup discussed above. Table 1 shows the number of simulations out of 50 simulations per n and c values for which it ended up in a single community for each pair of n and c . If the input graph is disconnected then obviously there will be multiple communities. Therefore, we also noted number of simulations for which the graph was connected and this number is shown in the brackets.

Table 1: This table shows simulations on Erdős-Rényi graphs $G(n, p)$ where $p = \frac{c \log n}{n}$. Each entry in the table shows the number of simulations out of 50 simulations per n and c values in which a single community is declared by MAX-LPA and number of simulations in which the graph $G(n, p)$ was connected is shown in brackets.

n	$c = 1$	$c = 1.2$	$c = 1.5$	$c = 1.7$
1000	44 (50)	47 (47)	50 (50)	50 (50)
2000	42 (46)	47 (50)	47 (50)	50 (50)
4000	45 (47)	49 (50)	50 (50)	50 (50)
8000	47 (48)	50 (50)	50 (50)	50 (50)
16000	49 (50)	50 (50)	50 (50)	50 (50)
32000	49 (50)	50 (50)	50 (50)	50 (50)
64000	50 (50)	50 (50)	50 (50)	50 (50)
128000	50 (50)	50 (50)	50 (50)	50 (50)

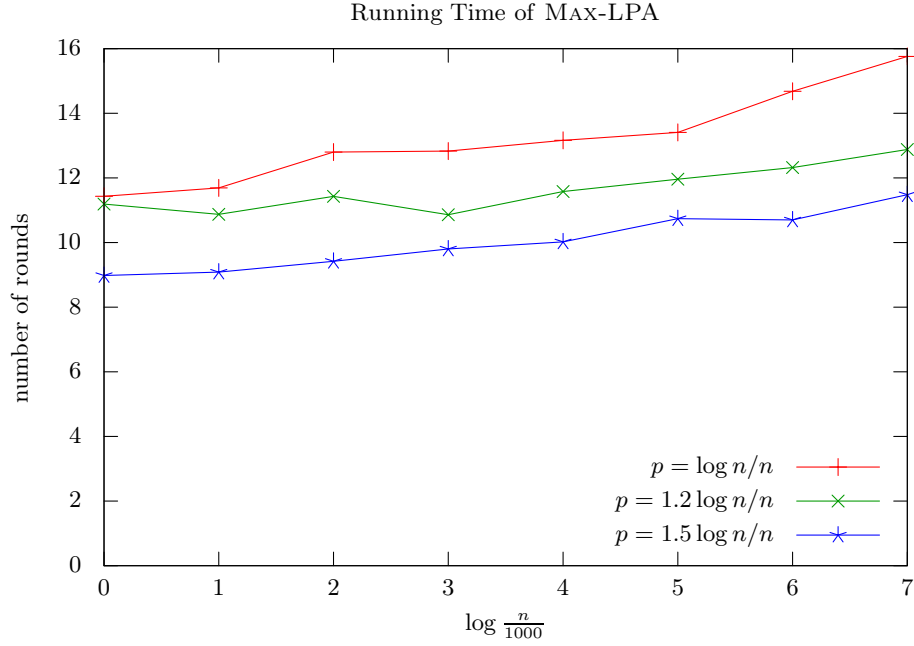


Fig. 2: Number of rounds for MAX-LPA when executed on sparse Erdős-Rényi (averaged over simulations where it ended with a single community out of 50 simulations per n and p).

It is well known that $p = \frac{\log n}{n}$ is a threshold for connectivity in Erdős-Rényi graphs and therefore we are getting few runs for $c = 1$ where the input graph was disconnected. From Table 1, we can say that MAX-LPA when executed on

Erdős-Rényi graphs with $p = \frac{c \log n}{n}$ and $c > 1$, with high probability, terminate with one community. It also seem to be the case that as c increases, we are getting more single community runs. This is because as c increases, the graph become more dense.

Figure 2 shows a plot of the number of rounds MAX-LPA takes to converge on $G(n, p)$ as n increases, averaged over all simulations which resulted in a single community at the end of the execution. The running time seems to grow in a linear fashion with logarithm of graph size. Also as c increases the running time decreases, which implies that as the graph becomes more dense MAX-LPA converges more quickly to a single community. Our results lead us to conjecture that when MAX-LPA is executed on Erdős-Rényi graphs $G(n, p)$ with $p = O(\frac{\log n}{n})$ it will, with high probability, terminate with a single community in $O(\log n)$ rounds.

Table 2 shows the number of simulations out of 50 simulations per n and c values for which MAX-LPA correctly identified the partition Π when executed on $G(\Pi, \pi, p')$ for $p' = \frac{0.6}{n}$. From previous results in Table 1, for $c = 1.5$ MAX-LPA declared a single community when executed on $G(n, p)$ w.h.p. Therefore in this experiments we started with $c = 1.5$. But for $c = 1.5$, the influence from the nodes from other partition is significant. As c increases this influence is not significant compared to the influence from nodes within the same partition.

Table 2: This table shows simulations of MAX-LPA on $G(\Pi, \pi, p')$ with $\Pi = (V_1, V_2)$, $|V_1| = |V_2| = n/2$, $\pi = (p, p)$, where $p = \frac{c \log n}{n}$ and $p' = \frac{0.6}{n}$. Each entry in the table shows, for particular n and c values, the number of simulations out of 50 in which MAX-LPA identified two communities V_1 and V_2 . The number of simulations in which graph was connected is shown in brackets.

n	$c = 1.5$	$c = 2$	$c = 4$
1000	22 (45)	39 (50)	50 (50)
2000	21 (39)	40 (50)	50 (50)
4000	22 (36)	47 (50)	50 (50)
8000	14 (38)	47 (50)	50 (50)
16000	26 (35)	49 (49)	50 (50)
32000	17 (33)	49 (49)	50 (50)
64000	26 (34)	46 (50)	50 (50)
128000	5 (35)	47 (47)	50 (50)

5 Future Work

We believe that with some refinements, the analysis technique used to show $O(\log n)$ -rounds convergence of MAX-LPA on paths, can be used to show poly-logarithmic convergence on sparse graphs in general, e.g., those with degree bounded by a constant. This is one direction we would like to take our work in.

At this point the techniques used in Section 3 do not seem applicable to more sparse clustered Erdős-Rényi graphs. But if we were willing to allow more time for MAX-LPA to converge, say $O(\log n)$ rounds, could we significantly weaken the requirements on the p_i 's and p' ? Specifically, could we permit an intra-community probability p_i to become as small as $c \log n/n$ for some constant $c > 1$? Similarly, could we permit the inter-community probability p' to come much closer to the smallest p_i , say within a constant factor? This is another direction for our research.

Acknowledgments. We would like to thank James Hegeman for helpful discussions and for some insightful comments.

References

- [1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.
- [2] Edward A Bender and E.Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296 – 307, 1978.
- [3] Herman Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):pp. 493–507, 1952.
- [4] G. Cordasco and L. Gargano. Community detection via semi-synchronous label propagation algorithms. In *Business Applications of Social Network Analysis (BASNA), 2010 IEEE International Workshop on*, pages 1–8. IEEE, 2010.
- [5] Ulrich Elsner. Graph Partitioning - A Survey, 1997.
- [6] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5(17), 1960.
- [7] G.W. Flake, S. Lawrence, and C.L. Giles. Efficient identification of web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160. ACM, 2000.
- [8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.
- [9] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.
- [10] S. Gregory. Finding overlapping communities using disjoint community detection algorithms. *Complex Networks*, pages 47–61, 2009.
- [11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.
- [12] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.
- [13] David R. Karger. Minimum cuts in near-linear time. *J. ACM*, 47(1):46–76, 2000.
- [14] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.
- [15] I.X.Y. Leung, P. Hui, P. Lio, and J. Crowcroft. Towards real-time community detection in large networks. *Arxiv preprint arXiv:0808.2633*, 2008.

- [16] X. Liu and T. Murata. How does label propagation algorithm work in bipartite networks? In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 5–8. IEEE Computer Society, 2009.
- [17] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.
- [18] M. Newman. *Networks: An Introduction*. OUP Oxford, 2010.
- [19] M. E. J. Newman. The spread of epidemic disease on networks. *Physical Review Letters*, 66:016128, 2002.
- [20] M.E.J. Newman. The Structure and Function of Complex Networks. *SIAM review*, 45(2):167–256, 2003.
- [21] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [22] D. Peleg. *Distributed computing: a locality-sensitive approach*, volume 5. Society for Industrial Mathematics, 2000.
- [23] S. Poljak and M. Sura. On periodical behaviour in societies with symmetric influences. *Combinatorica*, 3(1):119–121, 1983.
- [24] U.N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.
- [25] P.R. Suaris and G. Kedem. An algorithm for quadrissection and its application to standard cell placement. *Circuits and Systems, IEEE Transactions on*, 35(3):294–303, mar 1988.
- [26] L. Šubelj and M. Bajec. Unfolding network communities by combining defensive and offensive label propagation. In *Proceedings of the ECML PKDD Workshop on the Analysis of Complex Networks 2010 (ACNE '10)*, pp. 87–104, March 2011.
- [27] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, June 1998.