Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

Improved label propagation algorithms with node attribute and link strength for community detection

Mohsen Arab*[1] and Mahdieh Hasheminezhad†[2]

[1,2]Department of Computer Science,Yazd University, Yazd, Iran , Combinatorial and Geometric Algorithms Lab.

ABSTRACT

Community detection has a wide variety of applications in different fields such as data mining, social network analysis and so on. Label Propagation Algorithm (LPA) is a simple and fast community detection algorithm, but it has low accuracy. There have been presented some advanced versions of LPA in recent years such as CenLP and WILPAS. In this paper, we present improved versions of CenLP and WILPAS methods called CenLP+ and WILPAS+ respectively. Experiments and benchmarks demonstrate that while CenLP+ is as fast as CenLP, it outperforms CenLP on both synthetic and real-world networks. Moreover, while accuracy of WILPAS+ on synthetic networks comparable with that of WILPAS, on real-world networks, WILPAS+ excels WILPAS. In addition, whereas both presented methods CenLP+ and WILPAS+ show high accuracy on synthetic networks, on real-world networks they outperform remarkably all other tested label propagation based algorithms for community detection. Therefore, since

Keyword: Label Propagation, Community Detection, Node Attribute, Link Strength.

AMS subject Classification: 05C79.

ARTICLE INFO

Article history:
Received 25, February 2018
Received in revised form 08, May 2018
Accepted 28 May 2018
Available online 01, June 2018

*mohsen.arab63@gmail.com
†Corresponding author: M. Hasheminezhad. Email:hasheminezhad@yazd.ac.ir

1 Abstract Continued:

CenLP+ and WILPAS+ are both fast and accurate, specially on real-world networks, they can efficiently reveal community structures of mega-scale social networks.

2 Introduction

Community structure is considered as an important property of real-world networks. Despite the lack of unique definition of community, it is widely accepted that a community has more internal connections than external ones . Communities can be found in many complex systems such as social and biological networks, the internet, food webs and so on. Nodes of a community have often several characteristics in common.
A wide variety of different methods have been proposed for community detection. In 2002, Newman and Girvan introduced a divisive algorithms using centrality indices called edge betweenness to find community boundaries [10]. In 2004, a measure called modularity was introduced to assess the quality of detected communities of a network [24]. After that, so many methods were presented for modularity optimization [2, 6, 3]. In addition to modularity optimization strategies, graph partition-based methods [23, 8, 29], density-based methods[31, 27] and label propagation algorithm (LPA) [25] have been presented for community detection.
Among all the community detection methods, LPA is one of the fastest algorithms. LPA algorithm is simple and its time complexity is nearly linear time. However because of randomness, the detected communities have poor stability. That is, LPA may find different communities in different runs. In some runs, small communities are merged with big ones forming "monster" communities which is a drawback of LPA [17].
The LPA can be described as follows. Initially, each node is assigned a unique numeric label. At each iterative step, each node updates its label to the most frequent label from its neighbours in a random order. When there are multiple most frequent labels, the node will randomly pick one of them. Relabeling continues until the label of each node is its most frequent label among its neighbours. Finally, the nodes with the same label are considered in the same community. Because of two sources of randomness, LPA shows low accuracy. First source is random update order of nodes for label updating and the second one is randomly selecting one label when there are multiple most frequent labels to choose.
To increase accuracy, CenLP method eliminates these two types of randomness. First, CenLP replaces random order of nodes with one deterministic order of nodes. Second, CenLP provides a new label choosing mechanism when there are multiple most frequent labels to select. In CenLP, for each node u , its most similar neighbor of higher local density, if exists, is defined as its preference node or $p(u)$. Then, If a node u has an equal maximum number of neighbor labels and one of them equals the label of $p(u)$, then node u adopts that label. CenLP+ changes this strategy a little to improve accuracy. More accurately, CenLP+ attempts to choose label of $p(p(u))$, when there are multiple maximum neighbour labels to select, regardless of the content of the set of maximum

neighbour labels. The main reason for doing this in CenLP+ is that $p(p(u))$ is more likely to be the center of its community than $p(u)$. Therefore, adopting label of node $p(p(u))$ for node u, can lead to a much more accurate community detection.
WILPAS is another method which increase accuracy of community detection with replacing two source of randomness of LPA with two deterministic part. First, nodes are arranged such that more important nodes update their labels first. Second, each node adopts a neighbour label that has more influence on it. Influence of a neighbour label can be calculated as the summation of all influence of neighbour nodes holding that label. Influence of a node u on node v can be estimated as importance value of node u multiplied by strength of the link connecting these two nodes. Important value of node u is estimated by degree of node u. Moreover, strength of a link can be estimated by the similarity measure between its two endpoints. WILPAS+ attempts to find for a each node u, a 'follower' node such the follower node have both higher importance (higher degree) than u and maximum influence on u at the same time. Then each node will adopt the label of its follower node. The reason behind this strategy is that, in real networks, nodes with high degrees have important role in forming communities, spreading information, viral marketing and so on. Thus, follower node of a node u is more likely to be an important node in the community of node u. Therefore, these follower nodes can guide us to find ground-truth communites of real-world networks with higher accuracy.
This paper is structured as follows. In Section 3, related works in the field are listed. Some notions are defined in Section 4. In Section 5 the proposed methods CenLP+ and WILPAS+ are presented. Experimental results of comparing the proposed methods with some famous methods in this area are discussed in section 6. Finally, conclusion is given in Section 7.

3 Related Works

In 2007, Raghaval et al.[25] proposed Label Propagation Algorithm (LPA) for community detection. LPA can be summarized as four following steps:
1) Initialize every node with a unique label.
2) Arrange the nodes in a random order.
3) For every node in that random order, set its label with the one which is the most frequent label among its neighbours.
4) If every node has a label that the maximum number of their neighbours have, then stop the algorithm; else go to step 2.
Label of each node u in LPA is defined as follows:

$$l(u) = \underset{l}{\operatorname{argmax}} \sum_{v \in N^l(u)} 1 \,, \tag{1}$$

where $N^l(u)$ indicates the set of neighbours of node u with label l. This is LPA's asynchronous version. Since synchronous version has potential label oscillations as discussed

in [25], we will not consider this version. As discussed earlier LPA has two types of randomness. Unfortunately, randomness of LPA may result in missing small communities and even getting trivial solution in which all nodes are assigned the same label [17]. Moreover, it makes the algorithm unstable such that different communities may be detected in different runs of the algorithm.
Zhang et al. generalized LPA to weighted networks by calculating the probability value of every label [33]. The label updating formula in this case is changed as follows:

$$l(u) = \underset{l}{\operatorname{argmax}} \sum_{v \in N^l(u)} w(u,v) \ , \tag{2}$$

where $w(u,v)$ indicates the weight of the edge between nodes u and v.
Barber and Clark proposed modularity-specialized algorithm (LPAm) to constrain the label propagation process [5]. Their algorithm is near-linear time, but it may get stuck in poor local maxima in the modularity space. To scape local maxima, Liu et al. introduced an advanced modularity-specialized label propagation algorithm called LPAm+ [19]. LPAm+ combines LPAm with multistep greedy agglomerative algorithm to get higher modularity values. Thus, LPAm+ does not guarantee near-linear time complexity [34]. Xing et al. presented a node influence based label propagation algorithm called NIBLPA [30]. NIBLPA defines two concepts node influence and label influence for specifying node orders and label choosing mechanism respectively. Zhang et al. proposed a label propagation algorithm with prediction of percolation transition named LPAp [34]. They transformed the process of label propagation into network construction process. Using this prediction process of percolation transition, they tried to delay the occurrence of trivial solutions. Sun et al. proposed a centrality-based label propagation called CenLP [28]. They presented a new measure for computing the centrality of nodes. Based on these centrality values, one specific update order in addition to node preference values are specified in order to improve traditional LPA. Arab et al. presented a novel label propagation algorithm called WILPAS with specific update order and new mechanism for label updating [4]. WILPAS method considers both node importance and link weight during label propagation process in order to reveal real community structure, while avoiding forming monster communities.

4 Terminology

Let $G = (V, E)$ be an undirected network where V is the set of nodes and E is the set of links. The number of nodes and links of G is denoted by n and m respectively. That is $n = |V|$ and $m = |E|$. Let d_u be the degree of node u in the network. Degrees of node u within and outside of its community are denoted by d_u^{in} and d_u^{out} respectively. Mixing parameter μ for each node u is defined as $\frac{d_u^{out}}{d_u}$. The set of all neighbours of node u is denoted by $N(u)$. Internal and external links respectively refers to the links within and between communities. Moreover, $w(u,v)$ refers to the weight of the link between nodes u and v. Also $l(u)$ indicates the label of node u.

5 The Proposed Methods

In this section, we first analyze the structure of CenLP method. To do that, we present the definitions of local density ρ , the similarity δ with higher density neighbours, the centrality of nodes $\frac{\rho}{\delta}$ and also pseudo-code of CenLP. After that, we introduce the proposed improved version of CenLP called CenLP+. In fact, CenLP+ modifies the label choosing mechanism of CenLP in order to increase peformance.

5.1 CenLP

The CenLP method proposes a specified node order for label updating and also a new label choosing formula. The specified node order is such that nodes which are less likely to be centers of communities should update their labels first. These nodes called border nodes. The basic assumption for specifying centers of communities is that community centers are surrounded by neighbors with lower local density and they have a relatively low similarity with any nodes with a higher local density. The label choosing formula is such that a node u prefers to adopt the label of a neighbour node whose local density is higher than that of itself and their similarity is maximum among all neighbors of the node. If such node exists, it is called node preference of node u or $p(u)$. In the followings some basic definitions of CenLP are presented.

Definition (Structure Neighborhood). The structure neighborhood of a node u is the set $\Gamma(u)$ containing u and its adjacent nodes:

$$\Gamma(u) = N(u) \cup \{u\} \tag{3}$$

Definition (Strength). The strength of a node u is defined as

$$k(u) = \sum_{v \in N(u)} w(u, v). \tag{4}$$

Definition (Structural Similarity). The structural similarity between two nodes u and v is defined as

$$\sigma(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)||\Gamma(v)|}}, \tag{5}$$

where $||$ indicates the cardinality of a set.

Definition (local Density). For a node $u \in V$, the local density is defined as

$$\rho_u = \frac{k(u)}{n-1} \tag{6}$$

where $k(u)$ and n are the strength of node u and the number of nodes in the network respectively.

Definition (Similarity with Nodes of Higher Density). For a node $u \in V$, the similarity with nodes of higher density is defined as

$$\delta_u = max_{v \in N(u) \wedge \rho_v > \rho_u} sim(u, v), \tag{7}$$

where $sim(u, v)$ refers to the structural similarity $\sigma(u, v)$.
Definition (Centrality). Given a weighted undirected network $G(V, E, w)$, the centrality γ_u of a node $u \in V$ is defined as

$$\gamma_u = \frac{\rho_u}{\delta_u}. \tag{8}$$

Since centers of communities will have high ρ and low δ, their γ values will be higher than those of other nodes. Thus, these nodes can be recognized by this characteristic. Then, the nodes are sorted in ascending order of their γ values. This specific order of nodes is used for label updating in CenLP. Therefore, the border nodes should be updated first. Moreover, remaining nodes based on their preference nodes , update their labels.
Definition (Preference Node). For a node $u \in V$, the preference node $p(u)$ is defined as

$$p(u) = \{v \in V | \underset{v \in N(u) \wedge \rho_v > \rho_u}{\operatorname{argmax}} sim(u, v)\} \tag{9}$$

Note that preference node $p(u)$ may not exist for some nodes, specially for nodes which are centers of their community. The main part of label choosing mechanism of CenLP is as follows. If a node u has an equal maximum number of neighbor labels and one of them equals $l(p(u))$, then set: $l(u) = l(p(u))$; if the node does not have a preference node, select a candidate label randomly.

5.2 CenLP+ method: improved version of CenLP

The presented CenLP+ method changes the mentioned label choosing mechanism of CenLP such that the accuracy of community detection can be improved. In fact, label choosing mechanism of CenLP+ for each node u is as follows. If a node u has an equal maximum number of neighbor labels, then if $p(p(u))$ exists, then set $l(u) = l(p(p(u)))$, else if $p(u)$ exists, set $l(u) = l(p(u))$. But if neither $p(p(u))$ nor $p(u)$ exist, then select a candidate label randomly. Moreover, CenLP+ does not check the existing of the label of preference node among neighbour labels. The pseudo-codes of two methods CenLP and CenLP+ are presented as Algorithm 1 and Algorithm 2 respectively.
Fig 1 shows karate club network. This network is formed by 34 members of a karate club in the United States. Because of a disagreement between administrator and instructor of the club, a new club was formed by the instructor by taking about the half of the original club members. This network has two communities specified by shapes 'circle' and 'square' in Fig 1. The edges between nodes (members) of this network represent the social interactions between the members outside the club. Two nodes '34' and '1' represent the club president and the instructor respectively. Therefore, these two nodes are the centers of their own communities.
In Fig 1 red arrow connecting a node v to node u indicates that u is preference node of v, i.e. $u = p(v)$. Starting from each node, by following red arrows at several steps, finally either two nodes '34' or '1' are reached, which are the centers of their own communities. Except starting from node '10', by following red arrows in Fig 1, ultimately the center

```
1  foreach u ∈ updateOrder do
2      if u has an equal maximum number of neighbors then
3          if isExist (p(u)) ∧ candidateLabels.contains( l(p(u)) ) then
4              l(u) =l(p(u));
5          else
6              l(u) = randomly select from candidate labels;
7          end
8      else
9          l(u) = label with the highest frequency among neighbors;
10     end
11 end
```

Algorithm 1: CenLP method

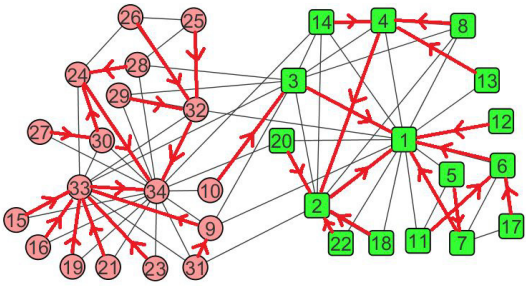

Figure 1: Preference nodes on karate club networks. Red arrow connecting a node v to node u indicates that u is preference node of v, i.e. $u = p(v)$.

of community of that node is reached. In experimental results we will see that, for each node v, adopting label of $p(p(v))$ as suggested by CenLP+, instead of choosing label of $p(v)$, as proposed by CenLP, will lead to more accurate community detection. This is because $p(p(v))$ is more close to the center of its community than $p(v)$.

```
1  foreach u ∈ updateOrder do
2      if u has an equal maximum number of neighbors then
3          if isExist (p(u)) then
4              if isExist (p(p(u))) then
5                  l(u) = l(p(p(u)) );
6              else
7                  l(u) = l(p(u));
8              end
9          else
10             l(u) = randomly select from candidate labels;
11         end
12     else
13         l(u) = label with the highest frequency among neighbors;
14     end
15 end
```

Algorithm 2: CenLP+ method

5.3 WILPAS method

The method WILPAS is based on this assumption that each node u has an effect on each of its neighbour node v according to both $w(u, v)$ as the strength of relationship between the two nodes and also d_u as the importance value of node u. Therefore, $w(u, v) * d_u$ can be thought as the influence value of node u on the neighbour node v which is denoted by $inf(u, v)$. Thus,

$$inf(u, v) = w(u, v) * d_u. \tag{10}$$

One can simply extend this definition to be used for influence of a label l on a node. More accurately, influence value of label l on a node v can be thought as the summation of the influence values of its neighbour nodes having label l. That is:

$$inf(l, v) = \sum_{u \in N^l(v)} inf(u, v) \tag{11}$$

In WILPAS, the extended importance value of each node is defined as $EI(v) = d_v + \sum_{u \in N(v)} d_u$. WILPAS has two stages. Stage one has specific node order for label updating and also one special label updating formula. Stage two of WILPAS is similar to LPA with random update order and randomly selecting most frequent labels. The specified node

order of stage one is based on descending order of EI values of nodes. In addition, label updating formula of stage one of WILPAS is such that each node v adopts a neighbour label l having maximum influence $inf(l, v)$. In other words, in stage one of WILPAS, the new label $l(v)$ for a node v is defined as follows:

$$l(v) = \underset{l}{\operatorname{argmax}}\ inf(l, v)\ . \tag{12}$$

The process of label updating continues in iterative steps untill labels of nodes do not change anymore. Stage two of WILPAS is injecting detected labels from stage one into ordinary label propagation agorithm (LPA). The stage two of WILPAS causes possible sub-communities to be merged to get real ones.

5.4 WILPAS+: improved version of WILPAS

The main goal for presenting WILPAS+ is to improve the quality of detected communities of WILPAS for real-world networks, while increasing its speed. In stage one of WILPAS, as discussed, each node adopts a label with maximum influence on it. But in WILPAS+, at first, each node v defines a neighbour node called following node $f(v)$. The node $f(v)$ has both high degree and high influence on v. Then, in stage one of WILPAS+, each node adopts the label of its following node. Stage two of WILPAS+ is similar to standard LPA with one difference. Once there are multiple most frequent labels to select, choose the one with maximum importance, i.e. adopt the one that their corresponding nodes have higher degrees.

The more accurate descriptions for WILPAS+ are presented as follows.

In WILPAS+ method, each node v attempts to find a following node $f(v)$ with two primary conditions: 1) $f(v)$ has high influence on v. 2) $f(v)$ should be more close to the center of its community than v. Therefore, one candidate for $f(v)$ is a neighbour node u of v such that $d_u \geq d_v$ and the influence $inf(u, v)$ is maximum. Consider two nodes which have maximum degrees of their own communities. If these two nodes are connected with an edge, there is a possibility that one of them becomes following node of another. This can result in merging the two communities by WILPAS+. To avoid that, the third condition for defining following node is necessary. At first, let define for a node v a neighbour node u with maximum influence on it and denoted it by $max\_inf(v)$:

$$max\_inf(v) = \underset{u \in N(v)}{\operatorname{argmax}}\ \ inf(u, v) \tag{13}$$

Then, the third condition can be described as follows. The 'following' node $f(v)$ should be such that $inf(v, f(v))$ should be equal or greater than $\alpha * max\_inf(v)$. The parameter α is a arbitrary threshold such that $0 < \alpha < 1$. If α is too small, the above problem of possibility of merging two communities may still exist. On the other hand, if α is very close to one, the second condition would be very hard to be satisfied. In experiment section we set $\alpha = 0.50$. For node v, following node $f(v)$ is defined as follows.

$$f(v) = \underset{u \in N(v)\ \wedge\ d_u \geq d_v\ \wedge\ inf(u,v) \geq \alpha * max\_inf(v)}{\operatorname{argmax}} inf(u, v) \tag{14}$$

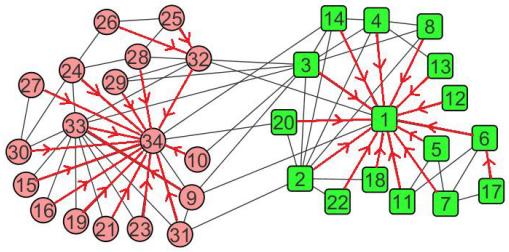

Figure 2: Following nodes on karate club network for WILPAS+ method. Red arrow connecting a node v to node u indicates that u is following node of v, i.e. $u = f(v)$.

Fig 2 shows following nodes of karate club network. Red arrow connecting a node v to node u indicates that u is following node of v, i.e. $u = f(v)$. Starting from each node, by following red arrows with at most two steps, finally either two nodes '34' or '1' are reached, which are the centers of their own communities.
WILPAS+ method has two stages. In first stage of WILPAS+, each node will get the label of its following node . There is an efficient way to do that. Consider red arrows in Fig 2 as edges. Let E' denote the set consisting of these edges. In graph $G(V, E')$ find connected components of nodes. Assign all nodes of each connected component the same label as a community. For example, by finding connected components of graph $G(V, E')$, where $G(V, E)$ is karate club network, two real communities of this network are detected. Finding connected components of $G(V, E')$ takes $O(|E'|)$ time using DFS algorithm. Since $|E'| = n$, stage one of WILPAS+ takes linear time complexity $O(n)$. The second stage of WILPAS+ is like ordinary label propagation algorithm with one difference: When there are multiple most frequent neighbour labels, the one is chosen which their corresponding neighbour nodes have higher degrees.

6 Experiments

This section evaluates the effectiveness and the efficiency of CenLP+ and WILPAS+. We conduct experiments on both artificial networks and famous real-world networks. We compare the performance of CenLP+ and WILPAS+ with LPA, CenLP, LPAp, LPAm, NIBLPA and WILPAS. All the simulations are carried out in a desktop pc with Pentium Core 2, 1.8 GHZ processor and 4GB of RAM under Windows 8.1 OS.
In this paper, we use normalized mutual information (NMI) as the evaluation measure which is currently widely used in measuring the quality of detected communities. NMI allows us to measure the amount of information common to two different network partitions. Accordingly, if real known partition matches detected ones, we have NMI=1, but when two partitions are independent of each other, we have NMI=0.

6.1 Test on synthetic networks

In this section, LFR benchmark networks is used which are currently the most commonly used synthetic networks in community detection [16]. The parameters of LFR benchmark networks are as follows: number of nodes n, the average degree k, maximum degree $maxk$, mixing parameter μ. Moreover, $minc$ and $maxc$ refer to the minimum and maximum values for community sizes respectively.
Two ranges for different community sizes indicated by the letters B (stays for big) and VB (stays for very big) are chosen. These wo ranges for letters B and VB are $[cmin, cmax] = [20, 100]$ and $[cmin, cmax] = [200, 1000]$ respectively. For each type of networks, we generate 10 samples and for each sample we run 10 times each tested label propagation-based algorithm. Then, the average of these 100 NMI values are reported as output.
Fig 3 shows the accuracy of the mentioned methods on the networks with size of 1000. We observe that for $n = 1000$, when $\mu \leq 0.55$, CenLP+ and WIPLAS+ methods get higher NMI values than ordinary CenLP and WILPAS methods. Moreover, when $\mu \leq 0.55$, CenLP+ and LPAm have the highest accuracy for community detection. WILPAS+ is third best accurate method for this range of μ on this network.
For $n = 10000$, it can be observed from Fig 4 that LPA, LPAp and NIBLPA has the lowest accuracy. Furthermore, three methods CenLP, CenLP+ and LPAm have approximately the same NMI results. WILPAS method has best accuracy on this network.

Fig 5 demonstrates the NMI results for the five most accurate tested label propagation methods i.e. WILPAS, CenLP, WILPAS+, CenLP+ and LPAm for a network with $n = 100000$, $k = 40$, $[cmin, cmax] = [200, 1000]$. As it can be observed from this figure, four methods WILPAS, CenLP, WILPAS+, CenLP+ have the same accuracy for $\mu \leq 0.65$. For $\mu = 0.70$, two methods WILPAS+ and CenLP+ obtain the same NMI result, but a little lower than that of WILPAS. Thus, in addition on the network with $n = 1000$ in Fig

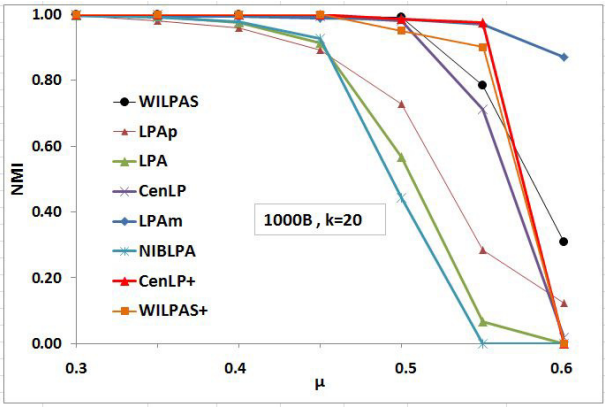

Figure 3: Comparing different label propagation-based algorithms on the network with $n = 1,000$.

3, CenLP+ excels ordinary CenLP method on network with $n = 100000$ as well (see Fig 5).
In summary, on synthetic networks, CenLP+ methods outperforms CenLP method. WILPAS+ has more accuracy than WILPAS on the synthetic network with $n = 1000$. However on networks with $n = 10000$ and $n = 100000$, for largest tested value of mixing parameter μ, accuracy of WILPAS is a little lower than that of WILPAS. But as we can see later, on real-world networks, WILPAS+ method outperforms ordinary WILPAS method.

6.2 Experiment on Real-world Networks

In this section, we are going to evaluate the above methods on real-world networks which their communities are already known. Zachary karate club [32], American college football

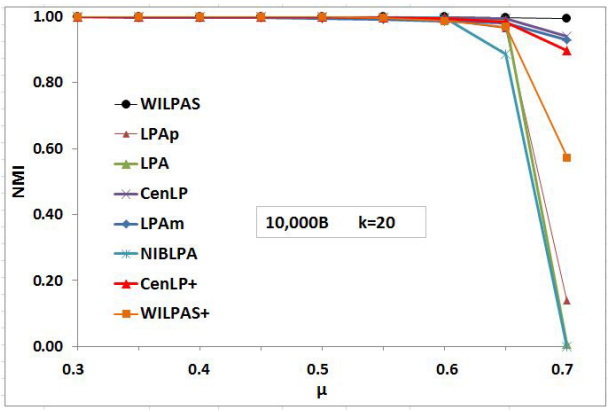

Figure 4: Comparing different label propagation-based algorithms on the network with $n = 10,000$.

[10], dolphin social network [22] and Polblog [1] are four famous networks in the field. The details of these networks are shown in Table 1. The NMI results of all tested label propagation-based methods are displayed in Table 2.
Each method is run 10 times on each real network, then the average NMI results are reported. The number in the {} for CenLP, NIBLPA and WILPAS, CenLP+ and WILPAS+ in Table 1 shows the number of found communities by these five deterministic methods. Since LPA, LPAp and LPAm detect different partitions on the same network for each run, we ignore them. The maximum resulted NMI values on each network has been bold in Table 2.

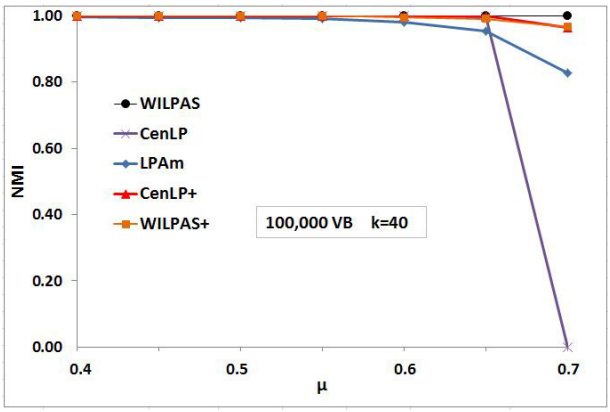

Figure 5: Comparing different label propagation-based algorithms on a network, when $n = 100,000$ and average degree $k = 40$.

Network	Nodes	Links	Communities
Karate [32]	34	78	2
Dolphin [22]	62	159	2
Football [10]	115	615	12
Polblog [1]	1490	16715	2

Table 1: Real-world networks with known community structures.

6.2.1 Zachary karate Club

The well-known karate club network of Zachary [32] is a standard benchmark for community detection. Zachary observed 34 members of a karate club in the United States over

two years. At some point, a disagreement between the club president and the instructor resulted in the split of the club into two separate groups. These two original communities are specified with shapes 'circle' and 'square' in Fig 6.
Fig 6 shows the output of two methods CenLP and CenLP+ on karate network. While CenLP divides this network into four communities, CenLP+ method detects two real communities of karate network approximately as it is. More accurately, CenLP+ just assign node 10 incorrectly to another community. WILPAS and WILPAS+ detect two original communities perfectly with NMI=1.

6.2.2 Dolphin social network

Dolphin network [22] shows the frequent associations between 62 dolphins living in Doubtful Sound, New Zealand. Nodes are dolphins and the edges between nodes shows that the two corresponding dolphins were seen together more than expected by chance. After leaving one of dolphins, they separated in two communities. These two original communities are specified by shapes 'circle' and 'square' in Fig 7. As it is observable from Fig 7, WILPAS+ has higher accuracy than WILPAS. Moreover, CenLP+ is more accurate than CenLP on this network. Among all the tested methods on this network, WILPAS+ gets the highest NMI value, as it can be seen from Table 2.
It is important to note that NMI measure is more sensitive to assignment of nodes to a wrong community rather than dividing a real communities into several sub-communities. For instance, if node 40 for WILPAS+ method in Fig 7.b was correctly assign to its real community (the community shaded with red), then instead of NMI=0.75, NMI=0.84 would be obtained. Moreover, CenLP in fact finds two different partitions with NMI=0.58 and NMI=0.64 respectively. The NMI value 0.61 for CenLP on dolphin network in Table 2 is the average of these two values in 10 runs. The detected partition shown in Fig 7.c is related to NMI=0.64 .

6.2.3 American college football

Another well known benchmark for community detection is American college football network compiled by Girvan and Newman [10]. This network represents Division I games for the 2000 season. Nodes represent teams and the edges represent the games between teams. This network has 12 communities.
As one can see from Table 2, WILPAS+ get highest NMI value 0.92 with finding 11 communities which is very close to 12 real communities of this network. CenLP+ is the second accurate method with resulted NMI=0.91.

6.2.4 Polblogs network

This network represents the links between weblogs about US politics preceding the US Presidential Election of 2004 [1]. The links were automatically extracted from a crawl of the front page of the weblogs. Each blog is labeled with '0' or '1' to indicate whether they are "liberal" or "conservative". This network can be considered both directed or

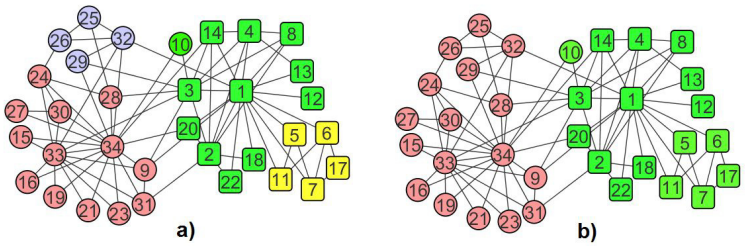

Figure 6: a) result of CenLP method on karate club network. b) result of CenLP+ method on karate club network.

undirected. In this paper, the undirected version of this network is considered which has 1490 nodes and 16715 links. Since nodes with degree zero makes this network disconnected, when comparing the performance of methods, these nodes are ignored. Thus, by removing 266 nodes with degree zero in addition to removing two nodes with degree one, a connected network with 1222 nodes is obtained. This resulted network is considered for testing and comparing community detection methods.

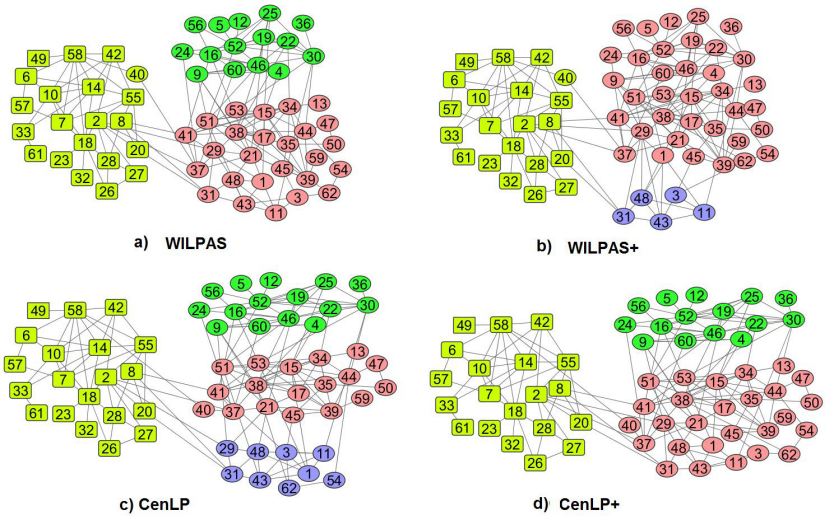

Figure 7: Detected communities of four methods WILPAS, WILPAS+, CenLP and CenLP+ on the dolphin network.

However CenLP and CenLP+ achieve maximum NMI value 0.71 on Polblog network, but the number of detected communities of WILPAS and WILPAS+ shows accuracy of these two methods in finding true number of original communities.

When dealing with community detection on real-world networks, WILPAS+ outperforms other methods on karate and dolphin and football network with highest obtained NMI value. CenLP+ is the second accurate method on real-world networks based on the resulted NMI values. Furthermore, on real-world networks, the numbers of detected communities of WILPAS and WILPAS+ is more close to the numbers of known communities of these networks.

network	LPAm	LPAp	WILPAS	LPA	NIBLPA	CenLP	CenLP+	WILPAS+
Karate	0.55	0.56	**1**,{2}	0.70	0.21 {3}	0.60, {4}	0.84, {2}	**1**,{2}
Dolphin	0.45	0.55	0.66,{3}	0.52	0.50 {5}	0.61,{4}	0.74, {3}	**0.75**,{3}
Polblog	0.45	0.61	0.70,{2}	0.70	0.20 {9}	**0.71** ,{3}	**0.71**, {3}	0.69,{2}
Football	0.89	0.88	0.90,{13}	0.87	0.78 {9}	0.90, {13}	0.91 , {14}	**0.92** ,{11}

Table 2: NMI results of the methods on four real networks with known community structures.

6.3 Efficiency analysis

To illustrate the running time of the proposed algorithms Cenlp+ and WILPAS+ and compare them with other algorithms, we produce 10 networks using LFR software, where the number of nodes $n = 100,000$ and the average degree $k = 40$ and $[minc, maxc] = [200, 1000]$ and mixing parameter $\mu = 0.40$. Figure 8 plots the average running time (in seconds) of our proposed methods CenLP+ and WILPAS+ on these 10 synthetic networks compared with other six label propagation algorithms: LPA, LPAm, LPAp, NIBLPA, CenLP and WILPAS. As we can see from Figure 8, while two methods CenLP and CenLP+ have the same running time, WILPAS+ is faster than WILPAS. More accurately, WILPAS+ is a little faster than CenLP and CenLP+, much faster than WILPAS and LPAm, but slower than LPA, LPAp and NIBLPA.
In summary, while the proposed method WILPAS+ has a little lower accuracy than that of WILPAS on synthetic networks, its accuracy on real-world network shows remarkable improvement in comparison to WILPAS. Moreover, another proposed method CenLP+ shows higher accuracy than ordinary CenLP method on both synthetic and real-world networks. In addition, two improved method CenLP+ and WILPAS+ preserve the speed of original methods CenLP and WILPAS.

7 Conclusion

In this paper, we propose two improved versions of label propagation-based algorithms CenLP and WILPAS denoted by CenLP+ and WILPAS+ respectively. Both of these

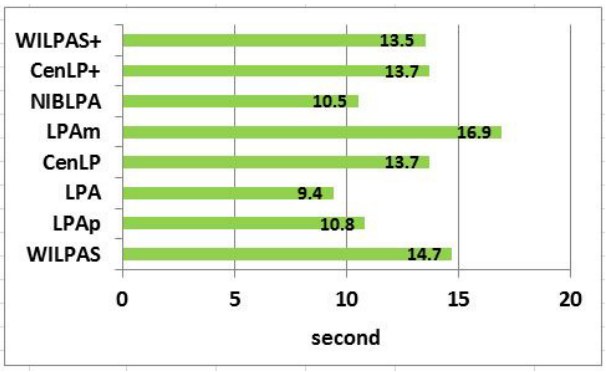

Figure 8: The execution times of different methods on a network with n=100,000, k=40, $[minc, maxc] = [200, 1000]$, $\mu = 0.40$.

two presented methods use node importance and link strength for community detection. Experimental results show that CenLP+ is more accurate than CenLP on both artificial and real-world networks, while preserving original speed of CenLP. Moreover, experiments show that while accuracy of WILPAS+ is comparable to that of WILPAS on synthetic networks, on real-world networks it demonstrates remarkable improvement in community detection.

In summary, both WILPAS+ and CenLP+ shows high accuracy in detecting true community structures of networks while preserving the speed of the original methods. On real-world networks, both WILPAS+ and CenLP+ outperform all other tested methods with gaining higher NMI values. In fact, experiments on several well-known real-world networks demonstrate that WILPAS+ is more capable of finding the number of known communities of these networks. Therefore, two presented methods WILPAS+ and

CenLP+ can be used for efficient community detection on large real-world social networks.

References

[1] Adamic L.A., and Glance, N., The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery, (2005) pp.36-43.

[2] Agarwal G., and Kempe, D., Modularity-maximizing graph communities via mathematical programming. The European Physical Journal B, 66 (2008), pp.409-418.

[3] Arab M., and Afsharchi, M., Community detection in social networks using hybrid merging of sub-communities. Journal of network and computer applications, 40 (2014), pp.73-84.

[4] Arab M. and Hasheminezhad, M., Efficient Community Detection Algorithm with Label Propagation using Node Importance and Link Weight International Journal of Advanced Computer Science and Applications(IJACSA), 9(5), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090566

[5] Barber M.J., and Clark, J. W., Detecting network communities by propagating labels under constraints. Physical Review E, 80(2009), p.026129.

[6] Bennett L., Liu, S., Papageorgiou, L.G., and Tsoka, S., A mathematical programming approach to community structure detection in complex networks. In Computer Aided Chemical Engineering 30 (2012) pp. 1387-1391.

[7] Danon L., Diaz-Guilera, A., Duch, J. and Arenas, A., Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, (2005) (09), p.P09008.

[8] Flake G.W., Lawrence, S., and Giles, C.L., Efficient identification of web communities. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. (2000) pp. 150-160.

[9] Fortunato S., and Barthelemy, M., Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104 (2007), pp.36-41.

[10] Girvan M., and Newman, M. E., Community structure in social and biological networks. Proceedings of the national academy of sciences, 99 (2002), pp.7821-7826.

[11] Guimera R., Sales-Pardo, M., and Amaral, L.A.N., Modularity from fluctuations in random graphs and complex networks. Physical Review E, 70(2004), p.025101.

[12] Lancichinetti A., and Fortunato, S., Community detection algorithms: a comparative analysis. Physical review E, 80(2009), p.056117.

[13] Lancichinetti A., and Fortunato, S., Limits of modularity maximization in community detection. Physical review E, 84(2011), p.066122.

[14] Lancichinetti A., Radicchi, F., Ramasco, J.J., and Fortunato, S., Finding statistically significant communities in networks. PloS one, 6(2011), p.e18961.

[15] Lancichinetti A., Fortunato, S. and Kertsz, J., Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11 (2009), p.033015.

[16] Lancichinetti A., Fortunato, S., and Radicchi, F., Benchmark graphs for testing community detection algorithms. Physical review E, 78(2008), p.046110.

[17] Leung I.X., Hui, P. Lio, P. and Crowcroft, J., Towards real-time community detection in large networks. Physical Review E, 79(2009), p.066107.

[18] Li S., Lou, H., Jiang, W., and Tang, J., Detecting community structure via synchronous label propagation. Neurocomputing, 151, (2015) pp.1063-1075.

[19] Liu X., and Murata, T., Advanced modularity-specialized label propagation algorithm for detecting communities in networks. Physica A: Statistical Mechanics and its Applications, 389(2010), pp.1493-1500.

[20] Liu Z., Li, P., Zheng Y., and Sun, M., Community detection by affinity propagation (2008). Technical Report.

[21] Lou H., Li, S., and Zhao, Y., Detecting community structure using label propagation with weighted coherent neighborhood propinquity. Physica A: Statistical Mechanics and its Applications, 392(2013), pp.3095-3105.

[22] Lusseau D., and Newman, M.E., Identifying the role that animals play in their social networks. Proceedings of the Royal Society of London B: Biological Sciences, 271 (Suppl 6), (2004) pp.S477-S481.

[23] Kernighan B.W., and Lin, S., An efficient heuristic procedure for partitioning graphs. The Bell system technical journal, 49 (1970), pp.291-307

[24] Newman M.E., and Girvan, M., Finding and evaluating community structure in networks. Physical review E, 69 (2004), p.026113.

[25] Raghavan U.N., Albert, R., and Kumara, S., Near linear time algorithm to detect community structures in large-scale networks. Physical review E, 76(2007), p.036106.

[26] Rosvall M., and Bergstrom, C.T., Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(2008), pp.1118-1123.

[27] Sun H., Huang, J., Han, J., Deng, H., Zhao, P. and Feng, B., gskeletonclu: Density-based network clustering via structure-connected tree division or agglomeration. In Data Mining (ICDM), 2010 IEEE 10th International Conference on (2010) pp. 481-490.

[28] Sun H., Liu, J., Huang, J., Wang, G., Yang, Z., Song, Q,. and Jia, X., CenLP: A centrality-based label propagation algorithm for community detection in networks. Physica A: Statistical Mechanics and its Applications, 436 (2015), pp.767-780.

[29] White S., and Smyth, P., A spectral clustering approach to finding communities in graphs. In Proceedings of the 2005 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, (2005) pp. 274-285.

[30] Xing Y., Meng, F., Zhou, Y., Zhu, M., Shi, M., and Sun, G., A node influence based label propagation algorithm for community detection in networks. The Scientific World Journal, (2014).

[31] Xu X., Yuruk, N., Feng, Z., and Schweiger, T.A., Scan: a structural clustering algorithm for networks. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. (2007) pp. 824-833.

[32] Zachary, W.W., 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research, 33(1977), pp.452-473.

[33] Zhang A., Ren, G., Cao, H., Jia, B., and Zhang, S., Generalization of label propagation algorithm in complex networks. In Control and Decision Conference (CCDC), 2013 25th Chinese (2013) pp. 1306-1309.IEEE.

[34] Zhang A., Ren, G., Lin, Y., Jia, B., Cao, H., J. and Zhang, S., Detecting community structures in networks by label propagation with prediction of percolation transition. The Scientific World Journal, (2014).