A Seed-Centric Community Detection Algorithm based on an Expanding Ring Search

Kwan Hui Lim **Amitava Datta**

School of Computer Science and Software Engineering
The University of Western Australia
Crawley, WA 6009, Australia
Email: kwanhui@graduate.uwa.edu.au, datta@csse.uwa.edu.au

Abstract

One common problem in viral marketing, counter-terrorism and epidemic modeling is the efficient detection of a community that is centered at an individual of interest. Most community detection algorithms are designed to detect all communities in the entire network. As such, it would be computationally intensive to first detect all communities followed by identifying communities where the individual of interest belongs to, especially for large scale networks. We propose a community detection algorithm that directly detects the community centered at an individual of interest, without the need to first detect all communities. Our proposed algorithm utilizes an expanding ring search starting from the individual of interest as the seed user. Following which, we iteratively include users at increasing number of hops from the seed user, based on our definition of a community. This iterative step continues until no further users can be added, thus resulting in the detected community comprising the list of added users. We evaluate our algorithm on four social network datasets and show that our algorithm is able to detect communities that strongly resemble the corresponding real-life communities.

Keywords: Community detection, clustering algorithm, social networks

1 Introduction

Most community detection algorithms aim to detect all community structures in the entire network graph, which is both tedious and computationally intensive due to the large scale of current social networks. For purposes such as viral marketing, counter-terrorism and epidemic modeling, we are most interested in the community surrounding a particular individual because he/she is determined to be influential in the spread of product information (viral marketing), at the heart of a terrorist organization (counter-terrorism), or a high-risk individual for an infectious disease (epidemic modeling). As such, it would be more efficient to focus directly on a community that is centered at this influential individual, compared to first detecting all communities followed by identifying the communities that this individual belongs to.

Hence, we propose a community detection algorithm that directly detects a community centered at an individual of interest. Our proposed algorithm starts from a seed user (i.e. the individual of interest) and performs an expanding ring search to iteratively include users into that community. Users are included into the community based on a metric of their number of links to other users in the community. This iterative adding of users continues until no further users satisfy the metric and could be added. Our main contributions include proposing this seed-centric community detection algorithm (Section 3) and evaluating this algorithm on three real-life social networks and the YouTube online social network (Section 5 and 6).

2 Related Work

There exists an extensive literature on community detection algorithms and we focus on those based on a set of seed nodes, as these algorithms are more closely related to our work. Andersen and Lang proposed an algorithm based on a series of random walkers, each traversing a limited number of steps starting from a set of seed nodes (Andersen & Lang 2006). This algorithm then uses network flow to clean up the results before returning the detected community based on a selection of nodes that the random walkers have traversed through.

Similarly, Andersen et al. proposed a local community detection algorithm based on a set of seed nodes using a modified version of the PageRank algorithm (Andersen et al. 2006). A series of random walkers start from this set of seed nodes and each node they traverse is considered for inclusion into the community based on the value of their resulting PageRank vector. Our proposed algorithm differs from the algorithms by Andersen and Lang, and Andersen et al. in that we detect communities surrounding a single seed node whereas they require a set of seed nodes. Also, our method differs in the definition of the metric that is used to determine whether a node should be included in a community.

Similarly, there are various algorithms for detecting communities using a single seed node. Clauset introduced the local modularity R which measures how much a node is on the boundary of the community (Clauset 2005). Clauset then starts from a seed node and iteratively adds neighbouring nodes into the community that maximizes the modularity R, resulting in the detection of a local community. Our proposed algorithm differs from Clauset's in our definition of modularity and the option to modify this modularity to detect communities of different strength.

In the same spirit as Clauset (i.e. the local maximization of modularity), Luo et al. proposed an algorithm that starts from a seed node and uses an iteration of adding and deleting nodes until the local maximization of modularity at the eventual commu-

nity (Luo et al. 2006). However, this method could potentially exclude the seed node and result in a detected community without the seed node. This potential exclusion of the seed node is the main difference with our algorithm, which ensures that the seed node is still included in the detected community.

3 Methodology

Most definitions of a community are generally based on the concept that the community comprises individuals who are more densely connected to each other in the community than to those outside the community. Specifically, Radicchi et al. introduced the concept of strong and weak communities where strong communities comprise *individual* users who each has more links within this community than outside, while weak communities comprise users who *collectively* have more links within this community than outside (Radicchi et al. 2004). In particular, we implemented a modified version of Radicchi et al.'s definition of a strong community by introducing a community strength factor for adjusting the size and strength of the community detected.

We first model the social network as an undirected, unweighted graph, $G = (N, E)$ where N and E respectively refer to the set of nodes/users and edges/links in the graph. Undirected links correspond to social links that are reciprocal and reflective of real-life friendships, thus our choice of undirected links for the algorithm. While our paper uses unweighted links, the algorithm could cater for weighted links by implementing a simple filtering scheme based on the weight of links. This filtering scheme would work in such a way that links below a certain threshold weight are excluded for consideration as part of the graph.

Each user $i \in N$ has k_i links, with each link pointing to another user either within or outside the community. The number of links pointing to users within the community is denoted as k_i^{in} and those outside the community as k_i^{out}. In addition, we introduce a community strength factor f that allows us to adjust the size and strength of the detected communities. Our definition of a community is as denoted:

$$k_i^{in} > k_i^{out} \times f \tag{1}$$

Our proposed algorithm differs from that of Radicchi et al. in two ways. Firstly, we introduce a community strength factor f to their original definition of a strong community, thus allowing us to adjust the strength and size of the community detected. Secondly, the method proposed by Radicchi et al. takes an entire graph and iteratively divides it until the separate communities emerge, whereas our algorithm starts from a single seed user and gradually builds up the community surrounding this user.

Our algorithm (as presented in Algorithm 1) can be broadly divided into the following steps:

1. Identify a user of interest as the seed node and include this user as part of the community.
2. Retrieve all neighbouring nodes of the seed node. Include these 1st degree (one-hop) neighbours as part of the community.
3. Retrieve all the 2nd degree (two-hops) neighbours of the seed node (i.e. neighbours of the neighbours of the seed node). Include them as part of the community if they fulfill our definition of a community as stated in Equation 1.

Algorithm 1 Seed-centric Community Detection

Input: $G = (N, E)$: An undirected, unweighted social network graph, $s \in N$: the seed node
Output: detectedCommunity: A list of nodes in the community centered at the seed node s

```
begin
Add Node s to detectedCommunity
for all Neighbour n_s of Node s do
   Add n_s to detectedCommunity
end for
for all Neighbour n_s of Node s do
   for all Neighbour m_n of Node n_s do
      Add m_n to listNeighbours
   end for
end for
while listNeighbours ≠ NULL do
   for all Node n in listNeighbours do
      if k_n^in > k_n^out × f then
         Add n to detectedCommunity
         Add n to listNewMembers
      end if
   end for
   listNeighbours = NULL
   for all Node n in listNewMembers do
      for all Neighbour m_n of Node n do
         Add m_n to listNeighbours
      end for
   end for
end while
return detectedCommunity
end
```

4. Repeat Step 3 for the 3rd, 4th, nth degree neighbours until no further nodes can be added to the community.
5. The eventual list of included nodes would be the community centered at the seed node.

As our algorithm aims to detect a community centered at an individual of interest, Step 1 is to identify such a user as the seed node s. In real-life, this seed node s can correspond to an individual with a large number of links to other users, or a person in a particularly influential position (e.g. the CEO of a company or the director of a research institute). Next, Step 2 includes all neighbours of the seed node s as part of his/her community, which is reasonable as these neighbours are one-hop friends of seed node s who he/she is more likely to interact with frequently. Following which, Steps 3 and 4 are basically iterative steps that continuously include nodes (which satisfy Equation 1) in an expanding ring search. This expanding ring search coupled with our definition of a community (Equation 1) ensures that the search does not propagate too far, as nodes that do not satisfy this definition will not further propagate the search.

4 Experimental Setup

In order to validate the correctness of the communities detected by our algorithm, it is important to evaluate our community detection algorithm on social networks where we know the ground truth (i.e. the real-life communities). For this purpose, we selected the Zachary Karate Club, Doubtful Sound Dolphins and Santa Fe Institute Collaboration datasets which have been used by many authors to establish the correctness of their community detection algorithms (Girvan & Newman 2002, Arenas et al. 2008).

The Zachary Karate Club and Doubtful Sound Dolphin datasets comprise 34 and 62 nodes respectively, where each dataset is further divided into two different communities (Zachary 1977, Lusseau et al. 2003). The Santa Fe Institute Collaboration dataset comprises 118 nodes which are further divided into four communities, each representing a different field of research (Girvan & Newman 2002). These datasets are chosen as we know the ground truth of the actual real-life communities and can compare them to the communities detected by our algorithm.

Next, we also evaluate our algorithm on a large-scale online social network based on YouTube. This dataset comprises 1.1 million nodes, 2.9 million edges and nodes may join any of the 47 different YouTube groups (Tang & Liu 2009).[1] The main challenge with evaluating community detection algorithms on online social networks is the verification of actual real-life communities (i.e. establishing the ground truth). In this case, we adopt the best approximation of ground truth by using the YouTube groups that the users belong to. Users who are members of the same YouTube group are inferred to be members of the same real-life community. In addition, we further validate our algorithm using network properties such as average clustering coefficient, average path length, average degree and diameter as measures of the topological structure of the detected communities.

5 Evaluation on Real-life Social Networks

We begin our evaluation on the three real-life social networks by first selecting the seed nodes for each social network. For the Zachary Karate Club, we chose the club president and instructor as the two seed nodes, who also have the highest number of links. Similarly, for the Doubtful Sound Dolphins, we chose two nodes with the highest number of links in their respective communities as the seed nodes. Likewise for the Santa Fe Institute Collaboration Network, we selected one seed node from each field of research who also have one of the highest number of links.

5.1 Overview of Results

We first evaluate the correctness of our algorithm by examining the precision and recall results on the three datasets. Precision refers to the number of correct nodes classified out of all nodes classified while recall indicates the number of correct nodes classified out of all actual nodes in the community. In terms of recall, our algorithm is able to detect almost all nodes ($\geq$98.5%) that belong to their respective communities. We were able to achieve 100% recall for both the Doubtful Sound Dolphins and Zachary Karate Club datasets. The recall rate for the Santa Fe Institute dataset was also high at 98.5%.

Similarly, the results for precision are also relatively high with our algorithm correctly classifying 97.6%, 87.2% and 84.9% of nodes into their actual communities for the Doubtful Sound Dolphins, Santa Fe Institute and Zachary Karate Club datasets, respectively. While the results for precision are high, it is worthwhile to further examine and understand why some nodes are incorrectly classified.

5.2 Further Analysis of Results

We now analyze the Zachary Karate Club dataset where Fig. 1 shows the communities detected (circled

[1]Tang and Liu have made this dataset publicly available at http://socialcomputing.asu.edu/datasets/YouTube2.

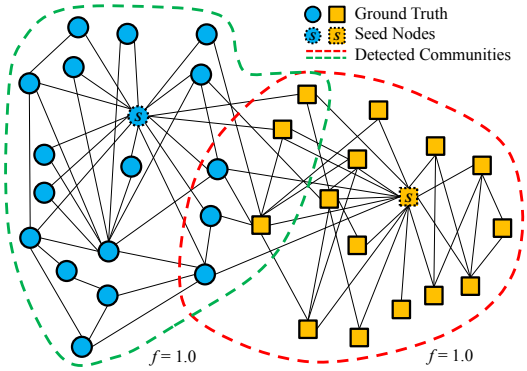

Figure 1: Zachary Karate Club

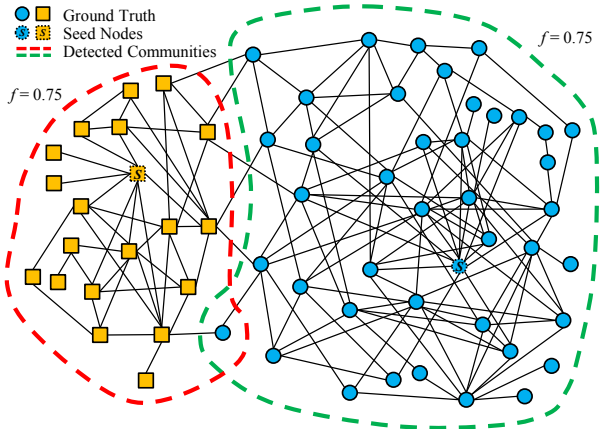

Figure 2: Doubtful Sound Dolphins

with a dashed line) by our algorithm compared to the ground truth of actual communities (indicated by different shapes and colour of the nodes). For the six nodes that were mis-classified into the wrong community, four of them (a two-third majority) had direct links to both seed nodes in the two respective communities (i.e. the club president and instructor). The remaining two nodes were directly linked to the seed node in one community while being one-hop away from the seed node of the other community. This close proximity of the mis-classified nodes to the seed nodes of the two communities show that the mis-classified nodes actually act as effective bridges or middle-men between the two communities. As such, they would be better classified as members of both communities rather than just belonging to a single community.

While the results are different from the ground truth, this is consistent with the observations of many authors that there are overlapping communities in social networks and individuals may belong to multiple communities (Palla et al. 2005). Furthermore, in Zachary's study of the karate club, he also noted that "not all individuals in the network were solidly members of one faction or the other", thus further supporting the results of our algorithm (Zachary 1977).

Similarly, the one mis-classified node for the Doubtful Sound Dolphins acts as such a bridge between the two communities. As shown in Fig. 2, this node is on the edge of both communities and have one link into each community. Hence, this node can easily belong to either community and would be better classified as belonging to both communities, considering its topological links and position in the network. Other authors also shared similar views that if a node has only a single link to a community, it should be

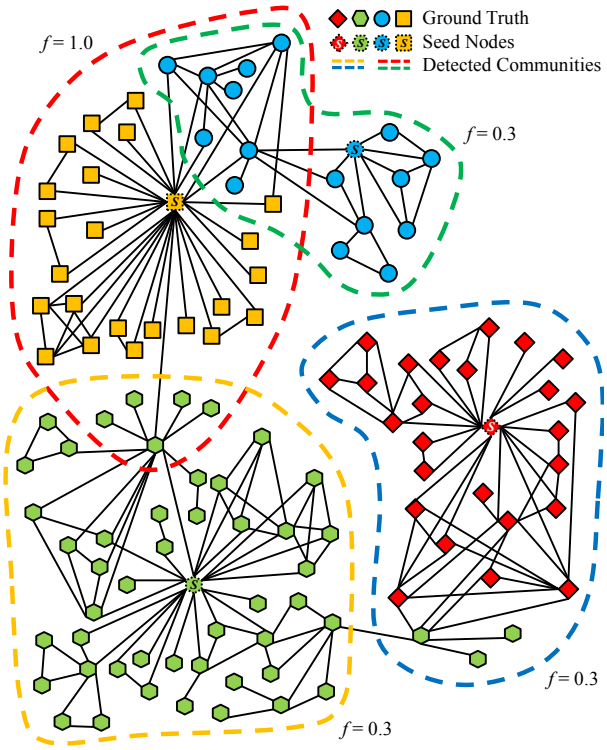

Figure 3: Santa Fe Institute Collaboration Network

classified as part of that community (Girvan & Newman 2002). These results show that while the precision of our algorithm does not fully match that of the ground truth, the detected communities are reasonable and meaningful groupings of the nodes.

While we achieved a high recall rate of 98.5% for the Santa Fe Institute dataset, the unsuccessful 1.5% is attributed to three (green, hexagon) nodes being mis-classified, as shown in Fig. 3. These three nodes were classified as part of the diamond community while the ground truth dictates that they belong to the hexagon community. However, an analysis of the actual topological links implies that these three nodes are better suited as members of the diamond community. Specifically, one of these nodes has four links to the diamond community but only one link to the hexagon community (while the other two misclassified nodes have only one link to this node). Based on this topological analysis, these nodes would be better classified as part of the diamond community.

Similarly, while we achieved a relatively high precision rate of 87.2% for the Santa Fe Institute dataset, the unsuccessful 12.8% was largely due to the misclassification of members of the hexagon and circle communities as part of the square community. Like the Zachary Karate Club dataset, almost half of these mis-classified (intermediate) nodes are directly linked to the seed node and thus should also be classified as part of the square community. For the remaining nodes, they are directly linked to these intermediate nodes and all of them have more links to these intermediate nodes than to other nodes. Therefore, they should also belong to the same community as these intermediate nodes (i.e. the square community).

6 Evaluation on YouTube Social Network

After evaluating our algorithm on three real-life social networks, we now evaluate it on the large-scale YouTube social network. The main challenge in this evaluation is the lack of an established ground truth of real-life communities, unlike the three real-life social networks previously evaluated. As such, we best approximate this ground truth using YouTube groups where users belonging to the same group are deemed to be in the same real-life community.

Table 1: Network Statistics of YouTube Dataset

Network Property	Detected Community			Control Group
	Min.	*Max.*	*Avg.*	
No. of Nodes	701	7241	1676	22180
YouTube Group Overlap	67.8%	93.7%	78.0%	N.A.
Avg. Degree of Links	3.54	13.97	8.25	8.66
Avg. Clustering Coeff.	0.14	0.36	0.28	0.13
Avg. Path Length	2.14	3.53	2.82	4.08
Diameter	4	7	5.4	11

As YouTube groups are an approximation of the ground truth of real-life communities, we further evaluate the communities detected by our algorithm using topological measures of average clustering coefficient, average path length, average degree and diameter. These are suitable metrics for evaluation as communities display typical characteristics of a high clustering coefficient and average degree with low average path length and diameter, especially when compared to the overall network.

6.1 Experiment Dataset and Control Group

In the YouTube social network dataset, there exists users who do not join any YouTube groups. Since YouTube groups serve as ground truth for our evaluation, we consider only users who have joined at least one YouTube group, in our experiments. Based on this criteria, there are 22,693 users who have joined at least one YouTube group. This set of users will be used to evaluate our algorithm as we are able to compare the detected communities with the actual YouTube groups they belong to.

As a control group for comparing network statistics, we selected the largest connected component from this set of 22,693 users (who have joined at least one YouTube group). This largest connected component comprises 22,180 users and would be used as the control group to compare against the detected communities (of our algorithm) in terms of average clustering coefficient, average path length, average degree and diameter. An ideal community detection algorithm would detect communities that exhibit a higher clustering coefficient, and shorter average path length and diameter compared to the overall network (i.e. our control group).

Similar to the selection of seed users for the three real-life social networks, we selected seed users for the YouTube dataset based on users with a high number of links. This selection criteria corresponds to the aim of our algorithm which is to detect communities centered at individuals of interest, such as influential or well-connected individuals. We first identify a set of users that are in the top 1% of the dataset, in terms of their number of links. From this set of users, we selected 10 users as the seed nodes for our algorithm. Using our algorithm, we then attempt to detect communities centered at each of these 10 users and measure the network statistics of the resulting 10 communities. In particular, we compare the average network statistics of these communities against that of the control group. Using the average result (from these 10 communities) avoids the effect of any

random or outlier results that may be unique to any particular community.

6.2 Comparison of Network Statistics

Table 1 shows the (minimum, maximum and average) network statistics of our detected communities compared to that of the control group. The YouTube group overlap measures how many other users in the detected community belong to the same YouTube group as the seed user. The high average result of 78% show that our algorithm is able to accurately detect communities where most of its users belong to the same YouTube group (as the seed user), an approximation of their real-life communities.

The YouTube group overlap result is not 100% due to the unique nature of YouTube groups where users who join such groups are producers/uploaders of videos related to that group. On the contrary, there are users who are only interested in viewing such videos but do not produce/upload videos. These users simply become friends with members of such groups and are able to be alerted about their new videos without having to join their YouTube groups. Even with such users, our algorithm is able to detect communities that are up to 93.7% accurate compared to the real-life communities

Despite the small average size of the detected communities, the average degree of links of these communities are very similar to that of the control group (differing only by 4.7%). This result shows that the detected communities comprise users who are well-connected among themselves (indicated by a high average degree of links), despite having an average community size that is less than 8% of the control group.

In addition to being well-connected, the detected communities are also highly cohesive based on an average clustering coefficient that is two times higher than that of the control group. Another observation is the lower average path length and diameter of the detected communities compared to that of the control group. A lower average path length and diameter means that nodes within these communities are able to reach each other in a smaller number of steps, which is also an indication of a cohesive and well-connected community.

Based on our approximation of ground truth, our proposed algorithm is able to detect communities that closely resemble real-life communities (up to 93.7%). The network statistics of these detected communities further illustrate the effectiveness of our algorithm. Specifically, the high clustering coefficient and average degree of links, and low average path length and diameter (of the detected community) indicate that our algorithm detects communities which are highly cohesive and well-connected, especially when compared to the control group.

7 Conclusion

We proposed a community detection algorithm for finding a community centered at an individual of interest, using an expanding ring search starting from this individual. At each progressive stage of the expanding ring search, we decide whether or not to add a user into this community based on our definition of a community. This definition is derived from the number of internal and external links of a user, coupled with an adjustable community strength factor. Our algorithm then continues iteratively until no further users can be added, thus resulting in the detected community comprising the list of added users.

In addition, we evaluated our algorithm on three real-life social networks to compare the detected communities to the ground truth of actual real-life communities. The results show that our algorithm is able to detect the actual communities at a high level of precision and recall rate of up to 97.6% and 100% respectively. Experiments on the YouTube social network also show that our algorithm is able to detect communities that closely resemble real-life communities (based on YouTube groups), up to an accuracy of 93.7%. Our evaluation of clustering coefficient, average path length, average degree of links and diameter also indicates that the detected communities are highly cohesive and well-connected.

8 Acknowledgments

Kwan Hui Lim was supported by the Australian Government, University of Western Australia (UWA) and School of Computer Science and Software Engineering (CSSE) under the International Postgraduate Research Scholarship, Australian Postgraduate Award, UWA CSSE Ad-hoc Top-up Scholarship and UWA Safety Net Top-Up Scholarship. The authors would also like to thank Amardeep Kaur for her comments on an earlier version of this manuscript.

References

Andersen, R., Chung, F. & Lang, K. (2006), Local graph partitioning using pagerank vectors, *in* 'Proc. of FOCS '06', pp. 475–486.

Andersen, R. & Lang, K. J. (2006), Communities from seed sets, *in* 'Proc. of WWW '06', pp. 223–232.

Arenas, A., Fernández, A. & Gómez, S. (2008), 'Analysis of the structure of complex networks at different resolution levels', *New Journal of Physics* **10**(5), 053039.

Clauset, A. (2005), 'Finding local community structure in networks', *Physical Review E* **72**(2), 026132.

Girvan, M. & Newman, M. E. J. (2002), 'Community structure in social and biological networks', *PNAS* **99**(12), 7821–7826.

Luo, F., Wang, J. Z. & Promislow, E. (2006), Exploring local community structures in large networks, *in* 'Proc. of WI '06', pp. 233–239.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E. & Dawson, S. M. (2003), 'The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting association.', *Behavioral Ecology and Sociobiology* **54**(4), 396–405.

Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005), 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature* **435**, 814–818.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. (2004), 'Defining and identifying communities in networks', *PNAS* **101**(9), 2658–2663.

Tang, L. & Liu, H. (2009), Relational learning via latent social dimensions, *in* 'Proc. of KDD '09', pp. 817–826.

Zachary, W. W. (1977), 'An information flow model for conflict and fission in small groups', *Journal of Anthropological Research* **33**(4), 452–473.