ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

A Study on Community Detection Based on Seed Nodes over Social Network

Monika Kasondra[1], Prof. Kamal Sutaria[2], Dipesh Joshi[3]

PG Student, Department of Computer Engineering, V.V.P. Engineering College, Rajkot, India[1]

Assistant Professor, Department of Computer Engineering, V.V.P. Engineering College, Rajkot, India[2]

Assistant Professor, Department of Computer Engineering, V.V.P. Engineering College, Rajkot, India[3]

ABSTRACT: Large-scale social networks emerged rapidly in the recent year. The social network has become more complex. The community is an important structure in the social network. So, community detection is required to define who is belonging to which community. In this paper, we use seed-centric approach for community detection. Seed node is a key node that has great influence in the social network. First, we find seed nodes over the network, then after, Second, using seed set, we detect communities over the social network.

KEYWORDS: detection, seed node, social network

I. INTRODUCTION

There are verities of the social networks, such as Facebook, Twitter, Amazone, LinkedIn, etc. We are growing with the information age. Now a day most of the people connected with this type of social network. With the development of the smartphones, more and more people log into their social networks through their smartphones and share text and multimedia information with their friends online[8].

The social network is usually modeled as graphs. A social network consists of a set of nodes along with edges connecting the nodes [1]. The nodes represent the object in social networks, such as people, commodities, etc. The edges represent the relationships between objects.

The community is an important structure in social networks. Basically, communities are set of nodes with higher edge density than the null model. In recent year, find seed nodes and how to use the seed nodes for community detection has become a hot topic in social network analysis and field of data mining.

In social network analysis, finding seed nodes have a wide range of applications. E.g., if we can dig out the most influential customers in marketing, then through the community structured by the seed nodes, a product brand can be rapidly promoted.

In this paper, we show the different methods, which describe how can detect communities based on the seed nodes over the social network.

II. COMMUNITY DETECTION

Detecting and evaluating the community structure of real-world graphs constitutes an essential task in the area of graph mining and social network analysis. The network contains many structures and tightly connected groups. Also referred as community, clusters, modules, etc.

The task of community detection is to find sets of nodes with lots of connections inside the sets and few edges outside the set.

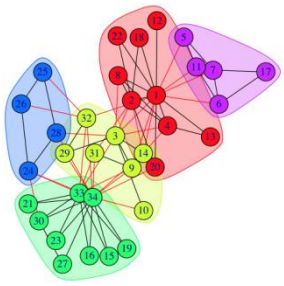
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016



Figure, 1. Communities in social network [9]

In figure 1, depicted the different communities over the social network. Different color defines different communities and number specify node id.

There are many available approaches for Community Detection:

1. *Group-based approach*
2. *Network-based approach*
3. *Propagation-based approach*
4. *Hierarchy-based approach*
5. *Seed-centric approach*

In this paper, we adopt a seed-centric approach to detect different communities over the social network.

III. LITERATURE SURVEY

Seeding phase [2]

In seeding phase, which consists of four phases: filtering phase, seeding phase, seed set expansion phase and propagation phase.

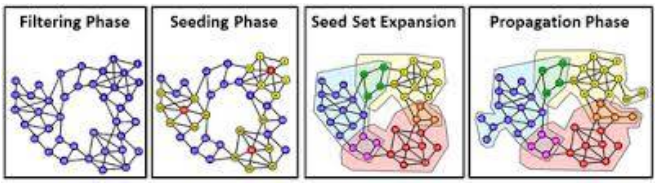


Figure 2. Seeding phase[3]

1. *Filtering phase*

In this phase, to remove regions of the graph that are trivially separable from the rest of the graph, so will not participate in overlapping.

The output of the filtering phase is the biconnected core graph where whiskers (subgraphs connected to the biconnected core) are filtered out. It removes regions of the graph that are clearly partitionable from the remainder. More importantly, there is no overlap between any of whiskers. This indicates that there is no need to apply overlapping community detection algorithm on the detached regions.

2. *Seeding phase*

After we get biconnected core graph, we find seeds in filtering phase. The goal of seeding strategy is to identify a diversity of vertices that lie within a cluster. The output of this phase is seed nodes over a graph.

3. *Seed set expansion*

In this phase, expand the seed set using a personalized PageRank clustering scheme. A personalized PageRank vector, followed by a sweep over all cuts induced by a vector, will identify a set of good conductance within the graph.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

The set identified via this procedure has a conductance that isn't too far away from the best conductance of any set containing that vertex.

4. *Propagation phase*

After seeding phase, propagation phase will apply. In this phase, once we get the personalized PageRank communities on biconnected core graph, we further expand each of the communities to the regions that we detached in the filtering phase. For each detached whisker connected via a bridge, we add that piece to all of the clusters that utilize the other vertex in the bridge. The output of this final phase of seed is different communities of the graph.

In [1] describe that they find seed node using the random walk. Seeds are finding based on the degree of the particular node. The degree of the node is defined as a total number of edges connected to that particular node. They decide some threshold value to find the seed nodes over the network. There are node r and t which is respect to simple node and threshold node.

If degree(r) is greater than a degree(t), and r is not any sub-region, then define r is a seed node.
If degree(r) is greater than a degree(t), but r is in sub-region, continue to find next node, which is not in any sub-region.
If degree(r) is greater than a degree(t), there is no any node then the procedure is finished.

In [4] they defined, Label Propagation algorithm does not require number and size of the community, so it is random, but not stable. To measure the importance of node, they use position probability of each node. Based on the importance of node, they define seed node.

They followed basically three steps:

1) Calculate the important nodes:

In this step, they first calculate transition matrix, based on this, they compute weights of the nodes in the graph.

2) Select community seeds and initialize core nodes:

If the weight of a node is greater than or equal to the threshold, then the node is core node. Otherwise, it is a free node. Then, first, core node directly added to set of community seed as a seed. Then jump to next core node. If the core node and existing community seeds have some common core neighbors, then sum up the common neighbors' weights. If the summation is greater than the weight of the core node, then add the core node to the sub-community centered at the community seed. If the summation is smaller or equal to the core node, then the core node become a new community seed.

3) Label Propagation and Community detection:

In this step, labels are initialized the same as the community seed id. Then nodes in the same community have the same label. For free nodes, their labels are equal to their ids. According to the algorithm, detect communities.

In [5], they find communities based on the degree of the vertex. The conductance of neighborhood communities shows similar behavior as network community profile computed with a personalized PageRank community detection method.

They defined cut-based measure using conductance parameter. The conductance of a cluster measures the probability that random edges leave the set.

They first select seed node based on an algorithm, after this, they detect communities using conductance parameter.

In this depicted figure, there is a small piece of the network, over which we have to find community. Black node defined seed node and green nodes are directly connected to the seed node. Then put cut over the community and calculate the conductance based on the following the formula:

$$\phi(S) = \frac{\text{cut}(S)}{\min\left(\text{vol}(S), \text{vol}(\bar{S})\right)}$$

(edges leaving the set)

(total edges in the set)

Here,
cut(S) is the total number of edges which leaving the set;
vol(S) is the sum of degrees of vertices in set S.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

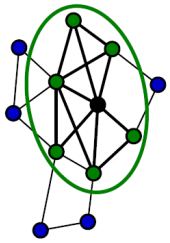
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

vol(¯S) is the sum of the degree of vertices remains in S.
Φ(S) is conductance of set S.



Cut(S) = 7
Vol(S) = 33
Vol(¯S) = 11
Φ(S) = 0.636

Based on this formula, they calculate the conductance of set S. after this they compare with other set's conductance and define the best community. Small conductance means good community.

There are main three ways of identifying a community with a good conductance score.

1) Fielder set
2) Personalized PageRank communities
3) Whisker communities.

One of the key problems with using personalized PageRank community algorithms is that finding a good set of seeds is not easy.

These communities represent the best of the neighborhood. We find that these locally minimal communities, of which there are much fewer than vertices in the graph, capture the local minimal in the network community profile plot. They can be enlarged using a local personalized PageRank community detection procedure. Afterward, the profile of this 'grown' neighborhood is strictly close to the profile of the PageRank communities when seeded with all vertices individually.

One explanation for the results with the PageRank seeds is that vertex neighborhoods from the base of any good community in the network.

In [2] they proposed an efficient overlapping community detection algorithm using a seed set expansion approach. They develop new seeding strategies for a Personalized PageRank scheme that optimizes the conductance community score. They give the key idea of the algorithm is to find good seeds, and then expand these seed sets using Personalized PageRank clustering procedure. Experimental results show that this seed set expansion approach outperforms other state-of-the-art overlapping community detection methods.

There consist of four type of phase like filtering phase, seeding phase, seed set expansion phase, propagation phase.

In filtering phase, we remove regions of the graph that are trivially separable from the rest of the graph, so will not participate in overlapping clustering. This phase is depicted in figure 2. In the seeding phase, we find seeds in the filtered graph, and in seed set expansion phase, we expand the seed sets using a PageRank scheme. Finally, in the propagation phase, we further expand the communities to the regions that were removed in the filtering phase.

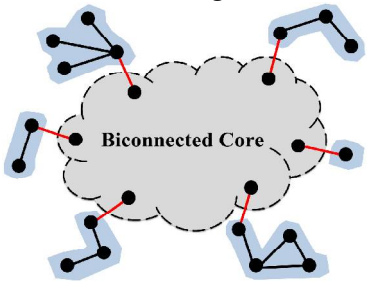


Figure 3: Biconnected core, whiskers, and bridges gray region indicates the biconnected core where vertices are densely connected to each other and blue components indicate whiskers.

In seeding phase there are two methods; seeding by Graclus Centers, seeding by Spread Hubs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

In spread hub algorithm, there are exploring hubs in the network. If there are several vertices whose degrees are the same, we take an independent set of those that are unmarked. This step may result in more than *k* seeds; however, the final number of returned seeds does not exceed the input *k* too much because there usually aren't too many high degree vertices.

Algorithm 1: Seeding by Spread Hubs

Input: graph G = (V; E), the number of seeds k.
Output: the seed set S.
1: Initialize S =Ø ;.
2: All vertices in V are unmarked.
3: while |S|< k do
4: Let T be the set of unmarked vertices with max degree.
5: for each t ϵ T do
6: if t is unmarked then
7: S = {t} ∪ S.
8: Mark t and its neighbours.
9: end if
10: end for
11: end while

In graclus centers algorithm, we get numbers of clusters. Then after we take each of the clusters as a seed it means that the center of a cluster id defined to the vertex, which is closest to the cluster centroid. It is described in step 7, of Algorithm 2. If there are several vertices whose distance is nearer to cluster, we include all of them.

Algorithm 2: Seeding by Graclus Centers

Input: graph G, the number of seeds k.
Output: the seed set S.
1: Compute exhaustive and non-overlapping clusters Ci
(i=1;…; k) on G.
2: Initialize S = Ø;.
3: for each cluster C_i do
4: for each vertex v ϵ C_i do
5: Compute dist(v, C_i) using (4).
6: end for
7: S = {fargmin dist(v, C_i)} ∪ S.
8: end for

In [6], they give the seeding algorithm which is parameter free, utilizes merely the local structure of the network, and identifies good seeds which span over the whole network. In order to find such seeds, our algorithm first computes similarity indices from local link prediction techniques to assign a similarity score to each node, and then a biased graph coloring algorithm is used to enhance the seed selection.

Experiments using large-scale real-world networks show that this algorithm is able to select good seeds which are then expanded into high-quality overlapping communities covering the vast majority of the nodes in the network using a personalized PageRank-based community detection algorithm. They also showed that using local seeding algorithm can reduce the execution time of community detection.

They defined two approaches of seed Algorithm like Global algorithm and Local algorithm, the comparison of this two seed algorithm are depicted as following:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Global Algorithm	Local Algorithm
Required Degree of All nodes in network	Required Degree of selected Good Seed nodes
It requires Minimum no. of seeds in advance	No prior Required
High Efficiency	Low Efficiency
Selecting the right number of seed is Hard	Selection of seed node is Easier

Table 1.comparison of Global and Local seed Algorithm

They select a node as a seed if it has the highest score among its neighbors. So the nodes with the highest local similarity score, which are expected to be good seeds, are assigned a specific color. Then the ties are broken at random so that no two adjacent nodes pick the same color. In the end, the nodes which received the specific color are selected as seeds.

The selected seeds are then expanded into overlapping communities using a personalized PageRank-based local community detection algorithm, which can be computed locally and is known to result in high-quality communities.

In [7], they evaluate a generic greedy algorithm which subsumes several previous efforts in the field. Experimental evaluation of multiple objectives functions and optimizations shows that the frequently proposed greedy approach is not adequate for large dataset.

As a more scalable alternative, we propose selSCAN, our adaptation of a global, density-based community detection algorithm. In a novel combination with algebraic distances on graphs, query times can be strongly reduced through preprocessing.

However, selSCANis very sensitive to the choice of numeric parameters, limiting its practicality. The random-walk-based PageRankNibble emerges from the comparison as the most successful candidate.

They first to target large complex networks in the order of 10^5 to 10^6 of edges, requiring scalable algorithms and implementations. After a review of previous work on the subject, algorithmic approaches are compared and classified. They identify one widely used approach which we call Greedy Community Expansion. They evaluate existing objective functions and optimizations.

In their experimental comparison, they include a reimplementation of the random-walk based PageRank-Nibble, representing an important class of approaches to the problem.

Furthermore, their main algorithmic innovation, they adapt a global algorithm to the selective scenario. A modification of density-based SCAN algorithm yields our variant selSCAN. The algorithm is generic with respect to a node distance measure and proposed algebraic distances an alternative to the original measure. The performance of the algorithm is experimentally evaluated with respect to accuracy, quality, community size, and running time.

IV. RESULT

Here the depicted result shows the run time of spread hubs and graclus centers seeding method on two different datasets like Amazone and DBLP [2].

We can see that run time of the spread hub is less than the graclus centers. Run time is defined in an hour.

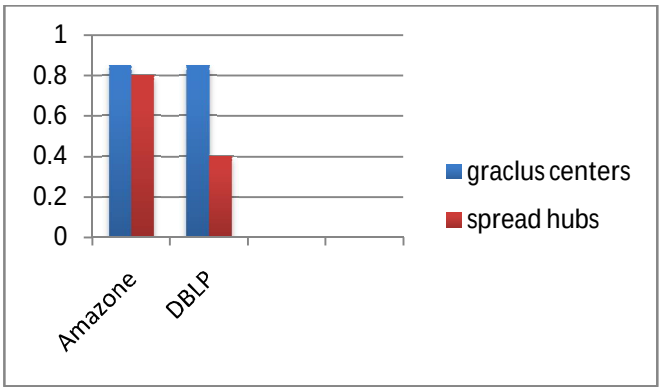
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016



In [1], they use several datasets to implement their algorithm. They use Karate club network, Dolphins network, American football network to analysis algorithm.

Zachary karate club

It is a popular network in terms of community structure. In the club, there is the administrator and the instructor. As a result, the administrator has left the club. Fig. 3 shows the communities which is detected by the algorithm. The result can be seen in table II.

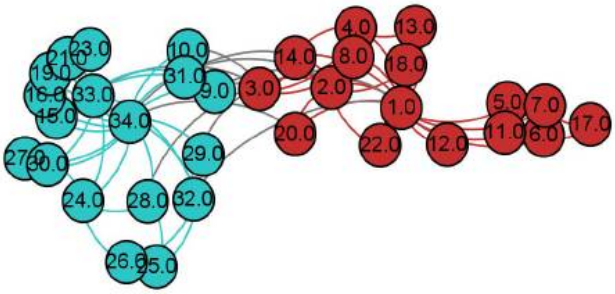


Fig. 4 Karate club network

Dolphins Network

In this network, it splits naturally into two large groups. Fig. 4 depicted the community structure detected by the described algorithm. The result can be seen in table II.

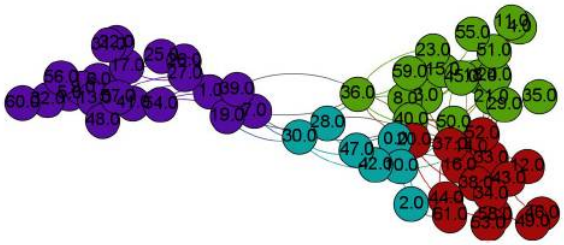


Fig. 5 Dolphin Network

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

US College Football

It is a network which representation of the schedule of regular season Fall 2000. In graph, teams are represent as vertices and season games between two team, which are connected to each other represented as edges. Fig 5 depicted community structure based on proposed algorithm. The result can see in table II.

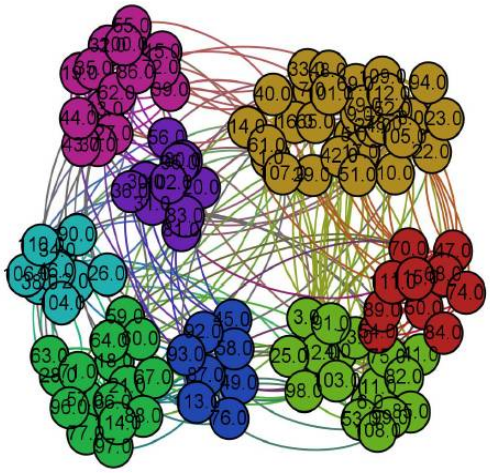


Fig. 6. American Football Network

Dataset	Community Numbers
Karate	2
Dolphins	4
Football	8

Table 2. Result of Experiment

V. CONCLUSION

Seed centric approach constitutes an emerging trend in the hot area of community detection in the social network. In this paper, we survey the different literature and understand different methodologies used for community detection. The result of this approach is defined good seed set, based on seeds after applying specific method or algorithm we gain good communities over the social network.

REFERENCES

[1] Chang Su, Yukun Wang, Lan Zhang,"A new method for community detection using seed nodes", IEEE International joint conferences on web intelligence and intelligent agent technology, 2014.
[2] Joyce JiyoungWhang, David F. Gleich, Inderjit S. Dhillon,"Overlapping community detection using seed set expansion", ACM-2013.
[3] https://www.google.co.in/search?q=seed+phase+in+social+network
[4] Su Chang, JiaXiaotao, XieXianzhong, Yu Yue,"A new random-walk based label propagation community detection algorithm", IEEE – International conferences on web intelligence and intelligent agent technology 2015.
[5] David F. Gleich, C. Seshadhri, "Vertex Neighborhoods, Low conductance cuts, and good seeds for local community methods", ACM-2012.
[6] FarnazMoradi, Tomas Olovsson, PhilippasTsigas, "A local seed selection algorithm for overlapping community detection", IEEE – International conference on advances in social networks analysis and mining – 2014.
[7] Christian L. Staudt, YassneMarrakchi, Henning Meyerhenke, "Detecting communities around seed nodes in complex networks", IEEE International Conference on Big Data, 2014.
[8] https://www.google.co.in/search?q=social+network
[9] https://www.google.co.in/search?q=community+detection

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

[10] Zhan Bu, Jiandong Wang, Zhengyou Xia, Chengcui Zhang, "community detection in very large dense network with parallel strategy", IEEE – 2013.
[11] Seyed Ahmad Moosavi, Mehrdad Jalali, "Community Detection in Online Social Network Using Actions of Users", IEEE Xplorer – 2014.
[12] RenjieWan ,JingyeCai, "Community Detection using an Optimized Label Propagation Algorithm", IEEE Xplorer -2014.
[13] Pili Hu, Wing Cheong Lau, " community classification in decentralized social network using local topological information. Globecom - 2014.
Fei Jiang, Yang Yang, Shuyuan Jin, Jin Xu, "Fast Search to Detect Communities by Truncated Inverse PageRank in Social Networks, IEEE International conference on Mobile service – 2015.