A Community Structure Detection Method based on Field Effect

Ru Zhang*, Zongwei Ren

School of Management, Harbin University of Commerce, Harbin, 150028, People’s Republic of China

Abstract

The information expressed by vertex is influenced by the environment in the semantic social networks, which is the result of natural and social factors. The field effect theory can explain the relationship between social environment and psychological environment. Therefore, a community discovery based on field effect is proposed by the perspective of pattern classification. The algorithm based on secondary classification ideology can be simply described as follows: the social networks are first divided into several original communities based on networks structure and the results of classification are assigned to each vertex of the networks as the label; secondly, the labels spread based on field effect that is computed by natural and social factors; ultimately the vertices which have same labels can be divided into a community. It is a process of secondary classification that can reduce uncertainty of the labels setting and randomness of labels propagation effectively. Experimental results show that the improved algorithm can get better information similarity based on field effect of vertex and make the inner node more closely.

Keywords: social networks; field effect; communities discovery; pattern classification

(Submitted on May 18, 2017; Revised on August 20, 2017; Accepted on September 15, 2017)

© 2017 Totem Publisher, Inc. All rights reserved.

1. Introduction

Today, people are expanding their social circle to the whole world by the Internet, which means people's mental live space is much larger than the “living space” and the social network of people are more complicated. But the human interaction has natural classification characteristics. There are some intensive sub-network in the social networks called a community. The community is very important for the social research and data mining work. How to detect the community is one of main problems that has been a subject that keeps attracting a great deal of interests [1,17].

In the past period of time, there are three stages for community discovery in social networks: hard communities division, overlapping communities division and semantic overlapping communities division, where the hard social networking community is to divide the social network into several disjoint sub-networks, such as GN (Givan-Newman) algorithm [3], FN(Fast Newman) algorithm [9] etc. The overlapping communities division is aim to divide the community structures that appear mutual tolerance relations, such as CPM (Clique percolation method) [11] and LFM (Lancichinetti Fortunato method) [5] and etc. Semantic overlapping community detection is based the information similarity expressed by vertex, which better reflect the community cohesion. Semantic information need to be derived by text, so the current semantic portioning algorithm communities mostly based on LDA (Latent Dirichlet allocation) model [2], such as SSN-LDA(Simple social network LDA) [5] and etc. However the expression of humans are always affected by the psychological field of their neighbor based on the research of human behavior, such as typical “herd behavior” or “anti-herding”, which means only the semantic information cannot decide the ownership of community especially in social website.

According to research we know that people formed communities based on common social goals, values, peer pressure and individual herd behavior. Therefore, natural and social factors are important for the human’s community ownership in the social networks. The online social networks are emphasized on meeting the psychological needs of people that the

* Corresponding author.
E-mail address: zhangru@hit.edu.cn.
psychological field effect is more effective in community structure detection of online social networks. We know that the field in natural involves following features: integrity, radiation, changeability, inertia and sensitivity, which would be used to detect the community. Therefore, a community detection algorithm based on field effect is proposed to improve the links of internal nodes and node’s semantic information similarity in community. The label propagation (LP) community discovery algorithm is used to model the field effect of neighbors [3] and the GN division results are taken as the description of social-physical attribute, which are assigned to the vertices as original labels. The field effects parameters are used to control the label propagate. The improved algorithm not only takes effects of networks structure into account, but also uses the people social attribute for the community discovery. Experiment results proved that it not only make up the shortcomings of traditional LP algorithm, but also improve the accuracy and stability of the community division results.

2. Related Theory

2.1. Community Discovery

It has been the consensus of researchers that social networks have the community structures. But what is the community? There is no precise mathematical definition. The generally description about community is the set of vertices in the networks, which has a closer connection among the vertices in the set and sparser connection out the set of vertices. In sociological perspective, the vertices have more similar nature within the community. In actual research, we need to divide the large and complex social networks such as Figure 1 into several portions, which are called community discovery. Figure 1 gives a generally description about community in social network. The social network is the set of vertices in the Internet, which has a closer connection among the vertices in the set and sparser connection out the set of vertices. In sociological perspective, the vertices have more similar nature within the community. In actual research, we need to divide the large and complex social networks such as Figure 1 into several portions, which are called community discovery. Many researchers have done a lot of work and most research mainly focused on the semantic information of the semantic social networks’ structure in the current time [7,8,12]. But reasons of community constructing are too sophisticated to explicated, the attribute of vertices and the weight of edges need to be considered [13,18].

Now we consider the community discovery on other scale: the graph with n vertices as input and the k classes vertices set as output. The community discovery also can be seen a classification problem. The classifier is defined by the characteristic of attribute. So the psychology of people and social attributes are more considered for community discovery in social networks, which is mentioned in the proposed algorithm.

2.2. Field Effect

Field theory in psychology was introduced by the Gestalt psychology from modern physics theories related to "field". It believed that the psychological phenomenon is not simply a collection of its constituent elements, but an integral component of a field, which interior has a dynamic of mutual relations. Psychologist Lewin developed the psychology field theory. He particularly stressed the following points:

- Field theory is the study of a power.
- Field theory emphasizes the study of mental processes.
- Field theory is look the environment and people as a whole.
- The psychological environment can be abstract as components.

Lewin applied the field of research in social psychology and drew the conclusion that group as an integral unit, where each member's mental activity are all in the others' psychological field of the group, which means the internal person’s area
is affected by the quasi-physical, quasi-concept, quasi-social and facts mapped from social environment [7]. Reference to Figure 2, the people’s behaviors were impacted by interpersonal and group relations which was very similar to the "linear force field". The behavior formula (1) describes the interaction and mutual influence, mutual restraint between the behavior of individual and population groups:

\[ B = F(PE) \]  

While \( B \) represents the behavior, \( F() \) refers to a function (also called a Law), \( P \) refers to a specific person, \( E \) refers to all of the interpretation of the psychological field of environment. \( P \) and \( E \) are two factors that constitute a living space groups. From equation (1), there are two forces interaction to decide the person’s action that one is the human’s ego state and the other is the psychological environment. Psychological field radiation characteristics are obvious, people know that psychological have very strong infectivity, such as: in a particular case, people excited, then the other entering, they will immediately have similar sentiments. People in the same community in a social network always have some similar individual performance, which is due to psychological field radiation caused in the community [14]. Therefore, we want to use radiation of psychology field to detect the structure of community.

![Figure 2. Living space pattern diagram](image)

### 2.3. Label Propagation Algorithm

LP is a heuristic algorithm that belongs to hierarchical clustering field, which is to study the community discovery from the perspective of social networks’ propagation characteristics and also takes networks structure into account. It has become an important algorithm for community discovery [13,18]. However, each vertex’s original label is randomly assigned and label updating is uncertain because of update strategy randomness impact in this algorithm. The superposition of random factors not only leads to a considerable different experimental result, but also makes the iterative process of the algorithm difficult to achieve a stable state.

The algorithm can be simply described as follows: for a given social networks \( G(S,F) \) with \( n \) vertices, which and \( S = \{s_1, s_2, ..., s_n\} \) indicate the vertices set and the label set respectively. For each \( \forall s_i \in S \), there are \( f_i \) indicates the label of vertex and \( N(s_i) = \{s_1, s_2, ..., s_k\} \) represents the set of \( k \) neighbors for \( s_i \). The label for each vertex defines the vertex’s community and original label is assigned randomly. The label of \( s_i \) is updated by labels of its connection point set in each asynchronous iterative. In classical LP algorithm the label is replaced by label that is the largest number in the neighbors set, such as (2):

\[ f_i = \arg \max_{f_j \in N(s_i)} \text{Num}(f_j) \]  

When some neighbor-labels present the same maximum value, the label \( s_i \) is replaced randomly. From these analyses, we can see that the mechanism of original label assigned and iterative label updating is full of random and uncertainty, which means the results of community discovering is erratic, such as Figure 3. There are two possible results for a networks discovery, which is the reason for the improved algorithm proposed. In order to avoid the label propagating randomly, the
radiation effects of psychological field such as pressure to conform, authority, intimacy and self-similarity are used to decrease the randomly label transforming.

2.4. Evaluation of Community Discovering

It is necessary to evaluate the algorithm performance better or worse by the quantitative indictor, which measures the result of community discovering. The modularity measure function Q proposed by Newman and Girvan in [4] is used to measure the quality of community discovery for the social networks which has unknown community structure. The definition for Q is as follows:

\[ Q = \sum_{s \in S} \left[ \frac{I_s}{E} - \left( \frac{2I_s + O_s}{2E} \right)^2 \right] \]

where \( I_s \) represent number of edges that two ends are all in community, \( O_s \) represent the number of edges that one end in community and the other out community, \( E \) is the total number of social networks. The \( Q \) is bigger, the quality and the modularity are better.

In order to evaluate the semantic information similarity of inner node, the \( SQ \) function is used that proposed in [6]:

\[ SQ = \frac{1}{X} \sum_{i} \sum_{j \in S} U(v_i, v_j) [A_{i,j} - \frac{R_i R_j}{X}] \]

where \( X \) is the total degree of social network, \( A \) is the adjacency matrix, the \( R_i \) is the degree of vertex of node \( i \), \( v_i \) is the attribute of vertex of node \( i \) and the \( U(v_i, v_j) \) is the similarity measure that is defined as (5):

\[ U(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i| |v_j|} \]

3. The Label Propagation Algorithm Based on Field Effect
3.1. The Idea of Algorithm

Internet technology has greatly weakened the geography effect of the community and made it possible for cross-regional connectivity. Common social goals and values lay the foundation for an online community. But the pressure to conform often makes the individual’s motives mix with community targets. As a result, an individual's attitude and purpose are homogenized with community and spontaneously strengthen this standard. These psychological interactions with the external environment can use the field effect to describe. The field in natural involves following features: integrity, radiation, changeability, inertia and sensitivity, which can be quantify as space factors and time factors. If the social networks were
viewed as unweight graphs, the number of vertexes and structure of subnet in social networks can be used as space factors, and the label propagating can be used to describe the radiation effect by psychological field with the time change. But the current label propagation is based on an unweight graph. It means the algorithm only considers the factor of the “with” or “without” relationship between people. From a sociological point of view, there are some psychological environment factors that affect the community formation, such as the intimacy, cognitive consistency and authority et al. So we should expand the dimension of community structure and rebuild the label update mechanism. If we look at community discovery as the classification problem, it is a process of two classifications, which one classifier is psychological attribute (psychological environment) in communication and the other is the physical relationship (social environment) in social networks. According to these analyses, we define the label decision function as (6):

$$f_{n}^{\alpha} = (1 - \alpha)\Omega(s) + \alpha \overline{f}_{s}$$ (6)

Where $\overline{f}_{s}$ that represents the physical relationship in social networks is the original label for the vertex $s$, $\Omega(s)$ represents the label affected by psychological field, which can be explained as psychological attribute and control the vertices’ label updating; $\alpha \in [0, 1]$ is an adjust parameter, which is used to control the balance between the physical relationship and psychological attribute. When $\alpha = 1$, it means only considered the physical relationship in the process of label propagating. When $\alpha = 0$, it means label propagating only according to the psychological attribute. So we can not only change the expression of psychological attribute and physical relationship in social networks on the basis of background according equation (6), but can also control their relationship through adjust parameter $\alpha$. It can improve the generalization capability of label propagation algorithm.

3.2. Algorithm Description

According to (6), the update labels are influenced by two factors: the original label representing the social environment and the update mechanism influenced by pressure to conform. If looking at community discovery as the classification problem, the original labels for the vertices can be decided by a networks classifier and the route of label propagating is influence by the psychological attribute. Since “convergence” is a promotion for community formatting in sociological perspective, the psychological attribute affected by psychological field can be described as the local similarity in intimacy, cognitive consistency and authority. In this paper, we used the linear fitting algorithm to describe the effect of psychological field, $\Omega(s)$ is defined as follow:

$$\Omega(s) = \sum_{j, j \in N(s)} w_{ij} f_{j}$$ (7)

where $w_{ij}$ is the weight of psychological field of neighbour vertex $s_{j}$ and must satisfy the follow conditions.

$$\left\{ \begin{array}{l} \min_{w_{ij}} \sum_{j, j \in N(s)} w_{ij} s_{j} \leq 1 \quad w_{ij} > 0 \\ \sum_{j, j \in N(s)} w_{ij} s_{j} > 1 \end{array} \right.$$ (8)

From the (8), we can see the $w_{ij}$ is mainly affected by the similarity of node. In this paper the Euclidean distance used to compute the similarity of nodes. In this paper, $\overline{f}_{s}$ is defined by construct of social networks, which is assigned by the Girvan-Newman (GN) community discovery algorithm:

$$\overline{f}_{s} = GN(s)$$ (9)

So the process of community structure detection algorithm is designed as Figure 4.
There are two classification processes in Figure 4. The purpose of preliminary classification for vertexes is to reduce the uncertainty of label propagation so it needs to consider the overall structure of social networks. At same time the classification results must be instructive. In this paper, the GN community discovery algorithm is used as the preliminary classification to get the original labels for the vertexes of social networks. In order to reduce the randomness of the label updating, linear fitting is also used to computing the local similarity. The details of algorithm are described as follows:

- Get original labels based on GN algorithm.

  GN algorithm is a heuristic algorithm and uses the L to control the degree of community segmentation [16]. It is important to control the splitting for the GN algorithm, which means the size of effective community. In this paper, we assume the control parameter L is defined by the minimum size of effective community \( N_{\text{min}} \) and size of networks \( \text{Num}(G) \):

\[
L = \frac{\text{Num}(G)}{N_{\text{min}}}
\]

According the process of algorithm in , the results of GN are assigned to the vertices as the original labels.

- Label propagation.

From (4),(5),(7), we can know the label updating mechanism as follows:

\[
f^{n+1}_{x_i} = (1 - \alpha) \sum_{j, x_j \in N(x_i)} w_{x_i x_j} f^{n}_{x_j} + \alpha \bar{f}_{x_i}
\]

### The Label Propagation Algorithm Based on Field Effect

**Input:** social networks \( G(S, E) \) and control parameter \( L \)

**Output:** the vertices of community \( \text{GN}(S) = \{1, 2, \ldots, L\} \)

1. Computing the betweenness of edge in \( E \);
2. Remove the edge that has the biggest value of betweenness;
3. Recalculate the betweenness of edge that is influenced by removed edge;
4. Repeat step b) until degenerate \( L \) community;
5. end

- Proof of convergence.

In order to ensure the community discovery results is stable, the iterative process must be convergence. \( Z(f_x) \) is defined as follow to represent the change rate of vertexes’ label in each iteration.

\[
Z(f_x) = \sum_{i=1}^{n} \sum_{j, x_j \in N(x_i)} w_{x_i x_j} [f_{x_i} - f_{x_j}]^2 + (\frac{1}{\alpha} - 1) \sum_{i=1}^{n} [f_{x_i} - \bar{f}_{x_i}]^2
\]
\[ f_S = \{f_s, \ldots, f_s\} \] represents the label set of all vertices in social networks. If \( Z(f_s) \) has a minimum value, it is the proof of label updating convergence. The minimum value of \( Z(f_s) \) can be compute as follow:

\[
\frac{\partial Z(f_s)}{\partial f_s} = [(I - W) + (I - W)^T]f_s + 2\left( 1 - \frac{1}{\alpha} \right)(f_s - f_s) = 0 \tag{13}
\]

Which

\[
[(I - W) + (I - W)^T]f_s \approx 2[(I - W)]f_s \tag{14}
\]

Bring (11) in (12) then get solution:

\[
f_s = (1 - \alpha)(I - \alpha W)^{-1}f_s \tag{15}
\]

Then it is proved that the proposed algorithm can obtain the stable community discovery results.

4. Simulation and Analysis of Experimental

Three data sets are selected to validate effectiveness of algorithm, which contains two benchmark data sets and a grabbing data set. They are used independently. The classic community detection algorithms include GN, FN, LP, LPAm, LFM and etc. The paper chooses the five algorithms as the comparison.

4.1. Experiments Based on Benchmark Datasets

This paper selects two benchmark data sets, which are ACF (American College Football) dataset [6] and ZKC (Zachary’s Karate Club) dataset [15]. The ZKC data set has 34 vertices and 78 edges consisting of two factions. In order to validate the local similarity in ZKC and ACF social networks, we generate two-dimensional random number for each vertex as vertex semantic information. ACF is a networks dataset, which contains the relationship data between American college football leagues in the 2000 regular season race. There are 115 teams that are divided into 12 groups. The networks comprise of 115 vertices, 616 edges, which some details can be seen in Table 1. These two data sets are taken from the real world and most often used in relation studies.

<table>
<thead>
<tr>
<th>Team</th>
<th>Size</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Atlantic Coast</td>
<td>9</td>
<td>5 33 37 45 89 103 105 109</td>
</tr>
<tr>
<td>2. Big East</td>
<td>8</td>
<td>19 29 30 35 55 79 94 101</td>
</tr>
<tr>
<td>3. Big Ten</td>
<td>10</td>
<td>2 6 13 15 32 39 47 60 64 100 106</td>
</tr>
<tr>
<td>4. Big Twelve</td>
<td>11</td>
<td>3 5 10 40 52 72 74 81 84 98 102 107</td>
</tr>
<tr>
<td>5. Conference USA</td>
<td>10</td>
<td>44 48 57 66 75 86 91 92 110 112</td>
</tr>
<tr>
<td>6. Independents</td>
<td>5</td>
<td>36 42 80 82 90</td>
</tr>
<tr>
<td>7. Mid-American</td>
<td>11</td>
<td>12 14 18 26 31 34 38 43 54 61 71 85 99</td>
</tr>
<tr>
<td>8. Mountain West</td>
<td>8</td>
<td>0 4 9 16 23 41 93 104</td>
</tr>
<tr>
<td>9. Pacific Ten</td>
<td>10</td>
<td>7 8 21 22 51 68 77 78 108 111</td>
</tr>
<tr>
<td>10. Southeastern</td>
<td>12</td>
<td>17 20 27 56 62 65 70 76 87 95 96 113</td>
</tr>
<tr>
<td>11. Sun Belt</td>
<td>7</td>
<td>11 24 50 59 63 69 97</td>
</tr>
<tr>
<td>12. Western Athletic</td>
<td>10</td>
<td>28 46 49 53 58 67 73 83 88 114</td>
</tr>
</tbody>
</table>

We verified the proposed algorithm on two benchmark data sets. Since the number of divisions of original community is related with \( n_{\text{min}} \), which is the size of minimum effective community, the control parameter \( L \) of GN is defined by \( n_{\text{min}} \). For the ZKC dataset, \( n_{\text{min}} \geq 10 \), so \( L = 3 \), for the ACF dataset, \( n_{\text{min}} \geq 5 \) which can be seen in fig 1, so \( L = 20 \) is use as indcitor to evaluate the performance of algorithm. Figure 5 and Figure 6 give the comparison results of algorithms.

Table 2 shows the values of evaluation measure. We can see that the algorithm based on field effect improved the similarity of communities and got better modularity.
A Community Structure Detection Method based on Field Effect

Figure 5. Community detection results of ACF

Figure 6. Community detection results of ZKC

Table 2. Comparison results of Algorithm based on ACF and ZKC

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$Q_{ACF}$</th>
<th>$SQ_{ACF}$</th>
<th>$Q_{ZCK}$</th>
<th>$SQ_{ZCK}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP</td>
<td>0.561</td>
<td>0.317</td>
<td>0.493</td>
<td>0.3202</td>
</tr>
<tr>
<td>GN</td>
<td>0.423</td>
<td>0.347</td>
<td>0.468</td>
<td>0.329</td>
</tr>
<tr>
<td>FN</td>
<td>0.394</td>
<td>0.339</td>
<td>0.334</td>
<td>0.291</td>
</tr>
<tr>
<td>LPAm</td>
<td>0.351</td>
<td>0.321</td>
<td>0.309</td>
<td>0.327</td>
</tr>
<tr>
<td>LFM</td>
<td>0.282</td>
<td>0.282</td>
<td>0.291</td>
<td>0.3172</td>
</tr>
<tr>
<td>Improved algorithm</td>
<td>0.59</td>
<td>0.391</td>
<td>0.52</td>
<td>0.375</td>
</tr>
</tbody>
</table>
4.2. Experiments Based on Grabbing Datasets

The grabbing dataset comes from the Arxiv website provided by Leskovec et al in [10]. It contains the proceedings data of cooperation on general relativity and quantum cosmology crawl topics. There are 5242 vertices represented by authors, 14,484 edges indicated by cooperation and two attributes described by paper title and Keywords for each vertex. In order to make the experiment more quickly, we used list scheduling in our distributed computing [17,18]. The communities of dataset are unknown, so we evaluate the performance of algorithm through four dimensions: the number of effective community and ineffective community, the mean size of the community and the modular measurement Q. At first, we also must determine the control parameters of GN. If we assume that a community size less than or equal to 10 is considered an invalid community, namely \( N_{\text{min}} > 10 \), we take \( L = 500 \) and set step as 0.05 in experiment. Results shown in Figure 7.

![Figure 7. Community discovery in Arxiv datasets](image)

From Figure 7, we can see that the proposed algorithm can get best results when \( \alpha = 0.45 \).

Table 2 shows the comparison of algorithm performance and that the improve algorithm can get best performance.

5. Conclusions

To improve the stability and accuracy of the results of community division, an improved label propagation algorithm based on field effect is proposed by analyzing the main factors affecting structure of community. The experimental results proved the effectiveness and stability of algorithm. At same time, a label decision frame is constructed, which give a systemic description for label updating and can help the others to research further in community discovery.

Acknowledgements

This project is supported by the National Natural Science Foundation of China (Grant No.71371061), Heilongjiang planning research project of philosophy and Social Sciences (Grant No. 17JLE333) and Key Laboratory of Data Science and Application of Hebei Province.

References

8. A. Lancichinetti and S. Fortunato, “ Consensus Clustering in Complex Networks”, Scientific Reports, vol. 2, no. 13, pp. 336-
13. X. Xiong, X. Niu, G. Zhou, K. Xu, and Y. Z. Huang, “Microgroup Mining on TSina via Networks Structure and User Attribute,” in Proc. of the 7th Int’l Conf. on Advanced Data Mining and Applications (ICADMA), pp. 138-151, Berlin, German, December 2011

Ru Zhang graduated from the School of Computer Science and Technology, Harbin Institute of Technology, for the degree of Master and Ph. D from 2002 to 2013. Now she is a lecturer of the School of Management in Harbin University of Commerce, Harbin, China. Her current research interests include computer networks and security, machine learning, and intelligent information processing.

Zongwei Ren received the Bachelor degree from Daqing Petroleum Institute, Daqing, China, in 2002, the Master degree and the Ph.D. degree from Harbin Institute of Technology. Now he is a professor of the School of Management, and the Director of the Logistics Engineering, Harbin University of Commerce, Harbin, China. He is a senior member of China Computer Federation and a member of IEEE. His current research interests include Wisdom medical, machine learning, and intelligent information processing.