A New Multi-view and Controllable Community-uncovering Algorithm

Huifeng LONG
Hunan City University, Yiyang, Hunan, 413000, China

Abstract: This paper introduces a new multi-view and controllable community-uncovering algorithm, an achievement of improving PageRank algorithm and Spin-glass model, which can avoid the overlapping community structure in the process of detecting communities by means of other algorithms and also helps to improve the usual community-expansion model. The process of uncovering communities by the introduced algorithms can be divided into three steps: first, identifying the nuclear one among nodes ranked by the advanced PageRank algorithms; Second, through using multi-view recognition modularity provided by Potts spin-glass model, optimizing the expansion model of local community that is found by applying the improved Iterative Greedy algorithm to eliminate the traditional modularity's limits in the resolution limit and the following negative effects. Finally, grasping the overlapping structure and notes carefully . By analyzing and comparing the two results of respectively using PRSGMFCA and traditional technical schemes in both computer simulation network and the real network , it proves that the former enjoys stronger stability and higher accuracy than the latter , and its computation complexity is also acceptable.

Keywords: Improved PageRank algorithms, Spin-glass model, Multi-view recognition and controllable, Iterative Greedy algorithms, Community-uncovering

1. Introduction

Social network reflects the social developing law. Analyzing its relevant activities and their laws has both important theoretical and practical meaning in promoting our social environment's healthy and continuous development, effectively resolving urgent events and shaping a good social morality.

Web (Web Community) refers to "a group of people connected by and communicating with each other in the network. They learn a little about each other by sharing some knowledge and information and care about each other like friends."Web community-uncovering can help people to conduct service management and social search, send social security alarm and introduce personalized products. Microblog is a simple form of Web Community, in which members are obviously characterized by "Homophily" and the new social communication model-"following" makes it possible for them to build link relations with each other without authorization. In microblog, members' interests and hobbies along with JSP links reflecting the social connection with each other plays a crucial role in promoting the communication between members. Facebook, the world's largest social network, originally is composed of several groups of students who study in the same school and enjoy the same interests. With more and more people, who are always the classmates or relatives of those original ones, becoming new members, Facebook finally develops into a Web Community, in which groups are classified by members' interests and members can build their own social connection. In recent years, scientists at home and abroad have made enormous efforts in Web Community-uncovering study and provided several algorithms. In HITS Algorithm, created by Kleinberg and others, Web Community is treated as a nuclear, linking to a central page and constituted by enormous authority pages. By using this algorithm we can find topics being ordered in a tree hierarchy, and this structure reflects the relationship between community and sub-community. In PageRank Algorithm , searching engine is assigned a weighting adaptive to all pages, which can automatically show us the importance of different pages by providing us a importance-dominated page rank after verifying them in a recursive way. SPB (shortest path betweens) is a community-uncovering algorithm developed from the network flow technology, which finds community by identifying the space without being limited by the shortest path between two random nodes with the largest network flow.

LDA (latent Dirichlet allocation) is a classic topic model, which applies a binary random variable between two hypothetically independent documents to reflect their underlying relations; Reference[9] points out that URL's markers reflect user's interests and those with high co-occurrence rate compose URL marker collections, showing different interests themes. So, people can uncover communities with the same interests theme by finding the same marker collections. Lin and his colleagues provide a algorithm, which helps people uncover communities by

analyzing users' perception. Lin believes that the forming of a community is determined by users' behavior and such behavior must be interactive. All these algorithms above cannot reflect microblog members not only have theme connection but also social connection, because they just simply analyze link relations, identify different themes or research users' behavior. By contrast, combining the results of researching trust relationships in e-commerce with users' social information and their interactive behavior, Joe and his colleagues create a new way to uncover Web Community: a network users' trust degree algorithm . But the confidentiality of most microblog users' personal information makes it impossible to get enough users' registration information in the process of calculating their trust degree, so this algorithm is not practical.

Now there are some studies about the local communities in microblog. Reference[12] introduces TwitterRank algorithm, created by analyzing Twitter users' homophily and advancing PageRank web page's weight, which because its ability to calculate users' weight can uncover the most dynamic user group in the microblog. Mr Wu and his colleagues give us a new algorithm-XinRank algorithm developed from the advanced TwitterRank algorithm. And then they use their algorithm to rank the users of Sina micrlog, China, according to their importance. Community reflects the characteristics of network users' behavior and the correlation between two users. Researching network community is crucial to get a clear eye on network function and its structure, and also makes it easier for people to detect the relationship of network participants. In Internet social apps Web Community is very common. For example, using community-uncovering algorithm in microblog can improve the effects of advertising campaign; and e-commercial users can use community-uncovering algorithm to build a more stable and precise recommended system, because by using this algorithm to study the records of searching, they can do a research and make a conclusion about the users' behavior, then as a result providing a more satisfactory searching results to the app users. However community-uncovering technology still has many defects such as community localization and community overlap. To solve the problems above, some scientists have gave us some algorithms such as CPM [4], GCE[5], LFM[6], MONC[7].

1.1. The Target Function of the Local Community Expansion

In this paper, Hamiltonian, explained clearly in his study by Reichardt, is the target function of the local community expansion, represented by formula 1 below:

$$\begin{aligned} H(C) = &-\sum_{i\neq j} a_{ij}A_{ij}\delta(C_i,C_j) \\ &+\sum_{i\neq j} b_{ij}(1-A_{ij})\delta(C_i,C_j) \\ &+\sum_{i\neq j} c_{ij}A_{ij}\left[1-\delta(C_i,C_j)\right] \\ &-\sum_{i\neq j} d_{ij}(1-A_{ij})\left[1-\delta(C_i,C_j)\right] \end{aligned} \quad (1)$$

In the formula above, "m" refers to the total number of margins and "c" as collection of community. More is the number of margins between two spinning nodes with the same direction and speed, fewer is the number of margins between two different spinning nodes and the higher is Hamiltonian. a_{ij}, b_{ij}, c_{ij} or d_{ij} respectively represents a weight. And here, the spinning nodes with the same direction and speed are in the same community. When $a_{ij} = c_{ij} = 1-\gamma\frac{d_i d_j}{2m}, b_{ij} = d_{ij} = \gamma\frac{d_i d_j}{2m}$,Hamiltonian can continue to be calculated in the formula 2 below:

$$H(C) = -\sum_{i\neq j}(A_{ij} - \gamma\frac{d_i d_j}{2m})\delta(C_i,C_j) \quad (2)$$

In this paper"2H(C)"is the target function of local community-uncovering.

1.2. Using the Advanced PageRank Algorithm to Choose the Nuclear Node

Here, because a local community is treated as a collection of a potential leader and its followers, uncovering the nuclear node, which is regarded as the seed of local community, can make local community-uncovering easier and more accurate.

Here, the advanced PageRank algorithm is chosen as the measurement, which is used to rank all network nodes. By this way we can find the nuclear node. And in order to make PageRank algorithm applicable in the ranking of nodes in the indirect graph, we must optimize the algorithm.

Definition 1 (Centrality). "G=(V, E, w)" represents an indirect weighted network, "w" weighting function, PRcen (i) the centrality of node and it can be calculated by the formula 3 bellow:

$$PRcen(i) = c\sum_{j\neq i} PRcen(j)\frac{w_{ji}}{\Sigma_{k\in adj[i]}w_{ik}} + \frac{(1-c)}{N} \quad (3)$$

In the formula above, "N" represents the number of G's nodes; "c" is a constant and its range can be represented by $.c \in (0,1).; adj[i]$ refers to the collection of all approximal points of "i". Here the limit of the range of "c" can not only accelerate the convergence of the algorithm, but also is helpful for the convergence caused by isolated nodes. In general, "c" is about 0.85. And "τ "is usually changed with the specific need. What is described above is the traditional PageRank algorithm, which is optimized here, aiming to enable it to use the link information of a

indirect graph network and to represent "PRcen"more accurately and timely. In the advanced algorithm, recursion is fitted in the specific condition here, and the centrality of every node is determined by the centrality of its approximal node. And every node's weight is calculated by the sum of weight of node and its approximate node multipling a certain proportion.

In other words, "Prcen" is used to described the centrality. The higher the centrality of a node is, the higher the centrality of its approximal node is. Weight is used to represent the strength of the junction between two nodes. The higher the weight is, the higher the relative centrality is. The formula for calculating it is as below:

Algorithm 1:

$Input: \tau, c$
$Output: PRcen$
1.$\forall i, PRcen_s[i] \leftarrow 1/N$
2.$while(res > \tau) do$
3.$\forall i, PRcen_t[i] \leftarrow 0$
4.$for\ all\ i\ do$
5.$for\ all\ j\ do$
6.$PRcen_t[j] \leftarrow PRcen_t[j] + PRcen_s[i] \times (w_{ij} / \Sigma j \in_{adj[i]} w_{ij})$
7.$end\ for$
8.$end\ for$
9.$\forall i, PRcen_t[i] \leftarrow c \times PRcen_t[j] + (1-c) \times (1/N)$
10.$res \leftarrow \| PRcen_s - PRcen_t \|$
11.$PRcen_s \leftarrow PRcen_t$
12.$end\ while$

The first several nodes in the rank are possible to be chosen as the nuclear node. Usually initialization is to a large extent decided the algorithm for dividing the network, so in algorithm 1 the algorithm can finish expansion rapidly if starting calculating with a correct nuclear node, but by contrast, if starting with the wrong nuclear node the result would be repeated, iterative and invalid. In order to avoid the problem algorithm 1 is an unusual community-uncovering technology, which has a beneficial effect on initialized enumeration and a high stability. The algorithm also largely reduces the possibility of emerging redundant community.

1.3. The Experimental Performance Analysis of the Simulated Network

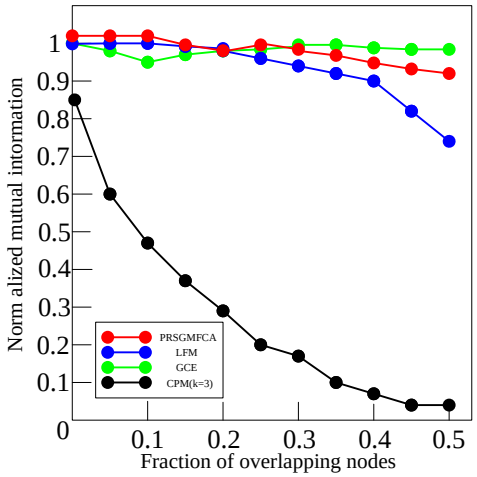
We use the LFR-benchmark method to build the computer simulated network. The building process involves the distribution characteristic between the node and the community and on the other side it will devote to the hierarchy and overlap among communities. During the experimental process, we set the network number as 50 based on these parameter settings specified in Table 1 while we use the NMI (Normalized Mutual Information) to evaluate the experiment results described in Diagram 1(Diagram (a) and (b) represents the results of the use of PRSGMFCA, CPM and GCE in the simulated network G1 and G2, while Diagram (c) and (d) represents the running results of the PRSGMFCA, FN and CPM in small community network G3 and G4. The error line shows the average error value when the algorithm runs 50 times).After comparing (a) and (b), we find that in a network (m=0.1) with clear structure most NMI of PRSGMFCA exceed 0.9 and have an excellent stability, while the stability of LFM is not good, for it chooses seeds so arbitrary that the results are obvious different. However, the network G2 (m=0.3) looks sparse and the internal margins of it are not denser than that in G1 while the margins between communities increase obviously. It also means that when a community structure is not so clear, the effectiveness of its algorithm will weaken along with the increasing of overlapping nodes. And CPM algorithm is easy to be affected by k (referred as the size of clique here), so the result calculated by it is not perfect. However, when the on value stays at a large number, the detecting of community structure from crowded overlapping communities performs well. But the use of improved PRSGMFCA is superior to other algorithm (CPM, LFM and GCE), which means it is very necessary to improve the strategies of choosing the original nodes and relative algorithms. On the other side, from the Diagram (c) and (d) we find the FN algorithm can not be detected structure effectively when the sizes of community G3 and G4 are lower than the recognized lower limit of Newman modularity, that is $\sqrt{2m} \approx 141$.And we can also noticed that for most parameters the results calculated by PRSGMFCA are about 0.9, which means that this improved model is not restricted by the resolution limit.

1.4. The Experimental Performance Analysis of the Real Network

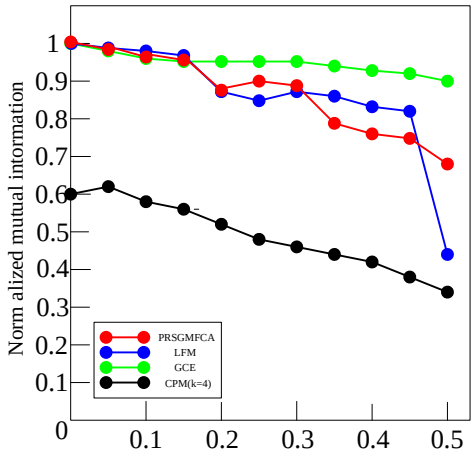
In addition to the condition discussed above, we also need to find some real network data to test the effectiveness of improved model. So we introduce several real experimental networks, such as Zachary's karate club, Dolphins' social network, Books about US politics Books and American College football described in Table 2. Meanwhile, the distribution of communities will be assessed by expanding modularity (EQ) described in.

Table 1. Parameter Settings of the Simulated Network.

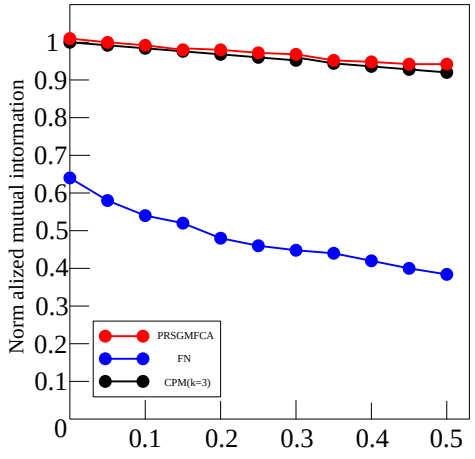
Network	N	k	$k_{\max}$	C_{mix}	$C_{\max}$	τ_1	τ_2	μ	O_n	O_m
G1	1000	20	50	20	100	-2	-1	0.1	0-500	2
G2	1000	20	50	20	100	-2	-1	0.3	0-500	2
G3	10000	20	50	20	100	-2	-1	0.1	0-500	2
G4	10000	20	50	20	100	-2	-1	0.3	0-500	2



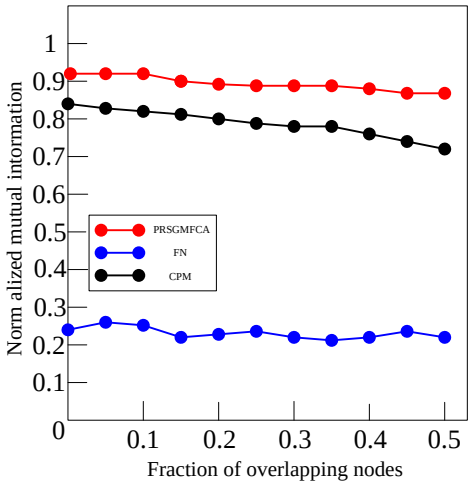
(A)Experimental result diagram of network G1.



(B)Experimental result diagram of network G2.



(C)Experimental result diagram of network G3.



(D)Experimental result diagram of network G4.

Figure 1. Experimental Result Diagram of Simulated Networks.

$$EQ = \frac{1}{2m}\sum_{C}\sum_{i,j\in C_k}\frac{1}{O_i O_j}[A_{ij} - \frac{k_i k_j}{2m}] \quad (4)$$

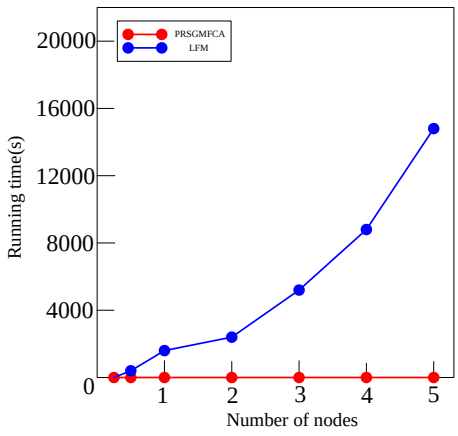
In A_{ij} represents any element of network adjacency matrix. If i links with j, then $A_{ij} = 1$ .If not, $A_{ij} = 0$. $m = \frac{1}{2}\sum_{ij} A_{ij}$ represents the total margins and $k_i = \sum_i A_{ij}$ shows the degree of i while the number of community i attached is defined as O_i .When a node belongs to no more than a community, we know EQ equaling to Q. However, when all nodes belong to the same community, $EQ = 0$. More obviously, the higher the value of EQ is, the more scientific the overlapping structure is.

In lists the results calculated by PRSGMFCA, CPM, LFM and GCE used in the real network. The EQ of PRSGMFCA is larger than that of CPM and LFM while the GCE does well in the Network football. We should know when $\gamma = 1$, we can use PRSGMFCA to find 2 communities in network karate; but when EQ ($\gamma = 1.2$) is the largest value, 4 communities can be found. We almost can't find 2 communities but can find 4 communities which are calculated repeatedly in LFM, for we choose seed nodes at random. PRSGMFCA has stable results while LFM does not run enough stably to get rid of the negative effect from the wrong original node which will result in many independent single-node node which will result in many independent single-node communities combined by many independent nodes. When we set the parameter k (the value of clique) ,we can get good results from CPM. However, when the network is not too dense or k is not suitable, the result will be bad.

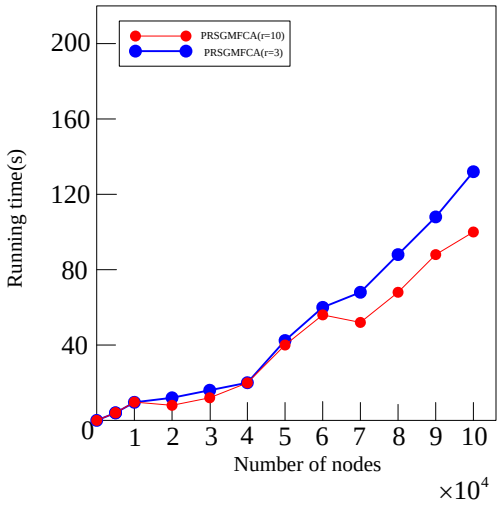
When we use the defaults, we find that the results from GCE rank between the results from PRSGMFCA and the results from LFM. But in the network football GCE performs better than PRSGMFCA while worse in the network karate. And in the network email LFM can not find correct community structure but only find a whole community and many independent nodes, which results from the realization of algorithm. From what we have discussed above, we know that PRSGMFCA can perform stably and get ideal results as well.

1.5. The Complexity Analysis of Algorithm Time

After inputting the network G and the relative parameters and through some procedures, it will output a set of local community structures. During the whole process, the worst complexity of algorithm time in node centrality's ranking is $O(n^2)$, in which n represents the total nodes and in conclusion the fewer nodes, the lower complexity. Therefore, in the real world the PageRank will be weakened in the linear time $O(\log n)$, however, it will be difficult to solve the time complexity of local community expansion, for its expansion depends on the variable γ under the dynamic change of it. So in this paper it first ensures the value of γ ,then uses $O(n_c^2)$ to represent the time complexity of constructing local community with n_c nodes. The worst condition is only to find 1 whole network community as the local community, at which the time complexity is $O(n^2)$. In fact, this will not usually be found in the ground-truth network. The algorithm runs so effectively and when the community is small enough, the time will approximate to linear time. From the Figure 2, it can be observed that the algorithm generates the running time, in which the number of nodes is controlled from 1000 to 100000.



(A)The running time analysis diagram between PRSGMFCA and LFM



(B)Running time of PRSGMFCA in different resolution

Figure 2. The Complexity Analysis Diagram of Algorithm Time.

2. Conclusion

To solve the problems uncovered in detecting the overlapping community structure, in this paper the optimized and improved PRSGMFCA Model algorithm was used. During the process of choosing the seed node of local community, it mainly used the optimized PageRank algorithm to rank, then constructed multi-view recognition modularity based on Spin-glass model to realize local community expansion and detected the structure with the improved Greedy Search Method to obtain the clear overlapping community structure finally. Meanwhile, after analyzing the difference between the optimized PRSGMFCA and the traditional technology method in the generated and real networks, it was found that the former owns better stability, higher correct rate and a complexity of algorithm time within an acceptable range.

References

[1] PORTER M A, ONNELA J P,MUCHA P J. Communities in networks[J].Notices American Mathematical Society, 2009, 56:(9): 1082-1097,1164-1166

[2] G. Palla, 1. Derenyi, I. Farkas, T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005,435(7043): 814-818.

[3] Yang J, Leskovec J. Structure and Overlaps of Ground-Truth Communities in Networks[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2014, 5(2): 26.

[4] A.Lancichinetti, S. Fortunato, J. Kertesz. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics. 2009,11(3):13-15.

[5] Mu C, Liu Y, Liu Y, et al. Two-stage algorithm using influence coefficient for detecting the hierarchical, non-overlapping and overlapping community structure[J]. Physica A: Statistical Mechanics and its Applications, 2014, 408: 47-61.

[6] Lei Pan, Chongjun Wang, Junyuan Xie. Detecting Link Communities based on Local Approach. In Proceedings of 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI11), 2011: 884-886.

[7] Rizman Žalik K, Žalik B. A local multiresolution algorithm for detecting communities of unbalanced structures[J]. Physica A: Statistical Mechanics and its Applications, 2014, 407: 380-393.

[8] J. Reichardt, S. Bornholdt. Statistical mechanics of community detection. Physical Review E. 2006, 74(1): 016110.