Parameter Searching in Attractor Algorithm for Community Detection— an Application in Pathway Enrichment Analysis

Panzhong Lu¹, Hanyi Zhang¹, Ying Liu¹, Yongze Wu¹, Xuan Qin²,³ and Jingbo Xia²,³,⁴*

¹College of Science, Huazhong Agricultural University, China
²Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, China
³College of Informatics, Huazhong Agricultural University, China
⁴Intramural Institute of Mathematics Research, Huazhong Agricultural University, China
*Corresponding author: xiajingbo.math@gmail.com; xjb@mail.hzau.edu.cn

Abstract. Generally, it was challenging to identify various complicated network structures in the real world, e.g., the dynamic network in social media. Given the underlying relationships among them was unveiled, it was efficient to achieve the goal of the knowledge curation and inference. As an existed community detection algorithm, Attractor built a successful dynamic system by representing the dynamic distance updating rules among relevant nodes through time changes, and it eventually partitioned nodes into different communities. Though Attractor was good at large-scale data handling and low time complexity, there was still room for improvement. By modifying its main iteration formulas and better controlling the iteration-related parameters, we achieved a better convergence rate for dynamic distance updating. Moreover, the scope of $\lambda$ parameter was tested intensively so as to find the best cutoff of the parameter setting. In the last section, the modified Attractor algorithm was applied to the gene enrichment, and the results were compared accordingly.

1. Introduction

Many network structures that contain lots of information existed in the real life, including the diverse relationships between nodes. Therefore, it was significant to discover the underlying micro-structures of the networks. Till now, there were a lot of algorithms being developed, like Normalized Cut, Minimum-cut Tree, Modularity and so on [1].

In this paper, Shao’s Attractor algorithm [2] was studied, which was based on a distance dynamical system. In the system, supposing a bulk of nodes composed a constructed graph and each of the nodes had an impact on its neighbors. These interactions caused the distance of the nodes changing. As a result of that, the nodes with closer relationship would be brought together and a community was formed. More specifically, the edge of each node was assigned with an initial weight at the beginning, then the relationship between each node was changing due to its interaction. At this point, the interaction could also be assessed by the differentiate connection between the nodes, and the weight of the distance would be obtained. Through repeating calculations, the distance between two nodes was going to approach to 0 or 1 in the end. Compared to other state-of-the-art algorithms, Attractor algorithm was more outstanding for its intuitive way to analyze community and its low time complexity. Besides, the algorithm represented the user-defined formulas to express the relationships between nodes, which indicated that there may be some improvements for it. In conclusion, to modify...
the formula and make Attractor more accurate, lots of work had been done. Furthermore, a case study in Gene Enrichment played an important role to make Attractor algorithm more practical.

1.1. Related Work

A couple of related work has been done to carry on the community detection algorithm [3, 4, 5, 6, 7]. Some of the representatives were listed as below:

- Minimum-Cut criterion. The minimum-cut criterion delineates K sets of disjoint vertices so that the sum of the weights of the edges between vertex sets is minimized. (The weights of the edges represent the similarity of the vertices.)
- Normalized Cut [3]. The NC is not capable of handling large-scale networks while it is suitable for the small ones.
- Markov Cluster algorithm [4]. MCL is based on the simulation of flow in graphs. Using Random Walks get the final probability matrix, but in the process of seeking, it also lost a lot of information. Its final probability has nothing to do with the position of the start point.
- Modularity [5]. Modularity is to measure the accurate result of community detection. The greater the value of Modularity, the better the result of community division. Unfortunately, it tends to fail on many real-world networks due to the “resolution limit”. The situation becomes worse especially when the network size increases.
- METIS [6, 7]. The METIS is a class of multi-level scalable partitioning techniques. METIS has high quality division result but has a low fill-in, thereby reduces storage load and calculations. Metis allows scaling up large-scale networks, but result suffer in coarsening.

2. Introduction

2.1. Dataset Analysis

Three different real-world networks with class labels were mainly tested in community detection algorithms, which were derived from the standard large network dataset collection (http://snap.standford.edu/data/) and The Koblenz Network Collection (http://konect.uni-koblenz.de/).

- American College football (Football): There are 115 nodes on behalf of American football teams and 616 edges on behalf of the games that teams played in season Fall 2000. First, the 115 teams were grouped into 12 conferences.
- Political blogs (Polblogs): The dataset consists of 1490 nodes represent users and the 19,090 edges represent links between blogs which were extracted from a crawl of the blogs.
- Zachary karate club (Karate): The network is generated from a karate club at a US university. The network is undirected, whose 34 nodes represent the 34 members of the club and 78 edges represent the relationships within these members.

2.2. Algorithm Introduction of Shao’s Attractor Algorithm [2]

Shao’s Attractor algorithm was based on a graph definition in the community $G = (V, E, W)$, where $V$ is the set of nodes, $E$ is the set of edges and $W$ is the relational set of weights, for instances, assuming there are two nodes $u$ and $v$, $e = \{u, v\} \in E$ represents the edge or the relationship between $u$ and $v$, stands for the weight of edge $e$.

Supposing an undirected graph $G = (V, E, W)$, $\forall u \in V$, the set $\Gamma(u)$ interferes the node $u$ and its adjacent nodes is: $\Gamma(u) = \{v \in V \mid \{u, v\} \in E\} \cup \{u\}$, and the Jaccard distance of two nodes is defined as: $d(u, v) = 1 - \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$ For a weighted graph, this distance value was accordingly modified to $d(u, v) = 1 - \frac{\sum_{x \in \Gamma(u) \cap \Gamma(v)} \omega(u, x) + \omega(v, x)}{\sum_{x \in \Gamma(u) \cup \Gamma(v)} \omega(x, y)}$. On the bases of the definitions above, to unveil the
laws of the dynamic distances, the algorithm discusses three different single interaction patterns. Previously, let $e = \{u, v\} \in E$ be an edge of the two nodes $u$ and $v$. After that, the algorithm elaborates how the three modes' distances changes differently.

The distances $d(u,v)$ can be changed under the interactions of direct linked nodes, where $f(\cdot)$ is a coupling function and $\sin(\cdot)$ is used here, and $1 - d(u,v)$ represents the similarity between the nodes $u$ and $v$, $\deg(u)$ is the degree of the node $u$. Shao [2] designed DI, EI and CI to represent the distance of two nodes in time $t$. Here $DI = - \left( \frac{1}{\deg(u)} \cdot f\left( 1 - d(x,u) \right) \cdot \left( 1 - d(x,v) \right) + \frac{1}{\deg(v)} \cdot f\left( 1 - d(x,v) \right) \cdot \left( 1 - d(x,u) \right) \right) \cdot d(u,v)$, where $CN = \left( (\Gamma(u) - u) \cap (\Gamma(v) - v) \right)$ stands for the common neighbors of nodes $u$ and $v$, they effect with other, which contributes to the distance of $d(u,v)$.

$EI = - \sum_{x \in EN(u)} \left( \frac{1}{\deg(u)} \cdot f\left( 1 - d(x,u) \right) \cdot \rho(x,u) \right) - \sum_{y \in EN(v)} \left( \frac{1}{\deg(v)} \cdot f\left( 1 - d(y,v) \right) \cdot \rho(y,v) \right)$, where

$$\rho(x,u) = \begin{cases} 
(1 - d(u,v)) & \text{if } (1 - d(u,v)) \geq \lambda \\
(1 - d(u,v)) - \lambda & \text{otherwise}
\end{cases}$$  \hspace{1cm} (1)

At last, after presenting the three main equations above, the dynamic system $d(u,v)$ can be established as the following function, where the parameter $t$ stands for the time step,

$$d(u,v,t+1) = d(u,v,t) + DI(t) + h(u,v)CI(t) + EI(t)$$  \hspace{1cm} (2)

For all of the formulas and mechanism, Shao's Attractor paper [2] is referred. As the parameter $\lambda$ in equation (1) is a key factor of the algorithm, a thorough testing of $\lambda$ was performed in this research.

2.3. Metric

There are some common strategies defined to compare different community detection algorithms through the evaluation of detected community.

- **Normalized Mutual Information (NMI):** Normalized Mutual Information (NMI) is a widely used method, which could consider varieties of mappings evenly. In the Information Theory, entropy, the information contained in a distribution is defined: $H(x) = - \sum_{x \in X} p(x) \log p(x)$.

The Mutual information further depicts the shared information between two distributions: $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$. Then Normalized Mutual Information (NMI) is defined as: $NMI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$, which could be regularly applied to clear the right mapping between the ground truth of the community structure and the clustered results directly.

- **Modularity:** In the real life, most network structures are full of complexity and has no standard classification tags. It is inconvenient for people to detect the true instruction of the network. Modularity is mainly used to solve this problem. It The greater the value of Modularity, the better the result of community division. Given a set of nodes C, strength of community effect is defined as: $\frac{1}{2m} \sum_{uv} (A_{uv} - \frac{k_{i}k_{j}}{2m}) \delta_{rs}$. Assign edges to the network
requires cutting each edge into two randomly. So that m edges will produce \[ 2m = \sum_{v=1}^{k} k_v \]

distal points. \[ A_{vw} - \frac{k_vK_w}{2m} \] is the difference between the actual number of nodes v and w and the expected number of random networks. \( \delta_v \) is used to indicate whether v is in a community, which equals to 1 or 0.

2.4. Possibility for Algorithm Improvement

2.4.1. Cohesion parameter setting of \( \lambda \) in formula (1). The original Attractor was mainly used to detect the community, but in the process of application, it was hard to determine the value of the \( \lambda \) parameter. To better understand the parameter setting, two evaluation methods of accuracy were applied to find the best \( \lambda \), where NMI (see 2.3) metric served for a manually labeled testing network which had similar structure of the targeted network, while Modularity (see 2.3) was employed directly to the targeted network. In both evaluations, parameter setting of \( \lambda \) was performed.

2.4.2. Modification of interactions in formula (2). During the analysis of the original algorithm, it was found that DI considers that when the \( u, v \) nodes are directly connected, the degree of \( u, v \) is greater, that is, the smaller the degree of influence of \( u, v \) is, the greater the value of DI is, D is also increased, and the \( u, v \) nodes do not approach quickly (D is the distance between two nodes). CI described \( u \) and \( v \) has common neighbours. The more common neighbours are, the smaller D is, and the two nodes gradually approach each other. EI describes \( u \) and \( v \) as having no common neighbours and considering their neighbours x one by one. (x is only a neighbour of \( u \) and \( v \)) If there are more respective neighbours, that is, they become communities with each other, then the larger EI is, the larger D becomes, and the two nodes do not approach quickly.

By analyzing the original algorithm, which is based on the original algorithm, the weight of the main iteration formula \( d(u, v, t + 1) \) is modified, and the correlation between the common neighbor and the independent neighbor is considered. After the positive correlation and negative correlation analysis, the weight function \( h(x) \) is introduced. The coupling function \( \cos(\cdot) \) in \( h(x) \) is to make it in the \( (0, 1) \) interval and controls the degree of influence.

\[
\begin{align*}
h(u,v) &= \cos\left(\frac{1}{\text{deg}(u)} + \text{deg}(v) + 1\right) + 1 \\
d(u,v,t+1) &= d(u,v,t) + DI(t) + h(u,v)CI(t) + EI(t)
\end{align*}
\] (3) (4)

3. Result

3.1. The Result of Cohesion Parameter Searching of \( \lambda \) in formula (1)

In the pattern of EI, \( \lambda \) is aimed to control the positive or negative influence of two nodes from the exclusive neighbours. In the algorithm, \( \lambda \) is adjusted higher generally with more communities while communities become lower in contrast, for which the number of clusters could be set as wanted. And according to the evaluation methods Normalized Mutual Information and Modularity, a discipline is uncovered that the more communities the network is divided into, the better the performance of the algorithm will be. But it doesn't mean larger \( \lambda \) is corresponded with better community detection. The valuation of \( \lambda \) depends on the demands of the detected networks division demands. Provided that community detected wanted to be larger, the \( \lambda \) better be higher.
Figure 1. $\lambda$ searching of Attractor on dataset Football

Figure 2. $\lambda$ searching of modified algorithm on dataset Football.

Figure 3. $\lambda$ searching of Attractor on dataset Polblogs
Figure 4. $\lambda$ searching of modified algorithm on dataset Polblogs

Figure 5. $\lambda$ searching of Attractor on dataset Karate

Figure 6. $\lambda$ searching of modified algorithm on dataset Karate
In this paper, a large number of experiments revealed the discipline of the $\lambda$ setting. When the $\lambda$ is set near 0 or 1, the numbers of communities would extremely small or big, which may cause the error of the running. Significantly, when $\lambda$ locates in 0.4 to 0.6, the NMI cover almost all the range and is able to perform the best, as the Figure 2 - Figure 7 revealed some part result of the experiments.

After the Attractor algorithm has been improved, the effect has been improved, but there are still room for improvement. First, to consider the existence of opinion leaders in the neighbours in terms with influence on the nodes. Second, to consider the positive/negative role of the neighbours.

### 3.2. The Result After the Modification of Interactions in Formula (2)

| Data Sets | | | NMI([2]) | NMI(ours) | Modularity([2]) | Modularity(ours) |
|-----------|---------------|-------------|-----------|-------------|-----------------|-----------------|
| Football  | | | 0.923 | 0.936 | 0.690 | 0.690 |
| Polblogs  | | | 0.559 | 0.598 | 0.639 | 0.640 |
| Karate    | | | 0.859 | 0.927 | 0.590 | 0.623 |

Considering some unreasonable parts in the Attractor strategies, here comes relative modifications put forward. Within the parameters of the distance dynamics observed, the system which is converged too fast needs to be corrected in avoid of the loss of the network's information.

With some coefficient updated, the performance becomes more excellent after the speed of distance's changing slowed. Simultaneously, the DI, CI and EI are considered making distinct difference in the system towards different kinds of dataset. When it come to a network whose nodes has generally large degrees, the CI and EI be regarded to make more difference of the vary of distance. For instance, in $\delta(u, v, t + 1)$, the coefficient $h(u, v)$, of the CI was set as Equation 3.

Also, more strategies were tried out in different situations. The $f(\cdot)$, a coupling function, has many different choices. With regard to the $f(\cdot)$, some analysis about the modification approaches needed to be talked about just like the data sets above.

From the contrast of between Figure 1 and Figure 2 towards the dataset football, the contrasts between Figure 3 and Figure 4 towards the dataset polblogs, the contrast between Figure 5 and Figure 6 go towards the dataset karate, an obvious improvement shown in Table 1 demonstrate the modification.

### 4. Case Study: Application in the Pathway Enrichment for Gene Communities

When it comes to the extraction of biological informatics, the information of gene-to-gene interaction is of great value in pharmacological researches. Till now, the text mining method has been used to extract these genes with interaction information and transforms these genes into a huge gene interaction network. Then it is easier for people to use the classification tool to enrich genes and aim to different purposes. In the paper, the object selected by text mining is a redirected drug which called Rapamycin, firstly, we retrieved 31118 abstracts which is related to the drug from Pubmed, and then we obtained relational information of the interaction among the large number of genes through TEES. In this process, on account of the point to classify this large gene network, we use the modified-Attractor algorithm to identify it. This classified sub-network obtains the corresponding drug related-activity gene network pathway.

Through this method, we get the results of the genes above Table 2. Then it is significant to analyze if the sub-network can be enriched in the pathway that we want. It can be a rule that the KEGG database is a standard gene network. We compared our sub-network results to the corresponding metabolic pathway information that provided in the KEGG database. In theory, different sub-networks of this large gene network should be related to different drug activities.
Table 2. The enrichment result of sub-network in Pancreatic Cancer pathway

| GeneRatio | BgRatio | Total | Rank | Method          | Jaccard distance |
|-----------|---------|-------|------|-----------------|------------------|
| 35/471    | 75/7430 | 106   | 7    | KEGG            | 0.068493151      |
| 20/153    | 75/7430 | 129   | 8    | Attractor(λ=0.30) | 0.096153846      |
| 15/121    | 75/7430 | 115   | 14   | Attractor(λ=0.44) | 0.082872928      |
| 15/118    | 75/7430 | 116   | 11   | Attractor(λ=0.50) | 0.084269663      |
| 14/108    | 75/7430 | 108   | 10   | Attractor(λ=0.61) | 0.082840237      |
| 14/98     | 75/7430 | 101   | 10   | Attractor(λ=0.69) | 0.088050314      |

Table 2 contains the GeneRatio column, which means the proportion of enriched pathway gene in the entire sub-network. The second column stands for the number of genes in the target pathway of the entire human body. Total column represents the number of the pathway that is enriched by the sub-network while Rank column stands for the level of the sub-networks enrichment's effects. Jaccard distance showed that the selection of \( \lambda \) is meaningful.

5. Acknowledgments
The authors like to express their gratitude to Dr. Junming Shao for kindly offering discussions and codes for Attractor research. This research is as well supported by the Fundamental Research Funds for the Central Universities of China (Project No. 2662018PY096).

6. References
[1] Newman M E and Girvan M 2004 Finding and evaluating community structure in networks (Physical Review E Statistical Nonlinear Soft Matter Physics) 026113: 6
[2] Shao Junming and Han Zhichao and Yang Qinli and Zhou Tao 2015 Community Detection based on Distance Dynamics.
[3] Jianbo Shi and Jitendra Malik 2000 Normalized cuts and image segmentation (IEEE TPAMI) 22(8): 888-905.
[4] Stijn Van Dongen 2000 A cluster algorithm for graphs (Report-Information systems) pp.1-40
[5] Mark EJ Newman 2006 Modularity and community structure in network (Proceeding of the National Academy of Sciences) 103(23): 8577-858
[6] George Karypis and Vipin Kumar 1998 A fast and high quality multilevel scheme for partitioning irregular graphs (SIAM Journal on scientic Computing) 20(1): 359-392
[7] George Karypis and Vipin Kuma 1998 Multilevelk-way partitioning scheme for irregular graphs (Journal of Parallel and Distributed computing) 48(1): 96-129.