31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Network Topological Reordering Revealing Systemic Patterns in Yeast Protein Interaction Networks

Xiaogang Wu, Ragini Pandey, and Jake Yue Chen, *Senior Member, IEEE*

Abstract— **Identifying candidate genes/proteins involved in human disease specific molecular pathways or networks has been a primary focus of biomedical research. Although node ranking and graph clustering methods can help identify localized topological properties in a network, it remains unclear how the results should be interpreted in biological functional context in systems-level. In complex biomolecular interaction networks, biomolecular entities may not have absolute ranks or clear cluster boundary among them. We presented Ant Colony Optimization Reordering (ACOR) method to examine emerging network properties. The task of reordering nodes is represented as the problem of finding optimal density distribution of "ant colony" on all nodes of the network. We applied ACOR method to re-analyze a yeast protein-protein interaction (PPI) network annotated with functional information (i.e., lethality), which revealed intriguing systems-level functional features.**

I. Introduction

The identification of candidate molecular biomarkers [2] or drug targets [3] involved in human diseases has been a primary focus of biomedical research. Conventional methods involve finding disease-associated genes through genetic linkage or mutation analysis. Experimental and computational techniques to analyze a gene/protein's sequence using phylogenic, biochemical, and molecular biology methods have led to profound understandings of the functions of genes/proteins based on their sequences and structure. In recent years, genomics, proteomics, and systems biology techniques have led to an influx of new molecular interaction data. The collective study of all molecular entities and their relationships - network biology - has led to new ways to characterizing biological functions of genes/proteins in their molecular interaction network and pathway context. For example, Jeong *et al.* were among the first to investigate two key molecular network concepts, *centrality* and *lethality*, in a yeast protein-protein interaction (PPI) network [1]. In this work, *centrality*, the topological connectedness of proteins to other interaction partners in a network, was applied to study *lethality*, a phenotype that indicates whether yeast could survive when the gene is knocked out. Studies that followed have discovered many important associations between topological and functional properties of PPI networks. For a recent review, refer to [4].

An increasingly popular biomolecular network analysis technique, known as *node ranking*, aims to find genes/proteins based on the gene/protein's network topological properties either entirely or partially. Weston *et al.* introduced a protein ranking algorithm (*RankProp*), which studied all similarity relationships among proteins in a sequence database by performing a diffusion operation on a pre-computed, weighted network [5]. Chen *et al.* developed a computational method to prioritize disease-related proteins in Nearest Neighbor Expanded (*NNE*) subnetworks and PPI-data-quality-adjusted protein ranking score [6]. Inspired by the *PageRank* algorithm of the Google search engine, Morrison *et al.* developed a gene ranking algorithm (*GeneRank*), which combined gene expression information with a network structure derived from GO co-annotations or correlated co-expression profiles [7]. Wang *et al.* presented a hybrid approach called *HykGene*, in which they selected gene biomarkers for phenotype classifications from microarray gene expression experiments by integrating gene ranking and hierarchical clustering analysis [8]. Ma *et al.* also developed a new approach called *CGI* for prioritizing genes in a disease by combining gene expression profiles and PPI data [9]. These methods enable biomedical researchers to filter hundreds of genes or proteins often derived from high-throughput experiments and help them hypothesize on which gene/protein may be used as biomarkers or therapeutic drug targets based on newly network biology knowledge.

Another task in biomolecular network analysis is to identify functional modules. Many types of *graph clustering* methods have been proposed to solve this problem in network biology. One simple method for graph clustering is to search for *minimal cuts* in a graph by using the *maximum flow* algorithm [10]. Conventional *hierarchical clustering* has been popular for partitioning large graphs that represent biomolecular interaction networks [11]. *Spectral graph clustering* can also be performed, by computing the eigenvectors corresponding to the second-smallest eigenvalue of the normalized Laplacian or an eigenvector of another matrix that represents the graph structure [11, 12].

Manuscript received April 23, 2009. This work was supported in part by Department of Defense (DOD) Breast Cancer Research Program (BCRP) Concept Award (W81XWH-08-1-0623) to Dr. Jake Chen.

X. Wu is with the School of Informatics, Indiana University, Indianapolis, IN 46202, USA, and Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN 46202 (e-mail: wu33@iupui.edu).

R. Pandey was with the School of Informatics, Indiana University, Indianapolis, IN 46202, USA. She is now with the Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN 46202 (e-mail: ragini.pandey@gmail.com).

J. Y. Chen is with the School of Informatics, Indiana University, Indianapolis, IN 46202, USA; Department of Computer and Information Science, Purdue University, Indianapolis, IN 46202; and the Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN 46202; (Phone: 317-278-7604; e-mail: jakechen@iupui.edu).

Heuristic methods such as *spring-force* or other *energy models* for network visualization can also be applied to graph clustering if the nodes in a network graph is within limits of computational resources allocated [11]. *Resistor network* of circuits that model each edge as a unit resistor may also be used to cluster the nodes based on the voltage potential differences by calculating the potentials at all of the nodes [13]. Recently, a robust *Markov clustering algorithm* based on *flow simulation*, a type of *random walk* methods, has also been proposed [14]. These graph clustering algorithms are perhaps related: cut-based methods are a special type of spectral graph partitioning methods, which in turn are related to random walks that can model the behavior of circuit networks and determine *betweenness* network topological features [15].

While node ranking and graph clustering methods can help identify localized network topological features, it remains unclear how the results should be interpreted in biological functional context. Complex biomolecular interaction networks are often characterized by *small-world* and *scale-free* properties, which suggest that bimolecular entities may not have "*absolute ranks*" or "*clear cluster boundary*" among them. Could there be more emerging biomolecular network properties for us to discover?

In this paper, we present a computational framework based on *Ant Colony Optimization* (*ACO*) [16] to reorder network nodes. The task of reordering nodes is represented as the problem of finding optimal density distribution of "ant colony" on all nodes of the network. This new framework - ACO Reordering (*ACOR*) - also enables us to examine emerging (globalized) properties in a biomolecular interaction network. We applied the ACOR method to re-analyze a yeast PPI data annotated with lethality information in [1]. Our results revealed intriguing systems-level functional features not previously reported.

II. Method

General ant colony optimization methods have been used to find shortest path in a graph or network. Here, we represent the problem of finding highly relevant nodes in a network as one in which simulated ants (*s-ant*) roam all possible network paths iteratively. By designing different strategies $\mathbf{F}_i$ of s-ants in each step "walking" in a network, the iteration process can be manipulated to get the density distribution $\mathbf{s}$ of s-ants crowding on each node, as shown in (1). According to this density distribution $\mathbf{s}$, the ranked adjacency matrix of the network will be shown as a map to reveal the system-level feature of the network.

$$\mathbf{s}_{i+1} = \mathbf{F}_i(\mathbf{M}_i)\times\mathbf{s}_i, \mathbf{s}_i \in R^n, \mathbf{M}_i \in R^{n\times n}, \quad \mathbf{F}_i \in R^{n\times n} \to R^{n\times n}, i = 0,1,...,N-1 \tag{1}$$

$\mathbf{s}_i$: i th step density distribution of s-ants crowding on each node.

$\mathbf{F}_i$: Strategies of s-ants for i th step taken to walk in a network.

$\mathbf{P}$: Adjacency Matrix of the network (in spite of directed vs. undirected or un-weighted vs. weighted).

$\mathbf{M}_i$: Matrix determined by both the network features P under analysis (including topology and function information) and the marks signed by s-ants.

$\mathbf{s}_0 = (1/n, 1/n, \ldots, 1/n)^T$: to describe the equivalence of each node in the network.

$\mathbf{c}_i$: Rank vector according to i th step density distribution $\mathbf{s}_i$.

$\mathbf{P}_{i+1} = \mathbf{P}_i(\mathbf{c}_i, \mathbf{c}_i)$: Ranked Adjacency Matrix of the network with rank $\mathbf{c}_i$.

In a simple case of the proposed scheme, s-ants never sign a mark on the network, and $\mathbf{M}_i$ is only determined by the network, which means it is invariable. Equation (1) can be reduced to the following:

$$\mathbf{s}_{i+1} = \mathbf{F}_i(\mathbf{M}_i)\times\mathbf{s}_i = \mathbf{F}_i(\mathbf{M})\times\mathbf{s}_i = \mathbf{F}_i \cdots \mathbf{F}_1 \cdot \mathbf{F}_0(\mathbf{M})\times\mathbf{s}_0 \tag{2}$$

For further simplification, s-ants can be modeled by the constraint of maintaining a constant walking strategy, and (2) can be reduced to the following:

$$\mathbf{s}_{i+1} = \mathbf{F}_i(\mathbf{M})\times\mathbf{s}_i = \mathbf{F}^{(i)}(\mathbf{M})\times\mathbf{s}_0 = \mathbf{M}^i \times\mathbf{s}_0 \tag{3}$$

Let $\mathbf{P}$ denote the adjacency matrix of the network (regardless of directed/undirected types or un-weighted/weighted types). In the event that s-ants fail to populate, $\mathbf{M}$ can be obtained by (4) below:

$$\mathbf{P} = \{p_{i,j}\}, \mathbf{M} = \{m_{i,j}\}, m_{i,j} = p_{i,j}/(1+\sum_j p_{i,j}), i, j = 1,2,...,n \tag{4}$$

We proved that the final density distribution $\mathbf{s}_N$ has a convergent limit as described by (5).

$$\lim_{N\to\infty} \mathbf{s}_N = S = \{s^{(i)}\}, s^{(i)} = \frac{1+\sum_j p_{i,j}}{n+\sum_i\sum_j p_{i,j}}, i, j = 1,2,...,n \tag{5}$$

If s-ants populate quickly, $\mathbf{M}$ can be simply evaluated as $\mathbf{M} = \mathbf{P}$. In this situation, a convergent property of this

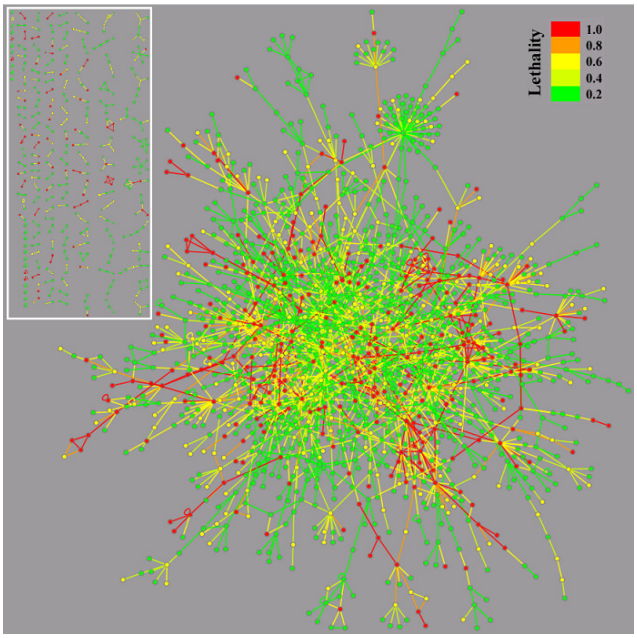

Fig. 1. Yeast PPI network (Protein: 1870; Interaction: 2277) layout were generated by Cytoscape. The main sub-network contains 1458 proteins and 1993 interactions. Other separated sub-networks are all shown in the upper left small windows. All the yeast PPI data and lethality data were provided by Dr. H. Jeong [1].

algorithm cannot be assured for all kinds of networks. In our experiments, it seems to be related with *scale-free* feature[1].

III. RESULTS

First, we define *Lethality Score* (*LS*) for each protein (node) as in (6).

$$LS(n_i) = \begin{cases} 1.0 & \text{Lethal} \\ 0.6 & \text{Unknown} \\ 0.2 & \text{Non-lethal} \end{cases} \quad (6)$$

Then define *LS* for each protein interaction (edge) as in (7).

$$LS(e_{i,j}) = [LS(n_i) + LS(n_j)]/2 = (0.2, 0.4, 0.6, 0.8 \text{ or } 1.0) \quad (7)$$

The yeast PPI network in [1] was redrawn in Fig. 1. Visually, it's difficult to find how centrality relates to lethality. Three different types of subnetworks (lethal, unknown, and non-lethal) are organized around the global network and difficult to be visually separated.

There are two modes for our ACOR algorithm. One is called *un-populated mode*, which is governed by (3) with **M** taking as (4), and another one is called *populated mode*, which is also governed by (3) while with **M = P**. ACOR in un-populated mode, where the s-ant population never changes for each iteration, is similar to the Google PageRank algorithm as seen in (4-5). These results can be compared with those in populated mode, where the s-ant population increases very rapidly, and it will accelerate the propagation of local topological information intuitively. That is why populated mode ACOR can reveal globalized network features, while un-populated mode ACOR can only show localized network properties. For adjacency matrix of network **P** = $\{p_{ij}\}$, p_{ij} = 0 if node *i* do not connect with node *j*, while p_{ij} = 1 (when functional information in (6) was not used) or p_{ij} = *LS* ($e_{i,j}$) (when functional information in (6) was used) if node *i* connect with node *j*.

By using ACOR in *un-populated* mode, the reordered adjacency matrices of the yeast PPI network without and

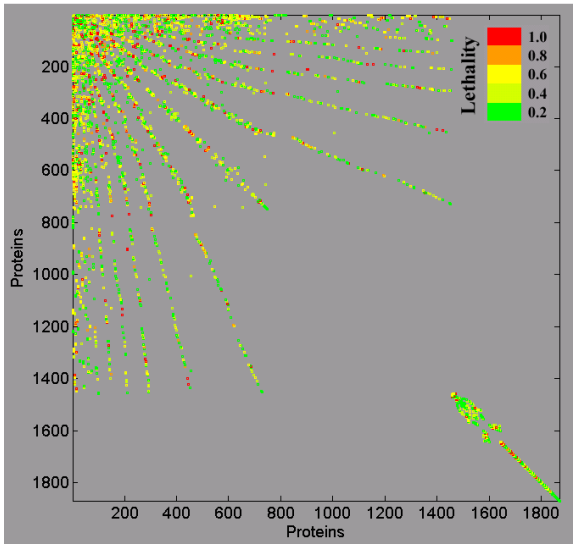

Figure 2. Yeast PPI network adjacency matrix reordered by ACOR in un-populated mode (Equation (3) with **M** as (4), *n*=256). Functional information *LS* ($e_{i,j}$) for each edge was not used in reordering.

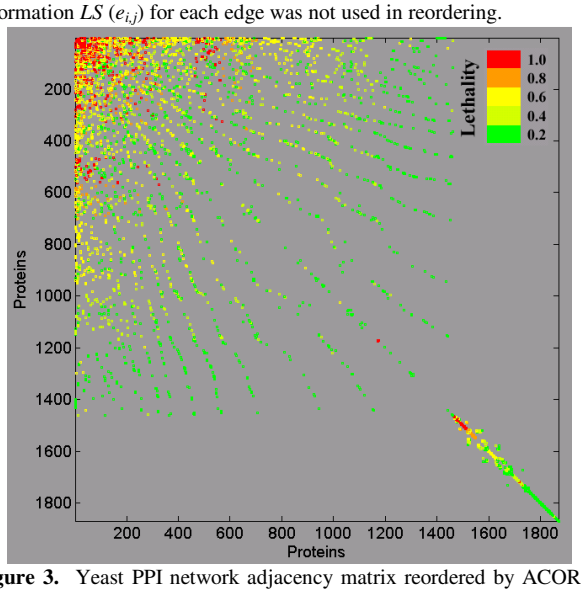

Figure 3. Yeast PPI network adjacency matrix reordered by ACOR in un-populated mode (Equation (3) with **M** taking as (4), *n*=256). Functional information *LS* ($e_{i,j}$) for each edge was used in reordering.

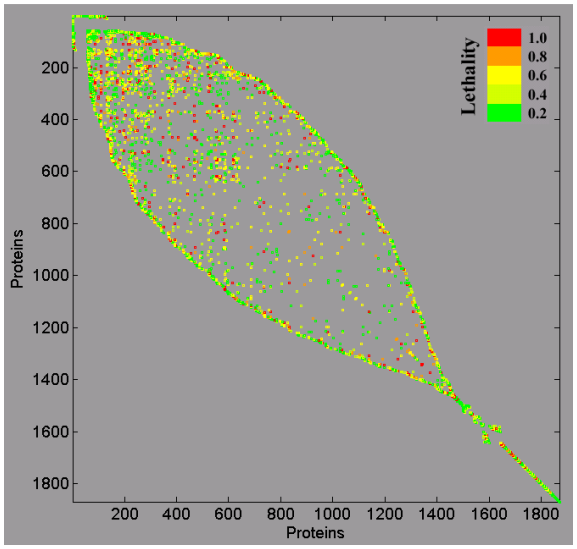

Figure 4. Yeast PPI network adjacency matrix reordered by ACOR in populated mode (Equation (3) with **M** = **P**, *n*=128). Functional information *LS* ($e_{i,j}$) for each edge was not used in reordering.

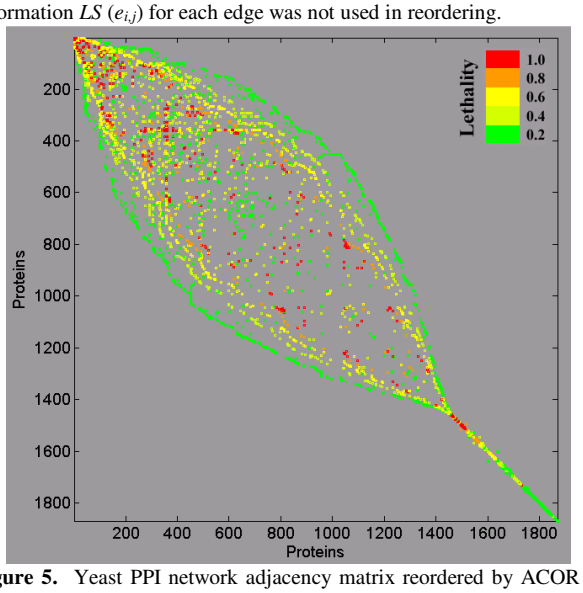

Figure 5. Yeast PPI network adjacency matrix reordered by ACOR in populated mode (Equation (3) with **M** = **P**, *n*=128). Functional information *LS* ($e_{i,j}$) for each edge was used in reordering.

with utilizing functional information (i.e., lethality) are shown in Fig. 2 and Fig. 3, respectively. Although the result from Fig. 2 only utilized topological information and showed interesting pattern, there is little concordance with lethality information. Neither are the results from Fig. 3.

In Fig. 4 and Fig.5, ACOR in populated mode reordered adjacency matrices of the yeast PPI network without and with utilizing functional information are shown, respectively. Here, we can see some distinct patterns. Comparing Fig. 4 with Fig. 5, different subnetworks (according to lethality score) in Fig. 5 were organized in "layers", which is characteristics of *fractals* (multi-scale self-similarity). To show this pattern more clearly, in Fig. 6, we separated subnetworks according to lethality scores for their interactions. We can see how similar those subnetworks are. Node degree distributions for the whole PPI network and each subnetwork are also shown (Fig. 7). We could confirm that these subnetworks all obey *power-law* distribution - characteristics of scale-free networks.

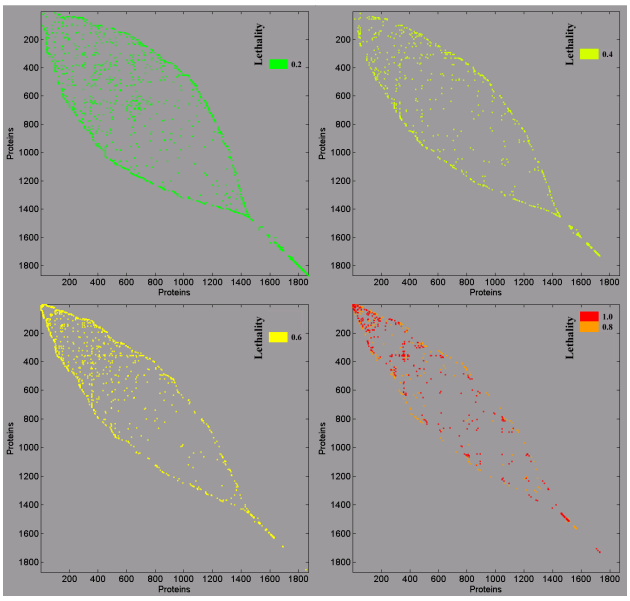

Figure 6. Four different subnetworks involved in Fig. 5, according to the interaction lethality (*LS* =0.2; *LS* = 0.4; *LS* = 0.6; and *LS* >=0.8)

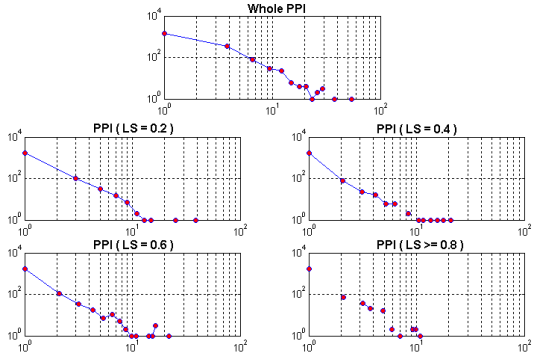

Figure 7. Node degree distribution for the whole yeast PPI network and four different subnetworks shown in Fig. 6.

IV. Conclusion

We developed a new ACOR framework that can efficiently extract network topological features while identifying global structures correlated with biomolecular functions as a whole. In complex biomolecular interaction networks without clear clusters or absolute ranks, our method can assign each node in PPI network a "relative rank" in "blurred clusters". The system-level functional patterns in yeast PPI networks are closely related to scale-free network features. We believe that this method could help unravel high-level "*orderness*" ultimately interpretable in biological contexts and represents a brand-new type of solution for future network biology studies.

Acknowledgment

We thank Dr. H. Jeong (KAIST) for providing the yeast gene lethality data and Drs. Z. Oltvai (University of Pittsburgh) and A.-L. Barabási (Northeastern University) for their helpful suggestions.

References

[1] H. Jeong, S.P. Mason, A.L. Barabasi, and Z.N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, (no. 6833), pp. 41-42, 2001.

[2] C.L. Sawyers, "The cancer biomarker problem," *Nature*, vol. 452, (no. 7187), pp. 548-552, 2008.

[3] D.C. Altieri, "Survivin, cancer networks and pathway-directed drug discovery," *Nature reviews. Cancer*, vol. 8, (no. 1), pp. 61, 2008.

[4] X. Wu and J.Y. Chen, "Molecular Interaction Networks: Topological and Functional Characterizations", A. Gil, B. Roseann, and R. Marco, "Automation in proteomics and genomics: an engineering case based approach", Wiley Publishing, 2009.

[5] J. Weston, A. Elisseeff, D. Zhou, C.S. Leslie, and W.S. Noble, "Protein ranking: From local to global structure in the protein similarity network," *Proceedings of the National Academy of Sciences*, vol. 101, (no. 17), pp. 6559-6563, 2004.

[6] J.Y. Chen, C. Shen, and A.Y. Sivachenko, "Mining Alzheimer disease relevant proteins from integrated protein interactome data," *Biocomputing 2007-Proceedings of the Pacific Symposium*, vol. 11, pp. 367-378, 2006.

[7] J.L. Morrison, R. Breitling, D.J. Higham, and D.R. Gilbert, "GeneRank: Using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, (no. 1), pp. 233, 2005.

[8] Y. Wang, F.S. Makedon, J.C. Ford, and J. Pearlman, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, vol. 21, (no. 8), pp. 1530-1537, 2005.

[9] X. Ma, H. Lee, L. Wang, and F. Sun, "CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data," *Bioinformatics*, vol. 23, (no. 2), pp. 215, 2007.

[10] U. Feige, D. Peleg, and G. Kortsarz, "The Dense k-Subgraph Problem," *Algorithmica*, vol. 29, (no. 3), pp. 410-421, 2001.

[11] S.E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, (no. 1), pp. 27-64, 2007.

[12] F.R.K. Chung, *Spectral Graph Theory*: American Mathematical Society, 1997.

[13] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, (no. 2), pp. 26113, 2004.

[14] S.M. van Dongen, Graph clustering by flow simulation, Ph.D. Thesis, Universiteit Utrecht, Utrecht, The Netherlands, May 2000.

[15] M.E.J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, (no. 23), pp. 8577-8582, 2006.

[16] M. Dorigo, E. Bonabeau, and G. Theraulaz, "Ant algorithms and stigmergy," *Future Generation Computer System*, vol. 16, (no. 8), pp. 851-871, 2000.