Community Structure Detection in Complex Networks with Applications to Gas-Liquid Two-Phase Flow*

Zhongke Gao and Ningde Jin

School of Electrical Engineering and Automation, Tianjin University,
Tianjin 300072, People's Republic of China
gaozhongke_1982@163.com

Abstract. We propose an algorithm to detect community structure in complex networks based on data field theory. The efficiency and accuracy of the algorithm for computer-simulated and real networks make it feasible to be used for the accurate detection of community structure in complex networks. Using the conductance fluctuating signals measured from gas-liquid two-phase flow dynamic experiments, we construct the flow pattern complex network. With the applications of the community-detection algorithm to the flow pattern complex network, we achieve good identification of flow pattern in gas-liquid two-phase flow. In this paper, from a new perspective, we not only present a new community-detection algorithm based on data field theory, but also build a bridge between complex network and two-phase flow.

Keywords: complex networks, community structure detection, data field theory, two-phase flow pattern identification.

1 Introduction

Complex networks, which have been observed to arise naturally in a vast range of physical phenomena, can describe any complex system that contains massive units (or subsystems) with nodes representing the component units and edges standing for the interactions between them. Since the publication of the seminal works of Watts and Strogatz [1] as well as Barabási and Albert [2], a lot of complex systems have been examined from the viewpoint of complex networks. Examples include World Wide Web [3], metabolic networks [4], protein networks in the cell [5], traffic networks [6], sexual networks [7], earthquake networks [8] and human electroencephalogram networks [9].

The description of the structure of complex networks has been one focus of attention of the physicists' community in recent years. The levels of description range from the microscopic (degree, clustering coefficient, etc., of individual nodes) to the macroscopic description in terms of statistical properties of the whole network (degree distribution,

* This work is supported by National Natural Science Foundation of China (Grant No. 50674070, 60374041) and National High Technology Research and Development Program of China (Grant No. 2007AA06Z231).

J. Zhou (Ed.): Complex 2009, Part II, LNICST 5, pp. 1917 – 1928, 2009.

total clustering coefficient, degree-degree correlations, etc.) [10–12]. Between these two extremes there is a “mesoscopic” description of networks that tries to explain its community structure. The general notion of community structure in complex networks was first pointed out in the physics literature by Girvan and Newman [13], and refers to the fact that nodes in many real networks appear to group in subgraphs in which the density of internal connections is larger than the connections with the rest of the nodes in the network. Because communities are relatively independent of one another structurally, it is believed that each of them may correspond to some fundamental functional unit. For example, a community in genetic networks often contains genes with similar functions and a community on the World Wide Web may correspond to web pages related to similar topics. Detecting and analyzing such communities from complex network, therefore, provides a means for functional dissection of the network and sheds light on its organizational principles. Furthermore, community structures may provide key insights into some uncharacterized properties of a system. For example, attempts have been made to identify and characterize communities (called functional modules sometimes) in biological networks, leading to in-silicon predictions of the functions of some genes [14–15].

In this paper, we first propose an algorithm to detect the community structure in complex networks based on data field theory. Then, through analyzing the algorithm for computer-simulated and real networks, we find the efficiency and accuracy of the algorithm make it feasible to be used for the accurate detection of community structure in complex networks. Finally, after constructing flow pattern complex network with the conductance fluctuating signals measured from gas-liquid two-phase flow dynamic experiment, we using the community-detection algorithm achieve good identification of flow pattern in gas-liquid two-phase flow.

2 Community-Detection Algorithm Based on Data Field Theory

Except the long-range action such as gravity and electromagnetic force, there is also short-range action such as nuclear potential which decays in a high speed. According to the field theory in physics, after introducing the interactivity of physical particles and the methods used to describe the field to the abstract number field space, W.Y.Gan et al.[16] proposed an hierarchical clustering method based on data fields. In this method, every object in the space is seemed as a particle with certain mass, around which exists a symmetrical dummy data field in spheric shape. Similar to the description of vector intensity function and scalar potential function in physical field, this method introduce the definition of potential function and field strength function. Let the potential function be defined as:

$$\varphi(x) = \sum_{i=1}^{n} \varphi_i(x) = \sum_{i=1}^{n} \left(m_i \times e^{-\left(\frac{\|x-x_i\|}{\sigma}\right)^2} \right) \quad (1)$$

where $\|x - x_i\|$ is the distance from object x_i to particle x; $m_i \geq 0 (i = 1, 2, \cdots, n)$ is the mass of each object, which satisfies normalizing condition, expressed by

$$\sum_{i=1}^{n} m_i = 1 \tag{2}$$

$\sigma \in (0, +\infty)$ is the influence factor which is used to control mutual action range between objects. The hierarchical partitioning of the original dataset is performed by iteratively simulating the interaction and movement of the data objects in the fields. Experimental results show that the proposed approach not only enjoys favorite clustering quality, but also has an approximately linear time complexity with respect to the size of dataset.

Enlightened by the hierarchical clustering method based on data fields, we propose an algorithm based on data field theory to detect community structure in complex networks. We defined the potential function as:

$$\varphi(x) = \sum_{i=1}^{n} \varphi_i(x) = \sum_{i=1}^{n} \left(m_i \times e^{-\left(\frac{\|x - x_i\|}{\sigma}\right)^2} \right) \tag{3}$$

where m_i and σ are constant; $\|x_j - x_i\|$ is the shortest path between node x_j and node x_i. The community-detection algorithm based on data field theory consists of the following steps.

(1) We first evaluate the number of community from the distribution map of correspondent elements of the first few non-trivial eigenvectors of the standard matrix, by ascertaining the quantity of groups into which most of the elements are clustered. Then we search for the key node of each community. Supposing there are k communities, we search for the k key nodes and put them in a set denoted as S .

(2) Let reciprocal of k , which is the number of elements contained in S , be the mass of each node. We calculate potential value from current node to each element of S , and the current node will belong to the element which has the largest potential value.

(3) If a node i belongs to more than one element of S , we will calculate potential value from nodes that have edges with i to each element of S . Supposing there are k_i nodes that have edges with i , we calculate sum of k_i potential value and the current node will belong to element which has largest sum of k_i potential value.

(4) We ergodicly calculate every node of the network until k communities appears.

In step (1) above, searching the key node of the community is very important to analyze the properties of the complex networks. There are many references to evaluate the importance of nodes and many ways can be used to search the key node, such as the approach based on node degree sequence arranging and the approach based on betweenness sequence arranging. The degree value of a node reflects the static structure characteristics of the topological model; betweenness reflects the flow of a node which is directly related to the movement of a node; the clustering coefficient of a node reflects the clustering state of other nodes around it. Therefore, we can assess node's importance based on node degree value, betweenness and clustering coefficient to find the key node in a community. The basic algorithm follows.

(a) We calculate every node's degree value and the average degree value of the network.

(b) Through thoroughly reviewing the node's pairs, we explore the shortest path between any pair of nodes and calculate the betweenness and clustering coefficient of every node contained in the network.

(c) Let y_i be defined as

$$y_i = \alpha k_i + \beta b_i + \gamma C_i (i = 1, 2, \cdots, N) \tag{4}$$

where α , β and γ are parameters; k_i is the degree value of node i ; b_i is the betweenness assigned to node i ; C_i is the clustering coefficient of node i .And N is the number of nodes contained in the network.

(d) According to the situation, we can select three parameters- α , β and γ to calculate y_i of all nodes and then select y_i from large to small in turn to get the k key nodes of k communities in the network.(k is the number of communities in a network and i is the serial number of a node).

3 Experiments and Analysis

First, as a controlled test of how well the algorithm performs, we have generated a network with known community structure. Then, we use a real network (The karate club network of Zachary) to test the algorithm. Finally, experiments show that the algorithm presented in this paper is of high accuracy with good performance.

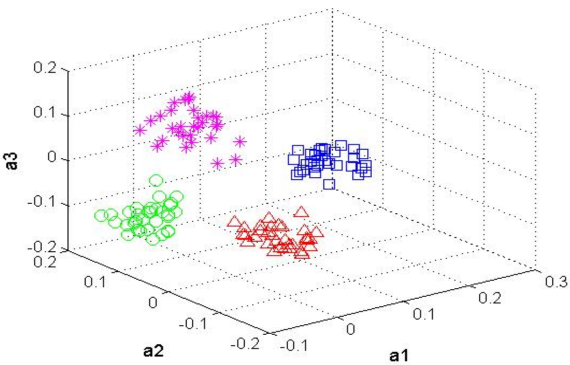
To test whether a particular division is meaningful, here we first introduce a quality function or "modularity" Q , proposed by Newman et al. [17]. Let e_{ij} be the fraction of edges in the network that connect vertices in community i to those in community j , and let $a_i = \sum_j e_{ij}$. Then

$$Q = \sum_i (e_{ii} - a_i^{\,2}) \tag{5}$$

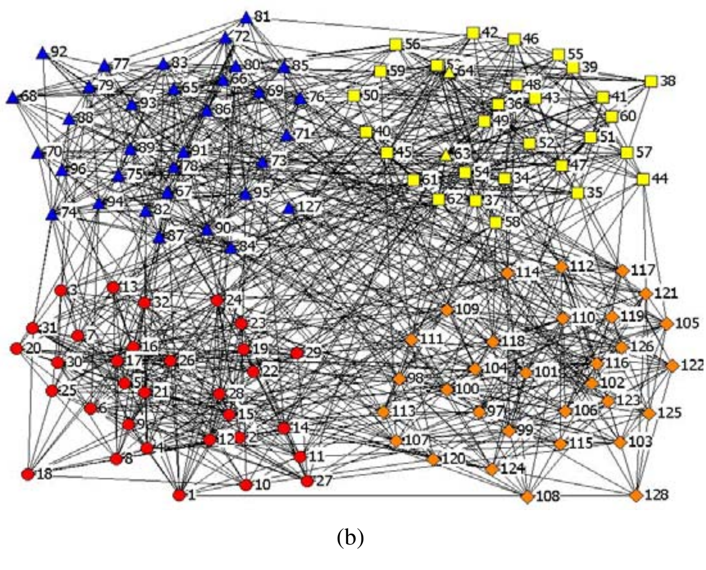
is the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure. If a particular division gives no more within-community edges than would be expected by random chance we will get $Q = 0$. Values other than 0 indicate deviations from randomness, and in practice values greater than about 0.3 appear to indicate significant community structure [17].

3.1 Computer-Simulated Network

To test the effectiveness of the algorithm on network without significant community structure, according to the network whose modularity can be tuned proposed by Newman et al., we generated using Matlab a network without significant community structure, which consists of 128 nodes divided into four communities of 32 nodes. The modularity of the network is 0.2768. We show in Figure 1(a) the distribution map of correspondent elements of the first three non-trivial eigenvectors drawn by Matlab. The community structure of network detected by algorithm based on data field theory is shown in Figure 1(b).



(a)

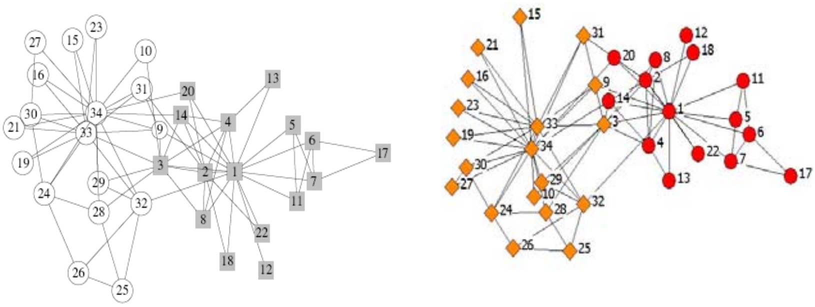


(b)

Fig. 1. Computer-simulated network: (a) Components of the first non-trivial eigenvector a1 are plotted versus those of a2 and a3. (b) Community structure of network detected by algorithm based on data field theory and the shapes of the vertices represent the four communities.

3.2 The Karate Club Network of Zachary

We now turn to application of the algorithm to real world network data. The example is taken from one of the classic studies in social network analysis. Over the course of two years in the early 1970s, Wayne Zachary observed social interactions between the members of a karate club at an American university. He constructed networks of ties between members of the club based on their social interactions both within the club and away from it. By chance, a dispute arose during the course of his study between the club's administrator and its principal karate teacher over whether to raise club fees, and as a result the club eventually split in two, forming two smaller clubs, centered around the administrator and the teacher. In Figure 2(a) we show a consensus network structure extracted from Zachary's observations [18], in which node 1 and 33 respectively represent administrator and the teacher. Feeding this network into our algorithm we find the results shown in Figure 2(b). For the karate club network, there is controversy over only one node, node 3, which is on the border between the communities and so it is understandable that it might be an ambiguous case. Compared with Figure 2(a), the community structure detected by our algorithm, as shown in Figure 2(b), well revealed the relationships among the members of the karate club.



(a) (b)

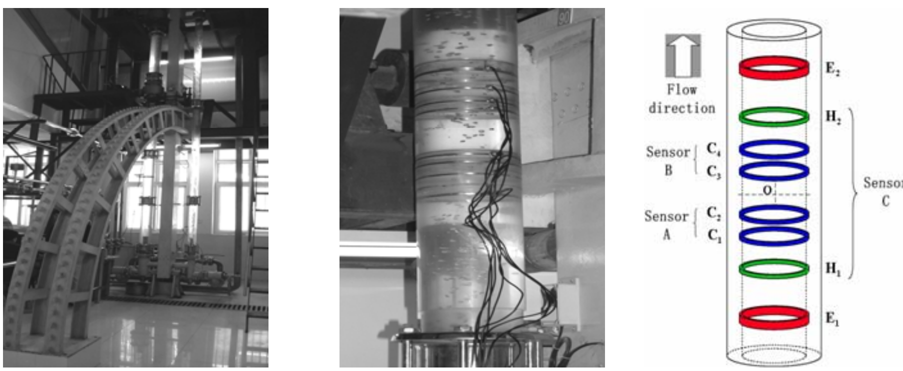
Fig. 2. The karate club network of Zachary [18], with numbered vertices representing the members of the club and edges representing friendships. The shapes of the vertices represent the two communities. (a) We show a consensus network structure extracted from Zachary's observations. (b) Community structure of network detected by algorithm based on data field theory.

4 Applications to Gas-Liquid Two-Phase Flow

Two-phase flow, a primary kind of multiphase system, is found widely in power, chemical, petroleum and other related industries. In the study of two-phase flow, flow patterns indicate how the phases are distributed and mixed due to the physical nature of the system. Two-phase flow patterns depend on the flow rates and direction, the conduit shape, size and inclination. Further, heat and mass transfer rates, momentum loss, rate of back mixing and pipe vibration all vary greatly with the flow patterns. Hence, it is quite important and necessary to recognize the patterns and discern their relationship with the flow properties. But due to the complex nature of two-phase flow, theoretical analyses have not been able to describe the system perfectly. Therefore, a technique is required to detect and identify the flow regime [19-20].

The data we use in this study are the conductance fluctuating signals measured from gas-liquid two-phase flow dynamic experiment in vertical upwards pipe of inner diameter 125mm, which was carried out in the multiphase flow loop of Tianjin University. The whole system is mainly composed of vertical multi-electrode array (VMEA) conductance sensor, high speed VCR, exciting signal model, signal modify model and data processing unit. The flow loop experiment facility is shown in Figure 3(a). The VMEA, researched and produced by our research team, is shown in Figure 3(b) and Figure 3(c). Reference [21] gives the detail description of gas-liquid two-phase flow dynamic experiments. Figure 4 shows the conductance fluctuating signals of five typical flow patterns in vertical upwards pipe. Detail description of flow pattern in gas-liquid two-phase flow can be found in Reference [22].

Due to the significant difference of electrical sensibility between gas phase and liquid phase, the random flow of gas phase will cause voltage fluctuation on the measuring electrode under a certain sinusoidal input, which implies that the conductance fluctuating signals measured from the VMEA conductance sensor are related to the flow transition. Thereby, we construct networks with the conductance fluctuating signal time series.



(a) (b) (c)

Fig. 3. (a)The multiphase flow loop. (b)The VMEA measurement section. (c)The geometry of VMEA. The VEMA consists of eight stainless steel ring electrodes of the material Cr1Ni18Ti9, which are flush mounted on the inside wall of the pmma pipe with 125mm inner diameter and divide into four pairs. Sensor A and Sensor B are respectively the upstream and downstream sensors. Sensor C is volume fraction sensor. E1-E2 is excitation electrode.

Now we start from the flow pattern complex network containing 90 nodes, which is constructed from the conductance fluctuating signals of 90 different flow work conditions, with each flow work condition represented by a single node and edge determined by the strength of correlation between nodes. Considering the nonlinear characteristics of the gas-liquid two-phase flow, we first extract 90 groups of new time series, which can reflect the essential feature of the dynamic system, from the

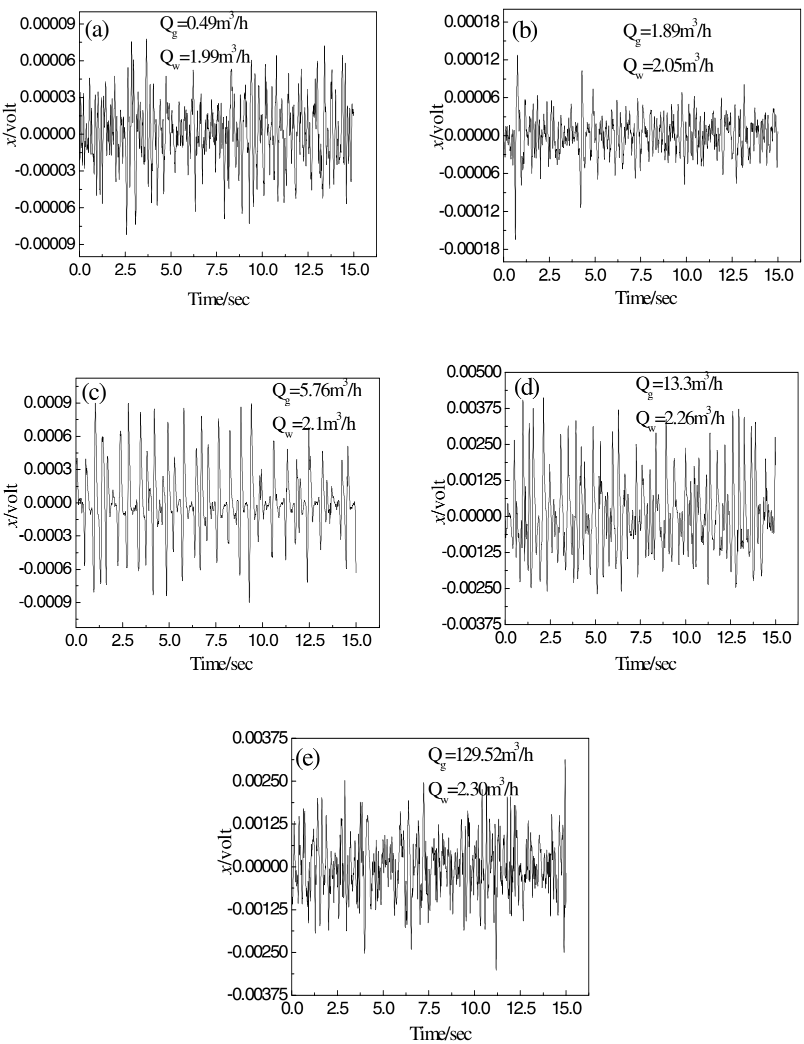


Fig. 4. Conductance fluctuating signals of five typical flow patterns in vertical upwards pipe, in which Q_g and Q_w represent gas flow and water flow respectively. (a) bubble flow; (b) bubble-slug transitional flow; (c) slug flow; (d) slug-churn transitional flow; (e) churn flow.

original conductance fluctuating signals using the method of Time-Delay Embedding [23]. And we use the C-C algorithm [24] to reckon the delay time τ of every group of the original conductance fluctuating signals and choose the specific delay time of Time-Delay Embedding, which can make the modularity largest, to establish the new time series. Then we extract six features in time domain from every group of new time series, i.e. the maximum, minimum, mean value, standard deviation, dissymmetry coefficient and kurtosis coefficient. That is, there are 90 time series vectors, each of which contains six elements, corresponding to the 90 nodes of the network. For each pair of time series vectors, $\mathbf{T_i}$ and $\mathbf{T_j}$, the correlation coefficient can be written as:

$$C_{ij} = \frac{\sum_{k=1}^{M}\left[\mathbf{T_i}(k)-\langle\mathbf{T_i}\rangle\right]\bullet\left[\mathbf{T_j}(k)-\langle\mathbf{T_j}\rangle\right]}{\sqrt{\sum_{k=1}^{M}\left[\mathbf{T_i}(k)-\langle\mathbf{T_i}\rangle\right]^2}\bullet\sqrt{\sum_{k=1}^{M}\left[\mathbf{T_j}(k)-\langle\mathbf{T_j}\rangle\right]^2}} \tag{6}$$

where M is the dimension of the time series vector and $\langle\mathbf{T_i}\rangle = \sum_{k=1}^{M}\mathbf{T_i}(k)/M$, $\langle\mathbf{T_j}\rangle = \sum_{k=1}^{M}\mathbf{T_j}(k)/M$. $\boldsymbol{C}$ is a symmetric matrix and C_{ij} describes the state of connection between node i and j. Choosing a critical value r_c, the correlation matrix $\boldsymbol{C}$ can be converted into adjacent matrix $\boldsymbol{D}$, the rules of which read:

$$D_{ij} = \begin{cases} 1, & (|C_{ij}| \geq r_c) \\ 0, & (|C_{ij}| < r_c) \end{cases} \tag{7}$$

The threshold value r_c determines the characteristics of the resulting network. But up to now, there has not been an acknowledged criterion to well determine the value of r_c. In this paper, we can expect a neighborhood of the threshold value, in which the network can capture the characteristics of the time series and the structure of the

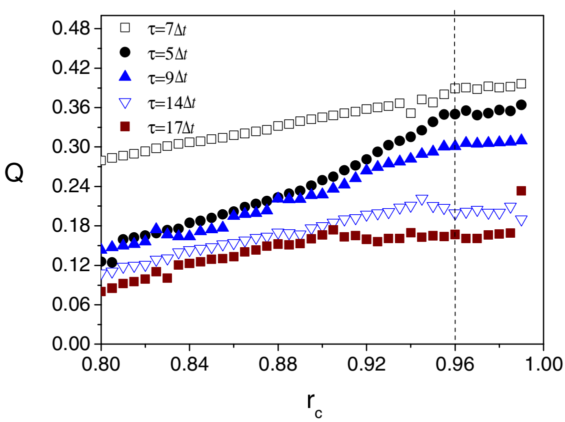
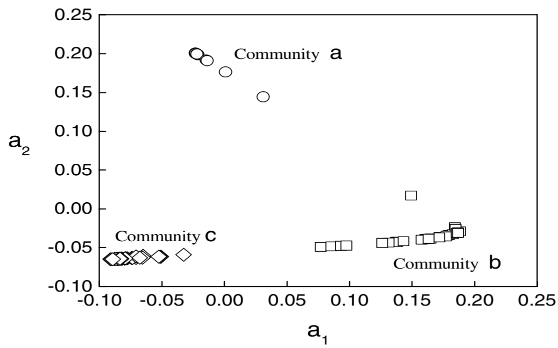
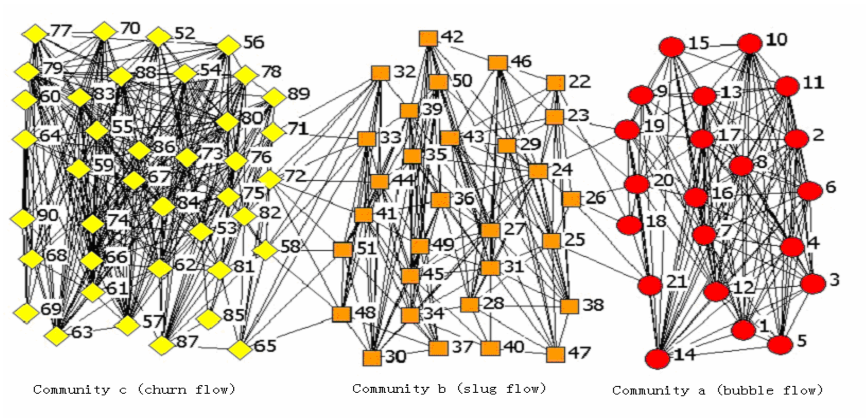


Fig. 5. Distribution map of the delay time, threshold value and modularity. For the flow pattern complex network, after achieving five different delay times calculated by C-C algorithm, we plot the curves of modularity with the increasing of threshold value r_c under every delay time and find that the modularity Q keep stable when r_c ranging from 0.965 to 0.985 and the modularity Q reach the largest when $\tau = 7\Delta t$.



(a)



(b)

Fig. 6. (a) Components of the first non-trivial eigenvector a_1 of the standard matrix are plotted versus those of a_2, from which the number of community can be identified. (b) Community structure of the flow pattern complex network, which shows that there are three communities corresponding to the bubble flow, slug flow and churn flow respectively, and the nodes of the network that connect tightly between two communities corresponding to the transitional flow.

resulting network can keep stable. That is, a proper threshold value can be found just by simulating a special dynamical process of the complex network, i.e., decreasing the number of connections by increasing the value of r_c while keeping the modularity of resulting network almost unchanged. According to the principles mentioned above, we select $\tau = 7\Delta t$ (Δt is the sample interval of original conductance fluctuating signals) and $r_c = 0.978$, as shown in Figure 5. Finally, all the nodes and their edges form the flow pattern complex network, and the topological structure of this network can be described with the adjacent matrix $\boldsymbol{D}$. The conditions $D_{ij} = 1$ and $D_{ij} = 0$ correspond to connection and disconnection, respectively.

After constructing the flow pattern complex network, we detect its community structure using the community-detection algorithm based on data field theory proposed in this paper. Three communities of 21,30 and 39 nodes are found, as shown in Figure 6, denoted as community *a*, community *b* and community *c*, respectively. Moreover, We find that community *a* mainly corresponds to the bubble flow, such as node 2 and node 16; community *b* mainly corresponds to the slug flow, such as node 31 and node 44; community *c* mainly corresponds to the churn flow, such as node 70 and node 90; the nodes that connect tightly between community *a* and community *b* correspond to the bubble-slug transitional flow, such as node 19 and node 26; the nodes that connect tightly between community *b* and community *c* correspond to the slug-churn transitional flow, such as node 32 and node 58. So, with the application of community-detection algorithm based on data field theory, we achieve good identification of flow pattern in gas-liquid two-phase flow.

5 Conclusion

In summary, we have presented a community-detection algorithm based on data field theory. Applying our algorithm to both computer generated and real-world networks with well defined community structures, we got reasonable results that it feasible to be used for the accurate detection of community structure in complex networks. Using the conductance fluctuating signals measured from gas-liquid two-phase flow dynamic experiments, we construct the flow pattern complex network, with each flow work condition represented by a single node and edge determined by the strength of correlation between nodes. With the applications of the community-detection algorithm to the flow pattern complex network, we find that there are three communities in the network, corresponding to the bubble flow, slug flow and churn flow respectively, and the nodes of the network that connect tightly between two communities corresponding to the transitional flow. In this paper, from a new perspective, we not only propose a new community-detection algorithm based on data field theory, but also build a bridge between complex network and two-phase flow by applying the algorithm to flow pattern identification in gas-liquid two-phase flow.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 50674070, 60374041) and National High Technology Research and Development Program of China (Grant No. 2007AA06Z231).

References

1. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' Networks. Nature 393, 440–442 (1998)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)

3. Albert, R., Jeong, H., Barabási, A.-L.: Diameter of the worldwide web. Nature 401, 130 (1999)
4. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.-L.: The large-scale organization of metabolic networks. Nature 407, 651–654 (2000)
5. Jeong, H., Mason, S., Barabási, A.-L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature 411, 41–42 (2001)
6. Li, X.G., Gao, Z.Y., Li, K.P., Zhao, X.M.: Relationship between microscopic dynamics in traffic flow and complexity in networks. Phys. Rev. E 76, 016110 (2007)
7. Liljeros, F., Edling, C.R., Amaral, L.A.N., Stanley, H.E., Aberg, Y.: The Web of Human Sexual Contacts. Nature 411, 907–908 (2001)
8. Sumiyoshi, A., Norikazu, S.: Complex earthquake networks: Hierarchical organization and assortative mixing. Phys. Rev. E 74, 026113 (2006)
9. Cai, S.M., Jiang, Z.H., Zhou, T., Zhou, P.L., Yang, H.J., Wang, B.H.: Scale invariance of human electroencephalogram signals in sleep. Phys. Rev. E 76, 061903 (2007)
10. Strogatz, S.: Exploring complex networks. Nature (London) 410, 268 (2001)
11. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002)
12. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. 45, 167–256 (2003)
13. Girvan, M., Newman, M.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. 99, 7821–7826 (2002)
14. Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. Proteins 54, 49–57 (2004)
15. Wilkinson, D.M., Huberman, B.A.: A method for finding communities of related genes. Proc. Natl. Acad. Sci. U.S.A 101, 5241–5248 (2004)
16. Gan, W.Y., Li, D.Y., Wang, J.M.: An hierarchical clustering method based on data fields. Acta Elec. Sin. 34, 258–262 (2006)
17. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133 (2004)
18. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33, 452–473 (1977)
19. Daw, C.S., Lawkins, W.F., Downing, D.J., Clapp Jr., N.E.: Choaic characteristics of a complex gas-solids flow. Phys. Rev. A 41, 1179 (1990)
20. Daw, C.S., Finney, C.E.A., Vasudevan, M., van Goor, N.A., Nguyen, K., Bruns, D.D., Kostelich, E.J., Grebogi, C., Ott, E., Yorke, J.A.: Self-Organization and Choas in a Fluidized Bed. Phys. Rev. Lett. 75, 2308 (1995)
21. Jin, N.D., Xin, Z., Wang, J., Wang, Z.Y., Jia, X.H., Chen, C.P.: Design and geometry optimization of a conductivity probe with a vertical multiple electrode array for measuring volume fraction and axial velocity of two-phase flow. Meas. Sci. Technol. 19, 045403 (2008)
22. Hewitt, G.F.: Measurement of two-phase flow parameters. Academic Press, London (1978)
23. Packard, N.H., Crutehfield, J.P., Farmer, J.D.: Geometry from a time series. Phys. Rev. Let. 45, 712–716 (1980)
24. Kim, H.S., Eykholt, R., Salas, J.D.: Nonlinear dynamics, delay times, and embedding windows. Physica D 127, 48–60 (1999)