COMMUNITY STRUCTURE IN A LARGE-SCALE TRANSACTION NETWORK AND VISUALIZATION

T. Iino\textsuperscript{a,\ast}, K. Kamehama\textsuperscript{a}, H. Iyetomi\textsuperscript{a}, Y. Ikeda\textsuperscript{b}, T. Ohnishi\textsuperscript{c}, H. Takayasu\textsuperscript{d} and M. Takayasu\textsuperscript{e}

\textsuperscript{a}Department of Physics, Niigata University, Niigata 950-2181, Japan
\textsuperscript{b}Hitachi Research Laboratory, Hitachi Ltd., Ibaraki 319-1221, Japan
\textsuperscript{c}Graduate School of Law and Politics, University of Tokyo, Tokyo 113-0033, Japan
\textsuperscript{d}Sony Computer Science Laboratories, Tokyo 141-0022, Japan
\textsuperscript{e}Department of Computational Intelligence and Systems Science, Yokohama 226-8502, Japan

E-mail: \textsuperscript{\ast}iino@ad.sc.niigata-u.ac.jp

\textbf{Abstract}. We analyze a transaction network of about 800 thousand Japanese firms to elucidate its community structure. Finding community in networks means the appearance of dense connected groups of vertices and sparse connections between groups. We adopt modularity as a quality function of communities introduced by Newman. The modularity optimization is one of effective approaches to find community.

We first use a bottom-up algorithm, which makes the optimization fast by using a greedy algorithm. For the community extraction, the greedy algorithm is widely used, however, may not sufficiently optimize modularity because the optimization tends to be trapped by a local maximum especially for large-scale networks. Alternatively we propose a top-down algorithm with implementation of an annealing method and compare effectiveness of the two algorithms.

We also compare the results of the community analysis with images of network structure visualized by molecular dynamics method. The vertices belonging to the same community are spatially located close to each other. The community structure determined by the modularity optimization is well reproduced in the network structure obtained by molecular dynamics.

\section{1. Introduction}

Many complex systems interested by scientific community can be represented as networks. Statistical methods for quantifying and theoretical modeling of networks have been applied with success to extensive networks, including the World Wide Web, citation networks, neural networks, and social network \cite{1, 2}. Real networks are often characterized by communities in which nodes are divided into densely connected groups and on the other hand those groups are sparsely connected. Communities frequently display different statistics from the network as a whole \cite{3}. Extracting communities can also provide coarse-graining view to networks for visualization or other purposes \cite{4}. Community structures are thus helpful for understanding the structure of large complex networks.

We analyze a transaction network consisted of about 800 thousand Japanese firms \footnote{The data used in this paper is collected by TSR and provided by RIETI.} from
a point of view of community structure and attempt to understand the exhaustive structure of Japanese economy. In the transaction network, each node corresponds to a firm and each link corresponds to a transaction relationship. The transaction network encompasses about all Japanese firms. The reference [5] has analyzed a partial network of manufactures in the same data. To find a community structure in networks, we need a criterion to evaluate the density of connections of a group for a given partition. Newman proposed modularity as such criterion and extracted community structures by finding the division at the highest modularity value. If a network $V$ is divided into $L$ subsets $\{V_1, V_2, \cdots, V_L\}$ which do not overlap each other and are not empty, the modularity $Q$ is defined as

$$Q = \sum_{i=1}^{L} Q_i = \sum_{i=1}^{L} (e_{ii} - a_i^2) = \left\{ \frac{1}{2M} \sum_{l \in V_i} \sum_{m \in V_i} A_{lm} - \left( \frac{1}{2M} \sum_{l \in V_i} \sum_{m \in V} A_{lm} \right)^2 \right\},$$

where $A_{lm}$ is the number of link(s) between node $l$ and node $m$: an element of adjacent matrix. The symbol $a_i$ denotes the number of links that connect into subset $V_i$. The symbol $e_{ii}$ represents the link density within subset $V_i$. When the subsets $\{V_1, V_2, \cdots, V_L\}$ are selected randomly, $e_{ii}$ is canceled out by $a_i^2$; the term $a_i^2$ gives the expectation value of the link density in subset $V_i$. Newman’s modularity thus compares the actual link density in a subset with its expectation value. A situation with $Q \approx 0$ indicates that the network has no statistically significant communities as compared with the randomly connected network and $Q \approx 1$ corresponds to a network which is partitioned into modules almost perfectly.

The modularity maximization has a fundamental problem; it is NP-complete task [6]. The number of ways to divide a network is given by the sum of the Stirling number of the second kind and so increases super-exponentially with the number of nodes. The true optimization of $Q$ over all possible divisions is very costly. Since the exhaustive search of all possible division would be actually impossible for such a large network as the transaction network with submillion nodes, approximate heuristic methods are usually employed, including greedy agglomeration [7, 8, 9], simulated annealing [10, 11, 12], extremal optimization [13] and spectral methods [3]. In this paper, we employ greedy agglomeration and simulated annealing because of computational performance for large-scale networks.

We explain two methods of finding the community structure briefly in Section 2. The results of the community analysis are given in Section 3. Subsequently, in Section 4, we compare the community structures with the visualized images of the transaction network obtained by molecular dynamics (MD) method 2. A visualization image gives us an intuitive idea on a structure of large complex networks. Therefore, it is desirable that there is tight relation between the obtained community structure and the visualized image. We report that the community structures are well reproduced in the picture visualized by MD. In Section 5, we elucidate to what extent the attributes (location, industry sector) of firms are reflected in communities of the transaction network. Finally we conclude this paper in Section 6.

2. Method of extracting community

We first use the method due to Clauset et al. [8], which makes the optimization fast by using a greedy algorithm. In the initial state of this method, each node belongs to an isolated subset. Every link is calculated the difference $\Delta Q$ between the modularity given when both ends of the link are merge to form a new subset and that given by the current subsets. We search the highest $\Delta Q$ and merge both ends of the link. This merging process is repeated until there is

---

2 ref. Kamehama et al. “Structure analyses of a large-scale transaction network through visualization based on molecular dynamics”
Figure 1. The greedy method proposed by Clauset et al. (ref. [8]).

Figure 2. The bisection method worked out here.

no link giving $\Delta Q > 0$. In this paper, this bottom-up algorithm is called by “greedy method”. The outline of the greedy method is shown in Figure 1. Though the greedy method is costless, it may not sufficiently optimize the modularity because greedy optimizations tend to be trapped by a local maximum especially for large-scale systems.

Alternatively, we work out a top-down algorithm that implements the simulated annealing. In the initial state of this method, all nodes in the network belong to one community. We search a dual partitioning that gives the highest modularity value. In the initial state, each node is assigned randomly subset $V_A$ or $V_B$. Then one node is selected randomly, i.e., belonging to $V_A$ and is tested to transfer it from $V_A$ to $V_B$. We calculate difference $\Delta Q$ between the modularity given when the selected node is transferred and that given by the current subsets. If node $l$ transfers from $V_A$ to $V_B$, the deference of modularity $\Delta Q$ is calculated as

$$\Delta Q = 2(e_{lB} - e_{lA}) - 2a_l(a_B - a_A + a_l), \quad (2)$$

where $e_{lA}$ ($e_{lB}$) denotes the number of links that connect between node $l$ and nodes in subset $V_A$ ($V_B$) as

$$e_{lA} = \frac{1}{2M} \sum_{m \in V_A} A_{lm}, \quad (3)$$

and $a_l$ denotes the number of links that connect to node $l$ as

$$a_l = \frac{1}{2M} \sum_{m \in V} A_{lm}. \quad (4)$$

The deference $\Delta Q$ is yielded from local state around node $l$. Consequently, this top-down algorithm can be applied to large-scale networks such as the submillion transaction network.
The probability of transition is given by

\[ P(\Delta Q, T) = \begin{cases} 
\exp\left(\frac{\Delta Q}{T}\right) & \text{if } \Delta Q < 0 \\
1 & \text{(otherwise)} 
\end{cases} \]  

(5)

where \( T \) plays a role of temperature. If \( T \) is high, the probability of transition tends to become high even if its transition decrease modularity. We repeat transition steps and cool down gradually. As \( T \) decreases, the subsets \( V_A \) and \( V_B \) are fixed. Eventually, we obtain a dual partitioning that gives the highest modularity value. Each of subset \( V_A \) and \( V_B \) is divided into two parts in the same way. If modularity do not increase any more by splitting any subset, the recursive sequence is terminated. This top-down algorithm is called “bisection method” in this paper. The bisection method is schematically shown in Figure 2.

### 3. Results

We are now ready to report results of the community structures in the Japanese transaction network. To compare the two optimization methods, we also analyze smaller sample networks. One of the sample networks is power grid network in a western part of US [14]. Nodes in the power grid network represent generators, transformers and substations. Links in the power grid network represent high-voltage transmission lines. The second sample network is a subset of the transaction network consisting of firms which have employees equal to or more than 500. The subset network can be regarded as the network of large-sized firms in Japanese society because the number of employees is one of the indices of the firm size.

The results of comparison as regards effectiveness of two optimization methods are shown in Table 1. In the results of the power grid network, both methods give almost same values of modularity that exceed 0.9. Furthermore, both methods give almost same numbers of communities.

On the other hand, in the results of the two transaction networks, the bisection method works better. These transaction networks are larger network size than the power grid network. The subset of the transaction network consisting large firms has nearly as many nodes and about four times as many links as the power grid network has. We consider that the greedy method tends to be trapped by a local maximum especially for large-scale networks. The bisection method overcomes this fault.

Next we show the difference of the community size distribution by the two optimization method. Figure 3 represents the cumulative distribution of community size in the power grid network. Difference between the distribution of the greedy method and the bisection method

### Table 1. Results of the community analysis on three networks.

|                      | Power Grid        | Transaction network (whole) | Transaction network (number of employees ≥ 500) |
|----------------------|-------------------|----------------------------|-----------------------------------------------|
| number of nodes      | 4,941             | 780,544                    | 4,851                                         |
| number of links      | 6,594             | 3,196,282                  | 25,832                                        |
| greedy method        |                   |                            |                                               |
| number of communities| 43                | 9,540                      | 38                                            |
| modularity           | 0.934             | 0.539                      | 0.504                                         |
| bisection method     |                   |                            |                                               |
| number of communities| 46                | 1,352                      | 18                                            |
| modularity           | 0.936             | 0.653                      | 0.548                                         |
is not detectable. On the other hand, in the whole transaction network, there is considerable difference between the greedy method and the bisection method as shown in Figure 4. The greedy method gives many small communities than the bisection method. Furthermore, the top three communities by the greedy method are exceptionally larger than other communities. The bisection method, however, suppressed such uneven partition of nodes.

In the transaction network data that this paper deal with, the firms has location, industry sector, sales, or other attributes. We characterize communities by checking attributes of firms in each community. Tables 2 and 3 show the fraction of firms’ attributes in major communities detected by the two methods.

**Table 2. Attributes of firms in major communities detected by the greedy method.**

| rank | size    | prefecture (fraction) | industry sector (fraction) |
|------|---------|-----------------------|----------------------------|
| 1    | 229,031 | Aichi (0.078)         | construction (0.593)       |
|      |         | Tokyo (0.070)         | wholesale/retail trade (0.160) |
|      |         | Osaka (0.065)         | manufacturing (0.147)      |
| 2    | 219,194 | Tokyo (0.190)         | wholesale/retail trade (0.459) |
|      |         | Osaka (0.084)         | manufacturing (0.252)      |
|      |         | Aichi (0.065)         | services (0.098)           |
| 3    | 165,521 | Tokyo (0.153)         | wholesale/retail trade (0.262) |
|      |         | Aichi (0.072)         | construction (0.249)       |
|      |         | Kanagawa (0.065)      | manufacturing (0.227)      |
| 4    | 37,789  | Hokkaido (0.697)      | construction (0.438)       |
|      |         | Tokyo (0.043)         | wholesale/retail trade (0.252) |
|      |         | Aichi (0.021)         | manufacturing (0.109)      |
| 5    | 23,784  | Kanagawa (0.355)      | construction (0.684)       |
|      |         | Tokyo (0.170)         | wholesale/retail trade (0.088) |
|      |         | Shizuoka (0.154)      | real estate (0.085)        |
Table 3. Same as Table 2, but by the bisection method.

| rank | size   | prefecture (fraction) | industry sector (fraction)                  |
|------|--------|------------------------|---------------------------------------------|
| 1    | 80,134 | Tokyo (0.111)          | wholesale/retail trade (0.511)              |
|      |        | Aichi (0.087)          | manufacturing (0.233)                       |
|      |        |                        | eating places/accommodations (0.094)        |
| 2    | 77,488 | Aichi (0.092)          | construction (0.702)                        |
|      |        | Shizuoka (0.091)       | wholesale/retail trade (0.106)              |
|      |        | Tokyo (0.090)          | manufacturing (0.071)                       |
| 3    | 72,094 | Tokyo (0.310)          | manufacturing (0.336)                       |
|      |        | Osaka (0.090)          | wholesale/retail trade (0.275)              |
|      |        | Kanagawa (0.081)       | information/communications (0.137)          |
| 4    | 62,729 | Tokyo (0.118)          | construction (0.628)                        |
|      |        | Aichi (0.109)          | wholesale/retail trade (0.156)              |
|      |        | Kanagawa (0.107)       | manufacturing (0.084)                       |
| 5    | 47,936 | Fukuoka (0.288)        | construction (0.636)                        |
|      |        | Kagoshima (0.140)      | wholesale/retail trade (0.152)              |
|      |        | Kumamoto (0.136)       | manufacturing (0.066)                       |

4. Comparison between the community analysis and the MD visualization

In this section, we compare the community structure detected by the modularity with the visualized image by MD method. Figures 5 and 7 distinguish the different communities using different colors. Figures 6 and 8 also show the visualized images, where we discriminate the largest community from others through change of the color. The nodes that belong to a same community tend to be located closely in the visualized image by MD. This result implies that the community structure determined by the modularity optimization is well reproduced in the network structure obtained by MD. This implication is given by both results of the two optimization method.

![Figure 5](image-url)  

(a) Community structures by the greedy method for (a) the whole transaction network, (b) the partial transaction network with $N_e \geq 500$, (c) the power grid. The base images are the optimized network structures due to MD visualization.
Figure 6. The largest community is illuminated for each of the networks in Figure 5.

Figure 7. Same as Figure 5, but by the bisection method.

Figure 8. The largest community for each of the networks in Figure 7, corresponding to Figure 6.
density of community $i$ in visualized images

$Q_i$ : contribution for modularity by community $i$

bisection (size $> 100$)
bisection (otherwise)

Figure 9. The correlation chart between the contribution for modularity and the density of nodes in visualized images for the whole transaction network: (a) by the greedy method, (b) by the bisection method.

To quantify the good agreement between the community analysis and the MD visualization, we will demonstrate the tendency that highly modularity groups make high density cluster on visualized images. Figure 9 shows the correlation between the contribution to modularity and the density of nodes in visualized images for the whole transaction network. The contribution to modularity is given by $Q_i$ in Equation (1). The density of nodes in visualized images is defined by

$$D_i = \frac{N_i}{L_i^3},$$

$$\langle x_i \rangle = \frac{1}{N_i} \sum_{l \in V_i} x_l, \quad \langle y_i \rangle = \frac{1}{N_i} \sum_{l \in V_i} y_l, \quad \langle z_i \rangle = \frac{1}{N_i} \sum_{l \in V_i} z_l,$$

$$L_i = \sum_{l \in V_i} \sqrt{(x_l - \langle x_i \rangle)^2 + (y_l - \langle y_i \rangle)^2 + (z_l - \langle z_i \rangle)^2},$$

where $x_l$, $y_l$ and $z_l$ represent the coordinate of node $l$ on visualized images, $N_i$ represents the number of nodes that belong to community $i$, $\langle x_i \rangle$, $\langle y_i \rangle$, $\langle z_i \rangle$ represent center of gravity of community $i$, and $L_i$ represents the mean distance from the gravity point about nodes that belong to community $i$. Thus we suppose that $L_i^3$ is the space occupation volume of community $i$. The Pearson correlation coefficient between $Q_i$ and $D_i$ by the greedy method is 0.975 and by the bisection method is 0.921. We observe that the communities which give larger contributions to modularity tend to form more strongly-agglutinated clusters in visualized images.

5. Groups classified by attribute of firms

In this section, we investigate whether classification due to the attributes (location, industry sector) of firms gives dense connected groups or not. Figures 10, 11, 12 and 13 show the visualized images color-coded by regions. In the whole transaction network, the nodes that belong to same region tend to be located closely in the MD images as shown in Figures 10 and 11. However, in the subset of the transaction network, the nodes that belong to same region do not tend to be located closely as shown in Figures 12 and 13.

Furthermore, we calculate the modularity of the partition given by industry sectors. The results are shown in Table 4. The modularity for partitioning according to industry sectors should be lower than the optimized modularity. Figures 14 and 15 show the visualization color-coded by industry sectors that nodes belong to. The firms of construction and finance/insurance
are located closely, however, the firms of real estate and some other industry sectors are dispersed as shown in Figures 15.

Figure 10. The MD images of the whole transaction network; color-coded by all prefectures, which are drawn from different view angles.

Figure 11. The MD images of the whole transaction network: in which (a) Tokyo, (b) Niigata, and (c) Okinawa are illuminated separately.

Figure 12. The MD images of the subset of the transaction network; color-coded by all prefectures, which are drawn from different view angles.
Figure 13. The MD images of the subset of the transaction network: in which (a) Tokyo, (b) Niigata, and (c) Okinawa are illuminated separately.

Figure 14. The MD images of the whole transaction network; color-coded by industry sectors, which are drawn from different view angles.

Figure 15. The MD images of the whole transaction network: in which (a) construction, (b) finance/insurance, and (c) real estate are illuminated separately.
Table 4. Modularity of the partition given by attribute.

|                         | Transaction network (whole) | (number of employees ≥ 500) |
|-------------------------|-----------------------------|-----------------------------|
| **prefectures**         |                             |                             |
| number of groups        | 47                          | 47                          |
| modularity              | 0.483                       | 0.109                       |
| **industry sectors**    |                             |                             |
| number of groups        | 18                          | 18                          |
| modularity              | 0.175                       | 0.134                       |

We calculate the modularity of the division by prefectures and confirm whether the firms which belong to same area have dense connections. The results are shown in Table 4. In the whole transaction network, the modularity of the division by prefectures is 0.483. The transaction network has highly-modularized groups that contain geographically close firms because the modularity value 0.483 is close to 0.539: the modularity of the community determined by the greedy method.

On the other hand, in the subset of the transaction network consisting large firms, the modularity of the division by prefectures is 0.109; this value is remarkably less than the modularity of the community determined by the greedy method, 0.504. Thus the large firms that employ many people tend to have nation-wide connection. We conclude the difference between Figures 11 and 13 is caused by the difference of the modularity. If a partition gives a high modularity, the MD image shows the partition as cluster. If a partition gives a low modularity, the visualization shows dispersed color configuration. The MD image is helpful to obtain an outline of the community structure.

We investigated the tendency that highly modularity groups makes high density cluster on visualized images as previous section. Figure 16 shows the correlation between the contribution for modularity $Q_i$ and the density of nodes $D_i$ in MD images for the whole transaction network. The Pearson correlation coefficient between $Q_i$ and $D_i$ by the prefectures is 0.976 and by the industry sectors is 0.880. The groups which give large contribution to modularity tend to make strongly agglutinated clusters in visualized images, in the same way as the community structure extracted. The firms of construction and finance/insurance are located closely on the visualized

Figure 16. The correlation between the contribution for modularity and the density of nodes in MD images for the whole transaction network: (a) by the prefectures, (b) by the industry sectors.
images because they give large contribution to modularity. However, some other groups of nodes belonging to same industry sectors give lower contribution to modularity compared to the result of prefectures. Eventually, the total modularity obtained by summing the partial contributions through all industry sectors is lower than prefectures.

6. Conclusion
We analyze the transaction network of submillion Japanese firms by detecting its community structure. For large networks, the method of community detection is restricted to costless algorithm. We compare the two optimizing method because the optimization may be trapped by a local maximum. As a result, the bisection method works better than the greedy method especially on large-scale networks as the nation-wide transaction network. The community structure obtained by the modularity optimization is reflected in the network structure given by molecular dynamics. A partition that has high modularity is visualized as a cluster. We confirm that the community structure is shown as clusters by the color-coded images.

In the visualized images of the whole transaction network by the molecular dynamics, the firms in a same prefecture or region tend to be located closely. Thus we elucidate the modularity of the partition by each firm’s address and confirm that the modularity is close to the value obtained by the optimizing method. The transactions between firms tend to be confined in same prefecture or region. On the other hand, in the visualized images of the subset of the transaction network consisting large firms, the firms in same prefecture or region tend to disperse. The modularity of the partition by each firm’s address in the subset of the transaction network is far from the value obtained by the optimizing method. Large firms are connected through transaction over local boundary. The visualization by molecular dynamics is helpful to illustrate the community structures of large-scale complex networks.

Acknowledgment
We thank the Research Institute of Economy, Trade and Industry (RIETI) for having provided us with the data on transaction relationship between firms in Japan. It was critical for us to carry out the present study.

References
[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. 74 (2002), 47.
[2] M. E. J. Newman, SIAM Rev. 45, 167 (2003).
[3] M. E. J. Newman, PNAS 103 (2006), 8577.
[4] M. E. J. Newman and M. Girvan, Phys. Rev. E 69 (2004), 026113.
[5] Y. Fujiwara and H. Aoyama, 0806.4280[arXiv].
[6] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Transactions on Knowledge and Data Engineering 20 (2008), 172.
[7] M. E. J. Newman, Phys. Rev. E 69 (2004), 066133.
[8] A. Clauset, M. E. J. Newman and C. Moore, Phys. Rev. E 70 (2004), 066111.
[9] V. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, J. Stat. Mech. p. P10008 (2008).
[10] R. Guimera and L. A. N. Amaral, Nature 433 (2005), 895.
[11] A. Medus, G. Acuña and C. O. Dorso, Physica A 358 (2005), 593.
[12] J. Reichardt and S. Bornholdt, Phys. Rev. E 74 (2006), 016110.
[13] J. Duch and A. Arenas, Phys. Rev. E 72 (2005), 027104.
[14] D. J. Watts and S. H. Strogatz, Nature, 393 (1998), 440.