Hubs and Authorities on Japanese Inter-Firm Network: Characterization of Nodes in Very Large Directed Networks

Takaaki Ohnishi,[1,*)] Hideki Takayasu[2] and Misako Takayasu[3]

[1]*Graduate School of Law and Politics, The University of Tokyo, Tokyo 113-0033, Japan*
[2]*Sony Computer Science Laboratories, 3-14-13 Higashigotanda, Tokyo 141-0022, Japan*
[3]*Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan*

Japanese inter-firm network which consists of about one million firms and four million directed links is analyzed by evaluating PageRank, and the authority and hub scores of HITS. We show scale-free degree distribution, power-law distribution of firm-size, fat-tailed distribution of growth rate, positive correlation between degree and firm-size, and similarities between degree and these scores. By comparing with randomized networks, we find that (i) the firm with large firm-size tends to have large PageRank, and small authority and hub scores, and (ii) PageRank correlates significantly with the growth rate which is hardly observed with other quantities.

§1. Introduction

To understand economic activity such as dynamics of firm's growth, system-level study of complex interactions is important. Essence of the system lies in dynamics and it cannot be described by merely investigating components of the system. Thus, beyond the characteristics of individual component, empirical analysis of the network structure of interactions is necessary. Nevertheless, it is only recently that these analyses have been performed by using huge amount of electronically stored data in the economic system.[1),2)]

Traditional measures of topology of complex network, such as clustering coefficient, the shortest path, and others, have been widely employed. These measures are intended for undirected network or ignore link directions. However, to understand the economic system deeper, information about the direction of interactions should be also taken into account. This can be made by Google's PageRank algorithm[3)] and the HITS algorithm of Kleinberg.[4)] Both of these algorithms are designed to compute the score of each node according the importance in the directed network. These quantities are important because they are associated with simple dynamic processes taking place on complex network and the principal eigenvector of the matrix describing the network structure.

In this paper, we make an empirical study on Japanese inter-firm network, which is very large directed network, by using both algorithms. Furthermore, to extract valuable and meaningful information, we compare the scores in real network with

*) E-mail: ohnishi@sat.t.u-tokyo.ac.jp

the corresponding random counterparts.

§2. Methods

2.1. *Data analyzed*

We analyze an exhaustive business dealings data of Japanese firms in 2005 provided by Tokyo Shoko Research, Ltd. (TSR). The data covers about one million firms, that is, practically all active firms in Japan. The data contains information about annual sales, annual incomes, annual profits, number of employee and list of suppliers and customers. In the list of each firm there are only up to 24 Japanese firms as suppliers or customers. However, by accumulating all firms' data there find big firms having several thousands of direct business partners. Note that relationships with general consumer, government or foreign firms are not included in this data. Accordingly, our study is intended for interrelationships among Japanese firms.

For sales, incomes and profits, the data for the years 2003 and 2004 are also available. In this paper, we show only the result of 2005, but we have confirmed almost the same results about 2003 and 2004.

2.2. *Adjacency matrix*

We define Japanese inter-firm network as a directed graph, in which the firms are the nodes and the transactions are the directed links. The network with n nodes is represented by $n \times n$ adjacency matrix whose the element A_{ij} is one if there is a money flowing from the node i to the node j, or equivalently if the node j buys something from the node i, and zero otherwise. We require no self-link $A_{ii} = 0$, but allow bidirectional links $A_{ij} = A_{ji} = 1$. In this paper, we consider the direction of money flow. Thus, indegree and outdegree of the node i are defined by $\sum_{j=1}^{n} A_{ji}$ and $\sum_{j=1}^{n} A_{ij}$, respectively.

2.3. *PageRank*

PageRank algorithm is based on the idea that a node should be highly ranked if other highly ranked nodes contain direct link to it. PageRank is defined as a stationary distribution of a random walk process on the network with additional random jumps. The initial node of the walker is chosen uniformly at random from all nodes. If the current node i has outlinks, the walker proceeds as follows: with probability $1 - \varepsilon_i$, follow one of outlinks with equal probability; with probability ε_i, select a random node out of all nodes. If the current node has no outlink, the walker moves to a randomly chosen node on the network. The damping factor ε_i weighs the mixture between random walk and random jump. Because our data does not contain all of the transaction relationships, the firm may transact with a firm that is not connected. The random jump imitates this effect that the firm sometimes transacts with a randomly chosen firm from all firms. Since exact value of ε_i is not obvious, we simply set it to $\varepsilon = 0.85$ for all i as is used in many studies.

This process is defined as a Markov chain on the network with transition matrix $(1 - \varepsilon)(M_{ij} + d_j) + \varepsilon/n$, where M_{ij} is row-normalized version of A_{ji}: $M_{ij} = 0$ if

node j has no outlink, otherwise $M_{ij} = A_{ji}(\sum_k A_{jk})^{-1}$, and $d_i = 1/n$ if node i has no outlink, otherwise $d_i = 0$. Since the transition matrix is row-stochastic, aperiodic, and irreducible, there exists a unique PageRank vector p_{a}, defined as the principal eigenvector such that $\sum_j ((1-\varepsilon)(M_{ij}+d_j)+\varepsilon/n)\, p_{\mathrm{a}j} = p_{\mathrm{a}i}$. By letting $p_{\mathrm{a}i} = nx_i(\sum_j x_j)^{-1}$, the PageRank can be rewritten as the solution of the following linear system of equations[5)]

$$x_i \;=\; (1-\varepsilon)\sum_j M_{ij}x_j \;+\; 1/n. \tag{2·1}$$

Note that in this paper PageRank is normalized so that its sum over all nodes of the network is n. To have large p_a, it is important to have not only large indegree but also in-neighbors (the nodes linking to it) with large p_a. Typically, PageRank is computed iteratively, and has been known to converge within very small iterations, even for extremely large networks.

The PageRank can be interpreted as fraction of time that the walker spends on each node. Assuming that money moves randomly on the network, we can regard p_a as the amount of money passing through firms. By considering the network in the opposite direction, namely, the direction of material and service, we can also calculate p_h interpreted as amount of material and service. In this way, by estimating the PagaRank value, we can speculate flow of money, and material and service that cannot be observed directly.

2.4. *HITS*

The HITS (hypertext induced topic selection) algorithm assigns each node two scores: authority a_i and hub h_i. These scores have the following interpretation: a node has a high authority value if it is linked to by many nodes with high hub value; a node has a high hub value if it links to many nodes with high authority value. This recursive definition leads to a set of linear equations

$$a_i \;= \sum_j A_{ji}h_j, \tag{2·2}$$

$$h_i \;= \sum_j A_{ij}a_j, \tag{2·3}$$

whose solutions are the principal eigenvectors of the symmetric positive definite matrices A^TA and AA^T, equivalently, the principal right and left singular vectors of A, respectively. In this paper, a and h are normalized so that its sum over all nodes of the network is n.

Unfortunately, certain network topologies cause this algorithm to return either nonunique or nonintuitive result.[6)] To avoid this problem, we replace A_{ij} by $(1-\varepsilon)A_{ij}+\varepsilon/n$. This corresponds to adding to all nodes a new set of links with small weights, and guarantees the existence of a unique result.

Substituting Eqs. (2·3) into (2·2), we obtain $a_i = \sum_{jk} A_{ji}A_{jk}a_k$, that is, the authority of a node is the sum of the authority of the out-neighbors (the nodes linked to by it) of the in-neighbors of the node. To have large a, it is important

that the suppliers of the customers of the firm have large a. Thus, large a means an inferiority as the supplier. Likewise, large h means that the customers of the suppliers of the firm have large h, implying an inferiority as the customer.

2.5. *Comparison with randomized network*

We evaluated the statistical significance of PageRank, authority and hub values in comparison with randomized networks having the same single node characteristics as the real network: Each node in the randomized networks has the same number of inlinks, outlinks and bidirectional links as the corresponding node has in the real network. This ensures that significant deviation reflects pure characteristics of network structure, which are independent of differences of the number of links.

The statistical significance test was performed by generating $1,024$ randomized networks and estimating the empirical p-value derived from the empirical distribution of the value. To generate randomized networks, we applied the Markov-chain Monte Carlo switching algorithm to the real network on which randomly chosen pairs of connections are repeatedly switched (X1→Y1, X2→Y2 are replaced by X1→Y2, X2→Y1) until the network is well randomized.[7] A value with $p \leq 0.05$ was considered statistically significant.

§3. Results

3.1. *Firm's basic properties*

Figure 1 shows cumulative distributions of firm-sizes, i.e. the probability that a value is larger than a threshold x, $P(> x)$. For all firm-sizes, we observe a power-laws behavior

$$P(> x) \ \propto \ x^{-\alpha} \tag{3·1}$$

for $x > x_{min}$. Table I shows the results for best fits of power-law form using maximum likelihood method and Kolmogorov-Smirnov statistic mentioned in Ref. 8). For sale, income and positive profit, the exponent α is close to unity, whereas for negative profit it is smaller than unity, and for the number of employee it is larger than unity. These scale-free properties are in good agreement with previous findings.[9]

Next, we calculate the annual growth rate

$$r_S \ = \ S(t)/S(t-1) \ , \tag{3·2}$$

where $S(t)$ and $S(t-1)$ are the firm-size in the year t and $t-1$. Here, we neglect the data with negative profit. As shown in Fig. 2, the probability density functions of the logarithmic growth rates display tails fatter than those of a Gaussian distribution;[10] for positive profit and income it appears to distribute according to a Laplace, and for sale it displays tails even fatter than those of a Laplace density.

3.2. *Network structure*

To characterize nodes from network structure, we consider six quantities: indegree, outdegree, p_a, p_h, a, and h. Figure 3 shows cumulative distributions of these quantities. Here, the distributions for randomized network introduced in the previ-

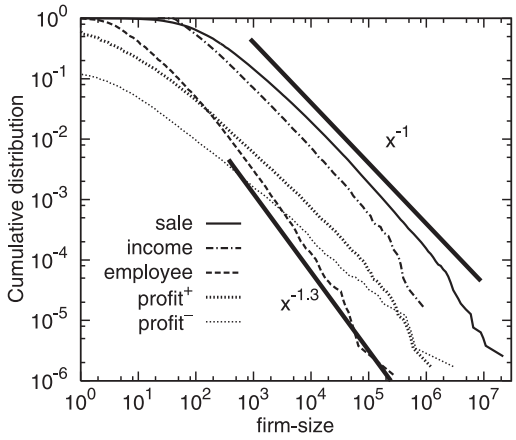

Fig. 1. Cumulative probability distributions of firm-sizes. The unit of sale, income, profit$^+$ (positive profit) and profit$^-$ (absolute value of negative profit) is million yen.

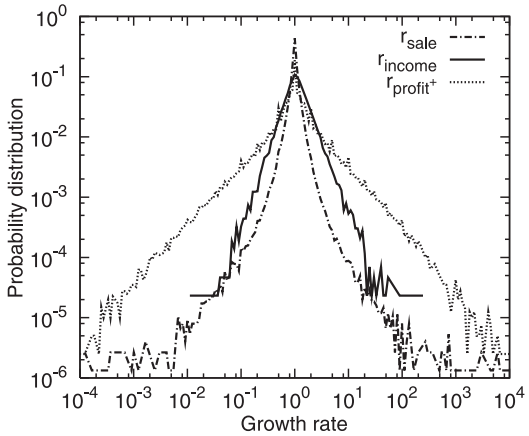

Fig. 2. Probability density distributions of growth rates.

Table I. Exponent value α and lower bound x_{min} calculated by using maximum likelihood method and Kolmogorov-Smirnov statistic mentioned in Ref. 8). The unit of sale, income, profit$^+$ and profit$^-$ is million yen.

Quantity	x_{min}	α
sale	13800	1.025 ± 0.010
income	424	0.967 ± 0.010
profit$^+$	209	0.891 ± 0.008
profit$^-$	25	0.746 ± 0.006
employee	344	1.298 ± 0.014

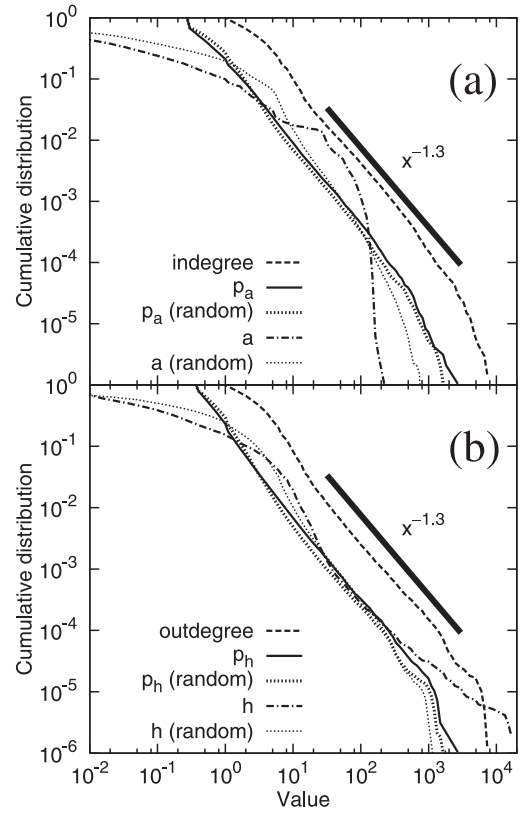

Fig. 3. Cumulative probability distributions of indegree, p_a, and a (a), and outdegree, p_h, and h (b). The distributions for randomized network are shown for comparison.

Table II. Exponent value α and lower bound x_{min} calculated by using the maximum likelihood method and the Kolmogorov-Smirnov statistic mentioned in Ref. 8).

Quantity	x_{min}	α
indegree	20	1.290 ± 0.009
outdegree	25	1.384 ± 0.013
p_a	2.52	1.400 ± 0.006
p_h	4.58	1.381 ± 0.010
h	90	1.015 ± 0.061
p_a (random)	6.82	1.321 ± 0.0004
p_h (random)	21.2	1.285 ± 0.001
a (random)	13.0	1.594 ± 0.001
h (random)	18.3	1.450 ± 0.001

Table III. Kendall's τ. * indicates the significance level at p-value $< 10^{-8}$.

	indegree	*outdegree*	p_a	p_h	a	h
outdegree	0.228 *	–	–	–	–	–
p_a	0.753 *	0.138 *	–	–	–	–
p_h	0.143 *	0.709 *	0.080 *	–	–	–
a	0.665 *	0.103 *	0.528 *	0.052 *	–	–
h	0.157 *	0.622 *	0.090 *	0.469 *	0.093 *	–
sale	0.251 *	0.358 *	0.208 *	0.342 *	0.180 *	0.197 *
income	0.144 *	0.239 *	0.131 *	0.255 *	0.080 *	0.160 *
profit$^+$	0.120 *	0.146 *	0.125 *	0.155 *	0.082 *	0.095 *
employee	0.248 *	0.314 *	0.206 *	0.288 *	0.201 *	−0.070 *
r_{sale}	0.062 *	0.033 *	0.066 *	0.028 *	0.057 *	0.028 *
r_{income}	0.042 *	0.012	0.056 *	0.012	0.051 *	0.029 *
$r_{profit+}$	0.004	0.003	0.026 *	0.008 *	−0.001	0.001

ous section are also shown for comparison. Both indegree and outdegree follow power law distribution with $\alpha \sim 1.3$ (Table II), indicating a directed scale-free network.

The distributions of p_a and $p_a(rand)$ are very similar to distribution of indegree, but there is a striking similarity between $p_a(rand)$ and indegree. This is in agreement with the recent findings[11] that p_a can be approximated by indegree if the degree-degree correlations are weak. Furthermore, as shown in Fig. 4(a), we find an almost linear behavior between p_a and indegree with some fluctuations at small degrees. This visual representation can be confirmed by Kendall's τ, that provides a distribution free test of independence and a measure for the strength of dependence between two variables. As listed in Table III, the correlation between p_a and indegree is larger than any other pair of p_a (p_a–outdegree, p_a–p_h, p_a–a, and p_a–h). Likewise, these features are also observed for p_h, $p_h(rand)$ and outdegree (Fig. 3(b), 4(d), and Table III). Therefore, p_a provides similar information as indegree, and p_h as outdegree.

On the other hand, there is similarity between $a(rand)$ and indegree, and between $h(rand)$ and outdegree, whereas distributions of a and h exhibit different behavior (Fig. 3). In the tail the distribution of a shows exponential decay, but h becomes broader. That is, in comparison with randomized network, there is no large a, but fairly large h exists, showing the asymmetry in the reverse direction. Because large a (or h) means inferiority as supplier (or customer), this indicates that there is no firm that is not at all relied on by customer, but there are some firms that are not at all relied on by some suppliers. However, when viewing the data as a whole, the correlation between a and indegree is larger than any other pair of a, and the correlation between h and outdegree is larger than any other pair of h (see Figs. 4(b),(e), and Table III). Therefore, again, a provides nearly the same information as indegree, and h as outdegree.

3.3. *Correlation between firm's properties and network structure*

We have also analyzed the relation between firm's properties and quantities of network structure: indegree, outdegree, p_a, p_h, a, and h. As shown in Table III,

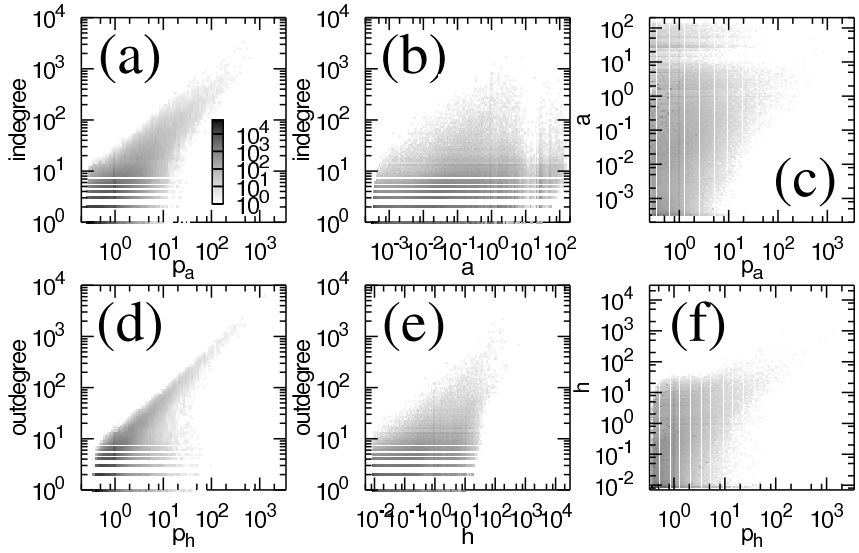

Fig. 4. Scatter-plot of (a) p_a and indegree, (b) a and indegree, (c) p_a and a, (d) p_h and outdegree, (e) h and outdegree, and (f) p_h and h of each node.

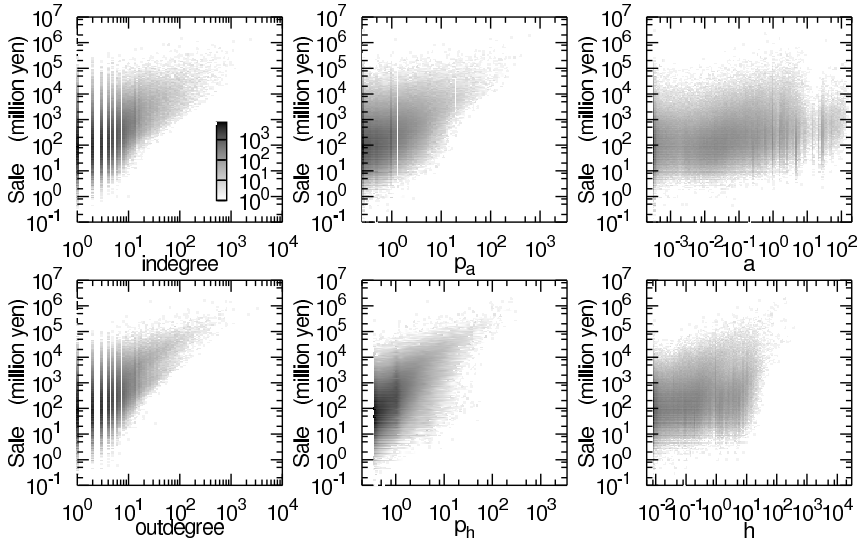

Fig. 5. Scatter-plot of sale and indegree (top left), p_a (top center), a (top right), outdegree (bottom left), p_h (bottom center), and h (bottom right) of each node.

almost all of quantities of network structure have significant positive correlations with firm-sizes. In paricular, outdegree and p_h show strong correlation. These correlations are also confirmed by scatter plots of Fig. 5. In contrast, the correlation between quantities of network structure and growth rates are very weak, though statistically significant.

Eventually, almost all of quantities of network structure show similar behavior in correlation with firm's properties. It is difficult, by authority and hub by them-

selves, to obtain additional information which cannot be captured from indegree and outdegree. This is because p_a and a are highly correlated with indegree, and p_h and h with outdegree.

3.4. *Significant deviation from randomized networks*

To eliminate the effect of the number of degree on firm's properties, we have evaluated the statistical significance of p_a, p_h, a and h by comparing with randomized networks, as described in the previous section. The nodes can be classified into following three types with statistical significance: $x >$ rand (significantly greater than expected value of randomized networks: greater-type), $x \sim$ rand (no significant differences), and $x <$ rand (significantly less than expected: less-type). Table IV lists the proportion of three types of node. By comparing the greater-type and the less-type, we find that firms have following features: small flux of money, large flux of material and service, the superiority as supplier, and the inferiority as customer.

Table IV. Proportion of three types of node.

	$< random$	$\sim random$	$> random$
p_a	43.6%	47.5%	5.2%
p_h	14.3%	61.5%	24.2%
a	56.6%	40.0%	3.4%
h	8.2%	67.0%	24.8%

To investigate the dependence of node type on firm's properties, we study cumulative probability distribution of firm-sizes for each node type. If there is no dependence, these distributions must be identical. Figure 6 shows distributions of sale of each node type for p_a, p_h, a, and h. For p_a and p_h the probability of the greater-type is larger than the other types, whereas for a and h the probability of the less-type is larger. Similar results are obtained for all other firm-sizes. Thus, the node with large firm-size tends to have large PageRank and small a and h in comparison with randomized networks. These results indicate that the firm with large firm-size has the following features: large flux of money, and material and service, superiority as supplier and customer compared with firms of the same number of degrees.

Furthermore, the dependence of node type is also seen in growth rates. Figure 7 shows that for p_h, a, and h the distributions of three types exhibit an identical shape, demonstrating no dependence on the node type, whereas for p_a the probability of the greater-type (or the less-type) is larger (or smaller) than the other types. This demonstrates that the firm with large flow of money tends to increase the growth rate, but the firm with small flow tends to decrease it. PageRank, authority and hub by themselves are independent of the growth rate. Nevertheless, we could observe this dependence by comparing with randomized network.

§4. Discussion and conclusion

The topological roles are important to characterize the nodes of complex network. We have investigated Japanese inter-firm network by evaluating PageRank, and the authority and hub scores of HITS. These measures deduce the importance of a node in a self-consistent way from its nearest and next-to-nearest neighbors by using information contained in link directions. We have seen that the measures

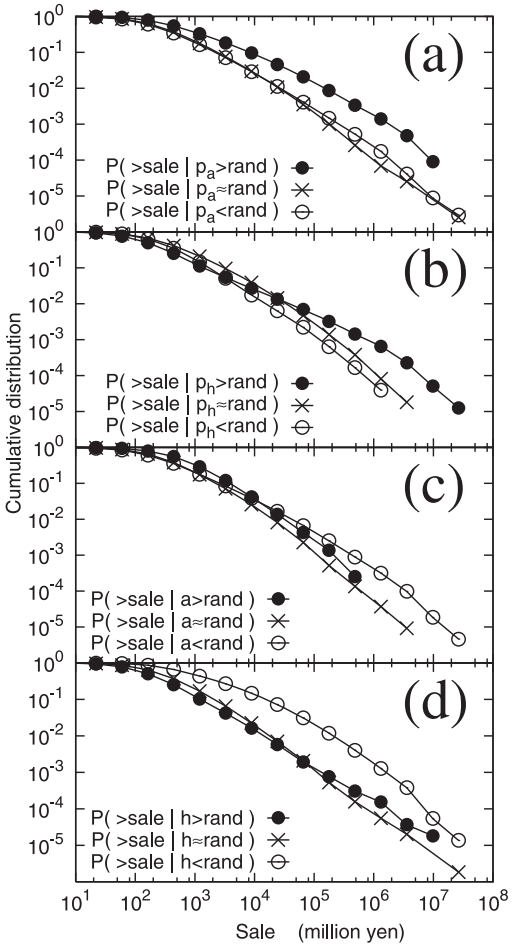

Fig. 6. Conditional cumulative probability distributions of sale for (a) p_a, (b) p_h, (c) a, and (d) h. Three conditions, $x >$ rand, $x \sim$ rand, and $x <$ rand are represented by different symbols. Similar results are obtained for all other firm-sizes.

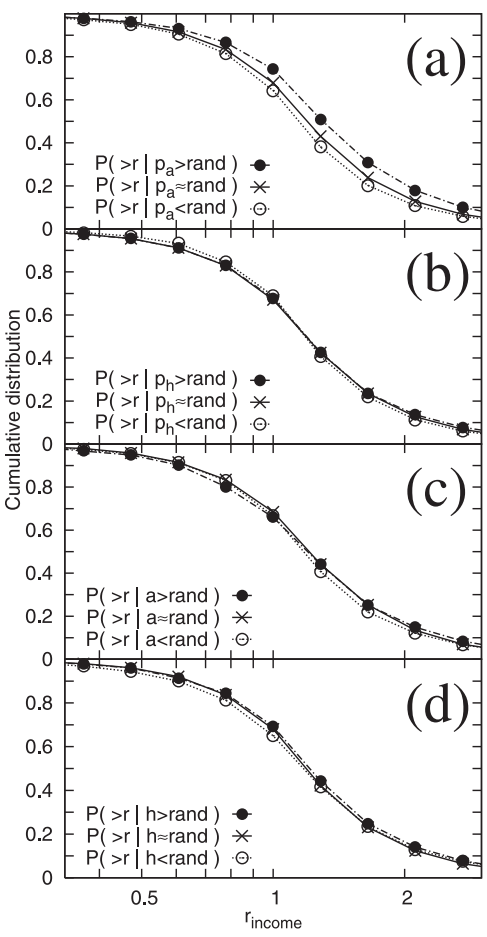

Fig. 7. Conditional cumulative probability distributions of r_{income} for (a) p_a, (b) p_h, (c) a, and (d) h. Three conditions, $x >$ rand, $x \sim$ rand, and $x <$ rand are represented by different symbols. Similar results are obtained for all other growth rates.

are strongly correlated with indegree or outdegree, showing that it is impossible to extract additional information not available to indegree and outdegree from the value of measures. To eliminate the effect of the number of degree, we considered the statistical significance of measures by comparing the value of them in real network with the randomized networks that preserve the number of degree. This allows us to obtain more information about the importance of nodes beyond the degree perspective.

One of the main results is that the firm with large PageRank (large flow of money, and material and service) and small authority and hub (superiority as supplier and customer) has large firm-sizes. A more remarkable result is that p_a (flow of money)

correlates significantly with the growth rate, which is hardly observed for other quantities. These empirical facts, new emergent properties that may arise from the systemic view, provide valuable insight into how firm's interactions give rise to the function and behavior of economic system. More detailed investigations are possible, including networks with weighted links, labels on nodes or links, and the others.

Acknowledgements

The authors are grateful to the Research Institute of Economy, Trade and Industry (RIETI) for allowing us to use the TSR data. This work was financially supported by a Grant-in-Aid for Young Scientists (B) No. 20760053 from the Ministry of Education, Culture, Sports, Science and Technology of Japan (T.O.) and by the Ken Millennium Corporation (T.O.). Numerical calculations were performed by the Earth simulator in Yokohama and Hitachi HA8000 at Information Technology Center, the University of Tokyo.

References

1) Y. U. Saito, T. Watanabe and M. Iwamura, Physica A **383** (2007), 158.
2) Y. Fujiwara and H. Aoyama, arXiv:0806.4280.
3) S. Brin and L. Page, Comput. Networks and ISDN Syst. **30** (1998), 107.
4) J. Kleinberg, J. of the ACM **46** (1999), 604.
5) C. P. C. Lee, G. H. Golub and S. A. Zenios, Technical report, Stanford University (2003).
6) A. Farahat, T. Lofaro, J. Miller, G. Rae and L. Ward, SIAM Journal on Scientific Computing **27** (2006), 1181.
7) R. Milo, N. Kashtan, S. Itzkovitz, M. Newman and U. Alon, cond-mat/0312028.
8) A. Clauset, C. Shalizi and M. Newman, arXiv:0706.1062.
9) K. Okuyama, M. Takayasu and H. Takayasu, Physica A **269** (1999), 125.
10) M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger and H. E. Stanley, Nature **379** (1996), 804.
11) S. Fortunato, M. Boguna, A. Flammini and F. Menczer, Lecture Notes in Computer Science **4936** (2008), 59.