Robustness of Centrality Measures against Link Weight Quantization in Social Network Analysis

Yukihiro Matsumoto
Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
y-matumt@ist.osaka-u.ac.jp

Sho Tsugawa
Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
s-tugawa@ist.osaka-u.ac.jp

Hiroyuki Ohsaki
Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
oosaki@ist.osaka-u.ac.jp

Makoto Imase
Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
imase@ist.osaka-u.ac.jp

ABSTRACT

Research on social network analysis has been actively pursued. In social network analysis, individuals are represented as nodes in a graph and social ties among them are represented as links, and the graph is therefore analyzed to provide an understanding of complex social phenomena that involve interactions among a large number of people. However, graphs used for social network analyses generally contain several errors since it is not easy to accurately and completely identify individuals in a society or social ties among them. For instance, unweighted graphs or graphs with quantized link weights are used for conventional social network analyses since the existence and strengths of social ties are generally known from the results of questionnaires. In this paper, we study, through simulations of graphs used for social network analyses, the effects of link weight quantization on the conventional centrality measures (degree, betweenness, closeness, and eigenvector centralities). Consequently, we show that (1) the effect of link weight quantization on the centrality measures are not significant to infer the most important node in the graph, (2) conversely, 5–8 quantization levels are necessary for determining both the most central node and broad-range node rankings, and (3) graphs with high skewness of their degree distribution and/or with high correlation between node degree and link weights are robust against link weight quantization.

SIMPLEX '12, April 17 2012, Lyon, France

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Science; G.2.2 [**Mathematics of Computing**]: Discrete Mathematics—*Graph Theory*

General Terms

Human Factors

Keywords

Social network, Link weight quantization, Centrality

1. INTRODUCTION

Research on social network analysis has been actively pursued [5, 23]. In social network analysis, individuals are represented as nodes in a graph and social ties among them, such as similarities, social relations, interactions, and flows, are represented as links [5, 23]. The strength of the social ties can be expressed as link weights. The resulting graph is then analyzed to provide an understanding of complex social phenomena that involve interactions among a large number of people.

Among the various indices proposed for social network analysis, centrality measures for the nodes in a graph (degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality) [3, 11] have been widely used in actual analyses [2, 4]. Centrality measures are indices that express the influence of one node on others, and have been used for several purposes, such as discovering a person who plays a central role in a community [2, 4], or inferring activity and leadership levels in a community [13, 20].

However, the graphs used for social network analyses generally contain several errors since it is not easy to accurately and completely identify individuals in a society or social ties among them [6, 8, 14, 16, 17]. For instance, unweighted graphs [2, 8] or graphs with quantized link weights [9] are used for conventional social network analyses since the existence and strengths of social

ties are generally known from the results of questionnaires given to the participants in the experiments. In many social network analyses [2,8,21], only the existence of a social tie is used and its strength is ignored. Even in social network analyses where the strength of a social tie is expressed as a link weight, the link weight may be quantized to take only a few discrete values [9, 10].

Several analyses on the robustness of centrality measures used for social network analyses against the imperfections of graphs (i.e., noise due to random addition and deletion of nodes and links) have been performed [6, 8, 14, 16, 17]. However, since unweighted graphs are used for the analyses in those studies, the effects of ignoring link weight and link weight quantization on centrality measures have not been explored.

In this paper, we study, through simulations of graphs used for social network analyses, the effects of link weight quantization on the conventional centrality measures (degree, betweenness, closeness, and eigenvector centralities).

The remainder of this paper is organized as follows. In Section 2, the experimental methods are explained. In Section 3, we present the experimental results, and discuss the effects of ignoring link weight and link weight quantization on the centrality measures. Finally, Section 4 contains our conclusions and a discussion of future work.

2. METHODOLOGY

We investigate how centrality measures of nodes differ between a weighted undirected graph G and a weighted undirected graph G_n, in which link weights are quantized to take n values.

We randomly generate a weighted undirected graph G by using three network generation models. Since there are several definitions of links (i.e., social ties among individuals) in social network analyses, the topological structures of the graphs used for the analyses are also different from each other. In this paper, we use the following three network generation models to generate graphs with different structural characteristics.

- Community Emergence (CE) model [15]

  The CE model is a network generation model that models the formation of community structure in a social network. A weighted undirected graph generated by the CE model consists of clusters of nodes, which are densely connected to each other by links with large weights. Between the clusters are a small number of links with small weights.

- Weighted Evolution (WE) model [1]

  The WE model is a network generation model that models the evolution of link weights and the topology created by the existing link weights. A weighted undirected graph generated by the WE model has the feature that the distributions of node degree and link weights follow power laws.

- WECS (Weighted Evolving with Community Structure) model [18]

  The WECS model is a network generation model that models formation of community structure in a social network by the evolution of link weights and the formation of the topology depending on the existing link weights. A weighted undirected graph generated by WECS model has a cluster structure, a power-law distribution of node degree, and a power-law distribution of link weights.

For comparison purposes, we also use the weighted undirected graph that is an Erdös - Rényi (ER) graph with link weights randomly assigned according to the Pareto distribution. In what follows, we call the model for generating this graph the "Random graph with Random link weight" (RR) model.

We use two methods for link weight quantization: linear and logarithmic quantization. In linear quantization we divide the range of link weights into n equal sections, and assign an integer between 1 and n to each section. The integer k associated with a weight w is then given by

$$k = \lceil n \times \frac{w}{w_{max}} \rceil, \tag{1}$$

where w_{max} is the maximum link weight in graph G. In logarithmic quantization we divide the range of link weights into n equal sections after a logarithmic transformation, and assign an integer between 1 and n to each section. The possible link weights in graph G_n are

$$w_{max}^{\frac{i}{n}}, \tag{2}$$

for $1 \leq i \leq n$ and the integer k associated with weight w is given by

$$k = \lceil n \times log_{w_{max}} w \rceil. \tag{3}$$

We consider four centrality measures: degree centrality [19], betweenness centrality [19], closeness centrality [19], and eigenvector centrality [3]. Degree centrality, betweenness centrality, and closeness centrality are indices that represent the influence of a node on others based on its degree, the proportion of shortest paths between all other node pairs passing through the node, and the shortest path lengths from the node to all other nodes, respectively. Eigenvector centrality is an index that represents the influence of a node on others based on the centrality of adjacent nodes.

We rank all the nodes in graphs G and G_n by sorting them in descending order of their centrality measures. Then, we calculate the consistency of node rankings [6] between graphs G and G_n. As measures of ranking consistency, we use Top_1, Top_3, $\mathrm{Top}_{10\%}$, $\mathrm{Overlap}_{10\%}$, and R^2 [6]. Top_m is 1 if the most central node in graph G lies in the top m most central nodes in graph G_n, and otherwise it is 0. $\mathrm{Top}_{p\%}$ is 1 if the most central node in graph G lies in the top $p\%$ of nodes in graph G_n, and otherwise it is 0. $\mathrm{Overlap}_{p\%}$ is the number of nodes in both the top $p\%$ of graph G and the top $p\%$ of graph G_n, divided by the number of nodes in either. $R^2$is the square of the Pearson correlation coefficient between centrality measures in graph G and those in graph G_n.

Using the four network generation models, we randomly generated 2,000 graphs, where the number of nodes is 100 and the average node degree is 5, and calculated the averages and 95% confidence intervals of the various ranking consistency indices.

The parameter values used in the network generation models are shown in Tab. 1. Since the 95% confidence intervals were sufficiently small in all cases, only the averages of the ranking consistency indices are shown in the following results.

3. RESULTS

3.1 Effect of Method for Link Weight Quantization and Quantization Level

First, we investigate how the centrality measures are affected by the method for link weight quantization (i.e., linear quantization or logarithmic quantization) and the quantization level.

Table 1: Parameter of network generation models

CE model		WE model		WECS model		RR model	
δ	1.0	δ	1	M	2	p	0.05
p_δ	8.85×10^{-3}	m	3	m_0	3	a	1.5
p_d	1×10^{-3}			α	0.15	b	1
p_r	1×10^{-3}			β	0.1		
time step	25,000			η	0.1		
				m	2		

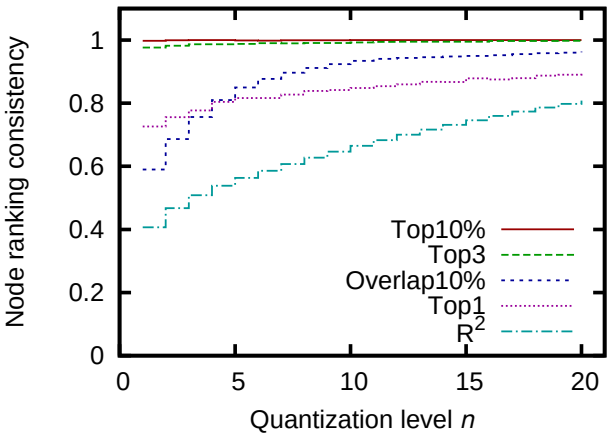

Figure 1: The relation between the quantization level n and the consistency of node ranking (WECS model, closeness centrality, linear quantization)

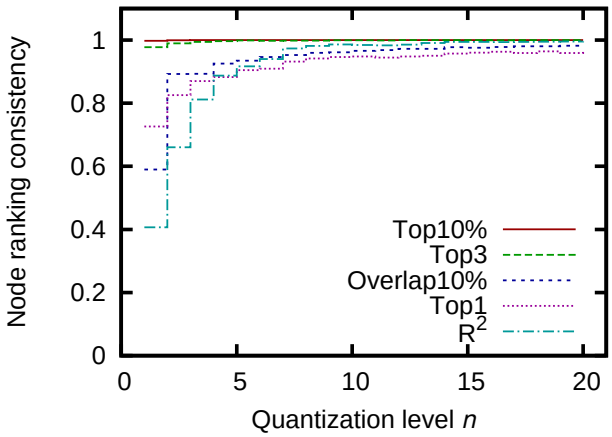

Figure 2: The relation between the quantization level n and the consistency of node ranking (WECS model, closeness centrality, logarithmic quantization)

In what follows, we show the results obtained when using the WECS model as the network generation model and closeness centrality as the centrality measure.

Figure 1 shows the relation between quantization level n and consistency of node ranking (Top_1, Top_3, $\mathrm{Top}_{10\%}$, $\mathrm{Overlap}_{10\%}$, and R^2) when using linear quantization. Note that $n = 1$ is equivalent to ignoring link weight. Figure 2 shows the results for logarithmic quantization.

Comparison of the results for linear quantization (Fig. 1) and for logarithmic quantization (Fig. 2) shows that logarithmic quantization is a more robust method than linear quantization. As we have discussed in Section 2, the distribution of link weights of a graph generated from the WECS model follows a power-law distribution. Hence, the quantization levels can be more effectively utilized by using logarithmic quantization rather than linear quantization and, as a result, logarithmic quantization is more robust. This suggests that, in order to use quantization levels effectively, it is important to design questionnaires appropriately to match the distribution of link weights in the graph used in the social network analysis.

Moreover, we can see in Fig. 2 that, while all indices measuring node ranking consistency are more than approximately 0.9 when the quantization level is five or more, Top_1, $\mathrm{Overlap}_{10\%}$, and R^2take small values when the quantization level is less than five. Since the average value of Top_1 is the proportion of times that the most central node in graph G is also the most central node in graph G_n, the average value of Top_1 is naturally smaller than the averages of Top_3 and $\mathrm{Top}_{10\%}$. In contrast, $\mathrm{Overlap}_{10\%}$ and $R^2$are broad-range ranking consistency indices that do not focus only on the most central node.

These results suggest that the effect of link weight quantization is not so significant when the purpose of social network analysis is to infer the most central node. However, these results also suggest that five to eight quantization levels are necessary for determining both the most central node and broad-range node rankings.

3.2 Effect of Graph Structure

Since there are several definitions of links (i.e., social ties among individuals) in social network analyses, the effect of link weight quantization on centrality measures in graphs with different structural characteristics should be investigated. We therefore generate graphs with different structural characteristics using four network generation models, and investigate the relation between the quantization level n and the consistency of node ranking.

In this investigation, closeness centrality is used as the centrality measure and logarithmic quantization is used as the quantization method.

Figure 3 shows the relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, in graphs produced by the four network generation models.

While the curves representing relations between quantization level n and $\mathrm{Overlap}_{10\%}$ are monotonically increasing regardless of the network generation model, their forms are significantly different.

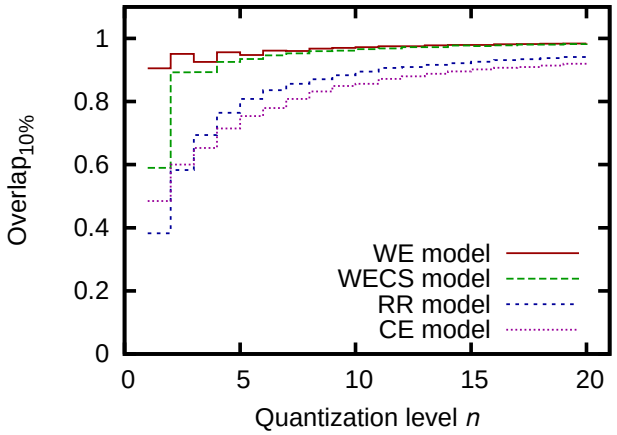

Figure 3: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, in the graphs generated by the four network generation models (closeness centrality and logarithmic quantization)

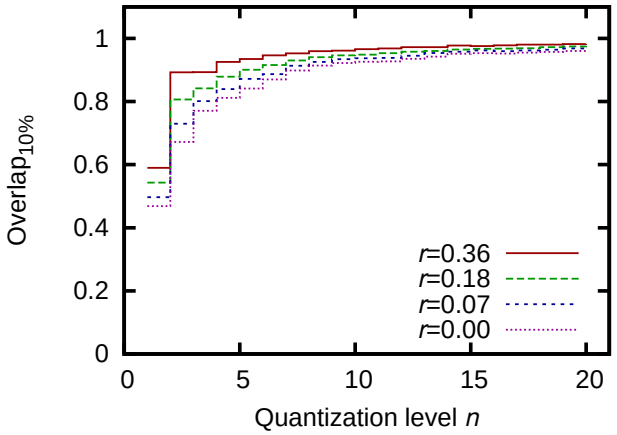

Figure 4: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$ in graphs produced by WECS model and in three kinds of graphs whose correlation between node degrees and link weights is weakened by replaced link weights randomly at graphs produced by the WECS model (closeness centrality and logarithmic quantization)

In particular, the values of $\mathrm{Overlap}_{10\%}$ when link weights are ignored ($n = 1$) are significantly different for each model. Furthermore, when we focus on quantization levels of two or more steps, we find that the four models can be classified into two categories: models in which the consistency of node rankings does not significantly decrease (WE and WECS models), and models in which the consistency of node rankings decreases rapidly (CE and RR models). This observation suggests that structural characteristics of graphs affect the robustness of centrality measures against link weight quantization.

Table 2 shows the averages and standard deviations for various structural characteristics of graphs and several statistics of the link weights calculated from 2,000 graphs randomly generated by the four network generation models. The average and standard deviation of the correlation between node degrees and link weights (i.e., correlation between $w_{i,j}$, which is the weight of link (*i*,*j*), and the sum of degrees of nodes *i* and *j*) are also shown in Tab. 2. The modularity of a graph with respect to some division of the graph into subgraphs measures how good that division is. The mean and standard deviation of the maximum modularity scores are given from clustering obtained by using modularity maximization [7].

Figure 3 and Tab. 2 show that if the skewness of the degree distribution is large and the correlation between node degrees and link weights is strong in graph G, then graph G is robust against link weight quantization. Thus, these results suggest that the effect of link weight quantization on centrality measures depends greatly on the characteristics of graphs used in social network analyses. It is intuitive that a strong correlation between node degrees and link weights in graph G results in the robustness of graph G against link weight quantization, since it is expected that the link weights contain little information in such graphs. Note that eigenvector centrality is reported to be robust against random link rewiring in scale-free networks, for which the skewness of the degree distribution is large [12]. Moreover, the four centrality measures are also known to be robust against the random addition and deletion of nodes and links in scale free networks [17].

To confirm that if the skewness of the degree distribution is large and the correlation between node degrees and link weights is strong in graph G, then graph G is robust against link weight quantization, we investigate the relation between the quantization level n and the consistency of node ranking in graphs whose correlations between node degrees and link weights are weakened. We generate graphs whose correlations between node degrees and link weights are weakened by swapping link weights randomly in graphs generated by the four network generation models. Figure 4 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ in graphs generated by the WECS model and in three kinds of graphs whose correlations between node degrees and link weights are weakened by swapping link weights randomly in graphs generated by the WECS model. This result shows that if the correlation between node degrees and link weights is strong in graph G, then graph G is robust against link weight quantization. Figure 5 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ in graphs whose correlation coefficient between node degrees and link weights are 0 produced by swapping link weights randomly in graphs generated by the four network generation models. This result shows that if the skewness of the degree distribution is large in graph G, then graph G is robust against link weight quantization even when the correlation coefficient between node degrees and link weights in graphs is 0.

These results suggest that graphs with high skewness of their degree distribution and/or with high correlation between node degrees and link weights are robust against link weight quantization.

3.3 Effect of the Type of Centrality Measure

Finally, we investigate that how the effects of link weight quantization differ for the four types of centrality measure. Figures 6 and 7 show the relation between the quantization level n and the

Table 2: The characteristics of graphs produced by four kinds of network generation models (average μ and standard deviation σ)

	CE model		WE model		WECS model		RR model	
	μ	σ	μ	σ	μ	σ	μ	σ
The characteristics of graphs								
average degree	5.47	0.42	5.88	0.00	4.07	0.06	4.95	0.38
average shortest path length	3.59	0.29	2.39	0.04	3.44	0.14	3.03	0.18
clustering coefficient [22]	0.38	0.03	0.11	0.01	0.10	0.01	0.05	0.01
modularity [7]	0.81	0.03	0.24	0.02	0.51	0.03	0.59	0.04
skewness of the degree distribution	1.16	0.45	3.62	0.39	2.70	0.48	0.38	0.25
kurtosis of the degree distribution	1.97	2.31	14.0	3.57	8.36	3.86	0.00	0.67
Statistics of the link weights								
average	482	68.3	1.98	0.00	3.69	0.46	3.02	1.82
standard deviation	898	190	1.73	0.09	8.69	2.26	8.22	27.2
median	155	27.8	1.37	0.04	1.03	0.16	1.59	0.07
skewness	4.13	1.30	3.47	0.39	5.75	1.29	7.90	3.33
kurtosis	24.8	19.5	14.3	3.67	40.0	19.5	84.1	63.0
maximum of link weights	7618	2988	12.8	1.49	78.3	29.8	110.4	2427.8
correlation coefficient between node degrees and link weights	0.18	0.08	0.57	0.04	0.36	0.07	0.00	0.07

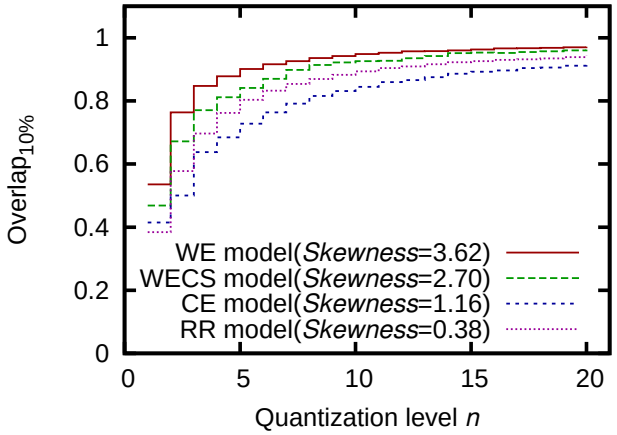

Figure 5: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$ in graphs whose correlation coefficient between node degrees and link weights is 0 by swapping link weights randomly at graphs produced by the four network generation models (closeness centrality and logarithmic quantization)

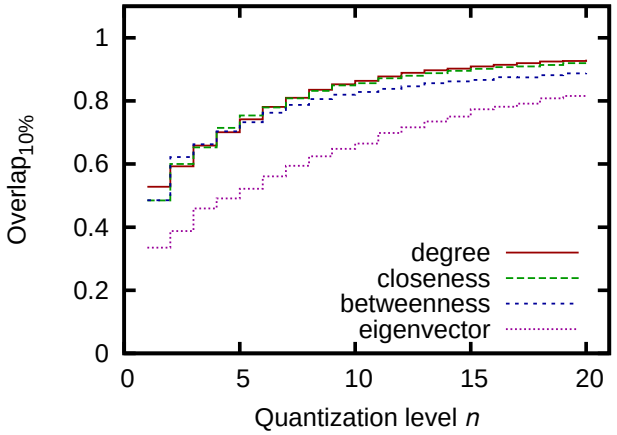

Figure 6: The relation between the quantization level n and the consistency of node rankings, $\mathrm{Overlap}_{10\%}$, in graphs produced by the CE model (logarithmic quantization)

consistency of the top 10% node rankings, $\mathrm{Overlap}_{10\%}$, in graphs generated by the CE model and the WECS model.

Figures 6 and 7 show that the relations between quantization level and node ranking consistency are quite similar for three of the four types of centrality measures. Hence, as we have discussed in Section 3.2, these results suggest that the effect of link weight quantization on centrality depends significantly on the characteristics of the graphs rather than on the type of centrality used in social network analysis. The four types of centrality measure also have a similar robustness against random addition and deletion of nodes and links [6].

However, Fig. 6 shows that when using eigenvector centrality, the consistency of the node rankings, $\mathrm{Overlap}_{10\%}$, is significantly smaller than for other centrality measures. Due to space limitations, the results for RR model are not shown, but we note that the robustness of eigenvector centrality in graphs generated by the RR model is found to be similar to that for graphs generated by the CE model. Further investigation is needed to determine the reason why eigenvector centrality is significantly affected by link weight quantization in graphs generated from the CE and RR models.

4. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the effect of link weight quantization on the centrality measures (i.e., degree, betweenness, closeness, and eigenvector centralities). Consequently we have shown that (1) the effect of link weight quantization on the centrality measures are not significant to infer the most important node in the graph, (2) conversely, 5–8 quantization level is needed to infer

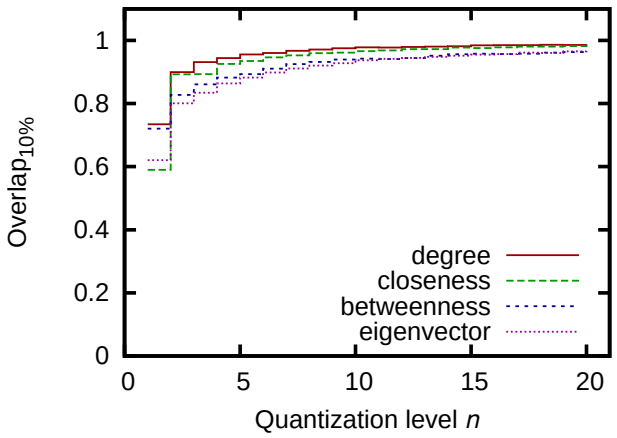

Figure 7: The relation between the quantization level n and the consistency of node rankings, $\mathrm{Overlap}_{10\%}$, in graphs produced by the WECS model (logarithmic quantization)

not only the most important node but also other important nodes since link weight quantization significantly affects the centrality measures, and (3) graphs with high skewness of their degree distribution and/or with high correlation between node degree and link weights are robust against link weight quantization.

We are planning to mathematically analyze the effects on the centrality measures of link weight quantization and of noise in link weights.

Acknowledgements

The authors would like to thank Prof. Masayuki Murata, Naoki Wakamiya, Harumasa Tada, and Yuki Koizumi for their kind support and valuable discussions.

This research was partly supported by "Global COE (Centers of Excellence) Program" of the Ministry of Education, Culture, Sports, Science and Technology, Japan and Grant-in-Aid for Scientific Research (B) (21300022).

5. REFERENCES

[1] A. Barrat, M. Barthélemy, and A. Vespignani. Modeling the evolution of weighted networks. *Physical Review E*, 70(6):66149, Dec. 2004.

[2] D. Batallas and A. Yassine. Information leaders in product development organizational networks: Social network analysis of the design structure matrix. *IEEE Transactions on Engineering Management*, 53(4):570–582, Oct. 2006.

[3] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[4] S. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34, Apr. 2006.

[5] S. Borgatti, A. Mehra, D. Brass, and G. Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, Feb. 2009.

[6] S. P. Borgatti, K. M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136, May 2006.

[7] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, Dec. 2004.

[8] E. Costenbader and T. Valente. The stability of centrality measures when networks are sampled. *Social networks*, 25(4):283–307, Sept. 2003.

[9] N. Creswick and J. Westbrook. Social network analysis of medication advice-seeking interactions among staff in an Australian hospital. *International Journal of Medical Informatics*, 79(6):116–125, Nov. 2010.

[10] R. Cross and A. Parker. *The hidden power of social networks: Understanding how work really gets done in organizations*. Harvard Business Press, 2004.

[11] L. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[12] G. Ghoshal and A. Barabási. Ranking stability and super-stable nodes in complex networks. *Nature Communications*, 2(394):1–7, July 2011.

[13] Y. Kamei, S. Matsumoto, H. Maeshima, Y. Onishi, M. Ohira, and K. Matsumoto. Analysis of coordination between developers and users in the apache community. *Open Source Development, Communities and Quality*, 275:81–92, Sept. 2008.

[14] P. Kim and H. Jeong. Reliability of rank order in sampled networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 55(1):109–114, Jan. 2007.

[15] J. Kumpula, J. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Model of community emergence in weighted social networks. *Computer Physics Communications*, 180(4):517–522, Dec. 2009.

[16] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review*, 73(1):016102, Jan. 2006.

[17] T. L.Frantz, M. Cataldo, and K. Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15:303–328, Dec. 2009.

[18] C. Li and G. Chen. Modelling of weighted evolving networks with community structures. *Physica A: Statistical Mechanics and its Applications*, 370(2):869–876, Oct. 2006.

[19] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, July 2010.

[20] S. Tsugawa, H. Ohsaki, and M. Imase. On inferring leadership in online development community using topological structure of its social network(in japanese). *IEICE Technical Report (IN2010-100)*, pages 19–24, Dec. 2010.

[21] T. Valente, S. Watkins, M. Jato, A. V. D. Straten, and L. Tsitsol. Social network associations with contraceptive use among Cameroonian women in voluntary associations. *Social Science & Medicine*, 45(5):677–687, Sept. 1997.

[22] D. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton Univ Pr, Nov. 2003.

[23] D. J. Watts. A twenty-first century science. *Nature*, 445(7127):489, Feb. 2007.