

Visualizing Japanese Co-Authorship Data

Gavin LaRowe, Ryutaro Ichise, Katy Börner

Indiana University, Bloomington, 47405, USA, National Institute of Informatics, Tokyo, Japan
glarowe@indiana.edu, ichise@nii.ac.jp, katy@indiana.edu

Abstract

This paper reports a bibliometric analysis of evolving co-author networks. Using 5,009 articles covering the years 1993 to 2005 from Transactions D. (Information Systems) of the Institute of Electronics Information and Communication Engineers (IEICE), we attempt to compare the network characteristics for each year, the co-author network characteristics for the entire time span, and the four major components of the entire data set. Finally, we analyze each of these in contrast to extant co-authorship network data and find that the pattern of co-authorship within Information Systems does not change significantly over this time period.

Keywords—co-authorship, network science, japanese

1 Introduction

Statistical and structural analysis of co-authorship networks can be a useful tool for analyzing relationships between various researchers within a scientific field or across fields[1]. Used in conjunction with robust collection of bibliometric data, these analyses can help shed light on social and institutional factors affecting co-authorship and, on a larger-scale, scientific collaboration.

In this paper, we study a sub-set of articles from Transactions D. (information systems) of the Institute of Electronics Information and Communication Engineers (IEICE) covering the years from 1993 to 2005 [3]. We subsequently generate various statistics and network visualizations of the co-authorship networks for the field of Information Systems within Japan over this time span; the Information Systems field per year; and the four largest components over this thirteen year period. Using these analyses, we then attempt to examine whether or not these networks are similar to other co-authorship networks and if any change can be noticed in the pattern of co-authorship within this sub-discipline.

1.1 Prior Research

Most of the prior research done on co-authorship between scientific researchers in Japan was performed by

public policy analysts in the mid-1990s[?]. Prior to a recent study by Diana Hicks in 1995, it was argued by a few scholars that collaboration in Japan between researchers within academia and industry were poor [2]. A few of the primary reasons cited were: administrative or regulatory impediments, the weak applicability of academic research within the commercial sector, insufficient funding that might help promote such collaboration due to short-term policy planning, and the inability to reform the *koza* relationship. The *koza* relationship is a fairly rigid relationship where most collaborative work takes place between full professors, associate professors, assistant professors, and their graduate students.

From a bibliometric perspective, there have also been a few studies done in Japanese regarding co-authorship amongst scientific researchers in Japan, though few place emphasis on the *koza* relationship, which may be taken for granted, when analyzing co-authorship data. Ryutaro Ichise, et Al. have studied co-authorship in Japanese research communities via data mining and graph visualization techniques [5]. Yasuda, et Al. have studied the importance conferences play in Japan regarding co-occurrence of artificial intelligence researchers from data harvested from the web [6]. For prior research regarding network theory, see studies by Newman, Albert, Barabási, and Wasserman [8].

2 Method

2.1 Data

The data for this study is from the Institute of Electronics Information and Communication Engineers (IEICE) covering the years 1993 to 2005. The Institute of Electronics Information and Communication Engineers is considered by many Japanese scholars to be the Japanese analogue to the Institute of Electrical and Electronics Engineers (IEEE).¹

IEICE transaction data is composed of four major areas, or 'Transactions': *A. Fundamentals*, *B. Communications*, *C. Electronics*, and *D. Information Systems*. Five-thousand and nine articles were selected

¹Data hosted by the National Institute of Informatics (NII), Japan.

from Transaction D. (Information Systems) for this study. All of the data was initially converted from the EUC_JP encoding to UTF-8. Below are figures regarding the author and article data extracted from Transactions D. of IEICE from the years 1993-2005:

Total # of years: 13
Total # of articles (1993-2005): 5009
Total # of authors (1993-2005): 12,337
Avg. # of authors per year: 949
Avg. # of articles per year: 385
Avg. # of papers per author (1993-2005): 2.47
Avg. # authors per paper (1993-2005): 3.25
Avg. # of collaborations per author (1993-2005): 3.00
Largest # of authors for a single paper: 18

For each year, unique authors were extracted. Initially, we attempted to extract the networks based on the English transliteration of the Japanese surnames found within the data. Those familiar with East Asian languages may realize this can lead to a many-to-one transliteration error via this approach.

Instead, focus was placed on the Japanese surnames, which are less prone to inaccuracies within this data set. It must be noted, however, that another problem encountered were prior transcription errors within the Japanese surname from when the data was initially input. This can cause some problems for Japanese language parsers and lexical analyzers which may parse data, for example, based on byte-code sequences versus having a fixed word delimiter such as a space in the English language.

2.2 Tools

The data was parsed using a Japanese language parser, fed into a Japanese language lexical analyzer, and placed into a relational database. Various functions to build the nodes, edges, and network files were used to generate the final network files within the database. These files were then loaded as graphs into the R statistical programming language. The *igraph*, *sna*, *cluster*, *e1701* libraries of the R programming language were then used to generate and analyze the statistical properties of all the networks described here.

3 Japanese Co-authorship Networks within Information Systems (IEICE)

3.1 Metrics

To gain some understanding of the underlying graph structures, the following metrics were examined for each of the four items mentioned in the introduction: (a) degree centrality; (b) closeness centrality; (c) betweenness centrality; (d) clustering co-efficient (transitivity); (e) average path length; (f) density; and (g) diameter. Many of these

may be familiar to the reader, though a brief explanation may help those not familiar with these concepts.

Degree centrality computes the in-, out-, or total-degree centrality of a graph, describing the size of a given vertex or node's neighborhood within the larger graph [7]. Closeness centrality describes the distance of each node to all other nodes. Betweenness centrality attempts to describe a node or vertex's position within the network in regard to the communication or *flow* it is able to control. The clustering co-efficient, or transitivity, measures the probability that the adjacent vertices of a vertex are connected [8]. The average path length calculates the average of the smallest number of hops between any two vertices in a graph. Density is the ratio between the number of edges and the number of possible edges in a fully connected graph of the same size. Diameter is the length of the longest geodesic, or the longest of the shortest path lengths between any of the vertices in the graph [9].

The results for these metrics can be found in Table 1. All singletons were removed from the graphs used for analysis. The mean value was determined for all of the centrality scores. The entire matrices for each were used to generate the plots below. To contrast our metrics, we also include co-authorship metrics from Newman and Barabási in Table 2. taken from Albert and Barabási's landmark paper on complex networks [10].

3.2 Analysis

The degree distribution of authors-to-articles does not deviate in character from other co-authorship networks for various other scientific fields and confirms the fat-tailed distribution commonly found in such co-authorship networks (Newman, 2004).

Computing the standard deviation for degree centrality (0.0001), closeness centrality (0.0002), betweenness centrality (0.0171), and average path length (0.139), we found very little deviation between these respective values over the thirteen year period.

The distribution of each centrality measure was examined to determine if any of the networks deviated from the standard normal distribution found in sample data used for the entire network. For this, we used a q-q plot which is commonly used to examine composite, dependent variables (see figures below). As expected, they show very little differentiation.

The clustering co-efficient and the average path length are also often analyzed for co-authorship networks. If a rapid increase in the average number of co-authorships per author or the number of co-authors per paper had occurred, the average path length should decrease (Albert and Barabási, 2004). Both the clustering co-efficients and the average path length remain relatively unchanged over the thirteen year period, i.e. the character of co-authorship for

<i>Network</i>	<i>Size/AU</i>	<i>AR</i>	<i>AC</i>	<i>DC</i>	<i>CC</i>	<i>BC</i>	<i>DE</i>	<i>DI</i>	<i>(k)</i>	<i>l</i>	<i>C</i>
1993	861	402	2.9	0.4599	0.00116	0.03019	0.00053	2	1.208	0.46	0.562
1994	758	377	2.87	0.5646	0.00132	0.00660	0.00053	2	1.045	0.57	0.6777
1995	733	327	2.77	0.3329	0.00136	0.00682	0.00075	2	1.076	0.33	0.6034
1996	900	406	2.94	0.3778	0.00111	0.00889	0.00045	2	1.086	0.37	0.5878
1997	1127	491	3.12	0.3762	0.00089	0.00887	0.00042	2	1.086	0.37	0.6912
1998	1125	466	3.24	0.3733	0.00089	0.00356	0.00033	2	1.037	0.37	0.6462
1999	995	415	2.95	0.2573	0.00100	0.00302	0.00033	2	1.045	0.26	0.6263
2000	1144	452	3.68	0.4336	0.00087	0.04808	0.00026	3	1.329	0.43	0.5745
2001	1194	454	3.66	0.3317	0.00084	0.00838	0.00038	2	1.092	0.33	0.6807
2002	681	257	3.14	0.2702	0.00147	0.05286	0.00028	3	1.514	0.27	0.6316
2003	722	343	3.16	0.3047	0.00139	0.00970	0.00040	2	1.113	0.3	0.6667
2004	924	276	3.63	0.4159	0.00108	0.00541	0.00042	2	1.05	0.42	0.6687
2005	1173	343	3.75	0.3444	0.00085	0.00597	0.00029	2	1.065	0.34	0.6866
Total	12,337	5,009	3.00	1.2900	0.00013	5.33230	0.00017	15	1.29	4.3	0.505
SPIRES	56,627								173	4.0	0.726
NCSTRL	11,994								3.59	9.7	0.496
Math	70,975								3.90	9.5	0.590
Neurosci.	209,293								11.5	6.0	0.760

Table 1: Metrics per year and totals: No. nodes (*Size*), no. articles (*AR*), avg. no. of collaborations per author per year (*AC*), *degree centrality (*DC*), *closeness centrality (*CC*), *betweenness centrality (*BT*), density (*DE*), diameter (*DI*), average degree (*k*), average path length(*l*), and clustering co-efficient (*C*); * = mean value indicated. Other metrics taken from p. 8 of Albert and Barabási’s paper, *Statistical Mechanics of Complex Networks*, in *Reviews of Modern Physics*, 74, 47 (2002).

this data, and perhaps this discipline, has not changed over time.

To supplement these results, we also compared our data to those found in the co-authorship networks analyzed by Newman and Barabási. The IEICE Information Systems data exhibit a high clustering co-efficient similar to those found in the other co-authorship networks. The average path length is very close to the SPIRES network, though smaller when compared to the NCSTRL, Math, and Neuroscience networks. Very similar to the SPIRES network, the degree distribution for the Information Systems network conforms to a power law $P(k) \propto k^{-2.216}$ exhibiting scale-free behavior, while the average path length remains relatively small, supporting a large number of highly-connected hubs in what is ultimately a small-world network [11].

4 Conclusions

Analysis of the co-authorship data shows similar characteristics, such as clustering co-efficient and average path length, found in other co-authorship networks from other scientific disciplines, as shown by Newman and Barabási. Although perhaps assumed and highly data dependent, this may indicate a similar pattern of co-authorship may occur within certain scientific fields, in this case Information Systems, regardless of the country of origin. When time-

series data is available, perhaps a function between cliques and the clustering co-efficient can be applied to characterize whether or not a given yearly co-authorship network is atypical or representative for a given discipline.

5 Acknowledgements

We would like to Thank the National Institute of Informatics (NII), Japan, for hosting our visit. We’d also like to acknowledge Russell Duhon and Soma Sanyal, Indiana University, for their help with this data.

References

- [1] Mark Newman. *Co-authorship networks and patterns of scientific collaboration*. In Proceedings of the National Academy of Sciences, 101: 5200-5205. (2004).
- [2] Diana Hicks. *University-industry research links in Japan*. . Policy Sciences. 26:4 pp.361-395, (1993).
- [3] *Information for authors*.. The Institute of Electronics, Information and Communication Engineers (IEICE). Retrieved March 02, 2007, from http://www.ieice.org/eng/shiori/mokuji_iss.html.
- [4] Shigeru Nakayama and Morris F. Low. *The research function of universities in Japan*. . Higher Education. pp. 245-258, (1997).

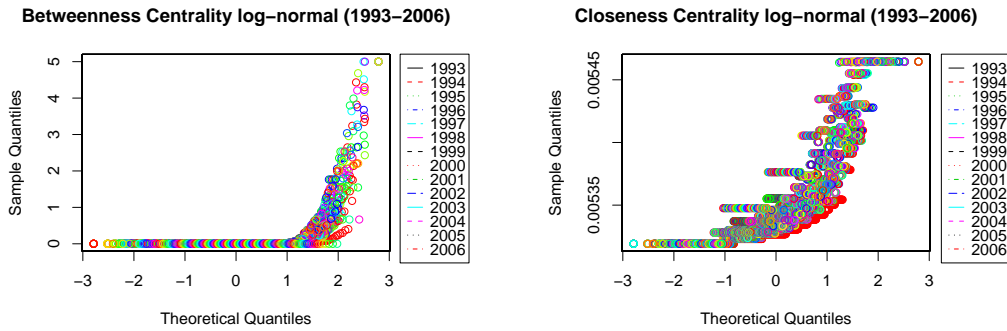


Figure 1: and Betweenness and Closeness Centrality Scores q-q plot (1993-2005)

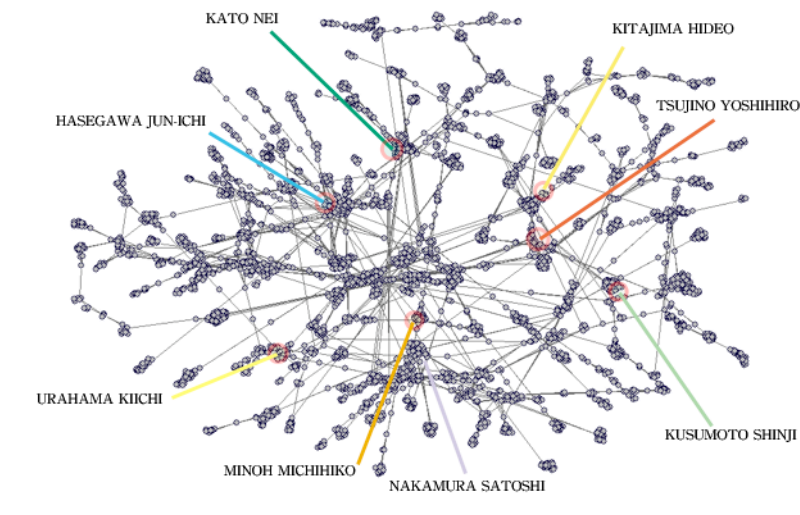


Figure 2: Largest connected component IEICE Transactions D. (1993-2005) with 3961 nodes showing eight of the top co-authors during this period.

- [5] Ichise, R., Takeda, H., and Muraki, T. *Visual Research Community Mining*. In Proceedings of the International Workshop on Risk Mining, pp. 25-34, (2006).
- [6] Yasuda Y., Matuo Y., and Takeda, H. *Analysis on Social Network Structure and Dynamism in AI community*. In Proceedings of the 20th Annual Conf. of JSAI, 1F2-1, in Japanese, (2006).
- [7] L.C. Freeman, *Centrality in Social Networks I.: Conceptual Clarification* Social Networks. 1 pp. 215-239, (1979).
- [8] Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications*. (1994) Cambridge: Cambridge University Press.
- [9] Börner, K., and Sanyal, S. *Network Science*. ARIST. 41:12 pp. 537-607, (2007).
- [10] Albert, R., and Barabási, A.L. *Statistical Mechanics of Complex Networks*. Reviews of Modern Physics. 74, 47 (2002).
- [11] Ebel, H., Davidsen, J., and Bornholdt, S. *Dynamics of Social Networks*. Complexity. 8:2, pp.24-27, (2002).

IEICE Information Systems
Component #2

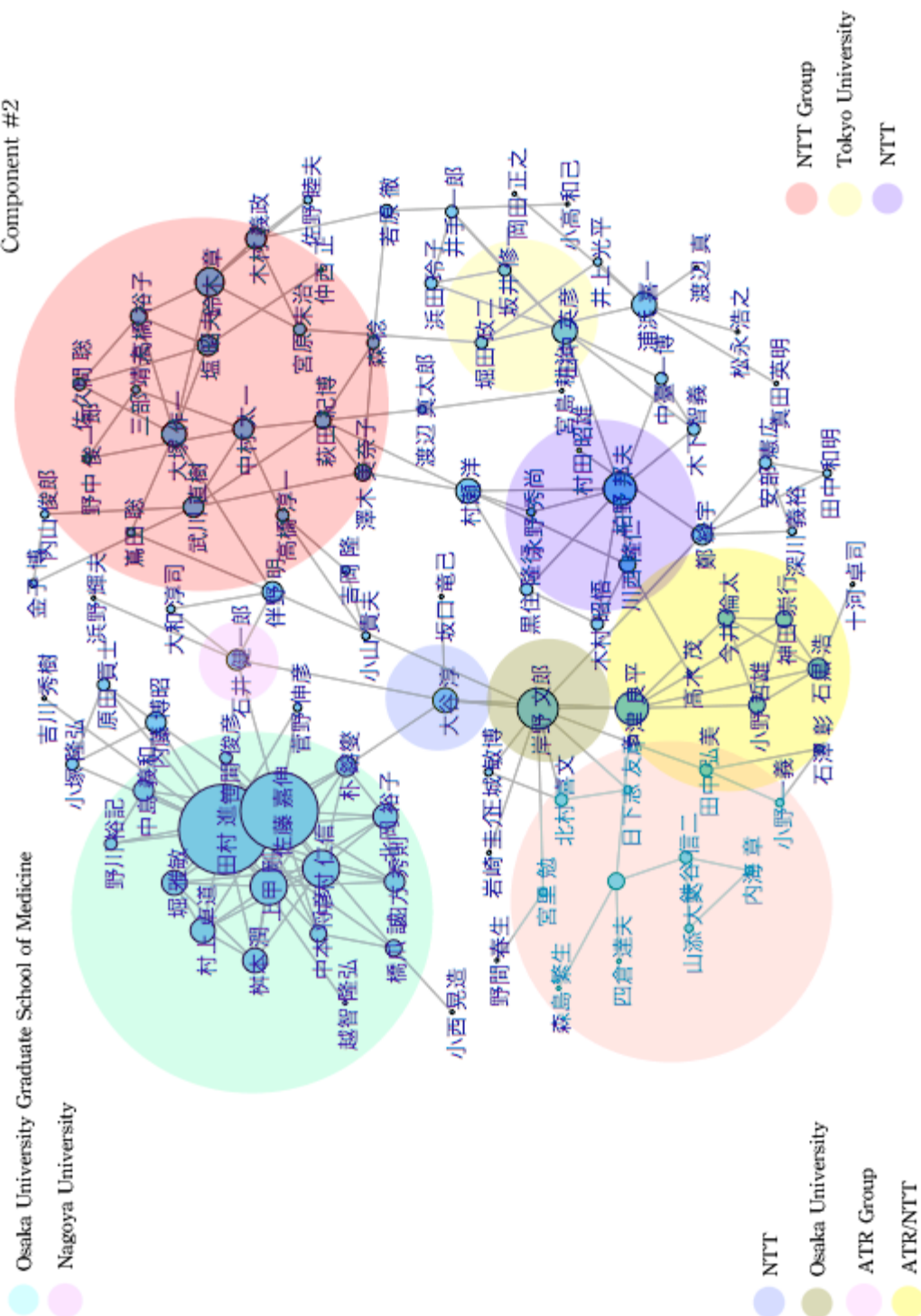


Figure 3: Largest components #2 IEICE Transactions D. (1993-2005; *ellipses denote general area.)

La Rowe, Gavin, Ichise, Ryutaro and Börner, Katy. (2007) Visualizing Japanese Co-Authorship Data. In Proceedings of the 11th Annual Information Visualization International Conference (IV 2007), Zurich, Switzerland, July 4-6, pp. 459-464, IEEE Computer Society Conference Publishing Services.

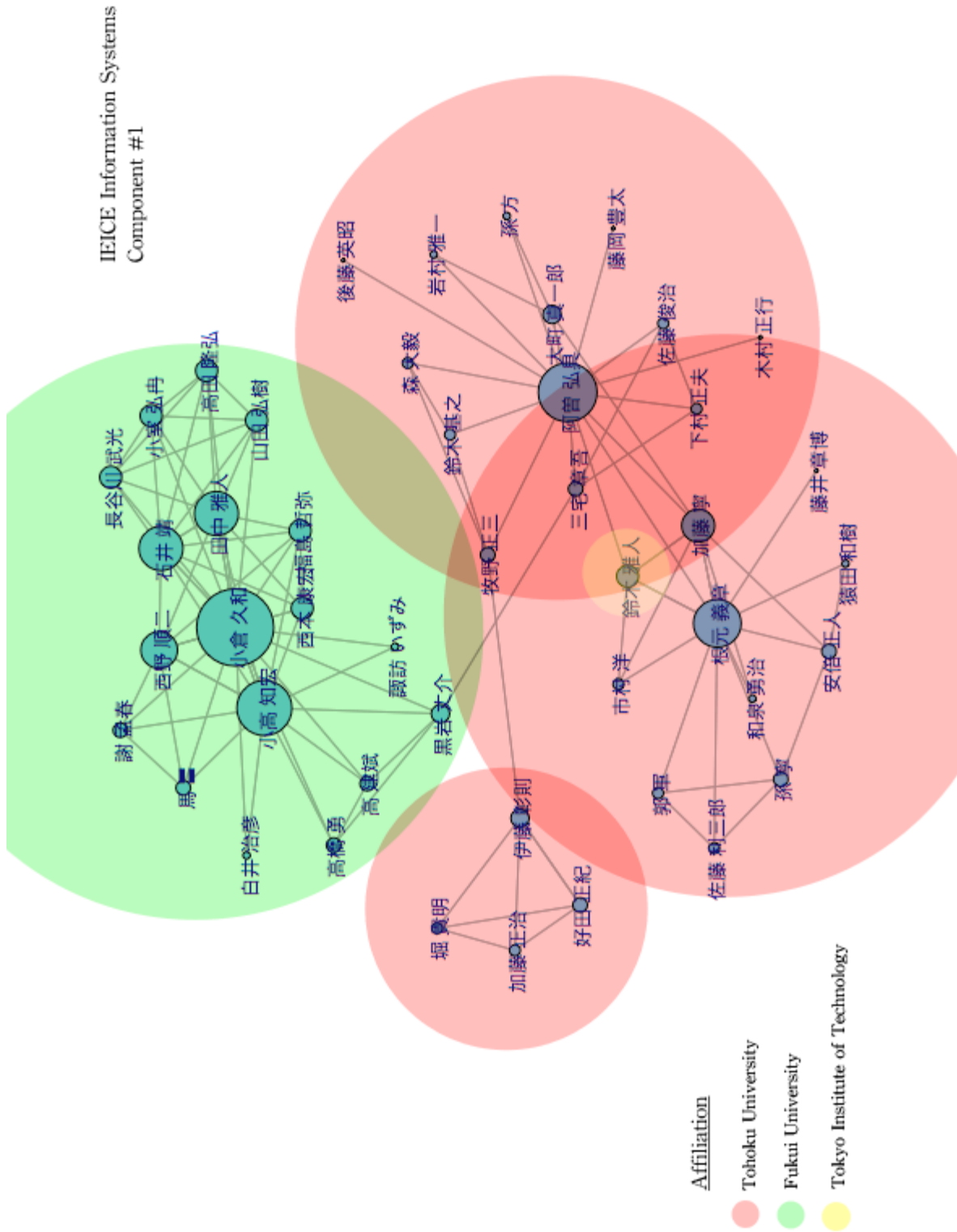


Figure 4: Largest components #1 IEICE Transactions D. (1993-2005; *ellipses denote general area.)