数理解析研究所講究録
第 1644 巻 2009 年 44-54

Investigating Web Structure by Cliques and Stars

Yushi Uno†, Tatsuya Kiyotani and Fumiya Oguri

Department of Mathematics and Information Sciences,
Graduate School of Science, Osaka Prefecture University, Sakai 599-8531, Japan.
†uno@mi.s.osakafu-u.ac.jp

Abstract. The link structure of the Web is generally viewed as the webgraph. In this paper, we investigate the structure of the Web through experiments by utilizing isolated cliques (i-cliques) and isolated stars (i-stars) in contracted webgraphs. The experiments are done into two directions: (i) web structure mining by enumerating i-cliques/i-stars in contracted webgraphs, and (ii) examine various properties of contracted webgraphs by their degree distributions and i-clique/i-star size distributions. As a result, we observed that i-cliques and i-stars can find hidden communities in the 'contracted' Web, which implies that they are still good candidate substructures even for contracted webgraphs. We also observed that i-clique size and i-star size distributions in contracted webgraphs show power-law. We assert that this kind of self-similarity implies the true scale-freeness.

Keywords. isolated cliques/stars, scale-freeness, self-similarity, webgraph, web structure mining.

1 Introduction

It is well known that the webgraph, which represents the link structure of the Web, is one of typical examples of the so-called 'scale-free networks'. There have been a great number of research into a direction of scale-freeness, e.g., investigating the scale-free properties of real world networks, constructing network models to explain these phenomena, and so on. For another direction, by regarding the Web as a huge database, it is extremely important not only to obtain primary information but to find hidden information that cannot be found by naive retrievals. It is often called 'web mining', and web structure mining aims to find hidden communities that share common interests in specified topics in the Web, etc. [4, 6, 8], by focusing on the webgraph.

On this model, a set of web pages of a community or its core is usually supposed to constitute a dense subgraph or a frequent inherent substructure in the webgraph, and web structure mining is actually realized by extracting them from the webgraph. As for significant substructures as communities, Kleinberg's hub-authority biclique model [6] is well known and attractive. Some experimental research for this approach try to enumerate (a subset of) bicliques from the webgraph and are successful for mining communities (or their cores) [7, 8, 9]. However, since there exist potentially enormous number of bicliques, it has become quite hard to carry out an exhaustive enumeration and to have effective outcome in the recent Web [11]. Another is a max-flow (or min-cut) approach that finds small cuts separating a specific set of seed pages [4]. However, this also has a drawback in the sense that we need specify seed pages in advance.

Recently, in the enumeration approach, substructures called *isolated clique* (*i-clique*) [5] and *isolated star* (*i-star*) [12] were proposed for structure mining. They are relatively easy to be enumerated mainly because those structures are essentially disjoint one another. This disjointness is reasonable in the sense that i-cliques and/or i-stars can stand for community cores since cores of different communities are usually supposed to be disjoint. A series of experiments on structure mining was performed in [11, 12], and it is reported that i-cliques and i-stars can find some candidates of communities. On the other hand, most of i-cliques and i-stars correspond to menu and index structures in single domains, respectively, and there proposed *i-clique contracted webgraph* and *i-star contracted webgraph* for further mining.

In the light of these preceding research, the main objective of this paper is to investigate properties of the webgraph and to execute web structure mining, by using i-clique and i-star in the i-clique/i-star contracted webgraphs. Section 2 gives a brief review of the preceding research and explains the outline of the experiments. In Section 3, we present a series of preliminary experimental results about the properties of contracted webgraphs and structure mining on them. Especially, we observe the fact of a new kind of scale-freeness, in which the distributions of i-clique/i-star sizes in contracted webgraphs show power-law, and we assert that this kind of 'self-similarity' implies the true scale-freeness. Finally, we conclude the paper and give some additional comments in Section 4. For some technical terms without precise definitions that appear in the rest of the paper, see [11, 12], for example.

2 The Webgraph and the Outline of Experiments

The *webgraph* is a directed graph whose nodes and arcs are (web) pages and (hyper)links among pages, respectively [3]. One of the most important properties of the webgraph is its *scale-freeness*, which implies that the in-degree distribution of nodes shows the power-law [1, 2]. Here, the *power-law distribution* is the one in which the probability that an stochastic variable X (in-degree of nodes, in this case) equals k is denoted by $\Pr(X = k) \propto 1/k^{\gamma}$ ($\gamma \geq 1$), and the index γ is called its *scaling exponent*. If we draw this distribution in log-log scale by setting in-degree and its frequency on x- and y-axis, respectively, it shows a straight line with its gradient $-\gamma$. It is known that many of the real-world networks have this property, and they are called *scale-free networks* in general.

For our experiments, we prepare a webgraph from the data collected by (The Stanford) WebBase Project [10], which is widely used for various kinds of web experiments. Here, since the graph constructed from the original web data is not necessarily simple, we apply two preprocesses: (1) remove loops (links from a page to the same page), and (2) identify multiple arcs with a single arc (links with the same origin/destination). Hereafter, we call the webgraph constructed in this way 'the' (*directed*) *webgraph*. Table 1 shows the information of the acquired web data and the webgraph.

Table 1: Information about the acquired web data and the constructed webgraph.

	webgraph	(undirected)
Host, Port#	WB1, 7008	
time of collection	Aug., 2003	
#domains	59,565	
#pages (nodes)	95,821,917	
#links (arcs)	1,737,732,518	345,699,858
intra-domain	1,591,587,293	345,514,732
inter-domain	146,145,225	185,126

According to preceding research [11, 12], in this paper, we consider (isolated) cliques and (isolated) stars as candidate substructures of communities of the webgraph, and use them for investigating the structure of the Web. Since these substructures are essentially defined on undirected graphs, we need regard the (directed) webgraph as undirected. There may be two simple alternatives to do it; (a) regard every arc as an undirected edge, or (b) regard bidirectional arcs as a single edge. From a viewpoint that mutual links can have significant information on the Web, we use a webgraph that consists only of mutual links as single undirected edges by discarding all the one-way links. We refer to this webgraph as the *undirected webgraph*. We give the information of this undirected webgraph also in Table 1.

From the webgraph constructed in this way, we can observe some interesting facts. Fig. 1 gives the in-/out-degree distributions of the webgraph (in log-log scale), and we can confirm that both in- and out-degree distributions show scale-freeness with different scaling exponent. Fig. 2 gives the degree distribution of the undirected webgraph, and we can see that the distribution roughly shows power-law even in the undirected webgraph that only has mutual links as their edges. Moreover, we can find a straight line of gradient 1 that stretches from around degree 200. This is known to be caused by huge cliques that was generated by making the graph undirected [11]. More than 1/3 of links in the original graph also have links of the reverse direction (i.e., mutual links) and that more than 99% of mutual links exist in single domains. This is another evidence that mutual links between different domains and the structures that contain those links can be expected to have significant information from the viewpoint of structure mining.

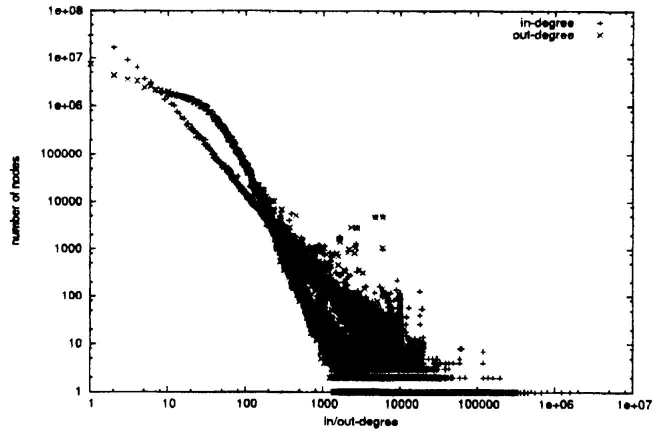

Figure 1: In- and out-degree distributions of the webgraph.

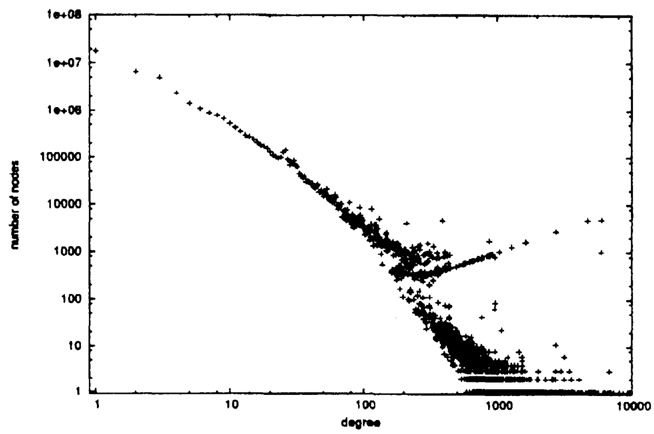

Figure 2: Degree distribution of the undirected webgraph.

A series of web structure mining experiments were performed by using i-cliques and i-stars in [11, 12]. Since we perform subsequent experiments motivated by their observations and results, we here briefly summarize them. For i-cliques, some *inter-domain i-cliques* stand for hidden communities, while almost all the *intra-domain i-cliques* stand for *menu structures* in single domains. Similarly, for i-stars, some *inter-domain i-stars* stand for hidden communities, while almost all the *intra-domain i-stars* stand for *index structures* in single domains. These facts imply the validity of i-cliques and i-stars as candidate substructures for structure mining. On the other hand, since menu structures and index structures seem useless for structure mining, *intra-domain i-clique contracted webgraph* (*ic-contracted webgraph*) and *intra-domain i-star contracted webgraph* (*is-contracted webgraph*) were proposed (for both directed and undirected webgraphs) for further mining and for compact representation of the webgraph. Here, we notice that all i-cliques (i-stars) can be contracted simultaneously, since their vertex sets are independent with very few exceptions.

Now it is quite natural to do experiments for investigating properties of ic-/is-contracted webgraphs and for structure mining by using i-cliques and i-stars in them. We explain the flow of our experiments. Let $\vec{G}_T^i$ and G_T^i be the (directed) webgraph and its undirected webgraph, respectively, after i-th simultaneous contraction of a set of target substructure T from the original (directed) webgraph $\vec{G}$ and its undirected webgraph G, respectively. Here the target substructures that we are interested in and we want to find are i-cliques (C) and i-stars (S) (so, C or S will be substituted in place of T). Then we do the following process:

0. Let $i := 0$ and initialize $\vec{G}_T^i := \vec{G}$.
1. Examine in-/out-degree distributions of $\vec{G}_T^i$.
2. Prepare G_T^i by making $\vec{G}_T^i$ undirected. Examine degree distribution of G_T^i.
3. Enumerate all target substructure Ts in G_T^i and examine the size distribution of T.
4. Examine the corresponding pages of inter-domain Ts as communities.
5. Contract all intra-domain Ts (as vertex sets) in $\vec{G}_T^i$, and let it $\vec{G}_T^{i+1}$. Go to 1.

We iterate the above process some appropriate times, for two target substructures i-clique and i-star. In the next section, we show some preliminary results of these experiments.

3 Properties of Contracted Webgraphs and Structure Mining

In this section, we show experimental results of investigating web structures by using i-cliques and i-stars in ic-contracted and is-contracted webgraphs, respectively. For each of ic-contracted and is-contracted webgraphs, we try to enumerate i-cliques and i-stars, to examine their size distributions, and to analyze their semantic meanings. Among those result, we can observe a new kind of scale-freeness in the size distributions of i-cliques and i-stars, and we emphasize that this imply 'self-similarity' structure of the Web.

3.1 Experiments by Isolated Cliques

First, we show the results obtained by i-cliques in ic-contracted webgraphs.

3.1.1 Properties of IC-contracted Webgraphs

Table 2 shows the number and the size of intra-/inter-domain i-cliques that are found in ic-contracted undirected webgraphs G_C^i ($i = 0, \ldots, 5$). In the first iteration (G_C^0), we can find an extremely large i-cliques such as size 5,940, and this is indeed a huge menu structure in domain `gostats.com`. In contrast, the largest size of inter-domain i-clique was 27. On the other hand, after the first iteration, the largest size of i-cliques found in G_C^i becomes quite small rapidly.

Figure 3 shows distributions of i-clique sizes in ic-contracted undirected webgraphs G_C^i ($i = 0, \ldots, 5$). It is worth noticing that each of these size distributions show power-law with slightly different scaling exponents. We consider that this kind of 'self-similarity' implies the true scale-freeness.

Table 3 shows the number of nodes and arcs in ic-contracted directed webgraphs. As we saw, there are quite large intra-domain i-cliques, for example of size 5,940, in G_C^0, and this clique has 35,277,660 arcs and 17,638,830 edges by itself, and it indeed occupy approximately 2% and 5.1% of the original directed and undirected webgraphs, respectively. Since we contract such intra-domain i-cliques in the first iteration, we can see that the number of arcs in G_C^1 becomes approximately 82% of the one in G_C^0.

Table 2: The number and the size of intra-/inter-domain i-cliques in ic-contracted undirected webgraphs.

	intra-domain			inter-domain		
	size ≥ 3	size = 2	max size	size ≥ 3	size = 2	max size
G_C^0	868,192	2,281,293	5,940	71	3,350	27
G_C^1	23,285	2,393,168	366	75	3,550	27
G_C^2	1,861	2,400,828	37	75	3,555	27
G_C^3	94	2,401,780	35	75	3,556	27
G_C^4	9	2,401,804	5	75	3,556	27
G_C^5	6	2,401,805	5	75	3,556	27

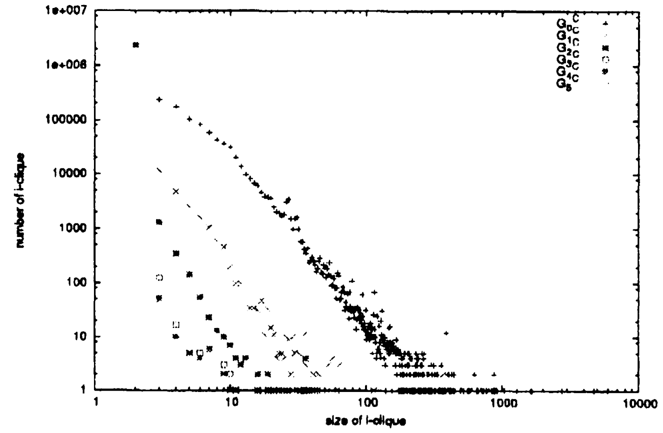

Figure 3: I-clique size distributions in ic-contracted undirected webgraphs.

Figure 4 shows in- and out-degree distributions of ic-contracted directed webgraphs G_C^0 and G_C^5. Although a lot of large i-cliques are contracted and the numbers of nodes and arcs decrease, the distributions are not affected much by them and the degree distributions in G_C^5 seem almost similar to the ones in G_C^0.

Table 3: Statistics of intra-domain ic-contracted directed webgraphs.

	#nodes	ratio	#arcs	ratio	max in-deg	max out-deg
$\vec{G}_C^0$	94,340,605	100.000%	1,737,732,518	100.000%	632,716	10,000
$\vec{G}_C^1$	89,182,661	94.532%	1,423,757,416	81.931%	582,313	10,783
$\vec{G}_C^2$	89,099,138	94.444%	1,420,934,272	81.769%	582,100	10,783
$\vec{G}_C^3$	89,094,059	94.438%	1,420,831,604	81.763%	582,100	10,783
$\vec{G}_C^4$	89,093,814	94.438%	1,420,826,873	81.763%	582,100	10,783
$\vec{G}_C^5$	89,093,794	94.438%	1,420,826,542	81.763%	582,100	10,783

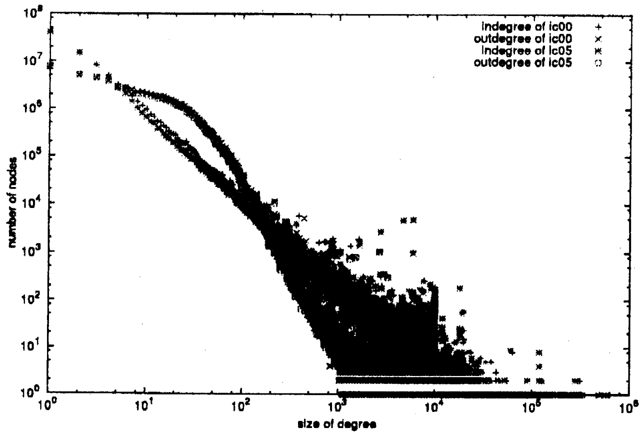

Figure 4: In- and out-degree distributions of ic-contracted directed webgraphs.

Table 4 shows the number of nodes and edges in ic-contracted undirected webgraphs. We can observe almost the similar facts as we observed in Table 3.

Table 4: Statistics of intra-domain ic-contracted undirected webgraphs.

	#nodes	ratio	#edges	ratio	max deg
G_C^0	42,708,162	100.000%	172,849,929	100.000%	9,667
G_C^1	37,525,258	87.864%	79,513,007	46.001%	9,737
G_C^2	37,438,408	87.661%	78,948,474	45.674%	9,737
G_C^3	37,432,773	87.647%	78,933,554	45.665%	9,737
G_C^4	37,432,481	87.647%	78,932,542	45.665%	9,737
G_C^5	37,432,460	87.647%	78,932,508	45.665%	9,737

Figure 5 shows degree distributions of ic-contracted undirected webgraphs G_C^0 and G_C^5. As we pointed out in Section 2, we can see for degree distribution in G_C^0 a straight line of gradient 1 that stretches from around degree 200 caused by huge cliques. We can observe that this straight line has almost diappeared in G_C^5. This is because we contract almost all of these cliques as intra-domain i-cliques.

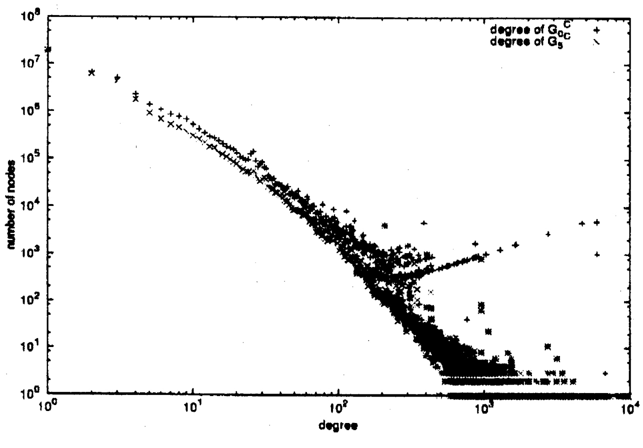

Figure 5: Degree distributions of ic-contracted undirected webgraphs.

3.1.2 Structure Mining on IC-contracted Webgraphs

As we can see in Table 2, we can find some inter-domain i-cliques in G_C^0, and some of them seem to correspond to a community [11]. In G_C^1, we can still find some new inter-domain i-cliques, and Table 5 shows an example of the result. It seems to correspond to a community. However, we cannot find any more new meaningful inter-domain i-cliques after the second iteration. In this sense, i-clique can be a candidate substructure for structure mining in ic-contracted webgraphs in relatively early iterations.

Table 5: An example of an inter-domain i-clique found in G_C^1 and its corresponding URLs. They seem to be related to fishing.

```
http://www.flyfisherman.com/latestissue/index.html
http://www.in-fisherman.com/
http://www.floridasportsman.com/
```

3.2 Experiments by Isolated Stars

We next show the results obtained by i-stars in is-contracted webgraphs.

3.2.1 Properties of IS-contracted Webgraphs

Table 6 shows the number and the size of intra-/inter-domain i-stars that are found in is-contracted undirected webgraphs G_S^i ($i = 0, \ldots, 8$). In the first iteration (G_S^0), we can find an extremely large i-stars such as size 9,664, and this is indeed a huge index strutcture in domain `www.faqts.com`. In contrast, the largest size of inter-domain i-clique was 340. It is somehow surprising that after the first iteration, the maximum size of inter-domain i-stars in G_S^1, size 16,655, once becomes larger than the one in G_S^0, and the maximum size remains relatively large after some more iterations. This is different from the case of i-cliques.

Table 6: The number and the size of inter-/intra-domain i-stars in is-contracted undirected webgraphs.

	intra-domain			inter-domain		
	size ≥ 3	size = 2	max size	size ≥ 3	size = 2	max size
G_S^0	1,034,855	1,108,921	9,664	1,344	1,471	340
G_S^1	105,226	1,026,414	16,655	1,499	1,510	3,675
G_S^2	19,112	1,011,812	12,926	1,563	1,497	3,675
G_S^3	5,012	1,008,385	8,686	1,588	1,495	3,675
G_S^4	1,849	1,006,864	8,867	1,595	1,495	3,675
G_S^5	793	1,006,156	6,623	1,597	1,495	3,675
G_S^6	426	1,005,807	6,055	1,598	1,495	3,675
G_S^7	233	1,005,723	2,688	1,599	1,495	3,675
G_S^8	142	1,005,616	2,394	1,599	1,494	3,675

Figure 6 shows distributions of i-star sizes in is-contracted undirected webgraphs G_S^i ($i = 0, \ldots, 8$). It is again worth noticing that each of these size distributions show power-law with slightly different scaling exponents. We emphasize that this kind of 'self-similarity' implies the true scale-freeness.

Table 7 shows the number of nodes and arcs in is-contracted directed webgraphs. As we saw, there are quite large intra-domain i-stars, for example of size 9,664, in G_S^0. In fact, the total number of satellite nodes of i-stars of sizes ≥ 3 in G_S^0 is 11,967,237, and since we contract such intra-domain i-stars in the

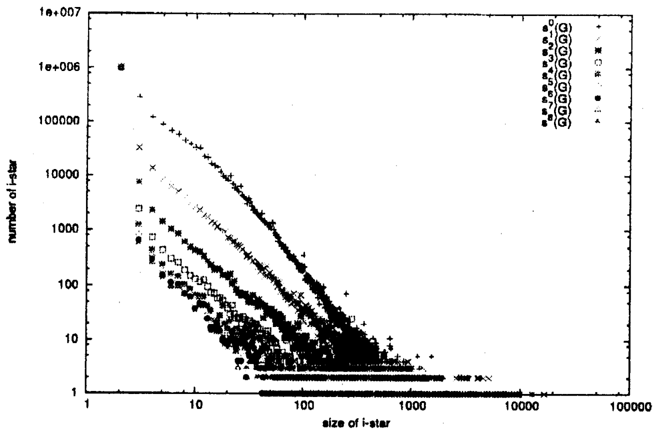

Figure 6: I-star size distributions in is-contracted undirected webgraphs.

first iteration, we can see that the number of nodes in G_S^1 becomes 82,386,313, which is approximately 87.3% of the one in G_S^0.

Table 7: Statistics of intra-domain is-contracted directed webgraphs.

	#nodes	ratio	#arcs	ratio	max in-deg	max out-deg
$\overrightarrow{G}_S^0$	94,340,605	100.000%	1,737,732,518	100.000%	632,716	10,000
$\overrightarrow{G}_S^1$	82,386,813	87.329%	1,513,579,077	87.101%	573,147	16,714
$\overrightarrow{G}_S^2$	79,448.674	84.214%	1,459,876,321	84.010%	560,105	12,929
$\overrightarrow{G}_S^3$	78,466,440	83.173%	1,444,961,271	83.152%	557,137	10,000
$\overrightarrow{G}_S^4$	78,206,046	82.897%	1,440,687,003	82.906%	556,595	10,000
$\overrightarrow{G}_S^5$	78,086,437	82.770%	1,439,095,356	82.815%	556,583	10,000
$\overrightarrow{G}_S^6$	78,026,017	82.706%	1,437,695,068	82.734%	556,577	10,000
$\overrightarrow{G}_S^7$	77,999,369	82.678%	1,437,127,801	82.701%	556,575	10,000
$\overrightarrow{G}_S^8$	77,987,165	82.665%	1,436,608,549	82.671%	556,575	10,000

Figure 7 shows in- and out-degree distributions of is-contracted directed webgraphs G_S^0 and G_S^8. Although a lot of large i-stars are contracted and the numbers of nodes and arcs decrease, the distributions are not affected much by them and the degree distributions in G_S^8 seem almost similar to the ones in G_S^0.

Table 8 shows the number of nodes and edges in is-contracted undirected webgraphs. We can observe almost the similar facts as we observed in Table 7.

Figure 8 shows degree distributions of is-contracted undirected webgraphs G_S^0 and G_S^8. As we pointed out in Section 2, we can see for degree distribution in G_S^0 a straight line of gradient 1 that stretches from around degree 200 caused by huge cliques. However, since we do not contract cliques in this experiment, we can observe that this straight line still remains in G_S^8.

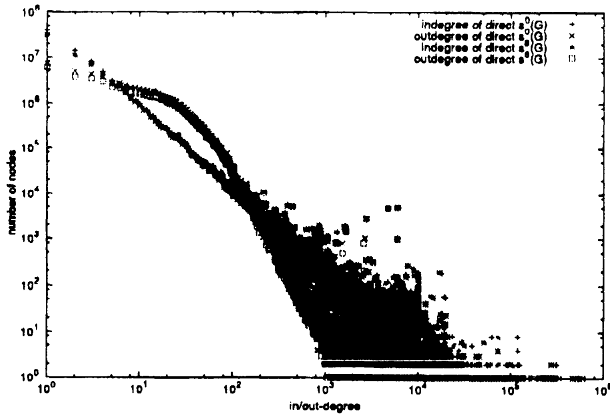

Figure 7: Degree distribution of is-contracted directed webgraphs.

Table 8: Statistics of intra-domain is-contracted undirected webgraphs.

	#nodes	ratio	#edges	ratio	max deg
G_S^0	42,708,162	100.000%	172,849,929	100.000%	9,667
G_S^1	33,547,032	78.549%	166,184,922	96.144%	16,687
G_S^2	31,631,837	74.065%	164,538,447	95.192%	12,926
G_S^3	30,925,377	72.410%	163,873,640	94.807%	9,748
G_S^4	30,791,005	72.096%	163,749,432	94.735%	9,748
G_S^5	30,735,039	71.965%	163,699,778	94.706%	9,748
G_S^6	30,703,318	71.890%	163,669,222	94.689%	9,748
G_S^7	30,688,964	71.857%	163,655,151	94.680%	9,748
G_S^8	30,685,137	71.848%	163,651,616	94.678%	9,748

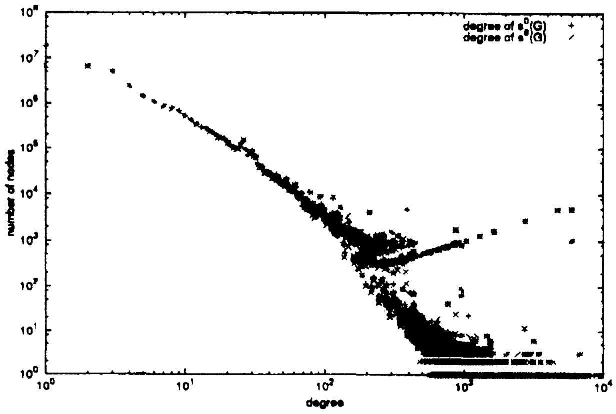

Figure 8: Degree distributions of is-contracted undirected webgraphs.

3.2.2 Structure Mining on IS-contracted Webgraphs

As we can see in Table 6, we can find quite a large number of inter-domain i-stars of sizes greater than 2 even in G_S^8. We already know that such inter-domain i-stars in G_S^0 can be candidates of communities [12], thus we will examine URLs of those structures found in $G_S^1, \ldots G_S^8$. Tables 9, 10, 11 and 12 are some examples of the results. We can see that an i-star, as a candidate of a community, can still be found in G_S^7. In this sense, i-star is very probable for structure mining even in is-contracted webgraphs.

Table 9: An example of an inter-domain i-star of size 4 found in G_S^1: (a) a page corresponding to the center node, and (b) pages corresponding to the satellite nodes. They are related to a research group on security and cryptography.

(a)	http://theory.lcs.mit.edu/~rivest/crypto-security.html
(b)	http://theory.lcs.mit.edu/~rivest/homepage.html
	http://www.cc.gatech.edu/classes/cs8113e_96_winter/
	http://www.uni-mannheim.de/studorg/gahg/PGP/cryptologl.html

Table 10: An example of an inter-domain i-star of size 5 found in G_S^3: (a) a page corresponding to the center node, and (b) pages corresponding to the satellite nodes. They are related to finite element method.

(a)	http://www-math.cudenver.edu/~lfranca/links/fem_people.html
(b)	http://www-math.cudenver.edu/~jmandel/
	http://www.duke.edu/~laursen/
	http://www.math.chalmers.se/~mohammad/
	http://www-math.cudenver.edu/~lfranca/links/rules.html

Table 11: An example of an inter-domain i-star of size 7 found in G_S^5: (a) a page corresponding to the center node, and (b) pages corresponding to the satellite nodes. They are related to Johns Hopkins University.

(a)	http://www.jhu.edu/
(b)	http://jhworld.jhu.edu/
	http://www.jhu.edu/news/home03/jul03/rubin.html
	http://www.jhu.edu/~asl/
	http://www.jhu.edu/~gazette/
	http://www.sais-jhu.edu/library/
	http://www.jhsph.edu/

Table 12: An example of an inter-domain i-star of size 9 found in G_S^7: (a) a page corresponding to the center node, and (b) pages corresponding to the satellite nodes. They are related to a research group on pigs.

(a)	http://mark.asci.ncsu.edu/SwineReports/2002/Contents.htm
(b)	http://www.ces.ncsu.edu/alleghany/
	http://netvet.wustl.edu/pigs.htm
	http://mark.asci.ncsu.edu/staff/seepub.htm
	http://mark.asci.ncsu.edu/Publications/internet_solutions/internetsolutions.htm
	http://mark.asci.ncsu.edu/tsee/seepub.htm
	http://mark.asci.ncsu.edu/STAFF/seepub.htm
	http://mark.asci.ncsu.edu/Staff/see.htm
	http://mark.asci.ncsu.edu/Staff/belstra.htm

4 Conclusion

In this paper, by using i-cliques and i-stars, we investigated the structure of the Web in i-clique and i-star contracted webgraphs proposed in [11, 12], and showed a series of preliminary experimental results. As a result, we observed that both i-cliques and i-stars are still strong candidate substructures for web structure mining even in i-clique or i-star contracted webgraphs. For example, we found a community in a webgraph after contracting i-stars seven times. As another result, both i-clique size and i-star size distributions in i-clique contracted and i-star contracted webgraphs, respectively, show power-law.

This is somehow amazing, and we again emphasize that this kind of 'self-similarity' implies the true scale-freeness. We remark that all the experiments are still ongoing, and more detailed results will be exposed.

Acknowledgment

This research was partially supported by the Scientific Grant-in-Aid from Ministry of Education, Science, Sports and Culture of Japan (Grant #0019500016).

References

[1] R. Albert, H. Jeong and A.-L. Barabási. Diameter of the World-Wide Web. *Nature* **401**, 130–131, 1999.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* **286**, 509–512, 1999.

[3] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. L. Wiener. Graph structure in the web. *Computer Networks* **33**, 309–320, 2000.

[4] G. W. Flake, S. Lawrence and C. L. Giles. Efficient identification of web communities. *Proc. 6th ACM Int'l Conf. on Knowledge Discovery and Data Mining*, 150–160, 2000.

[5] H. Ito, K. Iwama and T. Osumi. Linear-time enumeration of isolated cliques. *Proc. 13th Annual European Symposium on Algorithms*, 119–130, 2005.

[6] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM* **46**, 604–632, 1997.

[7] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A. S. Tomkins. The Web as a graph: measurements, models, and methods. *Proc. 5th Int'l Computing and Combinatorics Conference*, 1–17, 1999.

[8] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks* **31**, 1481–1493, 1999.

[9] L. Laura, S. Leonardi, S. Millozzi, U. Meyer and J. F. Sibeyn. Algorithms and experiments for the Webgraph. *Proc. 11th Annual European Symposium on Algorithms*, 703–714, 2003.

[10] The Stanford WebBase Project, `http://www-diglib.stanford.edu/~testbed/doc2/WebBase/`.

[11] Y. Uno, Y. Ota and A. Uemichi. Web structure mining by isolated cliques. *IEICE Transactions on Information and Systems*, Vol. E90-D (12), pp. 1998–2006, 2007.

[12] Y. Uno, Y. Ota and A. Uemichi. Web structure mining by isolated stars. *Lecture Notes in Computer Science*, Vol. 4936, pp. 149–156, W. Aiello, A. Broder, J. Janssen and E. Milios (Eds.), Springer, 2008.