An Exploratory Study on the Relationship Between OSS Project Popularity and Network Characteristics

SI708 Network Theory and Application – Term Paper
Min-seok Pang
noticeme@umich.edu

1. Introduction

In this term project, I have analyzed the networks of open source software(OSS), which refers to software whose source code is available to anyone and that is developed by the collaboration of voluntarily participating developers over the Internet. Any developer can initiate new OSS project, and developers can freely join and contribute to OSS development.

I have constructed networks of OSS projects hosted in SourceForge.net (http://www.sourceforge.net), an OSS-project hosting Web site. I utilized SourceForge.net research data archives maintained by University of Norte Dame F/OSS research group. (http://www.nd.edu/~oss/Data/) From this database, I have extracted a range of information of OSS projects, project participants, and several other data and conducted several analyses to investigate how OSS projects evolve over time and what kind of community structures emerge in these networks.

Further in this study, I have investigated the relationship between OSS project popularity and various network characteristics such as centrality and prestige. The intuition behind this is that popular projects tend to locate in the center of the network. Using a variety of popularity measures of OSS projects, I will find whether this proposition is true or not.

2. Network Construction

The networks of OSS projects are constructed as follows. First, I have constructed bipartite networks which consist of OSS projects and developers and where OSS project and participating developers are connected by arcs. University of Norte Dame database maintains SourceForget archives from Jan-2003 to Mar-2007, so I was able to build 11 bipartite networks as shown in Appendix 1. (Data between Feb-2003 and Oct-2004 were not available in the archives.) In building networks, inactive OSS projects and developers were excluded. The number of OSS projects and developers in each month are shown in Appendix 1.

Then, with Pajek, I converted these networks into two kinds of one-mode networks, OSS project networks and developers networks. In the former, if an OSS developer is participating two projects at the same time, two projects are connected with an edge. Likewise in the latter, two OSS developers were connected if both

of them are participating in developing more than one project.

Figure 1 shows part of OSS project network in Feb-2007, which consists of 1,000 most popular projects among 117,920 ones. The size of nodes represents the number of downloads of projects for the past seven days.

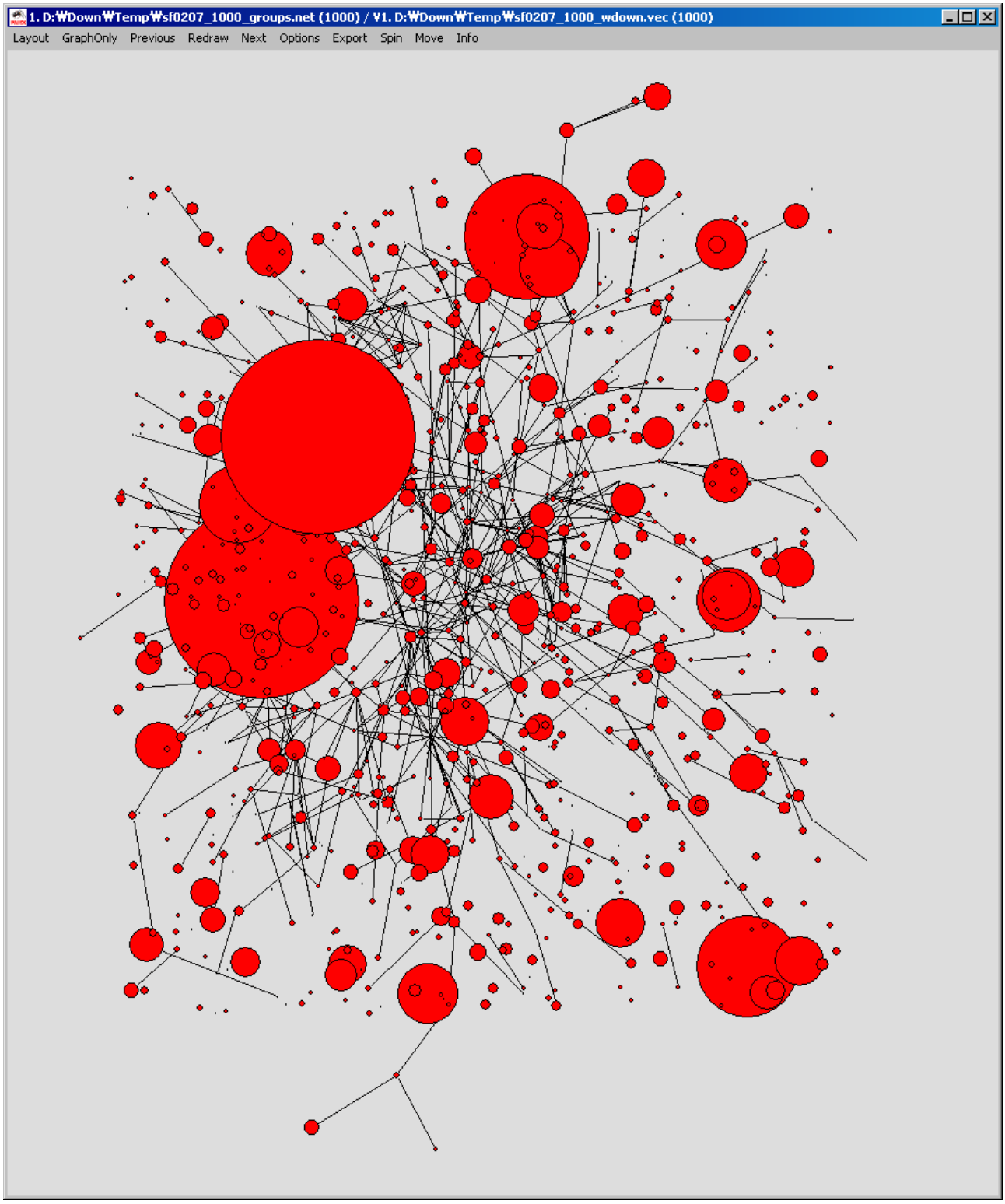

Figure 1. The Part of OSS Project Network in Feb-2007

Figure 2 shows the degree distribution of OSS project network in Feb-2007. To find whether it is distributed with a power law, a logarithm bin is used and the distribution is displayed in log-log slot. Figure 2 reveals that the degree has a distribution which is slightly different with a power distribution. Power low exponent calculated by the equation $\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}$ with $x_{\min} = 1$ is 2.117153. Figure 3 shows the degree distribution of OSS developer network in Feb-2007, showing that, in contrast to the project network, the degree of the developer network does not follow a power-law distribution.

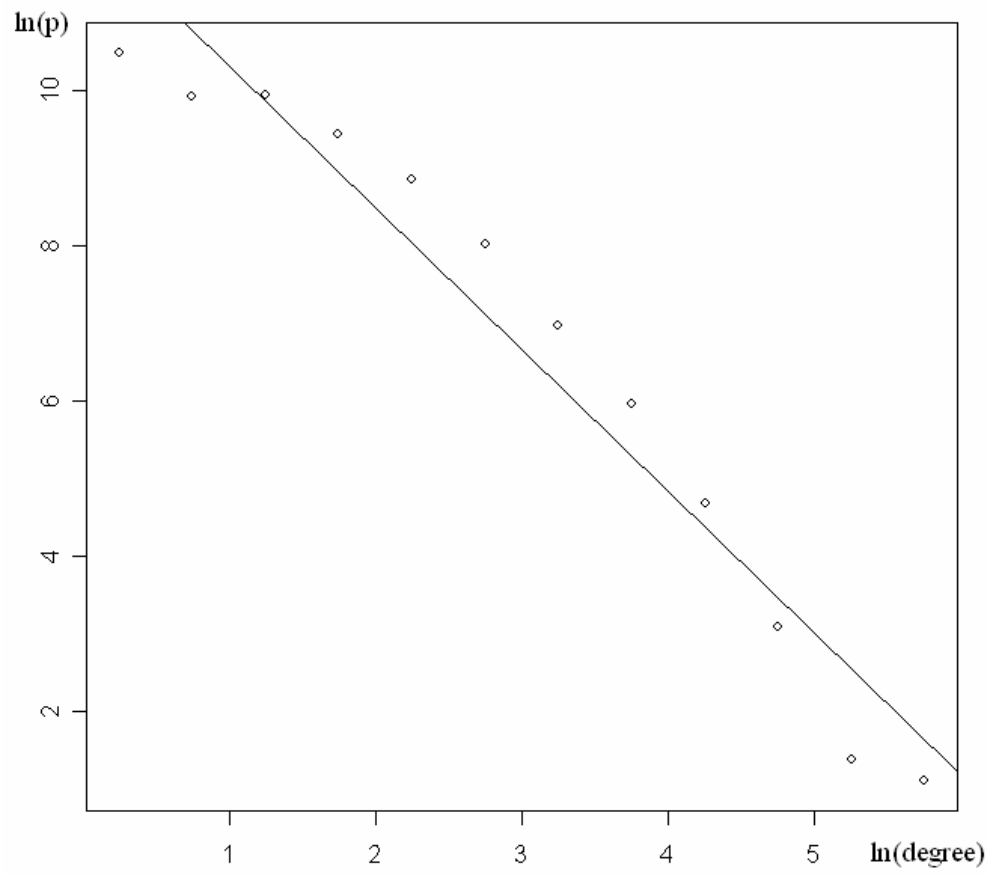

Figure 2. Degree Distribution of OSS Project Network in Feb-2007

3. The Evolution of OSS Networks

With longitudinal data from Jan-2003 to Feb-2007, I was able to examine how OSS project and developer networks evolve over time. Appendix 1 presents the basic characteristics of OSS project and developer networks.

The chart in Figure 4 shows that the size of OSS networks has increased monotonically except between Aug-2006 and Nov-2006. However, as shown in Appendix 1, the average number of developers per project and

that of projects that one developer joins do not change over time. The result shows that about two developers form a team for an OSS project and one developer works for about 1.4 projects in average.

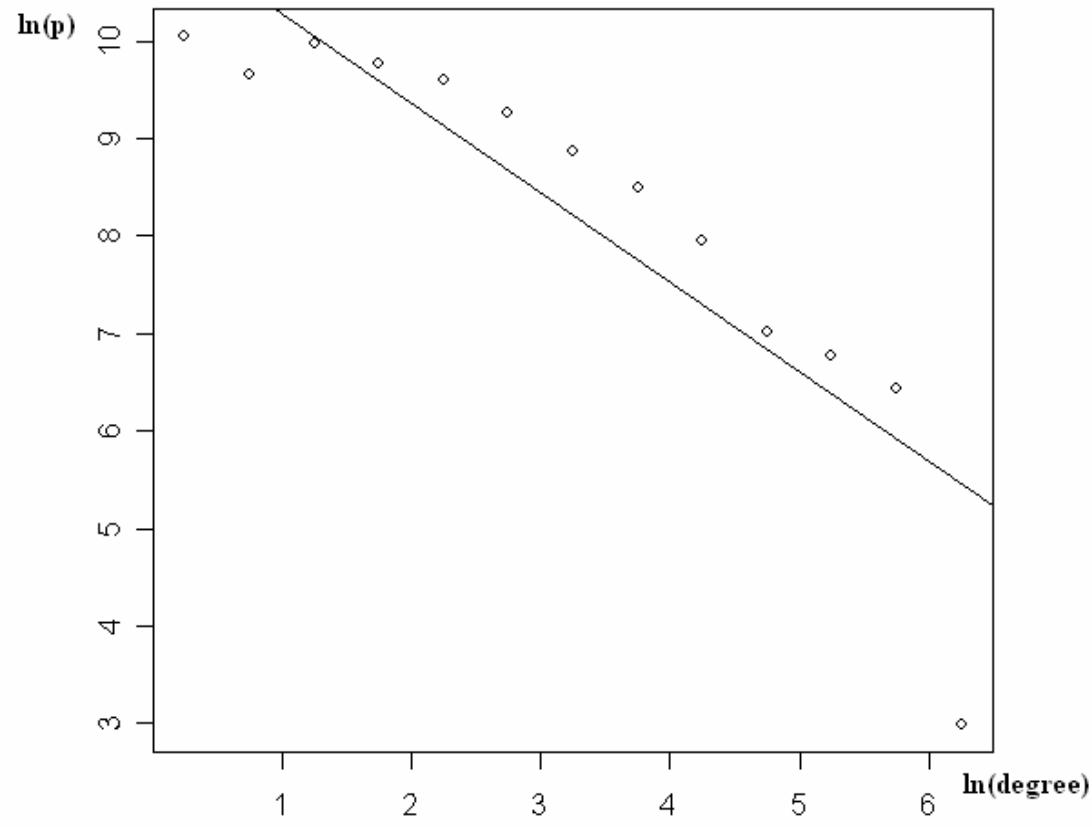

Figure 3. Degree Distribution of OSS Developer Network in Feb-2007

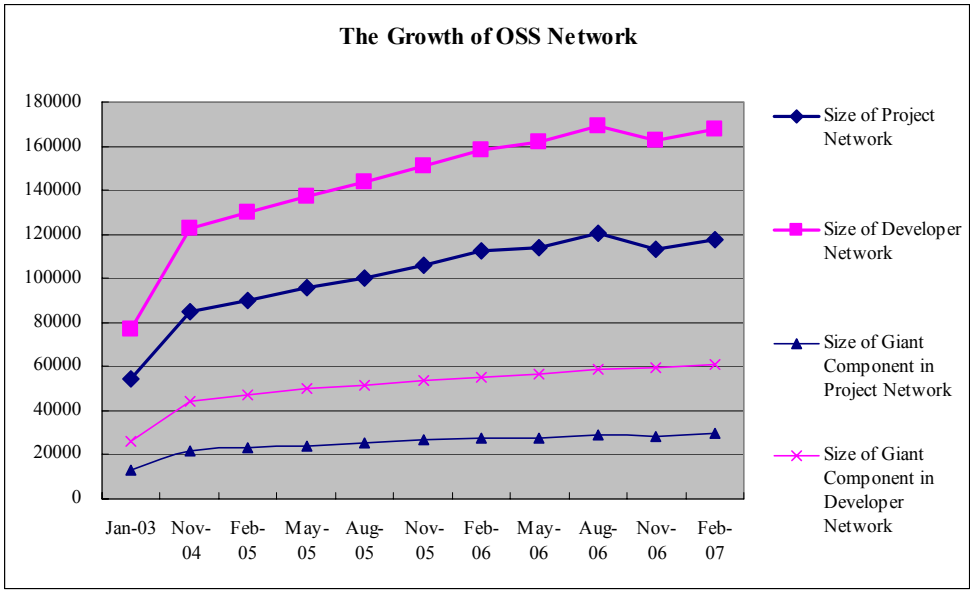

Figure 4. The Growth of OSS Network

The size of giant strong components has steadily increased as well, but its proportion in the whole network shown in Appendix 1 remains stable over the time with approximately 25% of project networks and 36% of

developer networks.

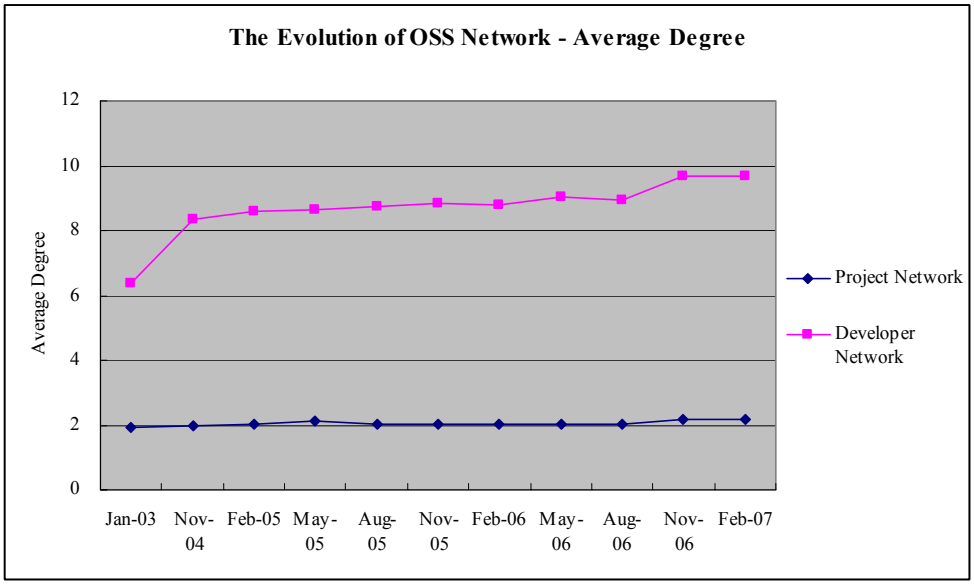

Figure 5. The Evolution of OSS Network in Average Degree

The chart in Figure 5 shows an interesting pattern. While the average degree of project networks almost remains stable with around 2.04, that of developer networks increases significantly over time from 6.39 in Jan-2003 to 9.7 in Feb-2007.

The chart in Figure 6 shows that clustering coefficients of OSS networks have increased slightly over time, even though the size of networks has also increased. Furthermore, Appendix 1 demonstrates that the diameter and average distance between reachable pairs have been stable over time. This result implies that, in OSS networks growing, newly added nodes tend to be connected with nodes which are located in the center, and existing nodes are more tightly connected with each other.

In addition, comparing OSS project networks with developer networks, I found that the developer networks have more nodes than the project networks and their density and clustering coefficient (6.19E-05 and 0.52 in average respectively) are greater than those of project networks (2.11E-05 and 0.28 in average respectively), but their diameter and average distance (22 and 7.13 in average respectively) are similar to those of project networks. This result demonstrates that OSS developers form denser networks than OSS projects.

Lastly, I have compared the project network in Feb-2007 with a random network with the same number of nodes (117920) and edges (127464). The analysis with Pajek found that the clustering coefficient and average distance of this random network are 5.1E-06 and 14.57, respectively. Its clustering coefficient is far less than that of the project network in Feb-2007 (0.28) and its average distance is larger than that of the project network (7.235). Base on this result, it can be concluded that the project network has a small-world property. (I could not conduct a similar analysis with the developer network, due to insufficient computing resource.)

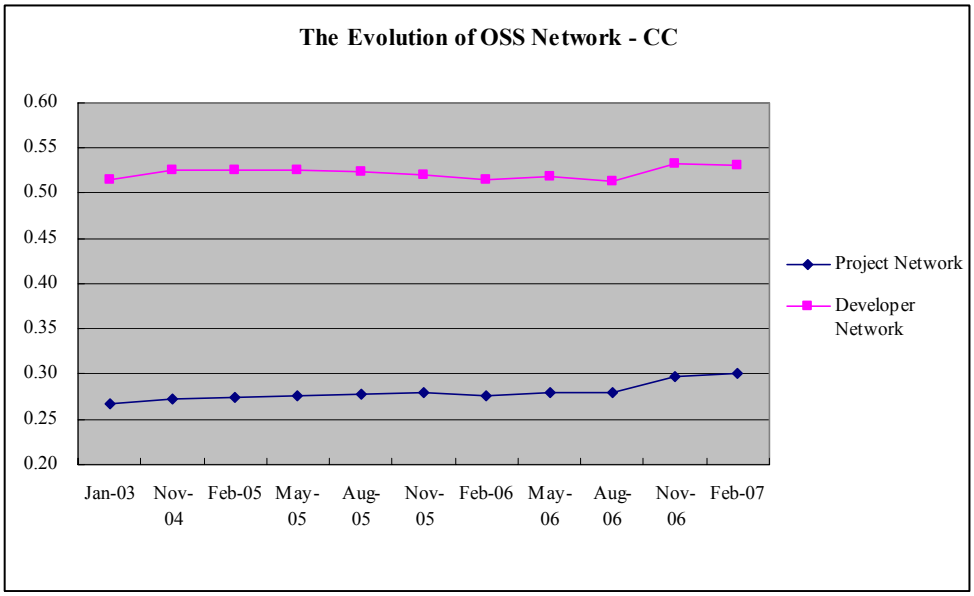

Figure 6. The Evolution of OSS Network in Clustering Coefficient

4. Community Structures of OSS Networks

In order to find community structures in OSS project and developer networks, I find *m*-slices with Pajek in the project network and developer network in Feb-2007. Figure 7 displays *m*-slices of the OSS project network in Feb-2007 whose *m* is greater than 10. This subnetwork consists of 147 nodes. From this figure, it can be found that there are three tight clusters of OSS projects, which consist of 5, 6 and 7 projects, respectively. Especially, Figure 8 exhibits that seven projects in 10-slices in the center are the closest community, and surprisingly the projects of No. 70 and No. 71 share 76 OSS developers and those of No. 21 and No. 71 have 63 common developers!

Figure 9 shows *m*-slices of the OSS developer network in Feb-2007 whose *m* is greater than 7. This subnetwork has 84 nodes. This figure reveals five communities of OSS developers with more than 4 developers. Figure 10 displays the largest cluster of OSS developers with 10 nodes and 33 edges!

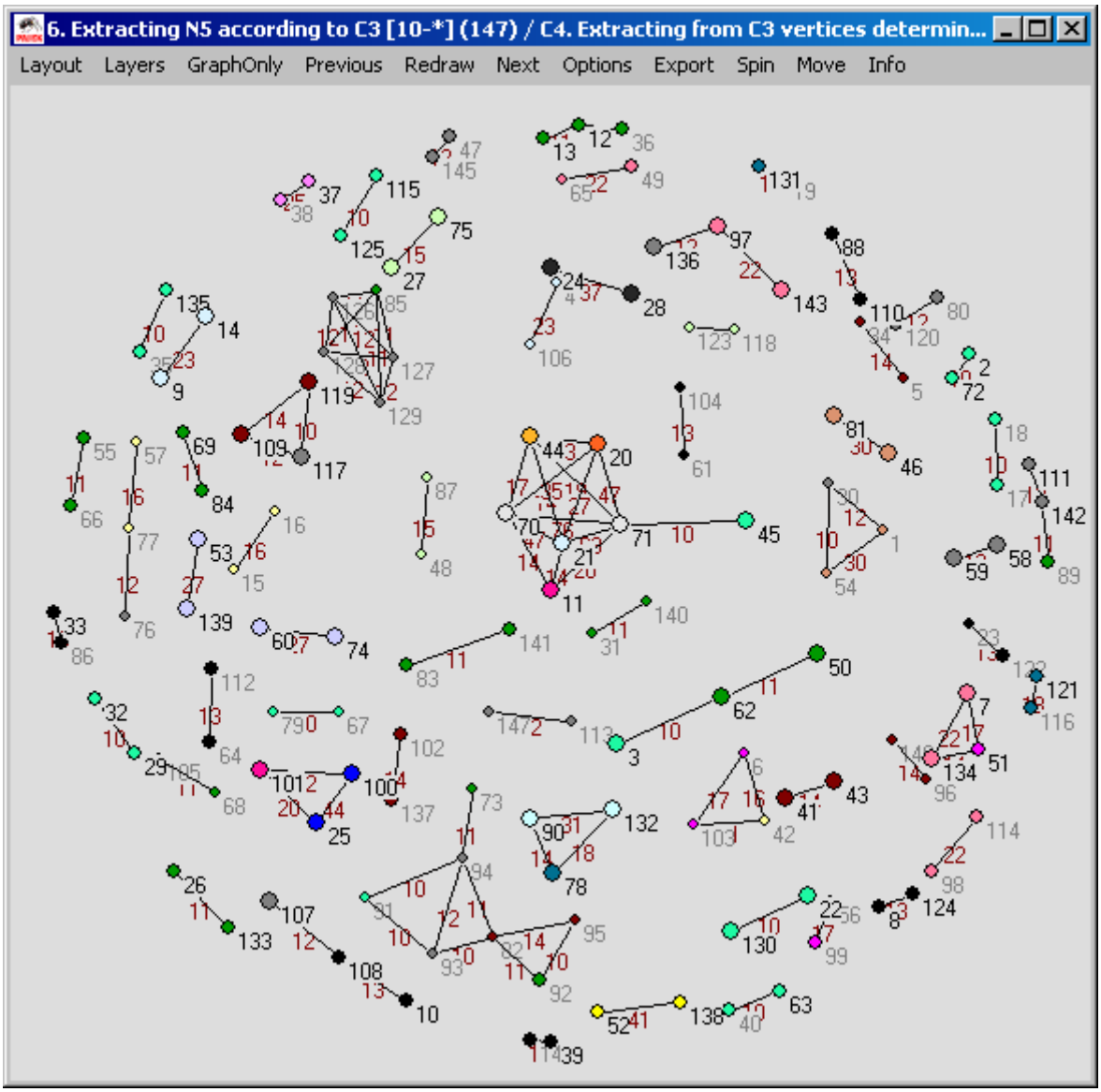

Figure 7. *m*-Slices in the OSS Project Network in Feb-2007

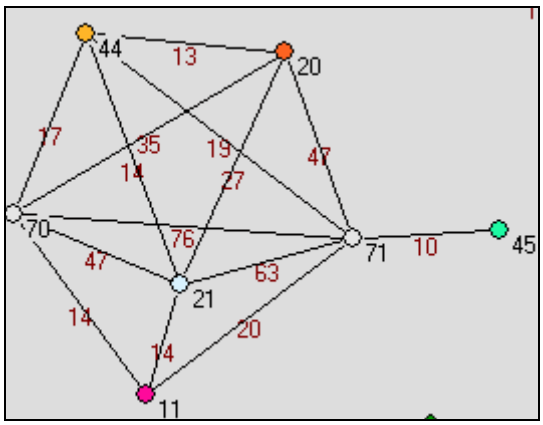

Figure 8. A 10-Slice in the OSS Project Network in Feb-2007

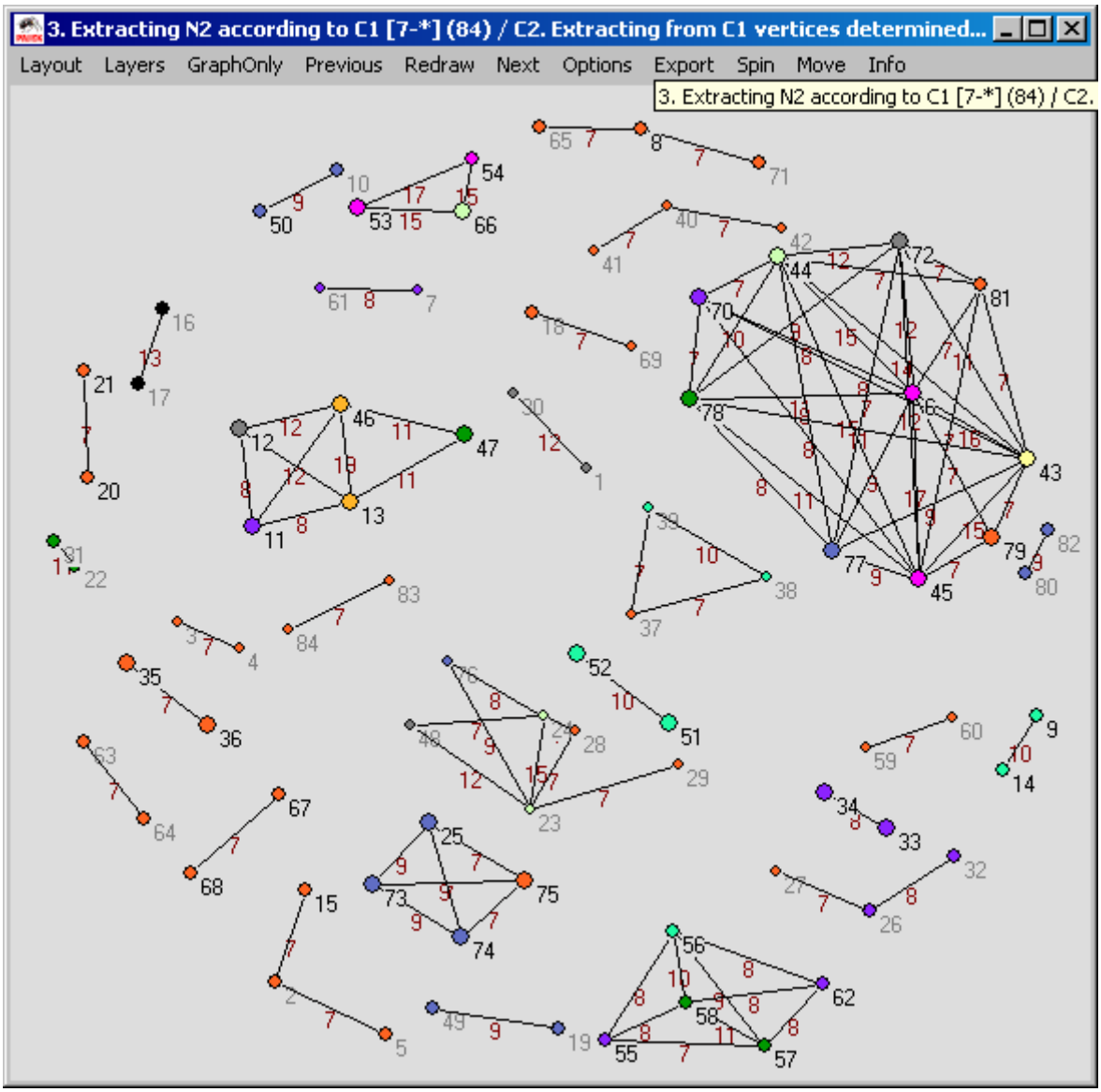

Figure 9. *m*-Slices in the OSS Developer Network in Feb-2007

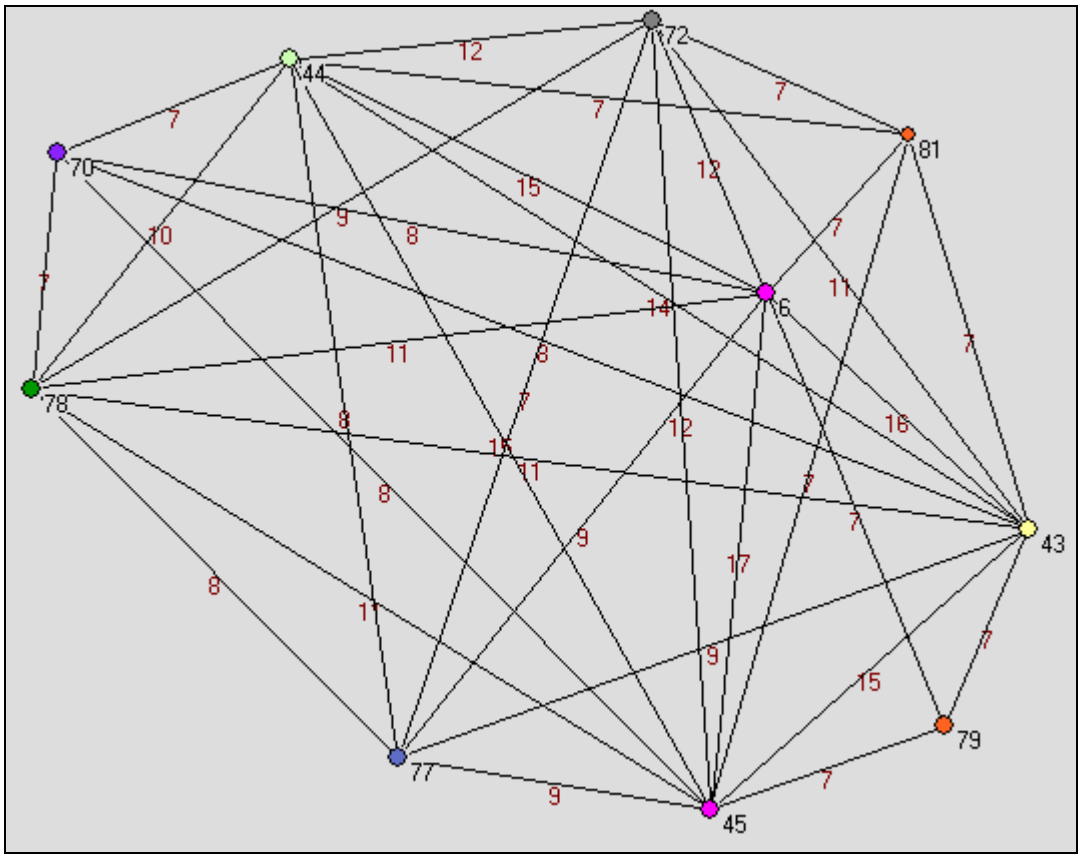

Figure 10. The Largest 7-Slice in OSS Developer Network in Feb-2007

5. Investigation on OSS Project Popularity

In order to examine the relationship between OSS project popularity and network characteristics, I have collected the following popularity measures from SourceForge.net archive as shown in Table 1.

Measures	Description
DALL	The number of total downloads from users since the inception of project
WDWN	The number of total downloads for the past 7 days
FCNT	The number of posts in project discussion forum for the past 7 days
PVWS	The number of pageviews in project homepage
MCNT	The number of mailing list subscribers

Table 1. OSS Project Popularity Measures

Additionally, the age of project (AGE, the number of days between register dates and Feb/28/2007), the number of participating developers (UCNT), and the number of commits (COMM) in CVS system, which reflects the extent to which the project is active, are collected.

For the network characteristics, I have calculated degree(DGRE), clustering coefficient(CC1), degree(DGRE), between(BETW), closeness(CLO), input domain prestige(INDO), and proximity

presige(PROX) of each node(project) using Pajek. Table 2 and Appendix 2 show descriptive statistics and the correlations of measures.

	Min	Median	Mean	Max	Standard Deviation
DALL	0	0	3762.00	84619648	250535.50
WDWN	9	0	36.26	917344	3313.61
FCNT	0	3	9.36	33520	218.08
PVWS	0	0	167.10	5630913	14608.18
MCNT	0	0	4.78	17274	91.35
AGE	8	1172	1201.00	2673	707.42
UCNT	0	1	1.73	351	2.93
COMM	0	0	51.19	140319	690.50
DGRE	0	1	2.14	362	4.65
CC1	0	0	0.31	1	0.44
BETW	0	0	0.000005565	0.01887	0.00008539
CLO	0	0.000019640	0.01264	0.08834	0.02381957
INDO	0	0.000009819	0.09001	0.40150	0.16742213
PROX	0	0.000009819	0.01264	0.08340	0.02382084

Table 2. Descriptive Statistics

In order to take a closer look, Table 3 shows the pairwise correlations between popularity measures and network characteristics with their significance (p-values). It shows that, except clustering coefficient, all of network characteristics have significant, positive correlations with popularity measures. For example, positive correlation between DGRE and DALL indicates that a project which users download frequently tends to show high degree centrality in OSS project network.

To further investigate the relationship, I have conducted OLS regressions between popularity measures and other variables. Popularity measures were regressed on network characteristic measures with control variables of age, the number of participating developers, and the number of commitments. Appendix 2 shows that there are high correlations between closeness, input domain, and proximity. Thus, to avoid instability in the regressions due to multicollinearity, input domain and proximity were not used in the regressions. Table 4 shows the results.

	DALL	WDWN	FCNT	PVWS	MCNT
DGRE	0.0407 (0.0000)***	0.0276 (0.0000)***	0.1038 (0.0000)***	0.0269 (0.0000)***	0.1544 (0.0000)***
CC1	0.0028 (0.2318)	0.0021 (0.3786)	0.0008 (0.7358)	0.0031 (0.1926)	-0.0005 (0.8310)
BETW	0.0514 (0.0000)***	0.0340 (0.0000)***	0.1013 (0.0000)***	0.0284 (0.0000)***	0.1740 (0.0000)***
CLO	0.0237 (0.0000)***	0.0182 (0.0000)***	0.0495 (0.0000)***	0.0161 (0.0000)***	0.0774 (0.0000)***
INDO	0.0201 (0.0000)***	0.0154 (0.0000)***	0.0423 (0.0000)***	0.0145 (0.0000)***	0.0651 (0.0000)***
PROX	0.0237 (0.0000)***	0.0182 (0.0000)***	0.0495 (0.0000)***	0.0161 (0.0000)***	0.0774 (0.0000)***

p-values are in parentheses. ***$p<0.001$

Table 3. Correlations between Popularity Measures and Network Characteristics

Table 4 provides several interesting results. First, all others being equal, old projects show high popularity except the weekly number of downloads. Second, projects which have a large number of participants and commitments are shown to be popular.

It is found that the network measures have varying relationships between popularity measures. Projects which have high degree centrality or high clustering coefficient tend to be popular in terms of the number of posts in discussion forums and the number of mailing list subscribers. On the other hand, those which have high betweenness and closeness centrality show high popularity as the number of downloads and the number of mailing list subscribers. It is also interesting to note that, while the number of pageviews have insignificant relationships with all of network measures, the number of mailing list subscribers have significant relationships with them.

Even though this result gives us some insights about OSS project network, it is unable to say that network characteristics explain the variances in popularity, since R^2 values are very low, even if the whole regression models are statistically significant. In addition, it needs further study to examine whether project popularity makes the project more central in the network, or a project which is located at the center becomes more popular.

Dependent Variables	DALL	WDWN	FCNT	PVWS	MCNT
(intercept)	-4228.0000 (0.0007)***	-46.1200 (0.0055)**	-10.9800 (0.0000)***	-263.6000 (0.0003)***	-5.6530 (0.0000)***
AGE	2.9640 (0.0004)***	0.0143 (0.2009)	0.0065 (0.0000)***	0.0863 (0.0798)*	0.0025 (0.0000)***
UCNT	1333.0000 (0.0000)***	29.0100 (0.0000)***	5.4500 (0.0000)***	150.4000 (0.0000)***	3.0290 (0.0000)***
COMM	16.8300 (0.0000)***	0.0846 (0.0000)***	0.0396 (0.0000)***	0.8088 (0.0000)***	0.0171 (0.0000)***
DGRE	-67.5100 (0.7382)	-2.5410 (0.3424)	1.7570 (0.0000)***	-4.4570 (0.7055)	0.7150 (0.0000)***
CC1	-1361.0000 (0.3918)	-9.2010 (0.6621)	-9.5210 (0.0000)***	27.2800 (0.7687)	-5.6320 (0.0000)***
BETW	83240000.0000 (0.0000)***	529000.0000 (0.0001)***	-4749.0000 (0.5800)	-118400.0000 (0.8396)	56370.0000 (0.0000)***
CLO	107900.0000 (0.0004)***	1263.0000 (0.0002)***	16.5200 (0.5260)	2144.0000 (0.2269)	71.9500 (0.0000)***
R^2	0.0053 (0.0000)***	0.0021 (0.0000)***	0.0344 (0.0000)***	0.0031 (0.0000)***	0.0603 (0.0000)***
N	182396	182396	182396	182396	182396

p-values are in parentheses. *p<0.1, **p<0.05, ***p<0.01

Table 4. OLS Regression Result

	Jan-03	Nov-04	Feb-05	May-05	Aug-05	Nov-05	Feb-06	May-06	Aug-06	Nov-06	Feb-07
No. of Active Projects	54234	84557	90022	95470	100514	106109	112359	114087	120368	113479	117920
No. of Active Developers	77050	122313	129934	137457	143822	151019	158301	161724	168852	162224	167802
Average # of Developers per Project	1.9700	2.0102	2.0092	2.0068	1.9988	1.9909	1.9693	1.9783	1.9600	2.0140	2.0072
Average # of Projects per Developer	1.3866	1.3897	1.3920	1.3938	1.3969	1.3989	1.3978	1.3956	1.3972	1.4088	1.4105
Project Network											
Density	0.0000355	0.0000236	0.0000223	0.0000221	0.0000201	0.0000191	0.0000179	0.0000177	0.0000169	0.0000190	0.0000183
Average Degree	1.9253	1.9955	2.0075	2.1099	2.0203	2.0267	2.0112	2.0193	2.0342	2.1561	2.1579
Size of Giant Component	13395	21583	22942	24254	25456	26643	27538	27784	28798	28523	29461
Proportion of Giant Component	0.2470	0.2552	0.2548	0.2540	0.2533	0.2511	0.2451	0.2435	0.2392	0.2514	0.2498
Clustering Coefficient	0.2678	0.2723	0.2743	0.2759	0.2778	0.2790	0.2769	0.2788	0.2799	0.2978	0.3004
Diameter	20	20	22	21	21	21	24	21	23	23	21
Mean Distance	7.3573	7.2706	7.2499	7.2657	7.2376	7.2189	7.2052	7.2080	7.2137	7.1945	7.1639
Developer Network											
Density	0.0000829	0.0000684	0.0000662	0.0000628	0.0000607	0.0000585	0.0000556	0.0000559	0.0000530	0.0000596	0.0000578
Average Degree	6.3874	8.3661	8.6016	8.6322	8.7299	8.8346	8.8015	9.0403	8.9491	9.6685	9.6989
Size of Giant Component	26398	44230	47126	49732	51819	54007	55343	56812	58643	59425	61111
Proportion of Giant Component	0.3426	0.3616	0.3627	0.3618	0.3603	0.3576	0.3496	0.3513	0.3473	0.3663	0.3642
Clustering Coefficient	0.5151	0.5258	0.5259	0.5257	0.5235	0.5210	0.5142	0.5183	0.5132	0.5330	0.5313
Diameter	20	21	22	21	22	21	23	22	24	24	22
Mean Distance	7.3029	7.1746	7.1588	7.1503	7.1366	7.1103	7.0913	7.0987	7.1033	7.0926	7.0466

Appendix 1. Several Characteristics of OSS Project and Developer Networks

	DLL	WDWN	FCNT	PVWS	MCNT	AGE	UCNT	COMM	DGRE	CC1	BETW	CLO	INDO	PROX
DALL	1													
WDWN	0.91237	1												
FCNT	0.16890	0.19811	1											
PVWS	0.80940	0.70662	0.07379	1										
MCNT	0.05850	0.03425	0.14737	0.05839	1									
AGE	0.01692	0.00835	0.04247	0.01067	0.04735	1								
UCNT	0.05055	0.04047	0.13356	0.04209	0.19702	0.07410	1							
COMM	0.06026	0.02962	0.15899	0.04791	0.18578	0.09951	0.31192	1						
DGRE	0.04075	0.02757	0.10384	0.02691	0.15441	0.13750	0.55479	0.24457	1					
CC1	0.00280	0.00206	0.00079	0.00305	-0.00050	0.10352	0.01807	0.00815	0.39391	1				
BETW	0.05144	0.03403	0.10131	0.02844	0.17399	0.05100	0.63690	0.26243	0.59191	-0.00731	1			
CLO	0.02370	0.01821	0.04948	0.01607	0.07740	0.17226	0.26820	0.11104	0.50696	0.42641	0.15442	1		
INDO	0.02011	0.01542	0.04230	0.01448	0.06513	0.16005	0.23761	0.09311	0.45914	0.43240	0.12125	0.98599	1	
PROX	0.02370	0.01821	0.04948	0.01607	0.07740	0.17225	0.26819	0.11104	0.50692	0.42633	0.15442	1.00000	0.98600	1

Appendix 2. Correlation Table