Caleb Carlson
Nicholas Fabrizio, iEng, MIET

CS-535 – Programming Assignment 1

1. Empirical Observation 1: Density of the graph

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
e(t)	170	2919	11519	30055	59236	98687	143301	201485	265655	334212	347414
n(t)	850	2862	5674	9047	12865	16985	21457	26211	31286	36345	37201

Table 1: Running total number of edges and nodes per year

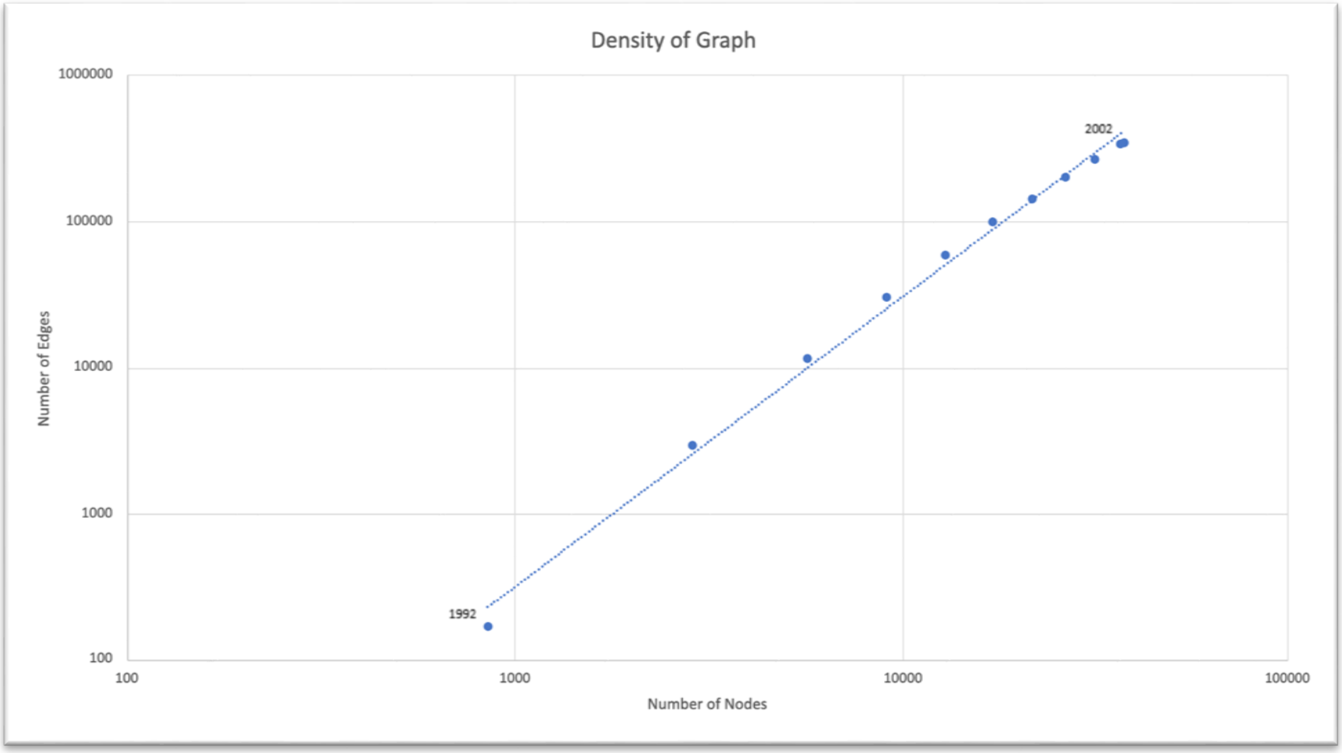

Figure 1: Log-Log plot of number of edges versus number of nodes by year

2. Empirical Observation 2: Diameter of the graph

Year	1992	1993	1994	1995	1996	1997
Effective diameter	7	11.38201	8.466838	6.905241	6.287455	6.905241

Table 2: Effective diameter by year for 90% of total nodes

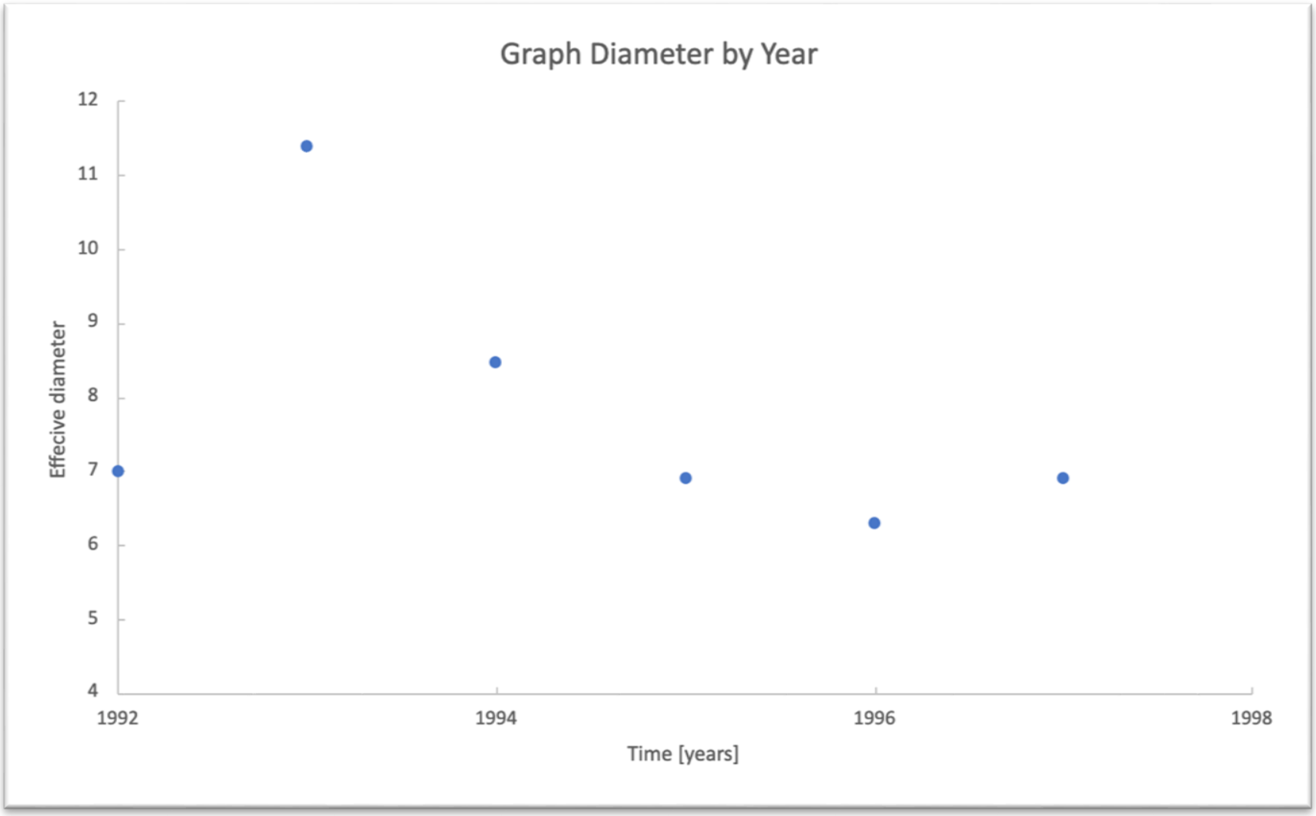

Figure 2: Effective diameter of the citation graph by year for 90% of total nodes

3. Analysis of the temporal citation network
   1) Analysis question 1

      In Figure 1, the log-log plot for the density of the arXiv citation network shows a general trend of a straight line with a positive slope, although the data points for 1992 as well as 2000 through 2002 are lower in terms of the number of edges causing an apparent divergence from the general trend line.

      Figure 1 does correspond to the exponent of the densification power law, which states that the slope of the log-log plot trend line should be between one and two [1]. For Figure 1, the values for years 1998 and 1999 are very close to being on the trend line. Using the values for these two data points and the formula for calculating the slope of a log-log plot line, the calculation yields 1.70. This value was calculated using the following formula.

$$\frac{\Delta \ln(y)}{\Delta \ln(x)}$$

      The divergence of the data point for 1992 can be explained by the phantom nodes and phantom edges concept described by Leskovec, Kleinberg and Faloutsos [1]. A manual check of the input data confirmed that there are quite a few nodes in the citations.txt file that have from or to nodes that are not represented in the published-dates.txt file. Phantom nodes and edges may also account for the slight variance from the trend line among other data points as well. The divergence from the trend line for the data points for the years 2000 through 2002 is obviously due to a drop in the rate of increase of both nodes and edges during that time frame. This could be caused by changes to the paper publishing process such as more stringent standards that resulted in fewer papers being published. It could also be related to the sharp recovery from the economic recession in those years leading to more people moving into private industry rather than working in academia.

2) Analysis question 2

   Observation of the plot in Figure 2 shows that the effective diameter of the arXiv citation graph increased dramatically from 1992 to 1993, and then decreased exponentially from 1993 through 1996. The observation also shows that there was a slight increase in effective diameter in 1997. Although it is difficult to state with certainty from the data in Figure 2, since the data point for 1997 deviates from the general trend of the plot from the preceding five years, it is possible that this data point is an outlier or that it is a result of phantom nodes and edges [1]. Analysis of the data from subsequent years is needed to determine the accuracy of this statement. Extrapolating this observation to the growth pattern of effective diameters in general would yield the statement that the effective diameters of temporal network graphs decrease over time.

   Although there are many possible explanations for the pattern displayed in Figure 2 for the citation network graph, one that is likely is that over time there will be papers published that have the attribute of being extremely insightful or controversial, reporting groundbreaking discoveries or being very popular of highly publicized. Published papers with these attributes will be highly likely to be cited by greater numbers of other papers. The greater the number of citations to one paper, the more papers it is directly connected to, and the more pairs of papers it links together. From a graph perspective, this paper is a node, and greater the number of edges it has, the greater the number of node pairs it links. Over time, these nodes become hubs with very dense connections, and their presence in the network graph decreases the effective diameter of the graph.

REFERENCES

[1] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05). Association for Computing Machinery, New York, NY, USA, 177–187. DOI:https://doi.org/10.1145/1081870.1081893