Programming Assignment 1 Report
CS 535 Big Data

Andrei Bachinin
Tanjim Bin Faruk

February 2023

1 Analysis question 1

1.1 Densification Power Law

The densification power law is a concept in the realm of temporal graph evolution. It states that networks become denser over time, with the average degree increasing (and hence with the number of edges growing super-linearly in the number of nodes). Moreover, the densification follows a power-law pattern.

In particular, the networks follow this relation:

$$e(t) \propto n(t)^a \tag{1}$$

where $e(t)$ and $n(t)$ denote the number of edges and nodes of the graph, at time t, and a is an exponent that generally lies strictly between 1 and 2. We refer to such a relationship as a *densification power law*, or *growth power law*. (Exponent $a = 1$ corresponds to a constant average degree over time, while $a = 2$ corresponds to an extremely dense graph where each node has, on average, edges to a constant fraction of all nodes.)[1][2]

1.2 Density of the graph

Up to the year 1999, the number of edges grew exponentially in the number of nodes, as evidenced by the value of a being greater than 1 (there were no new edges from after 1999). It demonstrates that the growth pattern of the graph follows the densification law.

Year	n(t)	e(t)	a
1992	850	152	0.745
1993	2826	2862	1.002
1994	5689	11379	1.082
1995	9069	29808	1.131
1996	12930	58927	1.160
1997	17116	98307	1.179
1998	21716	142934	1.189
1999	26631	201300	1.199

Table 1: Number of Nodes and Edges per Year with Exponent Values

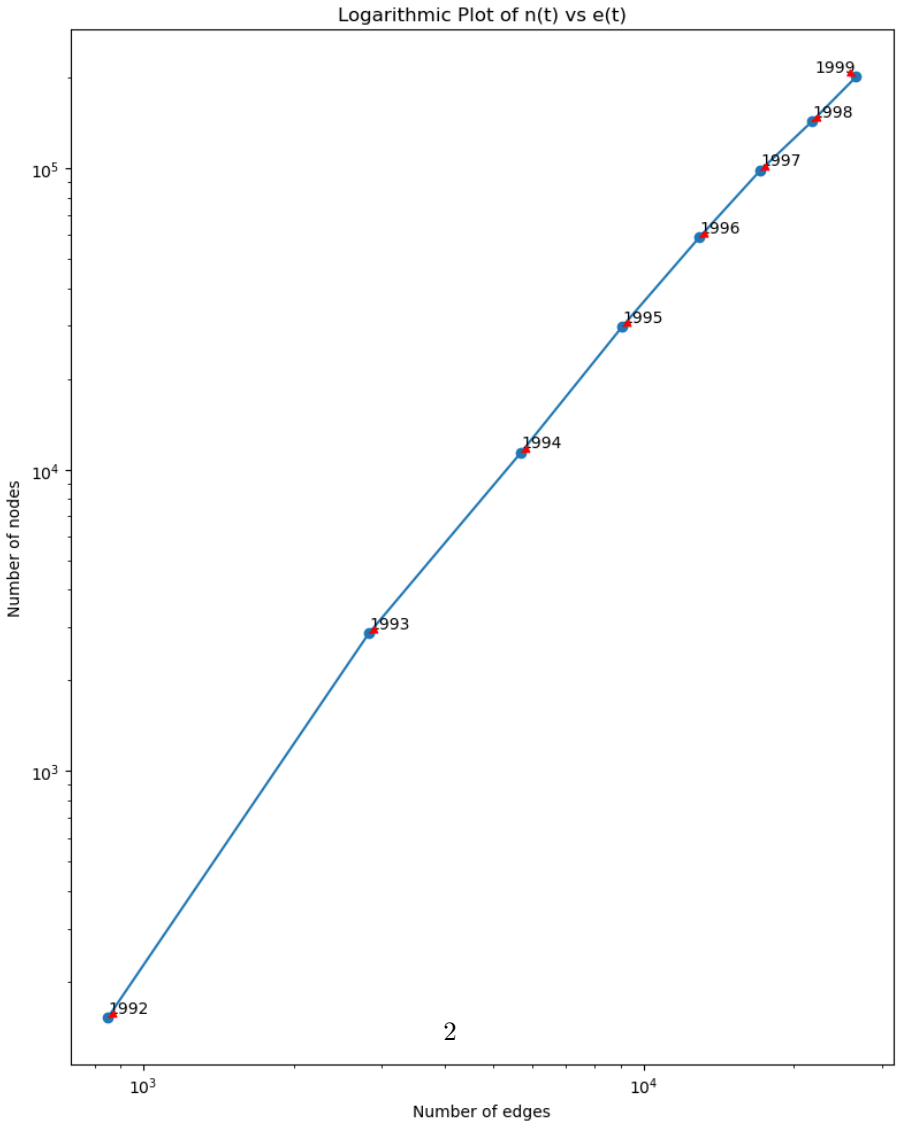

Figure 1: Log Scale Plot for Number of Edges versus Number of Nodes by Year

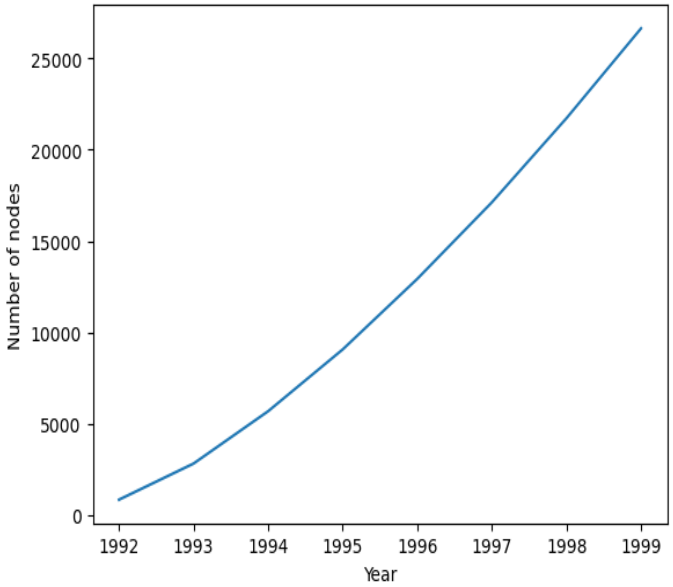

Figure 2: Number of nodes per year (1992 - 1999)

Figure 3 further demonstrates the exponential growth behavior of the edges. Compared to figure 2, the graph of edges exhibits exponential growth.

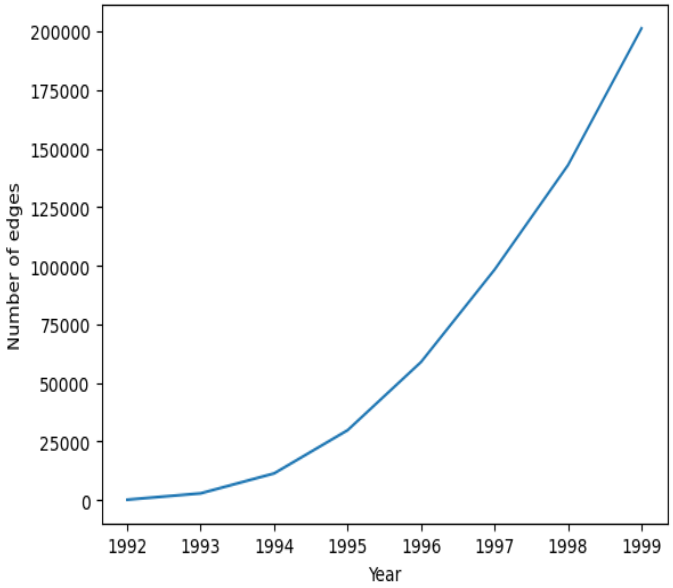

Figure 3: Number of edges per year (1992 - 1999)

2 Analysis question 2

2.1 Diameter of the graph

Generally, the more interconnected the graph is the higher the number of shortest paths exist between pairs of nodes. Theoretically, in a fully connected graph where every pair of nodes is directly connected by an edge, the number of shortest paths between any pair of nodes should be equal to one. If the degree of connectivity decreases, the number of shortest paths between pairs of nodes increases.

However, the relationship between the number of shortest paths and the pattern of connectivity is not that straightforward. For our given citation graph, papers and authors that are highly cited and influential are likely to have many incoming and outgoing edges. As a result, the shortest paths between these nodes are likely to be shorter, because they can be reached through a larger number of nodes. Additionally, authors from the same research area should tend to cite each other more frequently.

Furthermore, the age of papers in the citation network can also affect the number of shortest paths. As new papers are added to the network, they might cite older papers, leading to an increase in shortest paths between older and newer papers. As older papers continue to be cited over time, they may become more connected with other papers, causing the emergence of shorter paths.

Table 2 demonstrates that over the years every shortest path $g(n)$ is constantly increasing until the year 2000. Starting in the year 2000 we get the same values for the shortest paths because there are no new edges.

Usually, the higher n becomes, the bigger number of shortest paths emerge. We can observe that $g(1)$ increases slightly compared to other paths over the years: from 152 to 201074. At the same time, $g(4)$ increases dramatically from 610 to 115 million. It indicates that a large pair of nodes can reach each other within a small distance (< 4). We believe that such behavior fits the given *arXiv* citation network.

Year	# of Shortest Paths in g(1)	# of Shortest Paths in g(2)	# of Shortest Paths in g(3)	# of Shortest Paths in g(4)
1992	152	369	506	610
1993	2856	16059	45741	105112
1994	11368	96045	393339	1184069
1995	29791	341283	1833189	6068646
1996	58895	850673	5426701	18191226
1997	98256	1681592	11987796	40508832
1998	142834	2729844	20808492	70971611
1999	201074	4195366	33705585	115276212
2000	201074	4195366	33705585	115276212
2001	201074	4195366	33705585	115276212
2002	201074	4195366	33705585	115276212

Table 2: Number of shortest paths per year

References

[1] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 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, page 177–187, New York, NY, USA, 2005. Association for Computing Machinery.

[2] Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, USA, 2006.