

Coauthorship Networks: An Application of Graph Theory

Samarth Gupta and Ridham Bhagat

B.Tech. CSE IInd Year, Shiv Nadar University

December 14, 2020

Abstract

Co-authorship or research collaboration is an important way to link scientists from multidisciplinary backgrounds to produce a research output and to solve problems. Modelling the co-authorship networks helps to understand the emergence and propagation of thoughts in academic society. Co-authorship could be modelled as collaboration social networks. An in-depth analysis of these networks provides an opportunity to investigate its structure. Patterns of these relationships could reveal the mechanism that shapes our scientific community. Thus, an academic or research institution can benefit from building co-authorship networks as these can indicate the possibility of collaboration in an institution and may help in improving the research output. The purpose of this paper is to present the idea of co-authorship networks as an application of graph theory. We have also calculated Erdos numbers of some of the esteemed faculty of our department using the bibliography information available on Google Scholar.

1 Introduction

Graph theory is a branch of mathematics that concerns itself with networks of points connected by lines. Graph theory has applications

in chemistry, operations research, social sciences, and computer science Carlson [2020]. A graph or a network is a set of vertices, also called as points or nodes and of edges, also called as lines or arcs that connect the vertices Carlson [2020]. When appropriate, a direction may be assigned to each edge to produce what is known as a *directed graph*, or digraph. On the other hand, the edges in an *undirected graph* are undirected. One of the popular use of graphs is to model social structures representing relationships between people or groups. Such graphs or networks are known as social networks. In a social network, vertices represent individuals, organizations, or groups and edges represent relationships between them. For example, Facebook is a popular social network where each person represents a node and every interaction between these nodes like friend request, like, share, comment, tag represent edges. Similarly, in Twitter, persons are considered as nodes and if one person follows another then that is considered as the edge between the two. Facebook and Twitter are examples of a type of graph model of social network know as *friendship graphs*. Other popular graph models of social networks include collaboration graphs and influence graphs. *Collaboration graphs* are undirected graphs where two people are connected if they collaborate in a specific way. *Influence graphs* are directed graphs where there is an edge from one person to another if the first person can influence the second person. Examples of collaboration graphs include the Hollywood graph, co-authorship networks. In Hollywood graph models, vertices in the graph represent actors and an edge connects two vertices if the actors they represent have appeared in the same movie. Co-authorship graphs model the academic collaboration of researchers who have jointly authored a publication. Here, the vertices represent researchers and an edge connects the vertices representing two researchers who have co-authored a publication. All these example networks exist in the real world and are therefore termed as *real-world social networks*.

In this paper, we discuss the co-authorship networks as an application of graph theory. Section 2 gives the mathematical formulation of social networks as graphs. Section 3 discusses the concept of collaboration and co-authorship networks. Sections 4 and 5 briefly discuss the small world model and Erdos number in the context of co-authorship network.

2 Social Networks as Graphs

When we think of a social network, we think of a website that is allows for social interactions. Some of the popular social networks include Facebook, Twitter, and Google+ Leskovec et al. [2014]. A social network is a set of people or groups of people, with some pattern of interactions or relationships between them. Friendship networks among a group of individuals, business relationships between companies, and marriages between families are all examples of social networks Carlson [2020]. Other examples include information networks (documents, web graphs, patents), infrastructure networks (roads, planes, water pipes, powergrids), and biological networks (genes, proteins, food-webs of animals eating each other) Leskovec et al. [2014].

2.1 What is a Graph?

A graph represents a network which consists of a set of objects, mathematically called vertices or nodes. These vertices or nodes are interconnected with each other based upon some relation, with the help of edges or arcs. A graph consists of $(V(G), E(G))$, where $V(G)$ is a non-empty set called the vertices, and $E(G)$ is a set called the edges.

Graph can be classified as directed or undirected. In directed graphs the edges between two vertices have a particular direction; they are directed from one vertex to another. It is usually represented by an arrow. In undirected graphs the edges do not have any particular direction from one vertex to another; there is no difference between the two vertices connected via one undirected edge. It is usually represented by a straight line Chakraborty et al. [2018].

When any two vertices are joined by more than one edge, the graph is called a *multigraph*. A graph without loops and with at most one edge between any two vertices is called a *simple graph*. When each vertex is connected by an edge to every other vertex, the graph is called a *complete graph*. Two vertices that are connected by an edge between them are called as *neighbours* or *adjacent vertices*. *Degree* of a vertex is defined as the number of its neighbors, that is, the number of edges that enter or exit from it. A loop contributes two to the degree of its vertex Carlson [2020]. A *connected graph* is a graph where there is a path between every pair of vertices. A *path* in a graph is a finite

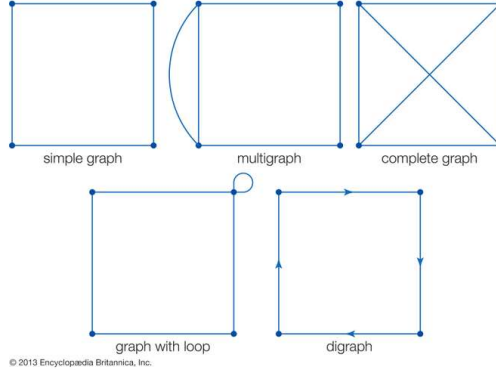


Figure 1: Figure taken from Carlson [2020]

or infinite sequence of edges that joins a sequence of vertices. Figure 1 shows some examples.

2.2 Real-world Graphs

The graphs occurring in the real-world usually exhibit the sparsity, small world phenomenon, network clustering, and power law degree distribution Newman et al. [2002], Yegnanarayanan and Umamaheswari [2011]. *Sparsity* means that the number of edges is within a constant multiple of the number of vertices. [Travers and Milgram, 1977] talked about the *small-world problem*, that is, the question of how two people can have a short connecting path of acquaintances in a network. The authors deduced that many pairs of apparently distant people are actually connected by a very short chain of intermediate acquaintances and named this phenomenon as the small-world effect Travers and Milgram [1977]. *Clustering* is the task of assigning a set of objects to groups, also called as classes, clusters, or categories so that the objects in the same cluster are more similar (according to a predefined property) to each other than to those in other groups Newman et al. [2002]. The degree to which the vertices in a graph cluster together is measured using its clustering coefficient. The *clustering coefficient* of a graph is a fraction of ordered triples of vertices (a, b, c) such that an edge between vertices b and c is present if the edges between vertices a and b , and between b and c are present. That is, the clustering coefficient measures the likelihood with which two neighbours of a vertex are themselves neighbours Yegnanarayanan and

Umamaheswari [2011]. A graph exhibits *power law distribution* if the number of vertices of degree d is inversely proportional to a power of d . That is, the number of vertices with degree d is proportional to $d^{-\beta}$ for some fixed constant β Yegnanarayanan and Umamaheswari [2011].

2.3 Mathematical Representation of Social Networks

Social networks are naturally modelled as graphs, and are also referred to as social graphs. The entities are the nodes, and an edge connects two nodes if the nodes are related by the relationship that characterizes the network. Social networks may be undirected or directed. For example, Facebook friends network is an undirected social network while the graphs of followers on Twitter or Google+ are directed social networks Leskovec et al. [2014]. A social network S can be modelled as a graph $G = (V, E)$, where $V = 1, \dots, n$ is the set of vertices or nodes representing the users in S and $E = (E_{ij})$, $i, j \in V$ is the set of edges that connect elements of V . The graph G may be directed or undirected, depending on the nature of the social network. $n = |V|$ denotes the number of nodes in G and $m = |E|$ denotes the number of edges in G . A graph G may be represented as an $n \times n$ adjacency matrix $A = (A_{ij})$, for $i, j \in V$, where,

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

and the number of edges (m) may be expressed as follows,

$$m = \begin{cases} \frac{1}{2} \sum_{i,j} A_{ij} & \text{for undirected graphs,} \\ \sum_{i,j} A_{ij} & \text{for directed graphs.} \end{cases} \quad (2)$$

3 Collaboration Networks

Scientific collaboration is a requirement in contemporary academic research with researchers with complementary skills and multidisciplinary approaches working together towards common goals. Researchers come together as combining different types of knowledge and

skills allows them to address complex problems. Research collaboration is a key mechanism that links distributed knowledge and competencies into novel ideas and research avenues (Heinze and Kuhlmann, 2008). A *collaboration graph or network* is a graph modelling some social network where the vertices represent participants of that network and where two distinct participants are joined by an edge whenever there is a collaborative relationship of some kind between them. Collaboration graphs are used to measure the closeness of collaborative relationships between the participants of that network Yegnanarayanan and Umamaheswari [2011].

3.1 Interesting Features of Collaboration Graphs

A collaboration graph is a simple graph and need not be a connected graph. The distance between two vertices in a collaboration graph is called *collaboration distance*. The collaboration distance between two distinct vertices is equal to the minimum number of edges on a path between them. If no such path exists, then the collaboration distance is defined to be infinite. In collaboration graph of mathematicians, the collaboration distance from a particular person to Paul Erdos, is called the Erdos Number of that person Yegnanarayanan and Umamaheswari [2011].

In a collaboration graph, generally, the number of edges is small, that is, a little more than the number of vertices. Collaboration graph of mathematicians were shown to have “small world topology”. That is, they have a large number of vertices, most with a small degree, that are highly clustered, and a “giant” connected component with small average distances between vertices. The degrees of the vertices in a collaboration graph follows a power law distribution. For collaboration graphs, the *clustering coefficient* is very high Yegnanarayanan and Umamaheswari [2011].

3.2 Co-authorship Networks

Co-authorship networks are collaboration networks where nodes represent authors who have published research papers, organizations or countries, which are connected when they share the authorship of a paper e Fonseca et al. [2016]. Co-authorship networks are constructed from databases of papers. The famous Erdos number, named for Paul Erdos, is the shortest distance between the authors and Erdos in a

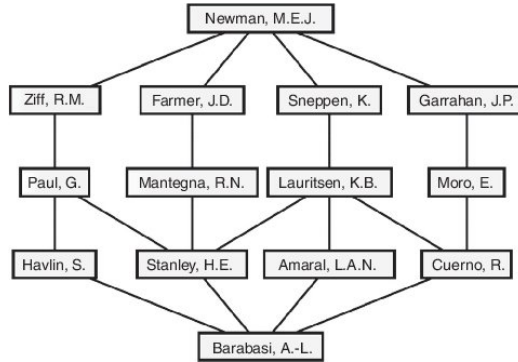


Figure 2: Figure depicting co-authorship network and the small world Phenomena taken from Kumar [2015]

co-authorship network Zhou et al. [2017].

Figure 2 shows an example co-authorship network.

4 Small World Model

The idea of social networks was made popular by the famous “Small world” experiment of Stanley Milgram in the 1960s Travers and Milgram [1977] Kumar [2015]. The experiment conducted by Stanley Milgram, asked test subjects, chosen at random from a Nebraska telephone directory, to get a letter to be reached to a target subject in Boston, a stockbroker friend of Milgram. The instructions were that the letters were to be sent to their addressee (the stockbroker) by passing them from person to person, but, that they could be passed only to someone whom the passer knew on a first-name basis. Because it was not likely that the initial recipients of the letters were on a first-name basis with a Boston stockbroker, their best strategy was to pass their letter to someone whom they felt was nearer to the stockbroker in some sense, either socially or geographically, perhaps someone they knew in the financial industry, or a friend in Massachusetts. A moderate number of these letters did eventually reach their destination, and Milgram discovered that the average number of steps taken to get them there was only about six, a result that has since passed into folklore and was immortalized later by John Guare in the title of his 1990 play, *Six Degrees of Separation* Yegnanarayanan and Umamaheswari [2011].

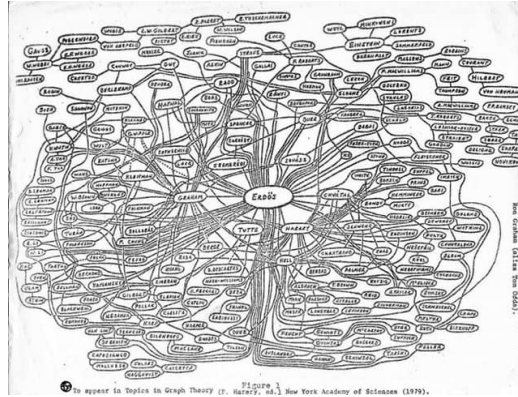


Figure 3: Figure taken from <https://www.mathscareers.org.uk/erdos-numbers/>

5 Erdos Number

One of the most prolific mathematicians of all time, Paul Erdos has written over 1400 papers with over 500 co-authors. Paul Erdos has epitomized the strength and breadth of mathematical collaboration. He travelled around the world for decades in search of new mathematics and new collaborators. His efforts are legendary in mathematical circles. This unparalleled productivity inspired the concept of the Erdos number. The Erdos number for a mathematician is an indicator of the quality and depth of that person's mathematics relative to the areas of mathematics that were of interest to Erdos himself. Erdos himself has Erdos number 0. People who had direct collaboration with Erdos himself have the Erdos number 1. Any person not assigned an Erdos number and who has written a joint mathematical paper with a person having Erdos number n earns the Erdos number $n + 1$. The tightly interconnected nature of the scientific community is visible as most publishing mathematicians, as well as many physicists and economists have a rather small Erdos number. Mathematicians, on noticing the importance of the network, modelled it with a collaborative graph, which two people are joined by edge as they collaborate. According to the graphs, among working mathematicians, Erdos number ranges up to 15 with a median of 5 Yegnanarayanan

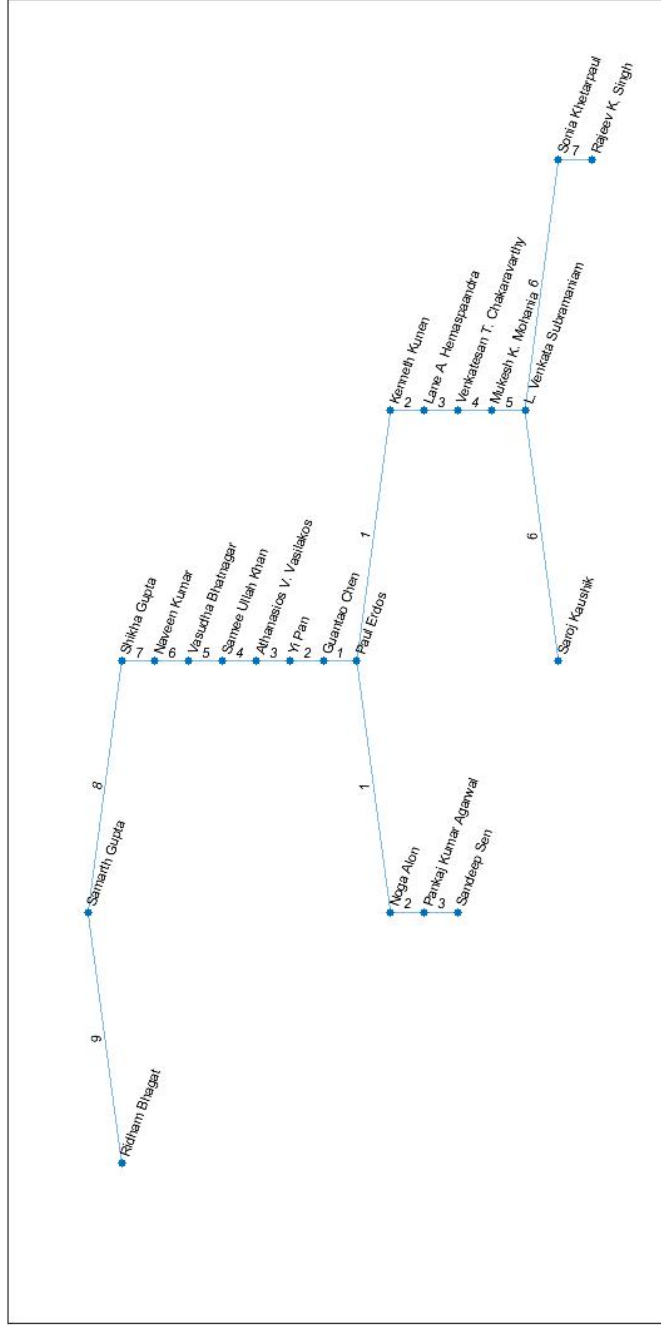


Figure 4: Erdos numbers of some people in our department

and Umamaheswari [2011], Barabási et al. [2002]. Figure 3 shows a small portion of the world’s Erdo’s network and Figure 4 shows Erdos numbers of some people in our department.

6 Conclusion

Co-authorship, publication by two or more researchers, is a form of collaboration between the researchers. One can say that an important result of collaborative research is use of the complementary skills to create new ideas and higher quality research that may not be achieved individually. Therefore, examining the co-authorships of research papers is a way to examine the research collaborations as some amount of acquaintance is expected amongst co-authors. Studies on co-authorship gained new interest after social network analysis began to be used to investigate the patterns in researcher collaboration networks. Social network analysis for co-authorship networks allows observation of their social structures to assess the status of an author in a particular field, and thus enhance the relations to get closer to the community core by identifying the most influential researchers. One of the most famous co-authorship networks is of the mathematician Paul Erdos, who has authored more than 1400 publications with more than 500 co-authors. The role of Erdos as a collaborator was so significant in the field of mathematics that the Erdos number is set to measure the proximity to Erdos through network co-authorship. A co-author of Paul Erdos has an Erdos number of 1. An author who collaborated with an author with Erdos number 1 has an Erdos number 2, and so on. A lower Erdos number indicates higher influence in the Mathematics research community. In this paper we have presented the idea of co-authorship networks as an application of graph theory. We have also calculated Erdos numbers of some of the esteemed faculty of our department (Figure 4) using the bibliography information available on Google Scholar. It is interesting to see that the Figure 4 exhibits “small world phenomenon”.

References

A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations.

- Physica A: Statistical mechanics and its applications*, 311(3-4):590–614, 2002.
- S. C. Carlson. Graph theory, 2020. URL <https://www.britannica.com/topic/graph-theory>.
- A. Chakraborty, T. Dutta, S. Mondal, and A. Nath. Application of graph theory in social media. *International Journal of Computer Sciences and Engineering*, 6:722–729, 2018.
- B. d. P. F. e Fonseca, R. B. Sampaio, M. V. de Araújo Fonseca, and F. Zicker. Co-authorship network analysis in health research: method and potential use. *Health Research Policy and Systems*, 14(1):34, 2016.
- S. Kumar. Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*, 2015.
- J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining Social-Network Graphs*, page 325–383. Cambridge University Press, 2 edition, 2014. doi: 10.1017/CBO9781139924801.011.
- M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002. ISSN 0027-8424. doi: 10.1073/pnas.012582999. URL https://www.pnas.org/content/99/suppl_1/2566.
- J. Travers and S. Milgram. An experimental study of the small world problem. In *Social Networks*, pages 179–197. Elsevier, 1977.
- V. Yegnanarayanan and G. Umamaheswari. A note on the importance of collaboration graphs. *Int. J. of Mathematical Sciences and Applications*, 1(3):1113–1121, 2011.
- G. Zhou, J. Li, and Z. Xie. A model of coauthorship networks. In *AIP Conference Proceedings*, volume 1890, page 040057. AIP Publishing LLC, 2017.