Int. J. of Mathematical Sciences and Applications,
Vol. 1, No. 3, September 2011

www.journalshub.com

A Note on the Importance of Collaboration Graphs

V.Yegnanarayanan[1] and G.K.Umamaheswari[2]

[1]Senior Professor, Department of Mathematics,
Velammal Engineering College,Ambattur-Red Hills Road,
Chennai - 600 066, India. Email id:prof.yegna@gmail.com
[2]Research Scholar, Research and Development Centre,
Bharathiar University, Coimbatore-641046, India.

Abstract

Numerous challenging problems in graph theory has attracted the attention and imagination of researchers from physics, computer science, engineering, biology, social science and mathematics. If we put all these different branches one into basket, what evolves is a new science called "Network Science". It calls for a solid scientific foundation and vigorous analysis. Graph theory in general and the collaboration graphs, in particular are well suited for this task. In this paper, we give a overview of the importance of collaboration graphs with its interesting background. Also we study one particular type of collaboration graph and list a number of open problems.

Keywords :collaboration graph, network science, erdos number

AMS subject Classification: 05XX, 68R10

1. Introduction

In the past decade, graph theory has gone through a remarkable shift and a profound transformation. The change is in large part due to the humongous amount of information that we are confronted with. A main way to sort through massive data sets is to build and examine the network formed by interrelations. For example, Google's successful web search algorithms are based on the www graph, which contains all web pages as vertices and hyperlinks as edges. There are all sorts of information networks, such as biological networks built from biological databases and social networks formed by email, phone calls, instant messaging, etc., as well as various types of physical networks.

Graph theory has two hundred years of history studying the basic mathematical structures called graphs. A graph G consists of a collection V of vertices and a collection E of edges that connect pairs of vertices. In the past, graph theory has been used in a wide range of areas. However, never before have we confronted graphs of not only such tremendous sizes but also extraordinary richness and complexity, both at a theoretical and a practical level. Numerous challenging problems have attracted the attention and imagination of researchers from physics, computer science, engineering, biology, social science, and mathematics.

1.1 Real world Graphs

The new area called "network science" emerged, calling for a sound scientific foundation and rigorous analysis for which graph theory is ideally suited. In the other direction, examples of real-world graphs lead to central questions and new directions for research in graph theory. These real-world networks are massive and complex but exhibits amazing coherence. Empirically, most real-world graphs have the following properties:

- **Sparsity** - The number of edges is with in a constant multiple of the number of vertices.
- **Small world phenomenon** - Any two vertices are connected by a short path. Two vertices having a common neighbor are more likely to be neighbors.
- **Power law degree distribution** - The degree of a vertex is the number of its neighbors. The number of vertices with degree j (or having j neighbors) is proportional to $j^{-\beta}$ for some fixed constant β.

1.2 Fundamental questions

To deal with these information networks, many basic questions arise:

- What are basic structures of such large networks?

- How do they evolve?
- What are the underlying principles that dictate their behavior?
- How are sub graphs related to the large (and often incomplete) host graph?
- What are the main graph invariants that capture the myriad properties of such large graphs?

To answer these problems, we first delve into the wealth of knowledge from the past, although it is often not enough. In the past thirty years there has been a great deal of progress in combinatorial and probabilistic methods, as well as spectral methods. However, traditional probabilistic methods mostly consider the same probability distribution for all vertices or edges while real graphs are uneven and clustered. The classical algebraic and analytic methods are efficient in dealing with highly symmetric structures, whereas real-world graphs are quite the opposite. Guided by examples of real-world graphs, we are compelled to improvise, extend and create new theory and methods. Here we will discuss the new developments in several topics in graph theory that are rapidly developing.

2. Collaboration graphs

A social network is a collection of people, each of whom is acquainted with some subset of the others. Such a network can be represented as a set of points (or vertices) denoting people, joined in pairs by lines (or edges) denoting acquaintance. One could, in principle, construct the social network for a company or firm, for a school or university, or for any other community up to and including the entire world.

Social networks have been the subject of both empirical and theoretical study in the social sciences for at least 50 years [1–3], partly because of inherent interest in the patterns of human interaction, but also because their structure has important implications for the spread of information and disease. It is clear, for example, that variation in just the average number of acquaintances that individuals have (also called the average degree of the network) might substantially influence the propagation of a rumor, a fashion, a joke, or the year's flu etc.

3. Six Degrees of Separation

Explicit networks of connections are constructed by using data drawn from a number of databases, including MEDLINE (Bio-Medical Research), the LOS Alamos e-print archive (physics) and NCSTRL (Networked computer science reference Library). These collaboration networks form "Small worlds" in which randomly chosen pair of scientists are typically separated by only a short path of intermediate acquaintances. Results for mean and distribution of numbers of collaboration of authors demonstrate the preference of clustering in the networks, and highlight a number of apparent differences in the patterns of collaboration between the fields studied.

A famous early empirical study of the structure of social networks, conducted by Stanley Milgram [12], 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's. 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 Milgram's 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 [9]. Although there were certainly biases present in Milgram's experiment—letters that took a longer path were perhaps more likely to get lost or forgotten, for instance [16]— his result is usually taken as an evidence for the ''small-world hypothesis,'' The following shortcomings limit the usefulness of these studies.

1. First, these studies are labor intensive, and the size of the network that can be mapped is therefore limited to a few tens or hundreds of people.
2. Highly sensitive to the subjective bias on the part of the interviewers: what is considered to be an acquaintance can differ considerably from one person to another.

To avoid these issues, a number of researchers have studied networks for which there exist more numerous data and more precise definition of connectedness. Examples of such networks are the electric power grid [14,15],the internet [4,7], and the pattern of air traffic between airports [5]. But these networks have a different problem. They are social in the sense that their structure in some way reflects the features of the society that built them, but they do not directly mean the actual contact between people. But researchers are still interested for their own sake. So if we want to study a genuine network of human acquaintances, that is large-containing over a million people and for which a precise definition of acquaintance is possible.

Then such a network is the network of scientific collaboration as documented in the papers scientists write. Famous CGs in variety of fields come from five data bases.

- **MEDLINE** (which covers published papers on bio medical research),
- **the Los Alamos e-Print Archive** (preprints primarily in theoretical physics),
- **SPIRES** (published papers and preprints in high-energy physics),
- **NCSTRL** (preprints in computer science).
- **Math Sci Net** (which covers published papers on mathematics and Science)

The idea of study of collaboration patterns by using data drawn from the publication record is not new. There is a substantial body of literature available in information science dealing with coauthor ship patterns and co citation patterns. However, no detailed reconstruction of an actual collaboration network has previously been attempted. The nearest thing to such a reconstruction comes not from Information Science at all, but from the mathematics community, within which the concept of the Erdos number, has a long story. In addition to distance between authors, there are many other interesting quantities to be measured on collaboration networks, including the number of collaborators of scientists, the number of papers they write, and the degree of clustering which is the probability that two of a scientist's collaborators have themselves collaborated.

4. Motivation for the study of Collaboration graph

Many people have the impression that Math research is largely a solitary pursuit. They imagine a mathematician squirreled away in some dingy garret, in a lonely wilderness cabin, Oblivious to everyday's concerns and focused on a single problem, scribbling inscrutable equations across scraps of paper and thinking long and hard before emerging with a eureka and a proof. But it not so. The dramatic announcement in 1993 by Andrew Wiles that he had proved Fermat's last theorem appeared to belong to this type of category. Although it is a fact that he had virtually separated himself from the rest of the mathematical community for nearly eight years to work on this problem, actually he had relied heavily on the previous work of other mathematicians who had tackled the same problem. He has occasionally tested his ideas on a few trusted experts in that area of mathematics relevant to his approach. And when reviewers later discovered a flaw in his original chain of logic, he obtained help from one of his former graduate students, Richard Taylor, to fill in the gap and complete the proof. At the same time, the relative isolation that he sought is certainly not the rule in mathematical research. Doing mathematics is really a remarkably social process. The abundance of meetings, conferences, workshops, colloquia, seminars and other gatherings of mathematicians attest to the importance of collaboration. E-communication speeds and facilitates such interaction.

Paul Erdos has epitomized the strength and breadth of mathematical collaboration. He traveled around the world for decades in search of new mathematics and new collaborators. His efforts have become legendary in mathematical circles, and mathematicians have taken a characteristically mathematical way of describing them by inventing a new quantity called an Erdos Number. 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. This score keeping system although has started years ago, it is Jerrold. W.Grossman at Oakland university in Rochester, Michigen., has become the complier and Guardian of the list of collaborators. Working with Patrick D.F.Ion, an editor of Mathematical Reviews in Ann.Arbor.Michigen, he continues to update, correct, and expand his files on Erdos Numbers.

When Mathematical Reviews first started compiling a record of all published mathematics work(about 75 years ago), more than 90% of papers were solo works. Now scarcely more than half are individual efforts, and the fraction of two author papers has risen from less than10% to more than one third. Also more than 10% of papers in mathematical sciences have 3 or more coauthors and more than 2% with 4 or more, Grossman says. One can also look at mathematics collaborations as a graph. The resulting collaboration graph CG is a monstrous tangle that snares nearly all mathematicians with tentacles reaching into computer science, physical & biological sciences, economics and even the social sciences. By determining the smallest number of edges linking any person to Erdos in this enormous CG, one can determine that person's Erdos number. Albert Einstein has Erdos number 2. Andrew wiles has Erdos number 4.

5. What is Collaboration graph?

In mathematics and social science, a collaboration graph is a graph modeling 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. CGs are used to measure the closeness of collaborative relationships between the participants of that network. By construction, it is a simple graph. CG need not be connected. That is., for example , people who never co-authored a joint paper represent isolated vertices in the CG of mathematicians. CG of mathematicians were shown to have "small world topology". That is., they have a large number of vertices, most of small degree, that are highly clustered, and a "giant" connected component with small average distances between vertices. The distance between two vertices in a CG is called a collaboration distance. Then the collaboration distance between two distinct nodes is equal to the smallest number of edges in an edge-path between them. If no such path exists, then the collaboration distance is defined to be infinite.

In CG of mathematicians, the collaboration distance from a particular person to Paul Erdos , is called the Erdos Number of that person. Math Sci Net has a free online tool for computing the collaboration distance between any two mathematicians as well as Erdos number of a mathematician. The tool also shows the actual chain of coauthors that realizes the collaboration distance. Some generalization of CG of mathematicians are for example, there is a hyper graph version, where individual mathematicians are vertices, and where a graph of mathematicians (not necessarily distinct) constitutes a hyper edge if there is a paper, where they were all coauthors. Another variation is a simple graph where two are joined by an edge if there is a paper with only two of them as co-authors. A multi graph version of a CG : two are joined by k-edges, if they authored exactly k-papers together. Another variation is a weighted collaboration graph, with rational weights, where two are joined by an edge with weight $\frac{1}{k}$ whenever, they have coauthored exactly k-papers together. This model leads to the notion of rational Erdos Number.

6. Salient features

Purists can argue over how to count papers with more than two coauthors. But here we will adopt the liberal attitude that each of the $\binom{k}{2}$ pairs of authors in a k-author paper are adjacent in the CG. 511 authors have Erdos Number 1. Most of them are still alive. So the number of people with Erdos Number 2 continues to grow rapidly, as does the order of the component of C containing these people, which is clearly the largest component of CG so far.

6.1 The people with most papers with Erdos

The people with most papers with Paul Erdos

Andras Sarkozy	57
Andras Hajnal	55
Ralph J.Faudree	46
Richardh.Schelp	40
Vera T.Sos	35
Cecil Clyde Rousseau	33
Alfred A.Renyi	32
Endre Szemeredi	30
Pal Turan	30
Stefan Andrus Burr	27
Ronald L.Graham	27
Joel H.Spencer	23
Miklos Simonovits	22
Carl Pomerance	21

6.2 The people with Erdos number 1 having the most coauthors

The people with Erdos number 1 having the most coauthors

Frank Harary	273
Noga Alon	161
Saharon Shelah	142
Ronald L.Graham	121
Charles J.Colbourn	120
Daniel J.Kleitman	115
Andrew M.Odlyzko	111

So far approximately CG has 4 lakh vertices and 7 lakh edges. In CG_{math} , the number of vertices of given degree is approximately proportional to a power of the degree. How to model and analyze social Interactions? The database of authored items give rise to a bipartite graph B_{math}, whose vertices of one type are papers and vertices of the other type are the coauthors. The B_{math} has about 2.8 million edges from which it follows that the number of authors per paper is ≈ 2 and the number of papers per author is 8.

6.4 Interesting features of CG(math)

1. Number of edges is small that is, a little more than the number of vertices. The average length of the path between vertices in the same component is small.
2. There is a giant component which accompanies a majority of the coauthors, and the remaining components are very tiny.
3. The degrees of the vertices in the CG follow a "Power law" pattern-the number of vertices of degree x is proportional to a (negative) power of x.
4.The clustering coefficient is very high where he clustering coefficient of a graph is a fraction of ordered triples of vertices a, b, c in which edges ab and bc are present that have edge ac present. In other words, how often two neighbours of a vertex are adjacent to each other.

Question What model of random graphical evolution will produce graphs with these properties of the CG?

7. More on Scientific Collaboration Networks

We study networks of scientists in which two scientists are considered connected if they have coauthored a paper together. This seems a reasonable definition of scientific acquaintance: most people who have written a paper together will know one another quite well. It is a moderately stringent definition, since there are many scientists who know one another to some degree but have never collaborated on the writing of a paper.

The fact that some of the databases used contain unrefereed preprints should not be regarded negatively. Although unrefereed preprints may be of lower average scientific quality than papers in peer-reviewed journals, they are, as an indicator of social connection, every bit is as good as their refereed counterparts. Estimating the true number of distinct authors in a database is complicated by two problems. First, two authors may have the same name. Second, an author may identify themselves in different ways on different papers, e.g., using first initial only, using all initials, or using full name. In order to estimate the size of the error introduced by these effects, all analyses reported here have been carried out twice. The first time we use all initials of each author. This will rarely confuse two different authors for the same person (although this will still happen occasionally), but sometimes misidentifies the same person as two different people, thereby overestimating the total number of authors. The second analysis is carried out using only the first initial of each author, which will ensure that different publications by the same author are almost always identified as such, but will with some regularity confuse distinct authors for the same person. Thus these two analyses give upper and lower bounds on the number of authors, and also give an indication of the expected precision of many of our other measurements.

7.1 Erdös Numbers and the Collaboration Graph

The following interesting facts about the collaboration graph and Erdös numbers are mostly based on information in the database of the American Mathematical Society's Mathematical Reviews (MR) as of July, 2004. Internet access to MR data is provided by the service MathSciNet. We gratefully acknowledge the assistance of the AMS in making this information available.

There are about 1.9 million authored items in the Math Reviews database, by a total of about 401,000 different authors. (This includes all books and papers in MR except those items, such as some conference proceedings, that do not have authors.) Approximately 62.4% of these items are by a single author, 27.4% by two authors, 8.0% by three authors, 1.7% by four authors, 0.4% by five authors, and 0.1% by six or more authors. There are four authors with more than 700 papers: Paul Erdös with 1416 (he actually wrote more papers than that, but these are just the ones covered by Math Reviews), Drumi Bainov with 823, Saharon Shelah with 760, and Leonard Carlitz with 730. Bainov's Erdös number is 4, SHELAH's is 1, and Carlitz's is 2. The other mathematicians with more than 500 papers listed in MathSciNet (and their Erdös numbers) are Hari M. Srivastava (2), Lucien Godeaux (infinite — actually he wrote only one joint paper), Ravi Agarwal (3), Edoardo Ballico (3), Frank Harary (1), Josip E. Pecaric (2), Shigeyoshi Owa (3), and Richard Bellman (2). The most prolific authors listed in the DBLP (dealing with computer science publications) can be found on a list at their website (DBLP), which is definitely worth exploring.

8. Author Name Disambiguation for Collaboration Graph

Andreas Strotmann, Dangzhi Zhao and Tania Bubela have propsed an algorithm for disambiguating author names of publications via deterministic clustering based on well-defined similarity measures between publications in which their names appear as authors. The algorithm was designed to be used for constructing a collaboration network, i.e., a graph of author nodes and co-author links. In this context, the goal was to produce a co-authorship graph with network characteristics that are close to those of the "true" collaboration network, so that meaningful network metrics can be determined from it. The algorithm is fairly easily comprehended as it does not depend on any black-box AI techniques. This is important in the context of policy studies, in which it was successfully applied and it enables policy makers to judge the soundness of the methodology with considerable confidence. It is also fast, making it possible to run large-scale analyses (here, in the order of a hundred thousand publications and the order of a million names to be disambiguated) on a moderately sized desktop computer within a few days. The algorithm is open to improvement via extensions that take into account additional kinds of fields in bibliographic records of publications in order to provide evidence that two occurrences of similar names belong to the same individual. We simply give the algorithm here for the sake of completeness.

8.1 Author Name Disambiguation Algorithm, Phase 1

In the first phase of the disambiguation process, the algorithm classifies name occurrences as referring to a single individual if:

1. All author names for this individual are mutually *compatible*.
2. There is *positive evidence* that indicates that all occurrences refer to the same individual.
3. There is *more* positive evidence for this particular choice than for potential alternatives.

In this algorithm, author names are considered *compatible* if

1. Their last names and first initials reduce to identical ASCII sub-sequences after Unicode compatibility decomposition normalization (referred to below as *normalized and reduced*).

2. Both sequences of first names consist of compatible first names in the same order, where two first names are compatible if

a. One of them is empty (i.e., the corresponding sequence of the other name is a compatible extension of this one); or if

b. One of them is an initial and that initial normalizes and reduces to the first letter of the reduced and normalized other first name (which may be an initial, too); or if

c. Neither first name is an initial or empty (i.e., they are both full first names) and they both normalize and reduce to the same sequence of ASCII characters.

A number of different types of evidence are considered as *positive evidence* that two occurrences of similar names refer to the same individual:

1. Equivalence (as opposed to compatibility) of full names, where equivalence is determined by (a) separately normalizing and reducing last names and first name sequences, (b) removing special characters from these reduced name parts, and (c) requiring equality of the resulting sequences.

2. The names occur in two (different) papers whose respective coauthor lists have more than one member in common (i.e., at least one member beyond the two names under consideration each). "In common" means that the names of those coauthors are compatible in the above sense.

3. There are common topics covered by the two papers in which the two similar names occur:

a. The papers appear in the same journal; or

b. The papers share a common major MeSH code assigned by PubMed.

Finally, when considering which of several potential candidate individuals to assign a particular name occurrence to, or when considering whether to consider two previously separate individuals as one, a *degree of similarity* between two individuals, i.e., between two groups of occurrences of similar names, is computed as follows:

1. For each of the different types of positive evidence described above, a list of common features between the two groups is compiled (e.g., common co-authors or common journal names). For a group of occurrences, the union of features of all members of that group is used.

2. Each of the common features is then weighted by the maximum number of times that that feature occurs in one of the groups, since a feature that occurs many times in a group may be considered characteristic of that group, and for two groups to have such a characteristic feature in common is strong evidence that they should be considered one group.

3. Finally, the similarity measure is computed by summing up the weights of all common features.

With these concepts in place, the first phase of the algorithm iterates over all the different normalized and reduced last-name-plus-first-initial combinations that occur in the data set. For each such combination, the algorithm assigns each and every occurrence of an author name that normalizes and reduces to this combination to one of possibly several individuals, as follows:

1. Initially, those groups of name occurrences that normalize and reduce to identical full names (i.e., to names that contain full first names rather than only initials as first names) are considered separate individuals, while initials-only name occurrences are left in a separate list for later processing.

2. For each of these initial individuals whose names contain initials, we pair this individual with another individual (if present) whose name is equivalent except for the initial, as these two individuals are likely candidates for merging into one. If no such pairings exist, and the list of initials-only name occurrences is empty, the result of the first step is considered the final result of the algorithm.

3. For each (“stripped”) individual that is paired in step 2 with (“expanded”) individuals that have additional name components, attempt to distribute that individual’s oeuvre over the expanded individuals. For each name occurrence of the “stripped” individual name in its oeuvre, therefore, we

a. Compute that name occurrence’s similarity to each of the “expanded” individuals.

b. If there is an “expanded” individual (i) with positive evidence (in the sense above) that the name occurrence might be that of the expanded individual’s; (ii) containing only compatible name occurrences; and (iii) with stronger positive evidence than any other expanded individual (if any), the “stripped” name occurrence is moved to that expanded individual.

4. Now process the list of initials-only name occurrences of step 1, using the same method as in step 3, but with the initials-only name occurrence as a “stripped” individual which is paired with each one of the individuals currently under consideration in this iteration.

5. For those initials-only name occurrences that could not be mapped to full name individuals in the previous step, build a graph of author name occurrences connected by an edge whenever there is (any) positive evidence that these two occurrences refer to the same individual. The connected components of this graph are then each considered as separate individuals to be added to the set of individuals as above.

Author Name Disambiguation Algorithm, Phase 2

Preliminary explorations of the resulting collaboration network showed that the clustering algorithm applied above was overly pessimistic, separating known authors into several

“individuals”. Therefore it is decided to balance the excess of false negatives (i.e., missed identifications of individual clusters) at the expense of possible false positives (i.e., spurious identifications of two distinct individuals as one) in the hope of gaining a network with statistical properties that are more like those of the (unattainable) “correct” network of individuals and their collaborations.

In this phase of the algorithm, rely exclusively on collaboration links: if two individuals have similar names and have at least one collaborator in common, but have never collaborated with each other, they are likely a single individual.

For all sets of author names with identical normalized and reduced last name and first initial, therefore

1. Construct a graph that links two individuals whenever they have a common co-author;

2. Determine the non-trivial connected components of this graph;

3. For each connected component, merge its member individuals into a single individual – unless that individual would end up being his or her own co-author by doing so.

This step is repeated twice instead of repeating it until it converges, in an attempt to keep a balance on potential false-positive errors introduced through this method.

Open Questions

It could be thought a priori that in order for a mathematician to make his or her entrance into the Erdo's component of C it is necessary to have many co-authors. But one of the important conclusions we can draw from the compilation of data for this article is that what really matters is not how many people you publish with but whom you publish with. That is, quality is more significant than quantity. We believe this is the case for the following two reasons. First, the most prolific mathematicians of the last century seem to have a finite Erdo's number and secondly, the Erdo's component has grown so enormous that it already embraces almost everyone. If that reasoning is not altogether convincing, consider a more dramatic example than the ones presented thus far, the case of the great Austrian logician Kurt Godel. In regard to the number of joint papers, Godel is at the other end of the spectrum from Erdos's: He wrote only one (see[10]), and that is a one-page note (in German) he wrote with Karl Menger and Abraham Wald[11] concerning Menger's approach to differential and projective geometry. It turns out that Wald's Erdo's number is 2.Hence, despite his precarious number of joint papers, Godel still makes his way into the Erdo's component of C with a rather small Erdo's number. We close this discussion with some open questions, which even in the era of super computing and worldwide information networks, are extremely difficult to answer. The first two were already put forward in [8] but no hint to their possible solution has yet surfaced.

- In the collaboration graph C, what is the second largest component (measured by the number of its vertices)? If we restrict ourselves to looking only at mathematicians, then the second largest component is probably not enough large, but it is conceivable that there are large components in other disciplines.
- What are the radius and diameter of the Erdo's component of C (in graph theoretical terms)? Again the question would be interesting both as applied to all researchers and when restricted to mathematicians.
- The Nobel-Erdo's number is, at a given moment, the number of Nobel prize laureates having a finite Erdo's number. This number changes as new prizes are awarded and more people enter into the Erdo's component of C. It was established that the Nobel-Erdo's number is ≥ 63 but its exact value is unknown. Surely the above bound is not nearly the best possible.
- The Erdo's span measures how far back in time the connection with Paul Erdo's extends. More precisely, we can define Erdo's span as the smallest number representing the year of birth of a person with a finite Erdo's number. All we can say for now is that this number is no greater than 1849, which is the birth of Gerorg Frobenius(1849-1917), the German algebraist who made major contributions to group theory. He developed the theory of finite groups of linear substitutions mostly in collaboration with Issai Schur (1875-1941) [6]. It turns out that Schur's Erdo's number is 2 because of his 1925 joint paper [13] with Gabor Szego, co-author of Erdo's. We do not know whether the Erdo's span can be traced further back into the early 1800s.What we can be sure of is that the Erdo's connection will extend forever in to the future.

References

[1]. D. Achlioptas, R. M. D'Souza, and J. Spencer, Explosive percolation in random networks, Science 323, no. 5920, (2009), 1453–1455.

[2]. W. Aiello, F. Chung, and L. Lu, A random graph model for massive graphs, Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, ACM Press, New York, 2000, pp.171–180.

[3]. M. Ajtai, J. Komlós, and E. Szemerédi, Largest random component of a k-cube, Combinatorica 2 (1982), 1–7.

[4]. Albert, R., Jeong, H. & Baraba´si, A.-L. (1999) Nature (London) 401, 130–131.

[5]. Amaral, L. A. N., Scala, A., Barthe´le´my. M. & Stanley, H. (2000) Proc. Natl.Acad. Sci. USA 97, 11149–11152. (First Published September 26, 2000; 10.1073ypnas.200327197)

[6]. M.Beckmann&T.Marschak, An activity analysis approach to location theory, in Proc.second symposium in linear programming(Washington,DC 1955) National Bureau of standards, Washington,D.C.,PP 331-379;382f.

[7]. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R.,Tomkins, A.& Wiener, J. (2000) in Computer Networks 33, 309–320.

[8] J. W. Grossman and P. D. F. Ion, On a portion of the well-known collaboration graph, Congressus Numerantium108, 129-131 (1995).
[9]. Guare, J. (1990) Six Degrees of Separation (Vintage, New York).
[10] Kurt Godel, Collected Works, Edited By S.Feferman ,Oxford University Press,1986.
[11]Kurt Godel, K.Menger&A.Wald,Diskussion iiber koordinatenlose,Differentialgemetrie,Ergebnisse eines mathematischen kollaquims 5(1933),25-26.
[12]. Milgram, S. (1967) Psychol. Today 2, 60–67.
[13]. I.Schur&G.Szego,Uber die abschnitte einer im Einheitskreise beschrankten potenzreihe , Sitzungsherichte der preussichen Akademie der Wissenschaften 1925,Physikalisch Mathematische Klasse,545-560.
[14]. Watts, D. J. (1999) Small Worlds (Princeton Univ. Press, Princeton, NJ).
[15]. Watts, D. J. & Strogatz, S. H. (1998) Nature (London) 393, 440–442.
[16]. White, H. C. (1970) Social Forces 49, 259–264.