2-2012

# The social network of software engineering research

Subhajit DATTA
*Singapore Management University*, subhajitd@smu.edu.sg

Nishant KUMAR

Santonu SARKAR

# The Social Network of Software Engineering Research

Subhajit Datta*
IBM Research
Bangalore
India
subhajit.datta@acm.org

Nishant Kumar
Accenture Technology Labs
Bangalore
India
kumar.nishant1@gmail.com

Santonu Sarkar
Infosys Labs
Bangalore
India
santonus@acm.org

## ABSTRACT

The social network perspective has served as a useful framework for studying scientific research collaboration in different disciplines. Although collaboration in computer science research has received some attention, software engineering research collaboration has remained unexplored to a large extent. In this paper, we examine the collaboration networks based on co-authorship information of papers from ten software engineering publication venues over the 1976-2010 time period. We compare time variations of certain parameters of these networks with corresponding parameters of collaboration networks from other disciplines. We also explore whether software engineering collaboration networks manifest symptoms of the small-world phenomenon, conform to the criteria of "social networks", and manifest increasing collaboration with time. In the light of these observations, we highlight some general characteristics of collaboration in software engineering research. The results presented in this paper facilitate understanding of the progression of software engineering from its infancy to maturity, and lay the foundation for developing theoretical models to explain the evolution of its research collaboration characteristics.

## Categories and Subject Descriptors

D.2.9 [**Software Engineering**]: Management—*Life cycle*; H.5.3 [**Information Systems**]: Group and Organization Interfaces—*Collaborative computing, Computer-supported cooperative work*; J.4 [**Social and Behavioural Sciences**]: Sociology

## General Terms

Experimentation

*Corresponding author


*ISEC'12* Feb 22-25, 2012, Kanpur, UP, India


## Keywords

software engineering research, collaboration, power law, social network analysis, clustering

## 1. INTRODUCTION

Research publications represent a significant outcome of scientific collaborations between individual researchers. In 1976 the first dedicated venues for publishing software engineering research were established – IEEE Transactions on Software Engineering (TSE) and International Conference on Software Engineering (ICSE). Since then several other specialized software engineering publication venues have come up, and the body of research literature in software engineering has grown significantly. Given a body of papers, a network can be constructed where the vertices (nodes) are the authors, and the edges (undirected links), represent co-authorship relations between the authors at either end of the edge. In other words, if there exists at least one paper jointly written by two authors, there will be an edge in the network between the corresponding vertices representing the authors. This type of networks have been used extensively to study scientific collaboration in many disciplines [4], [18], [6]. In this paper, we define corresponding co-authorship networks, which we call *software engineering research collaboration networks* (SRCN), based on a body of papers published in ten software engineering publication venues in the period 1976-2010.

Over the last few decades, software engineering research has assumed significance – and a unique character, as we seek to establish in this paper – in the light of the increasing penetration of software systems in many aspects of our individual and collective lives. We believe it can be illuminating to study how the characteristics of software engineering research collaboration have changed with time, in comparison with corresponding variations for other disciplines that have been reported. This paper reports the results of such a study carried out over a body of 11,429 papers involving 15,207 authors in the period from 1976 to 2010.

The next section discusses the contribution of the research followed by a brief overview of related work. Next we describe the methodology of the study. The results are presented next, followed by a discussion around the results, open issues and directions of future work, and conclusions from the study.

## 2. CONTRIBUTION OF THE RESEARCH

As discussed in the Related Work section, study of the

time variation of collaboration characteristics of co-authorship networks have been carried out at depth for other disciplines; and collaboration in computer science as well as a specific sub-area within software engineering have been studied to some extent. To the best of our knowledge, *this is the first detailed study of variations of the characteristics of co-authorship networks of software engineering publications for a period of time from the inception of specialized publishing venues for software engineering research to the present.* In terms of the ten publication venues selected, we seek to capture a significant body of published literature in software engineering in the last 34 years.

That the scope of our study starts from the very beginning of organized publication of research in the discipline has a subtle but significant implication. In many existing studies of research collaboration in other disciplines, the so-called problem of "missing past" [14] has influenced explanations of the observed characteristics [4]. In collaboration networks that are constructed from datasets which do not go back to the network's birth – that is, the very inception of recorded collaboration in the discipline – there is the problem of "phantom" vertices and edges, which invariably existed before the observation period started and end up impacting observations for the period being studied.

From our study we are able to infer certain general characteristics of software engineering collaboration that illuminate the progression of the discipline.

## 3. RELATED WORK

We consulted three detailed reviews on the science of complex networks which discuss at depth the characteristics of these systems, models proposed to describe them, the dynamics of their evolution, and diverse fields of their applications: *The structure and function of complex networks* by Newman [19], *Statistical mechanics of complex networks* by Albert and Barabasi [3], and *Evolution of networks* by Dorogovtsev and Mendes [10].

A pioneering work in the study of the evolution of scientific collaboration is reported by Barabasi et al. in [5]. The authors explore the co-authorship networks in mathematics and neuro-science in the period 1991-1998 and infer that the networks are scale-free and their evolution is governed by preferential attachment. They also suggest a model to explain the networks' evolution. We take the results of this paper as benchmarks in comparing the characteristics of software engineering research collaboration networks; please refer to discussion in the Results and subsequent sections for more details.

The structure of scientific collaborations has also been examined in detail by Newman; he established that such collaboration networks form small-worlds where pairs of randomly selected scientists are typically short distances away from one another and the networks show significant clustering [18]. Newman takes forward his exploration of scientific collaboration networks in two subsequent papers, where the statistical properties of these networks are studied, along with the existence and size of a giant component, and other non-local characteristics such as closeness and betweenness [16], [17]. Newman's work illuminates how scientific collaboration in different disciplines manifest subtly different patterns. Newman and Park study the innate differences between social networks and other types technological or biological networks in [21]; we use results from this paper to address some of our results in a subsequent section.

**Table 1: Inception of Venues**

| Venue | First Published |
|---|---|
| TSE | 1976 |
| ICSE | 1976 |
| SW | 1984 |
| OOPSLA | 1986 |
| ECOOP | 1987 |
| TOSEM | 1992 |
| FSE | 1993 |
| ISSTA | 1993 |
| ASE | 1997 |
| FASE | 1998 |

Evolution of research collaboration networks based on co-authorship information for the computer science discipline in the period 1980 to 2005 have been studied by Huang et al. [12]. They consider characteristics specific to six sub-categories within computer science – artificial intelligence, applications, architecture. database, system, and theory – to reach the conclusion that the database community is the best connected, while the artificial intelligence community is most assortative, and computer science as a field is more similar to mathematics than to biology. Additionally, the authors observe the small-world phenomenon and scale-free degree distribution accompanying the growth of the network. Interestingly, the authors have *not* studied software engineering as a sub-category within computer science.

Bird et al. construct a collaboration network using a snapshot of DBLP bibliographic data in computer science, define 14 sub-areas (including software engineering) and use topological measures to examine behaviours of individuals and collaboration patterns across areas in terms of how centralized, integrated and cohesive they are [6]. The authors conclude that data mining and software engineering are very "interdisciplinary", while cryptography and theory are not; cryptography is highly isolated within the computer science discipline as a whole, but densely connected internally.

Hassan and Holt study the collaboration networks based on co-authorship data from the proceedings of the Working Conference on Reverse Engineering (WCRE) for the period 1993-2002 and conclude that these have properties of small-world networks, though the small worlds of software engineering are usually bigger than in other small world networks [11].

The incidence of power laws in real world networks and the generation and detection mechanisms for power law behaviour have been investigated in [20], [8].

## 4. METHODOLOGY

The software engineering publication venues from which the papers were extracted are given in Table 1 with their respective years of inception. In the order listed, the abbreviations sand for: IEEE Transactions on Software Engineering (TSE), International Conference on Software Engineering (ICSE), IEEE Software (SW), Object-Oriented Programming, Systems, Languages and Applications (OOPSLA – recently renamed as SPLASH), European Conference on Object-Oriented Programming (ECOOP), ACM Transactions on Software Engineering and Methodology (TOSEM),

Foundations of Software Engineering (FSE), International Symposium on Software Testing and Analysis (ISSTA), Automated Software Engineering (ASE), and Fundamental Approaches to Software Engineering (FASE). We selected those venues which are exclusively focussed on general software engineering themes. Although we do not claim this to be an exhaustive list, we believe it covers most of the significant venues in software engineering. Purists may point out that IEEE Software is a *magazine* and not entirely focused on the dissemination of research results. The reason we chose to include IEEE Software is that several articles "expounding pioneering ideas that had a significant impact" [1] were published in this magazine over the years; [13], [15], [7], [9] to name just a few. A list of influential IEEE Software articles is available in [1].

Some of the conference venues we consider have had a number of workshops associated with them over the years. Bibliographic information from these co-located workshops have been ignored in our dataset. We also do not include peripheral tracks such as doctoral symposium etc. in the data extracted from the conference proceedings.

Bibliographic information for these venues is available in the public domain at Web based repositories such as ACM Portal, IEEE Xplore, and DBLP. The BibTex files obtained from these repositories are parsed and relevant information is persisted in a Derby[1] database for ease of retrieval. Each Bibtex field is mapped to a field in a database table. The database is then queried to generate an adjacency matrix $M$ where rows represent paper titles and the columns represent corresponding author(s). Let the papers be numbered from 1 to $i$ and authors from 1 to $j$. Then if author $y$ is (one of) the author(s) of paper $x$ then the corresponding entry in the adjacency matrix $M(xy) = 1$; otherwise it is 0. Using the duly completed adjacency matrix $M$, the final network is generated in a Pajek[2] file format.

If the same author has chosen to be identified by different variants of his/her name for different papers, the variants of the names are treated as different individuals. If two different authors happen to have the exact same name, they are treated as the same individual. This is a common problem in networks where vertices represent individuals [2]. From manual checks we have performed, we have reason to believe that though these ambiguities are present in our dataset, they constitute a very minor portion of the 15,207 unique author names considered.

For understanding the time variation of the software engineering collaboration networks, we have chosen 11 cumulative time-steps as defined in Table 2. We will have the occasion to consider non-cumulative time steps also; these are defined in Table 3. To denote the software engineering collaboration networks for a specific time-step, we use the notation SRCN(start_year-end_year). So, for example, SRCN(1976-1994) denotes a network based on all the papers cumulatively published between and inclusive of the years 1976 to 1994.

We define the set $S(c) = \{$SRCN(1976-1979), SRCN(1976-1982), ... , SRCN(1976-2010)$\}$ to be the set of all software engineering research collaboration networks over cumulative time-steps, and $S(nc) = \{$SRCN(1976-1979), SRCN(1980-1982), ... , SRCN(2007-2010)$\}$ to be the set of all software engineering research collaboration networks over non-cumulative time-steps.

**Table 2: Time-steps: Cumulative**

| Time-step No | Period |
|---|---|
| 1 | 1976-1979 |
| 2 | 1976-1982 |
| 3 | 1976-1985 |
| 4 | 1976-1988 |
| 5 | 1976-1991 |
| 6 | 1976-1994 |
| 7 | 1976-1997 |
| 8 | 1976-2000 |
| 9 | 1976-2003 |
| 10 | 1976-2006 |
| 11 | 1976-2010 |

**Table 3: Time-steps: Non-cumulative**

| Time-step No | Period |
|---|---|
| A | 1976-1979 |
| B | 1980-1982 |
| C | 1983-1985 |
| D | 1986-1988 |
| E | 1989-1991 |
| F | 1992-1994 |
| G | 1995-1997 |
| H | 1998-2000 |
| I | 2001-2003 |
| J | 2004-2006 |
| K | 2007-2010 |

After generating the networks, the parameters of interest (as reported in the Results section) are calculated using the tools Pajek, Gephi[3], and NodeXL[4]. Whether degree distributions follow a power law was verified using the methodology described in [8], and using programming resources available at `http://tuvalu.santafe.edu/~aaronc/powerlaws/`.

# 5. RESULTS

The entire software engineering research collaboration network considered in our study consists of 15,207 authors and 11,429 papers. To understand how the characteristics of software engineering research collaboration changes with time, we first look at the change in the number of new authors and papers in each non-cumulative time-step in Figure 1. This can be seen in the context of the time line of inception of the venues (Table 1) we have considered in building the collaboration networks.

We now move to the characteristics of $S(c)$. Figure 2 shows the degree distribution on a log-log plot for SRCN(1976-2010). As evident from the figure and the computed value of goodness of fit $p = 0.26$ ($p$ greater than 0.1 indicates power law characteristics), the degree distribution follows a power law; with values of the scaling parameter of $\alpha = 3.233$ and threshold value $x_{min} = 7.5$ [8]. The growth of the number of vertices(V) and edges(E) of $S(c)$ is shown in Figure 7. As evident, edges grow faster than vertices in the later time-steps

[1]http://db.apache.org/derby/
[2]http://pajek.imfm.si/doku.php
[3]http://gephi.org/
[4]http://nodexl.codeplex.com/

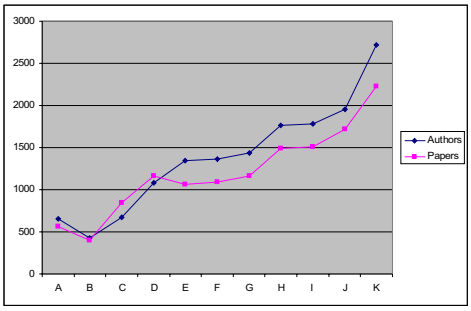


**Figure 1: New Authors and Papers Added for Non-Cumulative Time-steps Defined in Table 3**

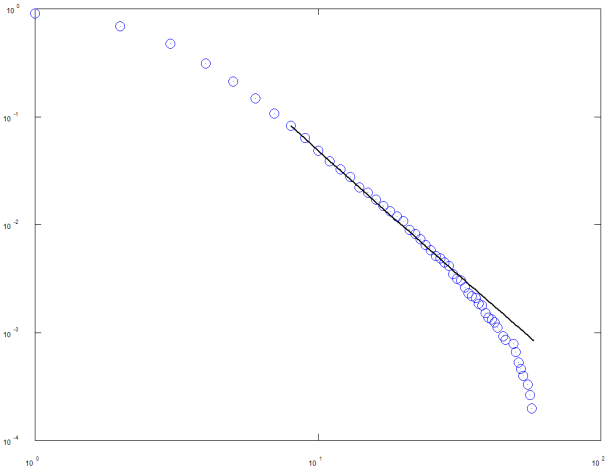

**Figure 2: Degree Distribution of SRCN(1976-2010)**

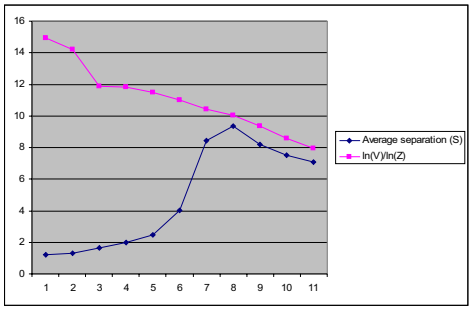


**Figure 3: Average Separation and Scaling Factor over Cumulative Time-steps Defined in Table 2**

than in the earlier ones.

The facility for two vertices $i$ and $j$ to contact one another depends on the length of the shortest path $l_{ij}$ between them. The average $l_{ij}$ over all pair of vertices is called the *average separation* of the network, and is denoted by $S$. Average separation is an indication of the interconnectedness of the network [4]. In Figure 3 we plot the variation of the average separation and a scaling factor $ln(V)/ln(Z)$ over time for $S(c)$ ($Z$ is the average degree). The average separation goes up and then shows a decreasing trend towards the later time-steps. The values of the scaling factor converges closer to the average separation in the later time-steps; the significance of this observation will be highlighted in subsequent discussion.

The diameter ($D$) of a network is represented by the length in number edges of the longest geodesic distance (that is, shortest path) between any two vertices [19]. It indicates the maximum distance that may need to be travelled for one vertex to contact any other vertex. The average degree $Z$ of a network is the number of edges per vertex; it reflects on the variation in the number of vertices and edges over time. The diameters(D) and average degrees(Z) of $S(c)$ are depicted in Figure 4. The average degree grows monotonically with time, as is expected in view of the higher rate of growth of the edges vis-a-vis the vertices in the network (Figure 7). However the diameter of $S(c)$ shows an interesting trend with the points of inflection in the latter time-steps.

A key difference between real world networks and completely random networks is the phenomenon of *clustering*. It is usually observed in social networks that two vertices that are linked to a third are more likely to be themselves linked. Intuitively, two of one's friends have a higher probability of being friends themselves. This is measured by the *clustering coefficient*. For a vertex $v$ with a degree $k_v$, there are $k_v$ neighbours of $v$. If all of these $k_v$ neighbours were linked, there would be $k_v$ choose 2 or $k_v * (k_v - 1)/2$ links between them. Let $N_v$ be the *actual* number of links between them. Then the clustering coefficient $C_v$ of node $v$ is defined as the ratio of the actual number of links and the maximum number of links between $k_v$ neighbours of $v$, and is given by $C_v = \frac{2*N_v}{k_v(k_v-1)}$. For the entire network, the clustering coefficient $CC$ is average of $C_v$ across all vertices [4]. For research collaboration network based on co-authorship of papers, if author A *independently* co-authored papers with authors B

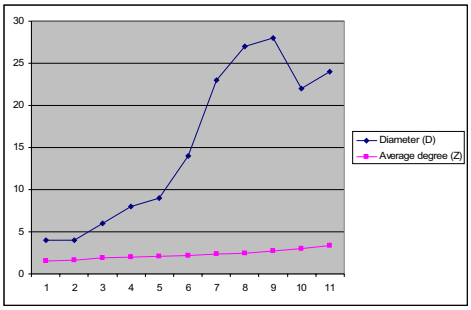


**Figure 4: Diameter and Average Degree over Cumulative Time-steps Defined in Table 2**

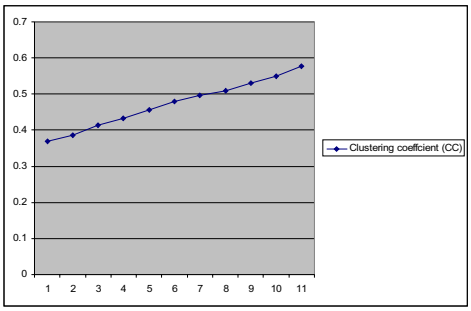


**Figure 5: Clustering Coefficient over Cumulative Time-steps Defined in Table 2**

and C, $CC$ represents the probability that authors B and C, co-authored a paper together. The variation of the clustering coefficient with time is plotted in Figure 5; it shows a monotonically steady increase.

Networks of scientific collaboration are usually fragmented in many clusters. This many be due to some scientists predominantly writing single-author papers, or inadequate data on past collaborations. With time, the clusters start getting connected with one another. To understand this progression, examining the relative size $L$ of the largest cluster (also known as the *giant component* [18]) as a percent of the number of vertices of the corresponding entire network is helpful [4]. Figure 6 shows the values of $L$ for $S(c)$ for the time-steps; till the fifth time-step it has a steady low value less than 5%; beyond the sixth time-step, it exhibits steady increase at a significantly high rate.

How do we interpret these results in the light of existing literature on collaboration networks, as well as the specific nature and context of software engineering research collaboration?

# 6. DISCUSSION

## 6.1 Degree distribution

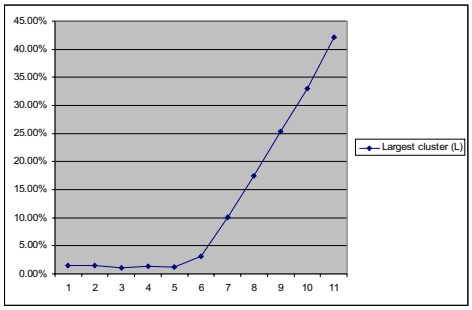


**Figure 6: Size of Largest cluster over Cumulative Time-steps Defined in Table 2**

In the study of networks, the *degree distribution* $P(k)$ – which represents the probability that a randomly selected vertex will have the degree of $k$; that is, $k$ edges incident on it – is an important parameter [4]. Networks having a power-law tail for $P(k)$ are known as *scale-free* networks; classical network models for random graphs such as the Erdos-Renyi model [3] or the Watts-Strogatz model [24], [22] have exponentially decaying $P(k)$ and are known as *exponential networks*.

The manifestation of power law (usually with an exponential cut-off) in the degree distribution of the collaboration networks of scientific research has been widely observed [4], [18]. Its implication is generally construed as: there are few individuals with many collaborators, and many individuals with few collaborators. This seems a plausible reflection of the reality of scientific collaboration; after all there are always the likes of Paul Erdos (the prolific mathematician who inspired the *Erdos number* metric) in every field. As a generative mechanism for power law in the context of scientific collaboration networks the idea of *preferential attachment* has been considered; it centres around the assumption that for a vertex, the likelihood of attracting new edges increases with the vertex's degree [3]. So according to preferential attachment, in a research collaboration network authors who have many collaborators in the past will have more collaborators in the future. To verify whether preferential attachment is indeed at the root of the power law degree distribution of SRCN(1976-2010), we decided to calculate the correlation between the degree of a vertex in time-step $(t-1)$ with the fraction of the new edges that attaches to that vertex in time-step $t$; repeating for all vertices in each successive pair of cumulative time-steps between 1976 to 2010. This method of checking whether preferential attachment holds is guided by the measurement of the probability that a vertex attaches to another vertex being proportional to the degree of the latter vertex, as outlined in [3]. The correlation coefficients for the ten pairs of preceding-succeeding time-steps ranged from 0.00005 to 0.11006, with a mean value of around 0.03. The latter time-steps show a higher correlation, with the last pair of time-steps having the largest value; but on the whole the correlation is only weakly positive. Thus preferential attachment being the *sole* driver of the power law degree distribution of SRCN(1976-2010) seems unlikely. The very low values of the correlation coefficients in the

earlier pairs of time-steps, indicated it may be interesting to check for the degree distributions for SRCN(1976-1979), ... , SRCN(1976-2006). We found that power law characteristics are *only* being manifested in SRCN(1976-2006) – in addition to SRCN(1976-2010). All other software engineering research collaboration networks in $S(c)$, *do not* have power law degree distribution. As the networks are cumulative, incremental accretion of some trend(s) appears to be reaching a critical point during the middle of the 2001-2010 decade, to reflect in power law characteristics of degree distribution in SRCN(1976-2006) and SRCN(1976-2010).

So in summary, software engineering research collaboration networks show power law in their degree distributions only in the last two cumulative time-steps. Although the indicator for preferential attachment is increasing over the time-steps, its low values do not justify the conclusion that preferential attachment predominantly guides the degree distribution in these networks.

## 6.2 Comparison of network parameters

In a seminal paper, Barabasi et al. study the evolution of the collaboration networks in the fields of mathematics and neuro-science over the period 1991-1998 and highlight the following observations: degree distribution follows power law, average separation decreases, clustering coefficient decays, size of the largest component increases, average degree increases, and attraction of edges to vertices is governed by preferential attachment [4]. We have already examined the degree distribution and relevance of preferential attachment for software engineering research collaboration context. Let us compare the other observations of Barabasi et al. for mathematics and neuro-science collaboration networks with $S(c)$.

In the networks Barabasi et al. studied, the average separation decreased over time; however for $S(c)$ it increases monotonically till the eighth time-step, and then starts going down. Barabasi et al., explain the decease in $S$ for mathematics and neuro-science due to two probable causes: increasing connectivity, leading to decreased diameter; and problem of the missing past with the databases they used to construct networks. For $S(c)$, the diameter in fact grows till the ninth time-step, after which it falls in the tenth time-step and rises again in the eleventh. Moreover, as stated earlier the problem of the missing past has been obviated in $S(c)$. So, the separation of $S(c)$ has a different characteristic than that of the networks studied in [4].

The decaying of clustering coefficients as reported in [4] have been explained by the authors as being "in agreement with the separation measurement"; they also point out that $CC$ converges to an asymptotic value in time – around 0.75 for neuro-science, and around 0.6 for mathematics. $S(c)$ shows a monotonically increasing value of $C$, going from 0.369 in the first time-step, to 0.578 in the last time-step. In the networks studied by Barabasi et al., $CC$ decreases monotonically towards settling at an asymptotic value.

Barabasi et al. report monotonic increase of the relative size of the largest cluster for both mathematics and neuro-science collaboration networks. For the former, $L$ starts close to 0 and goes up to around 0.7; while for the latter the range is from around 0.55 to 0.99. For both of these networks, the rate of increase of $L$ decreases in the later time-steps and seems to converge to some asymptotic value. In $S(c)$, $L$ increases very gradually till the sixth time-step (starting from 1.52% in the first time-step), and then goes into a regime of very rapid and steady increase, from the seventh time-step, ending at 42.11%, which is significantly lower than that reported for collaboration networks reported in other scientific disciplines [18].

The increase in average degree as reported by Barabasi et al. is also reflected in $S(c)$; for software engineering research collaboration networks, $Z$ increases from 1.545 to 3.335 across the eleven time-steps studied.

So summarizing the comparison between $S(c)$ and the mathematics and neuro-science collaboration networks as reported in [4], we can say the former is different from the latters in terms of the variation of average separation and clustering coefficient, and somewhat similar in terms of the variation of the relative size of the largest cluster, and the average degree.

## 6.3 Emerging Small World

In his exploration of the structure of scientific collaboration networks, Newman establishes that research collaboration networks form "small worlds", where pairs of randomly selected scientists are typically short distances away from one another and the networks show significant clustering [18]. Watts defines a small world graph as one with clustering coefficient lying between 0.5 and 0.8 and the average separation is approximately equal to $ln(N)/ln(Z)$ [23], [11]. With reference to Figure 5 we note that the value of the clustering coefficient first crosses 0.5 in the cumulative time-step 1976-2000 and goes up to 0.578 in the cumulative time-step 1976-2010. Figure 3 presents the variation of the average separation and the ratio $ln(N)/ln(Z)$; the values come closest to one another (less than 15% difference) in the eighth to the eleventh time-steps. So evidently, in the last four time-steps – from 2000 to 2010 – the software engineering research collaboration network shows symptoms of a small-world. It may be pointed out that the smallest separation of 7.066, recorded in the eleventh time-step is still about one degree greater than the famed "six degrees of separation" and close to the value of 7.1 reported for Computer Science co-authorship networks in [12]. The average separations of the seven scientific collaboration networks reported in [18] range from 4.0 to 9.7.

So, software engineering research collaboration networks appear to manifest small world characteristics in the decade of 2000-2010.

## 6.4 Social Networks of Collaboration

Although collaboration networks are often causally refereed to as "social networks", Newman and Park have pointed out that social networks differ from other types of networks in two important ways: non-trivial clustering (or network transitivity) and positive correlations (or assortative mixing) between the degrees of adjacent vertices [21]. We have already discussed the idea of clustering in the context of clustering coefficient; the likelihood of B collaborating directly with C, given A collaborates with B and C independently. Assortative mixing reflects on the tendency of higher degree vertices connecting with other higher degree vertices (and vice versa) in a social network. The variation of the clustering coefficient of $S(c)$ has been presented in Figure 5. To understand the level of degree correlations, we calculated the Pearson correlation coefficient between the vertices at the ends of each of the 25,511 edges at the last

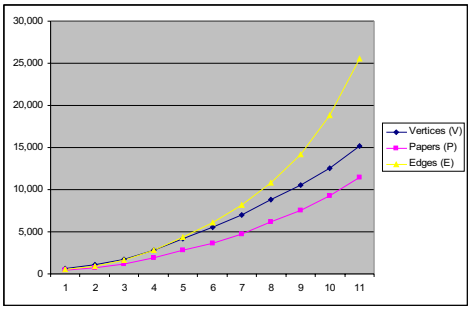


**Figure 7: Vertices and Edges over Cumulative Time-steps Defined in Table 2**

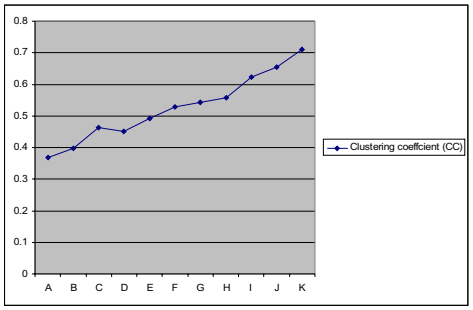


**Figure 8: Clustering Coefficient over Non-Cumulative Time-steps Defined in Table 3**

cumulative time-step of $S(c)$, as recommended by Newman in [19]. The Pearson correlation coefficient value is found to be 0.283. Thus SRCN(1976-2010) is moderately assortative as reflected by the moderate positive correlation between the degree of vertices connected by an edge. This value of the Pearson correlation coefficient, taken in conjunction with the $CC$ value of 0.578 in SRCN(1976-2010), indicates that the software engineering research collaboration network at the last time-step studied, displays moderate symptoms of being a social network, as specified in [21].

In summary, the cumulative network of software engineering research collaboration at 2010 manifests symptoms of social networks to a reasonable extent.

## 6.5 Towards Increasing Collaboration

Two interesting trends noted in Figure 5 and Figure 6 is the increasing values of the clustering coefficient and the relative size of the largest cluster. On the surface of it, both of these trends seem to point to increasing research collaboration in software engineering with the progression of time, which is also borne out by the faster increase in the number of edges vis-a-vis the number of vertices (Figure 7). However these figures are for networks over cumulative time-steps. Are previously published authors still collaborating in the new time-steps?

To address this question, we calculated the age-degree correlation of the vertices for SRCN(1976-2010). The vertex which appeared for the first time in the first cumulative time-step is ascribed the age of 11, and the one appearing in the last cumulative time-step has age of 1. Age and degree of vertices are found to have very weak positive correlation (correlation coefficient = 0.00062). Thus the fact that an author has been in the network for long does not necessarily imply (s)he has many collaborations. This seems to be plausible; it is unlikelythat the same authors would be active in research over the entire period of 34 years, or even major parts thereof.

To make a stronger case for increasing collaboration with time we need to check whether clustering coefficient and relative size of the largest cluster show similar characteristics of time variation across non-cumulative time-steps as they did for cumulative time-steps. In Figure 8 and Figure 9 we plot the time variation of clustering coefficient and relative size of largest cluster for the non-cumulative networks $S(nc)$.

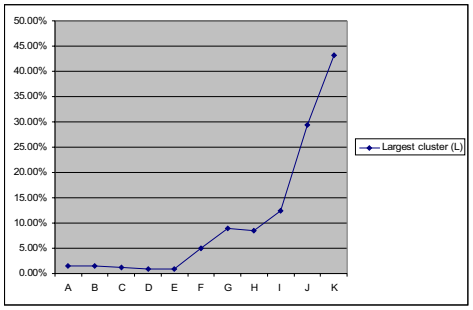


**Figure 9: Size of Largest cluster over Non-Cumulative Time-steps Defined in Table 3**

Comparing these figures with Figure 5 and Figure 6 respectively, we note that the overall trends are the same: steadily increasing clustering coefficient and rapidly increasing relative size of the largest cluster (in the later time-steps). So networks for individual time-steps taken in sequence show a characteristic similar to those of cumulative networks. The maximum clustering coefficient for $S(nc)$ is more than 22% higher (0.71 versus 0.578) than that of $S(c)$. The value of 0.71 is closer to the clustering coefficient value of other scientific collaboration networks as reported in [18]. The largest value of the relative size of the largest cluster is close to one another (42.11% versus 43.25%) for both $S(c)$ and $S(nc)$. This is surprising, as the value of $L$ has been found to be in the range of 80%-90% in collaboration networks of other disciplines [18]. In contrast, it appears that less than half of the authors in software engineering collaboration networks belong to the largest cluster.

The increasing values of $CC$ and $L$ reflect on the growing interconnectedness of software engineering collaboration network. The network grows by the publication of new papers. New papers necessarily mean the addition of new vertices, but not necessarily the addition of new edges. Hypothetically if all new papers in a time-step are written by single authors, then vertices will grow by the number of new papers, but edges will not grow at all. The growth of the

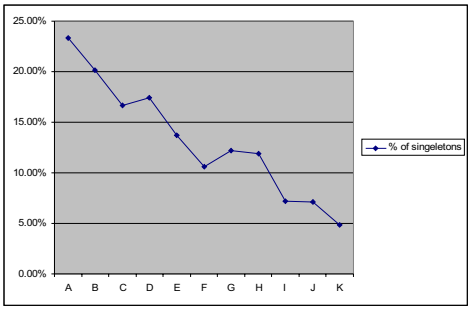


**Figure 10: Percent of Singleton Authors Out of New Authors over Non-Cumulative Time-steps Defined in Table 3**

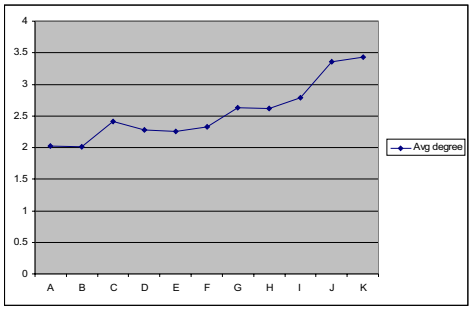


**Figure 11: Average Degree of Non-singleton Authors over Non-Cumulative Time-steps Defined in Table 3**

number of edges is proportional to the square of the number of co-authors; if $n$ is the number of co-authors in a new paper, it adds $n$ vertices and $n*(n-1)/2$ edges to the network. Thus the extent of co-authorship is a defining characteristic of collaboration of a discipline. In mathematics and theoretical sciences the number of joint authors for a paper is usually low, whereas in empirical sciences the authorship traditions are more "generous" – "... it is common, for example, for a researcher to be made a coauthor (sic) of a paper in return for synthesizing reagents used in an experimental procedure" [16]. Where is software engineering research collaboration positioned in the spectrum of co-authorship traditions?

To address this question we extracted the number of new authors for each non-cumulative time-step and the number of "singletons", that is authors with zero collaborators, among those new authors. Figure 10 shows the variation of the percent of singletons out of new authors in each non-cumulative time-step; starting with 23.37% in 1976-1979, it has come down (though non-monotonically) to 4.83% in 2007-2010. Evidently, less single author papers are being written in the later time-steps than in the earlier ones.

When we remove the singletons from the set of new authors in each non-cumulative time-step, we are left with authors who have had at least one collaborator. What is the trend of variation of the average degree of non-singleton authors across the same time-steps? With reference to Figure 11 we note that the average degree of non-singleton new authors increases (though non-monotonically) from 2.02 in the first time-step to 3.42 in the last time-step. Thus there is a net increase of more than one degree across the range of non-cumulative time-steps.

The significant decrease of the percent of new singleton authors and the notable increase in the average degree of new non-singleton authors give strong evidence of increasing collaboration in software engineering research with the progression of time.

Why is there evidence of increasing collaboration in software engineering research? Increasing researcher connectivity, led by the advent of the World Wide Web in the early 1990s is surely influential in facilitating more collaboration. But this is nothing unique to software engineering research; every collaborative enterprise should ideally be positively impacted by this facilitating factor. Is there something unique to software engineering research interests that has encouraged progressive increase of collaboration?

To address these questions we parsed the titles of all the 11,429 papers in our database and tokenized them into bag of words after stop words removal for each non-cumulative time-step. Figure 12 presents a word cloud of the tokenized titles. (The word cloud has been created using Wordle[5]) Expectedly, we see words of high import in software engineering research in commensurate boldness. However, the changing importance of certain words with time is not apparent from the word cloud. When we examine the frequencies of tokens in the bag of words for each time-step, we find that "object" has featured among the top four most frequent words in all five non-cumulative time-steps between 1984 to 2003 (barring common words like "software" etc); "object" has been the most frequent in two of those time-steps, 88-91 and 92-95. This only reflects the widely recognized spurt in the interest in object-orientation from the second half of the 1980s.

Acknowledging the risk of over-generalization, it may be said that a key theme of object-orientation is interaction; objects with streamlined responsibilities collaborate amongst themselves to collectively fulfil a system's functionality. Perhaps the interactive paradigm of object-orientation also influenced research around this paradigm to be more interactive, leading to heightened collaboration amongst software engineering researchers. We recognize this as a conjecture at this point of time; we intend to examine with more rigour in our future work.

In summary, symptoms of increasing collaboration with progression of time is evident in software engineering research.

## 7. OPEN ISSUES AND FUTURE WORK

Some of the factors which can pose as threats to the validity of our conclusions are:

- Non inclusion of publication venues which may not be fully dedicated to software engineering, but none the

[5]http://www.wordle.net/


**Figure 12: Word Cloud of Tokens from Paper Titles over Non-Cumulative Time-steps Defined in Table 3**

less publish significant software engineering research results.

- Ambiguity in the names of authors (confusing one individual into several or several individuals) if they happen to be authors with high degree of collaboration.
- Taking the time-steps as approximately around three years, which may mask micro-trends within the time-steps.
- Not considering the weights of collaborations in terms of the number of times two authors may have collaborated.

In our future work, we plan to address these threats to validity. The empirical data presented in this paper has led us to a number of observations on how collaboration characteristics of software engineering research vary over time. What we observed are symptoms of the underlying dynamics of software engineering research collaboration. Based on these, we plan to develop a model that would offer theoretical underpinnings to the empirical results presented in this paper.

## 8. CONCLUSIONS

In this paper we have extracted software engineering research collaboration networks from publicly available bibliographic information on co-authorship of papers in ten venues over 11 time-steps from 1976-2010. By exploring the time variations of various networks parameters we have made the following observations: power law characteristics in degree distribution only manifest in the last two cumulative time-steps and preferential attachment can not be said to predominantly influence this behaviour; software engineering research collaboration networks are different from corresponding networks in mathematics and neuro-science in terms of the time variation of average separation and clustering coefficient, and somewhat similar in terms of the variation of the relative size of the largest cluster and the average degree; software engineering research collaboration networks appear to manifest small world characteristics in the decade of 2000-2010 with a degree of separation around 7; the cumulative network of software engineering research collaboration at 2010 manifests symptoms of social networks to moderate extent in terms of non-trivial clustering and assortativity. There is also evidence of increasing collaboration in software engineering research with the progression of time. These characteristics as revealed in our study point to the distinct nature of research collaboration in software engineering vis-a-vis other disciplines. The empirical results presented in this paper establish a foundation for developing a theoretical model for the evolution of software engineering.

## 9. REFERENCES

[1] Ieee software's 25th-anniversary top picks. *IEEE Software*, 26:9–11, 2009.

[2] R. Alberich, J. Miro-Julia, and F. Rossello. Marvel universe looks almost like a real social network. *cond-mat/0202174*, Feb. 2002.

[3] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *cond-mat/0106096*, June 2001. Reviews of Modern Physics 74, 47 (2002).

[4] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *cond-mat/0104162*, Apr. 2001. Physica A 311, (3-4) (2002), pp. 590-614.

[5] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *cond-mat/0104162*, Apr. 2001. Physica A 311, (3-4) (2002), pp. 590-614.

[6] C. Bird, E. Barr, A. Nash, P. Devanbu, V. Filkov, and Z. Su. Structure and dynamics of research collaboration in computer science. In *SDM*, pages 826–837. SIAM, 2009.

[7] B. Boehm. Anchoring the software process. *IEEE Softw.*, 13(4):73–82, 1996.

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *0706.1062*, June 2007. SIAM Review 51, 661-703 (2009).

[9] L. L. Constantine and L. A. D. Lockwood. Usage-Centered engineering for web applications. *IEEE Softw.*, 19(2):42–50, 2002.

[10] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *cond-mat/0106144*, June 2001. Adv. Phys. 51, 1079 (2002).

[11] A. Hassan and R. Holt. The small world of software reverse engineering. In *Reverse Engineering, 2004. Proceedings. 11th Working Conference on*, pages 278–283, 2004.

[12] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: characterizing and modeling network evolution. In *Proceedings of the international conference on Web search and web data mining*, pages 107–116, Palo Alto, California, USA, 2008. ACM.

[13] P. Kruchten. The 4+1 view model of architecture. *IEEE Softw.*, 12(6):42–50, 1995.

[14] J. Leskovec, J. Kleinberg, and C. 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*, pages 177–187, Chicago, Illinois, USA, 2005. ACM.

[15] B. Meyer. Reusability: the case for object-oriented design. In *Software reuse: emerging technology*, pages 201–215. IEEE Computer Society Press, 1988.

[16] M. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):016131, 2001.

[17] M. Newman. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, 2001.

[18] M. E. J. Newman. The structure of scientific collaboration networks. *cond-mat/0007214*, July 2000. Proc. Natl. Acad. Sci. USA 98, 404-409 (2001).

[19] M. E. J. Newman. The structure and function of complex networks. *cond-mat/0303516*, Mar. 2003. SIAM Review 45, 167-256 (2003).

[20] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *cond-mat/0412004*, Nov. 2004. Contemporary Physics 46, 323-351 (2005).

[21] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *cond-mat/0305612*, May 2003. Phys. Rev. E 68, 036122 (2003).

[22] D. Watts. Networks, dynamics, and the Small-World phenomenon. *The American Journal of Sociology*, 105(2):527, 493, 1999.

[23] D. J. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 1st edition, Feb. 2003.

[24] D. J. Watts and S. H. Strogatz. Collective dynamics of /'small-world/' networks. *Nature*, 393(6684):440–442, June 1998.