The Small World of Software Reverse Engineering

Ahmed E. Hassan and Richard C. Holt
Software Architecture Group (SWAG)
School of Computer Science
University of Waterloo
Waterloo, Canada
{aeehassa,holt}@plg.uwaterloo.ca

ABSTRACT
Research in maintenance and reengineering has flourished and evolved into a central part of software engineering research worldwide. In this paper, we have a look at this research community through the publications of its members in several international conferences. We analyze our results using various graph and text mining techniques. We contrast our findings to other research communities.

1 INTRODUCTION

Publications in a research community give a picture of the progress of collaboration and emergence of topics in an active research field. The authorship details on each publications represent a social network of collaboration between researchers in the community. One would expect a high degree of collaboration in an academic community, in contrast to a lower degree of collaboration in commercial communities. Furthermore, the titles of these publications permit us to track the appearance of new research topics and areas of interest in the community and the computer industry as a whole. Such topics of interest may in some cases explain changes in the collaboration structure of a community and may shed some light on its evolution.

DBLP [2] tracks the publication history for several conferences in the areas of reengineering, maintenance and software engineering in general. The data is available as an XML file. It records for each year the title of the publications and the authors of these publications. The availability of this data has encouraged us to study the structure of collaboration and the evolution of areas of interest in the reengineering community as part of the larger community of software engineering research.

We examine the publications produced by researchers in the areas of software maintenance and reengineering in several international conferences. We develop a social collaboration network for the community using the co-authorship data for these conferences. In particular, we build a graph that has as nodes each author who published in these conferences. An edge exists between two nodes if they co-authored a paper together. Such a graph is shown in Figure 1. The figure was built using the co-authorship data for the Working Conference on Reverse Engineering (WCRE) from 1993 through 2002 inclusive. The size of each node in the graph is proportional to the number of publications by the node (author). Also weights were added to the edges to indicate the number of papers that two authors have written together. The layout of the figure was generated using a force based algorithm [5]. Thus author nodes in the layout are closer to other author nodes with which they interact the most. In the upper left corner of Figure 1, we show an overview of the full graph of co-authorship. The graph contains a single large connected component along with many smaller components that vary in size. In the main pane of the figure, we zoom to the center of the largest connected component and mark the author's names for some of the large nodes in it[1].

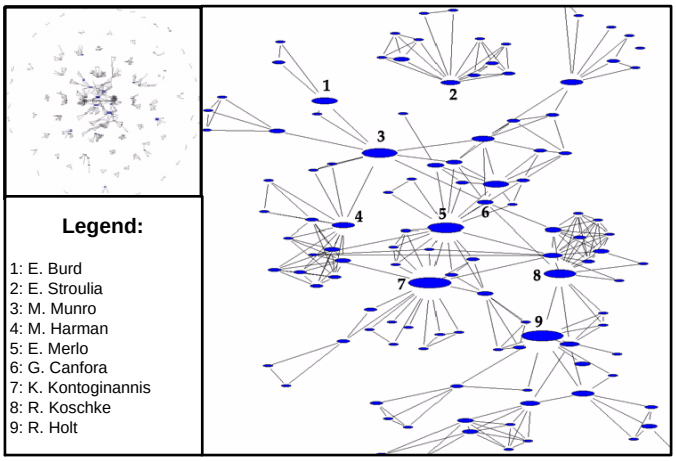

Figure 1: Co-Authorship Graph for WCRE (1993-2002)

Figure 2 shows the variation of the size of the largest components in the WCRE co-authorship graph from 1993 to 2002. For 2002, the largest component contains around 29% of all authors that ever published a paper in WCRE. The next largest component has always been considerably smaller - for 2002, it contains around 3% of all the authors. The years 1999 and 2000 saw large increases in the size of the largest component. This is due to the fact that a number of authors

[1] A more interactive view of the figure is available online as an SVG or GDL at:
http://plg.uwaterloo.ca/~aeehassa/home/pubs/wcreCoauthorsGraph.html.
The graph has recently been chosen as the graph of the month and is accessible at http://www.aisee.com/graph_of_the_month/wcre.html.

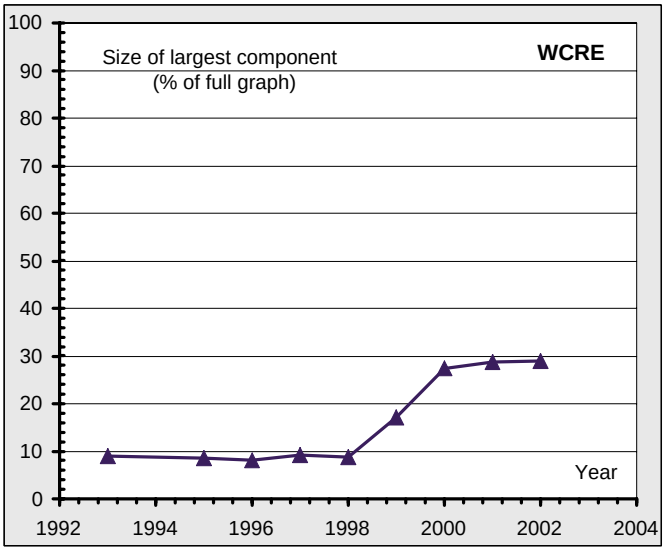

Figure 2: Changes to the Size of the Largest Component in the Co-Authorship Graph for WCRE (1993-2002)

in previously medium sized components in the collaboration graph have started collaborating with authors in the largest component of the graph.

2 SMALL WORLD SOCIAL COLLABORATION NETWORKS

We were interested in the reasons behind the variation of the size of the largest component and the evolution of collaboration in the WCRE community, so we decided to investigate if the WCRE co-authorship graph has the properties of a "small world graph".

The concept of small world graphs has been studied by Stanley Milgram. In the 1960s, Milgram [7] studied the hypothesis that members of any large social network are connected to each other through short chains of intermediate acquaintances. Milgram ran an experiment to measure the average number of intermediate acquaintances needed to deliver a letter addressed to a stockbroker in Pittsburgh. Letters were given to people in rural Nebraska. Each person was asked to hand the letter over to someone with whom they were on a first-name basis and whom they believed can eventually deliver the letter. Milgram measured the average number of links in the chain of people between Nebraska and Pittsburgh as six, hence the term "six degrees of separation". Recent work by Watts [12] and Kleinberg [6] has provided a formal presentation of this phenomena using graph theory concepts. The term "small world graph" has been used to describe networks that exhibit such behavior – large networks with rather short paths connecting each of their members. Collaboration networks that show such characteristics are probably good indicators of the ease of communication of discoveries and knowledge between the members of the network/community, due to the small number of people needed to transfer such information throughout the graph. In [11], Watts and Strogatz studied small world graphs. They defined two graph metrics to categorize a graph: the characteristic path length of the graph and its clustering coefficient.

The characteristic path length represents the average shortest distance from any node in the graph to any other node in the largest connected component of the graph. It measures on average how many individuals a researcher has to go through to reach another researcher in the community.

Characteristic Path Length L:
Given a graph with n nodes, let $D(i,j)$ be the length of the shortest path between the nodes i and j, then the *characteristic path length*, is $D(i,j)$ averaged over all $\binom{n}{2}$ pairs of nodes.

The clustering coefficient measures how collaborative are the co-authors of an author on average. For a given node, it is the ratio of the actual number of edges among the neighbors of that node to the maximum number of possible edges between these neighboring nodes. The clustering coefficient for a graph is defined as the average of the clustering coefficient of all its nodes.

Clustering Coefficient C:
The *clustering coefficient* for a node which has K neighbors (edges) is defined as $\frac{e}{\frac{(K*(K-1))}{2}}$, where e is the number of edges among neighbors of that node.

Watts and Strogatz define a small world graph as graph that:

- has a clustering coefficient that is much higher than a similarly sized random graph,
- yet it has a slightly longer characteristic path length than a similarly sized random graph.

A Small World Graph:
Watts [12] gives another definition of a small world graph. A small world graph is a graph with:

- a clustering coefficient, C, contained in the interval $[0.5, 0.8]$, and
- a characteristic path length, L, approximately equal to $\frac{ln(n)}{ln(k)}$, where n is the number of nodes and k is the average degree of a node in the graph.

At the end of 2002, WCRE has 267 papers written by 376 authors, with an average of 2.46 authors per paper, 1.76 papers per author, and 3.1 collaborators per author. Analysis of the largest connected component of the graph which has 29% of the authors reveals that the largest distance between two authors (the diameter of the graph) is 10. There are $n = 109$ authors in that component and they have 524 edges between them, so the average degree of a node is $k = \frac{524}{109} = 4.81$. Based on data up to year 2002, the WCRE co-authorship graph is a small world graph with a high clus-

tering coefficient of 0.76 and a characteristic path length of $4.3 (\approx \frac{ln(524)}{ln(4.81)} = 4.0)$.

Author Name	Centrality Score
Gerardo Canfora	2.76
Rainer Koschke	2.88
Ettore Merlo	2.94
Andre De Lucia	3.1
Richard C. Holt	3.2

Table 1: Listing of Authors with Smallest Centrality Scores

In comparison to a randomly generated graph with the same number of nodes and edges, the WCRE co-authorship graph has the properties of a small world graph with highly clustered neighborhoods which are connected using short paths.

The most central author in the component is currently Gerardo Canfora as he has the smallest centrality score. He has 2.76 as an average distance to other authors. This position has been held by Canfora since 2001, previously it was held by Ettore Merlo. Table 1 shows a listing of the five authors with the smallest centrality scores as of 2002.

3 TRACKING RESEARCH TRENDS AND DIRECTIONS

During our analysis of the WCRE co-authorship graph we noticed two interesting events:

- The first is the rapid growth in the years 1999 and 2000 of the largest component in the co-authorship graph as shown in Figure 2.
- The second is a dip in the length of the characteristic path in 2001 (see Figure 3).

We decided to use the titles of the papers in the proceedings to search for events that may explain some of these findings and to detect trends in the direction and focus of publications in WCRE as a proxy for the research interests in the community.

Using the paper titles stored in the DBLP XML file, we removed stopwords (such as “the” and “of”) and used a stemmer to derive the root of each word in the title of a paper (for example, truncating “extracting” to “extract”) . We then tracked the most popular terms used in the titles throughout time. The terms “reverse”,“engineering”, “program”, and “system” have been popular throughout most of the years. This is not a surprising finding given the focus of WCRE. To discover new trends and events we filtered this data using the following heuristic: we only report terms that have not been popular in the last two years. Thus if we detect that a term is popular we first check the previous two years to determine if it was popular back then as well. If it was popular back then, we do not report it as an emerging popular term. Table 2

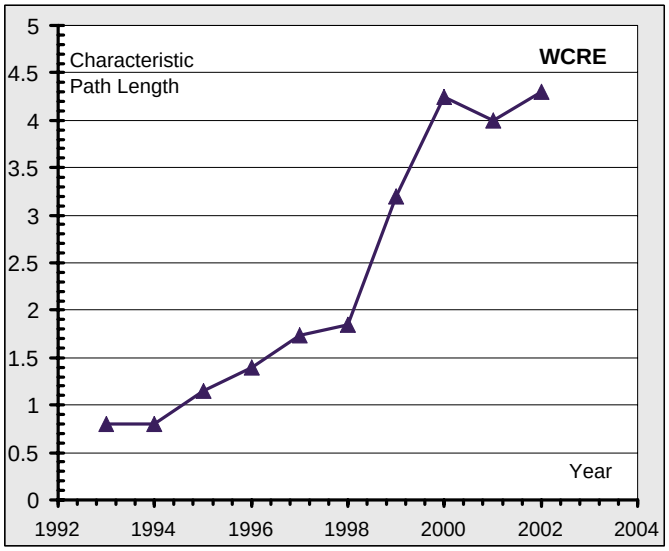

Figure 3: Variation in the Characteristic Path Over Time

lists the emerging popular terms for each year since 1993 till 2002. The usage of the two year popularity window causes the reappearance of terms if they become popular again such as the term “legacy” which is reported in 1995 and in 2000.

Year	Emerging Terms
1993	reverse, engineering, program, system, approach
1995	data, software, procedure, tool, reengineering, legacy, object, oriented, design, specification
1996	model, code, experiment, environment, architecture
1997	approach, understanding, databases
1998	system, requirements
1999	method, java, model, develop
2000	exchange, data, format, legacy, XFIG, web
2001	database, slicing, comprehension, decompliation
2002	pattern, analysis, source, extraction, static, XML

Table 2: Emerging Popular Terms

In the year 2000 there was a focus on the ideas of standardized schemas and data formats to facilitate exchange among researchers in the community. We believe that this focus on methods to facilitate exchange has caused more researchers to collaborate on proposals and experiments. We hypothesize that this collaborative initiative is the main cause for the decline of the characteristic path length in the following year 2001. A closer look at the publications in the 2001 papers reveals that by removing a paper [4] by Ferenc, Sim, Holt, Koschke, and Gyimothy titled “Towards a Standard Schema for C/C++”, the characteristic path for the WCRE graph rises in 2001 to 4.32, instead of its actual value of 3.94. We believe that these findings demonstrate the benefits of standard-

ized exchange formats on increasing collaboration in a research community and moving the community forward by focusing on more advanced and complex research questions and challenges.

4 BIGGER SMALL WORLDS IN SOFTWARE ENGINEERING

In the previous sections, we focused our analysis solely on the WCRE conference. The WCRE conference is one of several conferences devoted to the field of reengineering and software maintenance. Therefore there may be many collaborations and properties of the collaboration in the general community that are not visible through the WCRE publication history. Furthermore, it is interesting to contrast our findings for the reengineering and maintenance community to the wider software engineering community and to other research communities in computer science and other fields.

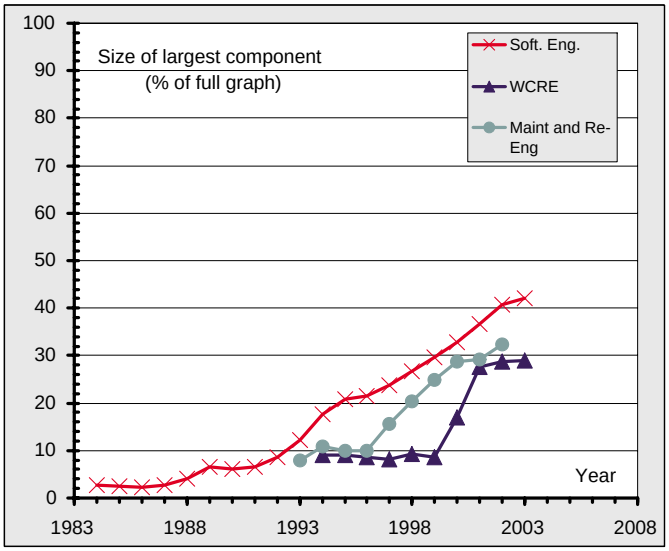

Figure 4: Size of the Largest Component Over Time for the Three Communities

The IWPC, CSMR, and ICSM are another set of conferences in the area of maintenance and reengineering (MR). These three conferences along with WCRE give a more complete and accurate representation of the larger community. As for the software engineering (SE) community, we examined the DBLP database and chose 21 conferences (ADSD, APSEC, ASE, KBSE, CAiSE, COMPSAC, COOTS, ECOOP, ESEC, FSE, ICSE, ICSR, METRICS, OOPSLA, PASTE, RE, SEKE and the four conferences in the MR category) as representatives of that large community. We compared some of our findings for the WCRE community to these larger communities. Unfortunately, the DBLP data is not available for each year of all the conferences we chose. Consequently, some of the results may not be as accurate as we would hope for but we believe the results still give a good indication of the state of collaboration in these communities.

Figure 4 shows the evolution of the size of the largest component over time for all three communities. All three communities start with a slow growth (almost constant) then begin a rather rapid growth. We hypothesize that the slow growth is due to the time it takes researchers to become acquainted with others and to start collaborating. Furthermore, the SE community has taken a considerable time for its growth to reach critical mass and grow fast. This may be attributed to the lack of information in the DBLP data for some conferences in these earlier years, the limited number of conferences in the area back then, and the large size and scope of the SE community in contrast to the other two smaller communities. It is interesting to note that the MR and SE communities have had a large period of growth since 1996, this coincides with the growth of popularity and accessability of the Internet and Electronic email which represent great collaboration mediums for authors worldwide. Another explanation of this rapid growth may be the effect of the growth of the MR community on the SE community. We believe that this is not the case due to the small size of the MR community relative to the SE community and because of the fact that the SE community was already in a growth phase even before the MR community started growing.

We were interested in finding the researcher who has the most central location in the co-authorship graph in these communities, so we reran our scripts on these other communities. Figure 5 summarizes our findings. Authors' names are displayed in the year they first held the most central location in the graph. The names are not repeated until another author has this central location. Ettore Merlo and Gerardo Canfora are still the most central researchers at the MR community level but this is not the case for the SE community. At the SE community level, currently, Premkumar T. Devanbu is the most central researcher. Grady Booch is the author which has been at the center of the graph the most times (5 non-consecutive years). Devanbu has been at the center of the graph for (4 non-consecutive years).

Table 3 summarizes various properties of the WCRE, RE, and SE graph. It also shows results from other publications which analyzed other research communities:

- The SIGMOD - Special Interest Group on Management Of Data - results have been published in [8].
- The SIGIR - Special Interest Group on Information Retrieval - analysis was published in [10].
- The ACM - Association for Computing Machinery - and GD - Graph Drawing - community results have been published in [3].
- The results for MedLine and SPIRES have been published in [9]. MedLine is a database of medical article citations, produced by the National Library of Medicine. SPIRES contains high-energy physics related articles, including journal papers, preprints, e-prints, technical reports, conference papers and theses.

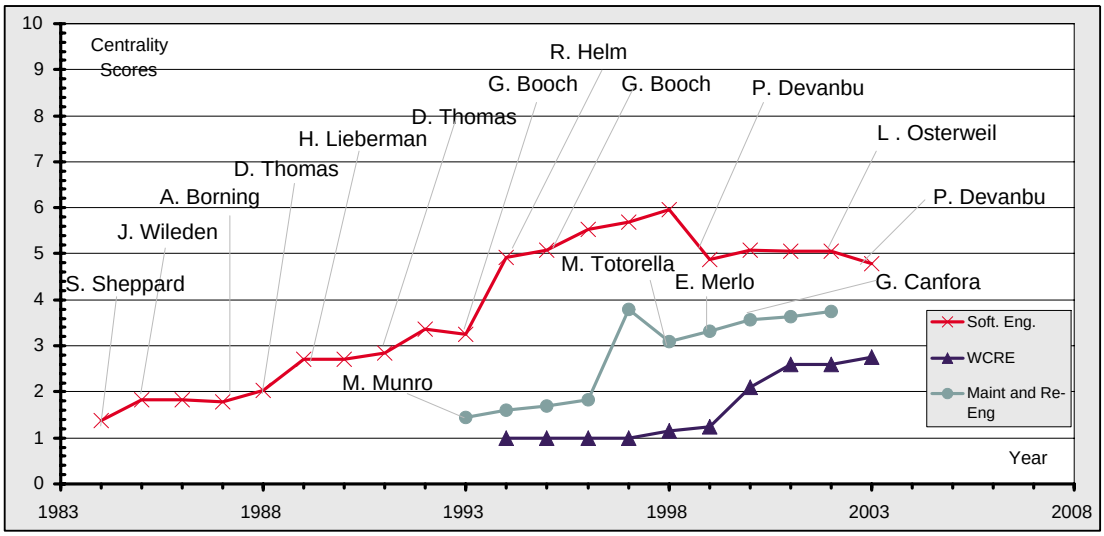

Figure 5: Centrality Score Over Time for the Three Communities

	WCRE	MR	SE	GD	SIGMOD	ACM	MedLine	SPIRES
Total papers	267	1145	6908	413	N/A	51503	2163923	66651
Total authors	376	1434	9343	502	2394	81279	1520251	56627
Authors per paper	2.47	2.45	2.41	2.54	N/A	2.32	3.74	8.96
Papers per author	1.76	1.96	1.78	2.09	N/A	1.80	6.4	11.6
Collaborators per author	3.08	3.29	3.52	3.74	N/A	3.36	18.1	173
Perc. of largest comp.	29	32	42	49	59	49	92.6	88.7
Clustering Coefficient	0.76	0.68	0.67	0.60	0.69	0.62	0.066	0.73
Avg. dist. (Char. path)	4.26	6.23	7.7	4.33	5.65	9.26	4.6	4.0
Maximum distance	10	16	20	10	15	30	24	19

Table 3: Summary of the Co-Authorship Graphs for Several Communities

It is interesting to note that the authors per paper, papers per author, and collaborators per author values for WCRE, RE, and SE are similar to the ACM values. The GD values are a good representation of another small conference in computer science like WCRE as they both have similarly a small number of authors and papers. The medical (MedLine) and physics (SPIRES) values vary widely in comparison to the other values - they show a large variation in the publication and collaboration patterns in these fields compared to the computer science field. For example, we see a considerably larger number of authors collaborating on a paper (9 authors on average for SPIRES). We also see a very large number of collaborators for each author (173 authors on average for SPIRES and 18.1 authors for MedLine). These rather larger numbers are apparently due to the different culture of assigning co-authors on a publication. Moreover these numbers reduce the value of simple analysis of the co-authorship graph for these communities, instead more elaborate technique are needed [9].

5 CONCLUSION

In this meta paper, we studied the publications of researchers in the WCRE conference, the maintenance and reengineering (MR) community and the software engineering (SE) community using data provided by the DBLP. We built co-authorship graphs for each community and showed that these graphs have properties of small world graph which indicate the ease of information and knowledge flow in these communities. Furthermore, we studied the emergence of trends and directions of research using the titles of publications. We then correlated these trends to events in the evolution of the co-authorship graph.

ACKNOWLEDGEMENTS

The idea of studying the publication history for WCRE was conceived during a session at WCRE 2003 in Victoria. It would not have been possible without the wireless access provided by the conference. The authors would like to acknowledge the suggestions by several of the WCRE atten-

dees who stopped by our demo on the last day. Data used for our analysis is due to the DBLP project which has been organized by Michael Levy. The aiSee [1] graph software was used to layout the graph shown in this paper.

REFERENCES

[1] aiSee Graph Layout Software Page. Available online at <http://www.aisee.com>

[2] DBLP Bibliography Home Page. Available online at <http://www.informatik.uni-trier.de/~ley/db/>

[3] C. Erten, P. Harding, S. Kobourov, K. Wampler, and G. Yee. GraphAEL: Graph Animations with Evolving Layouts. In *International Symposium on Graph Drawing (GD 2003)*, pages 49–58, Perugia, Italy, Sept. 2003.

[4] R. Ferenc, S. E. Sim, R. C. Holt, R. Koschke, and T. Gyimothy. Towards a Standard Schema for C/C++. In *Working Conference on Reverse Engineering*, pages 49–58, Stuttgart,Germany, 2001.

[5] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software Practice and Experience*, 21(11):1129–1164, 1991.

[6] J. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[7] S. Milgram. The Small World Problem. *Psychology Today*, 61(1), 1967.

[8] M. A. Nascimento, J. Sander, and J. Pound. Analysis of SIGMOD's co-authorship graph. *ACM SIGMOD Record*, 32(3):8–10, 2003.

[9] M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Physics Review*, E64, 2001.

[10] A. F. Smeaton, G. Keogh, C. Gurrin, K. McDonald, and T. Sodring. Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century. *SIGIR Forum*, 36(2):39–43, 2002.

[11] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 363:202–204, 1998.

[12] D. J. Watts. *Six Degrees: The Science of a Connected Age*. W.W. Norton and Company, 2003.