Genet Program Evolvable Mach (2007) 8:97–103
DOI 10.1007/s10710-006-9018-2

LETTER

The structure of the genetic programming collaboration network

Marco Tomassini · Leslie Luthi · Mario Giacobini · William B. Langdon

Received: 9 June 2006 / Revised: 26 June 2006 /
Published online: 19 October 2006

Abstract The genetic programming bibliography aims to be the most complete reference of papers on genetic programming. In addition to locating publications, it contains coauthor and coeditor relationships which have not previously been studied. These reveal some similarities and differences between our field and collaborative social networks in other scientific fields.

Keywords Genetic programming · Scientific collaboration · Social networks · Communities

1 Introduction

This paper is not about genetic programming (GP) results or methodology; it is about us, the actors of the genetic programming community. It is about the technical and social bonds between us. Naturally these are difficult to quantify, so we will do as other fields have done, and use joint published work to stand for social bonds. While this is obviously far from perfect, we have an immediate, (almost) comprehensive and quantitative data source in the genetic programming bibliography.

The genetic programming bibliography, created and maintained by one of us (WBL) and by S. Gustafson,[1] is a database that contains most of the papers published in the GP field since its inception. As such, it is a rich source of data that implicitly describes many aspects of the structure of the GP community. Searching the bibliography and looking at the images[2] immediately provides a lot of useful information about the field and its actors. However, a deeper analysis of the data, that goes beyond the mere pictorial aspect, provides a much

M. Tomassini (✉) · L. Luthi · M. Giacobini
Information Systems Department, University of Lausanne, Switzerland
e-mail: Marco.Tomassini@unil.ch

W. B. Langdon
Department of Computer Science, University of Essex, UK

[1] http://www.cs.bham.ac.uk/~wbl/biblio/

[2] http://www.cs.bham.ac.uk/~wbl/biblio/gp-coauthors/

more complete view. Indeed, in a very real sense, the coauthorship data is a social network because the act of collaborating in a research study usually requires personal acquaintance among the coworkers. Social network analysis, although it is an old discipline, has recently received new impetus and tools from the field of complex networks [1]. Analysis of social systems as complex networks has produced an amazing amount of important results in just a few years. An excellent survey of the main methods and results can be found in [2].

Only a few basic concepts are needed to understand the specific case of the GP collaboration network. An actor, or node, is an active GP researcher. That is, someone who has at least one entry in the bibliography. There is a connection between two actors if they have coauthored one or more papers. This gives rise to a network that can be described with the tools of graph theory. Similar studies have been recently performed on several other collaboration networks in disciplines such as physics, mathematics and medicine [3–5].

2 Fundamental features of the collaboration network

The GP coauthorship network comprises a total of $N = 2684$ connected nodes (authors) and a total of 11,005 edges (collaborations). There are 358 isolated vertices. These represent authors who have not collaborated with others to the extent of coauthoring a paper. (Isolated vertices are ignored in our graph statistics.) Due to the youth of GP, the graph is small compared to some previously studied collaboration networks which contain tens of thousands or even hundreds of thousands of authors [4]. This has both advantages and drawbacks. The main disadvantage is that, like any statistical study, more data allows deeper and more meaningful inferences to be drawn. In particular, studies of the form of the distributions (such as whether they follow exponential or power laws) requires a lot of data. The advantages are that the graph almost fully represents the state of the whole GP community (as of April 2006). This allows reliable characterisation of collaboration in the community. Also, the problems of homonymy, outliers, and different name spelling that plague the larger data sets, are very unlikely and easy to spot in our data. We included conference proceedings and edited books, as well as standard conference and journal articles. While edited books and some proceedings (e.g. EuroGP proceedings) usually do reflect personal acquaintance, some may not. Below, we present and discuss some basic measures that characterise our collaboration network.

2.1 Number of papers per author

The average number of papers per author is 3.5. The graph of the distribution $P(k)$ of the number k of papers per author is rather bumpy, especially in the tail of the distribution. Instead, we present in Fig. 1 the graph of the cumulative distribution $P(k \geq n)$ which is smoother and allows the same inferences to be made. The curve is rather well fitted by a straight line, and thus the distribution follows a power-law, with a calculated exponent of 2.5. This is in line with previous results on other coauthorship networks such as Medline and NCSTRL (see Table 1) which have exponents 2.8 and 3.4, respectively [4].

2.2 Number of collaborators per author and number of authors per paper

The average number of collaborators per author is 8.2. However, this includes a single paper with 108 coauthors in a nuclear physics journal. If we leave out this an anomalous entry, then the average drops to 4.06. This is close to the values for disciplines that follow similar collaboration patterns such as computer science (NCSTRL) or Mathematics (see Table 1).

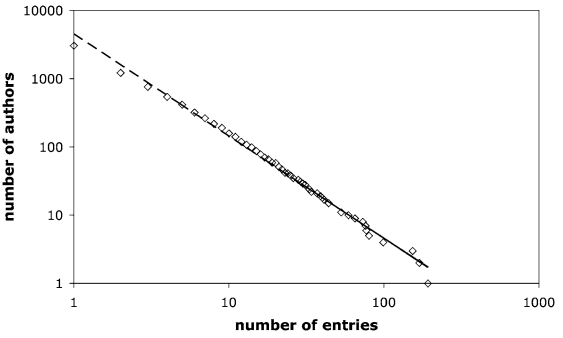
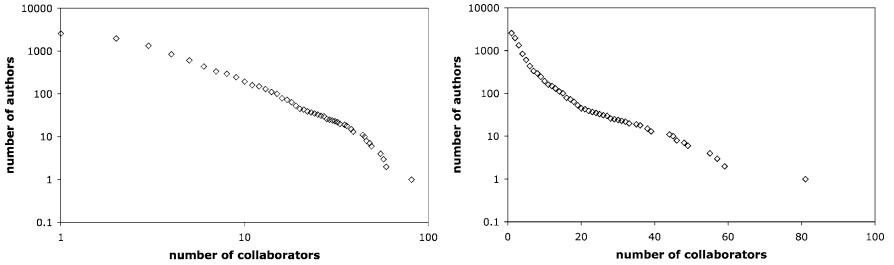


Fig. 1 Cumulative distribution of the number of papers per author. Log-log scale. The straight line is a mean-square fit



(a) (b)

Fig. 2 Cumulative distribution of the number of authors with a given number of collaborators on log-log scales (a). The same distribution on linear-log scales (b)

Figure 2 shows the cumulative distribution of the number of collaborators. From Fig. 2(a) one sees that the distribution is not a pure power-law, otherwise the points would approximately lie on a straight line, but rather there is a power-law regime in the first and middle parts, while the tail decays exponentially. This is confirmed by Fig. 2(b), where the tail can be approximately fitted by a straight line due to the use of a linear-log scale. That is, the whole network cannot be fitted by a power-law. This is quite common. In fact, many social networks do not follow a power-law degree distribution [6].

Table 1 Basic statistics for some scientific collaboration networks.

	GP	SPIRES	Medline	Mathematics	NCSTRL
Total number of papers	4139	66,652	2,163,923	1,600,000	13,169
Total number of authors	2684	56,627	1,520,251	253,339	11,994
Average papers per author	3.5	11.6	6.4	7	2.55
Average authors per paper	2.25	8.96	3.754	1.5	2.22
Average collaborators per author	4.06	173	18.1	2.94	3.59
Size of the giant component (%)	35.0	88.7	92.6	82.0	57.2
Clustering coefficient	0.669	0.726	0.066	0.15	0.496
Average path length	4.72	4.0	4.6	7.73	9.7

GP is our network. SPIRES is a data set of papers in high-energy physics. Medline is a database of articles on biomedical research. Mathematics comprises articles from *Mathematical Reviews*. NCSTRL is a database of preprints in computer science. Details on those databases can be found in [3, 4].

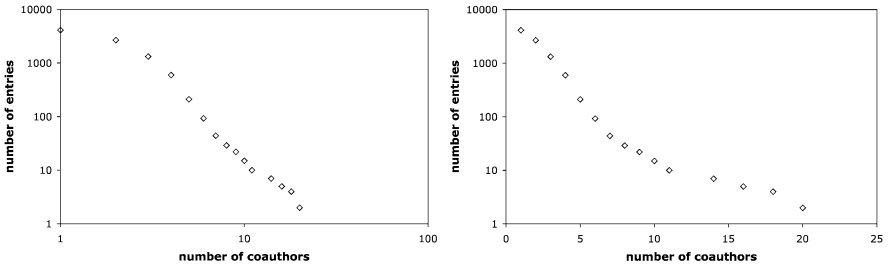


Fig. 3 Cumulative distribution of the number of papers with a given number of coauthors on log-log scales (a). The same distribution on linear-log scales (b)

Figure 3 shows the cumulative distribution of the number of papers written by a given number of coauthors. Here the distribution also has a tail that is fatter than that of a normal or exponential distribution, but it does not appear to follow a power-law.

2.3 Distribution of connected components

In the theory of Poisson random graphs, above a critical value of average degree $\langle k \rangle = 1$, one observes the sudden appearance of a so-called *giant component* of size $O(N)$. That is most vertices belong to that component and the other components are smaller with an exponential size distribution [2]. Although collaboration graphs differ from standard random graphs, they often also show this phenomenon. In our case, the size of the giant component is 942 authors, which represents 35% of the total graph size. For the actors belonging to the giant component the average number of collaborators per author is 5.54. The cumulative size distribution of the connected components is depicted in Fig. 4; from these data, the probability density function is well approximated by a power law exponent of -2.63. The existence of a big connected component has a social meaning. It means 35% of GP researchers are members of a single community, since those researchers are either directly connected via a collaboration or they are close to each other in a way that will be made clear in Section 3.

2.4 Clustering coefficient

The clustering coefficient of a graph indicates to what extent your friends are themselves friends. It can be seen as the fraction of connected triples in the graph that are triangles

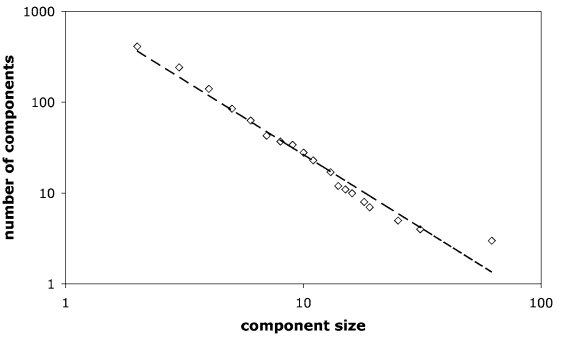


Fig. 4 Cumulative distribution of the number of components in the graph as a function of their size. Log-log scale. The straight line is a mean-square fit

[2]. Most biological and social networks (and some technological networks) have a much larger clustering coefficient than would be expected of a random graph with the same number of vertices and edges. Social networks are particularly clustered. For example, the clustering coefficient is 0.669 for the GP collaboration graph. (We would expect 0.0031 for the corresponding random graph). In terms of scientific collaborations, a high clustering coefficient means that people tend to collaborate in groups of three or more. This agrees with what we know of the GP field. It may mean that two researchers that collaborate independently with a third one may, in time, become acquainted and so collaborate themselves. Alternatively it might be due to collaborators coming from the same institution.

We summarise in Table 1 most of the results of this section, together with those for some other collaboration networks. Most GP statistics are similar to those of the larger databases. However one notable difference is the relative smallness of the largest component. This may be due to the comprehensive nature of the GP bibliography. It captures work done by smaller groups which does not get into major journals, whereas, perhaps, the other databases concentrate upon higher impact outlets where work is heavily cited but at the expense of ignoring less regarded authors. This may artificially inflate the fraction of authors within their giant component. Alternatively it may be due to the youth of the GP field, with many semi-isolated individuals and groups starting research independently. However as time goes by, one should observe small components progressively connecting themselves to the large one. The clustering is rather high, which shows that GP researchers know each other quite well within the large component, and the community is rather homogeneous. In contrast in biology and medicine or mathematics, where scientist from different sub-disciplines seldom collaborate, the clustering coefficient is much lower. Note also the high number of authors per paper, and especially the strikingly high number of collaborators per author in the nuclear physics community (SPIRES). Clearly, nobody can maintain an average of 173 scientific partners on a first-hand acquaintance basis and thus this figure is not socially meaningful.

3 Distances, centrality, and assortativity

A social network can be characterised by a number of measures that give an idea of "how far" people are from each other, or how "central" they are with respect to the whole community. These measures are well known in social network analysis. Here we shall concentrate on *mean path length* and on *betweenness*.

3.1 Mean path length

In many real networks, such as the Internet, the World Wide Web, social networks, and many others, including random graphs, the path length between any two vertices scales as a logarithmic function $O(\log N)$ of the number of vertices N. Such networks, if they also have a high clustering coefficient, are known as *small worlds* networks. Since, even for very large graphs, in contrast to regular lattices, any two nodes in a small world network are only a few steps apart. The mean path length[3] of the giant component of the GP collaboration graph is 4.72 and the longest among all the shortest paths (known as the diameter) is 14. Thus, unsurprisingly, the GP community, but only as far as its "core" component is concerned, is indeed a small world and is characterised by measures that are typical of these kinds of

[3] The mean path length is the average of the shortest paths between all pairs of nodes in the graph.

network (see Table 1). Being a small world means that information may circulate quickly and collaborations are easier to set up, which is advantageous for a research community. The connected components following the largest one are themselves small worlds. We expect over time some of them will merge with the largest component. (For this to happen, only a single new collaboration between two scientists each belonging to one of the components is needed.)

3.2 Betweenness

The betweenness $b(v)$ of a vertex v is the total number of shortest paths between all possible pairs of vertices that pass through this vertex. Nodes that have a high betweenness have more control, i.e. they are more central in the network, in that there is more "traffic" that goes through them. The first five authors (in decreasing order) in terms of centrality in the network are: W. Banzhaf, H. Iba, U.-M. O'Reilly, H. de Garis and W. B. Langdon. People who have a large value of betweenness play the role of intermediaries or "brokers" in a social sense. W. Banzhaf is also the researcher that has the highest number of different collaborators.

3.3 Degree correlations

One interesting aspect of real-life networks is the correlation between the degrees of neighbouring nodes, called *degree assortativity*. Most technological and biological networks have a negative correlation. That is, high-degree vertices are preferentially connected to low-degree vertices, while most measured social networks are assortative. Meaning highly connected nodes tend to be connected with other highly connected nodes [2]. Our collaboration network confirms this general observation with a correlation coefficient of 0.13 for the giant component, and 0.27 for the whole graph (excluding the single physicist's paper). These are close to the coefficients observed for other social networks (specifically 0.127 for Medline and 0.120 for Mathematics [7]).

In the mathematical literature, a popular game is the calculation of the "Erdös number" [3]. P. Erdös was an extremely prolific mathematician who published more than one thousand papers, most of them with other researchers. A mathematician's m Erdös number is the length of the path from m to Erdös in the collaboration graph. Just for fun, we did the same for three well known GP-ers: J. R. Koza, W. B. Langdon, and R. Poli. Our arbitrary criterion of choice of these three names was just that they have the highest number of published GP papers,

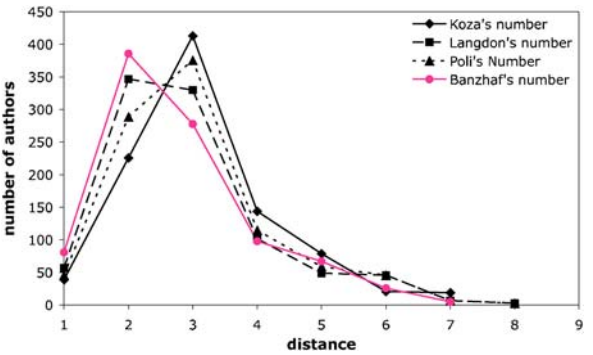


Fig. 5 Distance distributions for some authors in the collaboration graph

in that order. We also added a fourth distribution, corresponding to W. Banzhaf, because he had the highest betweenness. Figure 5 shows the distributions of Koza's, Langdon's, Poli's, and Banzhaf's numbers. We observe that W. Banzhaf has the shortest distances to most other actors, which is a confirmation of him having the highest betweenness. On the other hand, J. R. Koza, who is the person having the highest number of papers, is a bit more eccentric in the network. W. B. Langdon and R. Poli appear to have a rather similar distance distribution ranging between the previous two.

4 Conclusions and outlook

We have characterised the GP coauthor collaboration graph by means of a number of statistical measures. When doing so, several interesting features emerge that are analogous to those observed for other collaboration networks. Our community also has its peculiarities, mostly due to the its youth and to the fact that the discipline is more narrowly focused than, say, mathematics or physics. It should be obvious that the present "hard" approach to social network analysis can only provide some answers but not all of them. Typically, such "emotional" human aspects such as friendship in scientific collaboration or the geographic closeness of two research institutions will be buried in the sea of numbers and will never appear explicitly from such analyses. Nevertheless, we feel that our results are interesting and useful in the way that they characterise our community.

There are several other interesting observations that could be made, for which we do not have space here. One is the study of the sub-communities present in the large component. While the giant component is in itself a community with a clear boundary, one could search it for groups of closely connected researchers. Another aspect that would be revealing is the analysis of the time evolution of the collaboration graph. Perhaps there is not enough data for that study but it could highlight things about the community's evolution. Another interesting point concerns the act of collaboration itself. While an edge between two researchers in the graph means that they have collaborated at least once, it would be more correct to study the weighted graph. That is each edge could be weighted proportionally to the number of corresponding joint papers thus somehow expressing the "strength" of the collaboration. We are investigating these extensions.

References

1. D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.
2. M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.
3. J. W. Grossman, "The evolution of the mathematical research collaboration graph," *Congressus Numerantium*, vol. 158, pp. 201–212, 2002.
4. M. E. J. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," *Phys. Rev E*, vol. 64, pp. 016131, 2001.
5. M. E. J. Newman, "Scientific collaboration networks. II. shortest paths, weighted networks, and centrality," *Phys. Rev E*, vol. 64, pp. 016132, 2001.
6. L. A. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the National Academy of Sciences USA*, vol. 97, p. 11149–11152, 2000.
7. M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, p. 208701, 11 November, 2002.