Available online at www.sciencedirect.com
ScienceDirect
Physica A 385 (2007) 750–764
PHYSICA A
www.elsevier.com/locate/physa

Empirical analysis of the evolution of a scientific collaboration network

Marco Tomassini*, Leslie Luthi

Information Systems Department, University of Lausanne, Switzerland

Received 2 May 2007; received in revised form 19 June 2007
Available online 25 July 2007

Abstract

We present an analysis of the temporal evolution of a scientific coauthorship network, the genetic programming network. We find evidence that the network grows according to preferential attachment, with a slightly sublinear rate. We empirically find how a giant component forms and develops, and we characterize the network by several other time-varying quantities: the mean degree, the clustering coefficient, the average path length, and the degree distribution. We find that the first three statistics increase over time in the growing network; the degree distribution tends to stabilize toward an exponentially truncated power-law. We finally suggest an effective network interpretation that takes into account the aging of collaboration relationships.

PACS: 89.65.−s; 89.75.−k; 89.75.Fb

Keywords: Network evolution; Preferential attachment; Scientific collaboration; Social networks

1. Introduction and previous work

In recent years, thanks to the increasing availability of machine-readable data, many large networks have been empirically analyzed in detail in several disciplines including communications and information networks, biological, social, and technological networks. In many cases it has been found that these networks have small diameters and high clustering. In other words, any node is relatively close to any other node, and the local connection structure will not be random, but rather shaped by social or other forces [1,2]. The origins and the evolution of such networks have been the object of intensive research and there exist several models that can be used to explain the experimentally observed data. However, while models abound by now and the theory is rather well developed, the analysis has concentrated on static networks, i.e. networks that are, or are considered to be, in a steady state. However, to test models on network formation and evolution, one needs to study actual networks for which time-resolved data do exist, and these are more difficult to find. Some works have dealt with this kind of problem in the last few years. Noteworthy among them are the following

*Corresponding author.
E-mail address: marco.tomassini@unil.ch (M. Tomassini).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.07.028

investigations: Newman's study on scientific collaboration networks [3], Barabási et al. [4] and Jeong et al. [5] investigations on the growth of coauthorship, citation, Internet, and actor collaboration networks. Other interesting studies have targeted the web [6,7], potential energy surfaces [8], social interactions represented by e-mail exchanges [9] and the Internet encyclopedia Wikipedia [10]. Some of these networks are technological, such as the Internet, while others have a more social flavor. Citation networks, the web, and Wikipedia cannot be considered social networks in the proper sense, although they do support communication and information transmission in social contexts. On the other hand, scientific collaboration networks, e-mail networks, and the actor network usually imply underlying social ties with their associated costs and thus they are considered at least good proxis for social networks. Some networks are directed (the web, Wikipedia, citation networks) while others are undirected (coauthorship networks, actors, Internet, potential energy surfaces).

Most of the graphs in the above mentioned works as well as many others have a measured degree distribution $P(k)$[1] that is either a power-law $P(k) \propto k^{-\gamma}$, or a power-law with an exponential cutoff. This means that there is a non-negligible probability in these graphs that some vertices have high connectivity. Several growing models have been proposed to account for these topological features, most of them being based on some form of *preferential attachment*. Preferential attachment means that when new nodes join the graph linking to existing nodes j, the rate $\Delta k_j/\Delta t$ is an increasing function of the degree k_j of j. Some models assume this function to be linear [11], while in other cases it has been assumed to depend on a different power of k_j [7,12]. In general, we have that the probability $\Pi(k_j)$ with which an edge belonging to a new node connects to an existing node j of degree k_j will be

$$\Pi(k_j) = \frac{k_j^\alpha}{\sum_i k_i^\alpha},$$

where the sum is over all vertices i already present in the graph. Thus the rate of increase of node degree will be: $\Delta k/\Delta t \propto k^\alpha$.

For $\alpha = 1$ the rate is linear and the model reduces to the familiar Barabási–Albert construction [11] which yields a power-law degree distribution $P(k)$. For $\alpha < 1$ the preferential attachment is sublinear and $P(k)$ is a stretched exponential [12]. For $\alpha > 1$ a single node gets almost all the edges, with the rest having an exponential distribution of the degrees. Therefore, to know which kind of preferential attachment, if any, is at work in a particular growing network, one needs to study empirically networks for which the time at which new nodes entered the graph and new edges formed is known.

The following conclusions have been reached in Refs. [3,5,10]. First of all, preferential attachment appears to be present in all the studied networks, although some follow an almost linear growth (Internet, citations and Wikipedia [5,10]), while others appear to grow at a sublinear rate ($\alpha < 1$) and give rise to stretched exponential distributions. The latter case is present in some coauthorship networks and actor collaborations [3,5]. Since the coauthorship network studied in Ref. [4] shows a power-law degree distribution, the apparent contradiction has been explained in Ref. [5] by the presence of another linear preferential attachment mechanism involving the appearance of new internal edges among existing nodes.

From these results, it appears that social networks differ somewhat in their growing properties from information networks such as the web, Wikipedia and citation networks. This may be due to the fact that making new links in the latter is essentially free of cost, while there is some cost in the former related to the necessity of becoming acquainted with some other individual in the network before an association is possible.

In the present study we investigate the evolution of another scientific collaboration network, the genetic programming (GP) coauthorship network. The GP bibliography, created and maintained by W.B. Langdon and by S. Gustafson,[2] is a database that contains almost all the papers published in the GP field since its inception around 1986. This database is smaller than those that have been studied previously [4,13–15], but it has the advantage of being essentially complete. Moreover, different authors with the same initial and surnames and the same author spelled differently are rare occurrences, while this is a source of some error in the larger databases. We are thus in a position to study the growth of the collaboration network from the very

[1]There are two distinct degree distributions in directed networks: one for the incoming links $P(k_{in})$ and another for the outgoing links $P(k_{out})$.

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