Clustering and preferential attachment in growing networks

M. E. J. Newman

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

Abstract

We study empirically the time evolution of scientific collaboration networks in physics and biology. In these networks, two scientists are considered connected if they have coauthored one or more papers together. We show that the probability of scientists collaborating increases with the number of other collaborators they have in common, and that the probability of a particular scientist acquiring new collaborators increases with the number of his or her past collaborators. These results provide experimental evidence in favor of previously conjectured mechanisms for clustering and power-law degree distributions in networks.

I. INTRODUCTION

Many systems take the form of networks—sets of nodes, or vertices, joined together by links, or edges. The Internet, the power grid, social networks, food webs, distribution networks, and metabolic networks are commonly cited examples. Investigations of networks within the physics community fall loosely into two categories: (1) studies of static network structure [1–6] and dynamical processes taking place on fixed networks [7–9]; (2) studies of the dynamics of networks themselves—how and why their topology changes over time [1,2,10–12]. It is this second category that we address here, focusing on two properties which have received a large amount of attention in the literature—clustering and preferential attachment.

Sociologists have long known that social networks—networks of personal acquaintances, for example—display a high degree of transitivity, meaning that there is a heightened probability of two people being acquainted if they have one or more other acquaintances in common. In the physics literature this phenomenon is called "clustering." Watts and Strogatz [1] measured clustering in a number of real-world networks, including both social and physical networks, by calculating a clustering coefficient, equal to the probability that two vertices that are both neighbors of the same third vertex will be neighbors of one another. They found that in many networks the clustering coefficient is much higher than its expected baseline value, which is set by comparison with a random graph.

It has also been pointed out by a number of authors [3–5,13], particularly in studies of the Internet and the World-Wide Web, that real-world networks have highly skewed distributions of vertex degree. (The degree of a vertex is the number of other vertices to which it is connected.) In many cases, the degree distribution is found to follow a power law, a particularly telling functional form which often signifies an underlying process worthy of study.

Explanations have been put forward for both of these observations. In the case of clustering, it is conjectured that pairs of individuals with a common acquaintance (or several) are

likely to become acquainted themselves through introduction by their mutual friend(s) [2]. In the case of degree distributions, it is conjectured that, for a variety of reasons, vertices accumulate new edges in proportion to the number they have already, leading to a multiplicative process which is known to give power-law distributions [10–12]. This process is often called "preferential attachment." While both of these explanations are, in some contexts at least, perfectly plausible, there has been little if any empirical evidence in their favor—a glaring problem for two conjectures which have formed the foundation of a substantial body of research. The principal reason for this has been the lack of good time-resolved data on how networks grow.

In order to test a conjecture such as "people with many common friends are more likely to become acquainted than those with few or none," one needs to watch a network grow and see if the process described by the conjecture does indeed happen with significantly heightened frequency. Although data on the structure of networks are quite plentiful, data on how they grow have proved harder to come by. Recently, however, the author conducted some empirical studies of collaboration networks of scientists: networks in which pairs of scientists are linked together if they have coauthored one or more papers [14–16]. These collaboration networks are true social networks, since two scientists who have coauthored a paper will normally be acquainted with one another. (There are occasional exceptions—see Ref. [15].) They are also well documented, since there exist extensive machine-readable bibliographies of the scientific literature. What's more, as Barabási and co-workers have recently pointed out [17], these networks have excellent time resolution as well, because each paper comes with a publication or receipt date. As we now show, this allows us to test directly the clustering and preferential attachment conjectures.

In this study we look at collaboration networks derived from two bibliographic sources:

1. The Los Alamos E-print Archive, a database of preprints in physics, self-submitted by their authors;

2. Medline, a database of published papers in biology and medicine, whose entries are

3

professionally maintained by the National Institutes of Health.

While neither of these databases records the exact publication date of the papers they contain, both include a record of the sequence in which papers were added to the database. This is enough for our purposes: all that we need for our calculations is the order of the collaborations undertaken by each author in the database, and the order of the papers is a reasonable proxy for this—probably not correct in every case, but assumed to be correct in most. Two other databases that we studied previously [14] do not contain enough information to establish order of collaborations, recording publication or database entry of papers to the nearest year only. This creates ambiguity since many authors produce more than one paper a year, and so we did not use these databases for the current study.

Authors are identified by their full surname and all initials. As discussed previously [14,15], an author who gives their name differently on different papers may be confused for two people by this measure, while two people with identical surnames and initials may be confused for one. The error in the number of vertices in the network as a result of these problems was found to be on the order of 5%.

We study a six-year interval of time for both databases. (For the Los Alamos Archive we use 1995 to 2000 inclusive, for Medline 1994 to 1999.) Over this period the Los Alamos Archive records 58 342 distinct names, and Medline 1 648 660. In each of the calculations presented here, we use the first five of the six years to construct a collaboration network, and then examine how that network further changes in the remaining one year. Our assumption is that any scientist who is currently active will produce at least one paper during the initial five year period, as will any currently active collaboration between a pair of scientists, so that the network we have at the end of that period will be essentially complete. New vertices added in the sixth year represent, it is assumed, new individuals entering the field, and new edges represent genuine new collaborations. Of course there are some exceptions, such as established scientists who for one reason or another fail to publish anything for five years and then produce a paper in the sixth, and these will be misrepresented in our calculations.

4

We assume these are a small fraction of the total. There will also be some scientists who leave the field during the six years, to go into different fields or professions, or because they retire. We make no attempt to guess which individuals leave in this way: everyone whose name appears even once is considered a member of the network for the entire period of study thereafter. This will introduce some error into our calculations. However, it is straightforward to convince oneself that the correlations we are looking for in the present study will only be weakened by this error, not strengthened, so there is no danger of false positive results.

II. CLUSTERING

Let us consider first the question of clustering in the network. We already know that the clustering coefficient is high in our collaboration networks—0.45 for the Los Alamos Archive and 0.088 for Medline over a five-year period [14]. The calculation presented here improves on these results in two ways. First, the simple clustering coefficient includes contributions from collaborations between two authors which *preceded* their collaborations with any mutual acquaintances. By using time-resolved data we can exclude these collaborations from our measure of clustering. Second, we can determine whether the probability of two individuals collaborating increases as the number m of their previous mutual acquaintances goes up. If this is the case, then it suggests that the standard explanation of clustering—introduction of future collaborators to one another by common previous acquaintances—is correct, the probability of such an introduction presumably increasing with m. Other explanations, such as the institutional explanation proposed in Ref. [15], would be harder to justify.

Measuring the probability of collaboration between authors as a function of their number of mutual acquaintances is complicated by the fact that both the size of the graph and the numbers of mutual acquaintances themselves are changing over time. We consider the probability $P_m(t)$ that the two scientists connected by a link added at time t have m mutual acquaintances. (Time is somewhat arbitrary here. It can be real time, but it can also be

5

any other function which increases monotonically as papers are added to the database—only the order of the papers matters, not their precise timing. The links created by a paper with three or more authors are all considered to be added at the same instant.) We have

$$P_m(t) = \frac{n_m(t)}{\frac{1}{2}N(t)[N(t)-1]}R_m, \tag{1}$$

where $n_m(t)$ is the number of pairs with m mutual acquaintances immediately *before* the addition of the paper at time t, $N(t)$ is the current number of authors in the network, and R_m is the relative probability of collaboration between the two scientists connected by this link, i.e., the ratio between the actual probability of their collaborating and the probability of their collaborating in a network in which presence of mutual acquaintances makes no difference. We assume that the probability that two scientists with a given value of m collaborate at a particular time does not depend on the number of other scientists with that value of m, or on the size of the database in which the paper they write is archived, and hence that R_m is independent of t [18]. This makes it a suitable quantity to measure to test our clustering hypothesis. In a world with no clustering, we would have $R_m = 1$ for all m; in a world in which clustering arises through introductions, as above, it should increase with increasing m.

To measure R_m, one simply constructs a histogram of the value of m for each link added to the graph in which each sample is weighted by a factor of $\frac{1}{2}N(t)[N(t)-1]/n_m(t)$. In Fig. 1 we do this for the network of the Los Alamos Archive. As discussed above, we evaluate R_m for the last of our six years only, the previous five being used to establish the initial network for the calculation. As the figure shows, R_m does indeed increase with m, and is much greater than 1 for all $m > 0$. A pair of scientists who have five mutual previous collaborators, for instance, are about twice as likely to collaborate as a pair with only two, and about 200 times a likely as a pair with none. R_m increases roughly linearly for small m, perhaps indicating that each common collaborator of a pair of scientists is equally likely to introduce them. The curve appears to flatten off for higher m, although the data become poor for $m \gtrsim 8$, since the number of pairs of authors with this many common collaborators
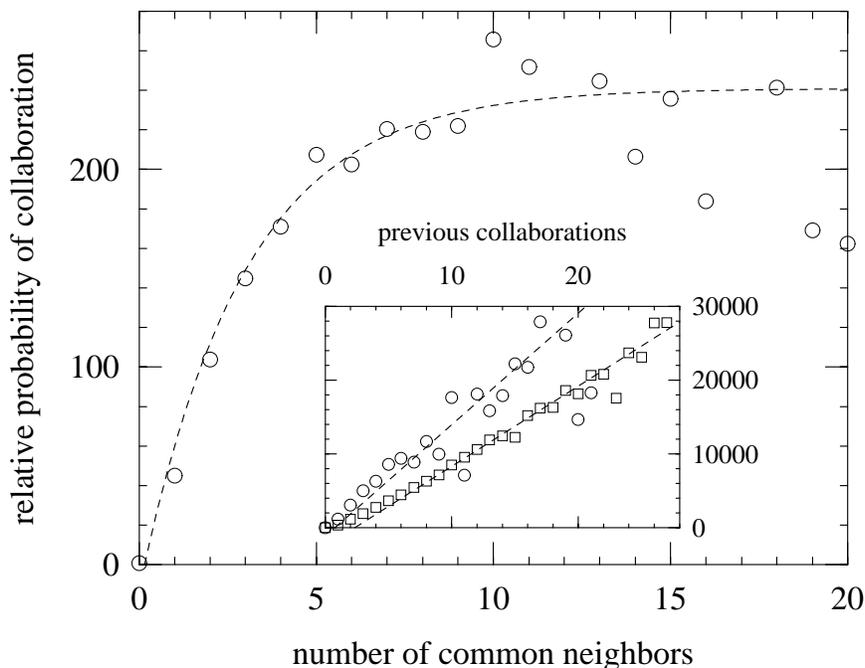
6

FIGURE 1 Probability of collaboration between scientists in the Los Alamos Archive as a function of their number of mutual previous collaborators. The dotted line is the best fit of the form (2). Inset: the relative probability of collaboration as a function of number of previous collaborations of the same scientists, for the Los Alamos Archive (circles) and Medline (squares). The dotted lines are the best straight-line fits to the data. The data for Medline have been divided by a factor of 50 vertically to fit the axes.

who have not already collaborated themselves is very small.

As well as supporting the standard explanation of clustering in social networks, our data for R_m might prove useful for modeling purposes. For example, in both the α-model of Watts [2] and the network growth model of Jin and Newman [19], a particular form is assumed for the probability of individuals becoming acquainted, as a function of their number of mutual friends. Fig. 1 provides a rough empirical guide for what that functional form should be. In the figure we give a fit to the data of the form

$$R_m = A - B\mathrm{e}^{m/m_0}, \tag{2}$$

where A, B, and m_0 are constants. This form appears to fit reasonably well and might be

7

suitable for use in the models.

III. REPEAT COLLABORATIONS

In the calculation described above, we included only newly appearing edges in the network. Repeat collaborations between authors who had collaborated before were excluded; we assume that such collaborations are more likely to be a result of previous acquaintance than the result of network structure. This however raises another interesting question: does probability of collaboration also increase with the number of times one has collaborated before? The answer is yes, as shown in the inset of Fig. 1, which measures the relative probability R_n (defined similarly to R_m above) of two coauthors collaborating if they have collaborated n times previously within the period covered by our study. If collaboration probability were independent of previous collaboration, we would have $R_n = 1$ for all n, but as the figure shows, R_n increases roughly linearly with n, indicating that number of past collaborations is a good indicator of the probability of future collaboration. However, one must bear in mind that this calculation may be influenced by varying frequencies of collaboration: regular collaborators who publish often will have more publications in the database as well as greater likelihood of publishing again in the last of our six years, producing a correlation just as seen in the figure. To eliminate this effect one would have to look at data for a longer period of time and compare collaborators with similar numbers of publications but different publication rates. Unfortunately, this is not practical with the data available to us at present.

IV. PREFERENTIAL ATTACHMENT

We can also use our data to test for preferential attachment in the collaboration network. Barabási *et al.* [17] have previously looked for preferential attachment in two collaboration networks derived from data for publications in mathematics and neuroscience. Papers in their databases were dated only to the nearest year, making the order in which collaborations

8

occur uncertain, as discussed above. To get around this, they restricted themselves to measuring the number of new papers each author in the network published in a single year, as a function of number of previous papers. This should be an increasing function if there is preferential attachment, or constant otherwise. Their results show a clear increase and hence favor preferential attachment.

Using our data we can measure preferential attachment (if any) in our networks directly by a method similar to the one we used to measure clustering above. We define a relative probability R_k that a link added at time t connects to the vertex representing a particular scientist who has collaborated previously with k others. By analogy with Eq. (1), the probability $P_k(t)$ that this link connects to any vertex with degree k is $P_k(t) = R_k n_k(t)/N(t)$, where $n_k(t)$ is the number of vertices with degree k immediately before addition of this paper. Then R_k can be estimated by making a histogram of the degrees k of the vertices to which each link is added in which each sample is weighted by a factor of $N(t)/n_k(t)$. If there is no preferential attachment, R_k should equal 1 for all k. If there is preferential attachment, it should be an increasing function of k, and the widely held belief is that it should in fact increase linearly with k. If it increases linearly, then the resulting degree distribution of the network will be a power law [10–12].

In Fig. 2 and its inset we show empirical results for R_k for the databases studied here. As the figure shows, the relative probability is in both cases close to linear for the initial part of the curve, but falls off once k becomes large. This is understandable: no one can collaborate with an infinite number of people in a finite period of time, so at some point $R(k)$ must start to decrease. This point appears to be around 150 collaborators in physics and 600 in biomedicine. Interestingly, these figures coincide roughly with the points at which the observed degree distribution in these networks starts to deviate from the power-law form [15], lending support to the theory that preferential attachment is the origin of the power law.

Our results differ somewhat from those of Barabási *et al.* [17], who found preferential attachment for their networks, but did not find linear behavior. In the language used here,
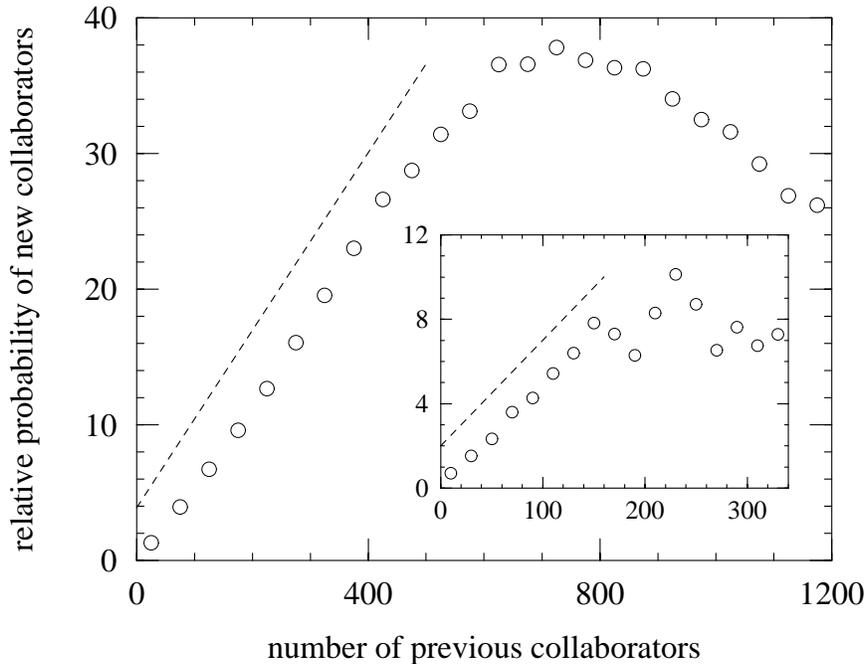
FIGURE 2   The relative probability that a new edge in the collaboration network will connect to a vertex of given degree. The main figure shows data from the Medline database, the inset data from the Los Alamos E-print Archive.

their finding was that $R(k) \sim k^{\nu}$, with $\nu \simeq 0.8$. Such a form does not fit our data well. A power-law fit to the increasing part of $R(k)$ for our data gives $\nu = 1.04 \pm 0.04$ for Medline and $\nu = 0.89 \pm 0.98$ for the Los Alamos Archive, both compatible with the conjecture of linear preferential attachment. In practice however, this difference may have little effect. As Krapivsky *et al.* [11] have shown, sub-linear preferential attachment gives rise to a stretched exponential cutoff in the resulting degree distribution, but we already have a similar cutoff in our distribution as a result of the deviation of $R(k)$ from linear behavior for large enough k.

V. CONCLUSIONS

To conclude, we have measured the probability of collaboration between scientists in two collaboration networks as a function of their number of mutual acquaintances in the network,

their number of previous collaborations, and their number of previous collaborators. We find that the probability of collaboration is strongly positively correlated with each of these, and for the latter two that the relationship is close to linear over a large part of its range. These results lend strong support to previously conjectured, but untested, theories about the way in which networks grow.

ACKNOWLEDGEMENTS

REFERENCES

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

[2] D. J. Watts, *Small Worlds,* Princeton University Press, Princeton (1999).

[3] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *Nature* **401**, 130–131 (1999).

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks* **33**, 309–320 (2000).

[5] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

[6] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, `cond-mat/0007235`.

[7] S. H. Strogatz, "Exploring complex networks," *Nature* **410**, 268–276 (2001).

[8] R. Monasson, "Diffusion, localization and dispersion relations on small-world lattices," *Eur. Phys. J. B* **12**, 555–567 (1999).

[9] C. Moore and M. E. J. Newman, "Epidemics and percolation in small-world networks," *Phys. Rev. E* **61**, 5678–5682 (2000).

[10] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science* **286**, 509–512 (1999).

[11] P. L. Krapivsky, S. Redner, and F. Leyvraz, "Connectivity of growing random networks," *Phys. Rev. Lett.* **85**, 4629–4632 (2000).

[12] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Phys. Rev. Lett.* **85**, 4633–4636 (2000).

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet

topology," *Comp. Comm. Rev.* **29**, 251–262 (1999).

[14] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci. USA* **98**, 404–409 (2001).

[15] M. E. J. Newman, "A study of scientific collaboration networks: I. Network construction and fundamental results," *Phys. Rev. E,* in press.

[16] M. E. J. Newman, "A study of scientific collaboration networks: II. Shortest paths, weighted networks, and centrality," *Phys. Rev. E,* in press.

[17] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the science collaboration network," preprint, Notre Dame University (2001).

[18] This assumption is debatable. Certainly scientific collaboration patterns have changed over the last century, so that R_m probably does vary with t on long time-scales. Whether this variation is visible over the course of a single year is unclear.

[19] E. Jin and M. E. J. Newman, "The structure of growing social networks," in preparation.