Games and Economic Behavior 79 (2013) 106–131

Contents lists available at SciVerse ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Sources of variation in social networks

Enghin Atalay

Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, United States

ARTICLE INFO

Article history:
Received 31 December 2011
Available online 9 February 2013

JEL classification:
D85

Keywords:
Degree distribution
Heterogeneous fitness
Network formation
Social networks

ABSTRACT

What explains the large variation in the number of contacts (degree) that different participants of social networks have: age, randomness, or some unobservable fitness measure? To answer this question, I extend the model presented in Jackson and Rogers (2007) to allow individuals to vary in their ability to attract contacts. I estimate the parameters of the extended model, using a social network of citations among high-energy physics papers, and find that the extended Jackson–Rogers model can parsimoniously fit the degree distribution of each age cohort. Moreover, both the length of time spent in the network and the unobservable fitness measure are important in explaining the observed variation in participants' degrees.

1. Introduction

Why do some participants of social networks have so many contacts, while most others have so few? How important are age and randomness in explaining the variation in the number of contacts (i.e., the *degree*) that participants have? What is the underlying process that produces the degree distributions that are repeatedly observed in studies of social networks?

I answer these questions by extending the framework introduced in Jackson and Rogers (2007). In that paper, the authors construct a simple model that fits degree distributions observed in different social networks. In their model, nodes enter the network one at a time and form links with existing nodes in a two-step process. First, in *random* meetings, the entrant uniformly samples from the population of incumbents. Second, in *network-based* meetings, the entrant samples from the contacts of the incumbents that it met. In each of these meetings, the entrant forms a directed link to the sampled incumbent. When network-based meetings are more prevalent, incumbents with many contacts are relatively more likely to gain additional contacts. At one extreme, when all links are formed via networking, nodes' degrees are Pareto distributed. At the other extreme, when all links are formed randomly, the degree distribution is that of an exponential random variable.[1]

In Sections 3.1 and 3.2, I review the arguments that are used to solve for the exact joint distribution of participants' ages and degrees. According to the baseline model, participants are only distinguished by their age (when they entered the network) and chance (whether they happened to be contacted by entrants more or less often than expected). As I show in Section 3.2, the role of chance in the model is limited: almost all of the variation in the number of contacts is due to differences in participants' dates of entry into the social network. I illustrate in Section 2 and Appendix A that this fact conflicts with what is actually observed in social networks. There is wide variation in the number of contacts of individuals within any age group. Therefore, there must exist some other factor, independent of age or chance, that contributes to the observed variation in nodes' degrees.

E-mail address: atalay@uchicago.edu.

[1] The model introduced in Jackson and Rogers (2007) closely resembles those presented in earlier articles, including Dorogovtsev et al. (2000), Pennock et al. (2002), and Vázquez (2003).

0899-8256/$ – see front matter
http://dx.doi.org/10.1016/j.geb.2013.01.009

I extend the baseline model by allowing nodes to differ in the rate at which they can expect to gain additional links. In Section 3.3, I define the *fitness* of a node as the probability that each of its meetings will generate a link. I cannot observe nodes' fitness measures. However, using the variation of degrees across nodes of a particular age, I can identify the variation of fitness. With more variability in fitness, there is more variability in the degree distribution of nodes of a particular age.

In Section 4, I use a dataset of citations among high-energy physics papers to show that heterogeneous node fitness is necessary to fit the observed within-cohort degree distributions. I estimate—via maximum likelihood—the network formation model, both with and without heterogeneous fitness. When nodes are assumed to have identical fitness, the model struggles to fit the relatively weak correlation between age and degree. By contrast, when nodes' fitness levels are drawn from a parametric distribution (that is the same for all age cohorts), I am able to parsimoniously fit the weak correlation between age and degree.[2] Maximum likelihood estimates indicate that a large fraction of the variation in degree is due to variation in fitness. For example, for a median-aged paper, the expected number of citations increases from 1.4 to 20.7 as fitness increases from the 25th to the 75th percentile. For a median-fitness paper, the expected number of citations increases from 2.7 to 13.6 as age increases from the 25th to the 75th percentile.

In Section 4.3, I show that, for most age–fitness combinations, a marginal increase in age has a smaller effect—compared to a marginal increase in fitness—on a node's expected number of contacts. For all but the youngest and lowest-fitness nodes, differences in the rate at which contacts are formed are due mainly to heterogeneity in fitness, rather than heterogeneity in the length of time spent in the network. Randomness in the link formation process plays a tertiary role in explaining the variation in nodes' degrees.

Apart from the shape of the degree distribution, the model introduced in Jackson and Rogers (2007) generates predictions on other features observed in social networks. First, their model predicts that the degrees of two linked nodes are positively correlated. Second, the probability that two nodes are linked is larger conditional on the two nodes having a common contact. In Section 5, I argue that introducing heterogeneity in nodes' fitness levels does not qualitatively alter either of these predictions of the Jackson and Rogers (2007) model. However, the theoretical prediction of the probability that two nodes form a link, conditional on the presence of a common contact, can be matched in the data only when the average fitness is substantially larger than what is estimated in Section 4.

Parsimony and tractability are two of the main strengths of the Jackson–Rogers model. With only two parameters, the model captures several features ubiquitous in social networks. The model that I present in this paper retains much of the tractability of the original Jackson–Rogers model.[3] In addition, it is able to capture not only the within-cohort degree distributions, but also the other characteristics of social networks that are studied in Jackson and Rogers (2007).

In many environments, the manner in which agents are linked to one another has important economic consequences. Jackson (2011, p. 512) catalogs a list of examples: Social networks play an important role by "...transmitting information about jobs, new products, technologies, and political opinions. They also serve as channels for informal insurance and risk sharing, and network structure influences patterns of decisions regarding education, career, hobbies, criminal activity, and even participation in micro-finance." The role that social networks play in economic activity makes it important to understand the mechanisms through which social networks form and the reasons why some agents have so many contacts while most others have so few.

1.1. Literature review

Before proceeding to Section 2, I discuss the theoretical literature from which the current paper borrows, and the empirical literature to which the current paper might lend some insights.

The two main theoretical ideas—first, that one can solve for the exact degree distribution using a mass-balance equation, and, second, that one can embed heterogeneous node fitness into a network formation model—are taken from earlier papers. The method of solving for the exact degree distribution is taken directly from Dorogovtsev et al. (2000). The recognition that one can embed heterogeneous node fitness into a network formation model is due to Bianconi and Barabási (2001) (in the context of a pure preferential attachment model) and Caldarelli et al. (2002) (in the context of a variant of the Erdös and Rényi, 1960 network formation model).

The primary contribution of the current paper is to apply the above-mentioned theoretical arguments to show that, for a social network of citations among high-energy physics papers, both age and fitness are important for explaining the observed variation of nodes' degrees. Related to this finding, allowing for heterogeneous node fitness results in a less prominent estimated role of network-based meetings versus random meetings. These observations have implications for the interpretation of Jackson and Rogers (2007) and its successors.

In an application of Jackson and Rogers (2007), Bramoullé et al. (2012) introduce a social network with different *types* of individuals and assume random meetings are more likely to occur between individuals of the same type. With this assumption, Bramoullé et al. generate clear, testable predictions on the relationship between an individual's degree and the

[2] These results are not specific to the network of citations among physics articles. In Appendix A, I show that the same patterns hold for two additional social networks, a network of citations among patents and a network of buyer–supplier relationships among publicly traded firms.

[3] Of course, allowing fitness to vary arbitrarily across nodes would allow one to explain 100% of the variation in nodes' degrees. The point of this paper is that, with only two extra parameters (which govern the first two moments of the distribution of nodes' fitness levels), the extended Jackson–Rogers model can better fit the degree distribution of each age cohort.