Social Networks 45 (2016) 78–88

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Alters as species: Predicting personal network size from contact diaries

Tso-Jung Yen[a,*], Yang-chih Fu[b], Jing-Shiang Hwang[a]

[a] Institute of Statistical Science, Academia Sinica, Taiwan
[b] Institute of Sociology, Academia Sinica, Taiwan

ARTICLE INFO

Article history:
Available online 7 January 2016

Keywords:
Personal networks
Ego and alters
Contact diaries
Alter accumulation curves
Mixture of Dirichlet processes

ABSTRACT

Like wildlife species in an ecological system, members within a personal network (or alters) constantly shift and often remain hard to count. Previous studies often estimated the size of such personal networks using information given by a focal person (or ego), who names a list of friends and acquaintances, or someone known or related, that meet certain specified criteria. In a search for alternative methods, we estimate the number of alters using contact diaries that help reveal active and comprehensive interactions, which enable us to predict personal network size from a longitudinal perspective. By exploring contact frequencies between ego and alters, we propose a modeling approach based on species accumulation curves from ecology. Under this approach, the contact frequency between ego and alter often turns out to be a mixture of binomial distributions, and the number of alters with whom ego may make contact in the future is assumed to follow a specified discrete distribution. We estimate the model with the Bayesian nonparametric method, in which the distribution of contact probabilities is assumed to be a mixture of Dirichlet processes. We then demonstrate this approach with a data set containing 48 contact diaries collected over three months and discuss how such an ecological analogy may enrich social network studies.

1. Introduction

As important as estimating the population size in an ecosystem, estimating the size of a personal network and how it changes over time has been a fundamental issue in social network studies (Killworth et al., 1990; Rogerson, 1997; McCarty et al., 2001; Feld et al., 2007; Roberts et al., 2009). Without knowing personal network size, much work about network structure and dynamics would remain tentative. Due to the nature of interpersonal networking and the lack of clear and universal operational definitions, however, it has been particularly challenging to estimate or predict network size at the personal level (Freeman, 2004; Kadushin, 2012). Facing such a challenge, can social network studies benefit or gain insights from an ecological analogy?

Members of a personal network (hereafter "alters") resemble species in an ecosystem, most of which are hidden and unobservable. Because people get connected in many possible ways, network members take turn in appearing within the sight of the network's focal person (hereafter "ego"), rarely showing up at the same time in an identical spot. Like an emerging or phasing-out species, moreover, alters are constantly being added or dropped as well. As biologists find it difficult to determine the population of interest, therefore, network researchers also often remain puzzled about how large a personal network can be.

To estimate the number of species and the size of a population in the wild, biologists and ecologists have applied various capture–recapture methods, along with inferential statistical techniques (Colwell et al., 2012). To estimate the size of a personal network, scholars also have employed different tools to help generate part of the network. In large-scale surveys, each of name generator, position generator, and resource generator aims to help compare the range of personal networks by eliciting a small number of network members (Wellman, 1979; Burt, 1984; Marsden, 1987; Lin, 2001; Lin et al., 2001; van der Gaag and Snijders, 2004). Experimental designs have also helped researchers to estimate personal networks by an array of techniques such as small-world, reverse small-world, first-name or last-name, and rare-subpopulations (Milgram, 1967; McCarty et al., 1997; Killworth and Bernard, 1978; Bernard et al., 1990; Zheng et al., 2006; McCormick et al., 2010). A series of studies, in particular, estimated personal network size

* Corresponding author at: Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan. Tel.: +886 2 981451477; fax: +886 2 27831523.
E-mail address: tjyen@stat.sinica.edu.tw (T.-J. Yen).

http://dx.doi.org/10.1016/j.socnet.2015.12.001
0378-8733/

based on how many people in different subpopulations (such as those who share specific first names, occupations, ethnicities, or who have uncommon characteristics like being held in a prison) that ego knew (Killworth et al., 2006; Zheng et al., 2006). One of the latest efforts (McCormick et al., 2010) extended earlier work by developing a non-random latent mixing model, using a modified scale-up estimator (Killworth et al., 1998) to estimate personal network size.

In this paper we propose an alternative approach, borrowed and revised from capture–recapture methods, for estimating and predicting personal network size. Using a series of contact diaries, our approach differs from others in two major aspects. First, while most network generators rely on ego's recalled or perceived assessment of social networking, we use the records that show active interactions in daily life. In particular, we extract the records from a series of contact diaries and identify those whom have had contact with ego as alters of the network. Second, previous studies focused on data collected from one-shot experiments or cross-sectional questionnaire surveys. Our data, in contrast, were recorded longitudinally (Fu, 2007). Each day some alters were contacted, acquired, or captured, for the first time in the diaries; others had been recorded earlier but were again contacted, acquired, or recaptured, by ego. Such a data structure enables us to draw a curve that tracks how unique alters accumulate over time in any given personal network, thus bringing us to the analogy of species accumulation curves in ecology.

Since the contact diary records active contacts between ego and alters on a daily basis, one can count the contact frequency between ego and the alter in any given period. With such information on contact frequencies, it is possible to estimate the accumulation curve. Then the estimated curve can be used to predict the number of new alters that ego will contact in the future. In other words, an *alter accumulation curve* can be constructed to predict ego's personal network size. Because we use longitudinal data to estimate the model, which in turn is used to predict future events, the nature of the proposed approach is predictive. Such a predictive approach further distinguishes our study from most previous social network studies.

The rest of the paper is organized as follows. Section 2 provides further descriptions on the contact diary data, their advantages in estimating personal network size, and the data set used in our study. Section 3 explains the alter accumulation curve, its mathematical representation, and its estimation based on Bayesian nonparametric statistics. Section 4 first shows some results from an exploratory data analysis and then presents the results of prediction based on the alter accumulation curve and the Bayesian estimation developed in Section 3. Section 5 discusses the strengths and limitations of the data and method we use, addresses future research directions, and highlights potential contributions to social network studies.

2. Contact diaries as a longitudinal data base

As an alternative approach to other methods of network data collection, the contact diary method asks people to record information about those whom they have actually contacted. Whereas other network generators sample a subset of alters from ego's acquaintances, contact diaries sample a subset of alters from those having contact with ego within a given period of time. In other words, contact diaries help yield records for reconstructing segments of social interactions that were active in everyday life (de Sola Pool and Kochen, 1978; Freeman and Thompson, 1989; Lonkila, 1999; Fu et al., 2013; Dávid et al., 2016). By following adequate guidelines about how to record contacts and interactions precisely, diary keepers may be more able to provide data that are less distorted by recalling and selection biases.

Although the diary method is more objective, diary keeping is labor-intensive, demanding, tedious, and thus more difficult to implement. As a consequence, the very few diary studies documented in the social network literature have been limited to small samples of egos. More recent studies have started collecting 24-h (or 48-h) contact diaries from large surveys that help epidemiologists understand how infectious diseases spread (Kretzschmar and Mikolajczyk, 2009; Mossong et al., 2008; Horby et al., 2011; Fu et al., 2012; Read et al., 2012). The limited length of these diaries, however, tends to restrain network researchers from estimating personal network size. When kept long enough, contact diaries contain rich and cumulative information about contact frequencies in a highly structured and longitudinal format. Such a data format in turn facilitates the statistical analyses and further modeling tasks that make the prediction more interpretable and allow us to better estimate personal network size. Therefore, while previous statistical estimations infer from a sample taken along a cross-sectional dimension, our estimation relies on a sample taken along a time frame. As elaborated later, such a longitudinal database is key to the specific method we use to uncover how ego may accumulate and build her personal network over time.

2.1. The data set

We draw upon data with ego–alter contacts from a comprehensive contact diary study (Fu, 2007). A total of 54 persons, who were recruited from a random sample and a convenience sample, participated in the study to keep a diary of active one-on-one contacts between March 1 and May 31, 2004. Participants were instructed to record key features of alter, the ego–alter relationship, and the contact situation as soon as each contact ended, and then finish the remaining details of the diary keeping at the end of the day. Of the 54 sets of three-month (i.e. 92 days) contact diaries, 48 were complete with identifiable alters that allow us to distinguish repeated contacts (contacts with alters recorded earlier) from new contacts (contacts with new alters) from day 1 to day 92 for each ego. We use these 48 sets of contact diaries to estimate 48 personal network sizes, excluding total strangers and those alters who were not identifiable by ego because the lack of such critical information prevents further analysis.

Fig. 1 shows the contact frequencies between alters and ego of the 48 personal networks in the first 10 days, the first 30 days, the first 50 days, and the first 70 days, respectively. In each plot, x axis represents the contact frequency, or the number of days that an alter is contacted in a given observation period, and y axis represents the number of alters corresponding to the contact frequency. These plots show that the contact frequencies range widely from only one day to almost the entire observation period. Most alters are only sparsely contacted, however, while a small proportion of alters are intensely contacted during the period. In addition, although contact frequencies are skewed to the right in these plots, they are heavily tailed, and more importantly, they exhibit some clustering behaviors, regardless of the length of the observation period. This finding gives us useful information in modeling the contact frequencies in Section 3. Fig. 2 shows the cumulative numbers of new alters contacted by egos of the 48 personal networks during the entire 92-day observation period. In the first 30 days, each ego had contact with 107 alters on average. By the end of the observation period, however, the average number of alters increased to 189. Fig. 3 shows the cumulative empirical distributions of the number of new alters acquired by an ego during different observation periods. The differences between these empirical distributions reflect our data collection procedure in that there was a rapid increase in the number of alters acquired by an ego during the early observation period. After a certain period (possibly 15–20 days), the distribution of such number gradually became stable.