An Efficient Network Generation Method: Interpersonal Networks and the Distribution of Links

Ben Klemens

Brookings Institution, Washington

Abstract

We describe a model of networks, that is both useful as a descriptive model for how interpersonal networks form, and as a tool for agent-based simulations that require artificially generated networks. It uses a latent space technique, but simplifies the standard computation using a principal component analysis, with no perceptible loss in fit. We test the method using friend networks in a sample of junior high school classrooms.

Keywords: Social networks, link density, latent space analysis

JEL Classification: Z13, C81

This paper describes a model of networks, that is both useful as a descriptive model for how interpersonal networks form, and as a tool for agent-based simulations that require artificially generated networks.

Interpersonal networks consistently demonstrate certain patterns, characterized by few very well-linked people and many sparsely-linked people. Sutton (1977) and Handcock and Jones (2003) give an overview of the various models intended to describe distributions of links that have such characteristics.

One thread in the literature describes networks via a latent space analysis (e.g., Hoff et al., 2002; Schweinberger and Snijders, 2003). The procedure is to first write down a likelihood function indicating the probability that two nodes in a space will link given the distance between them, and then, given a list of nodes and the links between them, find the configuration of nodes that maximizes the likelihood that the observed link distribution would occur.

[1] This paper is an offshoot from a larger project simulating smoking habits. Thanks to the rest of the team, notably Rob Axtell, Della Feher, Carol Graham, Jon Parker, Thom Valente, and Peyton Young.

Finding points via maximum likelihood is a maximally general, purely descriptive technique. It does not impose a structure on the distribution of nodes in the latent space or explain how networks similar to the given network would be formed.

As the papers cited above acknowledge, the likelihood search works well for small networks but does not scale to networks of thousands of agents, such as business-to-business networks or the Internet. By making no assumptions about the structure of the space, the maximum likelihood estimation must find the optimal location for tens of thousands of points, perhaps in a dozen or more dimensions, using a likelihood function that is on the order of n^2 terms. The papers above present methods to mitigate the computational burden, but the scaling problem persists for the most general version of the latent space search.

Thus, the method here imposes a specific structure upon the latent space. The additional assumptions provide more structure for comparison to other models and simplifies computation, but still leads to a very good fit to the data.

We project agents into a space via a principal component analysis (PCA, known in other fields as factor analysis or spectral decomposition), and assume a Normal distribution of nodes on each dimension in the projected space. Computing a PCA only requires finding the eigenvectors of a sparse matrix, and well-optimized computer packages make such computation tractable for data sets with many thousands of nodes. The computation of eigenvectors does not depend on the number of dimensions desired.

The additional constraints provide a more concrete explanation of how the network formed, and a procedure for forming new networks comparable to those in the given data set. However, more constraints will cause the node positions to be a worse fit to the data relative to an unconstrained maximum likelihood estimation. We compare the distribution of links for an artificial data set to the distribution observed in the data, and find that the assumptions produced no measurable distortions (see Figure 2).

1 The data

The method below is calibrated against data from a UCLA study by Thom Valente et al, which surveyed junior high school classrooms in Los Angeles. The surveys primarily focused on questions regarding students' attitudes toward smoking, but also asked students to list their five best friends. This paper uses

only the information on the choice of friends. [2]

The survey covered 86 classrooms in their entirety, save for students who were absent the day of the survey. For those students, there are data regarding who nominated them as friends, but not whom they would nominate. The methods below can use nominations of absent students without modification.

2 The factor analysis method

This section describes a network model based on first doing a principal component analysis of student preferences to place the students in an artificial space, and then probabilistically linking students based on their proximity in that space.

Once the parameters are written down based on the real-world data, the parameters can be used to produce new classroom networks whose characteristics match the original classrooms.

2.1 The parameters of the preference space

First, we generate a matrix listing who nominates whom to be a friend. That is, we converted the data for a thirty-student class into a thirty by thirty sparse matrix, where a one in position (m, n) indicates that student m nominated student n as a friend. Thus, each row has between zero and five ones, depending on the number of friends the student listed; the great majority listed five.

We then did a principal component analysis to find an appropriate basis for the matrix, and found that the best fit was with four dimensions, which explained over 90% of the variance in the data. These dimensions can be interpreted to represent underlying characteristics that determine a student's preferences. Since each dimension of a PCA is effectively orthogonal to the others, [3] the mean and standard deviation in each dimension is sufficient to describe the distribution in full.

Also, because the dimensions are orthogonal and only differences in position

[2] Just as standard latent space methods can acommodate relevant demographic or preference information by adding terms to the link likelihood function, the method here can accommodate more information by adding columns to the matrix of nominations.

[3] That is, with infinite data, the correlation would approach zero; with finite data, there is still small residual correlation between dimensions.

are relevant, we can define the distribution in each dimension as having $\mu \equiv 0$. We noted the standard deviation in each dimension and calculated the mean of standard deviations (σ) over the classrooms; the statistics are listed in Figure 1.

2.2 The Probit parameters

The next step is to determine when one student will choose to link with another. There are more than enough pairs of people who have much in common but are not friends, so a sensible model uses probabilistic linking, where the likelihood of linking increases as the distance between two people gets smaller, but where no link is guaranteed to be made or not made.

Let the utility from a link from student a to b be

$$U(\mathbf{a}, \mathbf{b}) = \beta_0 + \beta_1 D(a^1, b^1) + \beta_2 D(a^2, b^2) + \beta_3 D(a^3, b^3) + \beta_4 D(a^4, b^4) + \epsilon, \quad (1)$$

where a^n and b^n are the locations of the students in the nth dimension of the projected space, β_1 through β_4 are negative numbers to be estimated from the data, $\epsilon \sim N(0, 1)$, and β_0 is a fixed constant that ensures that the total number of links approximates the number of links in the original data set. We found values of β_n via maximum likelihood. The estimates are presented in Figure 1.

	Dimension			
	1	2	3	4
σ	0.093	0.076	0.064	0.059
β	-0.62	-0.62	-0.82	-0.81

class sizes: $\mu = 24.9$, $\sigma = 5.23$.
$\beta_0 = -0.66$

Fig. 1. The parameters needed to produce artificial networks.

3 Using the parameters

Given the parameters, one could repeat the steps above to generate a new classroom.

First, draw the size of the classroom, n. Figure 1 gives parameters that one

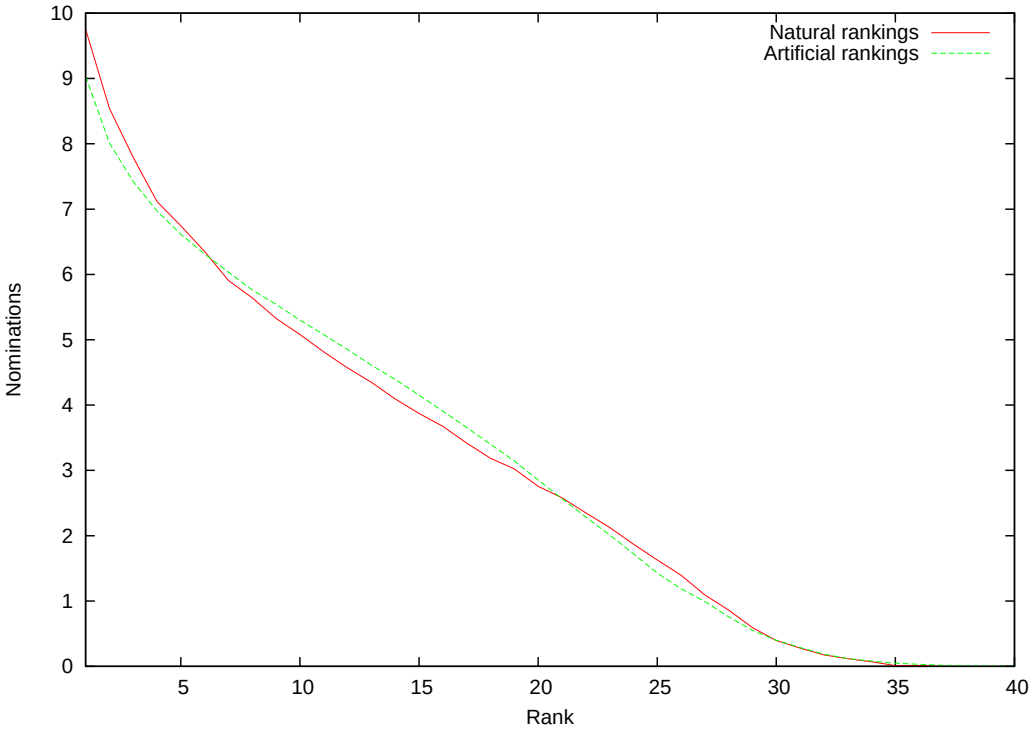

Fig. 2. The distribution of rankings for the natural and the artificial data.

may use to draw from a Normal distribution to select a class size. [4]

Then, generate n points in a four-dimensional space. Draw each dimension independently from four Normal distributions with mean zero and the variances listed in Figure 1.

Now that the artificial students have positions in the characteristic space, check for links. For each pair of students, calculate the value in Equation 1, where the βs are as in Figure 1 and $\epsilon \sim N(0, 1)$. If the value is greater than zero, then link the pair.

4 Comparing the natural and artificial data

This method of producing artificial networks is in no way based on the density of links in the real-world data. But the link distributions of the classrooms generated using the above algorithm prove to be remarkably close to the distribution of links in the natural data.

[4] The actual class sizes have distribution $\mu = 31.13$, $\sigma = 3.92$, but the distribution of class size in the sample is skewed downward. The recommended parameters in Figure 1 thus have mean 20% below the observed and a standard deviation $\frac{1}{3}$ larger.

Figure 2 shows the mean number of nominations for each rank of student for the natural data set and 1,000 artificial classrooms produced as above. A χ^2 test for goodness of fit confirms that the curves are very close: the test fails to reject the hypothesis that the artificial data is drawn from the actual with $> 99.999\%$ confidence. [5]

5 Conclusion

This paper presented a means of efficiently projecting a network onto a latent space. It may be used as a model of how the observed network formed, or as a means of generating classrooms whose link distribution matches those observed. The method consists of producing a cloud of points in a characteristic space and then probabilistically linking points in that space. This method does not depend on the iterative rich-get-richer means that have often been used to build networks, yet it generates a distribution of links with a distribution close to the real-world link distributions.

Further, the method is somewhat natural. It is reasonable to presume that friendships are formed based on underlying personal characteristics, that the distribution of those underlying characteristics generally form a bell curve, and that people who are close in personal characteristics are likely but not certain to be friends.

References

Breiger, R. L., Boorman, S. A., Arabie, P., Aug. 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology 12 (3), 328–383.

Handcock, M. S., Jones, J. H., January 2003. Likelihood-based inference for stochastic models of sexual network formation, working Paper 29, Center for Statistics and the Social Sciences, University of Washington.

Hoff, P. D., Raftery, A. E., Handcock, M. S., 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association 97 (460), 1090–1098.

Schweinberger, M., Snijders, T. A. B., 2003. Settings in social networks: A measurement model. Sociological Methodology 33, 307–341.

[5] The χ^2 statistic is $\sum_{i=1}^{40} \frac{(N_i - A_i)^2}{A_i} = 0.827$, where A_i is the mean nomination count for the ith ranked student in the artificial data set, and N_i is the same statistic for the natural data set.

Sutton, J., March 1977. Gilbrat's legacy. Journal of Economic Literature 35 (1), 40–59.