Network Science **1** (1): 16–30, 2013.
doi:10.1017/nws.2012.4

Friendship networks and social status

BRIAN BALL
Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
(*e-mail:* briball@umich.edu)

M. E. J. NEWMAN
Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109, USA
(*e-mail:* mejn@umich.edu)

Abstract

In empirical studies of friendship networks, participants are typically asked, in interviews or questionnaires, to identify some or all of their close friends, resulting in a directed network in which friendships can, and often do, run in only one direction between a pair of individuals. Here we analyze a large collection of such networks representing friendships among students at US high and junior-high schools and show that the pattern of unreciprocated friendships is far from random. In every network, without exception, we find that there exists a ranking of participants, from low to high, such that almost all unreciprocated friendships consist of a lower ranked individual claiming friendship with a higher ranked one. We present a maximum-likelihood method for deducing such rankings from observed network data and conjecture that the rankings produced reflect a measure of social status. We note in particular that reciprocated and unreciprocated friendships obey different statistics, suggesting different formation processes, and that rankings are correlated with other characteristics of the participants that are traditionally associated with status, such as age and overall popularity as measured by total number of friends.

Keywords: *friendship networks, social status, maximum likelihood, stochastic blockmodel*

1 Introduction

In this paper we consider social networks in which the ties represent friendship. Friendship networks have been the subject of scientific study since at least the 1930s (Wasserman & Faust, 1994). A classic example can be found in the study by Rapoport and Horvath (1961) of friendship among schoolchildren in the town of Ann Arbor, Michigan in the 1950s and 1960s, in which the investigators circulated questionnaires among the students in a school asking them to name their friends. Many other studies have been done since then, with varying degrees of sophistication, but most employ a similar questionnaire-based methodology. A crucial aspect of the resulting networks is that they are directed. Person A states that person B is a friend and hence there is a direction to the ties between individuals—it may be that B also states that A is a friend, but it does not have to be the case, and in practice it turns out that a remarkably high fraction of claimed friendships are not reciprocated. In the networks that we study in this paper the fraction of ties that are paired with a reciprocal tie in the opposite direction rarely exceeds 50% and can be as low as 30%.

This could be seen as a problem for the experimenter. One thinks of friendship as a two-way street—a friendship that goes in only one direction is no friendship

at all. How then are we to interpret the many unreciprocated connections in these networks? Are the individuals in question friends or are they not? One common approach is simply to disregard the directions altogether and consider two individuals to be friends if they are connected in either direction, or both—see Airoldi et al. (2011) for example. In this paper, however, we take a different view and consider what we can learn from the unreciprocated connections. It has been conjectured that, rather than being an error or an annoyance, the pattern of connections might reflect underlying features in the structure or dynamics of the community under study (Homans, 1950; Davis & Leinhardt, 1972; Doreian et al., 2000). Working with a large collection of friendship networks from US schools, we find that in every network there is a clear ranking of individuals from low to high such that almost all friendships that run in only one direction consist of a lower ranked individual claiming friendship with a higher ranked one. We conjecture that these rankings reflect a measure of social status and present a number of results in support of this idea. For instance, we find that a large majority of reciprocated friendships are between individuals of closely similar rank, while a significant fraction of unreciprocated friendships are between very different ranks, an observation consistent with qualitative results in the sociological literature going back several decades (Davis & Leinhardt, 1972). We also investigate correlations between rank and other individual characteristics, finding, for example, that there is a strong positive correlation between rank and age, older students having higher rank on average, and between rank and overall popularity, as measured by total number of friends.

The outline of the paper is as follows. First, we consider the general problem of a directed friendship network and describe a method for calculating rankings using a maximum-likelihood technique in combination with an expectation–maximization algorithm. We then apply this method to a collection of school friendship networks drawn from the US National Longitudinal Study of Adolescent Health (sometimes called the "AddHealth" study), a major survey of the social networks of school students conducted in the 1990s.[1] The study provides individual friendship networks for 84 different schools in the United States, and we apply our analysis separately to each network. As we show, a surprisingly uniform pattern emerges across essentially all of the networks, under which friendship patterns imply clear rankings that in turn are correlated with other measures and characteristics. In the final section of the paper we give our conclusions and discuss possible avenues for future research.

2 *Inference of rank from network structure*

Consider a directed network of friendships between n individuals in which a connection running from person A to person B indicates that A claims B as a friend. Suppose that, while some of the friendships in the network may be reciprocated or

[1] This work uses data from AddHealth, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due for Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from AddHealth should contact AddHealth, Carolina Population Center, 123 W Franklin Street, Chapel Hill, NC 27516–2524 (addhealth@unc.edu).

bidirectional, a significant fraction are unreciprocated, running in one direction only, and suppose we believe there to be a ranking of the individuals implied by the pattern of the unreciprocated friendships so that most such friendships run from lower to higher rank. One possible way to infer that ranking would be simply to ignore any reciprocated friendships and then use the remaining ones to rank individuals using any of several standard ranking methods based on pairwise comparisons, such as the methods commonly used for teams in sports leagues (Stefani, 1997; Callaghan et al., 2004). A simple example of such a method is the minimum violations ranking (Reinelt, 1985; Ali et al., 1986), in which one finds the ranking of network nodes that minimizes the number of connections running from higher ranked nodes to lower ranked ones. In practice this approach works quite well: for the networks studied in this paper the minimum violation rankings have an average of 98% of their unreciprocated friendships running from lower to higher ranks and only 2% running the other way. By contrast, versions of the same networks in which edge directions have been randomized have about 10% of edges running the wrong way on average. (Statistical errors in either case are 1% or less, so these observations are highly unlikely to be the results of chance.) Simple ranking methods such as this, however, are not ideal in the present case because their parametrization of the interplay between friendship direction and rank is rather crude, and in particular because they normally focus only on unreciprocated friendships. In most cases there are a substantial number of reciprocated friendships as well, as many as a half of the total, and they contain significant information about network structure and ranking. Moreover, it turns out that the statistics governing these reciprocated friendships are quite different from those for the unreciprocated ones and there are a number of other subtleties in the distribution of friendships that are potentially useful sources of information. To make full use of this information we need a more flexible and general method of analysis. In this paper we use a maximum likelihood approach defined as follows.

Mathematically we represent the distinction between reciprocated and unreciprocated friendships in the network using two separate matrices. The symmetric matrix $\mathbf{S}$ will represent the reciprocated connections such that $S_{ij} = S_{ji} = 1$ if there are connections both ways between nodes i and j, and zero otherwise. The asymmetric matrix $\mathbf{T}$ will represent the unreciprocated edges with $T_{ij} = 1$ if there is a connection to node i from node j (but not vice versa), and zero otherwise. The matrices $\mathbf{S}$ and $\mathbf{T}$ are related to the conventional adjacency matrix $\mathbf{A}$ of the network by $\mathbf{A} = \mathbf{S} + \mathbf{T}$.

Now suppose that there exists some ranking of the individuals, from low to high, which we will represent by giving each individual a unique integer rank in the range 1 to n. We will denote the rank of node i by r_i and the complete set of ranks by R. The probability of friendship between two individuals can, we assume, depend on their two ranks, or alternatively on any two independent combinations of the ranks. We have found it useful to view the problem in terms of the sum and difference of ranks; when represented in this way, we find that the dependence on the sum is weak, and most of the variation in probability is captured by the difference variable alone. In the work presented here, we therefore neglect the sum and treat the probability of friendship as a function of the difference of ranks only, which is an approximation, but a useful one that results in a significant simplification of the calculations. We do specifically allow the probability to be different for reciprocated and unreciprocated

friendships, which acknowledges the possibility that the two may represent different types of relationships as conjectured, for instance, in Davis and Leinhardt (1972) and Dijkstra et al. (2010). We define a function $\alpha(r_i - r_j)$ to represent the probability of having a pair of reciprocal edges between i and j and another $\beta(r_i - r_j)$ for an unreciprocated edge to i from j. Since $\alpha(r)$ describes reciprocal connections, it must be symmetric $\alpha(-r) = \alpha(r)$, but $\beta(r)$ need not be symmetric.

If we were not given a network but were given the probability functions α and β and a complete set of rankings on n vertices, then we could use this model to generate—for instance on a computer—a hypothetical but plausible network in which edges appeared with appropriate probabilities. In effect, we have a random graph model that incorporates rankings. In this paper, however, we want to perform the reverse operation: given a network we want to deduce the rankings of the nodes and the values of the functions α and β. To put this another way, if we are given a network and we assume that it is generated by our model, what values of the rankings and probability functions are most likely to have generated the network we observe?

This question leads us to a maximum likelihood formulation of our problem, which we treat using an expectation–maximization (EM) approach in which the ranks R are considered hidden variables to be determined and the functions α and β are parameters of the model. Maximum likelihood methods for inferring the values of hidden variables in networks, and particularly EM algorithms, have been the subject of a significant volume of recent research—see Goldenberg et al. (2009) or Snijders (2011) for a review.

The model presented here is, clearly, not a complete representation of the process of friendship formation. Unquestionably there are other factors besides rank that play into people's decisions to become friends. A realistic model of the process would certainly be more complicated and have many more parameters. Nonetheless, simplified models such as this one are not only common in data analysis but they also appear to work well in practice. If one is interested in estimating a set of hidden or latent variables, unobserved in the data but nonetheless affecting the observations, then a model incorporating only the effects of those latent variables plus a minimal set of other basic assumptions often turns out to produce useful estimates, and this perhaps surprising observation forms the foundation for a large part of modern statistics. Certainly it is possible to simplify models too far—the standard stochastic blockmodel of community structure in networks is an example of a model that in many cases does not incorporate enough basic features to produce good fits to real-world data for any parameter values and hence fails to extract latent structure even in simple and well-understood examples (Karrer & Newman, 2011). In many other cases, however, and particularly in the case of the present paper, a simplified model that identifies the crucial features and leaves out the rest, can impart substantial insight while avoiding unnecessary elaboration.

The particular model described in this paper is only one of many models that have been proposed to explain patterns of reciprocation in social network data. One of the best known previous models is the so-called p1 model of Holland and Leinhardt (1981), an early example of what would now be called an exponential random graph, which has parameters governing vertex degrees and a single parameter controlling the probability of reciprocated friendships. A variety of

extensions of this model, including some quite elaborate ones, have been subsequently proposed, such as those of Wang and Wong (1987) and Strauss and Ikeda (1990). See Wasserman and Pattison (1996) for a useful introduction to this area along with a wide range of examples. None of these models, however, employs ranking directly as a latent variable, and many of them lean in the opposite direction to our current goal of simplicity in model design, incorporating a range of elaborations that can increase the precision of the fit but also make interpretation more challenging.

Returning to the model proposed in this paper, we use a Poisson formulation in which the likelihood of generating a network G with rankings R, given the functions α and β, is

$$P(G,R|\alpha,\beta) = \prod_{i>j} \frac{[\alpha(r_i - r_j)]^{S_{ij}}}{S_{ij}!} \mathrm{e}^{-\alpha(r_i-r_j)} \prod_{i\neq j} \frac{[\beta(r_i - r_j)]^{T_{ij}}}{T_{ij}!} \mathrm{e}^{-\beta(r_i-r_j)}. \tag{1}$$

Note that we have excluded self-edges, since individuals cannot name themselves as friends. (We have also assumed that the prior probability on R is uniform over all sets of rankings, which is correct in the absence of any other rank information.)

The most likely values of the parameter functions α and β are now given by maximizing the marginal likelihood $P(G|\alpha,\beta) = \sum_R P(G,R|\alpha,\beta)$, or equivalently maximizing its logarithm, which is more convenient. The logarithm satisfies the Jensen inequality

$$\log \sum_R P(G,R|\alpha,\beta) \geqslant \sum_R q(R) \log \frac{P(G,R|\alpha,\beta)}{q(R)}, \tag{2}$$

for any set of probabilities $q(R)$ such that $\sum_R q(R) = 1$, with the equality being recovered when

$$q(R) = \frac{P(G,R|\alpha,\beta)}{\sum_R P(G,R|\alpha,\beta)}. \tag{3}$$

This implies that the maximization of the log-likelihood on the left side of Equation (2) is equivalent to the double maximization of the right side, first with respect to $q(R)$, which makes the right side equal to the left, and then with respect to α and β, which gives us the answer we are looking for. It may appear that expressing the problem as a double maximization in this way, rather than as the original single one, makes it harder, but in fact that is not the case.

The right-hand side of Equation (2) can be written as $\sum_R q(R) \log P(G,R|\alpha,\beta) - \sum_R q(R) \log q(R)$, but the second term does not depend on α or β, so as far as α and β are concerned we need consider only the first term, which is simply the average $\overline{\mathcal{L}}$ of the log-likelihood over the distribution $q(R)$:

$$\overline{\mathcal{L}} = \sum_R q(R) \log P(G,R|\alpha,\beta). \tag{4}$$

Making use of Equation (1) and neglecting an unimportant overall constant, we then have

$$\overline{\mathcal{L}} = \sum_R q(R) \sum_{i\neq j} \Big[\tfrac{1}{2} S_{ij} \log \alpha(r_i - r_j) + T_{ij} \log \beta(r_i - r_j) - \tfrac{1}{2}\alpha(r_i - r_j) - \beta(r_i - r_j) \Big], \tag{5}$$

where we have used the fact that $\alpha(r)$ is a symmetric function.

This expression can be simplified further. The first term in the sum is

$$\tfrac{1}{2}\sum_R q(R)\sum_{i\neq j} S_{ij}\log\alpha(r_i-r_j) = \tfrac{1}{2}\sum_z\sum_{i\neq j} S_{ij}q(r_i-r_j=z)\log\alpha(z), \tag{6}$$

where $q(r_i-r_j=z)$ means the probability within the distribution $q(R)$ that $r_i-r_j=z$. We can define

$$a(z) = \frac{1}{n-|z|}\sum_{i\neq j} S_{ij}q(r_i-r_j=z), \tag{7}$$

which is the expected number of reciprocated friendships in the observed network between pairs of nodes with rank difference z. It is the direct equivalent in the observed network of the quantity $\alpha(z)$, which is the expected number of edges in the model. The quantity $a(z)$, like $\alpha(z)$, is necessarily symmetric, $a(z)=a(-z)$, and hence Equation (6) can be written as

$$\tfrac{1}{2}\sum_R q(R)\sum_{i\neq j} S_{ij}\log\alpha(r_i-r_j) = \sum_{z=1}^{n-1}(n-z)a(z)\log\alpha(z). \tag{8}$$

Similarly, we can define

$$b(z) = \frac{1}{n-|z|}\sum_{i\neq j} T_{ij}q(r_i-r_j=z) \tag{9}$$

and

$$\sum_R q(R)\sum_{i\neq j} T_{ij}\log\beta(r_i-r_j) = \sum_{z=1}^{n-1}(n-z)\big[b(z)\log\beta(z)+b(-z)\log\beta(-z)\big], \tag{10}$$

where $b(z)$ is the expected number of unreciprocated edges between a pair of nodes with rank difference z. Our final expression for $\overline{\mathcal{L}}$ is

$$\begin{aligned}\overline{\mathcal{L}} = \sum_{z=1}^{n-1}(n-z)\big[&a(z)\log\alpha(z)-\alpha(z)\\ &+b(z)\log\beta(z)-\beta(z)+b(-z)\log\beta(-z)-\beta(-z)\big].\end{aligned} \tag{11}$$

Our approach involves maximizing this expression with respect to $\alpha(z)$ and $\beta(z)$ for given $a(z)$ and $b(z)$, which can be done using standard numerical methods. (Note that the expression separates into terms for the reciprocated and unreciprocated friendships, so the two can be maximized independently.) The values of $a(z)$ and $b(z)$ in turn are calculated from Equations (3), (7), and (9), leading to an iterative method in which we first guess values for $\alpha(z)$ and $\beta(z)$, use them to calculate $q(R)$ and hence $a(z)$ and $b(z)$, then maximize $\overline{\mathcal{L}}$ to derive new values of α and β, and repeat to convergence. This is the classic EM approach to model fitting.

To put this scheme into practice we need to specify a parametrization for the functions α and β, so that we can represent them on the computer. Since there are only a fixed number of values that the rank difference z can take in Equation (11) [$2(n-1)$ of them, to be precise] we could in principle represent the functions completely by specifying their value separately for each z. This would, however, probably overfit the data because in reality the functions must be quite smooth—we do not expect small differences in rank to make a big difference to the probability of friendship. For a smooth function, a natural parametrization is to use a Fourier

series, which is the choice we make in this paper. We have experimented with other parametrizations and find that our main conclusions are robust to the choice made, which suggests that the Fourier series is effective at capturing the form of the probability functions. But there is one aspect of the functions that it does not represent well. As discussed in the following section, we find that a substantial fraction of friendships in the networks analyzed in this paper run between individuals with closely similar rank, and we have found it necessary, in order to get a good fit to the model, to incorporate this observation into the parametrization by adding an additional term to both α and β consisting of a Gaussian peak of variable width, centered at the origin. With this addition we achieve robust fits to the model that are consistent across networks.

For the parametrization of the function α, which describes reciprocated friendships, we find good fits with only the central Gaussian peak plus a small uniform constant, which one can think of as the zeroth term in the Fourier expansion. For β, which represents the unreciprocated friendships, a more complicated form is needed—we use a five-term Fourier cosine series plus the central Gaussian peak. To some extent the choice of five terms is dictated by what is computationally feasible—our current numerical framework limits us to the five used here, but with improved computational resources it is possible that one could include more terms and achieve better fits. Nonetheless, the fits achieved appear robust—as mentioned above, other parametrizations produce similar functional forms.

Finally, we note that the sum in the denominator of Equation (3) is too large to be tractable numerically. In the calculations presented here, therefore, we approximate it using a Markov Chain Monte Carlo method. We generate complete rankings R in proportion to the probability $q(R)$ given by Equation (3) and average over them to calculate $a(z)$ and $b(z)$.

We have implemented the complete method in the C++ computer language. For those interested in applying it to their own calculations, a copy of the computer code can be found on-line.[2]

3 Results

We have applied the method of the previous section to the analysis of data from the AddHealth study, a large-scale multi-year study of social conditions for school students and young adults in the United States. Using results from surveys conducted in 1994 and 1995, the study compiled friendship networks for over 90,000 students in schools covering US school grades 7 to 12 (ages 12 to 18 years). Schools were chosen to represent a broad range of socioeconomic conditions. High schools (grades 9 to 12) were paired with "feeder" middle schools (grades 7 and 8) so that networks spanning schools could be constructed. (In three cases, the middle schools went up to grade 9 and high school started at grade 10 instead.)

To create the networks, each student was asked to select, from a list of students attending the same middle/high school combination, up to 10 people with whom

[2] http://www.umich.edu/~mejn/ranking

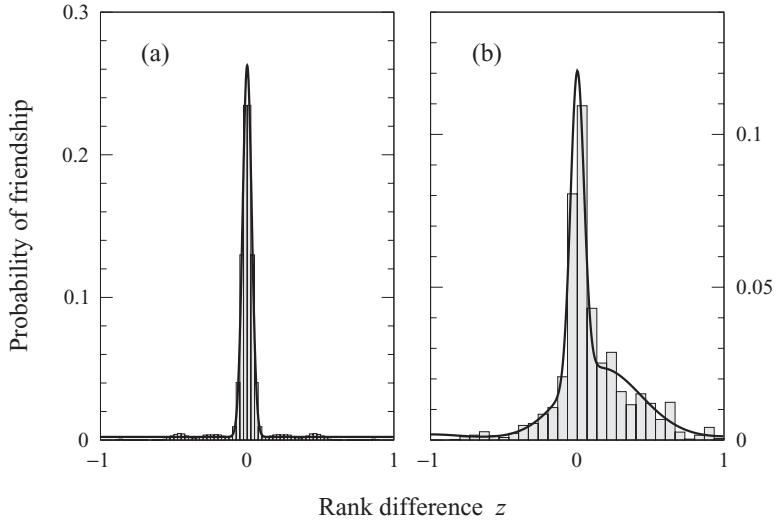

Fig. 1. (a) Probability of reciprocated friendships as a function of rank difference (normalized to run from −1 to 1). The histogram shows empirical results for a single example network; the solid curve is the fitted function $\alpha(z)$. (b) The equivalent plot for unreciprocated friendships. (Technically the fitted functions represent the expected number of edges rather than probability, but the two are asymptotically equivalent for the sparse networks studied here.)

they were friends, with a maximum of five being male and five female.[3] From these selections, 84 friendship networks were constructed ranging in size from tens to thousands of students. Each network is accompanied by data on the participants, including school grade, sex, and ethnicity. Some of the networks divide into more than one strongly connected component, in which case we restrict our analysis to the largest component only. We perform the EM analysis of the previous section on each network separately, repeating the iterative procedure until the rankings no longer change. Typically only a small number of iterations, around five or so, are needed for convergence, and in no case did the calculation require more than 15 iterations.

Figure 1 shows results for a typical network. In panel (a), the histogram shows the empirical probability of a reciprocated friendship between a vertex pair with rank difference z, given by the quantity $a(z)$ defined in Equation (7). The horizontal axis has been rescaled to run from −1 to 1 (rather than $-n$ to n). As the figure shows, the probability is significantly different from zero only for small values of z, with a strong peak centered on the origin. The solid curve shows the fit of this peak by the function $\alpha(z)$, which appears good. The fit is similarly good for most networks. The form of $a(z)$ tells us that most reciprocated friendships fall between individuals of similar rank: there is a good chance that two people with roughly equal rank will

[3] Students were also asked to list their best friends first and in principle there could be additional insight to be gleaned from the listing order. We have, however, ignored order in our analysis, treating all claimed friendships as equal. We do this in part because we wish to create methods that could be applied to other data for which ordering information is not available, and in part to demonstrate that the ordering is not in fact necessary for deducing useful rankings.

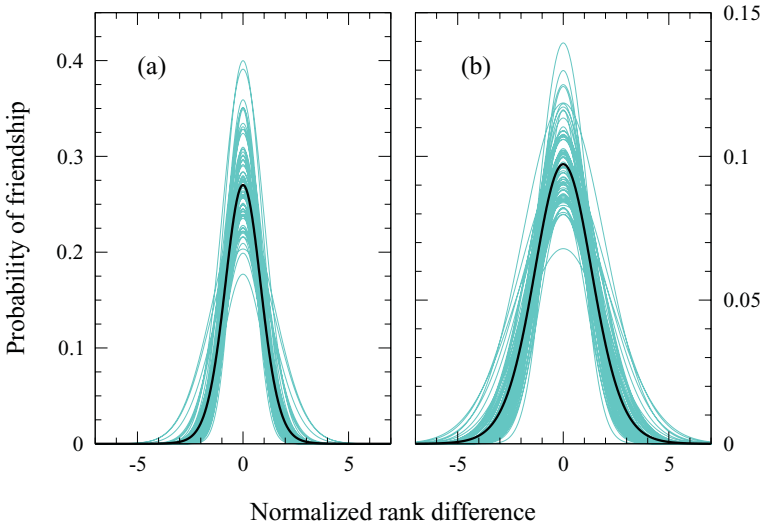

Fig. 2. The fitted central peak of the friendship probability distributions for (a) reciprocated, and (b) unreciprocated friendships. The horizontal axes are measured in units of absolute (unrescaled) rank difference divided by average network degree. Each light-colored curve is a network. The bold black curves represent the mean. (color online)

both claim the other as a friend, but very little chance that two people with very different ranks will do so. This result seems at first surprising, implying as it does that people must be able to determine their own and others' rank with high accuracy in order to form friendships, but previous studies have suggested that indeed this is true (Anderson et al., 2006).

Panel (b) in Figure 1 shows $b(z)$, Equation (9), for the same network, which is the probability of an unreciprocated edge between nodes with rank difference z. Again there is a strong central peak to the distribution, of width similar to that for the reciprocated edges, indicating that many unreciprocated friendships are between individuals of closely similar rank. However, the distribution also has a substantial asymmetric tail for positive values of the rank difference, indicating that in a significant fraction of cases individuals claim friendship with those ranked higher than themselves, but those claims are not reciprocated. The solid curve in the panel shows the best-fit form of the function $\beta(z)$ in the maximum-likelihood calculation.

The general forms of these distributions are similar across networks from different schools. They also show interesting scaling behavior. The widths of the central peaks for both reciprocated and unreciprocated connections, when measured in terms of raw (unrescaled) rank difference are, to a good approximation, simply proportional to the average degree of a vertex in the network. Figure 2 shows these peaks for 78 of the 84 networks on two plots, for reciprocated edges (panel (a)) and unreciprocated ones (panel (b)), rescaled by average degree, and the approximately constant width is clear. (The six networks not shown are all small enough that the central peaks for the unreciprocated edges can be fit by the other parameters of the model and thus a direct comparison is not appropriate.) This result indicates that individuals have, roughly speaking, a fixed probability of being friends with others close to them in

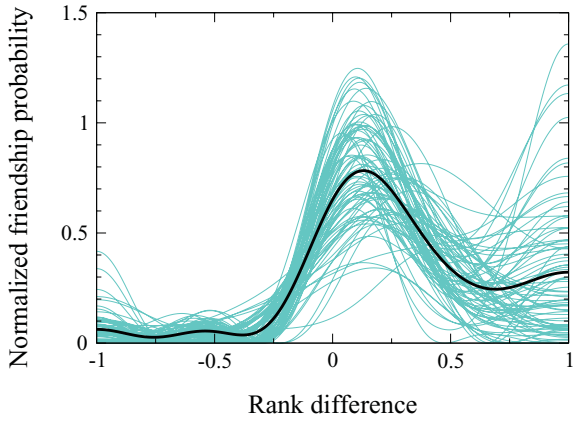

Fig. 3. The fitted probability function for unreciprocated friendships, minus its central peak. The horizontal axis measures rank difference rescaled to run from −1 to 1. Each light-colored curve is a network. The bold black curve is the mean. Note that the variation among the curves becomes larger as rank difference approaches 1, which happens because there is relatively little data in this region, so statistical fluctuations in the fits are more significant. The same is true in principle for the central peaks shown in Figure 2, but there is no corresponding variation between the curves in that case because all curves go to zero for large rank difference. (color online)

rank, regardless of the size of the community as a whole—as the average number of friends increases, individuals look proportionately further afield in terms of rank to find their friends, but are no more likely to be friends with any particular individual of nearby rank.

Outside the central peak, i.e., for friendships between individuals with markedly different ranks, there are, to a good approximation, only unreciprocated friendships, and for these the shape of the probability distribution appears by contrast to be roughly constant when measured in terms of the rescaled rank of Figure 1, which runs from −1 to 1. This probability, which is equal to the function $\beta(z)$ with the central Gaussian peak subtracted, is shown in Figure 3 for the same 78 networks, rescaled vertically by the average probability of an edge to account for differing network sizes, and again the similarity of the functional form across networks is apparent, with low probability in the left half of the plot, indicating few claimed friendships with lower ranked individuals, and higher probability on the right. The roughly constant shape suggests that, among the unreciprocated friendships, there is, for example, a roughly constant probability of the lowest ranked student in the school claiming friendship with the highest ranked, relative to other students, no matter how large the school may be.

The emerging picture of friendship patterns in these networks is one in which reciprocated friendships appear to fall almost entirely between individuals of closely similar rank. A significant fraction of the unreciprocated ones do the same, and moreover show similar scaling to their reciprocated counterparts, but the remainder seem to show a quite different behavior characterized by different scaling and by claims of friendship by lower ranked individuals with substantially higher ranked ones.

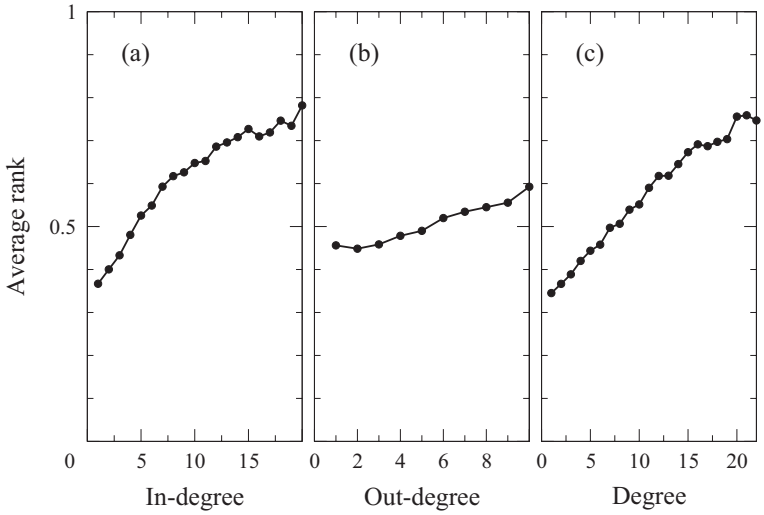

Fig. 4. Plots of rescaled rank versus degree, averaged over all individuals in all networks for (a) in-degree, (b) out-degree, and (c) the sum of degrees. Measurement errors are comparable with or smaller than the sizes of the data points and are not shown.

4 Discussion

Based on the results of the previous section, we conjecture that the rankings discovered by our analysis correlate, at least approximately, with social status. If we assume that reciprocated friendships—almost all of which fall in the central peak—correspond to friendships in the conventional sense of mutual interaction, then a further conjecture, on the basis of the similarity of the statistics, is that the unreciprocated friendships in the central peak are also mutual but, for one reason or another, only one side of the relationship is represented in the data. One explanation why one side might be missing is that respondents in the surveys were limited to listing only five male and five female friends, and so might not have been able to list all of their friendships.

By contrast, one might conjecture that the unreciprocated claims of friendship with higher ranked individuals, those in the tail of the distribution in Figure 1(b), correspond to "aspirational" friendships, hopes of friendship with higher ranked individuals that are, for the moment at least, not returned. Note also how the tail falls off with increasing rank difference: individuals are more likely to claim friendship with others of only modestly higher rank, not vastly higher.

One way to test these conjectures is to look for correlations between the rankings and other characteristics of individuals in the networks. For instance, it is generally thought that social status is positively correlated with the number of people who claim you as a friend (Hallinan & Kubitschek, 1988; Dijkstra et al., 2010). Figure 4(a) tests this by plotting average rank over all individuals in all networks (averaged in the posterior distribution of Equation (1)) as a function of network in-degree (the number of others who claim an individual as a friend). As the figure shows, there is a strong positive slope to the curve, with the most popular individuals being nearly twice as highly ranked on average as the least popular. Figure 4(b) shows the corresponding plot for out-degree, the number of individuals one claims as a friend,

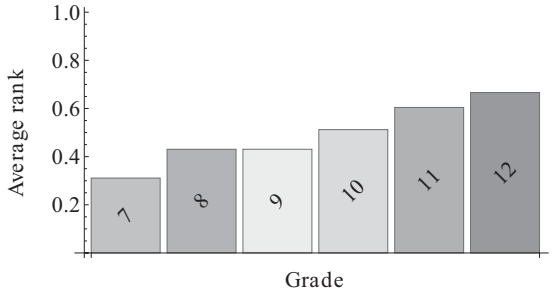

Fig. 5. Rescaled rank as a function of school grade, averaged over all individuals in all schools.

and here the connection is weaker, as one might expect—claiming many others as friends does not automatically confer high status upon an individual—although the correlation is still statistically significant. (Bear in mind that the out-degree has a largest possible value of 10 because survey participants were allowed to list at most 10 friends, which may reduce the variation of the out-degree and hence the size of its effect on rank.) Figure 4(c) shows rank as a function of total degree, in-degree plus out-degree, which could be taken as a measure of total social activity, and here again the correlation is strong. For all three panels the correlations are significant, with p-values less than 0.001. (Note, however, that rank and degree are not presumptively independent in the first place, since these are derived from the same data, a fact that should be borne in mind in interpreting these results.)

In addition to the network structure itself, we have some other data about each of the participants, including their age (school grade), sex, and ethnicity. The distributions of rank for each sex and individual ethnicities turn out to be close to uniform—a member of either sex or any ethnic group is, to a good approximation, equally likely to receive any rank from 1 to n, indicating that there is essentially no effect of sex or ethnicity on rank. (The Kolmogorov–Smirnov test does reveal deviations from uniformity in some cases, but the deviations are small, with Kolmogorov–Smirnov statistics $D < 0.08$ in all instances.) Age, however, is a different story. Figure 5 shows the rescaled rank of individuals in each grade from 7 to 12, averaged over all individuals in all networks, and here there is a clear correlation. Average rank increases by more than a factor of two from the youngest students to the oldest (a one-way ANOVA gives $p < 0.001$). Since older students are generally acknowledged to have higher social status (Coleman, 1961), this result lends support to the identification of rank with status. A further interesting wrinkle can be seen in the results for the 8th and 9th grades. Unlike other pairs of consecutive grades, these two do not have a statistically significant difference in average rank (a t-test gives $p > 0.95$). This may reflect the fact that the 8th grade is the most senior grade in most of the feeder junior-high schools, before students move up to high school. When they are in the 8th grade, students are temporarily the oldest (and therefore highest status) students in school and hence may have a higher rank than would be expected were all students in a single school together.

Finally, in Figure 6 we show an actual example of one of the networks that we analyze, with nodes arranged vertically on a scale of inferred rank and colored according to grade. The increase of rank with grade is clearly visible, as is the

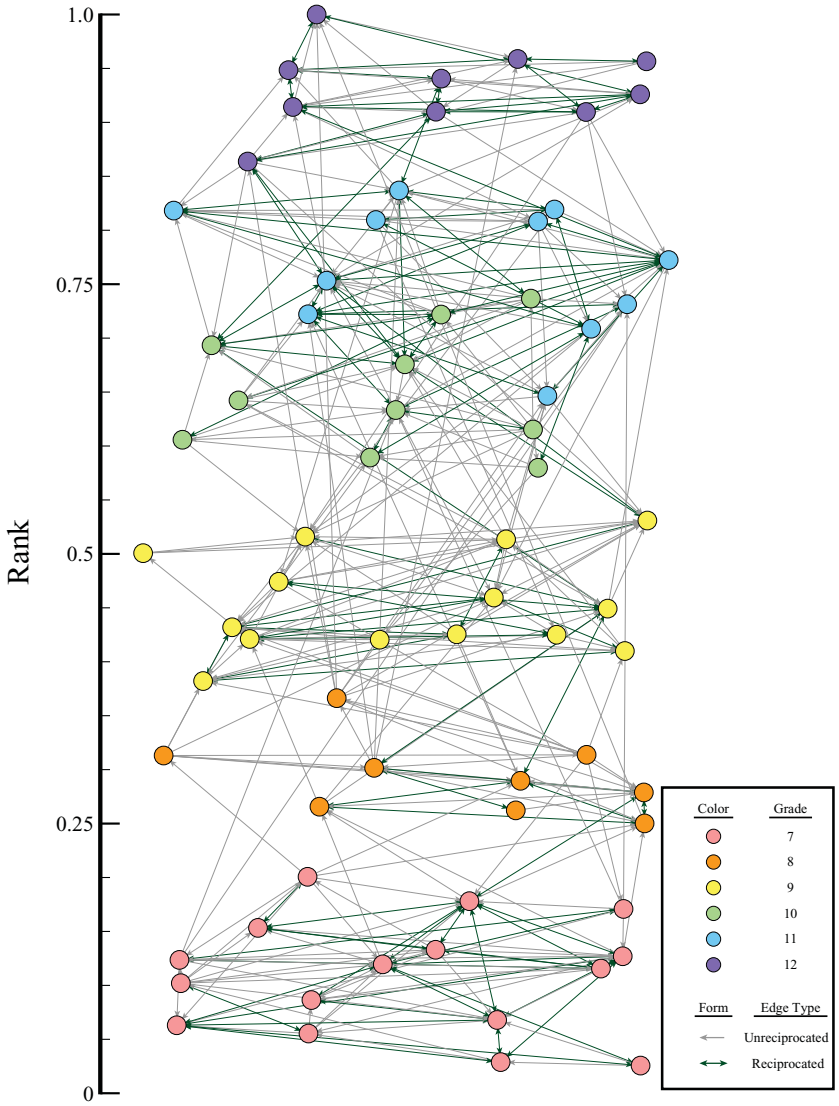

Fig. 6. A sample network with (rescaled) rank on the vertical axis, vertices colored according to grade, and reciprocated friendships colored differently from unreciprocated ones. Rank is calculated as an average within the Monte Carlo calculation (i.e., an average over the posterior distribution of the model), rather than merely as the maximum-likelihood ranking. Note the clear correlation between rank and grade in the network. (color online)

fact that most reciprocated friendships are between individuals of similar rank (and hence run horizontally in the figure).

5 Conclusions

In this paper we have analyzed a large set of networks of friendships between students in American high and junior-high schools, focusing particularly on the distinction between friendships claimed by both participating individuals and friendships claimed by only one individual. We find that students can be ranked from low to

high such that most unreciprocated friendships consist of a lower ranked individual claiming friendship with a higher ranked one. We have developed a maximum-likelihood method for inferring such ranks from complete networks, taking both reciprocated and unreciprocated friendships into account, and we find that the rankings so derived correlate significantly with traditional measures of social status such as age and overall popularity, suggesting that the rankings may correspond to status. At the same time, the rankings seem to be essentially independent on average of other characteristics of individuals such as sex or ethnicity.

There are a number of questions unanswered by our analysis. At present we have access only to limited data on the personal characteristics of participants. It would be interesting to test for correlation with other characteristics. Are rankings correlated, for instance, with academic achievement, number of siblings or birth order, after-school activities, personality type, body mass index, wealth, or future career success? Some of these questions will probably never be answered, but some additional variables, including some of those above, were measured by the experimenters during the original study and could form the basis for future investigations.

There is also the question of why a significant number of apparently close friendships are unreciprocated. One idea that has appeared in the literature is that some unreciprocated connections may correspond to new, temporary, or unstable friendships, which are either in the process of forming and will become reciprocated in the future, or will disappear over time (Sørensen & Hallinan, 1976; Hallinan & Kubitschek, 1988). Evidence suggests that in practice about a half of the unreciprocated friendships do the former and a half the latter, and it is possible that the two behaviors correspond to the two classes of connections we identify in our analysis. A test of this hypothesis, however, would require longitudinal data—successive measurements of friendship patterns among the same group of individuals—data which at present we do not possess. Finally, there are potential applications of the statistical methods developed here to other directed networks in which direction might be correlated with ranking, such as networks of team or individual competition (Stefani, 1997; Callaghan et al., 2004) or dominance hierarchies in animal communities (Drews, 1993; De Vries, 1998).

Acknowledgments

The authors thank Carrie Ferrario, Brian Karrer, Cris Moore, Jason Owen-Smith, Bethany Percha, and Claire Whitlinger for useful comments and suggestions. This work was funded in part by the National Science Foundation under grant DMS-1107796 and by the James S. McDonnell Foundation.

References

Airoldi, Edoardo M., Choi, David S., & Wolfe, Patrick J. (2011). Confidence sets for network structure. *Statistical Analysis and Data Mining*, **4**, 461–469.

Ali, I., Cook, W. D., & Kress, M. (1986). On the minimum violations ranking of a tournament. *Management Science*, **32**, 660–672.

Anderson, C., Srivastava, S., Beer, Jennifer S., Spataro, Sandra E., & Chatman, Jennifer A. (2006). Knowing your place: Self-perceptions of status in face-to-face groups. *Journal of Personality and Social Psychology*, **91**, 1094–1110.

Callaghan, T., Mucha, P. J., & Porter, M. A. (2004). The bowl championship series: A mathematical review. *Notices of the American Mathematical Society*, **51**, 887–893.

Coleman, James S. (1961). *The Adolescent Society: The Social Life of the Teenager and Its Impact on Education*. Westport, CT: Greenwood Press.

Davis, James A., & Leinhardt, S. (1972). The structure of positive interpersonal relations in small groups. *Sociological Theories in Progress*, **2**, 218–251.

De Vries, H. (1998). Finding a dominance order most consistent with a linear hierarchy: A new procedure and review. *Animal Behaviour*, **55**, 827–843.

Dijkstra, J. K., Cillessen, Antonius H. N., Lindenberg, S., & Veenstra, R. (2010). Basking in reflected glory and its limits: Why adolescents hang out with popular peers. *Journal of Research on Adolescents*, **20**, 942–958.

Doreian, P., Batagelj, V., & Ferligoj, A. (2000). Symmetric-acyclic decompositions of networks. *Journal of Classification*, **17**, 3–28.

Drews, C. (1993). The concept and definition of dominance in animal behaviour. *Behaviour*, **125**, 283–313.

Goldenberg, A., Zheng, Alice X., Feinberg, Stephen E., & Airoldi, Edoardo M. (2009). A survey of statistical network structures. *Foundations and Trends in Machine Learning*, **2**, 1–117.

Hallinan, Maureen T., & Kubitschek, Warren N. (1988). The effect of individual and structural characteristics on intransitivity in social networks. *Social Psychology Quarterly*, **51**, 81–92.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**, 33–50.

Homans, G. C. (1950). *The Human Group*. San Diego, CA: Harcourt Brace.

Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review* E, **83**, 016107.

Rapoport, A., & Horvath, W. J. (1961). A study of a large sociogram. *Behavioral Science*, **6**, 279–291.

Reinelt, G. (1985). *The Linear Ordering Problem: Algorithms and Applications*. Berlin, Germany: Heldermann.

Snijders, Tom A. B. (2011). Statistical models for social networks. *Annual Review of Sociology*, **37**, 131–153.

Sørensen, Aage B., & Hallinan, Maureen T. (1976). A stochastic model for change in group structure. *Social Science Research*, **5**, 43–61.

Stefani, R. (1997). Survey of the major world sports rating systems. *Journal of Applied Statistics*, **24**, 635–646.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, **85**, 204–212.

Wang, Yuchung J., & Wong, George Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8–19.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis*. Cambridge, UK: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov random graphs and p*. *Psychometrika*, **61**, 401–426.