Ego-centered networks and the ripple effect

M.E.J. Newman

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

Abstract

Recent work has demonstrated that many networks have broad distributions of vertex degree. Here we show that this has a substantial impact on the shape of ego-centered networks and on concepts and methods based on ego-centered networks, such as snowball sampling and the “ripple effect”. In particular, we argue that one’s acquaintances, one’s immediate neighbors in the acquaintance network, are far from being a random sample of the population, and that this biases the numbers of neighbors two and more steps away. We demonstrate this concept using data on academic collaboration networks.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Ego-centered networks; Personal networks; Degree distribution; Ripple effect; Collaboration networks

1. Introduction

In social network parlance, an ego-centered network (sometimes also called a personal network) is a network centered on a specific individual (generically “actor”), whom we call the ego (Killworth et al., 1990; Wellman, 1993; Wasserman and Faust, 1994; Scott, 2000). For example, Sigmund Freud and all his friends would form an ego-centered network. This network would have radius 1, meaning we include everyone within distance 1 on the friendship network of the central individual, Freud in this case. If we also included friends of friends in the network, it would have radius 2. In Fig. 1, we show a radius-2 ego-centered network of scientific collaborations. The ego in this case is my own: the central vertex in the figure represents me, the first ring of vertices around that my co-authors on papers and books published within the last 10 years, and the second ring their co-authors. As the figure shows, networks of this type can grow very rapidly with radius.

Ego-centered networks are of interest for a number of reasons. For instance, in two recent papers, Bernard et al. (1991, 2001) address the following question. Consider some subset of
the population, consisting of $e$ people. They could be people in a particular demographic or social group, or the people involved in a particular event. How many of these $e$ people, if any, is the typical person likely to know? As Bernard et al. show, this is easy to calculate. If the total population who might be involved in the event is $t$, then each member of that population has a probability $p = \frac{e}{t}$ of being involved. If the average person knows $c$ other people, then the average number of those people who were involved is simply $m = cp = \frac{ct}{c}$. Bernard et al. take the example of the population of the US, for which they estimate from previous empirical studies that the average person has a social circle of about $c = 290$ people (Killworth et al., 1990; Bernard et al., 1991), and for which the total population currently stands at around 280 million. Thus, the ratio $tc \simeq 1,000,000$ in this case, giving the simple rule of thumb:

$$m = \frac{e}{1,000,000}.$$  \hfill (1)
Simply stated, this equation says that the average individual living in the US is acquainted with about one person in a million out of the country’s total population.

As an example, let us apply the method to the problem of estimating how many HIV positive individuals the average person in the US knows. At the time of writing, there were about 800,000 known cases of HIV in the US (including those who have died). The number of actual cases is probably substantially greater than this and is estimated to be somewhere between 1.0 and 1.5 million. To take a conservative figure, let us suppose that the actual total is \( e = 1 \) million. From Eq. (1), we then estimate that on average each member of the US population as a whole has or had one acquaintance who is or was HIV positive. It must be emphasized that this is an average figure. HIV positive individuals are not a uniform sample of the population. Nonetheless, Eq. (1) is expected to give a correct population average of \( m \).

Now we want to extend this calculation one step further. If a person has no immediate friends in the group under consideration, how many of their friends’ friends are in this group? Alternatively, one could rephrase the question and inquire how many people in the population as a whole have one or more friends of friends in the specified group: one can visualize a group or event as the center of a set of ever widening circles of influence in the social network. Colloquially, this is what we call the “ripple effect”. The two questions here are equivalent, but not identical. In this paper we speak in the language of the former, which focuses our attention on the calculation of the number of actors two steps away from the ego in an ego-centered network.

Unfortunately, the calculation of this number is not simple. The crucial point to notice is that in many networks there exists a small number of actors with an anomalously large number of ties. While it may appear safe to ignore these actors because they form only a small fraction of the population, we show that in fact this is not so. Because of the way the ripple effect works, this small minority has a disproportionately large influence, and ignoring them can produce inaccurate estimates for the figures of interest. We show here how to perform calculations that take these issues into account correctly.

The topic of this paper is also of interest in some other areas of social network theory. One such area is “snowball sampling”, an empirical technique for sampling social networks that attempts to reconstruct the ego-centered network around a given central actor (Erickson, 1978; Frank, 1979). In this technique, the central actor is first polled to determine the identities of other actors with whom he or she has ties. Then those actors are polled to determine their ties, and so forth, through a succession of generations of the procedure. The statistical properties of snowball samples have been studied using Markov chain theory (Heckathorn, 1997) and the technique has been shown to give good (or at least predictable) samples of populations in the limit where a large number of generations of actors is polled. Unfortunately, in most practical studies only a small number of generations is polled, and in this case, as we will see, the sample may be biased in a severe fashion: snowball samples, like calculations of the ripple effect, are highly sensitive to the presence in the population of a small number of actors with an unusually large number of ties.

The calculations given here are also, to some extent, relevant to the spread of disease, which can be viewed in terms of ripples of infection spreading in ever wider circles from an initial disease carrier. Applications in this area are discussed elsewhere (Newman, 2002).

The outline of this paper is as follows. In Section 2 we calculate exactly the expected number of network neighbors at distance 2 from a central individual, in a network with-
out transitive triples. In Section 3 we show how the resulting expression is modified when the network has transitivity, and in Section 4 we apply our theory to two example networks, showing that in practice it appears to work extremely well. In Section 5 we give our conclusions.

2. Friends of friends

So how do you estimate the number of people who are two steps away from you in a social network (or indeed in a network of any kind)? Bernard et al. (2001) suggest the following simple method. If each actor in a network has ties to \( c \) others on average, and each of those has ties to \( c \) others, then the average number of actors two steps away is approximately \( c^2 \). A slightly better estimate would be \( c(c - 1) \), which allows for the fact that one of the \( c \) ties your neighbor has is to you yourself, so only \( c - 1 \) are to other people.

There are some problems with this estimate, however. First, as pointed out by Bernard et al., people who know one another tend to have strongly overlapping circles of acquaintance, so that not all of the \( c \) people your friend knows are new to you—many of them are probably friends of yours. In other circles this effect is called network transitivity (Wasserman and Faust, 1994) or clustering (Watts and Strogatz, 1998), and it is also related to the concept of network density (Granovetter, 1976). Typically, the mean number of people two steps away from an actor can be reduced by a factor of two or so by transitivity effects. Bernard et al. allow for this by including a “lead-in factor” \( \lambda \) in their calculation. We discuss transitivity in more detail in Section 3.

Even if we ignore the effects of transitivity, however, there is a substantial problem with the simple estimate of the number of one’s second neighbors in a social network. By approximating this number as \( c(c - 1) \), we are assuming that the people we know are by and large average members of the population, who themselves know average numbers of other people. But, as pointed out by Feld (1991), we would be quite wrong to make such an assumption. The people we know are anything but average.

Consider two (fictitious) individuals. Individual A is a hermit with a lousy attitude and bad breath to the point where it interferes with satellite broadcasts. He has only 10 acquaintances. Individual B is erudite, witty, charming, and a professional politician. She has 1000 acquaintances. Is the average person equally likely to know A and B? Absolutely not. The average person is 100 times more likely to know B than A, since B knows 100 times as many people. Extending this argument to one’s whole circle of friends, it is clear that the people one knows will, overall, tend to be people with more than the average number of acquaintances (Feld, 1991). This means that the total number of their friends—the people two steps away—will be larger than our simple estimate would suggest. And as we will show, it may be very much larger.

Understanding this issue fully will require us to indulge in some moderately involved mathematics. Readers who find such things distressing may wish to skip Eqs. (2)–(7). Our principal results are summarized in Eqs. (8), (12) and (14).

---

1 Here and elsewhere in this paper, we assume that ties between actors are symmetric. In other words, if A is friends with B, then we assume also that B is friends with A.
The fundamental concept that we need to capture here is that not all people have the same number of acquaintances. In the language of social network analysis, there is a distribution of the degrees of vertices in the social network. (Recall that the degree of a vertex is the number of other vertices to which it is directly connected.) Let $k$ denote the degree of a vertex and $p_k$ the degree distribution, i.e. the probability that a vertex chosen uniformly at random from the network will have degree $k$. Thus, for example, the mean degree $c$ of a vertex is:

$$c = \bar{k} = \sum_{k=0}^{\infty} kp_k.$$  

(2)

Degree distributions have been measured for a variety of networks, and in many cases are found to show great variation (Albert et al., 1999; Faloutsos et al., 1999; Amaral et al., 2000; Newman, 2001; Jeong et al., 2001; Liljeros et al., 2001). It is certainly not true that vertices always have degree close to the mean, although they may in some cases (e.g. Marsden, 1987). An example of a network with a broad degree distribution can be seen in Fig. 1, where some vertices have degree only 1, while at least one has degree greater than 100.

Now the ideas above can be expressed mathematically as follows. The distribution of the degrees of the vertices to which a random vertex is connected is not given by $p_k$. The probability that you know a particular person is proportional to the number of people they know, and hence the distribution of their degree is proportional to $kp_k$ and not just $p_k$ (Feld, 1991; Newman et al., 2001). The correctly normalized distribution is thus:

$$q_k = \frac{kp_k}{\sum_{k=0}^{\infty} kp_k}.$$  

(3)

Now consider the number of vertices two steps away from a given vertex. The probability $P(k_2|k_1)$ that this number is $k_2$, given that the number of vertices one step away is $k_1$, is:

$$P(k_2|k_1) = \sum_{m_1=1}^{\infty} \sum_{m_2=1}^{\infty} \cdots \sum_{m_{k_1}=1}^{\infty} \delta \left( \sum_{i=1}^{k_1} (m_i - 1), k_2 \right) \prod_{j=1}^{k_1} q_{m_j},$$  

(4)

where $\delta(m, n)$ is 1 if $m = n$ and 0 otherwise. Note the occurrence of $m_i - 1$ in this expression; the amount that your $i$th neighbor contributes to the total number of your second neighbors is one less than his or her degree, because one of his or her neighbors is you. The overall probability that the number of second neighbors is $k_2$ can then be calculated by averaging Eq. (4) over $k_1$:

$$P(k_2) = \sum_{k_1=0}^{\infty} p_{k_1} P(k_2|k_1).$$  

(5)

We want the mean value of $k_2$, which we will denote $c_2$, and this is given by:

$$c_2 = \bar{k}_2 = \sum_{k_2=0}^{\infty} k_2 P(k_2).$$  

(6)
Combining Eqs. (3)–(6), we thus arrive at the quantity we are interested in:

\[
c_2 = \sum_{k_2=0}^{\infty} \sum_{k_1=0}^{k_2} p_{k_2} P(k_2|k_1) = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{k_1} k_2 P(k_2|k_1)
\]

\[
= \sum_{k_1=0}^{\infty} p_{k_1} \sum_{m_1=1}^{k_1} \cdots \sum_{m_{k_2}=1}^{k_1} k_{2} \delta \left( \sum_{i=1}^{k_1} (m_i - 1), k_2 \right) \prod_{j=1}^{k_1} q_{m_j}
\]

\[
= \sum_{k_1=0}^{\infty} p_{k_1} \sum_{i=1}^{k_1} \sum_{m_i=1}^{\infty} \cdots \sum_{m_{k_2}=1}^{m_i-1} (m_i - 1) \prod_{j=1}^{k_1} q_{m_j}
\]

\[
= \sum_{k_1=0}^{\infty} k_1 p_{k_1} \left[ \sum_{m_i=1}^{\infty} q_{m_i} \right] \prod_{j=1}^{k_1-1} \sum_{k=0}^{\infty} (k-1)q_k = \sum_{k=0}^{\infty} k(k-1)p_k = \bar{k}^2 - \bar{k}.
\]

(7)

Here we have performed the sum over \( k_2 \) in the second line to get rid of the \( \delta \)-function and all of the sums over the variables \( m_i \) in the third line (noting that the factor \( (m_i - 1) \) depends on only one of them). The result:

\[
c_2 = \bar{k}^2 - \bar{k},
\]

(8)

is the correction we were looking for to the simple estimate of the number of vertices two steps away.\(^2\) The number of vertices two steps away is given by the mean square degree minus the mean degree. The important point to notice is that this expression depends on the average of the square of a vertex’s degree, rather than the square of the average, as the simple estimate assumes. If the degrees of vertices are narrowly distributed about their mean, then these two quantities will be approximately equal and the simple estimate will give roughly the right result. As mentioned above, however, many networks have broad degree distributions, and in this case the average of the square and the square of the average will take very different values. In general, we can write:

\[
c_2 = \bar{k}^2 - \bar{k} + (\bar{k}^2 - \bar{k}^2) = c(c - 1) + \sigma^2,
\]

(9)

where \( \sigma^2 \) is the variance of the degree distribution and \( c \) is, as before, its mean. Thus, the difference between the simple estimate \( c(c - 1) \) and the true value of \( c_2 \) is equal to the variance. In Section 4 we give some examples of real networks for which the variance is large—much larger than \( c(c - 1) \)—and hence for which the simple estimate gives poor results.

3. Transitivity and mutuality

The calculation of the previous section is incomplete for a number of reasons. Chief among these is that it misses the effect of network transitivity or clustering. In most social networks, adjacent actors have strongly overlapping sets of acquaintances. To put this

\(^2\) An alternative derivation of this result using probability generating functions can be constructed using the results of Newman et al. (2001).
Fig. 2. An illustration of the calculation of the clustering coefficient for a small network. Vertex A has three paths of length 2 leading from it, as marked. Similarly vertices B, C and D have 3, 2 and 2 such paths, for a total of $3 + 3 + 2 + 2 = 10$. There is one triangle in the network. Hence, from Eq. (10), the clustering coefficient is $6 \times (1/10) = 3/5$. Alternatively, one can count the number of connected triples of vertices, of which there are five, one each centered on vertices A and B, three on vertex C, and none on vertex D. Using Eq. (11), the clustering coefficient is then $3 \times (1/5) = 3/5$ again.

Transitivity can be measured by the quantity:

$$C = \frac{6 \times \text{number of triangles in the network}}{\text{number of paths of length 2}}.$$  \hspace{1cm} (10)

Here paths of length 2 are considered directed and start at a specified vertex. A “triangle” is any set of three vertices all of which are connected to each of the others. The factor of six in the numerator accounts for the fact that each triangle contributes six paths of length 2 to the network, two starting at each of its vertices. This definition is illustrated in Fig. 2. Simply put, $C$ is the probability that a friend of one of your friends will also be your friend.

The quantity $C$ has been widely studied in the theoretical literature, and its value has been measured for many different networks. Watts and Strogatz (1998) have dubbed it the clustering coefficient. It is sometimes also known as the “fraction of transitive triples” in the network. Eq. (10) is not in the form of the standard definition, and so may not be immediately recognizable as the same quantity discussed elsewhere. The most commonly used definition is:

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}},$$  \hspace{1cm} (11)

where a “connected triple” means a vertex that is connected to an (unordered) pair of other vertices. It takes only a moment to convince oneself that the two definitions are equivalent—see Fig. 2 again. (Note that paths are ordered in (10) and triples are unordered in (11), which accounts for an apparent difference of a factor of two between the two definitions.)

What effect does clustering have on our calculation of the number $c_2$ of second-nearest neighbors in the network? Consider a vertex with degree $m$ lying in the first “ring” of our ego-centered network, i.e. one of the immediate neighbors of the central vertex. Previously we considered all but one of this vertex’s $m$ neighbors to be second neighbors of the central vertex. (The remaining one is the central vertex itself.) This is why the term $m - 1$ appears in Eq. (7). Now, however, we realize that in fact an average fraction $C$ of those $m - 1$ neighbors are themselves neighbors of the central vertex and hence should not be counted as second neighbors. Thus, $m - 1$ in Eq. (7) should be replaced with $(1 - C)(m - 1)$.

Making this substitution in Eq. (7) we immediately see that:

$$c_2 = (1 - C)(\bar{k}^2 - \bar{k}).$$  \hspace{1cm} (12)
This result is in general only approximate, because the probability of a vertex having a tie to another in the first ring is presumably not independent of the degrees $m_i$ of the other vertices. As we show in the following section, however, Eq. (12) gives considerably better estimates of $c_2$ than our first attempt, Eq. (8).

But this is not all. There is another effect we need to take into account if we are to estimate $c_2$ correctly. It is also possible that we are over-counting the number of second neighbors of the central individual in the network because some of them are friends of more than one friend. In other words, you may know two people who have another friend in common, whom you personally do not know. Such relationships create “squares” in the network, rather than the triangles of the simple transitivity. To quantify the density of these squares, we define another quantity which we call the mutuality $M$:

$$M = \frac{\text{mean number of vertices two steps away}}{\text{mean number of paths of length 2 to those vertices}}$$

In words, $M$ measures the mean number of paths of length 2 leading to your second neighbor. Because of the squares in the network, Eq. (12) overestimates $c_2$ by exactly a factor of $1/M$, and hence our theory can be fixed by replacing $m - 1$ in Eq. (7) by $M(1 - C)(m - 1)$.

But now we have a problem. Calculating the mutuality $M$ using Eq. (13) requires that we know the mean number of individuals two steps away from the central individual. But this is precisely the quantity $c_2$ that our calculation is supposed to estimate in the first place. Our entire goal here is to estimate $c_2$ without having to measure it directly, which would in any case be quite difficult for most networks. Instead, therefore, we introduce an approximation which allows us to make reasonably accurate estimates of $c_2$ using only knowledge about the degree distribution and the clustering coefficient.

Consider the two configurations depicted in Fig. 3a and b. In (a), the ego, denoted E and shaded, has two friends A and B, both of whom know F, although F is a stranger to E. The same is true in (b), but now A and B are friends of one another also. The approximation we introduce is to count only configurations of type (b) and not those of type (a). As we will show, we can estimate the frequency of configurations of type (b) from a knowledge of the clustering coefficient. In some networks we find that (a) occurs only rarely, in which case our approximation gives excellent results. In others (a) may not be rare and our approximation will then give a partial correction to Eq. (12), but will still tend to overestimate $c_2$ somewhat.

Consider Fig. 3c. The central actor E has a tie with A, who has a tie with F. How many other paths of length 2 are there from E to F? Well, if E has $k_1$ neighbors, as before, then by the definition (11) of the clustering coefficient, A will have ties to $C(k_1 - 1)$ of them on average. The tie between actors A and B in the figure is an example of one such. But now A has ties to both B and F, and hence, using the definition of the clustering coefficient again, B and F will themselves have a tie (dotted line) with probability $C$. Thus, there will on average be $C^2(k_1 - 1)$ other paths of length 2 to F, or $1 + C^2(k_1 - 1)$ paths in total, counting the one that runs through A. This is the average factor by which we will over-count the number of second neighbors of E because of the mutuality effect. Substituting into Eq. (7), we then

---

3 We choose the name “mutuality” because $M$ is also the average number of mutual friends of two actors who are distance 2 apart in the network.
conclude that our best estimate of $c_2$ is:

$$c_2 = M(1 - C)(\bar{k} - \tilde{k}),$$

(14)

where the mutuality coefficient $M$ is given by:

$$M = \frac{\bar{k}/[1 + C^2(k - 1)]}{\bar{k}}.$$

(15)

Notice that both $1 - C$ and $M$ tend to 1 as $C$ becomes small, so that Eq. (14) becomes equivalent to Eq. (8) in a network where there is no clustering, as we would expect.

In essence what Eq. (15) does is estimate the value of $M$ in a network in which triangles of ties are common, but squares that are not composed of adjacent triangles are assumed to occur with frequency no greater than one would expect in a purely random network.

To summarize, if we know the degree distribution and clustering coefficient of a network—both of which can be estimated from knowledge of actors’ personal radius-1 networks—then we can estimate the number $c_2$ of friends of friends the typical actor has using Eqs. (8), (12) and (14). These three equations we expect to give successively more accurate results for $c_2$. Because we have neglected configurations of the form shown in Fig. 3a and because of approximations made in the derivation of Eqs. (12) and (14), we do not expect any of them to estimate $c_2$ perfectly. As we will see in the following section, however, Eq. (14) provides an excellent guide to the value of $c_2$ in practice, with only a small error (less than 10% in the cases we have examined).

4. Example application

In this section we test our theory by applying it to two networks for which we can directly measure the mean number of second neighbors of a vertex and compare it with the predictions of Eqs. (8), (12) and (14).
Academic co-authorship networks are one of the best documented classes of social networks. In these networks the vertices represent the authors of scholarly papers, and two vertices are connected by an edge if the two individuals in question have co-authored a paper together. With the advent of comprehensive electronic databases of published papers and preprints, large co-authorship networks can be constructed with good reliability and a high degree of automation. Co-authorship networks are true social networks in the sense that two individuals who have co-authored a paper are very likely to be personally acquainted. (There are exceptions, particularly in fields such as high-energy physics, where author lists running to hundreds of names are not uncommon. We will not be dealing with such exceptions here, however.)

We examine two different co-authorship networks:

1. A network of collaborations between 1.5 million scientists in biomedicine, compiled by the present author (Newman, 2001) from all publications appearing between 1995 and 1999 inclusive in the Medline bibliographic database, which is maintained by the National Institutes of Health.

2. A network of collaborations between a quarter of a million mathematicians, kindly provided to the author by Jerrold Grossman and Patrick Ion (Grossman and Ion, 1995; de Castro and Grossman, 1999), who compiled it from data provided by the American Mathematical Society.

In Fig. 4 we show the degree distributions of these networks. As the figure shows, neither is narrowly distributed about its mean. Both in fact are almost power-law in form, with long tails indicating that there are a small number of individuals in the network with a very large

---

4 In fact, co-authorship networks are most correctly represented as affiliation networks (i.e. bi-partite graphs) of authors and papers, with edges connecting authors to the papers that they wrote or co-wrote (Newman et al., 2001). Here, however, we deal only with the one-mode projection of the co-authorship network onto the authors.
Table 1
Summary of results for collaboration networks of mathematicians and biomedical scientists

<table>
<thead>
<tr>
<th>Network</th>
<th>Actors</th>
<th>Mean degree</th>
<th>Clustering</th>
<th>Estimate of $c_2$</th>
<th>Actual $c_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Simple</td>
<td>Good</td>
</tr>
<tr>
<td>Mathematics</td>
<td>253,339</td>
<td>3.92</td>
<td>0.150</td>
<td>11.4</td>
<td>47.9</td>
</tr>
<tr>
<td>Biomedicine</td>
<td>1,520,251</td>
<td>15.53</td>
<td>0.081</td>
<td>225.6</td>
<td>2006.0</td>
</tr>
</tbody>
</table>

number of collaborators. In the network of mathematicians, for instance, a plurality (about one-third) of individuals who have collaborated at all have degree 1, i.e. have collaborated with only one other. But there is one individual in the network, the legendary Hungarian Paul Erdős, who collaborated with a remarkable 502 others (Hoffman, 1998). (This number is a lower bound; even though Erdős died in 1996, new collaborations of his are still coming to light through publications he co-authored that are just now appearing in print.)

We have calculated the number of second neighbors of the average vertex in these networks in five different ways: using the simple estimate discussed in the introduction, using the three progressively more sophisticated estimates, Eqs. (8), (12) and (14), developed here, and directly by exhaustive measurement of the networks themselves. The results are summarized in Table 1. As expected, the simple method of estimating $c_2$, which assumes it to be equal to $c(c - 1)$, gives an underestimate for both networks, by a factor of more than three for the mathematicians and more than five for the biomedical scientists. Moreover, we have been quite generous to the simple method in this calculation, omitting from the formulas any correction for transitivity, such as the lead-in factors discussed in the introduction. Including such a correction gives estimates of $c_2 = 8.1$ and 152.8 for the two networks. These estimates are too low by factors of over 4 and 8, respectively—large enough errors to be problematic in almost any application.

By contrast, the new method does much better. The “good” and “better” estimates, Eqs. (8) and (12), give figures of the same general order of magnitude as the true result, and provide good rule-of-thumb guides to the expected value of $c_2$. But the best estimate, Eq. (14), making use of Eq. (15) to calculate the mutualty coefficient $M$, does better still, giving figures for $c_2$ that are within 8% and 9% of the known correct answers for the mathematics and biomedicine networks, respectively. Clearly this is a big improvement over the simple estimate. Eq. (14) appears to be accurate enough to give very useful estimates of numbers of friends of friends in real social networks.

5. Conclusions

There are a number of morals to this story. Perhaps the most important of them is that your friends just are not normal. No one’s friends are. By the very fact of being someone’s friend, friends select themselves. Friends are by definition friendly people, and your circle of friends will be a biased sample of the population because of it. This is a relevant issue for many social network studies, but particularly for studies using ego-centered techniques such as snowball sampling.
In this paper we have not only argued that your friends are unusual people, we have also shown (in a rather limited sense) how to accommodate their unusualness. By careful consideration of biases in sampling and correlation effects such as transitivity in the network, we can make accurate estimates of how many people your friends will be friends with. We have demonstrated that the resulting formulas work well for real social networks, taking the example of two academic co-authorship networks, for which the mean number of a person’s second neighbors in the network can be measured directly as well as estimated from our equations.

It is important to note, however, that application of the formulas we have given requires the experimenter to measure certain additional parameters of the network. In particular, it is not enough to know only the mean number of ties an actor has. One needs to know also the distribution of that number. Measuring this distribution is not a trivial undertaking, although some promising progress has been made (Granovetter, 1976; Killworth et al., 1990; Bernard et al., 1991). One must also find the clustering coefficient of the network, which requires us to measure how many pairs of friends of an individual are themselves friends. This may require the inclusion of additional questions in surveys as well as additional analysis.

To return then to the question with which we opened this paper, can we estimate how many friends of friends a person will have on average who fall into a given group or who were involved in a given event? If the number involved in the event is $e$ as before, and the total population is $t$, then the number we want, call it $m_2$, is given by $m_2 = c_2 e / t$. Thus, once we have $c_2$ we can answer our question easily enough. Using figures appropriate for the US and the simple estimate of $c_2$ that it is equal to $c(c - 1)$ where $c$ is the number of acquaintances the average person has, we get $c_2 = 290 \times 289 = 83,810$, $t = 280$ million, and $m_2 = e/3340$. As we have seen here, however, this probably underestimates the actual figure considerably. The real number could be a factor of five or more greater than this formula suggests. Unfortunately, as far as we know, the necessary data have not been measured for typical personal acquaintance networks to allow us to estimate $c_2$ by the methods described here. In particular, measurements of the clustering coefficient are at present lacking. We encourage those involved in empirical studies of these networks to measure these things soon.

Acknowledgements

The author thanks Peter Dodds, Michelle Girvan, Duncan Watts, and Elizabeth Wood for useful conversations, and Jerry Grossman, Oleg Khovayko, David Lipman, and Grigoriy Starchenko for graciously providing data used in the examples of Section 4. This work was funded in part by the National Science Foundation under grant number DMS-0109086.

References