

Project Description

We propose to develop new survey and analysis methods to measure and study various dimensions of polarization in social networks in the United States. Specifically, we propose to extend the theory and practice of “How many X’s do you know” surveys, an innovative idea of Killworth, McCarty, et al. (1998b). The **intellectual merit** of our proposal is: **(1)** Developing new methods of analyzing “How many X’s do you know” surveys¹ to learn about properties of the social network and of individuals and groups within this network; **(2)** Developing and improving the design of these surveys so they can be efficiently used as a general research tool for network analysis; **(3)** Designing, conducting, and analyzing a new survey to study social and political attitude polarization (hereafter referred to as attitude polarization) in social networks. Our analysis will measure the diversity of individuals’ networks, the social distribution of groups, by which we mean the variation in the frequencies of people in these groups being known by members of the general population, and the social factors that potentially explain this variation.

Broader impacts. We anticipate contributions in a number of areas in statistics and the social sciences. In statistics, the project will contribute to sampling methods, network analysis, and multilevel regression of structured data. In sociology, it will have a potentially important impact on the study of social capital, social networks, attitude polarization, and segregation. In political science, it will contribute to the measurement of attitude polarization. For example, the methods developed in the proposed study can provide answers to the following questions: To what extent are Americans divided between Democrats and Republicans, liberals and conservatives, secularists and believers? To what extent are divisions arranged in social networks of family, friends and acquaintances? How segregated are the patterns of social interaction of Americans on dimensions like church going, nontraditional family forms, or political ideology in comparison to race? The answers to these questions are highly relevant to an overall assessment of the level of social integration in contemporary America.

1 Polarization of social and political attitudes in social networks

1.1 Social capital, network polarization, and attitude polarization

There is a widespread perception that American society is becoming more heterogeneous over time along social, economic, and political dimensions. Taking a broad historical perspective, scholars argued that the life course was becoming more standardized as a consequence of urbanization and modernization up through the middle of the twentieth century (Modell, Furstenberg, and Hershberg 1976). By the 1950s, Americans had developed a consensus about the ideal family, a husband/father who worked, a housewife/mother and two children (Cherlin 1992). But since that time, “de-standardization” has been the norm (Mayer 2005) as the rates of cohabitation and fertility outside of marriage have increased, age at marriage has increased, the marriage rate for those with less education has declined, and unions have become more unstable (Held 1986, Buchmann 1989, Bianchi and Casper 2000). At the same time, income inequality at the household level has grown dramatically, which has been attributed to changes in labor markets, the growing dispersion of family forms, and assortative mating at marriage. Accompanying these trends has been a polarization of American politics when measured as the ideological distance between Democratic and Republican members of Congress (McCarty, Poole, and Rosenthal 2002).

These trends are related. McCarty et al. (2002) have shown that political partisanship is increasingly stratified by income. Meanwhile, our own calculations using data from the General Social Surveys have demonstrated that political partisanship is also increasingly stratified by family type and by religiosity. Many commentators have argued that these trends are creating a growing “values divide” in the United States (White 2003, Frank 2004, Green et al. 1996) that is articulated in terms of attitudes toward the family, sexuality, and gender. Brooks (2002), for example, found growing dispersion over the question of whether the family is in decline over time, and found that opinions were largely structured by regular attendance at church. There has been recent debate among political scientists as to whether opinions of the general public have become more polarized (McCarty, Poole, and Rosenthal 2005) or have remained centrist (Fiorina 2005).

The trends in family forms, in religiosity, and in social attitudes and political behavior are certainly not yet fully understood (e.g., Seltzer et al. 2005). One important component of these explanations almost certainly involves diffusion through social networks. Rogers and Kincaid (1981), Montgomery and Casterline (1993), and Kohler (2001) have all argued that fertility change occurs in part through the diffusion of information about methods of fertility control and information about the behavior of others that changes perceptions of social

¹In the following, we will refer to these surveys and corresponding survey questions as “How many X’s” surveys and questions.

acceptability and preferences. Nazio and Blossfeld (2003) have argued that social modeling accounts for the spread of cohabitation within European countries. Similarly, Brooks (2002) argued that high rates of church attendance created exposure to church based communications and influence and thereby led to relatively high rates of concern about family decline by those who regularly attend church. More recently, Rindfuss et al. (2004) measured whether Japanese respondents knew anyone who was engaging in “innovative” family behavior (using formal childcare, cohabiting, having a nonmarital birth, etc.), and found that knowing someone who has engaged in a less traditional family practice is strongly associated with having less traditional attitudes toward that practice. Verba et al. (1995) studied connections between demographics, attitudes, personal resources, and political behavior. While not definitive, all these results suggest the possibility of mutual reinforcement between attitudes and behavior through the mechanism of social ties.²

In a study of changing social attitudes in the United States, DiMaggio et al. (1996) argued that attitude polarization involves at least three distinct dimensions. The first is some notion of variance, which contains the idea that a large portion of the population has opinions that differ to some considerable extent from those of another large portion of the population.³ The second idea is one of constraint, which links back to the notion of ideological cohesion that was articulated in the classic essay of Converse (1964) on the nature of belief systems, and implies that attitudes on ideologically related issues would be associated with each other; knowing someone’s position on one attitude allows a prediction about their position on other attitudes. DiMaggio et al. note the connection to Coleman’s (1958) idea of “cross-issue contagion,” in which opinions on one issue propagate to related issues as their links are made clear through some process of communication or social learning. Finally, the third dimension is “consolidation,” the extent to which opinion trends move together across different groups characterized by social variables such as occupation, ethnicity, gender, or religion (what Page and Shapiro 1992 referred to as “parallel publics”).

DiMaggio et al. (see also Mouw and Sobel 2002, and Evans 2002, 2003) focused their attention on social attitudes and on trends in social attitudes. Our interest is instead in the social bases for polarization. We agree with other scholars that attitudes may be “consolidated” on the basis of social status and that social learning, reinforcement and attitude “contagion” can occur through social interaction. Consistent with existing research (e.g., Rindfuss et al. 2004), we hypothesize that if two individuals have different attitudes, they are likely to have different patterns of interaction. Patterns of interaction and attitudes are mutually reinforcing: attitudes diffuse through networks, and people with specific attitudes tend to seek out others whose attitudes and behavior are compatible with their own. It follows from this logic that individuals whose views on social attitudes are relatively extreme are likely to have *patterns of social interaction* that reinforce and are reinforced by these extreme attitudes; to restate: their social networks are likely to be relatively homogeneous with respect to the attitude or associated behavior in question, and furthermore their social networks are likely to differ substantially from those whose social attitudes are sharply different.

By patterns of interaction, we mean three specific qualities: the level of trust toward those who potentially could be in one’s network, the structural locus of interaction, and *overdispersion* of specific ties, by which we mean variation in the relative propensity of people to “know” members of a specific social group. (We shall consider overdispersion in more detail in Section 1.2 and in our statistical model in Section 2.) Putnam (1995, 2000) considers social trust to be a precondition for the social interaction that builds social capital. The structural locus of interaction is the amount of interaction in the specific dimensions contained in Putnam’s concept of social capital (political participation, religious participation, civic participation, and informal social connections), which have a variety of positive externalities (Helliwell and Putnam 2004). The dimensions of social capital in turn can be divided into what Putnam refers to as “bonding” social capital, the interaction with people like oneself that tend to reinforce exclusive identities—creating strong in-group loyalties—[but] may also create strong out-group antagonism (Putnam 2000), and “bridging” social capital which involves connections that “are outward looking and encompass people across diverse social cleavages” (Putnam 2000). This distinction between bonding and bridging social capital applies directly to the concept of overdispersion. Bonding social capital tends to involve homophilous social ties, while bridging social capital more likely involves social ties with those who are different from oneself. In a world where social ties were almost always of the bonding type, one would expect overdispersion to be very high: one would have many ties to people like

²McPherson et al. (2001), however, conclude from their extensive review that these patterns are generally a consequence more of social selection than of social influence. Another possibility is that some unmeasured factor affects both attitudes and behavior. However, experimental research has established that values do affect patterns of interaction (Huston and Levinger 1978).

³DiMaggio et al. also argued that polarization should involve the relative absence of people whose views are intermediate between the two more extreme groups, and thus in this way parallels conceptions that have been proposed for income inequality (a “hollowing out” of the middle class).

oneself, and few or no ties to people who were different from oneself.⁴ Putnam could not directly measure overdispersion and therefore could find “no reliable, comprehensive, nationwide measures of social capital that neatly distinguish ‘bridgingness’ and ‘bondingness,’ ” which caused him to de-emphasize this distinction in his empirical analysis more than he would have preferred (Putnam 2000). Our project, in contrast, will provide a direct measure of these two forms of social ties.⁵

1.2 Studying polarization by modeling overdispersion

In this project, we will develop new survey and analysis methods to measure network polarization. By surveying a random sample of Americans and asking them about the people they know (where the meaning of “know” is discussed later in the proposal), we will obtain partial information about the hundreds of persons in their social network, which will allow us to measure overdispersion of ties to both relatively large populations (e.g., regular churchgoers) and also small populations (e.g., gay couples who attend conservative Christian churches, or liberal Democrats who live in a rural area) that would be too costly using existing methods.

Overdispersion of ties can be viewed as a measure of polarization of social networks since it measures the variation in the relative propensity for a person to form a tie with persons of type X. A person’s relative propensity to know persons of type X can be defined as the ratio of the expected number of ties to this type from this person and the number of ties that would occur by chance given the degree of a person’s network and the population size of type X. Overdispersion is a more general conception of nonrandomness than is homophily, which was originally defined by Lazarsfeld and Merton (1954) and is now a frequently researched topic in the study of networks (McPherson, Smith-Lovin, and Cook 2001). In contrast to homophily, which involves the tendency to associate with people like oneself (which corresponds to Putnam’s notion of bonding social capital), overdispersion can be defined with respect to patterns of association with those unlike oneself (which corresponds to Putnam’s concept of bridging social capital) and can further measure variation in bonding and bridging social capital both within and between specific social groups.

Our ability to develop measures of the overdispersion of ties to people of different groups allows for the creation of interactional measures of segregation that differ from the residential measures that are typically used. To illustrate, if the relative propensity to know African-Americans were constant in the population, then interactional segregation would be zero. It is, of course, certain that this relative propensity varies (this fact is roughly equivalent to Putnam’s assertion that bonding social capital exists in the world, or to the assertion that people’s networks tend to be homophilous). Other scholars have recently argued for the importance of measuring segregation as the tendency for individuals of different races to come into contact with each other (Echenique and Fryer 2004). Complementing that proposed indirect method, our approach provides a direct measure of this form of segregation. In addition, we can estimate a model for the structural basis for this association-based segregation. Thus, for example, our approach allows estimation of the overdispersion that is residual to socioeconomic status or geographic location. We can use techniques of multilevel regression to decompose the variation in the overdispersion of ethnic minorities and other population subgroups.

Our approach allows for the development and testing of hypotheses about the relationships between attitudes and social ties, and about the relationship between the relative propensity for one type of social tie to the relative propensity for other types of social ties. For example, knowing someone who is living with a gay partner is probably associated with (and arguably influences) one’s attitudes towards homosexuality or towards gay marriage. More generally, we predict that people with relatively low propensities to know people different from themselves are more likely to have extreme social and political attitudes, which is similar to Putnam’s (2000) conjecture (untested, because he did not have suitable data) that bridging social capital should be positively related to civic tolerance. We further hypothesize that if high levels of attendance at an evangelical church is

⁴As elaborated below, overdispersion measures more than simply whether bridging social capital is high or low. We would expect overdispersion to be lower with a combination of high bridging social capital and low bonding social capital than with a combination of low bridging social capital and high bonding social capital. However, overdispersion can be created by mechanisms that are not captured by Putnam’s distinction. If 12% of an average white person’s network ties were to black people, overdispersion would be lower than if only 1% of an average white person’s network ties were to black people. However, overdispersion could still exist because of variation in the relative propensity within the white population to know black people, and because of variation in the relative propensity within the black population to know black people.

⁵Putnam’s more recent Social Capital Community Benchmark Survey (2000) includes extensive questions about social ties but does not have comprehensive direct measures of “bridging” and “bonding” social capital. The only question in that new survey that would provide such information asks how many times in the past month the respondent has been in the home of a friend of a different race or had them in the respondent’s home. Our project, in contrast, will collect data and provide methods for estimating overdispersion across an array of social groups defined by demographic status, religious behavior, and political attitudes.

associated with concerns about family decline (which includes concerns about gay marriage; see Brooks 2002), then relative propensity to know evangelicals would be related to the relative propensity to know gay couples as well as to attitudes towards gay marriage. Finally, we can make hypotheses at aggregate levels, such as cities or states. We hypothesize that overdispersion involving family forms, religion, race, and other salient social statuses in a state is strongly related to that state’s level of attitude polarization. For example, the level of overdispersion of ties with gay couples should be associated with the level of polarization of attitudes towards gay marriage.

Consideration of network polarization provides insights into the nature and sources of attitude polarization. To start, measurement of network polarization provides a basis for more accurately measuring attitude polarization. To put it simply, two individuals whose measured positions on a social attitude are a certain distance apart are likely to be more distant in reality (and also in the future) if their network ties to statuses that predict this attitude (i.e., “consolidated statuses”) are dissimilar. Furthermore, the consequences of attitude polarization for voting, for political and social conflict, and for further rises in polarization are likely to be greater when the salient social networks are more strongly overdispersed than when they are more random. The aspects of social status that predict attitudes (and thus form the basis for attitude “consolidation”) are the natural dimensions along which network overdispersion should be measured. Our approach also allows the estimation of a network version of DiMaggio et al.’s (1996) notion of constraint, namely the extent to which extreme relative propensities for relevant consolidated statuses tend to be correlated. We further generalize by estimating variation in these correlations as a function of the relative propensities of knowing people with salient consolidated statuses (e.g., is the probability of having constrained social and political attitudes affected by one’s relative propensity to know people who are different from oneself?), instead of simply addressing whether attitudes on different social issues are correlated. Finally, we can estimate the extent to which the degree of one’s network (the number of people that one “knows”) is by itself predictive of social attitudes.

Similarly, we expect both trust and the level of social capital along Putnam’s enunciated dimensions to structure one’s propensity to interact with diverse types of people. We would expect that high levels of trust increase one’s relative propensity to know people unlike oneself and therefore also reduce the extent to which one’s attitudes on social issues tend to be extreme. Conversely, low trust is predicted to be associated with more skewed social ties and more extreme positions on social and political attitudes.

Standard surveys like the General Social Surveys and the National Election Studies already contain measures of trust and of social capital as well as social attitudes. Our approach is distinctive because it allows for the measurement of network polarization and thus allows us to study the linkages between network polarization and attitude polarization. While existing studies such as Rindfuss et al. (2004) allow one to assess the correlation between knowing certain types of people and having certain attitudes, they do not provide the information needed for measuring important attributes of these ties, specifically the degree of one’s network, and the extent to which one’s ties to people of a certain type exceeds or falls short of what one would predict from chance alone based on the size of this type in the general population. Without this information, Rindfuss et al. (2004) are not able to establish the extent to which ties to people of a particular type are polarized (i.e., have high overdispersion). Their data do not allow one to determine the connection between having an extreme position on a particular attitude and one’s relative propensity to know people who (based on their objective statuses) are likely to hold similar positions on this attitude. They cannot tell whether attitude polarization is linked with polarization of social networks.

1.3 New survey focusing on social network and attitude polarization

We propose to conduct a survey to learn about polarization of traits and opinions, as manifested as overdispersion of groups in the social network. Our key tool will be “How many X’s” questions, which can be used to study overdispersion of groups, as we explain in Section 2. We propose to incorporate a module containing such questions into the 2006 General Social Survey (please see the attached letter of support from Tom Smith, the director of the GSS). The GSS offers several advantages for the type of research that we propose here. First, it offers the possibility of collecting high-quality data through face-to-face interviews and a high response rate. Second, the core GSS questionnaire contains questions about most of the important demographic and socioeconomic information that we would need to measure in order to carry out the proposed research. Third, the GSS contains a number of standard questions about social and political attitudes and behavior. Fourth, the GSS contains measures of social capital measured in terms of organizational affiliation, including membership in a church and membership in various voluntary civic and political associations. It also contains measures

of social trust. These variables are important potential predictors of a person’s actual level of bonding and bridging social capital as defined by Putnam. Fifth, the GSS is a highly visible survey. Its visibility would ensure that the results of our research are widely disseminated and would stimulate further research into both the core methodological and core substantive aspects of our proposed research. Sixth, the GSS has been used on two previous occasions (in 1985 and 2004) to collect more conventional network data (specifically, about five individuals with whom the respondent has discussed “important matters” in the past six months). While these previous efforts do not allow estimation of overdispersion because only a small portion of respondents’ total networks were assessed, we expect to gain important insights about the value of our proposed methodology by making comparisons of the relationship between networks and attitudes as it is measurable with “how many X’s” questions in comparison with the previous GSS approaches to social network measurement.

The topical module that we propose to include in the GSS would primarily consist of “How many X’s” questions. They would include the following questions.

- “How many X’s” questions with names (e.g., “How many persons do you know named Stephanie?”), to use for normalization and also possibly to learn about ethnic clustering.
- “How many X’s” questions with ethnic, occupational, political, and social categories. For example, doctors, truck drivers, partisan Democrats and Republicans, evangelical Christians, prisoners, immigrants, members of specific racial or ethnic groups, gay parents, etc.
- At least for some questions, we would further restrict the “How many X’s” questions to specific subsets of a person’s network. In these cases, we would ask about how many X’s do you know among respondent’s relatives, friends, neighborhood, workplace, or in the local community. This approach would allow separate estimates of overdispersion in specific subnetworks. It would also allow us to explore the role of geography in producing overdispersion. For example, regardless of the level of interactional segregation we found between whites and nonwhites, or between churchgoers and nonchurchgoers, it is important to establish the extent to which this segregation occurs because of the geographic clustering of social groups as opposed to the segregation that occurs after this geographic clustering is taken into account.
- For some questions involving ties to moderate-size groups, we will use subsetting strategies in order to obtain accurate information. One subsetting strategy, which was just described, is to restrict the question to specific aspects of a person’s network. Another subsetting strategy involves asking about intersections between two statuses. For example, the African-American population is too large for many people (especially African-Americans) to recall accurately the number they know. But they might more accurately recall the number of African-Americans named Michael that they knew. We will employ subsetting strategies such as these in order to assess the network degree involving ties to relatively large groups such as African-Americans or evangelical Christians. We discuss this issue more in Section 2.2.

• In addition to learning indirectly about social networks through overdispersion and regression models, we plan to ask some direct questions about polarization of respondents’ ego networks. For example, “Consider the political affiliations of your friends. Would you say they are mostly Democrats, more Democrats than Republicans, about evenly divided, more Republicans than Democrats, or mostly Republicans?” Or ask about liberal/conservative, or ask about friends’ views on particular issues (e.g., abortion, Iraq, minimum wage). A comparison between these answers and the model estimates from the “How many X’s” questions (described in more detail below) will provide important information about the utility of direct as opposed to indirect measures of network polarization.

To give a sense of the potential of our modeling strategy, we display in Figure 1 the estimated coefficients of a regression model fit to responses to the question asked in a national survey, “How many males do you know in state or federal prison?” Using methods described in Section 2, our analysis controlled for the estimated degree (total number of acquaintances) of respondents, so that these coefficients represent comparisons in the relative propensity to know a prisoner. It is no surprise that being male, nonwhite, young, unmarried, less educated, unemployed etc., are associated with knowing more males than expected in state or federal prison. However, the R^2 of the regression model with these

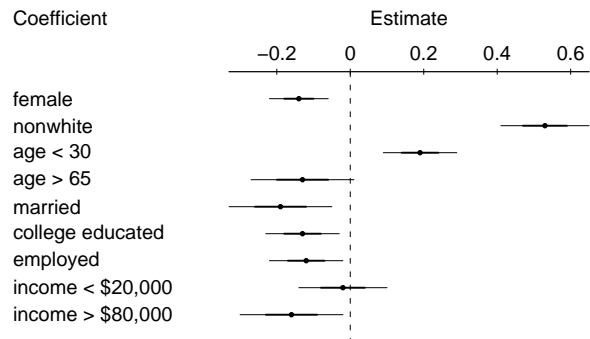


Figure 1: Coefficients of the regression of residuals for the “how many males in federal prison do you know” question on personal characteristics.

demographic predictors is only 11%. Our new survey will include questions on attitudes, occupation, and other factors, which along with geography should allow us to better understand the variation in propensity to know people of a given group.

In addition to measures of relative propensities and overdispersion of subgroups, there can be interest in the direct information on the proportion of the population who know 0, 1, 2, or more persons of different groups. For example, the contact hypothesis holds that contacts between members of the general population and a subgroup will tend to improve knowledge of and attitudes toward the subgroup (Williams 1947; see Jackson 1993 and Lee et al. 2004 for recent reviews). We can address this hypothesis by asking “How many X’s” along with relevant attitude questions, for example to see the changes associated with knowing two persons, compared to one, compared to none, of a particular group.

In summary, we plan to ask “How many X’s” questions about a range of subgroups of interest in order to study network polarization. We will also directly ask questions about attitudes and relevant background variables, and be able to use these along with geographic and demographic predictors to understand network polarization and its relationship with attitude polarization.

2 Models, data analysis, and design

We now turn to our proposed research in the design and analysis of “How many X’s” data. We first lay out our models and ideas for data analysis, illustrating with the survey of McCarty et al. , which introduced this type of survey. In particular, we show how this sort of data can be used to learn about overdispersion of groups in the social network. We then discuss research issues in designing such surveys, along with practical issues in implementing our own proposed study.

2.1 Analyzing “How many X’s” surveys to learn properties of the social network

“How many X’s” surveys were introduced by Killworth, McCarty, et al. as a tool for estimating (1) the distribution of individuals’ network size, defined to be the number of acquaintances, in U.S. population (this could also be called the *degree distribution*) and (2) the sizes of certain subpopulations, especially those that are hard to count using regular survey results (Killworth et al. 1998a,b). In this section, we sketch the proposed sophisticated analysis of such data to learn about properties of individuals and groups in the social network. We **demonstrate the feasibility of our modeling approach by reanalyzing data** from the survey conducted by McCarty et al. (2001): responses from 1370 adults in the United States (selected by random digit dialing) to questions of the form, “How many persons do you know⁶ in group X?”

We demonstrate that these data can be used to learn about the overdispersions of groups within the general social network, and to study how individual and community-level characteristics can predict the propensity of individuals to know persons in specified groups. This form of data collection is thus particularly appropriate for studying polarization in social positions and attitudes. The Killworth, McCarty, et al. surveys did not directly ask about groups defined in terms of social and political attitudes (e.g., liberals or conservatives), but we anticipate that similar analytical techniques will allow us to learn about these topics in our new survey.

Models of social connections between individuals and group members. We first introduce general notation for the links between persons i and j in the population⁷ (with groups k defined as subsets S_k of the population), with a total population size of N :

$$\begin{aligned}
 p_{ij} &= \text{probability that person } i \text{ knows person } j, & (1) \\
 a_i &= \sum_{j=1}^N p_{ij} = \text{the “gregariousness” (expected degree) of person } i \\
 B &= \sum_{i=1}^N a_i = \text{expected total degree of the population} = 2 \cdot (\text{expected \# links}) \\
 B_k &= \sum_{i \in S_k} a_i = \text{expected total degree of persons in group } k \\
 b_k &= B_k/B = \text{prevalence parameter or the proportion of total links that involve group } k \\
 \lambda_{ik} &= \sum_{j \in S_k} p_{ij} = \text{expected number of persons in group } k \text{ known by person } i \\
 g_{ik} &= \lambda_{ik}/(a_i b_k) = \text{individual } i\text{'s relative propensity to know a person in group } k.
 \end{aligned}$$

With complete network data, it would make sense to model the probabilities p_{ij} given properties of person

⁶The respondents were told, “For the purposes of this study, the definition of knowing someone is that you know them and they know you by sight or by name, that you could contact them, that they live within the United States, and that there has been some contact (either in person, by telephone or mail) in the past two years.” We plan to consider alternative definitions, as discussed on page 13 of this proposal.

⁷We implicitly assume the acquaintanceship to be symmetric, which is consistent with the wording of the survey question.

i , person j , and interactions of the dyad (see Handcock and Jones 2004, Hoff 2003, and Hoff, Raftery, and Handcock 2002). From a random sample of individuals, however, we can estimate the propensities g_{ik} of person i to know a person in group k . Such an analysis gives us flexibility, in particular in that we can learn about rare groups in the population even without having many (or even any) of them in our sample. (For example, McCarty et al. (2001) ask about prisoners and fatal accident victims, two groups which will not be reached at all by a telephone survey.)

The parameter b_k is not the proportion of *persons* in the population who are in group k ; rather, b_k is the proportion of *links* that involve group k in the acquaintance network (for this purpose, counting a link twice if it connects two members of group k). If the links are assigned completely at random, then $b_k = N_k/N$, while N_k is the number of individuals in group k . Realistically, and in our model, the values of b_k will not be proportional to the N_k 's. If b_k is higher than the population proportion of group k , it indicates that individuals from group k are more popular than an average person or that the trait defining group k is more visible or easier to recall. A careful inspection reveals that g_{ik} is the ratio of the proportion of the links that involve group k in individual i 's network, divided by the proportion of the links that involve group k in the population network. If $g_{ik} > 1$, then one would expect the percentage of acquaintances with people from group k will be higher in individual i 's personal network, compared to the population average. This is why we have termed g_{ik} the *relative propensity*.

We use the following notation for our survey data: n survey respondents and K population subgroups under study; for the McCarty et al. data, $n = 1370$ and $K = 32$. We label the individual responses as y_{ik} : y_{ik} = number of persons in group k known by person i . We consider three models that can be written in statistical notation as $y_{ik} \sim \text{Poisson}(\lambda_{ik})$, with increasingly general forms for λ_{ik} :

$$\begin{aligned} \text{Erdős-Renyi model:} & \quad \lambda_{ik} = ab_k \\ \text{our null model:} & \quad \lambda_{ik} = a_i b_k \\ \text{our overdispersed model:} & \quad \lambda_{ik} = a_i b_k g_{ik}. \end{aligned}$$

The Erdős and Renyi (1959) random graph model is the mathematical model for completely random formed acquaintances, which leads to an equal expected degree for all individuals. The null model goes beyond the Erdős-Renyi model by allowing the gregariousness parameters to differ between individuals (a_i) and prevalence parameters between groups (b_k). The overdispersed model generalized further by allowing different individuals to differ in their relative propensities to form ties to people in specific groups (g_{ik}). As described below, when fitting the overdispersed model, we do not attempt to estimate all the individual g_{ik} 's; rather, we estimate certain properties of their distributions.

For each subpopulation k , we let the multiplicative factors g_{ik} follow a gamma distribution with a value of 1 for the mean and a value of $1/(\omega_k - 1)$ for the shape parameter, which yields,

$$\text{overdispersed model: } y_{ik} \sim \text{Negative-binomial}(\text{mean} = a_i b_k, \text{overdispersion} = \omega_k). \quad (2)$$

Setting $\omega_k = 1$ corresponds to setting the shape parameter in the gamma distribution to ∞ , which in turn implies that the g_{ik} 's have zero variance, reducing to the null model. Higher values of ω_k correspond to overdispersion—that is, more variation in the distribution of connections involving group k than would be expected under the Poisson model, as would be expected if there is variation among respondents in the relative propensity to know someone in group k .

Our primary goal in fitting model (2) is to estimate the overdispersions ω_k and thus learn about the “nonrandomness” that exist in the formation of social networks. With a hierarchical (multilevel) model and Bayesian inference (see, e.g., Snijders and Bosker 1999, Raudenbush and Bryk 2002, and Gelman et al. 2003), we also estimate the gregariousness parameters a_i 's and the group prevalence parameters b_k 's.

We fit the negative binomial model to the McCarty et al. (2001) data, achieving approximate convergence ($\hat{R} < 1.1$; see Gelman et al. 2003) of three parallel chains after 2000 iterations. We present our inferences for the gregariousness parameters a_i , the prevalence parameters b_k , and the overdispersion parameters ω_k , in that order. We fit the model first using all the data and then separately for the male and female respondents.

The distribution of social network sizes a_i . The estimation of the distribution of social network sizes, the a_i 's, has troubled researchers for some time. Good estimates of this basic social parameter have remained elusive despite numerous efforts. Some attempts have included diary studies (Gurevich 1961, Pool and Kochen 1978), phone book studies (Pool and Kochen 1978, Freeman and Thompson 1989, Killworth et al. 1990), the reverse small-world method (Killworth and Bernard 1978), and the summation method (McCarty et al. 2001).

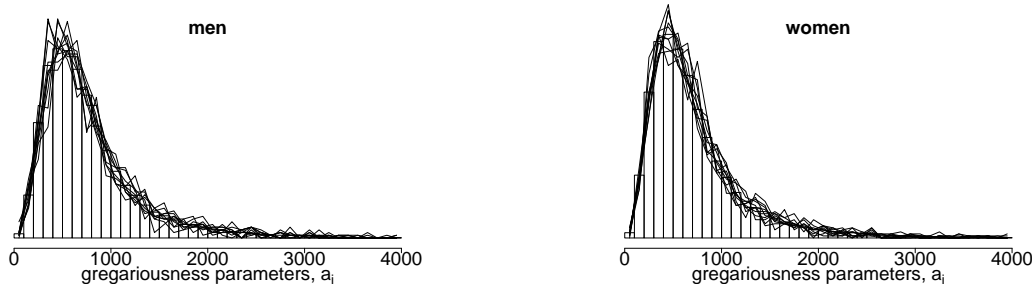


Figure 2: Estimated distributions of “gregariousness” or expected degree, a_i , from the fitted model. The overlain lines are posterior simulation draws indicating inferential uncertainty in the histograms.

Despite a large amount of work, this body of research offers little consensus. Our estimates of the distribution of the a_i 's, which extend the ideas of Killworth et al. (1998a,b), shed further light on this question of estimating the degree distribution of the acquaintanceship network. Further, we are able to go beyond previous studies by using our statistical model to summarize the uncertainty of the estimated distribution, as shown in Figure 2. Future researches are proposed in the next section on how to calibrate the estimation of the degree distribution through a more thoroughly designed survey.

Figure 2 displays estimated distributions of the gregariousness parameters a_i for the survey respondents, showing separate histograms of the posterior simulations from the model estimated separately to the men and the women. The similarity between the distributions for men and for women is not an artifact of our analysis but instead seems to be telling us something interesting about the social world. The spread in each of the histograms of Figure 2 almost entirely represents population variability. The model allows us to estimate the individual a_i 's to within a coefficient of variation of about $\pm 25\%$. When taken together this allows us to estimate the distribution very precisely. This precision can be seen in the solid lines overlain on Figure 2 that represent inferential uncertainty.

Figure 3 presents a simple regression of some of the factors predictive of $\log a_i$, using the data on the respondents in the McCarty et al. survey. These predictors are relatively unimportant in explaining social network size: the regression summarized in Figure 3 has an R^2 of only 10%. The strongest patterns are that persons with a college education, a job outside the home, and high incomes know more people, and persons over 65 and those having low incomes know fewer people. These factors all have effects in the range of 10%–20%.

We now consider the group-level parameters. The left panels of Figure 4 show the subpopulations k and the estimates of b_k , the proportion of links in the network that go to group k . The right panel shows the estimated overdispersions ω_k . The sample size is large enough that the 95% error bars are tiny for the b_k 's and reasonably small for the ω_k 's as well. (It is a general property of statistical estimation that mean parameters (such as the b_k 's in this example) are easier to estimate than dispersion parameters such as the ω 's.)

Relative sizes b_k of subpopulations. Considering the b_k 's first, the clearest pattern in Figure 4 is that respondents of each sex tend to know more people in groups of their own sex. We can also see that the 95% intervals are wider for groups with lower b_k 's, which makes sense since the data are discrete, and for these groups, the counts y_{ik} are smaller and provide less information.

Another pattern in the estimated b_k 's is the way that they scale with the size of group k . One would expect an approximate linear relation between the number of people in group k and our estimate for b_k : that is, on a graph of $\log b_k$ vs. $\log(\text{group size})$, we would expect the groups to fall roughly along a line with slope 1. However, as can be seen in Figure 5, this is not the case. Rather, the estimated prevalence increases approximately with square root of population size, a pattern that is particularly clean for the names. This relation has also been observed by Killworth et al. (2003).

Discrepancies from the linear relation can be explained by difference in average degrees (e.g., as members

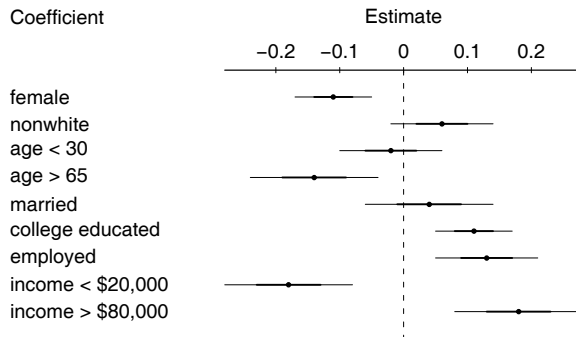


Figure 3: Coefficients of the regression of estimated log gregariousness parameters $\log a_i$ on personal characteristics.

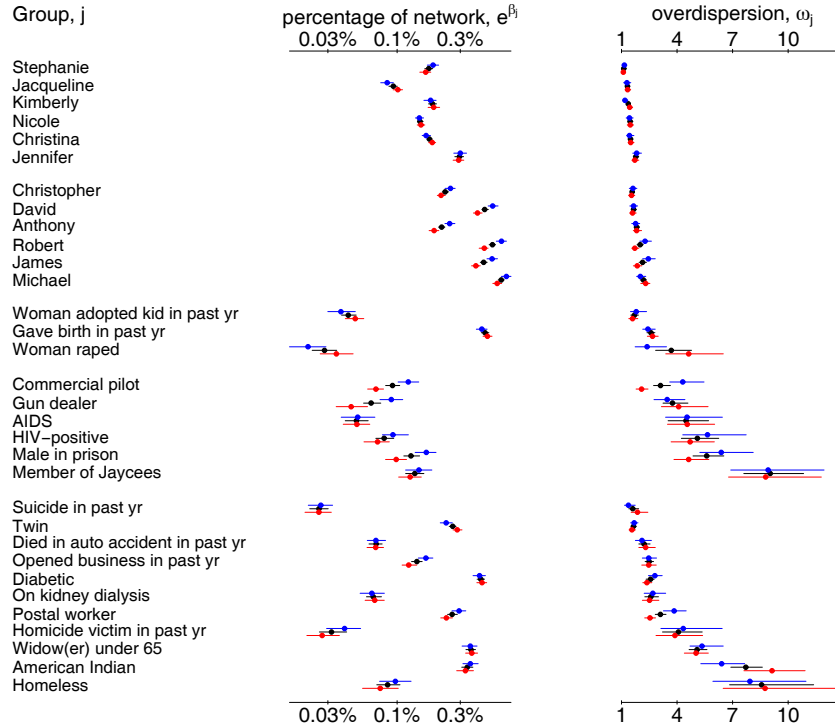


Figure 4: For each group k in the survey of McCarty et al. (2001), we plot the estimate (and 95% interval) of b_k and ω_k . The estimates and uncertainty lines are clustered in groups of three; for each group, the top (blue), middle (black), and bottom (red) dots/lines correspond to men, all respondents, and women, respectively. The groups are listed in categories—female names, male names, female, male (or primarily male), and mixed-sex groups—and in increasing average overdispersion within each category.

of a social organization, Jaycees would be expected to know more people than average, so their b_k should be larger than another group of equal numbers), inconsistency in definitions (e.g., what is the definition of an American Indian?), and ease or difficulty of recall (e.g., a friend might be a twin without you knowing it, whereas you would probably be know whether she gave birth in the last year).

This still leaves unanswered the question of why square root (i.e., a slope of $1/2$ in the log-log plot), rather than linear (a slope of 1)? Killworth et al. (2003) discuss various explanations for this pattern. As they note, it is easier to recall rare persons and events, whereas more people in more common categories are easily forgotten. You will probably remember every Ulysses you ever met, but it can be difficult to recall all the Michaels and Roberts you know even now.

This reasoning suggests that acquaintance networks are systematically underestimated, and hence when social network size is estimated in this way (derived from McCarty et al. 2001), it is more appropriate to normalize based on the known populations of the rarer names (e.g., Jacqueline, Nicole, and Christina in this study) rather than on more common names such as Michael or James, or even on the entire group of twelve names in the data.

Another pattern in Figure 5 is that the line for the names is higher than for the other groups. We suppose that is because, for a given group size, it is easier to recall names than characteristics. After all, you know the names of most your acquaintances, but you could easily be unaware that a friend has diabetes, for example.

Overdispersion parameters ω_k for subpopulations. The overdispersion is intended to estimate the variability in respondents' relative propensities to form ties to members of different groups. For groups where $\omega_k = 1$, we can conclude that there is no variation in these relative propensities, so that persons in group k appear to be randomly distributed in the social network. However, for groups where, ω_k is much greater than 1 , the null model is a poor fit to the data, and persons in group k do not appear to be uniformly distributed in the social network.

The right panel of Figure 4 displays the estimated overdispersions ω_k , and they are striking. First, we observe that the names have overdispersions of between 1 and 2 —that is, indicating very little variation in relative propensities. In contrast, the other groups have a wide range of overdispersions, ranging from near

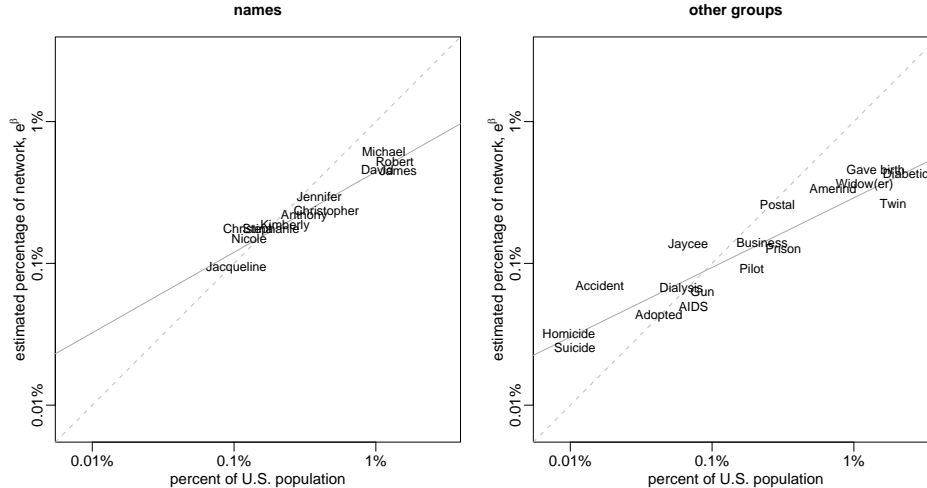


Figure 5: Log-log plot of estimated prevalence of groups in the population (as estimated from the “How many X’s” survey) plotted vs. actual group size (as determined from public sources). Names and other groups are plotted separately, on a common scale, with fitted regression lines shown. The solid lines have slopes 0.53 and 0.42, compared to a theoretical slope of 1 (as indicated by the dotted lines) that would be expected if all groups were equally popular, on average, and equally recalled by respondents.

1 for twins (which are in fact distributed nearly at random in the population), to 2–3 for diabetics, recent mothers, new business owners, and dialysis patients (who are broadly distributed geographically and through social classes), higher values for more socially localized groups such as gun dealers and HIV/AIDS patients (and demographically localized groups such as widows/widowers), and even higher values for Jaycees and American Indians, two groups with dense internal networks. Overdispersion is highest for homeless persons, who are both geographically and socially localized.

These results are consistent with our general understanding and also potentially reveal patterns that would not be apparent without this analysis. For example, it is no surprise that there is high variation in the propensity to know someone homeless, but it is perhaps surprising that AIDS patients are less overdispersed than HIV-positive persons, or that new business owners are no more overdispersed than new mothers.

Analysis using residuals. Further features of these data can be studied using residuals from the overdispersed model. A natural object of study is correlation: for example, do people who know more Anthonys tend to know more gun dealers (after controlling for the fact that social network sizes differ)? For each survey response y_{ik} , we can define the standardized residual as $r_{ik} = \sqrt{y_{ik}} - \sqrt{a_i b_k}$, the excess people known after accounting for individual and group parameters. (It is standard to compute residuals of count data on the square root scale to stabilize the variance; see Tukey 1972.)

For each pair of groups k_1, k_2 , we can compute the correlation of their vectors of residuals; several patterns can be seen in the correlation matrix (not shown). First, there is a slight positive correlation within male and female names. Second, perhaps more interesting sociologically, there is a positive correlation between the categories can be considered negative experiences—homicide, suicide, rape, died in a car accident, homelessness, and being in prison. That is, someone with a higher relative propensity to know someone with one bad experience is also likely to have a higher propensity to know someone who had a different bad experience. The strength of this correlation is a potentially interesting measure of inequality.

In addition to correlations, one can model the residuals based on individual predictors. For example, Figure 1 on page 5 shows the estimated coefficients of a regression model fit to the residuals of the null model for the “how many males do you know in state or federal prison” question. We used this example in Section 1.3 to demonstrate the potential for such survey questions for studying the determinants of overdispersion of isolated subpopulations. As with the correlation analysis, by performing this regression on the residuals and not the raw data, we are able to focus on the relative number of prisoners known, without being distracted by the total network size of each respondent (which we have separately analyzed in Figure 3).

Analysis using hierarchical regression models. The model presented so far includes network properties of gregariousness, subgroup prevalences, and overdispersion of groups, but does not explicitly relate these to individual characteristics. We have shown some results by partitioning the data (separately analyzing men and women; see Figures 2 and 4) and by post-processing (running regressions of the estimated parameters α_i

and residuals r_{ik} on individual-level predictors; see Figures 1 and 3); however it would be more statistically efficient to include these as part of a larger model.

We propose to extend the overdispersed model $\lambda_{ik} = a_i b_k g_{ik}$ (see (2)) in a hierarchical regression framework given P individual-level predictors X_i :

$$\alpha_i = \log a_i = X_i \phi + \eta_i, \quad \gamma_{ik} = \log g_{ik} = X_i \psi_k + \epsilon_{ik}. \quad (3)$$

Here, ϕ is a vector of regression coefficients for modeling gregariousness, and, for each group k , ψ_k is a vector of coefficients for modeling relative propensities. This requires the estimation of $P \times (K + 1)$ regression coefficients, which we will model hierarchically. The model can also be written, extending (2), as

$$y_{ik} \sim \text{Negative-binomial}(\text{mean} = e^{X_i \phi + \eta_i + \beta_k + X_i \psi_k}, \text{overdispersion} = \omega_k).$$

The parameter ω_k represent residual overdispersion in group k , beyond what is explained by the predictors X .

At the very least, such a regression model would be an improvement upon separate regressions such as in Figures 1 and 3. The real gain, however, would come from including predictors that go beyond mere demographics to include information such as occupation, community characteristics, and social and political attitudes, that relate to social networks and can be included in a multilevel model (as was done, for example, by Sampson et al. 1997). The result will be a fairly elaborate model (with all these coefficients for each of K groups) for which we will use up-to-date methods in fitting, understanding, and checking multilevel Bayesian models (e.g., Gelman 2004a,b, 2005a,b).

We will be interested in the regression coefficients, the residual overdispersion ω_k , and the total overdispersion from the original model. For example, suppose that the group defined by having the name Jose shows overdispersion in the simple model, but with the overdispersion disappearing after controlling for respondent ethnicity and state of residence. Then it would be interesting that there is no residual overdispersion, but would also be relevant that the group of persons named Jose is overdispersed in the social network as a whole. This sort of analysis is similar to what is done in studying racial and ethnic segregation. To what extent are different groups separated in the social network, even after adjusting for their geographic segregation?

To evaluate theories of attitudes in social networks, we propose to extend this model by fitting linear and logistic regressions of attitudes on network parameters, including gregariousness, relative propensities to know persons of certain groups, and actual counts of number of persons of group X known. For example, the contact hypothesis (Lee et al. 2004) could manifest itself as a positive effect of the number of persons of group X , after controlling for gregariousness and relative propensity. We will incorporate such models into the hierarchical regression framework, thus treating predictors such as gregariousness and relative propensity as missing data, by analogy to measurement-error models in regression.

We plan to perform the computations using an adaptive Metropolis algorithm implemented in Umacs, a set of R functions that is the subject of our concurrent research (Kerman and Gelman 2004). Compared to alternative software such as Bugs and MLwin, using our own R program gives us more flexibility in implementing a reasonably fast algorithm that can handle the large number of parameters that arise in hierarchical models with individual-level parameters. R is open-source, and we will make our software available as an R package, as we have done before (Sturtz, Ligges, and Gelman 2005). Specific techniques we plan to use to speed the algorithm include redundant additive and multiplicative parametrization (Bafumi, Gelman, and Park 2005) and adaptive scaling of Metropolis jumping rules (Pasarica and Gelman 2004). Specific techniques we plan to use to understand the fitted model include Bayesian analysis of variance (Gelman 2005a, Gelman and Pardoe 2005a), average predictive effects (Gelman and Pardoe 2005b). We will evaluate and compare models using graphical predictive checks (Gelman, 2004a), Bayes factors (Raftery, 1995), and predictive error measures (Spiegelhalter et al. 2002).

Other application areas. Sociologists are not the only scientists interested in the structure of networks. Methods presented here can be apply to a more generally defined network, as any set of objects (nodes) connected to each other by a set of links (edges). Areas of application include epidemiology, science and technology (collaboration networks, the Internet backbone, the World Wide Web, the power grid), and biological networks (metabolic networks, protein interaction networks, neural networks, the food web); for examples and reviews see Morris (1993, 2004), Morris and Kretzchmar (1995), Strogatz (2001), Newman (2002, 2003a,b), Watts, Dodds, and Newman (20002), and Watts (2002, 2004). We anticipate that the methods we develop for analyzing social networks using random sampling of nodes will be applicable to other network studies where

subgroups of nodes are of interest, or where it is practical or desirable to study groups of nodes by looking at the characteristics of their immediate neighbors.

2.2 Design and implementation of “How many X’s” surveys

Full or partial data: a recall dilemma. Our approach relies crucially on having count data, so that we can measure departures from our null model of independent links, hence the Poisson model on counts. However, several previous studies have been done in which only dichotomous data were collected. Examples include position generator studies (for a review, see Lin 1999) and resource generator studies (Van der Gaag and Snijders 2003) both of which attempt to measure individual-level social capital. In these studies, respondents were asked if they know someone in specific category—either an occupational group (doctor, lawyer, etc.) or resource group (someone who knows how to fix a car, someone who speaks a foreign language), and responses are dichotomous. Since several studies have collected data using these instruments, it would be helpful to be able to use such data to estimate the variation in popularities of individuals, groups, and overdispersions of groups—the a ’s, b ’s, and ω ’s in our model.

First, with some additional information about the group prevalence, it would be possible from binary response data to learn about overdispersion, which is identified by having too many high counts given the mean. In fact, the two-way structure in the data can be used to estimate overdispersion from mere yes/no data, given reasonable estimates of b_k ’s. However, good informative estimates of b_k are not always available. Without them, estimates from binary data are found to be extremely noisy and not particularly useful.

More encouragingly, if questions are asked of the form, “Do you know 0, 1, or more than 1 person named Michael?”, it is straightforward to fit the overdispersed model from these censored data, with the only change being in the likelihood function. If y_{ik} is the number of acquaintances in group k known by person i , we can write the censored data as $z_{ik} = 0$ if $y_{ik} = 0$, 1 if $y_{ik} = 1$ and 2 if $y_{ik} \geq 2$. The likelihood for z_{ik} is then simply the negative binomial density at 0 and 1 for the cases $z_{ik} = 0$ and 1, and $\Pr(z_{ik} \geq 2) = 1 - \sum_{m=0}^1 \Pr(y_{ik} = m)$ for $z_{ik} = 2$, the “2 or more” response, with the separate terms computed from the negative binomial density. (An alternative approach would be to treat the original y_{ik} ’s as missing data and use data augmentation (Tanner and Wong 1987), but in this case it is simpler to just compute the partial sums of the negative binomial distribution.) A multiple-choice question would capture less information than an exact count but would perhaps be less subject to the recall biases discussed in Section 2.1.

To illustrate the fitting of the model from partial information, we artificially censor the McCarty et al. (2001) data, creating a “yes/no” dataset (converting all responses $y_{ik} > 0$ to yeses), a “0/1/2/3+” dataset, and a “0/1/2/3/4/5+” dataset, fitting the appropriate censored-data model to each, and then comparing the parameter estimates to those from the full dataset. Censoring at 3 or 5 preserves much but not all of the information for estimation of b_k and ω_k , while censoring at 1 (yes/no data) gives reasonable estimates for the b_k ’s but nearly useless estimates for the ω_k ’s. In addition to having wider confidence intervals, the estimates from the censored data differ in some systematic ways from the complete-data estimates. Most notably, the overdispersion parameters ω_k are generally lower when estimated from censored data, suggesting some problems with the negative-binomial model.

In general, then, any organization performing a “How many X’s” survey has the choice of where to censor: 0/1/2+, 0/1/2/3+, 0/1/2/3/4+, . . . , or simply to ask the total number as was done by McCarty et al. (2002). As the censoring point rises, more is learned but more of the survey respondent’s time is taken, thus reducing the total number of questions that can be asked in the survey. As part of this proposal, we plan to perform a study to provide general recommendations on this issue: the theoretical/computational part of the study will determine the statistical efficiency of estimates based on data from different censoring points, and the empirical part will measure the speed and accuracy of survey respondents when the question is asked in different ways.

Names and normalization to estimate the degree distribution. As learned by Killworth, McCarty et al. (2003), the estimated average degree depends on which names are used for the normalization step: using more common names such as Michael and James gives estimates of around 300, whereas using rarer names such as Jacqueline and Stephanie gives estimates of around 700 (as in Figure 2). The discrepancy arises not because of the sex of the names but because respondents seem to search deeper into their social networks when asked about rarer names. (The discrepancy also does not appear to arise simply from difficulty with high count values y_{ik} , since it also comes up in the analysis of the data censored at 0/1/2/3+.)

The use of names to normalize also requires specific names to be chosen. McCarty et al. used names that do not have obvious ethnic connections (and the lack of overdispersion in these data imply that these names are indeed close to randomly distributed in the population), but this choice is not necessary, especially for

collecting data to be analyzed by a regression including ethnic and geographic information on respondents. A split-sample design is also possible, where different sets of names are assigned randomly to respondents.

Sizes of groups to be studied. As noted earlier, one of the advantages of a “How many X’s” survey is its ability to learn about groups representing less than 1%, even less than 0.1%, of the general population (see Figure 4). In general, the optimal design for such a survey will depend on group size. For example, if group X represents 10% of the population, we would not want to ask, “Do you know 0, 1, or 2+ members of group X?”, since almost all the respondents would probably answer “2+” (unless it is an extreme example of overdispersion). In this case, one would either have to ask about higher numbers or, probably better, to restrict the question in some way. For example, composite questions: How many black people do you know named Michael, etc? How many black teachers do you know? Or restricting the universe: How many black people did you speak with in the past week? These issues are important since some of the groups in which we are interested (for example, African Americans, Latinos, Democrats, Republicans, evangelical Christians, gays) represent 10% or more of the population.

Efficiency, sample size, and power calculations. To determine statistical efficiency, power, and related issues, we have to define what’s being studied. Any given survey will have sufficient power to answer some questions but not others. Estimands of interest here include the degree distribution, group sizes and overdispersions, and the coefficients ϕ and ψ in the regression model (3). As part of this project, we plan to perform simulations based on the McCarty et al. data (and then again based on our own survey) to determine estimation uncertainties for these quantities as a function of sample size. Future researchers using these tools will then be able to determine how large a sample size will be necessary to estimate effects of a specified size.

Definitions and scope of “knowing.” We will define the state of “knowing” someone more precisely in the survey. We will pilot-test different prompts similar to that used by Flap et al. (2004) such as the following. “In the following questions, we will ask about different types of people that you know. By ‘know’ we mean that you know the person’s name and that you would stop for at least a moment and talk to if you happened to meet the person (on the street, in a shopping mall, etc.).” We will ask about different depths of the social network by varying the extent of recalling. For example, instead of asking “How many people do you know who are living together but are not married,” we would specify the allowable answers as “Is it zero, one, two, or three or more?” Responses to such questions still allow estimates of the key parameters.

Two concerns need to be addressed in our pilot study: (1) the definition should appropriately include the right scope of acquaintances that is mostly related to both network and attitude polarization, and (2) the definitions should be clear to the respondents, especially in how far the respondent should recall into their social networks. We will use our pretest to decide what depth of the social network to ask about (e.g., zero to two or more, or zero to 5 or more X’s). It will also allow us to decide how to specify the scope of the questions, e.g. to people you talked with last week, or close friends, people who live nearby, people at work, etc. Our decisions will be based in part on the relative accuracy of different question strategies, and partly on the value of alternative scope conditions for providing the data needed to estimate key model parameters.

Implementation of the survey. We plan to collect the data described above by adding approximately 50 questions to the 2006 wave of the GSS (these questions will be asked of the 1500 respondent panel of the GSS that contains the core demographic and attitude questions). Incorporating our questions into the GSS has a number of advantages as stated in Section 1.3 on page 4.

We have identified a pool of questions for inclusion in the study as stated in Section 1.3 on page 4. During the summer of 2005, we will select the final set of questions to be included in GSS from that list and determine the appropriate response categories for each. As indicated above, we will also finalize the definition of “knowing” that is used in the social network module we are proposing to add. Piloting of the instrument will be conducted by the Social Indicators Survey Center (SISC), which Teitler directs, in close collaboration with the four investigators. SISC is part of Columbia University’s School of Social Work and is housed in a new building, shared with the Statistics Department. The pilot work will be implemented using RDD phone surveys and post-survey debriefing, in-person cognitive interviews, and focus groups.

While our request to add approximately 10 minutes of questions to the 2006 wave of GSS has been very favorably received (see attached letter of intent from Tom Smith), a final decision by the PI’s of the GSS will not be made until August, 2005. In the unlikely event that it is not possible for us to add questions to the next wave of GSS, we will collect data either by including questions to the next wave of the National Election Study, or by conducting our own survey, through SISC.

We plan to analyze the data using hierarchical regression models with overdispersion, as discussed at the end of Section 2.1. If GSS is used, the data will come from a cluster sample, which makes analysis more

complex but is ultimately an advantage because we can then learn about community-level predictors. The appropriate way to handle cluster sampling in a Bayesian analysis is to include cluster indicators and then model their regression coefficients hierarchically (Gelman et al. 2003, chapter 7).

Summary of potential benefits of our approach. Attitude polarization and a possible decline in social capital are important concerns in the United States today, and there is broad interest in the social sciences in studying these phenomena in the context of networks. “How many X’s” surveys, when analyzed using hierarchical generalized linear models, can be used to estimate the overdispersion of named or described subgroups and their relative propensities to be known by persons in the general population. Survey questions about attitudes, behaviors, and group membership can be used to study the factors predictive of overdispersion and differential propensities of knowing, and for certain subgroups it should be possible to distinguish network overdispersion from related phenomena such as homophily and geographic clustering.

Our proposal has several features that allow us to go beyond the extensive existing work on social capital and networks. We are able to learn about small groups (less than 1% or even less than 0.1% of the population; see Figure 4). We are able to learn about groups that cannot be directly surveyed (for example, prisoners, small children, people who have recently died). We can learn about groups defined by others’ perceptions of them (for example, vocal Democrats or Republicans). Using multilevel regression modeling with demographic and geographical information, we can study the determinants of overdispersion. By using different definitions of knowing, we can look at different depths of the social network. Our normalization process will give an estimate of individuals’ network sizes which can then be included as a predictor in the hierarchical regression.

In summary, this survey gathers information that would not easily be available from either a network sample or a typical population survey. By sampling 1500 people, each with an average of 750 acquaintances, we can learn information about 1 million people in the social network.

3 Results from prior NSF support

Gelman: The recent NSF grants of Gelman closest to this proposal are SES-99-87748, “Bayesian analysis of sample surveys” (2000–2003, \$254,675 total costs), PI’s Andrew Gelman and John B. Carlin, and “Multilevel modeling for the study of public opinion and voting” (2003–2006, \$214,914 total costs). Research on Bayesian multilevel (hierarchical) modeling that is relevant to the current proposal includes Gelman, Van Mechelen et al. 2004, Gelman 2004a,b, 2005a,b, Gelman and Pardoe 2005a,b, Pasarica and Gelman 2005). We have innovatively applied these ideas in areas including the study of voting and elections (Gelman and Huang 2005, Bafumi et al. 2005, Park, Gelman, and Bafumi 2004), Gelman, Katz, and Tuerlinckx 2002), the law (Gelman, Liebman, et al. 2004, Gelman, Fagan, and Kiss 2005), biology (Gelman, Chew, and Shnaidman 2004, Brochot et al. 2004), and psychometrics (Berkhof, Van Mechelen, and Gelman 2003, and Meulders et al. 2005). The key innovations of our approach to Bayesian data analysis are the explicit connections between modeling, model checking, and exploratory data analysis (Gelman, Meng, and Stern 1996, Gelman 2003, 2004). These steps are particularly relevant for the proposed research in which new models will need to be developed for studying social structure, especially given the justifiable skepticism that many researchers in empirical social science feel toward complex model-based inference. Also as part of our recent NSF support we have made a number of contributions in methods for sample surveys (Lu and Gelman 2003, Gelman, Stevens, and Chan 2003, Reilly, Gelman, and Katz 2001, Gelman and Carlin 2001).

DiPrete: DiPrete’s last four NSF grants (SBR 96-31944, SBR 94-11509, SES92-09159, and SES90-12619) have each involved dynamic analysis of social mobility over the life course involving either job, family, or financial status, using panel data from the U.S. and various western European countries. His recent work has focused more directly on family as an important basis for social inequality, and on the causes of variation and change in family forms and family behaviors. His most recent work on this topic has been funded by NICHD (R03 HD41035-01 and N01-HD-3-3354). DiPrete is currently part of a research group that has been awarded a multiyear contract from NICHD to develop a new interdisciplinary research program to advance scientific understanding about the factors and processes that produce family change in populations over time and that influence variation in family change and behavior across societies and groups defined by race, ethnicity, culture and gender. DiPrete is currently the head of the working group on variation and change in rates, timing, and type of union formation and dissolution for this project. Publications from his recent research include DiPrete (2005), DiPrete and Gangl (2004), DiPrete and Engelhardt (2004); DiPrete, Morgan, Engelhardt and Pacalova (2003), DiPrete (2002), DiPrete, Maurin and Goux (2002), DiPrete, Maurin, Goux, and Tahlin (2001), McManus and DiPrete (2001), and DiPrete and McManus (2000).

Teitler: Teitler’s research has been funded by the National Institutes of Health and other organizations

other than NSF. His recent and ongoing research focuses on studies of social policies and family behaviors, primarily using survey data (Reichman, Teitler, and Curtis, forthcoming; Teitler, Reichman, and Nepomnyaschy 2004; Teitler 2002, Teitler 2001, Teitler, and Weiss 2000). Teitler also conducts research on survey research methodology (Teitler, Reichman, and Sprachman 2003, Reichman et al. 2001, Cabrera et al. 2003), focusing particularly on reporting and non-response biases. His work in this area has challenged the survey research industry to think about tradeoffs in various types of procedures that can improve data quality, and was prominently referred to by the last two presidents of the American Association for Public Opinion Research in their presidential addresses (see Schulman 2003 and Martin 2004). Teitler has designed and implemented numerous surveys including the New York Social Indicators Survey (an ongoing telephone survey of New York City residents), the Fragile Families and Child Wellbeing study (a multi-mode ongoing panel study of mostly unwed parents), the Survey of Adults and Youth (a national survey of Adults and 10–18 year old youth in the US), and the Parent and Teen Survey (a study of teen sexual and fertility behavior in Philadelphia).

Zheng: The primary research area of Zheng has been in computational biology and statistical genetics. Supported by NIH grant R01 GM070789, Zheng developed novel methodologies to study the complicated association structure of biological systems (Lo and Zheng 2002, Lo and Zheng 2004), and reliability of studies on large numbers of objects (Zheng and Lo 2005). Zheng has been working on statistical modeling and inference for social networks in collaboration with Gelman recently. This project is the result of this newly-established collaboration.

4 Education and training related to the proposed work

NSF-funded research is intended ultimately to be disseminated to a larger audience. We have achieved that in several ways.

1. We have developed a new course in multilevel modeling, offered jointly by the Statistics and Political Science departments at Columbia. This course was highly rated and attracted doctoral and postdoctoral students from several departments and schools at the university. A book on regression and multilevel modeling aimed at the social science audience is completing (Gelman and Hill 2005).

2. We wrote an innovative book on teaching statistics (Gelman and Nolan 2002) that also resulted in a new approach to training graduate students in the teaching of statistics (Gelman 2005c). The book has received very positive reviews and is the subject of a session this spring at a workshop at Smith College on the teaching of quantitative methods in political science.

3. We recently developed an M.A. program in Quantitative Methods in Social Sciences at Columbia University. This program, jointly administered with the departments of Economics, History, Political Science, Psychology, and Sociology, has been highly successful. Now in its sixth year, it has been spun off to the university's Institute for Social and Economic Research and Policy.

The proposed project involves important and interesting research topics in statistics, political sciences, and sociology. The proposed project will involve sociology, social work, and statistics students in all phases of data collection. Furthermore, it will bring these students together in a collaborative multidisciplinary endeavor. Graduate students and postdoctoral fellows will be encouraged to develop or further their thesis research along the lines of the proposed work. At the statistics department of Columbia, undergraduates are encouraged to participate in ongoing faculty research projects, while all senior sociology majors are required to do either a one-semester research project or a two semester senior thesis. Drs. Gelman and Zheng will involve interested undergraduates from statistics or political science, and Dr. DiPrete will similarly involve interested undergraduates from sociology. All four of the senior researchers are connected with the QMSS program and will encourage students from this program to use the proposed project as a vehicle for masters theses. We are also involved with local schools. This year, a student working with survey data under the supervision of Dr. Gelman was a finalist in the Intel competition for high school student science projects. Teitler is currently mentoring another high school student as part of the Intel science competition. The Social Indicators Survey Center has an explicit pedagogical mission, which it achieves by providing new sources of data for instructional use, by involving students in data collection activities, and by providing internship opportunities of students at the School of Social Work. Furthermore, the Center is in the process of creating more formal linkages with the Qualitative Methods in the Social Sciences M.A. program, to provide students with hands-on experiences in survey methodology.