"How many people do you know in prison?": Using overdispersion in count data to estimate social structure in networks[*]

Tian Zheng[†]   Matthew J. Salganik[‡]   Andrew Gelman[§]

August 22, 2005

Abstract

Networks—sets of objects connected by relationships—are important in a number of fields. The study of networks has long been central to sociology, where researchers have attempted to understand the causes and consequences of the structure of relationships in large groups of people. Using insight from previous network research, Killworth et al. (1998a,b) and McCarty et al. (2001) developed and evaluated a method for estimating the sizes of hard-to-count populations using network data collected from a simple random sample of Americans. In this paper we show how, using a multilevel overdispersed Poisson regression model, these data can also be used to estimate aspects of social structure in the population. Our work goes beyond most previous research on networks by using variation, as well as average responses, as a source of information. We apply our method to the McCarty et al. data and find that Americans vary greatly in their number of acquaintances. Further, Americans show great variation in propensity to form ties to people in some groups (e.g., males in prison, the homeless, and American Indians), but little variation for other groups (e.g., twins, people named Michael or Nicole). We also explore other features of these data and consider ways in which survey data can be used to estimate network structure.

Keywords: negative binomial distribution, overdispersion, sampling, social networks, social structure

1 Introduction

Recently a survey was taken of Americans, asking, among other things, "How many males do you know incarcerated in state or federal prison?" The mean of the responses to this question was 1.0. To readers of this journal that number may seem shockingly high. We would guess that you probably don't know anyone

[†]Department of Statistics, Columbia University, New York, New York
[‡]Department of Sociology, Columbia University, New York, New York
[§]Department of Statistics and Department of Political Science, Columbia University, New York, New York

in prison. In fact, we would guess that most of your friends don't know anyone in prison either. This number may seem totally incompatible with your social world.

So how was the mean of the responses 1? According to the data, 70% of the respondents reported knowing zero people in prison. However, the responses show a wide range of variation, with almost 3% reporting that they know at least 10 prisoners. Responses to some other questions of the same format, for example "How many people do you know named Nicole?" show much less variation.

This difference in the variability of responses to these "how many X's do you know" questions is the manifestation of fundamental social processes at work. Through careful examination of this pattern, as well as others in the data, we can learn about important characteristics of the social network connecting Americans, as well as the processes that create this network.

This analysis also furthers our understanding of statistical models from two-way data, by treating overdispersion as a source of information not just an issue that requires correction. More specifically, we include overdispersion as a parameter that measures the variation in the relative propensities of individuals to form ties to a given social group, and allow it to vary among social groups. Through such modeling of the variation of the relative propensities, we derive a new measure of social structure that only uses survey responses from a sample of individuals not data on the complete network.

1.1   Background

Understanding the structure of social networks, and the social processes which form them, is a central concern of sociology for both theoretical and practical reasons (Wasserman and Faust, 1994; Freeman, 2004). Social networks have been found to have important implications for the social mobility (Lin, 1999), getting a job (Granovetter, 1995), the dynamics of fads and fashion (Watts, 2002), attitude formation (Lee et al., 2004), and the spread of infectious disease (Morris and Kretzchmar, 1995).

When talking about social networks, sociologists often use the word "social structure," which in practice has taken on many different meanings, sometimes unclear or contradictory. In this paper, as in Heckathorn and Jeffri (2001), we generalize the conception put forth by Blau (1974) that *social structure* is the difference in affiliation patterns from what would be observed if people formed friendships entirely randomly.

Sociologists are not the only scientists interested in the structure of networks. Methods presented here can be applied to a more generally defined network, as any set of objects (nodes) connected to each other by a set of links (edges). In addition to social networks (friendship network, collaboration networks of scientists, sexual networks), other examples include technological networks (the Internet backbone, the World Wide Web, the power grid) and biological networks (metabolic networks, protein interaction networks, neural networks, food webs); for reviews see Strogatz (2001), Newman (2003b), and Watts (2004).

1.2 Overview of this paper

In this paper we show how to use "how many X's do you know" count data to learn about the social structure of the acquaintanceship network in the United States. More specifically, we can learn to what extent people vary in their number of acquaintances, to what extent people vary in their propensity to form ties to people in specific groups, and also to what extent specific subpopulations (including those that are otherwise hard to count) vary in their popularities.

The data used in this paper were collected by McCarty et al. (2001) and consist survey responses of 1370 individuals on their acquaintances with groups defined by name (Michael, Christina, Nicole, . . .), occupation (postal worker, pilot, gun dealer, . . .), ethnicity (Native American), or experience (prisoner, auto accident victim, . . .); for a complete list of the groups see Figure 4. Our estimates come from fitting a multilevel Poisson regression with variance components corresponding to survey respondents and subpopulations and an overdispersion factor that varies by group. We fit the model using Bayesian inference and the Gibbs-Metropolis algorithm, and identify some areas in which the model fit could be improved using predictive checks. Fitting the data with a multilevel model allows individual and subpopulation effects to be separated. Our analysis of the McCarty et al. data gives reasonable results, which is a useful external check on our methods. Potential areas of further work include more sophisticated interaction models, application to data collected by network sampling (Heckathorn, 1997, 2002, Salganik and Heckathorn, 2004), and application to count data in other fields.

2 The problem and data

The original goals of the McCarty et al. surveys were (1) to estimate 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) to estimate the sizes of certain subpopulations, especially those that are hard to count using regular survey results (Killworth et al., 1998a,b).

The data from the survey are responses from 1370 adults (survey 1—796 respondents, January 1998; survey 2—574 respondents, January 1999) in the United States (selected by random digit dialing) to a series of questions of the form, "How many people do you know in group X?[1]"; see Figure 4 for a list of the 32

---

[1]The 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."

In addition, there are some minor complications with the data. For the fewer than 0.4% of responses that were missing, we followed the usual practice with this sort of unbalanced data of assuming an ignorable model (that is, constructing the likelihood using the observed data). Sometimes responses were categorized, and then we use the central value in the bin (for example, imputing 7.5 for the response "5–10"). To correct for some responses that were suspiciously large (for example, a person claiming to know over 50 Michaels), we truncate all responses at 30. (Truncating at value 30 affects 0.25% of the data. As a sensitivity analysis, we tried changing the truncation point to 50; this had essentially no effect on our results.) We also inspected the data using scatterplots of responses, which revealed a respondent who was coded as knowing 7 persons of every

groups asked about in the survey). In addition to the network data, background demographic information including sex, age, income, and marital status was also collected.

Killworth et al. (1998a,b) summarize the data in two ways. First, for their first goal of estimating the social network size for any given individual surveyed, they use his or her responses for a set of subpopulations with known sizes and scale up using the sizes of these groups in the population. To illustrate, suppose you know 2 persons named Nicole. At the time of the survey, there were 358,000 Nicoles out of 280 million Americans. Thus, your 2 Nicoles represent a fraction $\frac{2}{358,000}$ of all the Nicoles. Extrapolating to the entire country yields an estimate of $\frac{2}{358,000} \cdot (280 \text{ million}) = 1560$ people known by you. A more precise estimate can be obtained by averaging these estimates using a range of different groups. This is only a crude inference since it assumes that everyone has equal propensity to know someone from each group. However, as an estimation procedure, it has the advantage of not requiring a respondent to recall his or her entire network, which typically numbers in the hundreds (McCarty et al., 2001).

The second use for which this survey was designed is to estimate the size of certain hard-to-count populations. To do this, Killworth et al. (1998a,b) combined the estimated network size information with the responses to the questions about how many people the respondents know in the hard-to-count population. For example, the survey respondents know, on average, 0.63 homeless people. If it is estimated that the average network size is 750, then homeless people represent a fraction of $\frac{0.63}{750}$ of an average person's social network. The total number of homeless people in the country can then be estimated as $\frac{0.63}{750} \cdot (280 \text{ million}) = 0.24$ million. This estimate relies on idealized assumptions (most notably, that homeless persons have the same social network size, on average, as Americans as a whole) but can be used as a starting point for estimating the sizes of groups that are difficult to measure directly (Killworth et al., 1998a,b).

In this paper we demonstrate a new use of the data from this type of survey to reveal information about social structure in the acquaintanceship network. We use the variation in response data to study the heterogeneity of relative propensities for people to form ties to people in specific groups. Additionally, we provide support for some of the findings in McCarty et al. (2001) and Killworth et al. (2003).

---

category. We removed this case from the dataset.

3 Formulating and fitting the model

3.1 Notation

We introduce a general notation for the links between persons i and j in the population (with groups k defined as subsets S_k of the population), with a total population size of N:

$$p_{ij} = \text{probability that person } i \text{ knows person } j, \tag{1}$$

$$a_i = \sum_{j=1}^{N} p_{ij} = \text{gregariousness parameter or the expected degree of person } i,$$

$$B = \sum_{i=1}^{N} a_i = \text{expected total degree of the population} = 2\cdot(\text{expected \# link,})$$

$$B_k = \sum_{i \in S_k} a_i = \text{expected total degree of persons in group } k,$$

$$b_k = B_k/B = \text{prevalence parameter or the proportion of total links that involve group } k,$$

$$\lambda_{ik} = \sum_{j \in S_k} p_{ij} = \text{expected number of persons in group } k \text{ known by person } i,$$

$$g_{ik} = \lambda_{ik}/(a_i b_k) = \text{individual } i\text{'s relative propensity to know a person in group } k.$$

We are implicitly assuming acquaintanceship to be symmetric, which is consistent with the wording of the survey question (see footnote 1 on page 3). To the extent that the relation is not symmetric, our results still hold if we replace the term "degree" by "in-degree" or "out-degree" as appropriate.

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 (for this purpose, counting a link twice if it connects two members of group k). If the links in the acquaintance network are assigned completely at random, then $b_k = N_k/N$, where N_k is the number of individuals in group k. Realistically, and in our model, the values of b_k may not be proportional to the N_k's. If b_k is higher than the population proportion of group k, it indicates that the average degree of individuals from group k is higher than the average degree of the population.

For the parameter g_{ik}, a careful inspection reveals that

$$g_{ik} = \frac{\sum_{j \in S_k} p_{ij} / \sum_{j=1}^{N} p_{ij}}{b_k} \tag{2}$$

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. In other words, $g_{ik} > 1$ if individual i has higher propensity to form ties with individuals from group k than an average person in the population. This property of g_{ik} is why we have termed it 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 y_{ik} as the response[2] of individual i to the question, "How many people do you know of subpopulation k?"; that is, $y_{ik} = $ number of persons in group k known by person i. We now discuss three increasingly general models for the data y_{ik}.

---

[2]We implicitly assume, in this section, that the respondents have the perfect recall of the number of acquaintances in each subpopulation k. Issues of imperfect recalling are observed and discussed in section 4.2.

Erdös-Renyi model. We study social structure as departures from patterns that would be observed if the acquaintances are formed randomly. The classical mathematical model for completely randomly formed acquaintances is the *Erdös-Renyi model* (Erdös and Renyi, 1959), under which the probability p_{ij} of a link between person i and person j is the same for all pairs (i, j). Following the notation above, this model leads to equal expected degrees a_i for all individuals and relative propensities g_{ik} that all equal to 1. The model also implies that the set of responses y_{ik} for subpopulation k should follow a Poisson distribution.

However, if the expected degrees of individuals were actually heterogenous, then we would expect super-Poisson variation in the responses to "how many X's do you know" questions (y_{ik}). Since numerous network studies have found large variation in degrees (Newman, 2003b), it is no surprise the Erdös-Renyi model is a poor fit to our y_{ik} data—a χ^2 goodness-of-fit test values 350,000 on $1369 \times 32 \approx 44,000$ degrees of freedom.

Null model. To account for the variability in the degrees of individuals, we introduce a *null model* in which individuals have varying gregariousness parameters (or the expected degrees) a_i. Under this model, for each individual, the acquaintances with others are still formed randomly. However, the gregariousness may differ from individual to individual. In our notation, the null model implies $p_{ij} = a_i a_j / B$, and relative propensities g_{ik} are still all equal to 1. Departure from this model can be viewed as evidence of structured social acquaintance networks. A similar approach was taken by Handcock and Jones (2004) in their attempt to model human sexual networks, although their model is different because it does not deal with two-way data.

In the case of the "how many X's do you know" count data, this null model fails to account for much of social reality. For example, under the null model, the relative propensity to know people in prison is the same for a reader of this journal and a person without a high school degree. The failure of such an unrealistic model to fit the data is confirmed by a χ^2 goodness-of-fit test that values $160,000$ on $1369 \times 31 \approx 42,000$ df.

Overdispersed model. The failure of the null model motivates a more general model that allows individuals to vary not only in their gregariousness (a_i), but also in their relative propensity to know people in different groups (g_{ik}). We call this the *overdispersed model* since variation in these g_{ik}'s results in overdispersion in the "how many X's do you know" count data. As is standard in generalized linear models (e.g., McCullagh and Nelder, 1989), we use "overdispersion" to refer to data with more variance than expected under a null model, and also as a parameter in an expanded model that captures this variation.

Comparison of the three models using "How many X's do you know" counts data. Figure 1 shows some of the data—the distributions of responses, y_{ik}, to the questions, "How many people named Nicole do you know?" and "How many Jaycees do you know?", along with the expected distributions under the Erdös-Renyi model, our null model, and our overdispersed model. We chose these two groups to plot because they are close in average number known (0.9 Nicoles, 1.2 Jaycees) but have much different distributions; the distribution for Jaycees has much more variation, with more zero responses and more responses in the upper tail.[3]

---

[3] "Jaycees" are members of the Junior Chamber of Commerce, a community organization of people between the ages of 21
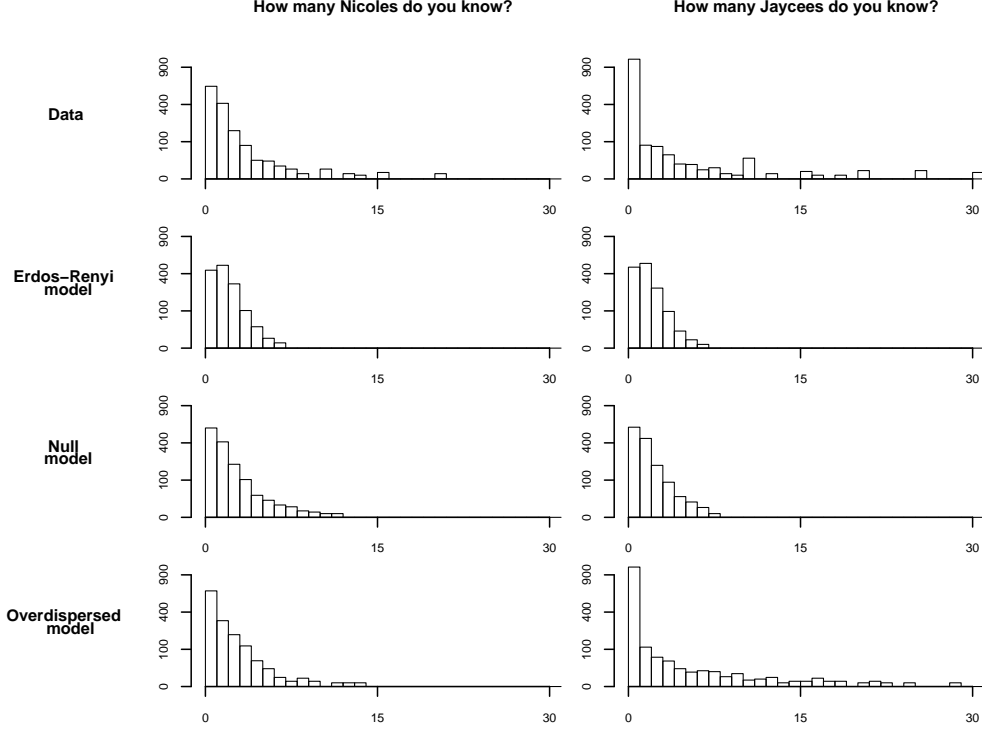
Figure 1: Histograms (on the square-root scale) of responses to "How many persons do you know named Nicole?" and "How any Jaycees do you know?" from the McCarty et al. data and from random simulations under three fitted models: the Erdös-Renyi model (completely random links), our null model (some people more gregarious than others, but uniform relative propensities for people to form ties to all groups), and our overdispersed model (variation in gregariousness and variation in propensities to form ties to different groups). Each model shows more dispersion than the one above, with the overdispersed model fitting the data reasonably well. The propensities to form ties to Jaycees show much more variation than the propensities to form ties to Nicoles, and hence the Jaycees counts are much more overdispersed. (The data also show minor idiosyncrasies such as small peaks at the responses 10, 15, 20, and 25. All values greater than 30 have been truncated at 30.) We display on square-root scale to more clearly reveal patterns in the tails.

The three models can be written as follows in statistical notation as $y_{ik} \sim \text{Poisson}(\lambda_{ik})$, with increasingly general forms for λ_{ik}:

$$\text{Erdös-Renyi model:} \quad \lambda_{ik} = ab_k$$
$$\text{our null model:} \quad \lambda_{ik} = a_i b_k$$
$$\text{our overdispersed model:} \quad \lambda_{ik} = a_i b_k g_{ik}.$$

Comparing the models, the Erdös-Renyi model implies a Poisson distribution for the responses to each "How many X's do you know" question, whereas the other models allow for more dispersion. The null model

and 39. Because the Jaycees are a social organization, it makes sense that not everyone has the same propensity to know one—people who are in the social circle of one Jaycee are particularly likely to know others.

7

turns out to be a much better fit to the Nicoles than to the Jaycees, indicating that there is comparably less variation in the propensity to form ties with Nicoles that with Jaycees. The overdispersed model fits both distributions reasonably well and captures the difference between the patterns of the acquaintance with Nicoles and Jaycees by allowing individuals to differ in their relative propensities to form ties to people in specific groups (g_{ik}). As we shall show, using the overdispersed model, both variation in social network sizes and variations in relative propensities to form ties to specific groups can be estimated from the McCarty et al. data.

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.

3.2   The overdispersed model

Overdispersion in these data can arise if the relative propensity for knowing someone in prison, for example, varies from respondent to respondent. We can write this in the generalized linear model framework as,

$$\text{overdispersed model:} \quad y_{ik} \sim \text{Poisson}(e^{\alpha_i + \beta_k + \gamma_{ik}}), \tag{3}$$

where $\alpha_i = \log(a_i)$, $\beta_k = \log(b_k)$ and $\gamma_{ik} = \log(g_{ik})$. In the null model, $\gamma_{ik} \equiv 0$. For each subpopulation k, we let the multiplicative factor $g_{ik} = e^{\gamma_{ik}}$ follow a gamma distribution with a value of 1 for the mean and a value of $1/(\omega_k - 1)$ for the shape parameter.[4] This distribution is convenient because then the γ's can be integrated out of (3) to yield,

$$\text{overdispersed model:} \quad y_{ik} \sim \text{Negative-binomial}(\text{mean} = e^{\alpha_i + \beta_k}, \text{overdispersion} = \omega_k). \tag{4}$$

(The usual parametrization (see, e.g., Gelman et al., 2003) of the negative-binomial distribution is $y \sim$ Neg-bin(A, B), but for this paper it is more convenient to express in terms of the mean $\lambda = A/B$ and overdispersion $\omega = 1 + 1/B$.) 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.

The overdispersion parameter ω can be interpreted in a number of ways. Most simply, it scales the variance: $\text{var}(y_{ik}) = \omega_k \text{E}(y_{ik})$ in the negative binomial distribution for y_{ik}. A perhaps more intuitive interpretation uses the probabilities of knowing exactly zero or one person in subgroup k. Under the negative binomial distribution for data y,

$$\Pr(y = 1) = \frac{\Pr(y = 0)\text{E}(y)}{\text{overdispersion}}. \tag{5}$$

---

[4]If we wanted, we could allow the mean of the gamma distribution to vary also; however, this would be redundant with a location shift in β_k; see (3). The mean of the gamma distribution for the $e^{\gamma_{ik}}$'s cannot be identified separately from β_k, which we are already estimating from data.

Thus, we can interpret the overdispersion as a factor that decreases the frequency of people who know exactly one person of type X, as compared to the frequency of people who know none. As overdispersion increases from its null value of 1, it is less likely for a person to have an isolated acquaintance from that group.

Our primary goal in fitting model (4) is to estimate the overdispersions ω_k and thus learn about biases that exist in the formation of social networks. As a byproduct, we also estimate the gregariousness parameters $a_i = e^{\alpha_i}$ representing the expected number of persons known by respondent i, and the group prevalence parameters $b_k = e^{\beta_k}$, which is the proportion of subgroup k in the social network.

We estimate the α_i's, β_k's, and ω_k's with a hierarchical (multilevel) model and Bayesian inference (see, e.g., Snijders and Bosker, 1999, Raudenbush and Bryk, 2002, and Gelman et al., 2003). The respondent parameters α_i are assumed to follow a normal distribution with unknown mean μ_α and standard deviation σ_α, which corresponds to a lognormal distribution for the gregariousness parameters $a_i = e^{\alpha_i}$. This is a reasonable prior given previous research on the degree distribution of the acquaintanceship network of Americans (McCarty et al., 2001). We similarly fit the group-effect parameters β_k with a normal distribution $N(\mu_\beta, \sigma_\beta^2)$, with these hyperparameters also estimated from the data. For simplicity, we assign independent Uniform(0,1) prior distributions to the overdispersion parameters on the inverse scale: $p(1/\omega_k) \propto 1$. (The overdispersions ω_k are constrained to the range $(1, \infty)$, and so it is convenient to put a model on the inverses $1/\omega_k$, which fall in $(0, 1)$.) The sample size of the McCarty et al. dataset is large enough that this noninformative model works fine; in general, however, it would be more appropriate to model the ω_k's hierarchically also. We complete the Bayesian model with a noninformative uniform prior distribution for the hyperparameters $\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta$. The joint posterior density can then be written as,

$$p(\alpha, \beta, \omega, \mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta | y) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \binom{y_{ik} + \xi_{ik} - 1}{\xi_{ik} - 1} \left(\frac{1}{\omega_k}\right)^{\xi_{ik}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{y_{ik}} \prod_{i=1}^{n} N(\alpha_i | \mu_\alpha, \sigma_\alpha^2) \prod_{k=1}^{K} N(\beta_k | \mu_\beta, \sigma_\beta^2),$$

where $\xi_{ik} = e^{\alpha_i + \beta_k}/(\omega_k - 1)$, from the definition of the negative binomial distribution.

Normalization

The model as given has a nonidentifiability. Any constant C can be added to all the α_i's and subtracted from all the β_k's, and the likelihood will remain unchanged (since it depends on these parameters only through sums of the form $\alpha_i + \beta_k$). If we also add C to μ_α and subtract C from μ_β, then the prior density also is unchanged as well. It would be possible to identify the model by anchoring it at some arbitrary point—for example, setting μ_α to zero—but we prefer to let all the parameters float, since including this redundancy can speed the Gibbs sampler computation (van Dyk and Meng, 2001).

However, in summarizing the model we would like to identify the α and β's so that each $b_k = e^{\beta_k}$ represents the proportion of the links in the network that go to members of group k. We identify the model in this way by renormalizing the b_k's for the rarest names (in the McCarty et al. survey, these are Jacqueline, Christina, and Nicole) so that they line up to their proportions in the general population. We renormalize to the rare names rather than to all 12 names because there is evidence that respondents have difficulty recalling

all their acquaintances with common names (see Killworth et al., 2003, and also Section 4.2 below). Finally, since the rarest names asked about in our survey are female names—and people tend to know more persons of their own sex—we further adjust by adding half the discrepancy between a set of intermediately-popular male and female names in our dataset.

This procedure is complicated but is our best attempt at an accurate normalization for the general population (which is roughly half women and half men) given the particularities of the data we have at hand. In the future, it would be desirable to gather data on a balanced set of rare female and male names. The left panel of Figure 5 illustrates how after renormalization, the rare names in the dataset have groups sizes equal to their proportion in the population. This specific procedure is designed for the recall problems that exist in the McCarty et al. dataset. Researchers working with different datasets may have to develop a procedure that is appropriate to their specific data.

In summary, for each simulation draw of the vector of model parameters, we define the constant

$$C = C_1 + \frac{1}{2}C_2, \tag{6}$$

where $C_1 = \log\left(\sum_{k \in G_1} e^{\beta_k}/P_{G_1}\right)$ adjusts for the rare girls' names, and $C_2 = \log\left(\sum_{k \in B_2} e^{\beta_k}/P_{B_2}\right) - \log\left(\sum_{k \in G_2} e^{\beta_k}/P_{G_2}\right)$ represents the difference between boys' and girls' names. In these expressions, G_1, G_2, and B_2 are the set of rare girls' names (Jacqueline, Christina, and Nicole), somewhat popular girls' names (Stephanie, Nicole, and Jennifer), and somewhat popular boys' names (Anthony and Christopher), and $P_{G_1}, P_{G_2}, P_{B_2}$ are the proportion of people with these groups of names in the U.S. population.

We add C to all the α_i's and to μ_α, and subtract it from all the β_k's and μ_β, so that all the parameters are uniquely defined. We can then interpret the parameters $a_i = e^{\alpha_i}$ as the expected social network sizes of the individuals i and the parameters $b_k = e^{\beta_k}$ as the sizes of the groups as a proportion of the entire network, as in the definitions (1).

3.3   Fitting the model using the Gibbs-Metropolis algorithm

We obtain posterior simulations for the above model using a Gibbs-Metropolis algorithm, iterating the following eight steps:

1. For each i, update α_i using a Metropolis step with jumping distribution, $\alpha_i^* \sim N(\alpha_i^{(t-1)}, (\text{jumping scale of } \alpha_i)^2)$.

2. For each k, update β_k using a Metropolis step with jumping distribution, $\beta_k^* \sim N(\beta_i^{(t-1)}, (\text{jumping scale of } \beta_k)^2)$.

3. Update $\mu_\alpha \sim N(\hat{\mu}_\alpha, \sigma_\alpha^2/n)$, where $\hat{\mu}_\alpha = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$.

4. Update $\sigma_\alpha^2 \sim \text{Inv-}\chi^2(n-1, \hat{\sigma}_\alpha^2)$, where $\hat{\sigma}_\alpha^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \mu_\alpha)^2$.

5. Update $\mu_\beta \sim N(\hat{\mu}_\beta, \sigma_\beta^2/n)$, where $\hat{\mu}_\beta = \frac{1}{K}\sum_{k=1}^{K} \beta_k$.

6. Update $\sigma_\beta^2 \sim \text{Inv-}\chi^2(K-1, \hat{\sigma}_\beta^2)$, where $\hat{\sigma}_\beta^2 = \frac{1}{K}\sum_{k=1}^{K}(\beta_k - \mu_\beta)^2$.
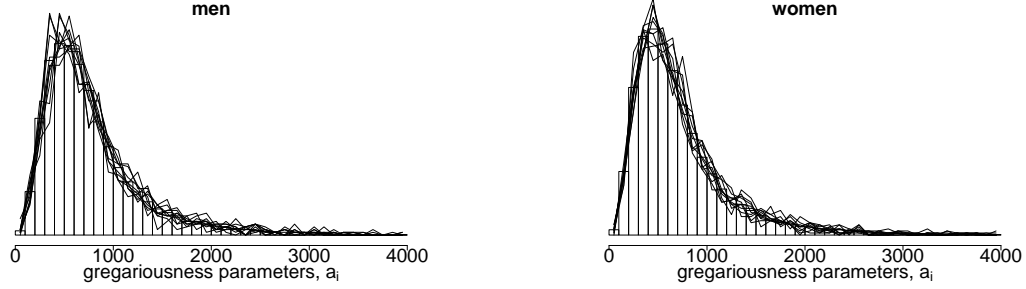
Figure 2: Estimated distributions of "gregariousness" or expected degree, $a_i = e^{\alpha_i}$ from the fitted model. Men and women have similar distributions (with medians of about 610 and means about 750), with a great deal of variation among persons. The overlain lines are posterior simulation draws indicating inferential uncertainty in the histograms.

7. For each k, update ω_k using a Metropolis step with jumping distribution, $\omega_k^* \sim \mathrm{N}(\omega_k^{(t-1)}, (\text{jumping scale of } \omega_k)^2)$.

8. Rescale the α's and β's by computing C from (6) and adding it to all the α_i's and μ_α and subtracting it from all the β_k's and μ_β, as discussed at the end of Section 3.2.

We construct starting points for the algorithm by fitting a classical Poisson regression (the null model, $y_{ik} \sim \mathrm{Poisson}(\lambda_{ik})$, with $\lambda_{ik} = a_i b_k$) and then estimating the overdispersion for each subpopulation k using $\frac{1}{n} \sum_{i=1}^{n} (y_{ik} - \hat{a}_i \hat{b}_k)^2 / (\hat{a}_i \hat{b}_k)$. The Metropolis jumping scales for the individual components of α, β, ω are set adaptively so that average acceptance probabilities are approximately 40% for each scalar parameter (Gelman, Roberts, and Gilks, 1996).

4 Results

We fit the overdispersed 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 $a_i = e^{\alpha_i}$, the prevalence parameters $b_k = e^{\beta_k}$, 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 (582 males and 784 females, with 4 individuals excluded due to missing gender information). Fitting the models separately for men and women makes sense since many of the subpopulations under study are single-sex groups. As we shall see, men tend to know more men and women tend to know more women, and more subtle sex-linked patterns also occur. Other interesting patterns arise when we examine the correlation structure of the model residuals, as we discuss in Section 4.5.

4.1 The distribution of social network sizes a_i

The estimation of the distribution of social network sizes, the distribution of the a_i's in our study, is a problem that 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), the scale-up method described earlier in this paper (Killworth et al., 1998a, b), and the summation method (McCarty et al., 2001). Despite a large amount of work, this body of research offers little consensus. Our estimates of the distribution of the a_i's 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.

Figure 2 displays estimated distributions of the gregariousness parameters $a_i = e^{\alpha_i}$ for the survey respondents, showing separate histograms of the posterior simulations from the model estimated separately to the men and the women. Recall that these values are calibrated based on the implicit assumption that the rare names in the data have the same average degrees as the population as a whole (see the end of Section 3.2). The similarity between the distributions for men and for women is intriguing. This similarity is not an artifact of our analysis, but instead seems to be telling us something interesting about the social world.

We estimate the median degree of the population to be about 610 (650 for men and 590 for women), with an estimated 90% of the population having expected degrees between 250 and 1710. These estimates are a bit higher than that of McCarty et al. (2001), for reasons we discuss near the end of Section 4.2.

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 precisely. This precision can be seen in the solid lines which are overlaid on Figure 2 and represent inferential uncertainty.

Figure 3 presents a simple regression analysis estimating some of the factors predictive of $\alpha_i = \log(a_i)$, using the data on the respondents in the McCarty et al. survey. These explanatory factors 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.

4.2 Relative sizes b_k of subpopulations

We now consider the group-level parameters. The left panels of Figure 4 show the 32 subpopulations k and the estimates of e^{β_k}, the proportion of links in the network that go to a member of group k (Be^{β_k} is the total degree of group k). The right panel displays the estimated overdispersions ω_k. The sample size is large enough that the 95% error bars are tiny for the β_k's and reasonably small for the ω_k's as well. (It is a

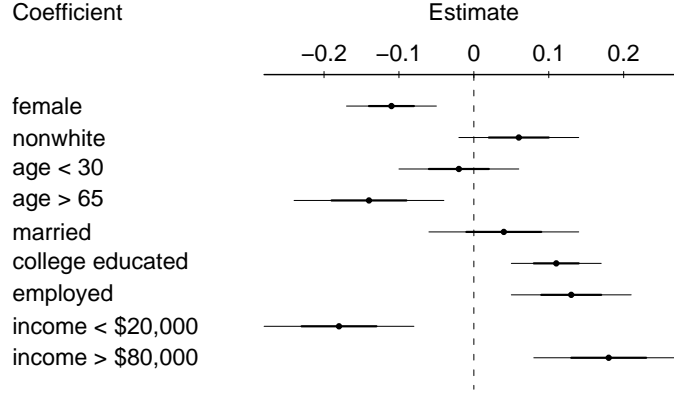| Coefficient	Estimate

Figure 3: Coefficients (and ± 1 standard error and ± 2 standard error intervals) of the regression of estimated log gregariousness parameters α_i on personal characteristics. Because the regression is on the logarithmic scale, the coefficients (with the exception of the constant term) can be interpreted as proportional differences: thus, with all else held constant women have social network sizes 11% smaller than men, persons over 65 have social network sizes 14% lower than others, and so forth. The R^2 of the model is only 10%, indicating that these predictors explain little of the variation in gregariousness in the population.

general property of statistical estimation that mean parameters (such as the β's in this example) are easier to estimate than dispersion parameters such as the ω's.) The figure also displays the separate estimates from the men and women.

Considering the β'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 β'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 of a social organization, Jaycees would be expected to know more people than average, so their b_k should be larger than an average group of an equal size.), 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 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
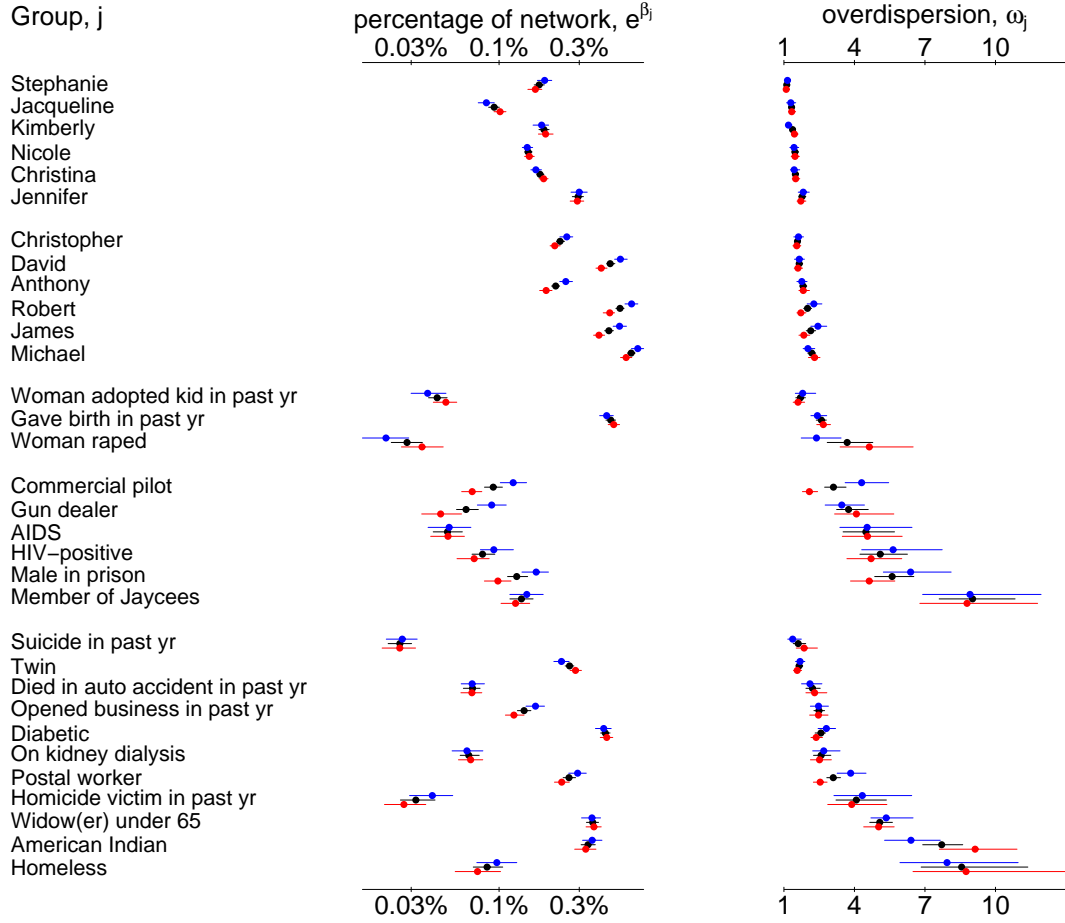
Figure 4: Estimates (and 95% intervals) of b_k and ω_k, plotted for groups X in the "How many X's do you know?" survey of McCarty et al. (2001). The estimates and uncertainty lines are clustered in groups of three; for each group, the top (blue), middle (black), and bottom (red) dots/lines correspond to men, all respondents, and women, respectively. The groups are listed in categories—female names, male names, female groups, male (or primarily male) groups, and mixed-sex groups—and in increasing average overdispersion within each category.

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 this scale-up method is used to estimate social network size, 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. We discussed the particular renormalization we use at the end of Section 3.2. This also explains why our estimate of the mean of the degree distribution is 750, as compared to 290 as estimated from the same data by McCarty et al. (2001).

Another pattern in Figure 5 is that the slope of the line for the names is steeper than for the other groups.
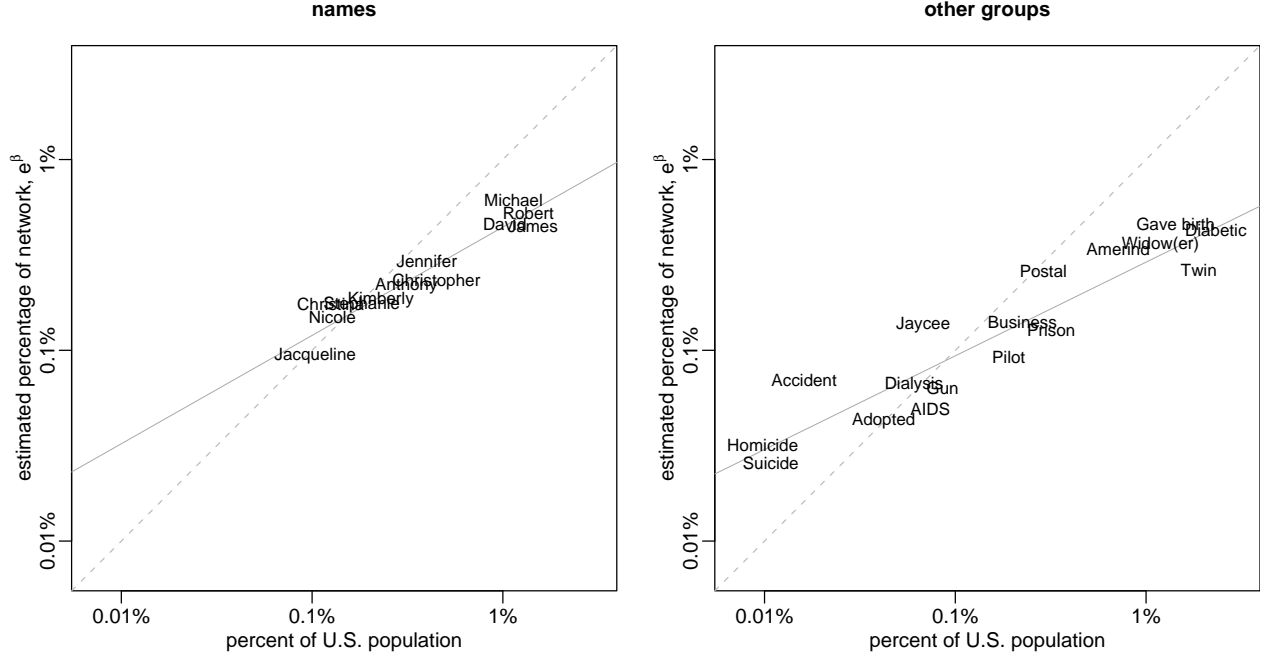
names

other groups

estimated percentage of network, e^β

1%

0.1%

0.01%

Michael
Robert
David James
Jennifer
Christopher
Anthony Kimberly
Christina Stephanie
Nicole
Jacqueline

0.01%   0.1%   1%
percent of U.S. population

estimated percentage of network, e^β

1%

0.1%

0.01%

Gave birth Diabetic
Widow(er)
Amerind
Postal   Twin
Jaycee   Business Prison
Pilot
Accident   Dialysis Gun
Adopted AIDS
Homicide
Suicide

0.01%   0.1%   1%
percent of U.S. population

Figure 5: Log-log plot of estimated prevalence of groups in the population (as estimated from the "How many X's do you know?" 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 and equally recalled by respondents.

We suppose that is because, for a given group size, it is easier to recall names than characteristics. After all, you know the name of almost all your acquaintances, but you could easily be unaware that a friend has diabetes, for example.

4.3   Overdispersion parameters ω_k for subpopulations

Recall that we introduced the overdispersed model to attempt to estimate the variability in individuals' 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. However, larger values of ω_k imply variation in individuals' relative propensities.

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 little variation in relative propensities. In contrast, the other groups have a wide range of overdispersions, ranging from near 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 who is 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.
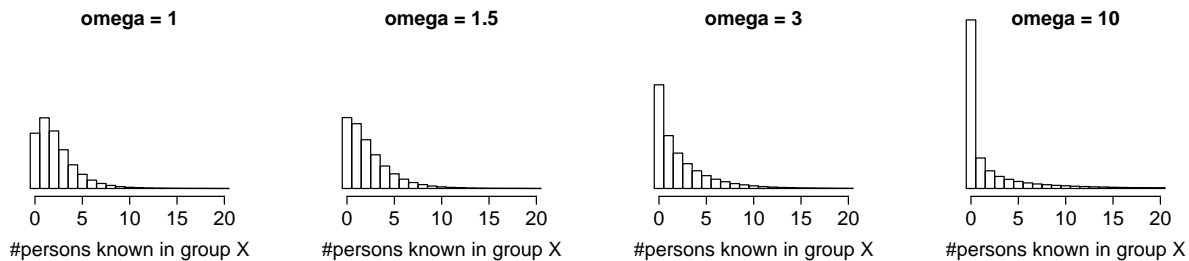

Figure 6: Distributions of "how many X's do you know" count data simulated from the overdispersed model corresponding to groups of equal size (representing 0.5% of the population) with overdispersion parameters 1 (the null model), 1.5, 3, and 10. All the distributions displayed here have the same mean. However, we can see that as the overdispersion parameter (ω) increases, we observe broader distributions with more 0's, more high values, and fewer 1's.

One way to understand the parameters ω_k in the data, which range from about 1 to 10, is to examine the effect these overdispersions have on the distribution of the responses to the question, "How many people do you know of type X?" The distribution becomes broader as the a_i's vary and as ω increases. Figure 6 illustrates for several values of ω: as the overdispersion parameter increases, we expect to see increasingly many 0's and high values, and fewer 1's (as expressed analytically in equation (5)).

4.4 Differences between men and women

A more subtle pattern in the data involves the differences between male and female respondents. Figure 7 plots the difference between men and women in the overdispersion parameters ω_k, vs. the "popularity" estimates b_k, for each subpopulation k. For names and for the other groups, there is a general pattern that overdispersion is higher among the sex for which the group is more popular. This makes some sense: overdispersion occurs when members of a subgroup are known in clusters, or more generally when knowing one member of the subgroup makes it more likely that you will know several. For example, on average, men know relatively more airline pilots than women, perhaps because they are more likely to be pilots themselves, in which case they might know many pilots, yielding a relatively high overdispersion. We do not claim to understand all the patterns in Figure 7, for example that Roberts and Jameses tend to be especially popular and overdispersed among men, compared to women.

4.5 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, so that anyone who knows more X's will tend to know more Y's)? For each survey response y_{ik}, we can define the standardized residual as

$$\text{residual:} \quad r_{ik} = \sqrt{y_{ik}} - \sqrt{a_i b_k}, \tag{7}$$

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; Tukey, 1972.)

For each pair of groups k_1, k_2, we can compute the correlation of their vectors of residuals; Figure 8 displays the matrix of these correlations. Care must be taken when interpreting the figure. At first, it may appear that the correlations are quite small. However, this is in some sense a natural results of our model. That is, if the correlations where all positive for group k, then the popularity b_k of that group would increase.

Several patterns can be seen in Figure 8. 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
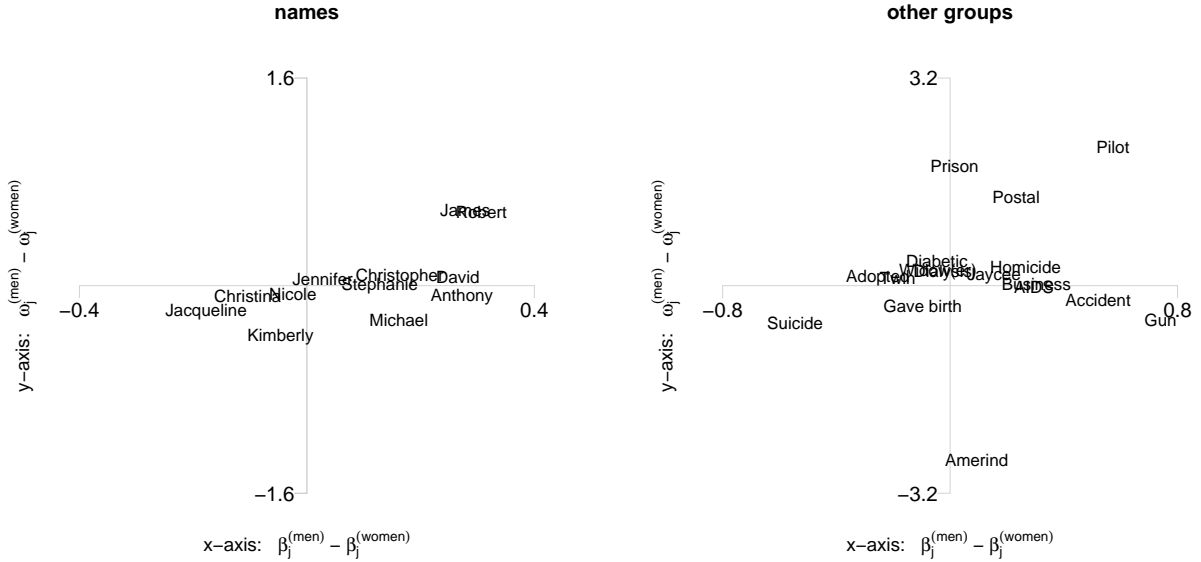

Figure 7: Differences between men and women in the overdispersion parameter ω_k and log-prevalence β_k, for each group k. In each graph, the y-axis shows the estimate of $\omega_j^{(\text{men})} - \omega_j^{(\text{women})}$, the difference in overdispersions among men and women for group j, and the x-axis shows $\beta_j^{(\text{men})} - \beta_j^{(\text{women})}$, the difference in log-prevalences among men and women for group j. Names and other groups are plotted separately on different scales. In general, groups that are more popular among men have higher variations in propensities for men. A similar pattern is observed for women.
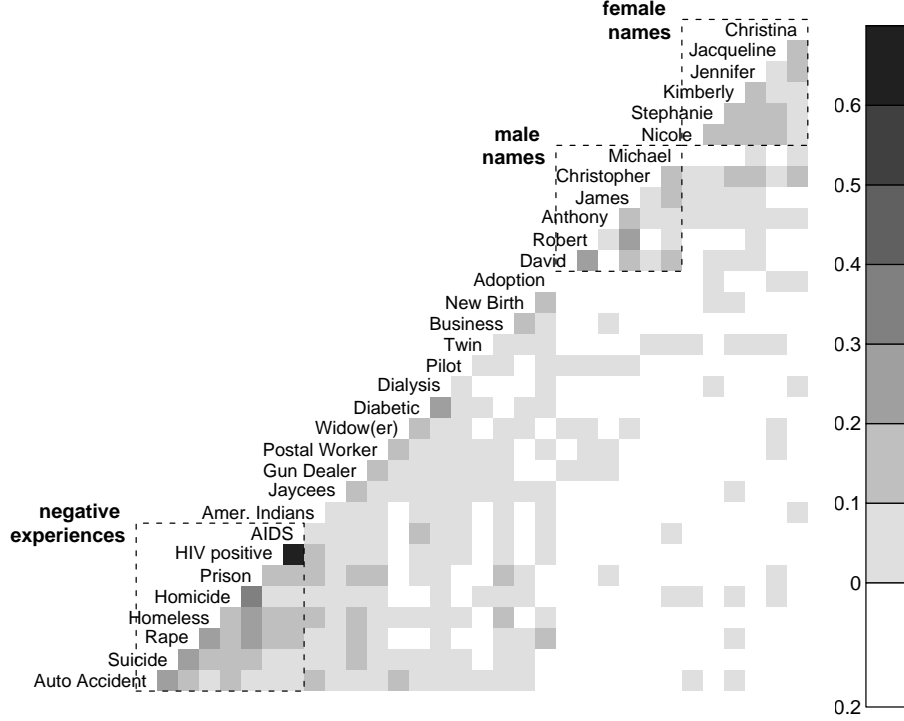
Figure 8: Correlations of the residuals r_{ik} among the survey respondents: people who know more HIV-positive persons know more AIDS patients, etc. The groups other than the names are ordered based on a clustering algorithm that maximizes correlations between nearby groups.

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. Another pattern is the mostly positive correlations among the names and mostly positive correlation among the non-name groups, but not much correlation between these two general categories. One possible explanation is that for some individuals, names are easier to recall, while for some others non-name traits (such as new births) are more memorable.

Instead of correlating the residuals, we could have examined the correlations of the raw data. However, these would be more difficult to interpret because we would find positive correlations everywhere, for the uninteresting reason that some respondents know many more people than others, so that if you know more of any category of person, you are likely to know more in just about any other category.

Another alternative would be to calculate the correlation of estimated interactions γ_{ik} (the logarithms of the relative propensities of respondents i to know persons in group k) rather than the residuals (7). However, estimates of the individual γ_{ik} are extremely noisy (recall that we focus our interpretation on their distributional parameter ω_k) and so not very useful. However, as was seen in Figure 8, the residuals still gives us useful information.

In addition to correlations, one can attempt to model the residuals based on individual level predictors. For example, Figure 9 shows the estimated coefficients of a regression model fit to the residuals of the null
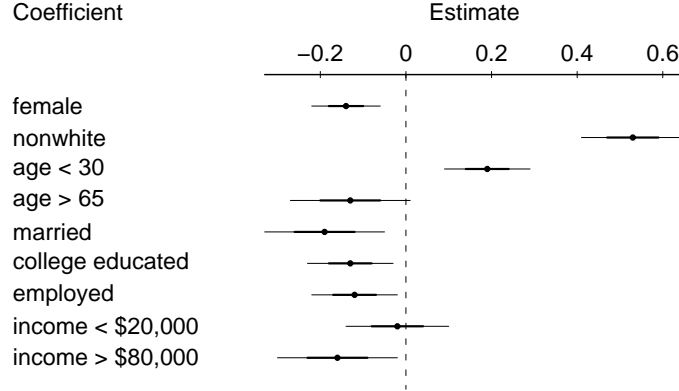
Figure 9: Coefficients (and ± 1 standard error and ± 2 standard error intervals) of the regression of residuals for the "How many males do you know incarcerated in state or federal prison?" question on personal characteristics. Being male, nonwhite, young, unmarried, etc., are associated with knowing more people than expected in federal prison. However, the R^2 of the regression is only 11%, indicting that most of the variation in the data is not captured by these predictors.

model for the "how many males do you know in state or federal prison" question. It is no surprise that being male, nonwhite, young, unmarried, etc., are associated with knowing more males than expected in state or federal prison. However, somewhat surprisingly, the R^2 of the regression model is only 11%.

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).

4.6 Posterior predictive checking

We can also check the quality of the overdispersed model by comparing posterior predictive simulations from the fitted model to the data (see, e.g., Gelman et al., 2003, chapter 6). We create a set of predictive simulations by sampling new data y_{ik} independently from the negative binomial distributions given the parameter vectors α, β, ω drawn from the posterior simulations already calculated. We can then examine various aspects of the real and simulated data, as illustrated in Figure 10. For now, just look at the bottom row of graphs in the figure; we return in Section 5 to the top three rows. For each subpopulation k, we compute the proportion of the 1370 respondents for which $y_{ik} = 0$, $y_{ik} = 1$, $y_{ik} = 3$, and so forth. These values are then compared to posterior predictive simulations under the model. On the whole, the model fits the aggregate counts fairly well but tends to under-predict the proportion of respondents who know exactly one person in a category. In addition, the data and predicted values for $y = 9$ and $y = 10$ show the artifact that persons are more likely to answer with round numbers (which can also be seen in the histograms in Figure 1). This phenomenon, often called heaping, was also noted by McCarty et al. (2001).
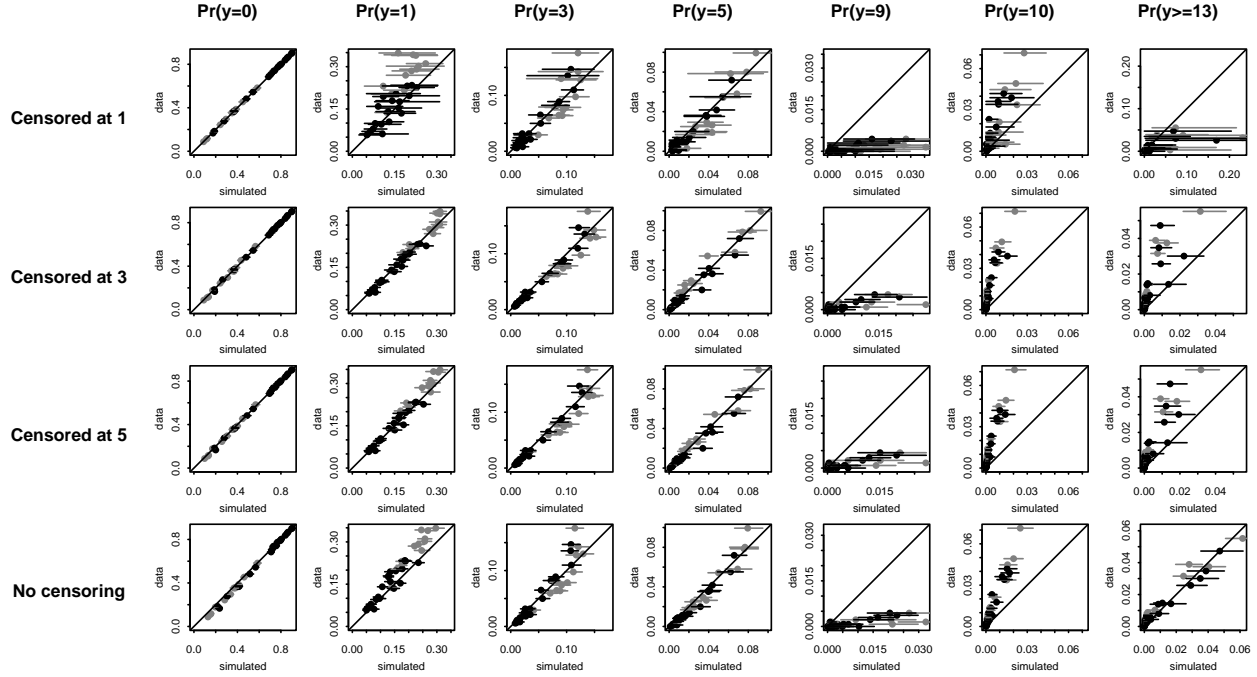
Figure 10: Model checking graphs: observed vs. expected proportions of responses y_{ik} of 0, 1, 3, 5, 9, 10, and ≥ 13. Each row of plots compares actual data to the estimate from one of four fitted models. The bottom row shows our main model, and the top three rows show models fit censoring the data at 1, 3, and 5, as explained in Section 5. In each plot, each dot represents a subpopulation, with names in gray, non-names in black, and 95% posterior intervals indicated by horizonal lines.

5   Measuring overdispersion without complete count data

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 was collected. Examples include the position generator studies (for a review, see Lin, 1999) and the resource generator studies (Van Der Gaag and Snijders, 2005) which both attempt to measure individual-level social capital. In these studies, respondents are 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. It would be important to know if we could use such data to estimate the variation in popularities of individuals, groups, and overdispersions of groups—the α_i's, β_k's, and ω_k's in our model.

First, 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, we find that by slightly increasing the response burden on respondents and collecting data of the type 0, 1, 2, 3-or-more, researchers would be able to make reasonable estimates of overdispersion,

even with such censored data. Such multiple-choice question would naturally capture less information than an exact count but would perhaps be less subject to the recall biases discussed in Section 4.2.

5.1   Theoretical ideas for estimating the model from partial data

We briefly discuss, from a theoretical perspective, what information is needed to estimate overdispersion from partial information such as yes/no data or questions such as, "Do you know 0, 1, 2, or more than 2 persons of type X?". We illustrate with the McCarty et al. data in the next subsection.

   With simple yes/no data ("Do you know any X's?"), overdispersion can only be estimated if external information is available on the b_k's. However, overdispersion can be estimated if questions are asked of the form, "Do you know 0, 1, ..., c or more person named Michael?", for any $c \geq 2$. It is straightforward to fit the overdispersed model from these censored data, with the only change being in the likelihood function. From the negative binomial model, $\Pr(y_{ik} = 1) = \exp\left(\log a_i + \log b_k - \log \omega_k - \frac{\log \omega_k}{\omega_k - 1} a_i b_k\right)$, and with information on b_k's, b_k and ω_k can be separated. If y_{ik} is the number of acquaintances in group k known by person i, we can write the censored data (say, for $c = 2$) 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.

5.2   Empirical application with artificially censored data

To examine 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. We compare the estimated group prevalence parameters β_k and overdispersion parameters ω_k from each of the three censored datasets with the estimates from the complete (uncensored) data. From the results (not shown), we conclude that censoring at 3 or 5 preserves much but not all of the information for estimation of β_k and ω_k, while censoring at 1 (yes/no data) gives reasonable estimates for the β_k's but nearly useless estimates for the ω_k's. In addition, the Gibbs-Metropolis algorithm is slow to converge with the yes/no data.

   Along with 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.

   To understand this phenomenon better, we repeat our procedure—fitting the model to complete and censored data—but using a fake dataset constructed by simulating from the model given the parameter estimates (as was done for posterior predictive checking in Section 4.6). Our computation (not shown) reveals that the estimation procedure seems to be working well with the fake data when censored at 3 or

5. Most notably, no underestimation for the overdispersion parameters ω_k is observed due to the censoring. However, the nonidentification shows up when estimating from yes/no data.

A comparison of the results using the real data and the results using the simulated fake data reveals that some of the changes obtained from fitting to censored data arise from poor fit of model to the data. To explore this further, we compute the expected proportions of $y_{ik} = 0$, $y_{ik} = 1$, etc., from the model, as fit to the different censored datasets. The top three rows of Figure 10 show the results. The censored-data models fit the data reasonably well or even better than the non-censored data for low counts but do not perform so well at predicting the rates of high values of y, which makes sense since this part of the distribution is being estimated entirely by extrapolation.

6   Discussion

6.1   Connections to previous work

We have developed a new method for measuring one aspect of social structure which can be estimated from sample data—variation in the propensities for individuals to form ties with people in certain groups. Our measure of overdispersion may seem similar to, but is in fact distinct from, previous measures that have attempted to uncover deviations from random mixing such as homophily (McPherson et al., 2001) and assortative mixing (Newman, 2002, 2003a).

Originally defined by Lazersfeld and Merton (1954), homophily represents the tendency for people to associate with those who are similar. Later Coleman (1958) developed a way of quantifying this tendency which is common in current use (see for example, Heckathorn, 2002). Newman's measures of assortative mixing are another attempt to measure the tendency for vertices in networks to be connected to other similar vertices.

Our object of study is different because we are estimating the variation in propensities of respondents to form ties to people in a specific group, whether or not the respondents are actually in the group themselves. That is, we are looking at how contact with a group is distributed throughout the population (group members and non-group members) whereas homophily and assortative mixing only focus on the tendency for group members to form ties to other group members. For example, people with certain diseases may not necessarily associate with each other, but they could have a higher propensity to know health care workers.

From the McCarty et al. data we estimate overdispersion for groups that do not appear in our sample (for example, homeless, death by suicide, death by auto accident, homicide victims, and males in prison). We estimate varying degrees of overdispersion for these groups without the need, or even the possibility, of measuring the homophily or assortative mixing of these groups.

We are able to makes estimates about these groups that are not included in our sample because our method of detecting social structure is indirect. By surveying a random sample of 1370 Americans and then asking about all their acquaintances, we are gathering partial information about the hundreds of thousands

of persons in their social network (using our estimate of the mean of the degree distribution, the survey potentially gathers information on $1370 \times 750 = 1$ million individuals), thus allowing us to learn about small and otherwise hard-to-reach groups.

Further, by explicitly focusing on variation among people, our method differs from many existing network measures which tend to focus on measures of central tendency of group behaviors. Our method also differs from many statistical models for count data which treat super-Poisson variation as a problem to be corrected and not a source of information itself. We suspect that this increased attention to variation could yield useful insights on other problems.

6.2   Future improvements and applications of these methods

Our model is not perfect, of course, as can be seen from the model-checking graphs of Figure 10. For one thing, the model cannot capture *underdispersion*, which could be thought of as an increased probability of knowing exactly one person of type X (see equation (5)), which could occur, for example, with occupational categories where it is typical to know exactly one (e.g., dentists). To model this phenomenon, it would be necessary to go beyond the negative binomial distribution, with a natural model class being mixture distributions that explicitly augment the possibility of low positive values of y.

A different way to study variance in the propensity to forms ties to specific groups would be to classify links using the characteristics of the survey respondents, following the ideas of Hoff, Raftery, and Handcock (2002), Jones and Handcock (2003), and Hoff (2005) adapting logistic regression models to model social ties. For example, the McCarty et al. data show that men, on average, know nearly twice as many commercial pilots than do women—and the two sexes have approximately the same average social network size, so this difference represents a clear difference in the relative propensities of men vs. women to know an airline pilot. The nonuniformity revealed here would show up in simple yes/no data as well. (For example, 35% of men in our survey, compared to 29% of women, know at least one airline pilot.) So we can see at least some patterns without the need to measure overdispersion.

Given the complexity of actual social networks, however, there will in practice always be overdispersion even after accounting for background variables. A natural way to proceed is to combine the two approaches by allowing the probability of a link to a person in group k to depend on the observed characteristics of person i, with overdispersion after controlling for these characteristics. This corresponds to fitting regression models to the latent parameters α_i and γ_{ik} given individual-level predictors X_i. Regressions such as displayed in Figures 3 and 9 would then be part of the model, thus allowing more efficient estimation than could be obtained by postprocessing of parameter estimates and residuals. Controlling for individual characteristics also allows poststratified estimates of population quantities (see, e.g., Lohr, 1999, Park, Gelman and Bafumi, 2004.)

For the goal of estimating social network size, it would make sense to include several rare names of both sexes to minimize the bias, demonstrated in Figure 5 and discussed in Killworth et al. (2003), of under-recall

for common categories. Using rarer names would increase the variance of the estimates, but this problem could be mitigated by asking about a large number of such names. Fundamentally, if recall is a problem, the only way to get accurate estimates of network sizes for individuals is to ask many questions.

In this paper, we have fit the overdispersion model separately for men and women. One would also expect race and ethnicity to be important covariates, especially for the recognition of names whose popularity varies across racial groups. We have run (not shown) analyses separately for the whites and the nonwhites. The differences for most estimated parameters were not statistically significant. A difficulty is that there were only 233 nonwhites in the survey data.

6.3   Understanding the origins and consequences of overdispersion

Perhaps the biggest unanswered questions that come from this paper do not deal with model formulation or fitting, but with understanding the origins and consequences of the phenomena that we have observed.

We find a large variation in individual propensities to form ties to different groups, but we do not have a clear understanding of how or why this happens. In some cases, the group membership itself may be important in how friendships are formed, for example, being homeless or being a Jaycee. However, for other groups, for example people named Jose (a group that unfortunately was not included in our data), there might be variation in propensity not caused by the group membership itself, but by associated factors such as ethnicity and geographic location. Sorting out the effect of group membership itself, versus its correlates, is an interesting problem for future work. Insights in this area may come from the generalized affiliation model of Watts, Dodds, and Newman (2002). Understanding how social institutions like schools help to create this variation in propensity is another area for further research.

In addition to trying to understand the origins of overdispersion, it is also important to understand its consequences. A large amount of research in psychology has shown that, under certain conditions, intergroup contact affects opinions (for a review see Pettigrew, 1998). For example, one could imagine that a person's support for the death penalty is affected by how many people they know in prison. These psychological findings imply that the distribution of opinions in the society are at least partially determined by the social structure in a society, and not simply by the demographics of its members. That is, we could imagine two societies with exactly the same demographic structures but with very different distributions of opinions only because of differences in social structure.

Overdispersion, which means that acquaintanceship counts with specific subpopulations have more zeroes and more high numbers than expected under the null model, will have the effect of polarizing opinions on issues. If contact with the specific subpopulation were more evenly distributed in the population, we might see a different, more homogeneous, distribution of opinions about that group.

In addition to changing the distribution of opinions, overdispersion can influence the average opinion as well. For example, we can consider support for the rights of indigenous people in the two hypothetical populations in Figure 11. The left panel of the figure shows the distributions of the number of American
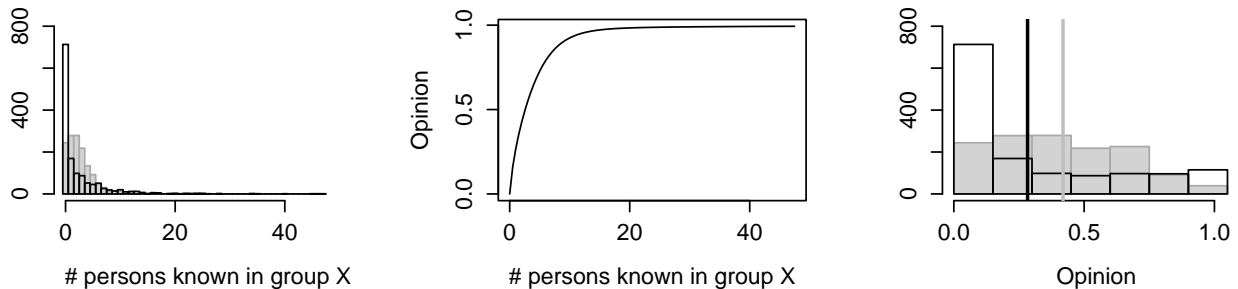
Figure 11: Illustration of the effect of overdispersion on mean opinion in a population. The left panel shows two different populations each with the same number of people and the same mean number of connections to a specific group but different overdispersions ($\omega = 1$ (gray bars) and $\omega = 7.7$ (empty bars)). The center panel shows a function, which applies to all individuals in both populations, that maps the number of persons an individual knows in a specific group to that individual's opinion on a specific issue. The right panel shows the resulting distribution of opinions. The population with no overdispersion has substantially higher mean opinion (0.42 vs 0.28, indicated in the graph). Thus, observed differences in opinion distributions across different societies could potentially be attributed entirely to differences in social structure rather than any differences between individuals.

Indians known. In both distributions the mean is the same (as our estimate from McCarty et al. data), but the distributions differ in their overdispersion. In one population there is no variation in relative propensities to form ties to the American Indians ($\omega = 1$), while in the other population there is substantial variation in relative propensities, in this case $\omega = 7.7$ which matches our estimate with respect to the American Indians in the acquaintanceship network of Americans.

The center panel of Figure 11 shows a hypothetical function that maps the number of people known in a specific group to an opinion (on a scale of 0 to 1, with 1 being most positive) on a specific issue—in this example, a map from the number of American Indians known, to a composite score measuring an individual's support for the rights of indigenous people. Here we assume that the function is increasing, monotonic, and nonlinear with diminishing returns (derivative and second derivative that approach zero). In this case, the change in a subject's opinion caused by knowing someone who is American Indian is likely to be larger if that person previously knew zero American Indians than if the subject previously knew 10 American Indians. In our simplified example this mapping is the same for everyone in both of these populations so the two populations can be thought of as being made up of identical people.

Even though the people in both populations are identical, the right panel of Figure 11 shows the distributions of opinions in the populations are substantially different. In the population without overdispersion ($\omega = 1$) there is much more support for the American Indians than in the population with overdispersion ($\omega = 7.7$). One way to think about this difference is that in the population where contact with the American Indians is overdispersed, the impact of the contact is concentrated in fewer people so each contact is likely to have less of an effect.

The difference in mean support for the rights of indigenous people (0.42 vs. 0.28 on a scale of 0 to 1) in the

two populations can be attributed *entirely* to differences in social structure. In both cases the populations are made up of identical individuals with identical mean amount of contact with the American Indians; they differ only in social structure. This hypothetical example indicates that it is possible that certain macro-level sociological differences between societies are not attributable to differences between individuals in these societies. Rather, macro-level differences of opinion can sometimes be attributed to micro-level differences in social structure.

In this paper we have shown that Americans have varying propensities to form ties to specific groups and we have estimated this variation for a number of traits. Future empirical work could explore this phenomena for groups other than those included in the McCarty et al. (2001) data or explore this phenomena in other countries. Important work also remains to be done in understanding the origins and consequences of this aspect of social structure.

References

Blau, P. M. (1974). Parameters of social structure. *American Sociological Review* **39**, 615–635.

Coleman, J. S. (1958). Relational analysis: the study of social organization with survey methods. *Human Organization* **17**, 28–36.

Erdös, P., and Renyi, A. (1959). On random graphs. *Publicationes Mathematicae* **6**, 290–297.

Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science.* Vancouver: Empirical Press.

Freeman, L. C., and Thompson, C. R. (1989). Estimating acquaintanceship volume. In *The Small World*, ed. M. Kochen. Ablex Publishing.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: Chapman and Hall.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford University Press.

Gurevich, M. (1961). *The Social Structure of Acquaintanceship Networks.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Granovetter, M. (1995). *Getting a Job: A Study in Contacts and Careers*, second edition. University of Chicago Press.

Handcock, M. S., and Jones, J. (2004). Likelihood-based inference for stochastic models of sexual network formation. *Theoretical Population Biology* **65**, 413–422.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems* **44**, 174–199.

Heckathorn, D. D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* **49**, 11–34.

Heckathorn, D. D., and Jeffri, J. (2001). Finding the beat: using respondent-driven sampling to study jazz musicians. *Poetics* **28**, 307–329.

Hoff, P.D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association* **100**, 286–295.

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

Jones, J., and Handcock, M. S. (2003). An assessment of preferential attachment as a mechanism for human sexual network formation. *Proceedings of the Royal Society London, B* **270**, 1123–1128.

Killworth, P. D., and Bernard, H.R. (1978). The reverse small-world experiment. *Social Networks* **1**, 159–192.

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., and McCarty, C. (1990). Estimating the size of personal networks. *Social Networks* **12**, 289–312.

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A., and Bernard, H. R. (1998a). A Social Network Approach to Estimating Seroprevalence in the United States. *Social Networks* **20**, 23–50.

Killworth, P. D., McCarty, C., Bernard, H. R., Johnsen, E. C., Domini, J., and Shelley, G. A. (2003). Two interpretations of reports of knowledge of subpopulation sizes. *Social Networks* **25**, 141–160.

Killworth, P. D., McCarty, C., Bernard, H. R., Shelly, G. A., and Johnsen, E. C. (1998b). Estimation of seroprevalence, rape, and homelessness in the U.S. using a social network approach. *Evaluation Review* **22**, 289–308.

Lazarsfeld, P. F., and Merton, R. K. (1954). Friendship as a social process: a substantive and methodological analysis. In *Freedom and Control in Modern Society*, ed. M. Berger, 11–66. New York: Van Nostrand.

Lee, B. A., Farrell, C. R., and Link, B. G. (2004). Revisiting the contact hypothesis: the cases of public exposure to homelessness. *American Sociological Review* **69**, 40–63.

Lin, N. (1999). Social networks and status attainment. *Annual Review of Sociology* **25**, 467–487.

Lorh, S. L. (1999) *Sampling: design and analysis*, Duxbury Press.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., and Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization* **60**, 28–39.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. London: Chapman and Hall.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology* **27**, 415–444.

Morris, M., and Kretzchmar, M. (1995). Concurrent partnerships and transmission dynamics in networks. *Social Networks* **17**, 299–318.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters* **89**, article number 208701.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Physical Review E* **67**, article number 026126.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Review* **45** (2), 167–256.

Park, D. K., Gelman, A. and Bafumi, J. (2004) Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis* **12**, 375-385.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology* **49**, 65–85.

Pool, I. d. S., and Kochen, M. (1978). Contacts and influence. *Social Networks* **1**, 5–51.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models*, second edition. Thousand Oaks, Calif.: Sage.

Salganik, M. J., and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology 34, 193-239.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel Analysis*. London: Sage.

Strogatz, S. H. (2001). Exploring complex networks. *Nature* **410**, 268–276.

Tukey, J. W. (1972). Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft. Iowa State University Press.

Van Der Gaag, M. and Snijders, T. A. B. (2005). The Resource Generator: social capital quantification with concrete items. Social Networks 27(1), 1-29.

Van Dyk, D. A., and Meng, X. L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.

Wasserman, S., and Faust K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences USA* **99**, 5766–5771.

Watts, D. J., Dodds, P. S., and Newman, M. E. J. (2002). Identity and search in social networks. *Science* **296**, 1302–1035.

Watts, D. J. (2004). The "new" science of networks. *Annual Review of Sociology* **30**, 243–270.