Measuring social distance using indirectly observed network data[*]

Tyler H. McCormick[†‡] Amal Moussa[†‡] Johannes Ruf[†‡]
Thomas A. DiPrete[§] Andrew Gelman[¶] Julien Teitler[‖] Tian Zheng[‡**]

Abstract

Aggregated Relational Data (ARD), or "How many X's do you know?" questions, are an increasingly common tool for learning about social networks through standard surveys. Recent statistical advances (Zheng et al. (2006), for example) present social scientists with new options for analyzing such data. In this paper, we present guidelines for learning about network processes and a template to aid practitioners. We begin by proposing ARD be viewed as a measure of "social distance" between a respondent and a subpopulation (individuals named Kevin, those in prison, or serving in the military). We then present common methods for analyzing these data and associate each of these methods with a specific way of measuring distance. We examine the implications of using each of these social distance measures to social scientists using an internet survey about contemporary political issues provided by the Cooperative Congressional Election Study.

1. Introduction

Though the overwhelming majority of methods for analyzing network data assume complete network data are available, these data are financially or logistically impossible to collect, an issue Burt (1982) calls "the single factor most restricting structural theory." Aggregated Relational Data (ARD) are one increasingly popular alternative for measuring networks indirectly. ARD, introduced by Killworth et al. (1998a), are answers to questions of the form "How many X's do you know?" where "X" represents a subpopulation of interest. Thus, instead of measuring direct relationships between actors as in the complete network case, we observe the frequency with which an actor interacts with a particular group. ARD are often used to predict characteristics of populations that are difficult to reach using standard surveys (Killworth et al., 1998b) and more recently to learn about polarization and segregation (DiPrete et al., 2010). These data do not require any specific sampling technique and are easily integrated into standard surveys.

Since we measure network features indirectly, the standard network statistics for complete network data are no longer well defined. Recent advances in statistical techniques have also suggested new statistics for learning about network features from these data. We address these two sources of ambiguity by conceptualizing a measure of social distance and comparing two ways of defining distance. In the forthcoming sections we first demonstrate how ARD can be viewed as a social distance measure. We then demonstrate that our two statistics yield different measures of social distance and explore the implications of these definitions for predicting a

[*] We thank the National Science Foundation for partial support of this research.

[†] The first three authors contributed equally to this paper.

[‡] Columbia University, Department of Statistics, 1255 Amsterdam Avenue, New York, NY 10027

[§] Columbia University, Department of Sociology, 1180 Amsterdam Avenue, New York, NY 10027

[¶] Columbia University, Department of Statistics and Department of Political Science, 1255 Amsterdam Avenue, New York, NY 10027

[‖] Columbia University, School of Social Work, 1255 Amsterdam Avenue, New York, NY 10027

[**] Corresponding author. E-mail: tzheng@stat.columbia.edu

respondent's political opinions. We use these distinctions to suggest guidelines for data analysis.

The natural starting point is the frequency of interactions between an actor and members of a given subpopulation, which is measured directly by ARD. Actors who interact more frequently with a subpopulation should, intuitively, be more socially proximate than those who interact infrequently. A social scientist who includes an ARD count as a covariate in a regression model, therefore, adjusts for the level of exposure to the subpopulation of interest.

Zheng et al. (2006) present an overdispersed Poisson regression model for ARD adjusting for respondent degree, or total network size, and the size of the subpopulation. Using the residuals from this model, therefore, yields an ARD-based statistic which measures the connectivity of the respondent to a given subpopulation in excess of what would be expected for a person with a similar network size. For two respondents who report knowing the same number of people on welfare, for example, we would conclude that the respondent with an acquaintanceship network of size 500 would be socially closer to the welfare population than a respondent with an 800 person network.

Given the different information in these two ARD-based statistics, a pertinent question for social science researchers is which of the two to include in an analysis. To address this issue, we compare the two measures as indicators of social influence, specifically as predictors of respondent opinions regarding contemporary political and social issues. The study of social influence, which explores the connection between the structure of actors' social relations and their behaviors or opinions, is a common framework for understanding the role of the social environment. Such research exhibits what Laumann (1979) considers "the hallmark of network analysis . . . to explain, at least in part, the behavior of network elements . . . by appeal to specific features of the interconnections among the elements" (p.394). The primary challenge, as we discuss in the remainder of this section, is defining the "specific features of the interconnections" of interest and explaining how these features manifest social influence.

When studying influence from a network perspective, a fundamental assumption is that the substantive bases of social influence can be represented as nearness in the respondent's social network. Specifically, the paradigm of most modern studies of social influence is that influence is proportional to nearness in the respondent's social network (Burt, 1987) with two primary definitions of nearness—social cohesion and equivalence. Social cohesion defines proximity in terms of the ties between actors, two actors being proximate if the length (or strength) of ties between them meets a particular standard. Equivalence, in contrast, considers the pattern of two actors' network relations and considers two actors proximate if they interact with the network in similar ways (have the same friends, for example).

Social cohesion and equivalence both relate the substantive foundations of influence to structural features of the network in the presence of dyadic data, or the complete network. Dyadic data encodes the absence or presence of a relationship between each pair of actors in the network. In the most restrictive case of structural cohesion, for example, two actors are proximate if a tie exists between them.

Without information about specific members of the network, as in the ARD case, the notion of cohesion introduced in Burt (1987) is difficult to define. We can, however, conceive both the counts and residuals as a form of equivalence, which we term *weak equivalence*. The key distinction in our definition is that we define proximity in terms of the expected distance between a respondent and an entire group of actors, rather than any specific actor.

In the coming sections we compare two potential proximity measures defined by ARD, counts and residuals. By counts we mean the raw responses to our ARD questions, which are currently the most common way which indirect network measures are included in analyses. We contrast the raw counts with residuals from an overdispersed Poisson regression model based on Zheng et al. (2006). We give evidence that the two methods reveal different aspects of social structure and contrast their performance as predictors in a standard regression analysis of factors influencing respondent opinions. The counts, therefore, represent a respondent's exposure, or level of knowledge, of the subpopulation group. In contrast, the residuals represent connectedness with the subpopulation as deviations from the expected. The residuals indicate social structure more directly, because they measure connectednes with a subpopulation net of what would be expected simply from the size of the subpopulation and the respondent's network size.

In Section 2 we describe the dataset and the overdispersed Poisson regression model and we define the residuals which we consider as one possible measure of social structure. We use this model to present results regarding social structure and opinion formation in Section 3. Section 4 discusses the main findings of our work.

2. Data and model

In this section we describe the dataset, the overdispersed Poisson regression model, the parameter estimation, and our procedure for computing residuals.

2.1 Data

The dataset is provided by the Cooperative Congressional Election Study (CCES), a large national online survey created by thirty universities (MIT Web, 2007). Each university has created a module of about 120 questions for 1000 respondents. The survey was conducted by Polymetrix in October and November 2006. For each survey of 1000 persons, half of the questionnaire is developed by an individual research team, and half of the questionnaire is given by Common Content (MIT Web, 2007), which consists of approximately 60 questions, 40 in the pre-election wave about general political attitudes, various demographic factors, voting choices, and political information, and 20 in the post-election wave. These questions are included on all 30 surveys leading to a 30000 person national sample survey. In addition to these questions, Polimetrix provides demographic indicators, party identification, ideology, and validated votes obtained after the 2006 election. Our dataset comes from Columbia University's module.

Polymetrix uses a random sample matching methodology to produce "representative" samples from non-randomly selected samples of respondents: first a target random sample is drawn from the US population, then each member of the target sample is matched with a respondent by minimizing a distance function on a large set of variables so that the respondent is as similar as possible to the selected member of the target sample. Thus, the matched sample has similar characteristics to the target sample.

Using an Internet survey can be a problem for generating a representative sample: there tend to be fewer elderly Internet users than young Internet users, however among the Internet users the propensity for participation in survey research is higher for elderly users than for young users (MIT Web, 2007). Thus, there is a pre-selection effect that can generate a misrepresentative sample where certain groups are under-represented. The survey could be also biased towards politically active people since 89% of the respondents claim to have voted in the 2006 elections, while the overall voter turnout was substantially lower.

Respondents were asked various questions related to their socioeconomic and personal characteristics (e.g., race, gender, education, income), to their political opinions (e.g., approval of a timetable in Iraq, approval of George W. Bush's handling of Iraq), and to their social network (e.g., how many Kevins or Brendas they know, how many unemployed persons and people on welfare they know). "To know" in our study is defined as knowing a person's name and being willing to stop and to talk to him for at least a moment. For $n = 994$ respondents (six of the original 1000 respondents did not answer any of the questions) and $K = 13$ subpopulations we are provided with the number of persons each respondent knows in each subpopulation. More precisely, the respondents could choose between with five possible choices: 0, 1, 2 to 5, 6 to 10, and more than 10 persons, which constitutes an interval-based dataset.

To illustrate, the answers for "How many Rachels do you know?" and "How many people on welfare do you know?" are summarized in Figure 1. The histograms show an overdispersion for the distributions of both groups, meaning that most people know few Rachels and few people on welfare, however some people know many Rachels and many people on welfare, with a more pronounced overdispersion for the "Welfare" group than for the "Rachel" group.

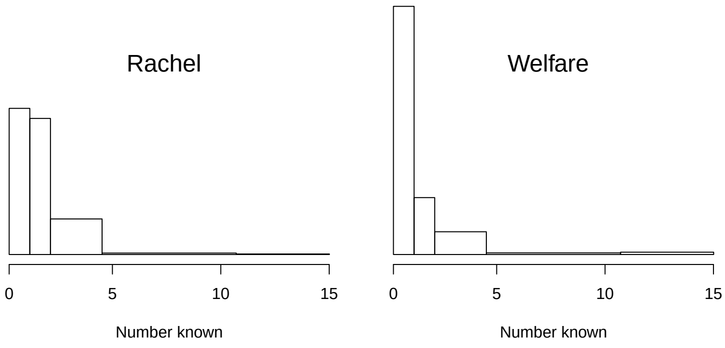

Figure 1: Histograms showing the distribution of the answers of two typical questions: "How many Rachels do you know?" and "How many people on welfare do you know?" The heights of the last two bins on the left histogram correspond to 23 and 8 respondents, on the right histogram to 27 in both bins. The larger ratio of the heights of the first and second bar in the right histogram compared to the left one indicates a higher overdispersion in the "Welfare" group. The estimated overdispersion parameters ω are here 1.2 for the Rachel group and 5.4 for the "Welfare" group.

2.2 Model

The standard Erdös-Renyi model for social links, which assumes that links between people in the population are formed completely randomly (i.e., the probability that two persons get to know each other is the same whoever those persons are), implies that the number $v_{i,k}$ of persons in subpopulation k that respondent i knows follows a Poisson distribution with intensity $\bar{a}b_k$, where $\bar{a}$ is the expected network size of a randomly selected member of the population and b_k is the expected number of links involving subpopulation k divided by the total expected number of social links (popularity of subpopulation k).

In this model, all individuals have the same expected degree. This is clearly not the case in the population at large, however, since some individuals are in more socially exposed positions (members of the clergy or politicians, for example) and there is natural variability in gregariousness. Furthermore, numerous previous studies using ARD have also found variation in excess of what would be expected under a Poisson model (Newman, 2003). This super-Poisson variation is known as overdispersion and results from the propensity for individuals to know either no members of a subpopulation or multiple members. Previous ARD studies, for example, found the mean number of individuals known who are currently incarcerated is around one (McCarty et al., 2001). This is misleading, however, since the majority of the population does not have any acquaintances who are currently incarcerated. On the other side, an individual who knows one person currently in prison, most likely knows multiple. We demonstrate this phenomena in our data in Figure 1.

To address these two issues, Zheng et al. (2006) propose an overdispersed model where individual i has also an individualized propensity $g_{i,k}$ to know people from the subpopulation k, formally: $v_{i,k} \sim \text{Poisson}(a_i b_k g_{i,k})$. This propensity $g_{i,k}$ follows a Gamma distribution with mean 1 and shape parameter $1/(\omega_k - 1)$, where ω_k is the overdispersion parameter, whose simplest interpretation is a scale of the variance: $\text{Var}(v_{i,k}) = \omega_k \text{E}(v_{i,k})$. The overdispersion accounts for the extra variance of the data that the null model can not account of (in the null model, $\text{Var}(v_{i,k}) = \text{E}(v_{i,k})$). The probability distribution of $v_{i,k}$ is negative binomial with mean $a_i b_k$ and overdispersion parameter ω_k. Reasonable priors for the gregariousness parameters a_i and for the popularity parameters b_k are lognormal distributions: $a_i \sim \exp N(\mu_\alpha, \sigma_\alpha^2)$ and $b_k \sim \exp N(\mu_\beta, \sigma_\beta^2)$. We assume a uniform prior on $(0,1)$ for the inverse of the overdispersion parameter and finally we complete the Bayesian hierarchical model by putting a noninformative prior for the parameters μ_α, σ_α, μ_β, and σ_β.

As we discuss in Section 2.1, the responses are interval-based. The relevant likelihood has been worked out by DiPrete et al. (2010). Denoting

$$L_{i,k}(y) := \binom{y + \xi_{i,k} - 1}{\xi_{i,k} - 1} \left(\frac{1}{\omega_k}\right)^{\xi_{i,k}+2} \left(\frac{\omega_k - 1}{\omega_k}\right)^{y}$$

and

$$\begin{aligned} p_{i,k} \quad &:= \quad L_{i,k}(0)\mathbf{1}_{\{v_{i,k}=0\}} + L_{i,k}(1)\mathbf{1}_{\{v_{i,k}=1\}} + \sum_{y=2}^{5} L_{i,k}(y)\mathbf{1}_{\{2\le v_{i,k}\le 5\}} \\ &+ \quad \sum_{y=6}^{10} L_{i,k}(y)\mathbf{1}_{\{6\le v_{i,k}\le 10\}} + \sum_{y=11}^{\infty} L_{i,k}(y)\mathbf{1}_{\{11\le v_{i,k}\}}, \end{aligned}$$

the joint posterior density can be written as

$$p(a, b, \omega, \mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta | v) = \prod_{i=1}^{n}\prod_{k=1}^{K} p_{i,k} * \prod_{i=1}^{n} N(\alpha_i|\mu_\alpha, \sigma_\alpha^2) * \prod_{k=1}^{K} N(\beta_k|\mu_\beta, \sigma_\beta^2),$$

where

$$\xi_{i,k} = \frac{a_i b_k}{\omega_k - 1}, \ \alpha_i = \log(a_i), \ \beta_k = \log(b_k),$$

and n denotes the number of respondents and K the number of subpopulations.

2.3 Parameter estimation

We estimate the parameters a_i, b_k, and w_k by the Metroplolis-Hastings algorithm. The Metropolis-Hastings algorithm is a rejection sampling algorithm used to generate a correlated sequence of draws from the target density that may be difficult to sample by a classical independence method. The algorithm is described in details in Zheng et al. (2006). As a starting value, we estimate the parameters a_i and b_k by the usual Poisson regression:

$$\begin{aligned}\hat{a}_i^0 &= \left(\sum_{k=1}^{K} c_k\right)^{-1} \sum_{k=1}^{K} c_k y_{i,k} \text{ with } c_k = \frac{1}{\sqrt{y_{\cdot,k}}},\\ \hat{b}_k^0 &= \left(\sum_{i=1}^{n} d_i\right)^{-1} \sum_{i=1}^{n} d_i y_{i,k} \text{ with } d_i = \frac{1}{\sqrt{y_{i,\cdot}}},\end{aligned}$$

where we define the "count" variable $y_{i,k}$ as

$$y_{i,k} := \mathbf{1}_{\{v_{i,k}=1\}} + 3.5 * \mathbf{1}_{\{2\leq v_{i,k}\leq 5\}} + 8 * \mathbf{1}_{\{6\leq v_{i,k}\leq 10\}} + 15 * \mathbf{1}_{\{10<v_{i,k}\}}. \tag{1}$$

That is, we map each interval on its midpoint and interpret this point as the number of people, which a certain respondent knows in a subpopulation. For the largest interval, namely more than 10 people, we assign the value 15.

We estimate the overdispersion parameter ω_k as the empirical scaled variance of the data, that is

$$\hat{\omega}_k^0 = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_{i,k} - \hat{a}_i^0 \hat{b}_k^0)^2}{\hat{a}_i^0 \hat{b}_k^0}.$$

The model presents a nonidentifiability problem: if a_i is multiplied by a constant and b_k is divided by the same constant, the likelihood does not change. We identify the parameters so that b_k represents the total links that involve subpopulation k. To achieve this, we normalize our parameters so that total number of links involving the name groups, that is, the sum of the b_k's associated to the name groups, equals their true proportion in the US population. We refer readers to Zheng et al. (2006) for further details regarding the normalization and to McCormick et al. (2009) for demographic information about the subpopulations.

2.4 Residuals

Residuals in standard regression are defined as the difference between the observed and expected values. In this study we define the residuals as the difference between the square-root of actual responses and their expected values under the null model, that is,

$$r_{i,k} := \sqrt{y_{i,k}} - E(\sqrt{Y_{i,k}}), \tag{2}$$

where $y_{i,k}$ (and equivalently $Y_{i,k}$) are defined as in Equation (1) as the midpoints of the possible responses. The square-root is introduced to stabilize the variance.

A small residual means that a respondent knows about as many people in a certain subpopulation as would be expected from her network size. In an Erdös-Renyi type model, all residuals would be expected to be very small. While the absolute counts, $y_{i,k}$, depend not only on the social network structure but also on the respondent's network size, the residuals $r_{i,k}$ tend to capture more the network structure.

3. Results

In the previous sections we presented two candidate methods of measuring proximity using indirectly observed network data, counts and residuals. Here, we explore the different types of information about the network given by each of these methods. We contend that, in accounting for degree, the residuals measure the respondent's average social proximity to members of the subpopulation. Without adjusting for degree, the counts measure directly the frequency of interaction between a respondent and members of a subpopulation. We then examine these measures as predictors of opinions, and in doing so further explore the impact of social influence on opinion formation.

We analyze the information about social structure in counts and residuals in Section 3.1 and compare their predictive power on opinion formation in Section 3.2. Last, we discuss in Section 3.3 a way to detect sampling bias in a survey relying on ARD questions.

3.1 Measuring social structure with ARD

Using ARD, we observe only the aggregated number of ties between a respondent and a particular subpopulation. The indirect nature of our data make the standard network-based measures of social structure inapplicable. Instead, we compare two measures based on ARD. Our raw counts reflect the number of interactions between a respondent and the subpopulation of interest. In general, we posit that respondents with higher frequency of interaction with a subpopulation are more proximate. While it does not require any additional modeling, the raw counts do not account for the total volume of a respondent's ties. In contrast, residuals as defined in Section 2.4 adjust for the respondent's network size and the relative size of the subpopulation. The remaining information represents the tendency for a respondent to know someone in the subpopulation in excess of what would be expected for someone with their network size. We contend that the model residuals more reasonably represent social structure than the raw counts, which do not separate the structure of overdispersion from the degree distribution and relative size of the subpopulation in question. Instead, the raw counts indicate a more coarse level of knowledge of, or exposure to, the subpopulation.

To understand the information contained in these measures, we first analyze patterns in the measures across the thirteen subpopulations. We consider hierarchical clustering (see Venables and Ripley (2002), for example) using Kendall's Tau as a distance measure. We apply the clustering algorithm to both the residuals and the counts and consider both the final pairing and the levels at which particular subpopulations break in the dendrograms as evidence in similarity in profiles. Subpopulations breaking at lower levels of the tree, for example, are considered more similar. After standardizing the counts and residuals, we also apply multidimensional scaling (see Hastie et al. (2001)). We use two dimensions for an easily interpretable visual display of the similarity between profiles for the subpopulations.

Figure 2 shows multidimensional scaling and hierarchical clustering based on the (standardized) residuals and counts of number known in each of the subpopulations. Though the primary comparison for the multidimensional scaling plots is still within each graph, we use a common center and rotation to facilitate comparison of the general pattern across graphs.

In Figure 2, the names are also more widely spaced for the residuals than for the counts. The similarity of the position of the names within each graph is a

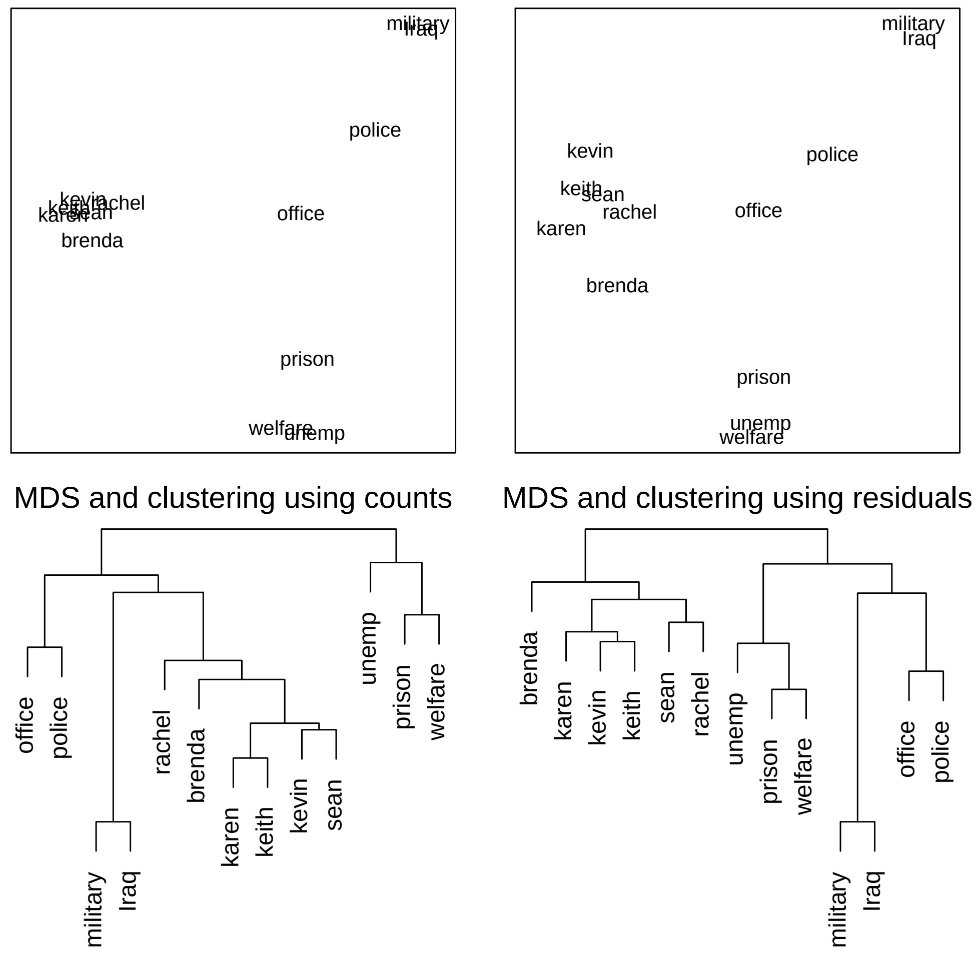

Figure 2: Hierarchical clustering and multidimensional scaling (MDS) using raw counts and model residuals. For ease of comparison, we use a common orientation rotated based on the first names and centered using the number known who have run for office. The spacing of the names is more appropriate when using residuals than counts, indicating that the information about social structure contained in the counts is confounded with degree.

bit misleading. Aside from the male names being closer to the male names and female names being closer to other female names there is little reason to believe that the names should be socially close. This result is consistent, however, with the finding of Zheng et al. (2006) that the residuals for the names are slightly correlated. One possible explanation is that some people remember names better than events. Nonetheless, the six names are nearly on top of one-another for the counts. For the residuals, the names are still close together, but noticeably less than for the counts. The counts are confounded by degree, or network size, and thus display less resolution than the residuals. The distance between the names in the residual plot (right) is more reasonable, where the “military” and “Iraq” populations are closer together than “Keith” to “Kevin,” for example. Similarly, the names are farther away from one another and from the subpopulations on the dendrogram for the residuals, whereas their position is more similar to the subpopulations on the dendrogram for the counts.

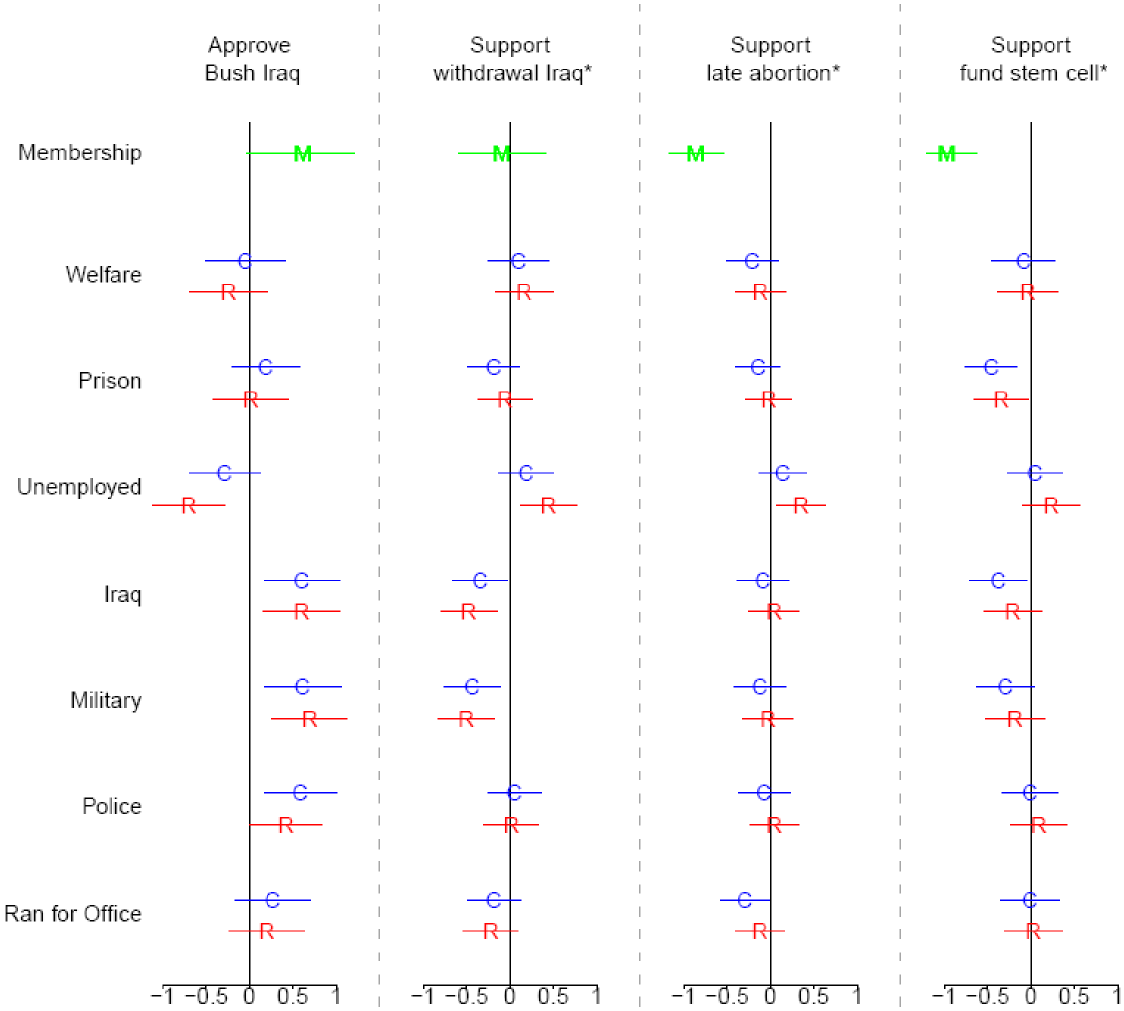

Figure 3: Coefficients and standard errors for counts, residuals and memberships. These are obtained from the regression of the respondent's opinion once on counts and once on residuals for each of the subpopulations, and once on the membership. "C" is a coefficient for counts, "R" for residuals and "M" for membership. Membership for the Iraq questions means that the respondent or an immediate family member is an active member of the military. For the remaining two questions a respondent was considered a member if they attend church regularly. For the three questions with a star, respondents could answer in the affirmative, the negative or say "not sure." For the remaining question, respondents could answer either in the affirmative or the negative. Overall the signal is most pronounced for the military and unemployed subpopulations. In both cases the coefficients for residuals tend to be more extreme than those for the counts.

The distinction between the left and right panels of Figure 2 indicates that the counts and residuals convey different information about the responses. In controlling for degree, the right plot of Figure 2 represents additional structure once total network volume has been accounted for. In the counts, much of the information about social structure is masked by degree. There is, in essence, no way of knowing if a respondent who reports knowing a large number of members of a subpopulation could be proximate to the subpopulation, or could simply have a very large network.

The counts, therefore, represent a respondent's exposure, or level of knowledge, of the subpopulation group. In contrast, the residuals represent connectedness with the subpopulation in excess of the expected, which indicates social structure more directly.

3.2 Social structure and political ideology

We now consider the implications of the distinct information the two qualities provide for predicting respondents' opinions. Figure 3 presents coefficients and standard errors for regression models predicting opinions. For each of the subpopulations, the respondent's opinion is regressed once on the counts and once on the residuals. In each model, we also control for the respondent's political ideology, political party, and demographic characteristics.

Opinion formation, as with influence processes in general, depends on a highly dependent series of interactions between connected actors. Any one actor's opinion about stem cell research, for example, will likely be influenced by the opinions of others in the actor's network. Exactly how this influence manifests as a change in opinion, the so-called substantive nature of influence (Burt, 1987), is a challenging problem. Certainly, not all actors in the respondent's network will have equal influence over the respondent.

We begin, therefore, by establishing that the composition of an individual's social network impacts their opinions using our most direct measure of social proximity—group membership. Those who attend church regularly were more likely to hold views consistent with major religious groups. The regression shows a significant negative relationship between the frequency of church attendance and being in favor of late abortion (standardized coefficient -.86 and standard error .16). Regressing being in favor of federal funding for stem cell research on the frequency of church attendance indicates that frequent churchgoers were less likely to be in favor of stem cell research (standardized coefficient -.97 and standard error .17). Also, knowing that an individual is serving in the military or has immediate family members in the military, for example, establishes that many of the respondent's closest contacts are also members of the service. Having immediate family members currently serving in the military has a positive influence on approving Bush's handling of Iraq (standardized coefficient .60 and standard error .30), for example.

These results are consistent with the work of Erickson (1988), and the preceding work of Moscovici (1985), which suggest that an individual's decisions are guided by their comparison to groups of similar peers. In these examples with group membership, similarity is literal—churchgoers unite through similar beliefs, service members have a common set of institutionalized experiences. Both counts and residuals measure a more generalized similarity based on the types of people with which a respondent interacts and measured by their social proximity. In the context of marriage, Kalmijn (1994) contends that the cultural aspects of social status are more influential than economic aspects in the assortative matching process. That is, an individual who is culturally more similar to members of a higher social class is a more valuable partner than one who is of a higher economic class. In our context, we use social distance as a proxy for cultural similarity. We would expect, therefore, to see that those who are not members of the service themselves, but who are socially close to members of the armed services would have opinions more similar to service members than others outside the service, even thought they share membership status with non-service members. In our sample, there were very few (< 5) individuals who were socially close to the military but were not service members themselves, so we were unable to examine this relationship directly. Instead, we consider the impact of being socially close to a subpopulation in general, whether that distance comes from membership in the group or other sources. We suggest that future work in this area consider this question when designing survey instruments and computing sample size.

We now turn our attention to our two measures of social proximity, counts and residuals. Using both measures, we found evidence of a relationship between social proximity and respondent opinion. For example, being more socially proximate to those serving in the military was associated with approving of George W. Bush's handling of the war in Iraq (standardized coefficient for residuals .70 and standard error .22 and for counts .61 with standard error .22) and being more likely to support withdrawal from Iraq (standardized coefficient for residuals -.50 and standard error .16 and for counts -.43 with standard error .16). Though the differences are not statistically significant, the magnitudes for military-related opinions were smaller for those socially close to individuals serving in Iraq than for those close to the military in general (standardized coefficient for approving of Bush was, using residuals, .60 with standard error .22 and with counts a coefficient of .60 and standard error .22 for Iraq for example). For both of these subpopulations residual coefficients were generally more extreme than those for counts for the majority of questions. We posit that the counts are confounded by the respondent's degree. With counts, an individual who has a high degree and high number of associations in the military receives the same social distance as someone who has a very small degree and a high number of associations. We do not expect degree to be associated with an individual's opinion, and thus we would expect the opinions of the individual with a larger degree to be less homogenous, which would result in coefficients of smaller magnitude. In other words, people with high residuals who support a position are socially close to the subpopulation in question. Individuals with high counts may be socially close but they may also simply have a high degree.

We found similar patterns in counts and residuals when measuring social distance to the unemployed, even though the signal direction is nearly the exact opposite of military. We found that those socially close to the unemployed were less likely to support George W. Bush's plan for Iraq (standardized coefficient for residuals -.70 and standard error .21 and for counts -.29 with standard error .20) and more willing to support abortion rights (standardized coefficient for residuals .35 and standard error .14 and for counts .14 with standard error .14).

3.3 Latent sampling bias

Despite the overlap in the subpopulations, the signal is much weaker for the welfare group than for the unemployed. Even more surprising is that the coefficients for counts in the prison subpopulation are typically more extreme than those of the residuals. We believe that the weaker signal in these subpopulations could be partially attributable to network-based sampling bias, which we now describe.

Since this is an internet survey, there were additional efforts to ensure a representative sample, as discussed in Section 2.1, we found that individuals who are socially close to individuals in prison or on welfare were under-represented. This fact is perhaps not surprising since both of members of both of these subpopulations are often impoverished and thus they, or individuals they are socially close to, may have difficulty accessing the internet. These observations reveal a latent bias in the sampling procedure of this internet survey. Figure 4 displays the actual fractional subpopulation size (see the Appendix of McCormick et al. (2009)) against the fractional subpopulation size estimated by the Zheng et al. (2006) model. The majority of the subpopulations are reasonably estimated; yet, prison and welfare are significantly underestimated.

Our results indicate that the survey includes too few individuals who are socially close to those on welfare and in prison. Since the residuals measure this social closeness, the residuals for these two subpopulations should be rather uninformative since the people who are truly tied to these subpopulations are not in the survey. Figure 4 also indicates that unemployed people have smaller social networks, or people cannot always identify their unemployed acquaintances as such. Unemployment is also a transient status at many levels of society, making it more likely that a respondent would interact with someone who is unemployed than in the more segregated subpopulations of welfare or prison. Transmission errors (Killworth et al., 2003, 2006) may also contribute to the under-representation of individuals who are socially close to those in prison and on welfare. Such errors occur when a respondent knows a member of a particular subpopulation but is unaware that the person belongs to the subpopulation. Given the stigma associated with belonging to these subpopulations, individuals may be unlikely to discuss their membership with anyone besides their most trusted confidants.

4. Discussion

We propose the use of "How many X's do you know?" surveys (aggregated relational data, ARD) to measure the connections between individuals and subpopulations of interest. Using a continuous measure like proximity rather than discrete categories to represent social structure introduces ambiguity from multiple definitions of social distance (Bottero and Prandy, 2003). We have demonstrated this ambiguity in the ARD context. We contend that the raw counts reflect a respondent's exposure to a subpopulation while the residuals, in adjusting for degree, more reasonably represent social structure.

As a template, we also consider the impact of social structure on political opinions using ARD. From a social influence perspective, both the residuals and the counts are measures of social proximity and underlying the social proximity is a substantive process that influences opinions. Being socially close to some subpopulations may influence the opinions of some respondents more than others, for example, because of the type of "social power" represented by the tie as in French and Raven (1959). Respondents who are in the military, for example, may be particularly likely to be influenced by being socially close to others in the military because they feel empathy based on their common experiences.

We also discovered evidence of latent sampling bias in our survey. The key feature of this type of bias is that it is based on the social proximity of a respondent to a particular type of group and not on demographic characteristics. A potentially lucrative direction for future work would involve using ARD to detect hidden sampling bias. More importantly, if one could reliably estimate the bias then a re-weighting scheme could be proposed to correct for it.

In this paper we have addressed how statistics derived from ARD represent information about social structure. There are also additional issues associated with the quality of ARD that we have not addressed. Respondents may know someone who is a member of a subpopulation (diabetics, for example) but not know the

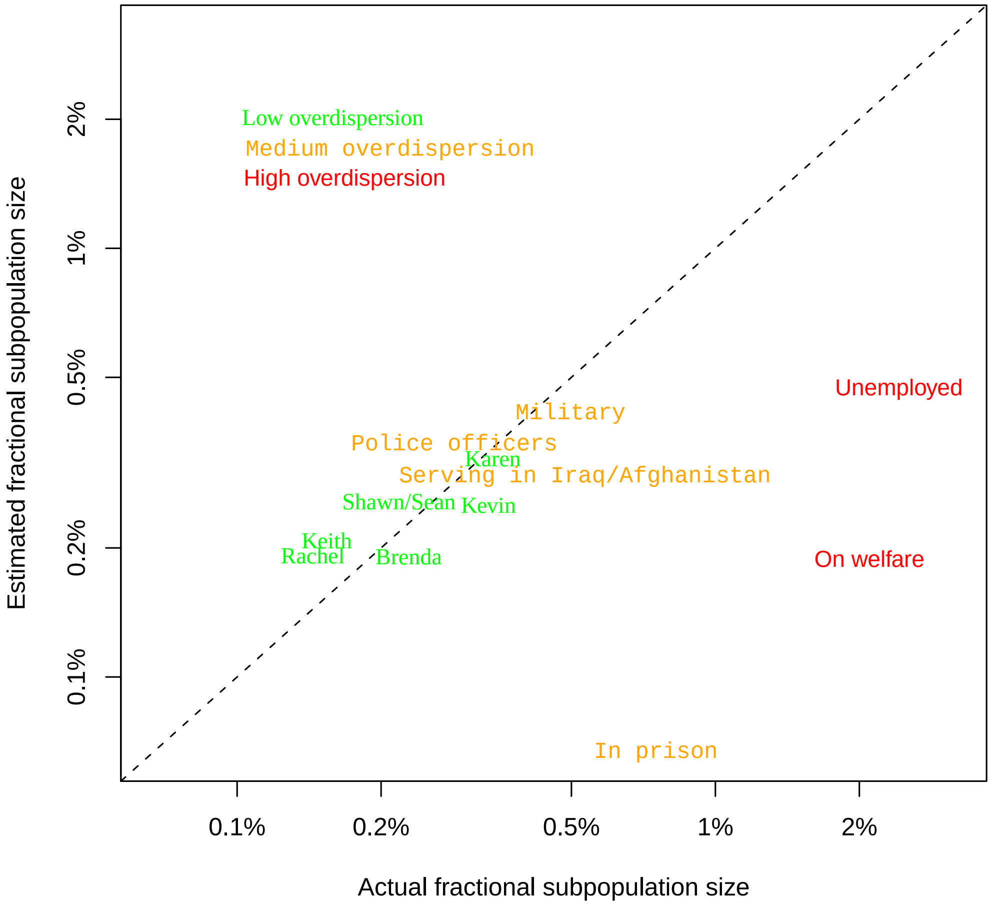

Figure 4: Comparison (on a logarithmic scale) of the estimated subpopulation sizes to their actual sizes in the US. The figure suggests the existence of a hidden sample bias; although the sample is representative using several demographic characteristics it includes too few individuals who are socially close to those in prison and on welfare.

person is a member of the subpopulation. We refer readers to McCormick et al. (2010) for a review of these issues and recent work to address them.

References

Bottero, W. and Prandy, K. (2003). Social interaction distance and stratification. *The British Journal of Sociology*, 54(2):177–197.

Burt, R. S. (1982). Studying status/role-sets using mass surveys. In Burt, R. S. and Minor, M. J., editors, *Applied Network Analysis*, chapter 5, pages 100–118. Sage Publication.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92:1287–1335.

DiPrete, T. A., Gelman, A., McCormick, T. H., Teitler, J., and Zheng, T. (2010). Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*, To Appear.

Erickson, B. H. (1988). The relational basis of attitudes. In Wellman, B. and Berkowitz, S. D., editors, *Social Structures: A Network Approach*. Cambridge University Press.

French, J. R. P. and Raven, B. H. (1959). The social basis of power. In Cartwright, D., editor, *Studies in social power*. University of Michigan Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Kalmijn, M. (1994). Assortative mating by cultural and economic occupational status. *American Journal of Sociology*, 100(2):422–452.

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelly, 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 Shelly, 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.

Killworth, P. D., McCarty, C., Johnsen, E. C., Bernard, H. R., and Shelley, G. A. (2006). Investigating the variation of personal network size under unknown error conditions. *Sociological Methods & Research*, 35(1):84–112.

Laumann, E. (1979). Network analysis in large social systems: Some theoretical and methodological problems. In Holland, P. and Leinhardt, S., editors, *Perspectives on social network research*, pages 379–402. Academic Press.

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

McCormick, T. H., Moussa, A., Ruf, J., DiPrete, T. A., Gelman, A., Teitler, J., and Zheng, T. (2009). Comparing two methods for predicting opinions using social structure. In *Proceedings of the Joint Statistical Meetings*. American Statistical Association.

McCormick, T. H., Salganik, M. J., and Zheng, T. (2010). How many people do you know?: Efficiently estimating network size. *Journal of the American Statistical Associaton*, To Appear.

MIT Web (2007). CCES. `http://web.mit.edu/polisci/portl/cces/index.html`.

Moscovici, S. (1985). Social influence and conformity. In Lindzey, G. and Aronson, E., editors, *The Handbook of Social Psychology*, volume 2, pages 347–412. Random House.

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

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics With S*. Springer.

Zheng, T., Salganik, M. J., and Gelman, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estimate social structure. *Journal of the American Statistical Association*, (101):409–423.