Comparing two methods for predicting opinions using social structure

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

Abstract
We explore the influence of social structure on opinions about contemporary political issues. We use Aggregated Relational Data (ARD), introduced by Killworth et al. (1998a) as "How many x's do you know?" questions, on an internet survey provided by the Cooperative Congressional Election Study (CCES). Here, "x" represents a subpopulation of interest such as individuals in prison or those serving in the military. We measure the social distance between respondents and these subpopulations in two ways: (i) based on exposure and (ii) based on social closeness. In defining social closeness we use residuals from the overdispersed Poisson model presented in Zheng et al. (2006). We explore the information about social structure these measures contain and examine their implications for respondents' political opinions. As a byproduct of our analysis, we also propose a method for detecting sampling bias in internet surveys using ARD.

Key Words: Aggregated relational data, hierarchical model, opinion formation, overdispersed Poisson distribution, sample survey, social network

1. Introduction

Opinion formation is a complicated, multi-faceted process where individual characteristics interact with a constantly evolving, complex social environment. 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.

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. French and Raven (1959) link influence to the broadly defined concept of "social power." Some forms of social power, such as the ability to reward, pertain to behaviors but not necessarily opinions. A manager or employer, for example, has significant influence over an employee's actions but

*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, School of Social Work, 1255 Amsterdam Avenue, New York, NY 10027
¶Corresponding author. E-mail: tzheng@stat.columbia.edu

not necessarily over their opinions. Instead, other manifestations of social power likely have a greater impact on opinion formation, such as a perceived position of legitimacy or expertise.

Asch (1956) and many subsequent authors suggest a different view of the influence process. Asch (1956) contends that opinion formation involves evaluating a "social consensus" arrived at by others (Friedkin and Johnson, 1990). Erickson (1988)'s work, and the preceding work of Moscovici (1985), suggest a more complicated process where only certain individuals are members of the reference group. In ambiguous situations, individuals' decisions are guided by their comparison to groups of similar peers. Though the comparison to a reference group underlies both perspectives, group membership is much more selective in Erickson (1988)'s framework and defining similarity between actors becomes fundamental.

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.

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 in many social science applications. Observing network relations indirectly through Aggregated Relational Data (ARD) is one increasingly popular alternative. 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., 2009). ARD do not require any specific sampling technique and are easily integrated into standard surveys.

Since ARD measure network features indirectly, the standard representations of proximity for complete network data are no longer well defined. Instead, we compare characteristics of indirectly observed network data which could indicate social proximity. A natural starting point is the frequency of interactions between an actor and members of a given subpopulation, which is measured directly by ARD. We also consider residuals from a variation of the model in Zheng et al. (2006) described in Section 2.4. We consider both types of data as potential measures of social proximity and examine the relative information about social structure each quantity provides. In both cases the fundamental assumption of social influence—that more proximate actors are more influential—remains unchanged. With ARD,

however, we assess the expected proximity to a specific subpopulation.

In the coming sections we explore the social structure described by the two potential proximity measures, counts and residuals. We also consider the influence of this structure on respondents' opinions. We describe the data in Section 2.1 and present in Sections 2.2 and 2.3 the modified overdispersed Poisson regression model. In Section 2.4 we define the residuals which we consider as one possible measure of social structure. Section 2.5 describes the regression techniques we use to relate our measures of social proximity to respondents' opinions. In Section 3 we discuss our results. Section 4 uses our findings to explore substantive underpinnings of the influence modeled by our two types of proximity measures and provides directions for future work. In the appendix we present the sources from which we have obtained the true group sizes in the US population.

2. Dataset and model

2.1 Dataset

The dataset is provided by the Cooperative Congressional Election Study (CCES)[1], a large national online survey created by thirty universities. 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. Common Content 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. 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.

Respondents are 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 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

[1] See `http://web.mit.edu/polisci/portl/cces/index.html`.

defined as knowing a person's name and being willing to stop and to talk to him for at least a moment. For $n = 1000$ respondents[2] and $K = 13$ subpopulations we are provided with the number of persons the respondent i knows in the subpopulation k. More precisely, the respondents were presented with five possible choices: 0, 1, 2 to 5, 6 to 10, and more than 10, which constitutes an interval-based dataset.

For illustration, 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 a lot of Rachels and a lot of people on welfare, with a more pronounced overdispersion for the "Welfare" group than for the "Rachel" group.

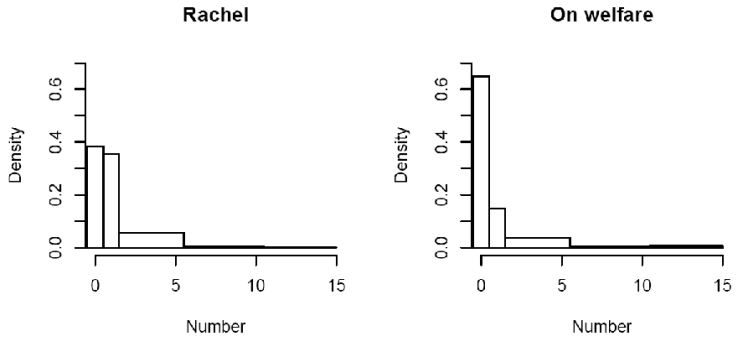

Figure 1: Histograms showing the distribution of the answers of two typical questions: "How many Rachels do you know?" and "How many people on welfare do you know?" The larger ratio of the heights of the first and second bar in the right histogram compared to the left one indicates a higher overdispersion in the "Welfare" group.

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 is Poisson distributed with intensity $a_i b_k$, where a_i is the expected number of persons known by respondent i and b_k is the expected number of social links involving subpopulation k divided by the total expected number of social links (popularity of subpopulation k).

However, such a model is too simplistic for the purpose of our study because it assumes that all individuals have the same propensity to form social links. Our survey dataset has a strong potential for overdispersion, that is, the answers display more variability than predicted by the Erdös-Renyi model which is mainly due to natural clusters in the data (socioeconomic groups, gender groups, ...). In other words, there are groups where many respondents do not know anyone in this group but some respondents know many people in this group, see our discussion of Figure 1. Therefore, we use the more sophisticated model developed by Zheng et al. (2006).

[2]Six of the respondents did not answer any of the questions. So we are left with 994 respondents.

They 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 $\frac{1}{\omega_k - 1}$, 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 model[3] by putting a noninformative prior for the parameters μ_α, σ_α, μ_β, and σ_β.

Denoting[4]

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

and

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

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

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

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

[3]For an overview of hierarchical models, see Gelman and Hill (2006).

[4]Our likelihood differs from the one presented in Zheng et al. (2006) because we have interval-based data and therefore have to adjust the likelihood slightly. The likelihood was worked out by DiPrete et al. (2009).

by the usual Poisson regression:

$$\begin{aligned}\hat{a}_i^0 &= \frac{1}{\sum_{k=1}^{K} c_k} \sum_{k=1}^{K} c_k y_{i,k} \text{ with } c_k = \frac{1}{\sqrt{\overline{y_{.,k}}}} \\ \hat{b}_k^0 &= \frac{1}{\sum_{i=1}^{n} d_i} \sum_{i=1}^{n} d_i y_{i,k} \text{ with } d_i = \frac{1}{\sqrt{\overline{y_{i,.}}}},\end{aligned}$$

where we set

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

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.[5]

2.4 Residuals

The classical definition of residuals is the difference between the observed and expected values. In this study, in order to capture the overdispersion of the data, we define the residuals as the difference between the square-root of actual responses and their expected value under the null model, that is,

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

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. By factoring out the mean network size of each individual, residuals can be used to avoid confounding with the network size.

2.5 Two measures of social structure

We describe in this section the methodology we follow to study the influence of social proximity on political opinion formation. More precisely, we compare the predictive power of two measures of social proximity: counts and residuals.

To understand the information contained in these measures, we first analyze patterns in the measures across the thirteen subpopulations. We consider hierarchical clustering (Venables and Ripley, 2002) using Kendall's Tau as a distance measure. We apply the clustering algorithm to both the residuals and the counts and consider

[5]For details about the normalization, see Zheng et al. (2006).

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 (Hastie et al., 2001). We use two dimensions for an easily interpretable visual display of the similarity between profiles for the subpopulations.

We then turn to the study of political opinion formation. We select a set of political opinions among the ones asked in the survey, as for example whether the respondents approve Bush's policy in Iraq, if they support the Patriot Act, or if they are in favor of a time table for withdrawal from Iraq. We predict the opinions by performing various ordinal logistic regressions, always controlling for demographic and social factors such as income, gender and political ideology. Missing data in the control variables are implemented based on a simple regression on the other predictors. We regress each opinion once on the counts for each of the subpopulations (Rachel, unemployment, welfare, etc.) and once on the corresponding residuals. In total 26 regressions are performed for each opinion, 13 using counts and 13 using residuals.

We perform all computation using **R** (R Development Core Team, 2009).

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 then examine these measures as predictors of opinions, and in doing so further explore the impact of social influence on opinion formation.

Using ARD, we observe only the aggregated number of ties between a respondent and a particular subpopulation. Our raw counts reflect the number of interactions between the 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 use the Zheng et al. (2006) model to adjust for the respondent's network size. The remaining information, then, 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 while the raw counts indicate a more coarse level of knowledge of, or exposure to, the subpopulation.

In exploring the distinction between the information in the raw counts and model residuals, we first examine response profiles from the two types of data. Figure 2 shows multidimensional scaling (Hastie et al., 2001) 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 used a common center and rotation to facilitate comparison of the general pattern across graphs. The similarity of the position of the names within each graph is a 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 Zheng et al. (2006)'s finding that there is a slight correlation amongst the residuals for the names. One possible explanation is that

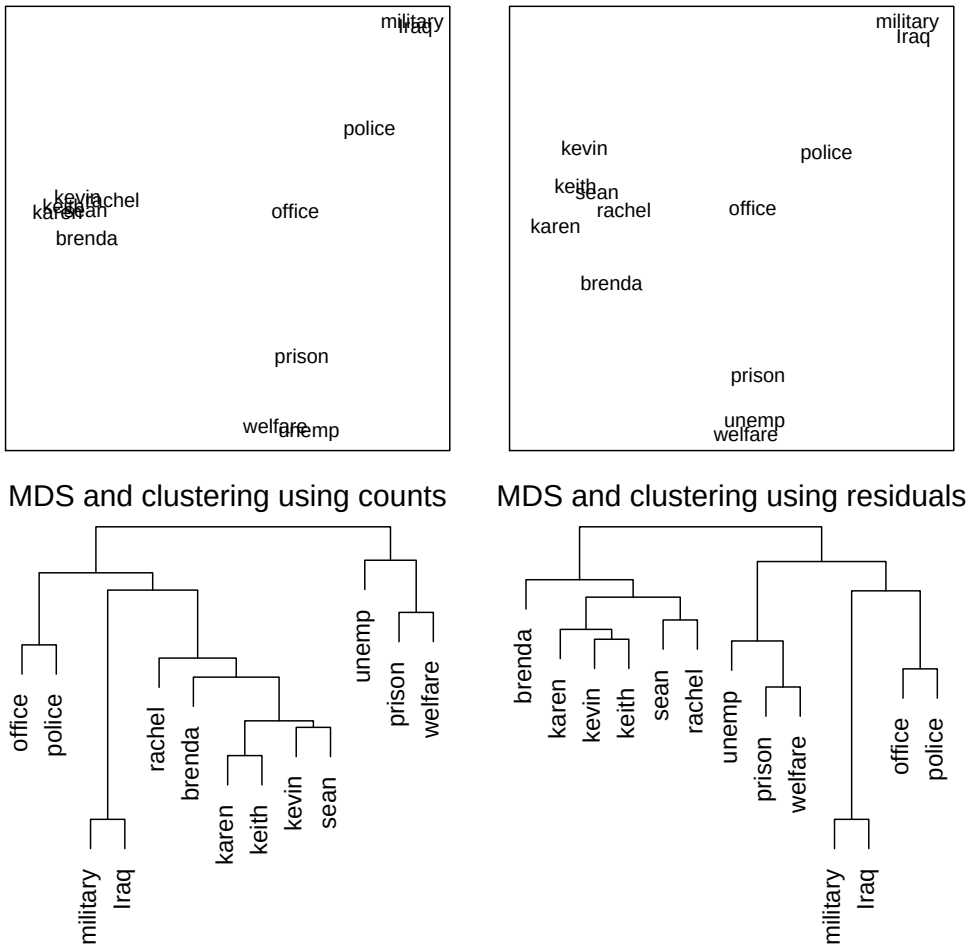

Figure 2: Hierarchical clustering and multidimensional scaling using raw counts and model residuals. The multidimensional scaling plots have a common center and rotation. The spacing of the names is more appropriate using residuals than with counts, indicating that the information about social structure contained in the counts is confounded with degree.

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. 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

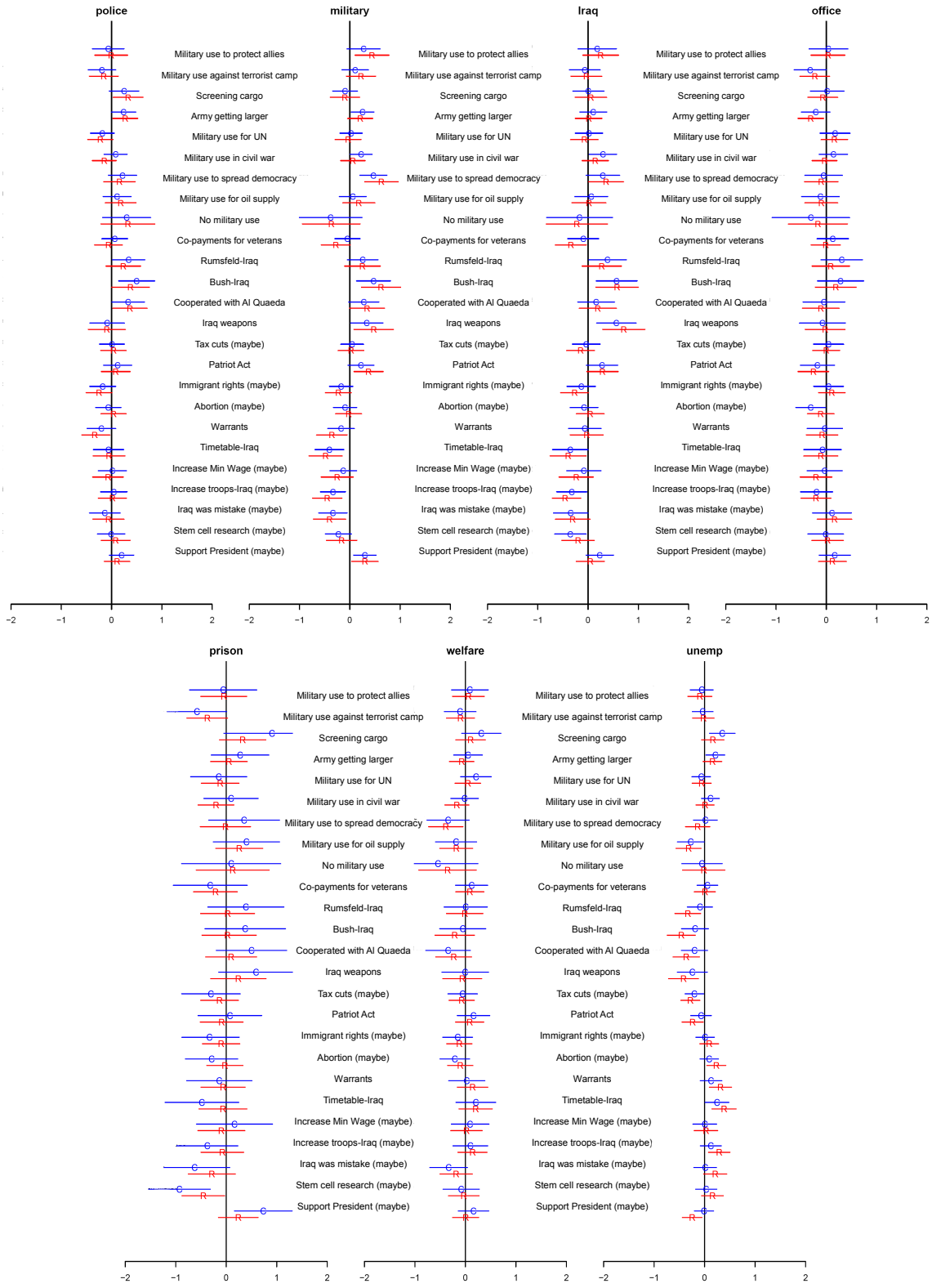

Figure 3: Coefficients and standard errors for regression coefficients. "C" is a coefficient for counts and "R" is for residuals. Overall the signal is most pronounced for the military and unemployed subpopulations. In both cases the coefficients for residuals tend to be more extreme than those for the counts.

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. 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, we regressed both the counts and the residuals on the respondent's opinion. In each model, we also control for the respondent's political ideology, political party, and demographic characteristics. With 25 opinion questions and seven subpopulations, not including first names, displaying the data in a way which facilitates interpretation is in itself challenging. We used hierarchical clustering to group subpopulations and opinion questions with similar response profiles and standardized all covariates to ensure coefficients are on the same scale.

Overall, we found that a respondent's political party and ideology were both highly influential. Even accounting for these factors, however, indirect measures of social structure revealed information of potential interest to social scientists. Individuals who were more connected with the unemployed, for example, were more likely to oppose policies currently associated with the conservative political establishment. That these effects persist even after accounting for characteristics of the individual suggest that the respondents' social network could have a significant role in directing the formation of the their opinions. We also found a distinction between the opinions of those who were connected to individuals in the military and those serving in Iraq. This distinction could be of particular interest given the overall contentiousness of the war efforts. With respect to the distinction between the information captured by counts and residuals, we first consider the two subpopulations where the overall signal is strongest—those serving in the military and the unemployed. The direction of the signal, however, in military is nearly the exact opposite of unemployed, which again may itself be of interest to social scientists. For both of these subpopulations residual coefficients were more extreme than those for counts for the majority of questions. In the subpopulations with a weaker signal the residuals and counts appear to perform about equally.

Given the strong signal and improved performance of the residuals for the unemployed, the results for two related groups, welfare and prison, are surprising. 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.

These observations reveal a latent bias in the sampling procedure of this internet survey. Figure 4 displays the actual fractional subpopulation size of the subpopulations under consideration 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. Since this is an internet survey, there were additional efforts to ensure a representative sample, as discussed in Section 2.1. Nonetheless, our results indicate that the survey includes too few individuals who are socially close to those one 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 is underestimated. We posit that the residuals preserve a strong signal in this case,

however, because the current dismal economic climate has dispersed the unemployed throughout all social strata. 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 underrepresentation of individuals who are socially close to those in prison and on welfare. Such errors occur when a respondent knows a member of 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.

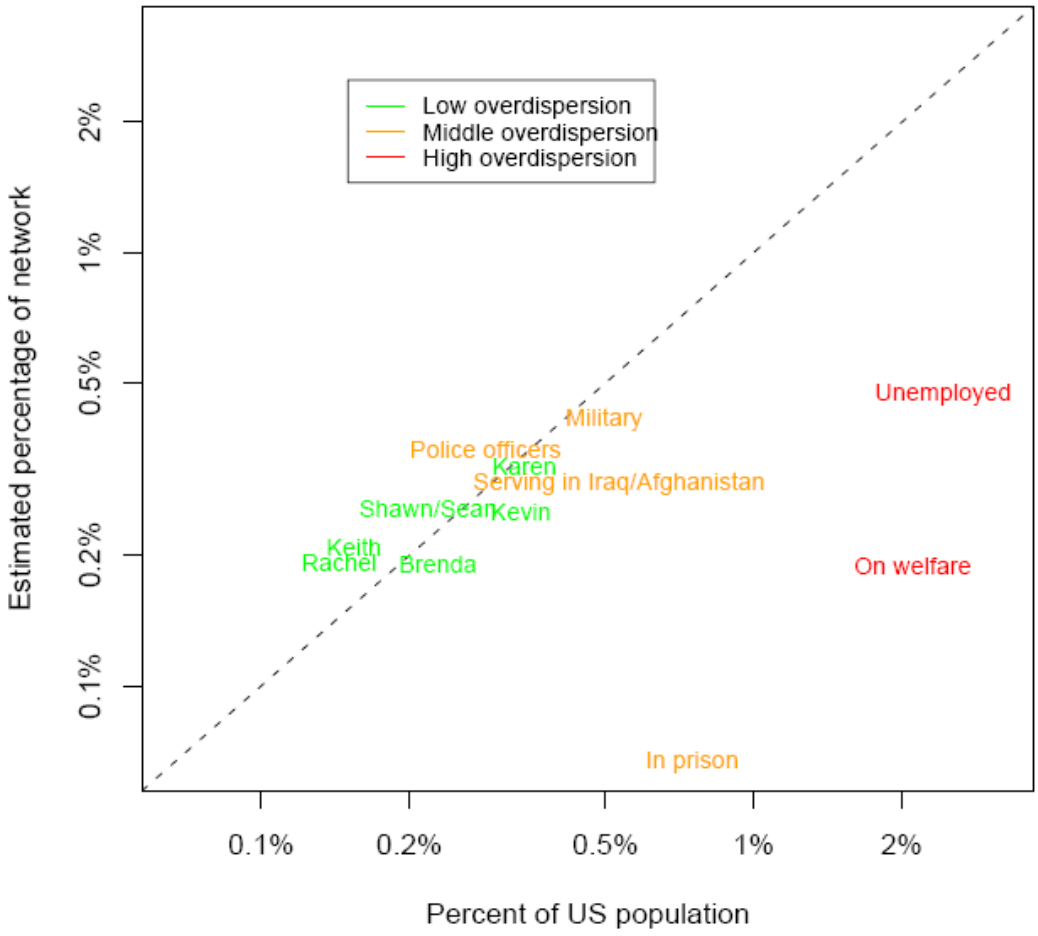

Figure 4: Fractional subpopulation sizes representing hidden sampling bias. Although the sample is representative using several demographic characteristics it includes too few individuals who are socially close to those in prison and on welfare.

4. Discussion

We consider the impact of social structure on political opinions using Aggregated Relational Data. We use counts and residuals to represent social structure and contend that the model residuals more reasonably represent social structure while the raw counts indicate a more coarse level of knowledge of, or exposure to the subpopulation.

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 French and Raven (1959)'s "social power" represented by the tie. 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.

During the course of our analysis, we also discovered evidence of sampling bias in this survey. Although measures were taken to ensure that the sample is representative across numerous characteristics, 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. 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.

A. Model calibration

In order to be able to compare our estimates of the group sizes $\hat{b}_k$ of links involving group k to the true US proportions in Figure 4, we need the true numbers. Except for the group "run for office" this information is publicly available.[6] For sake of convenience we used 281 Million as the number for the size of the U.S. population. In the following we present our data sources:

- Names: U.S. Census Bureau (retrieved from www.census.gov/genealogy/names/dist.female.first and www.census.gov/genealogy/names/dist.male.first, different spellings are taken into account)
- Unemployed: 6,800,000, U.S. Department of Labor, Bureau of Labor Statistics (retrieved from http://www.bls.gov/news.release/empsit.nr0.htm)
- Welfare: 5,901,000, U.S. Department of Health and Human Services, Administration for Children and Families (retrieved from http://www.acf.dhhs.gov/news/tables.htm)[7]
- Police: 799,320, U.S. Department of Labor, Bureau of Labor Statistics (retrieved from http://www.bls.gov/oes/current/naics3_999000.htm)
- Prison: 211,031,000, U.S. Department of Justice (obtained from Prisoners in 2006, 2007)
- Military: 1,400,000, U.S. Census Bureau (retrieved from http://www.census.gov/Press-Release/www/2003/cb03-ff04se.html)

[6]Most of the information can be found on the internet. The number for people serving in Iraq and Afghanistan was kindly provided by the Office of the Assistant Secretary of Defense.

[7]This is a number from 2000, which is the last available number. It was strongly declining and might have been much lower by 2006. This would be in favor of our model-based estimator which estimates a smaller number.

- Iraq / Afghanistan: 1,500,000, U.S. Department of Defense[8]

References

Asch, S. E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs*, 70.

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. (2009). Segregation in social networks based on acquaintanceship and trust. *Columbia Population Research Center Working Paper*, 09-09.

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.

Friedkin, N. E. and Johnson, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–205.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

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

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.

[8]This number is estimated as the difference of the cumulative number of soldiers up to October 2007 and a fraction of the soldiers who are serving in Iraq, Afghanistan and the area around there in 2007. We only reduce it by a fraction since some of the soldiers right now have already been deployed.

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.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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.