NBER WORKING PAPER SERIES

INFERRING THE IDEOLOGICAL AFFILIATIONS OF POLITICAL COMMITTEES VIA FINANCIAL CONTRIBUTIONS NETWORKS

Yiran Chen
Hanming Fang

Working Paper 24130
http://www.nber.org/papers/w24130

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2017

We would like to thank Ben Connault, Frank DiTraglia, Yue Hou, Karam Kang, Michael Leung, Tengyuan Liang, Sarah Moshary, Andrew Shephard, Rakesh Vohra, Yiqing Xu and participants in the North American Econometric Society Summer Meeting (2017, St. Louis) for helpful discussions and suggestions. All remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Inferring the Ideological Affiliations of Political Committees via Financial Contributions Networks
Yiran Chen and Hanming Fang
NBER Working Paper No. 24130
December 2017
JEL No. D72,D85,P16

ABSTRACT

About two thirds of the political committees registered with the Federal Election Commission do not self identify their party affiliations. In this paper we propose and implement a novel Bayesian approach to infer about the ideological affiliations of political committees based on the network of the financial contributions among them. In Monte Carlo simulations, we demonstrate that our estimation algorithm achieves very high accuracy in recovering their latent ideological affiliations when the pairwise difference in ideology groups' connection patterns satisfy a condition known as the Chernoff-Hellinger divergence criterion. We illustrate our approach using the campaign finance record in 2003-2004 election cycle. Using the posterior mode to categorize the ideological affiliations of the political committees, our estimates match the self reported ideology for 94.36% of those committees who self-reported to be Democratic and 89.49% of those committees who self reported to be Republican.

Yiran Chen
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104
yiranc@sas.upenn.edu

Hanming Fang
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104
and NBER
hanming.fang@econ.upenn.edu

1 Introduction

Campaign finance is an integral part of the U.S. politics, and political committees are major participants in campaign finance related activities. We use the term *Political Committees* (henceforth, PCs) to refer to federal political action committees (PACs), party committees, campaign committees for presidential, house and senate candidates, as well as issue-based groups or organizations, including lobbyists or fundraisers. They collect contributions from individual donors, make contributions to other committees and candidates, and spend money for or against candidates.

U.S. campaign finance laws mandate that political committees disclose all financial transactions including the contributions they receive and their expenditures, thus rendering numerous data for analyzing their campaign related activities. However, the PCs are *not* mandated to file their party (or ideological) affiliations. Indeed, as we will detail in Section 3, nearly 60% of PCs' party affiliations are unreported. This missing data problem may generate obstacles in the study of important issues related to campaign finance. For example, for researchers who want to study the patterns of individual contributions using the individual contributions data provided by the Federal Election Commission (FEC), it is important to know the ideological affiliations of the PCs to which individual donors contribute. Nevertheless, in the Contributions by Individuals (Year 2003-2004) data released by the Federal Election Commission (FEC), about 24%, both in terms of the instance and the amount, of individual contributions go to PCs with unknown party affiliations.

In this paper, we contribute to the methodologies that aim to address the missing ideological affiliations of the PCs. We propose a method of inferring PCs' ideological affiliations from the financial transactions network among the PCs. The Contributions to Committees from Committees data, also administered by the FEC, contains the universe of all records of financial contributions among all registered PCs. We use this data set of the contribution activities to construct a financial transactions network of the PCs, where each PC is a vertex of the network, and the money flows between the PCs form the (weighted) edges. The basic idea of our method of inferring ideological affiliations of the PCs is simple. If the PCs tend to contribute more frequently to other PCs with similar ideology, then the PCs that actually filed their party affiliations, together with the structure of the observed financial transactions network, should provide information about the ideologies of the PCs with unknown ideological affiliations. Specifically, we build an economic model of link formation and transfer amount, and use the ideas of "community detection" first developed in the *stochastic block* model literature where contribution decisions (both the link formation and the weights) depend on both the observed characteristics and the *potentially* latent (for the PCs that do not file their party affiliations) ideologies of the potential contributing and receiving PCs.

Our model incorporates several new features that are absent in the existing stochastic block model literature. First, we introduce weights and rich heterogeneity in the edge formation process: the decision to make a contribution and its contribution amount is not only governed by the latent political ideologies of the PCs, but also depends on the vertex-level contextual information such as financial and institutional characteristics. Second, we model the reported party affiliations of those PCs that do self report their party affiliations as noisy measurements of their true latent ideologies. Thus, our methodology allows us to estimate the latent ideologies of all PCs, including those that self reported party affiliations.

We use three publicly available data sets in our analysis. Two are from the campaign finance record in 2003-2004 election cycle, namely the Contributions to Committees from Committees, and the Committee Master File, both maintained by the FEC. We use the first data set to construct the financial transactions network of the PCs, and we use the second data set to obtain the party affiliations of some PCs (if they self report), as well as the designations and types of all PCs. The third data set is additional industrial breakdown information of the PCs which we collected from `OpenSecret.org`.[1] Our data sets cover the *universe* of all PCs engaging in transactions with other PCs in 2003-2004. This feature of our network data has both advantages and disadvantages. An advantage is that it avoids potential bias arising from analyzing a partially sampled network, but a major disadvantage is the computational burden associated with the large network.[2] There are 5,858 vertices (i.e., PCs) in the financial transactions network, 3,727 of which did not report their party affiliations. Thus the number of potential ideological configurations is enormous, specifically, 2^{3727} even if we just allow for the binary ideology of Democrat or Republican. For similar reasons, in a Bayesian approach that delivers a probability distribution over different ideologies for each vertex (see Section 4 for details), exact estimation of the posterior mode is infeasible. We instead propose a Gibbs sampler algorithm to approximate the solution. In Monte Carlo Simulations, we demonstrate that our estimation algorithm achieves very high accuracy in recovering the latent party affiliations provided that the pairwise difference in ideology groups' connection patterns satisfy what is known as the Chernoff-Hellinger divergence criterion. We illustrate our approach using the campaign finance record in the 2003-2004 election cycle. Using the posterior mode to categorize the party affiliations of the PCs, our estimated ideological affiliations match the self reported ideology for 94.36% of those committees who self reported to be Democratic and 89.49% of those committees who self reported to be Republican.

[1] `https://www.opensecrets.org/pacs/list.php`

[2] Chandrasekhar and Lewis (2011) shows that bias arises when one works with *partially* sampled network.

Related Literature. This paper is closely related to two strands of existing literature. In terms of the research question, our paper is related to the political economy literature on the measurement of political ideologies. In terms of methodology, it is related to the statistics literature on community detection.

We first discuss the literature on the measurement of political ideology. In a seminal paper, Poole and Rosenthal (1985) proposed a measure of the ideology points (NOMINATE score) of federal legislators using the roll call data. In their paper, legislators' ideology points can differ by election cycle, and bridge legislators and bridge bills are used to ensure that the measures are comparable across time.[3] Note that Poole and Rosenthal's NOMINATE score is only available for members of Congress, which is a very small sample (around 500 legislators in each election cycle); and it is tied to voting behavior. Subsequently, ideology measurement for other political actors are proposed based on the NOMINATE score. For example, McCarty, Poole and Rosenthal (2006) combined NOMINATE score and the amount of contribution from PACs to the legislators to estimate PACs' ideology. Their proposed measure is money-weighted average of the NOMINATE scores of the legislators to whom the PAC has contributed. These measures could be biased because they do not account for PACs' contribution to losing candidates and other political actors. McKay (2008) and McKay (2010) combined NOMINATE score and the preferred votes on key roll calls published by interest groups to estimate interest groups' ideologies. Her proposed measure is the average of the NOMINATE scores of "perfectly scoring" legislators whose votes are exactly the same as the preferred votes of the particular interest group. These measures are only available for interest groups which publish their preferred votes. Additionally, they could be biased: if an interest group publishes many key votes, the number of "perfectly scoring" legislators could be too small and leads to inaccuracy; if an interest group publishes only a few key votes, the number of "perfectly scoring" legislators could be too large and artificially draws the measure toward the center.

There are other proposed methods which do not rely on the NOMINATE score. Some studies use the campaign finance data to jointly estimate candidates and PACs' ideologies. For example, McCarty and Poole (1998) proposed a measure based on PAC's contribution decision between incumbent-challenger pairs, excluding unchallenged and open seat elections which account for a significant fraction in the federal elections. Their measures are not available for candidates in these elections; and are potentially biased for PACs which contributed in these elections. More recently, Bonica (2013) proposed a method using the contributions from PACs to candidates. This

[3] "Bridge legislators" are legislators who serve multiple terms; and "bridge bills" are bills that are considered at different legislative cycles but with similar contents.

approach, from the perspective of network study, restricted the sample to a directed bipartite graph, excluding connections among PACs or candidates, as well as connections from candidates to PACs. In our paper, in contrast, we use an unrestricted network incorporating all financial connections. Moreover, he uses maximum likelihood estimation method which requires multiple observations, so he pooled observations over a period of 30 years (1980-2010) and further restricted the sample to PACs that have given to 30 or more unique candidates, and candidates who have received from 30 or more unique PACs. Pooling data over time is potentially problematic. It requires the assumption that the actor's ideology is fixed in a span of 30 years. Differently, in our paper, we make inference about a PC's ideology out of a single observation of financial transactions network: for any PC, separate ideology measures can be calculated for each election cycles. As a result, our measures are well suited for the study of the time trend of political activities. Moreover, his sample selection excludes candidates and PACs with small numbers of financial transactions. However, in our paper, we cover a more extensive scope of political actors: a PC is included as long as it has at least one financial transaction with another PC.

Other studies use data from the social media platforms. For example, Barberá (2015) used the Twitter "following" links to estimate the ideology of the political elites (accounts of candidates, parties, media, and journalists) and the mass (other individual accounts) in multiple countries on the same scale. Similar to Bonica (2013), he restricted the sample to a directed bipartite graph focusing on the mass following the elites. He used a Bayesian method and a two-stage estimation strategy that exploits the bipartite structure. In the first stage, he used a sub-network induced by a random subsample of the mass accounts and all the elite accounts, and jointly estimated their ideologies. In the second stage, holding the first stage ideology measures of the elite accounts fixed, he estimated the ideology for all the mass accounts using the full network. This two stage procedure reduces the computational intensity; but using a partially sampled network may lead to bias in the first stage, which may be further propagated in the second stage. Our approach is different in that we simultaneously estimate the ideologies of all PCs using the full network. Finally, one key difference between Bonica (2013) and Barberá (2015) and our paper is that both of these papers use spatial models. Spatial models assume *a priori* homophily: political actors with closer ideology points have higher propensity to connect. This has an especially strong implication on the centrists' behavior. It rules out the possibility for the centrists to have low propensity to connect with other centrists, and high propensity to connect with the center-left and the center-right. In contrast, our model can accommodate non-homophilic patterns. In fact, according to our estimates, the Independent PCs do have higher propensity to form financial connections with Democratic and Republican PCs than other Independent PCs.

We now discuss the literature of community detection. The main task in this literature is to classify vertices in a network into different groups based on observed connections. When repeated observations of the network are available, canonical statistical tools can be directly applied. For example, Trebbi and Weese (2015) used generalized method of moments to estimate, for each district (vertex), the fraction of insurgents in different groups. The number of parameters is of order n, the number of vertices. Though large, it is fixed when the number of network observations T grows. Therefore, standard asymptotic results and inference tools are valid.

When only one observation of the network is available, which is the case for our application, the nature of the problem is changed. In this case, the only hope for asymptotic result is to have network size $n \to \infty$; however, when the network size grows, the number of parameters (of order n) also grows, which renders canonical statistical tools invalid. Some popular approaches circumvent this issue by using model-free heuristics: minimum-cut method in Stoer and Wagner (1997) minimizes the number of edges between communities; modularity maximization method in Newman (2006) maximizes the difference between the fraction of the edges within groups and its expected value if edges were formed at random; convergence of iterated correlations (CONCOR) method in Breiger, Boorman and Arabie (1975) bisects the adjacency matrix by iteratively calculating correlation coefficients among rows (or columns); and spectral method in Newman (2013) extracts information of graph partition from the top few eigenvectors of the adjacency matrix. Though widely used in practice, these approaches have the following limitations. First of all, the first three methods all assume a priori assortative communities, i.e., denser within-community connection than between-community connection. They are not appropriate for problems where some group may have higher external connectivity. For example, in our application, the Independent PCs may engage in more transactions with Democratic and Republican PCs than with other Independent PCs. We need a model which does not assume away this possibility. Second, in the absence of a statistical model, none of these methods allows for statistical inference: we cannot state how confident we are in the obtained classification. Third, these methods cannot incorporate vertex level or vertex pair level heterogeneity beyond the latent community.

Therefore, we took a model-based approach instead - we build on the stochastic block model (SBM) initiated by Snijders and Nowicki (1997) and Nowicki and Snijders (2001). In the SBM, a network is randomly formed conditional on the underlying community structure, and the community structure itself is also stochastically generated. This model is widely accepted as a canonical model for community detection and its estimators have desirable properties. Recent studies characterized the information-theoretic threshold for exact recovery, i.e. conditions on the data generating process such that one observation of the network embodies enough information to exactly recover the community structure. Mossel, Neeman and Sly (2014), Abbe, Bandeira and Hall (2014), and Abbe and Sandon (2015) studied this problem in unweighted networks. Jog and Loh (2015) and Yun and Proutiere (2016) extended previous results to weighted networks. In particular, they show that if the Chernoff-Hellinger divergence of community pair's connection patterns is above a particular threshold, the probability of correctly recovering the entire community structure converges to 1 as $n \to \infty$. Kanade, Mossel and Schramm (2016) and Cai, Liang and Rakhlin (2017) studied this problem when a fraction of the vertices's community affiliations are revealed. Finally, Abbe (2017) provided a detailed review of the recent development in the research of exact recovery in stochastic block models. As explained earlier, exact solution to the Bayesian posterior mode is infeasible for large network, and various algorithms are proposed to approximate the solution. For example, the spectral algorithm can be viewed as an approximation to posterior mode.[4] The Gibbs sampler algorithm we use in this paper is an approximation to first obtain the posterior distribution and then the posterior mode. From the perspective of empirical implementation, a weakness of the generic SBM is the lack of heterogeneity beyond community affiliations, so we introduce pair-wise heterogeneity in a similar way as Peng and Carvalho (2015*a*) and Peng and Carvalho (2015*b*). These papers proposed a degree correction strategy to capture vertex-level heterogeneity: they use additional latent variable or observed degree to capture the popularity of a vertex, and a "popular" vertex has higher propensities to connect to all other vertices. In this paper, we allow heterogeneity at *vertex pair* level, by incorporating interactions of the observable vertex level characteristics, such as location, industry, and budget. Apart from richer specification, the use of observable characteristics has additional benefits: it entails less computational intensity than the use of latent variable, and it avoids the endogeneity problem from the use of observed degrees.

The remainder of the paper is structured as follows. In Section 2, we introduce the basic framework for the financial network among political committees, and describe the statistical model of random network formation that we will empirically implement; in Section 3, we describe the data sets used in our analysis, as well as several "naive" methods that we attempted; in Section 4, we provide the details of the Bayesian estimation procedure and arguments for identification; in Section 5, we present the Monte Carlo results; in Section 6, we describe the specifications used in our empirical implementation, and present the main empirical results; finally, in Section 7, we conclude and discuss directions for future research.

[4] The original discrete parameter space is relaxed to a sphere, and the spectral method is a solution to the relaxed problem, and is derived through derivatives. A detailed description can be found in Newman (2013).

2 Financial Network of Political Committees

We represent the money-flow network among political committees as a static, weighted and undirected graph.[5] A graph $\mathcal{G}$ consists of vertices $\mathcal{V}$ and edges $\mathcal{E}$. In our model, each vertex $i \in \mathcal{V} = \{1, ..., n\}$ represents a PC. In our application, n will represent the total number of PCs registered with the FEC, and is 5,858 for the 2003-2004 election cycle. Each edge $(i, j) \in \mathcal{E}$ is an unordered pair in $\mathcal{V} \times \mathcal{V}$. In our application, $(i, j) \in \mathcal{E}$ if there exists money flow, either unilateral or bilateral, between committees i and j. A weighted graph also includes, for each edge $(i, j) \in \mathcal{E}$, a corresponding weight w_{ij}, which in our application will correspond to the sum of money flows between the two PCs. Equivalently, a weighted graph $\mathcal{G}$ can be represented by a weighted adjacency matrix $\mathbf{y}$ where

$$y_{ij} = \begin{cases} w_{ij} & \text{if } (i, j) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

In addition to the network structure described above, each committee $i \in \{1, ..., n\}$ is characterized by the following attributes: a unidimensional latent ideology x_i, and a multi-dimensional observable characteristics $\mathbf{z}_i$, which captures the financial and institutional characteristics of the PC. The details of the variables contained in the vector $\mathbf{z}_i$ will be described in Section 6 when we present our empirical specification.

Ideologies of Vertices. We assume that there are m discrete categories of ideologies where $m \geq 2$. In our application, m will be equal to 3, corresponding to Democratic, Republican and Independent respectively. We denote the *true* ideology of vertex $i \in \mathcal{V}$ by $x_i \in \{1, ..., m\}$. We assume that, for all vertices, their x_i's are latent and unobserved to us. However, For a *subset* of vertices $\mathcal{V}^o \subset \mathcal{V}$, there exists a *noisy* measure, denoted by $\hat{x}_i \in \{1, ..., m\}$, of the latent ideology x_i; but for vertices in $\mathcal{V} \backslash \mathcal{V}^o$, we do not have this noisy measure. In our application, the size of $\mathcal{V}^o$ is typically about 1/3 of the size of $\mathcal{V}$.[6]

To summarize, the framework for the financial network of PCs can be represented by:

$$\langle \mathbf{y}, \mathbf{x}, \hat{\mathbf{x}}, \mathbf{z} \rangle = \langle \{y_{ij}\}_{1 \leq i,j \leq n}, \ \{x_i\}_{1 \leq i \leq n}, \ \{\hat{x}_i\}_{i \in \mathcal{V}^o}, \ \{\mathbf{z}_i\}_{1 \leq i \leq n} \rangle . \tag{2}$$

Note, however, since we do not observe $\mathbf{x}$, the data consists of

$$\text{DATA}: \ \langle \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z} \rangle = \langle \{y_{ij}\}_{1 \leq i,j \leq n}, \{\hat{x}_i\}_{i \in \mathcal{V}^o}, \{\mathbf{z}_i\}_{1 \leq i \leq n} \rangle . \tag{3}$$

[5] While both the model and the estimation strategy can be straightforwardly extended to a directed graph, an undirected graph is adopted for computational tractability.

[6] In Appendix H, we also implement a version of the model in which we assume that $\hat{x}_i = x_i$ for all $i \in \mathcal{V}^o$.

The goal of our empirical exercise is to make *inference* about the latent ideology $\mathbf{x}$ based on data $\langle \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z} \rangle$.

2.1 A Statistical Model

As we explained in the introduction, the vector of latent ideologies $\mathbf{x} = (x_1, ..., x_n)$ can be high-dimensional. To circumvent the high-dimensionality problem, we adopt a Bayesian approach by assuming that each vertex can, *a priori*, be of ideology $k \in \{1, ..., m\}$ with probability θ_k where $\boldsymbol{\theta} = (\theta_1, ..., \theta_m) \in \Delta^{m-1}$; and then we use its observed links with other vertices and the weights of the links to render a *posterior* probability distribution $\mathbf{p}_i \in \Delta^{m-1}$ over these categories. We will use the *mode* of the posterior distribution $\mathbf{p}_i$ as our best guess for i's political or ideological affiliation.[7]

Formally, our model of the financial network formation among PCs is based on the *stochastic block* models of random network formation. In the model, the number of vertices n, the number of blocks m, and vertices' observable characteristics $\mathbf{z}$ are assumed to be exogenous and fixed. The adjacency matrix $\mathbf{Y}$, the ideology vector $\mathbf{X}$, and the noisy measure $\hat{\mathbf{X}}$ are assumed to be random.[8]

Prior. For any PC $i \in \mathcal{V}$, its latent ideology X_i has the following marginal prior distribution:

$$\mathbb{P}(X_i = k) = \theta_k \text{ for } k = 1, ..., m \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_m) \in \Delta^{m-1}$.

Measurement Error. For a PC $i \in \mathcal{V}^o \subset \mathcal{V}$, we observed i's reported political affiliation $\hat{X}_i$, which we interpret as a noisy measure of its true ideology X_i. Specifically,

$$\mathbb{P}(\hat{X}_i = k | X_i = x_i) = \begin{cases} 1 - \epsilon & \text{for } k = x_i \\ \frac{\epsilon}{m-1} & \text{for } k \in \{1, ..., m\} \backslash \{x_i\}, \end{cases} \tag{5}$$

where $\epsilon \in [0, 1)$ captures *the rate of measurement error*.[9]

[7] Of course, the whole posterior distribution $\mathbf{p}_i$ itself is of interest beyond its mode: $\mathbf{p}_i$ will allow us to provide a more continuous measure of $i's$ ideology spectrum. For example, two vertices may end up with the same ideological categorization based on their posterior mode, while one vertex could be more central than the other.

[8] Throughout the paper, generic random variables are denoted by capital letters $\mathbf{Y}$, $\mathbf{X}$, and $\hat{\mathbf{X}}$, whereas their realizations and specific configurations are denoted by small letters $\mathbf{y}$, $\mathbf{x}$, and $\hat{\mathbf{x}}$.

[9] ϵ can be pre-specified as 0 if there is good reason to believe that there is no error in report (perhaps by the nature of the data set). Otherwise, a positive ϵ is more flexible because it allows for measurement error, which makes it possible for the report $\hat{x}_i$ to differ from the true ideology x_i. We will implement a version of the model restricting ϵ to be zero in Appendix H, where we instead hold out a subset of the vertices with self-reported ideologies in our estimation, and use the holdout sample to validate our estimation results.

Edge Formation. Conditional on the true ideologies, X_i, X_j, and observable characteristics $\mathbf{z}_i, \mathbf{z}_j$, entries Y_{ij}'s in the weighted adjacency matrix are assumed to be independently generated across (i, j) pairs:

Assumption 1 (Conditional Independence). For any pairs of PCs, (i, j) and (i', j'), with either $i \neq i'$ or $j \neq j'$,

$$Y_{ij} \perp Y_{i'j'} \big| \left(X_i, X_j, \mathbf{z}_i, \mathbf{z}_j, X_{i'}, X_{j'}, \mathbf{z}_{i'}, \mathbf{z}_{j'}\right).$$

An economic interpretation of this assumption is as follows. A PC designs its general principle in contributions according to its political ideology as well as its financial and institutional characteristic. Following the principle, its staffs make decisions on whether to contribute to a specific committee, and the idiosyncratic factors in each of these decisions are unobserved by the researcher and are assumed to be independent. This conditional independence assumption is a potentially strong assumption, because it abstracts away from various possible strategic considerations that a political committee may have in its contribution decisions; of course, the restrictiveness of this assumption in practice depends on how complete the vector of characteristics $\mathbf{z}$ is. It is important to note that Assumption 1 is stated *conditional on the latent ideologies* of the relevant PCs. Thus, it allows for what we believe to be the first-order role of ideologies in political contributions. Other characteristics of the PC, for example, its previous connections, are also allowed to affect its contribution decisions, to the extent that such information is contained in its characteristics $\mathbf{z}$. In addition, it is well acknowledged that it is difficult to establish asymptotic results in a network model in the presence of widespread correlation (see, e.g., Leung (2016)); as a result, valid statistical inference using data from a single network (or a small number of networks) necessitates restrictions on the degree of correlation. Finally, conditional independence is a maintained assumption in the literature of stochastic block model (Snijders and Nowicki (1997) and Nowicki and Snijders (2001)), and the literature of classification in general (see, e.g., Koller and Friedman (2009)).[10]

Specifically, we assume that the edge formation process is as follows. For any pair of vertices $(i, j) \in \mathcal{V}^2$ with $i \neq j$, conditional on $(X_i, X_j) = (k, l) \in \{1, ..., m\}^2$, and $\mathbf{z}_i, \mathbf{z}_j$,

$$\begin{aligned} Y_{ij} &> 0 \text{ if } \beta_{0,kl} + \boldsymbol{\beta_{1,kl}}(\mathbf{z}_i + \mathbf{z}_j) + \boldsymbol{\beta_{2,kl}}\mathbf{z}_i\mathbf{z}_j + e_{ij} > 0; \\ Y_{ij} &= 0, \text{ otherwise}, \end{aligned} \tag{6}$$

where $\left(\beta_{kl,0}, \ \boldsymbol{\beta_{1,kl}}, \boldsymbol{\beta_{2,kl}}\right)$ are the parameters governing the edge formation probability, and $\mathbf{z}_i\mathbf{z}_j$ is a shorthand (with some abuse of notation) for interaction terms of $\mathbf{z}_i$ and $\mathbf{z}_j$. Because we deal

[10] For example, the naive Bayes model is widely used as a spam filter to classify emails into normal vs. spam emails, and it assumes that conditional on text class, the presence of words are independent.

with an undirected graph, we further assume that the process (6) satisfies parameter symmetry over (k, l) , i.e.,

$$\begin{aligned} \beta_{0,kl} &= \beta_{0,lk}, \\ \boldsymbol{\beta_{1,kl}} &= \boldsymbol{\beta_{1,lk}}, \\ \boldsymbol{\beta_{2,kl}} &= \boldsymbol{\beta_{2,lk}}, \end{aligned} \tag{7}$$

and that $e_{ij} \sim$ i.i.d. $\mathcal{N}(0, 1)$.

To simplify notation in the likelihood function in Section 4, we will denote the conditional probability of two vertices i and j forming an edge as η_{ij} and express it in a compact form as:[11]

$$\begin{aligned} \eta_{ij}(X_i, X_j) &= \mathbb{P}(Y_{ij} > 0 | X_i, X_j, \mathbf{z}_i, \mathbf{z}_j) \\ &= \Phi\left(\boldsymbol{\gamma}(X_i, X_j, \mathbf{z}_i, \mathbf{z}_j)' \boldsymbol{\beta}\right), \end{aligned} \tag{8}$$

where

$$\begin{aligned} \boldsymbol{\gamma}(X_i, X_j, \mathbf{z}_i, \mathbf{z}_j) &= \mathbf{D}(X_i, X_j) \bigotimes [1,\ \mathbf{z}_i{+}\mathbf{z}_j,\ \mathbf{z}_i\mathbf{z}_j]', \\ \mathbf{D}(X_i, X_j) &= \left[\mathbb{1}_{(X_i=k, X_j=l) \vee (X_i=l, X_j=k)}\right]_{1 \le k \le l \le m}, \\ \boldsymbol{\beta} &= \left[\beta_{0,kl},\ \boldsymbol{\beta_{1,kl}},\ \boldsymbol{\beta_{2,kl}}\right]'_{1 \le k \le l \le m}. \end{aligned}$$

Weights of the Edge. Once two vertices form an edge, we assume that the weight of their edge, which in our application will correspond to the total amount of financial transactions between the two vertices, is drawn from a Q-valued discrete distribution with probabilities that depend on ideological proximity. Specifically, conditional on $Y_{ij} > 0$, if $X_i = k, X_j = l$, the Y_{ij} takes values $(w_1, ..., w_Q)$ with probabilities

$$\mathbf{h}_{kl} \equiv (h_{kl,1}, ..., h_{kl,Q}) \in \Delta^{Q-1}. \tag{9}$$

In our estimation we will pre-fix Q and the values $(w_1, ..., w_Q)$. Again because we deal with an undirected graph, we impose the natural symmetry assumption that

$$\mathbf{h}_{kl} = \mathbf{h}_{lk}. \tag{10}$$

[11] Note that we omit the covariates $\mathbf{z}_i$ and $\mathbf{z}_j$ in the arguments of expression (8) for η_{ij} for notational convenience.

2.2 Discussions

First, we would like to note that the edge formation process as specified by (6) can be interpreted as resulting from a model of matching. Two committees decide whether to establish a financial connection based on their joint surplus from forming a match. The surplus has a deterministic component $\beta_{0,kl}+\boldsymbol{\beta_{1,kl}}(\mathbf{z}_i+\mathbf{z}_j)+\boldsymbol{\beta_{2,kl}}\mathbf{z}_i\mathbf{z}_j$ as well as a stochastic component e_{ij}. Parameters differ by ideology pair (k,l). $\beta_{kl,0}$ captures the direct effect of ideology proximity, and $\boldsymbol{\beta_{1,kl}}$ and $\boldsymbol{\beta_{2,kl}}$ capture the effect of committee specific characteristics interacting with ideology proximity. In other words, ideology influences the edge formation probability through both the constant term and the coefficients.[12]

Second, financial and institutional characteristics $\mathbf{z}_i$ and $\mathbf{z}_j$ are included in the specification (6) not only because they may be important factors in the PCs' contribution decisions, but also, technically, it is our strategy for *degree correction.* Without these covariates, the model is essentially the same as Snijders and Nowicki (1997), where edge formation probability is governed only by ideology proximity. In their model, variations in degree are attributed only to random shock and ideology proximity. Moreover, the implied degree distribution is a mixture of $m(m+1)/2$ binomial distributions. However, when the number of ideology categories is small, the model-implied degree distribution may not be able to capture the empirical degree distribution. Therefore, we include vertex specific characteristics to introduce richer heterogeneity in the edge formation probability. This is a variant of the degree correction strategy introduced in Peng and Carvalho (2015*a*) and Peng and Carvalho (2015*b*). In addition to proximity of the latent ideology, Peng and Carvalho (2015*a*) assume that the edge formation probability also depends on additional latent variables $\xi_i, \xi_j \in \mathbb{R}$, capturing vertex specific popularity. Peng and Carvalho (2015*b*) replace the latent popularity variables with the quantile ranks of the vertices' observed degrees $q_i, q_j \in \{1, 2, ..., Q\}$. Our paper differs from Peng and Carvalho (2015*b*) in that the characteristics we use do not involve any features of the network, rendering a much cleaner model with which to perform statistical inference because the potential problem of endogeneity is avoided. Compared to Peng and Carvalho (2015*a*), our paper has a faster convergence rate and lower computational intensity because it does not require inferring about additional latent variable ξ_i.

Finally, note that conditional on the latent ideologies of the vertices, our model of network formation is equivalent to a pair-wise matching model. However, since the vertices' latent ideologies are not known, our edge formation process as specified in (6) allows for correlation *conditional on*

[12] In an empirical application, additional constraints can be added: for example, restricting some coefficients in $\boldsymbol{\beta_{1,kl}}$ and $\boldsymbol{\beta_{2,kl}}$ to be 0, or to be the same for different ideology pairs. We provide the details of the additional restrictions in Section 6.

observables $\left\{\hat{X}_i\right\}_{i\in\mathcal{V}^o}$ and $\mathbf{z}$. Moreover, the correlation is spread all over the connected component of the network via the latent ideologies. As a result, our network is not decomposable based on observables.

3 Data and Descriptive Statistics

Three data sets are used in our study, two of which are administrative data sets from the Federal Election Commission (FEC) in 2003-2004 election cycle: the “ Committee Master File”, and the “Contributions to Committees from Committees” data; and the third data set is collected from `OpenSecrets.org`.[13]

The “Committee Master File” contains basic information about all the PCs registered with the FEC. The PCs in this data set include federal political action committees (PACs), party committees, campaign committees for presidential, house and senate candidates, as well as groups or organizations such as lobbyists or fund raisers. This data set also contains information on the PC’s geographical location, institutional characteristics, and the self-reported party affiliation, if available.

The “Contributions to Committees from Committees” data contains the *universe* of the contribution records between PCs. We observe the universe of records because the campaign finance laws mandate the disclosure of all transactions and expenditure related to federal election activities. In this data set, for each contribution, it lists the contributor, recipient, and the amount of contribution. We will use this data set to construct the *complete* network of financial transactions among the PCs.

Finally, the `OpenSecrets.org` data set contains the industry categorization for the PCs. The data sources corresponding to each of the variables we use are summarized in Table 1.

Variable	Data Source
$\mathbf{y}$	Contributions to Committees from Committees
$\hat{\mathbf{x}}$	Committee Master File
$\mathbf{z}$	Committee Master File (location and institutional characteristics), Contributions to Committees from Committees (imputed financial budget), `OpenSecrets.org` (industry categorization).

Table 1: Data Source

[13] The website is maintained by the Center for Responsive Politics.

3.1 Vertex Characteristics

In the 2003-2004 election cycle, 5,858 PCs participated in contribution activities with other PCs and formed a financial network, among a total of 7,559 active PCs (defined as PCs with either positive total receipts or positive total disbursement).[14] We do not include in our study the PCs that do not participate in financial contributions with other PCs, though they could alternatively be viewed as isolated vertices in the network. This is not particularly worrisome because the PCs excluded from our study are financially insignificant. Among all PCs with positive receipts, the PCs in the financial network accounted for 96.83% of total amount of receipts; and among all PCs with positive disbursement, these PCs accounted for 96.87% of total amount of receipts.

Since the campaign finance laws do not mandate that political committees report their party affiliations, the Committee Master File has *incomplete* information on the PCs' political affiliation. In Table 2, we summarize the distribution of *reported* party affiliations of the PCs in the Committee Master File. Respectively 17.4% and 17.12% of the PCs reported to be Democratic and Republican, while 1.01% of the PCs reported to be Independent, and 0.85% of the PCs reported other affiliations such as Labor Party or Conservative Party. Importantly, 63.62% of the PCs did not report their party affiliations. In terms of financial significance, contributions sent by the PCs without self reported political affiliations accounted for 15.25% of the total amount of contributions among the PCs, and contributions received by them accounted for 37.38% of the total contributions.

Reported Affiliation	Frequency	% of PC	% of Contribution From	% of Contribution To
Democratic	1,019	17.40	45.25	34.74
Republican	1,003	17.12	38.64	25.78
Independent	59	1.01	0.53	1.63
Other	50	0.85	0.33	0.46
Missing	3,727	63.62	15.25	37.38
Total	5,858	100	100	100

Table 2: Tabulation of Reported Party Affiliation

In Table 3 we describe the non-ideological characteristics included in our analysis. Information on state and industry enables us to capture the effect of geographical and industrial proximity on political contribution. Figure 1 shows that the District of Columbia has the highest number of political committees (896), followed by California (575), Virginia (436), Texas (301), New York

[14] Information on total receipts and disbursement is obtained from data sets "Candidate Summary (All Candidates)" and "PAC & Party Summary" released by the FEC. This information is missing for 89 PCs in the observed financial network. We define them as active, and use their budget in the financial network as a proxy for their total receipts and total disbursement.

Characteristics	Type	Range
State	Categorical	55 distinct categories
Industry	Categorical	46 distinct categories
Committee Type	Categorical	6 distinct categories
Committee Designation	Categorical	3 distinct categories
National	Dummy	{0,1}
Budget (in $1,000)	Continuous	[0, 104064.2]

Table 3: Non-ideological Characteristics
Note: States include unincorporated territories.

(284), Pennsylvania (273) and Illinois (226). In Table 4, we provide a coarse industrial breakdowns of the political committees. Besides "Other" , which aggregates many small industrial categories, Finance/Insurance/Real Estate is the most represented industry (9.12%), followed by Miscellaneous Business (6.96%), Health (5.43%) and Labor (5.19%). In our empirical analysis, we drop the ideological and party affiliated sectors, and use a finer breakdown for sectors such as Misc Business and Other. We provide a detailed description of the construction of the industry variable in Appendix A.

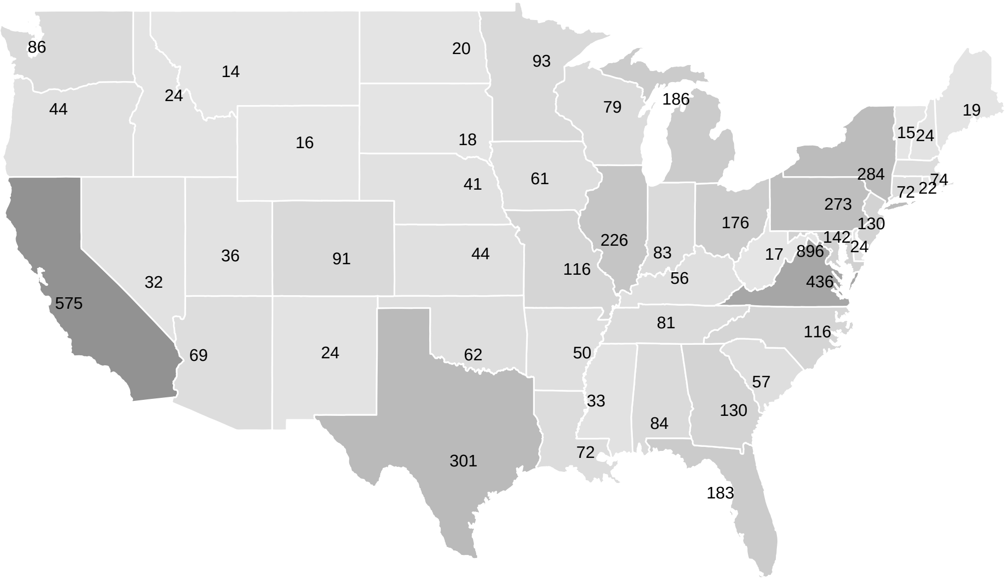


Figure 1: Geographical Distribution of Political Committees

Committee type, committee designation, and national dummy are institutional characteristics that are potentially related to contribution patterns. There are six distinct categories of committee type in the Committee Master File: House campaign, Senate campaign, Presidential campaign,

	All PCs		Report Dem		Report Rep		Report Ind		Missing/Report Other	
Sector	Number	%	Number	%	Number	%	Number	%	Number	%
Agribusiness	285	4.87	0	0	0	0	4	6.78	281	7.44
Communic/Electronics	186	3.18	0	0	0	0	3	5.08	183	4.85
Construction	124	2.12	0	0	0	0	1	1.69	123	3.26
Defense	54	0.92	0	0	0	0	1	1.69	53	1.40
Energy/Natural Resource	274	4.68	1	0.10	0	0	3	5.08	270	7.15
Finance/Insurance/ Real Estate	534	9.12	0	0	0	0	6	10.17	528	13.98
Health	318	5.43	0	0	0	0	4	6.78	314	8.31
Labor	304	5.19	1	0.10	0	0	9	15.25	294	7.78
Lawyers & Lobbyists	186	3.18	0	0	0	0	0	0	186	4.92
Transportation	157	2.68	0	0	0	0	2	3.40	155	4.10
Misc Business	408	6.96	0	0	0	0	6	10.17	402	10.64
Ideological/Single-Issue	2,404	41.04	782	76.74	812	80.96	19	32.20	791	20.94
Joint Candidate Committee	164	2.80	81	7.95	65	6.48	0	0	18	0.48
Party Committee	359	6.13	153	15.01	124	12.36	0	0	82	2.17
Other	24	0.41	0	0	0	0	1	1.69	23	0.61
Unknown	77	1.31	1	0.01	2	0.20	0	0	74	1.96
Total	5,858	100	1,019	100	1,003	100	59	100	3,777	100

Table 4: Industrial Sectors of Political Committees

qualified PAC, qualified Party and others.[15] Table 5 shows that 27.72% of the committees are campaign committees of either House, Senate or Presidential elections, 45.43% are qualified PACs, and 4.52% are qualified Party committees, and the rest ("Other") are mainly non-qualified committees. In Table 6, we provide information about the committee designations. There are three distinct categories of committee designation in the Committee Master File: authorized by a candidate, joint fund-raiser, and others.[16] Table 6 shows that 1.89% of the committees are authorized by a candidate, 2.83% are joint fund-raisers, and the majority of the rest are either principal campaigns or committees not authorized by a candidate. We do not list principal campaigns separately because it would be redundant given the more detailed categorization in committee type. Finally, the national dummy listed in Table 3 takes value 1 if and only if the committee is one of the following six committees: the Democratic National Committee (DNC), the Democratic Senatorial Campaign Committee (DSCC), the Democratic Congressional Campaign Committee (DCCC), the Republican National Committee (RNC), National Republican Senatorial Committee (NRSC), and National Republican Congressional Committee (NRCC).

We compute a PC's budget from its contribution record. It is constructed as the *sum of its contributions to* all other PCs within this election cycle. Typically, this amount is lower than a committee's total receipts or total disbursements. We use this measure because it is most powerful in explaining a PC's probability of making political contribution, and thus its total number of connections. In fact, the correlation between a PC's total receipts or disbursements and its number of contributions to other PCs is low. The distribution of budget is given in Table 7 and Figure 2. It has a wide range and a fat tail: 25.7% of the PCs have budget less than or equal to $1,000, while 8% of the PCs have budget more than $500,000.

3.2 Descriptive Statistics of the Financial Transactions Network

The Contributions from Committees to Committees data set records 411,106 transactions among 5,858 political committees in the 2003-2004 election cycle. Figure 3 is a graphical representation of the network using graphing software Gephi.[17] Each vertex represents a PC, and the color of the vertex represents reported affiliation: *blue* for Democratic, *red* for Republican, *green* for

[15] Qualified PACs need to have been in existence for six months and received contributions from 50 people and made contributions to five federal candidates. Qualified party committees need to have been in existence for at least six months and received contributions from 50 people or are affiliated with another party committee that meets these requirements.

[16] A committee is designated as "*authorized by a candidate*" if it is authorized by a candidate in writing to receive contributions or make expenditures on behalf of the candidate, but is not her principal campaign committee. A committee is designated as "*a joint fundraiser*" if it is created by two or more candidates, PACs or party committees to share the costs of fundraising, and split the proceeds.

[17] See https://gephi.org.

	All PCs		Report Dem		Report Rep		Report Ind		Missing/Report Other	
Committee Type	Number	%	Number	%	Number	%	Number	%	Number	%
House	1,275	21.77	613	60.16	644	64.21	8	13.56	10	0.26
Senate	309	5.27	140	13.74	158	15.75	2	3.39	9	0.24
Presidential	40	0.68	20	1.96	9	0.90	6	10.17	5	0.13
Qualified PAC	2,661	45.43	20	1.96	13	1.30	43	72.88	2,585	68.44
Qualified Party	265	4.52	127	12.46	114	11.37	0	0	24	0.64
Other	1,308	22.33	99	9.72	65	6.48	0	0	1,144	30.29
Total	5,858	100	1,019	100	1,003	100	59	100	3,777	100

Table 5: Types of the Political Committees

Note: The 22.33% other committees mainly consist of non-qualified PACs and non-qualified Party committees.

	All PCs		Reported Dem		Reported Rep		Reported Ind		Missing/Reported Other	
Committee Type	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Authorized by a Candidate	111	1.89	42	4.12	65	6.48	4	6.78	0	0
Joint Fundraiser	166	2.83	81	7.95	66	6.58	0	0	19	0.50
Other	5,581	95.27	896	87.93	872	86.94	55	93.22	3,758	99.50
Total	5,858	100	1,019	100	1,003	100	59	100	3,777	100

Table 6: Committee Designation

Note: The 95.27% other committees mainly consist of principal campaigns and committees that are not authorized by a candidate.

Quantile	All PCs	Report Dem	Report Rep	Report Ind	Missing/Report Other
Min	0.00	0.00	0.00	0.00	0.00
25%	1.00	2.98	4.00	5.00	0.70
50%	10.00	32.75	48.07	20.00	6.08
75%	75.90	353.30	456.15	88.87	30.00
Max	104,064.19	104,064.19	55,873.70	2,386.53	9,180.75
Obs.	5,858	1,019	1,003	59	3,777

Table 7: Quantiles of Budget in $1,000

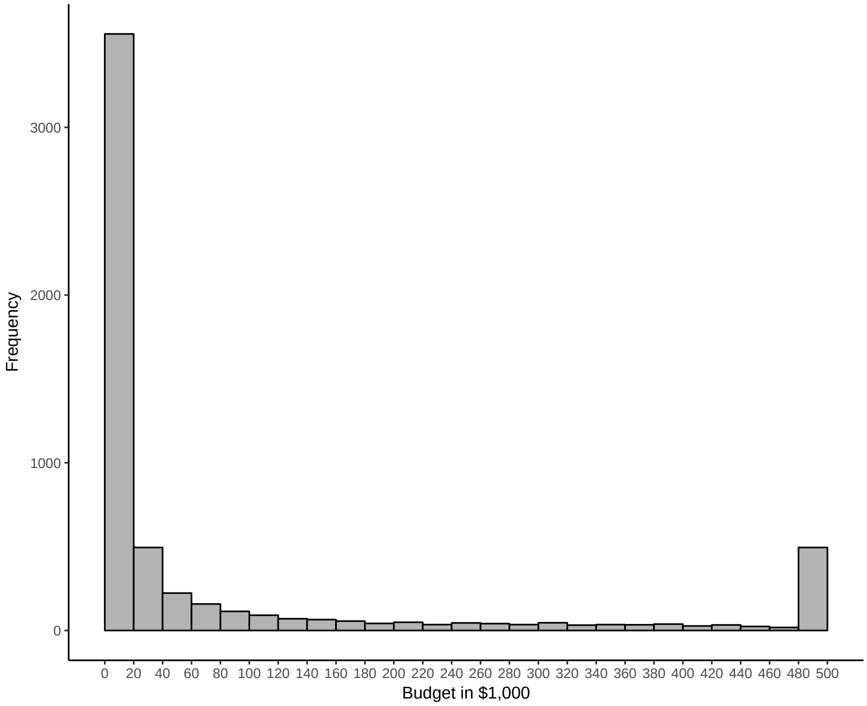


Figure 2: Distribution of Budget

Note: Observations with budget higher than $500,000 are plotted at $500,000.

Independent, and *yellow* for missing/other. For the purpose of this graph, we collapse multiple contributions with the same direction between any pair of PCs into one edge: each edge represents a *directed* financial connection; and we use the contributor's color to represent the color of the edge. Thus there are a total of 164,529 edges in Figure 3. This network has 20 disconnected components.[18] Nevertheless, the component in the center is disproportionately large with 5,806 vertices. The subsequent description focuses on this giant component.

Network Level Statistics. In the rest of the analysis we will consider the network as undirected, thus we further collapse transactions with different directions between any pair of PCs. We derive

[18] A component is a maximum connected sub-network.

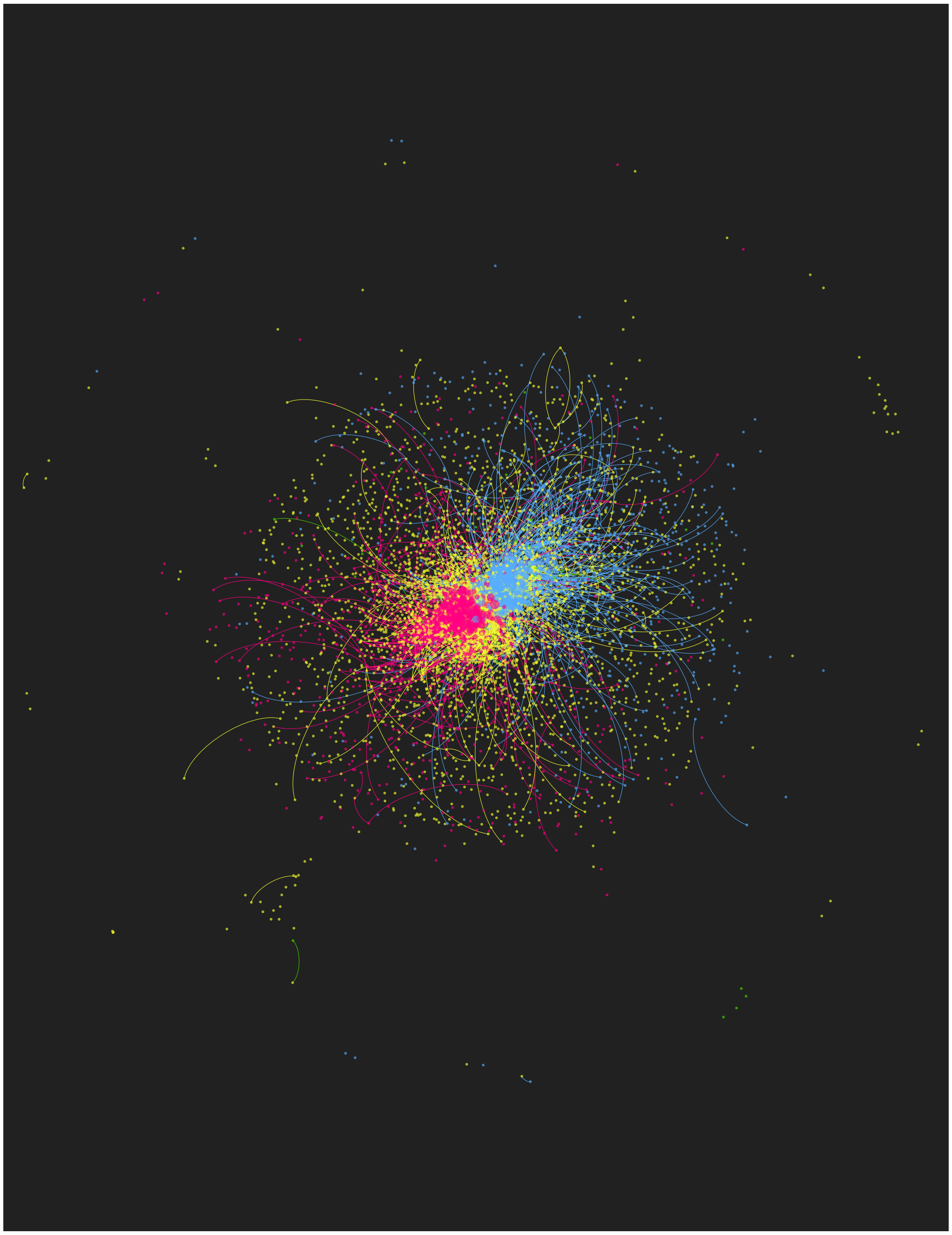


Figure 3: Political Contribution Network

Note: Vertex color represents reported affiliation: blue for Democratic, red for Republican, green for Independent, and yellow for missing. Edge color is the same as contributor's color.

a financial network with 5,806 vertices and 145,406 edges. Table 8 presents some of the key statistics for the financial transaction network among the PCs. First, the network is sparse, yet well-connected. On the one hand, the average degree is 50.09, i.e. on average, a PC is connected to only 0.86% of other PCs.[19] This indicates sparsity because the number of edges is only a small fraction of all the vertex pairs. On the other hand, it has a diameter of 10, and an average distance of 2.99.[20] Both statistics indicate that the network is well-connected, which poses a challenge for our study. Given the connectedness, there is no straightforward approach to structurally decompose this giant component into separate sub-networks, despite the visual patterns of clustering. Moreover, there is no other natural way of decomposition. A common practice in the network applications is to partition the full network into geographically disjoint sub-networks and assume no interaction across sub-networks. However, this practice is not suitable in our application because political contributions are not concentrated at local levels.

Number of vertices	5,806
Number of edges	145,406
Average degree	50.09
Diameter	10
Average distance	2.99

Table 8: Network Statistics

Distributions of Degrees and Edge Weights. To explore beyond the network-level summary statistics, we further investigate the distribution of degrees. Figure 4 shows that the degree distribution has a large spread and a fat tail: the highest degree is as large as 978. An important feature is that the shape of this distribution is inconsistent with a mixture of a small number of binomial distributions–an implication in the standard stochastic block model. To capture the observed degree distribution, we introduce rich heterogeneity in our model. Specifically, we include budget whose distribution has similar patterns to account for the variations in degree.

Additionally, we study the sub-networks induced by reported Democratic, Republican and Independent PCs, and analyze connection patterns conditional on reported affiliations. In Figure 5, we have three graphs for self-reported Democratic, Republican, and Independent PCs respectively. In each graph, we present the distribution of the PCs' numbers of connections with the three groups

[19] *Degree* is a vertex-level statistics. It is the number of direct connections a vertex has. Following the notation introduced in the previous section, vertex i's degree is given by $d_i = \sum_{j \neq i} \mathbb{1}(y_{ij} > 0)$.

[20] The *diameter* of a network is calculated as the maximum length of the (finite) shortest paths among vertex pairs. The *average distance* of a network is calculated as the average length of the (finite) shortest paths among vertex pairs.

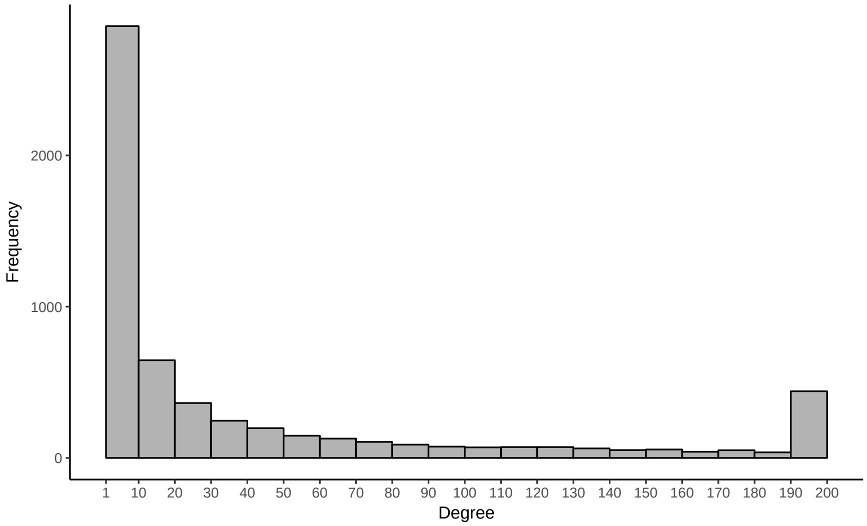
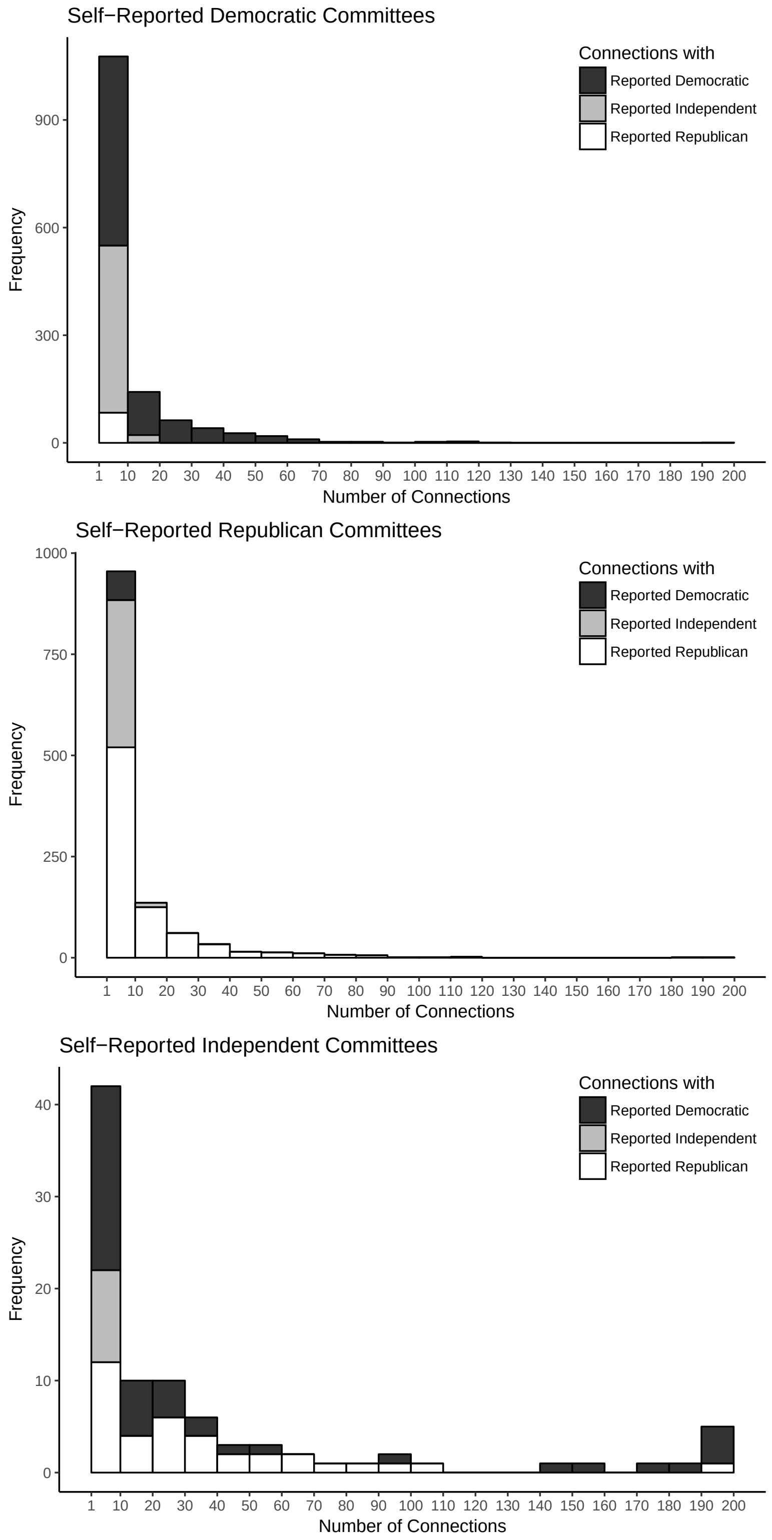


Figure 4: Histogram of degree

Note: Observations with degree higher than 200 are plotted at 200.

of self identified PCs. Self-reported Democratic and Republican PCs show evidence of homophily. On the one hand, many PCs are financially connected with PCs affiliated to same party, and a sizable of them have more than 20 such connections; on the other hand, only a small number of PCs are financially connected with PCs affiliated to the other party, and most of them have less than 10 such connections. Self-reported Independent PCs show similar connection pattern with the self-reported Democratic vs. Republican PCs. Many are financially connected with both parties, and a sizable of them have more than 20 such connections. A small fraction are financially connected with other self-reported Independent PCs, and most of them have less than 10 such connections - partially due to the small number of self-reported Independent PCs. Note that we cannot claim, based on the connection patterns between PCs with self reported affiliations, that this is an evidence of heterophily. It is possible that they are connected with a large number of Independent PCs without self-reported affiliations. Furthermore, even if Independent PCs exhibit heterophily, this behavior does not invalidate either our model or estimation. Identification only requires that PCs with different political ideologies have different contribution patterns. This is at least true for the PCs with self-reported affiliation, which is reassuring.

Finally, in order to infer about the unknown ideologies from the PCs with self reported ideologies, an implicit assumption in our model is that the PCs without self reported affiliations do not act systematically differently from the PCs with self reported affiliations. Therefore, we reproduce the degree distribution in Figure 6 with the composition of each bar marked by color. The degree distribution of PCs without self reported affiliations is similar to that of the PCs with self reported affiliations. This evidence is consistent with the assumption, though insufficient. In Section 6, we



Self−Reported Democratic Committees

Self−Reported Republican Committees

Self−Reported Independent Committees

Figure 5: Distributions of Number of Connections with Different Committees

Note: These figures only include observations with positive number of connections. Observations with more than 200 connections are plotted at 200.

will present more evidence on this after we have presented our estimation results.

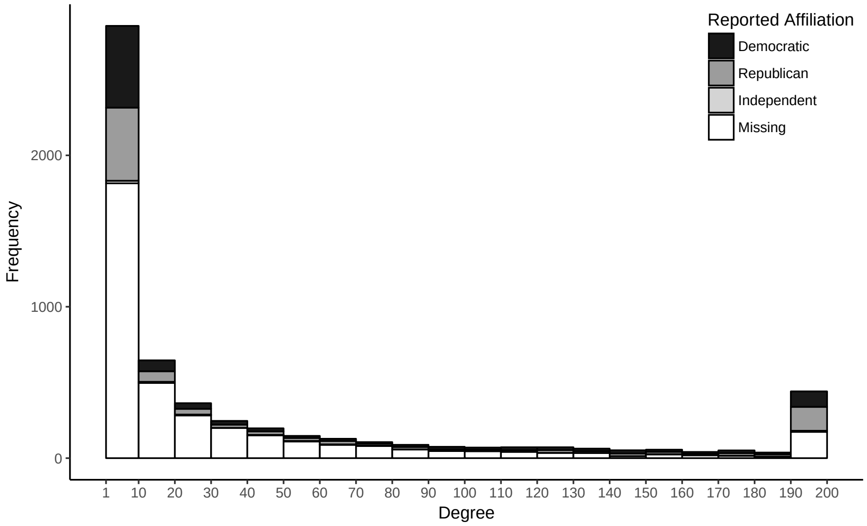


Figure 6: Decomposition of Degree Distribution

Note: The same as Figure 4, but the composition of each bar, i.e., the composition of PCs with a certain rage of degrees, is presented with different colors.

The financial network also provides information on transfer amount (the sum of the contribution amount) between pairs of connected PCs. Its distribution is given in Table 9 and Figure 7. Most of them are small because of the regulations on contribution limits (detailed description in Appendix B). In addition, we present the distribution of transfer amount conditional on reported ideology in Table 9 and Figure 8. On average, the transfer amount is high between PCs with the same reported affiliation, and low for the other cases. A caveat in interpreting these statistics is that we exclude all the contributions involving PCs without self reported affiliations, so they do not necessarily give the full picture of the contribution pattern.

3.3 Issues with Naive Alternative Methods

In this subsection, we briefly discuss the issues with some naive alternative methods that we have attempted, and explain why they are invalid. We first define

$$\hat{x}_i = \begin{cases} -1 \text{ if committee } i \text{ is reported to be Democratic} \\ \;\;\,0 \text{ if committee } i \text{ is reported to be Independent} \\ \;\;\,1 \text{ if committee } i \text{ is reported to be Republican.} \end{cases}$$

Quantile	All	Dem,Dem	Rep,Rep	Ind, Ind	Other/Missing,Other/Missing	
Min	1.00	6.00	10.00	5,000.00	11.00	
25%	1,000.00	1,000.00	1,000.00	10,000.00	2,000.00	
50%	2,000.00	1,000.00	1,000.00	10,000.00	5,000.00	
75%	5,000.00	2,000.00	4,028.00	29,143.00	10,000.00	
Max	47,190,000.00	47,190,000.00	22,542,215.00	223,000.00	12,367,982.00	
Mean	8,990.74	76,034.83	58,189.32	48,920.67	9,544.85	
Obs.	145,406	5,582	5,151	6	13,979	
Quantile	Dem,Rep	Dem,Ind	Dem,Other/Missing	Rep,Ind	Rep,Other/Missing	Ind,Other/Missing
Min	10.00	51.00	1.00	50.00	2.00	5.00
25%	1,000.00	1,500.00	1,000.00	1,000.00	1,000.00	2,000.00
50%	1,000.00	4,000.00	2,000.00	2,000.00	2,000.00	5,000.00
75%	2,000.00	7,500.00	5,000.00	5,000.00	4,000.00	10,000.00
Max	150,000.000	75,000.00	2,966,933.00	60,000.00	1,638,000.00	1,000,000.00
Mean	9,059.11	5,209.31	3,833.99	3,677.08	3,468.98	18,441.87
Obs.	121	2,238	48,132	1,359	68,281	557

Table 9: Quantiles of Transfer Amount (in $1.00)

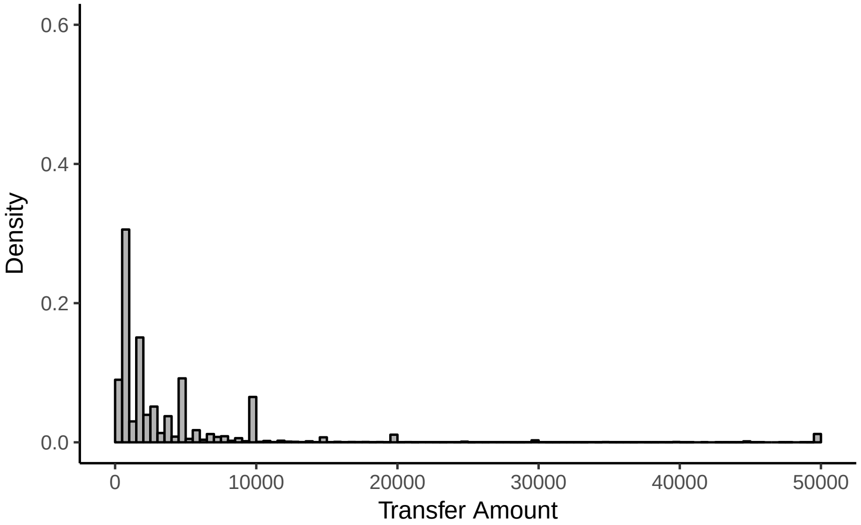


Figure 7: Empirical Distribution of Transfer Amount

Note: Bin size is 500. Observations with transfer amount higher than $50,000 are plotted at $50,000.

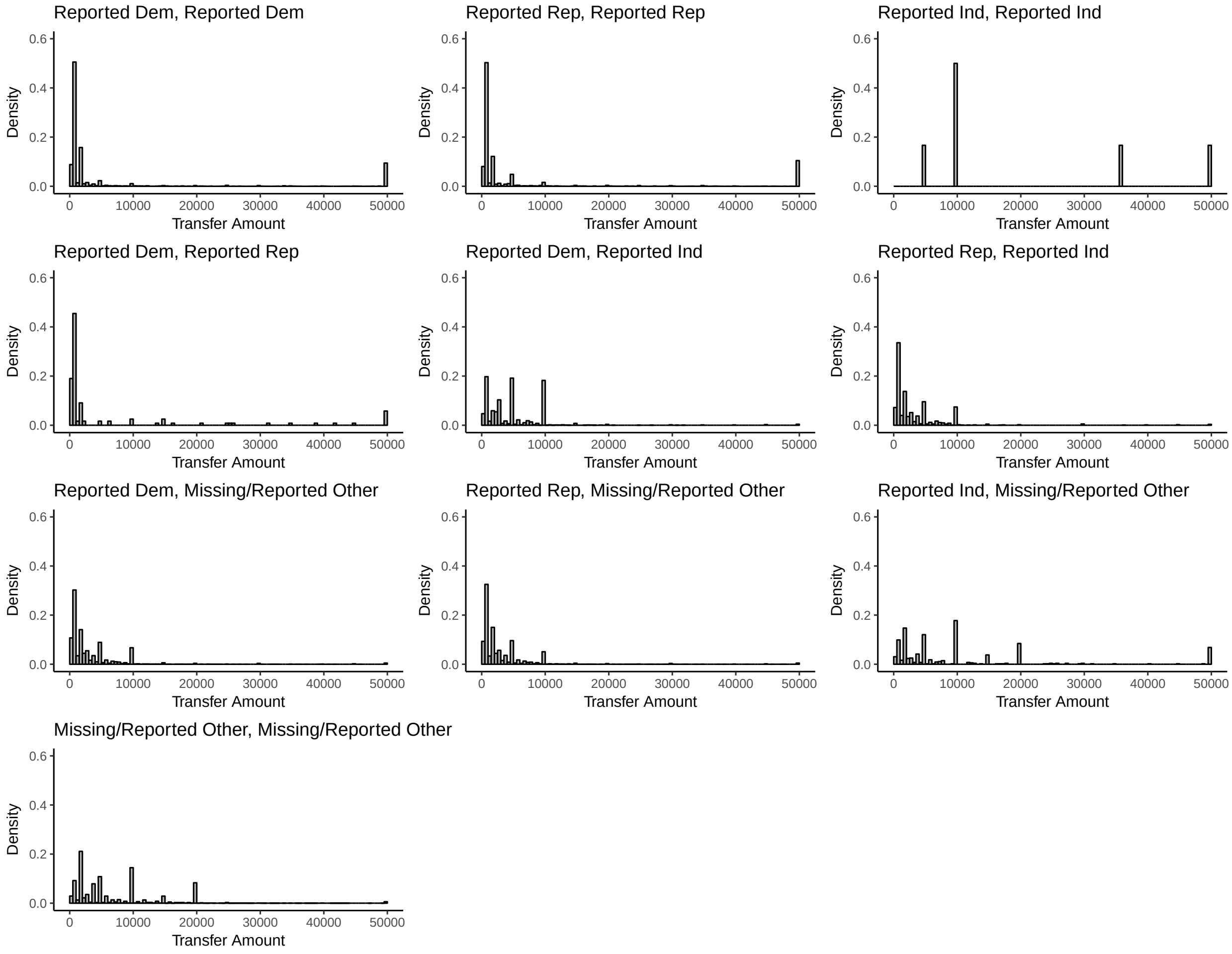


Figure 8: Empirical Distribution of Transfer Amount Conditional on Reported Ideology

Note: Bin size is 500. Observations with transfer amount higher than $50,000 are plotted at $50,000.

The naive method tries to find $\mathbf{x}^*$ solving the following fixed point problem as a solution to the ideology recovery problem:

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \mathbf{x}^* \end{bmatrix} = \operatorname{sign}\left(\mathbf{y}\begin{bmatrix} \hat{\mathbf{x}} \\ \mathbf{x}^* \end{bmatrix}\right). \tag{11}$$

In other words, a PC is Democratic (Republican) if it is connected with more Democratic (Republican) PCs than Republican (Democratic) ones, or Independent if it is connected with an equal number of Democratic and Republican PCs. However, neither existence nor uniqueness of the solution is guaranteed. We attempted to solve this problem by iteration method, but failed. The following example in Figure 9 illustrates the reason. According to the categorization rule described above, the two "unknown" vertices should be assigned Democratic. However, this assignment generates inconsistency in the "Republican" vertex's behavior: as a Republican PC, it is connected with two Democratic PCs. There does not exist a categorization which can reconcile such inconsistency, so the naive method proposed in (11) does not guarantee a well-defined solution.

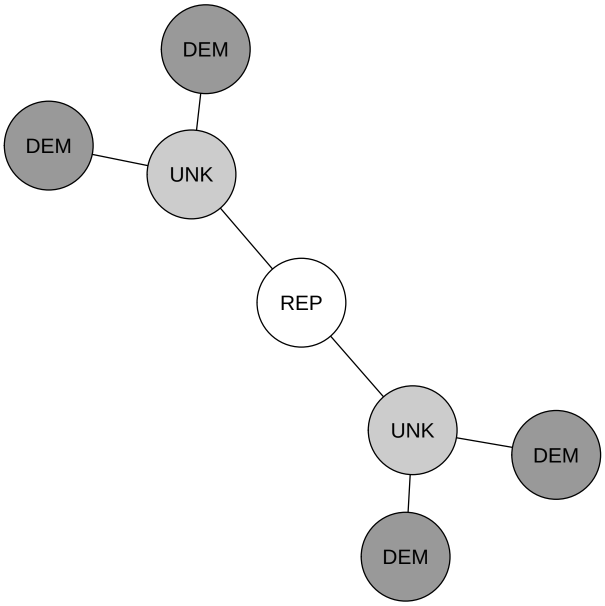


Figure 9: Non-Existence of Solution with Naive Method

Since the solution concept above requires too strong a coherency in categorization, a less restrictive method was also attempted:

$$x_i = \sum_{j \in \mathcal{V}^o} y_{ij} \hat{x}_j. \tag{12}$$

In this case, a PC's ideology is defined by its connected PCs with self reported affiliation. There are two major problems with this method. First, when a PC is only connected to PCs with unreported affiliations, its ideology is not defined. If we try to address this problem by iteratively applying the equation above, we go back to the previous method. Second, categorization has very low precision for PCs connected with few PCs with self reported affiliations and mostly with PCs with unknown affiliations. The poor performance of naive methods necessitate the use of a more sophisticated method.

4 Estimation

Given the model description, the likelihood of $\mathbf{y}, \hat{\mathbf{x}}$ conditional on $\mathbf{x}, \mathbf{z}$ is given by

$$\begin{aligned} & \mathbb{P}\left(\mathbf{y}, \hat{\mathbf{x}} | \mathbf{x}, \mathbf{z}; \epsilon, \boldsymbol{\beta}, \mathbf{h}\right) \\ = \; & \mathbb{P}\left(\hat{\mathbf{x}} | \mathbf{x}; \epsilon\right) \mathbb{P}\left(\mathbf{y} | \mathbf{x}, \mathbf{z}; \boldsymbol{\beta}, \mathbf{h}\right) \\ = \; & (1-\epsilon)^{n_t} \left(\frac{\epsilon}{m-1}\right)^{n_e} \prod_{1 \le i < j \le n} \left[\eta_{ij}(x_i, x_j) h_{x_i x_j}(y_{ij})\right]^{\mathbb{1}(y_{ij}>0)} \left[1 - \eta_{ij}(x_i, x_j)\right]^{\mathbb{1}(y_{ij}=0)}, \end{aligned} \tag{13}$$

where $n_t = \sum_{i \in \mathcal{V}^o} \mathbb{1}(x_i = \hat{x}_i)$ is the number of vertices in $\mathcal{V}^o$ whose $\hat{x}_i$'s coincide with x_i's, $n_e = \sum_{i \in \mathcal{V}^o} \mathbb{1}(x_i \neq \hat{x}_i)$ is the number of vertices in $\mathcal{V}^o$ whose $\hat{x}_i$'s differ from x_i's, and $h_{x_i x_j}(y_{ij}) = h_{x_i x_j, q}$ if $y_{ij} = w_q$ where $h_{x_i x_j, q}$ is defined in (9).

We use the Maximum A Posteriori (MAP) estimator to recover the latent ideology. It is a Bayesian estimator that equals the mode of the posterior probability. Specifically, it solves

$$\max_{\mathbf{x} \in \{1,\ldots,m\}^n} \mathbb{P}\left(\mathbf{y}, \hat{\mathbf{x}} | \mathbf{x}, \mathbf{z}; \epsilon, \boldsymbol{\beta}, \mathbf{h}\right) \mathbb{P}\left(\mathbf{x}; \boldsymbol{\theta}\right). \tag{14}$$

Note that the Maximum Likelihood Estimator (MLE) solves

$$\max_{\mathbf{x} \in \{1,\ldots,m\}^n} \mathbb{P}\left(\mathbf{y}, \hat{\mathbf{x}} | \mathbf{x}, \mathbf{z}; \epsilon, \boldsymbol{\beta}, \mathbf{h}\right), \tag{15}$$

and that MAP is equivalent to MLE under uniform prior $\boldsymbol{\theta} = (\frac{1}{m}, ..., \frac{1}{m})$.

We now argue for the validity of the MAP estimator, and propose a Bayesian algorithm to obtain an approximate solution. Theoretically, statistical inference is non-standard in our model. First of all, we have only one observation of the network, i.e. one realization of $\mathbf{y}, \hat{\mathbf{x}}$. Second, the number of parameters (the number of latent political ideologies) grows with the network size. Therefore, canonical asymptotic theory is not applicable. For example, law of large numbers cannot be directly applied. As a result, in the recent literature, new concepts and tools are introduced to study this problem. The following subsection summarizes the key concepts and results mostly related to our study.

4.1 Threshold for Exact Recovery

In this subsection we describe the theoretical results in Yun and Proutiere (2016) that justify our estimation using one observation of the network. The standard stochastic block model was first introduced in Snijders and Nowicki (1997) and Nowicki and Snijders (2001), and its exact recovery problems were studied in Mossel, Neeman and Sly (2014), Abbe, Bandeira and Hall (2014), and Abbe and Sandon (2015). The labeled stochastic block model introduced weights on edges, and its exact recovery problems were studied in Jog and Loh (2015) and Yun and Proutiere (2016). First, we provide a formal definition of the labeled stochastic block model.

Definition 1 (Labeled Stochastic Block Model LSBM $(n, \boldsymbol{\theta}, \frac{\log(n)}{n}\boldsymbol{W})$). The LSBM generates an n-vertex random graph with community affiliation $\boldsymbol{X}$ and weighted adjacency matrix $\boldsymbol{Y}$ according to the following process. Each vertex is assigned a community affiliation $X_i \in \{1, 2, ..., m\}$ independently under probability $\boldsymbol{\theta} = (\theta_1, ..., \theta_m) \in \Delta^{m-1}$. Conditional on community affiliations, the edges are drawn independently. The edges Y_{ij}'s take discrete values $\{0, w_1, w_2, ...w_Q\}$ where 0 represents no edge and w_q represents an edge with the specified (non-zero) weight. The distribution of edges is governed by an m-by-m-by-Q matrix $\boldsymbol{W}$. Specifically, vector $\boldsymbol{W}(k, l; \cdot)$ characterizes edge distribution between a pair of vertices in communities k and l:

$$\begin{aligned} \mathbb{P}(Y_{i,j} = w_q | X_i = k, X_j = l) &= \frac{\log(n)}{n} W(k, l; q) \qquad \text{for } q = 1, .., Q; \\ \mathbb{P}(Y_{i,j} = 0 | X_i = k, X_j = l) &= 1 - \frac{\log(n)}{n} \sum_{q=1}^{Q} W(k, l; q). \end{aligned} \tag{16}$$

Note that $\boldsymbol{W}$ is not indexed by the network size n, i.e. it does not change with n. This implies that the distribution described above will change with n. More precisely, the probability of having an edge scales as $\Theta(\frac{\log(n)}{n})$ and the degree scales as $\Theta(\log(n))$. This logarithmic growth rate of degree with respect to the network size is called the logarithmic degree regime. The literature on exact recovery studies this regime because it is dense enough that the graph is connected with high probability; yet it is still sparse enough that the conditional independence condition yields asymptotic independence of the failures of the component-MAP for different vertices.

Next, we provide the definition of exact recovery. Exact recovery is an asymptotic requirement in the context of the SBM - a counterpart of consistency in the classical statistical problems.[21]

Definition 2 (Exact Recovery). Exact recovery is solved if there exists an algorithm such that $\mathbb{P}(\boldsymbol{X}^{est} = \boldsymbol{X}) \to 1$ as $n \to \infty$ where $\boldsymbol{X}^{est}$ is the estimated community affiliation.[22]

In other words, exact recovery requires that for a large enough network, the probability of correctly recovering the entire community structure (i.e. no misclassification) is almost 1. The most promising estimator to solve exact recovery is the MAP estimator because it minimizes $\mathbb{P}(\boldsymbol{X}^{est} \neq \boldsymbol{X})$ - if MAP fails in solving exact recovery, no other algorithm can succeed (see, e.g., Abbe (2017)).

In order to describe the condition for exact recovery, we first define the Chernoff-Hellinger (CH) divergence (Abbe and Sandon (2015)).

Definition 3 (Chernoff-Hellinger divergence $D(\boldsymbol{\theta}, \boldsymbol{W})$).

$$D(\boldsymbol{\theta}, \boldsymbol{W}) = \min_{k,l:k\neq l} D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l)) \tag{17}$$

where $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l))$ is given by

$$D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l)) = \max_{\lambda\in[0,1]} \sum_{q=1}^{Q} \sum_{j=1}^{m} \theta_j [(1-\lambda)W(k,j;q)+\lambda W(l,j;q)-W(k,j;q)^{1-\lambda}W(l,j;q)^{\lambda}] \tag{18}$$

Intuitively, $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l))$ measures the difference in connection patterns between a pair of communities, k and l; and thus the CH-divergence $D(\boldsymbol{\theta}, \boldsymbol{W})$ is the minimum of such difference between any pair of distinct communities. In the following, we explain more precisely the meaning of difference in connection pattern. Note that mathematically $\theta_j[(1-\lambda)W(k,j;q) + \lambda W(l,j;q) - W(k,j;q)^{1-\lambda}W(l,j;q)^{\lambda}]$ measures the difference between $\theta_j W(k,j;q)$ and $\theta_j W(l,j;q)$. Moreover, $\theta_j W(k,j;q)\log(n)$ gives, for a vertex in community k, the expected number of w_q-weighted edges with community j; and $\theta_j W(l,j;q)\log(n)$ gives, for a vertex in community l, a similar number. Therefore, $\theta_j[(1-\lambda)W(k,j;q) + \lambda W(l,j;q) - W(k,j;q)^{1-\lambda}W(l,j;q)^{\lambda}]$ measures the difference between communities k and l in terms of the number of w_q-weighted edges with community j. Finally, summing over different communities j and different edge weights w_q delivers the expression in (18). $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l))$ is non-negative. Larger value of $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l))$ represents larger differ-

[21] Exact recovery is sometimes referred to as strong consistency, reflecting the resemblance to consistency.

[22] The equivalence is up to group permutation of $\boldsymbol{X}^{est}$ with respect to community names.

ence in connection patterns between communities k and l, and $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l))$ is zero if and only if the two communities have identical connection patterns, i.e., $W(k,j;q) = W(l,j;q), \forall j, q$.

We further illustrate the definition of CH divergence in a special case of a homogeneous model (Jog and Loh (2015)), where its expression is significantly simplified. In a homogeneous model, vertices are assigned to different communities with equal probabilities; and the distribution of an edge only depends on whether the pair of vertices belong to the same communities:

$$\theta_j = \frac{1}{m}, \forall j; \tag{19}$$

$$W(k,l;\cdot) = \begin{cases} W_{within}(\cdot) \text{ if } k = l \\ W_{between}(\cdot) \text{ if } k \neq l. \end{cases} \tag{20}$$

Under homogeneity, the CH divergence reduces to the Hellinger divergence (corresponding to $\lambda = \frac{1}{2}$), measuring the difference in within-community and between-community connection patterns:

$$\begin{aligned}
& D(\boldsymbol{\theta}, \boldsymbol{W}) \\
=& D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(k), \boldsymbol{W}(l)) \quad \forall k \neq l \\
=& \max_{\lambda\in[0,1]} \sum_{q=1}^{Q}\sum_{j=1}^{m} \theta_j[(1-\lambda)W(k,j;q) + \lambda W(l,j;q) - W(k,j;q)^{1-\lambda}W(l,j;q)^{\lambda}] \\
=& \max_{\lambda\in[0,1]} \frac{1}{m}\sum_{q=1}^{Q}\sum_{j\in\{k,l\}} [(1-\lambda)W(k,j;q) + \lambda W(l,j;q) - W(k,j;q)^{1-\lambda}W(l,j;q)^{\lambda}] \\
=& \max_{\lambda\in[0,1]} \frac{1}{m}\sum_{q=1}^{Q} W_{within}(q) + W_{between}(q) - W_{within}(q)^{1-\lambda}W_{between}(q)^{\lambda} - W_{between}(q)^{1-\lambda}W_{within}(q)^{\lambda} \\
=& \frac{1}{m}\sum_{q=1}^{Q}(\sqrt{W_{within}(q)} - \sqrt{W_{between}(q)})^2.
\end{aligned} \tag{21}$$

The first equation follows symmetry, the second equation directly applies the definition in (18), the third equation uses (19) and $W(k,j;\cdot) = W(l,j;\cdot) = W_{between}(\cdot)\ \forall j \neq k,l$, the fourth equation applies (20), and the last equation results from $\lambda = \frac{1}{2}$ being the maximizer.

Now we can state the main theoretical results. Combining Theorem 3 and Claim 4 in Yun and Proutiere (2016), we have:

Theorem 1 (Threshold for Exact Recovery). Exact recovery is solvable for LSBM $(n, \boldsymbol{\theta}, \frac{\log(n)}{n}\boldsymbol{W})$) if $D(\boldsymbol{\theta}, \boldsymbol{W}) > 1$.

The theorem shows that: if the difference in connection patterns of any pair of communities is large enough, we can correctly recover the entire community structure from one observed network with high probability. This theorem provides theoretical foundation for the use of the MAP estimator (because it is the "optimal" algorithm in terms of exact recovery), and it is very similar to the consistency results in classical statistics.

Some comment on the case where $D(\boldsymbol{\theta}, \boldsymbol{W}) < 1$ is useful. In this case, the MAP estimator fails exact recovery (i.e. has misclassification) with strictly positive probability. This result should not be interpreted as discouraging: although the probability of having misclassification does not vanish with the growth of the sample size, the misclassification rate defined as the proportion of vertices misclassified could still be low. In our Monte Carlo simulations, we observe that even when the CH divergence is below 1, we still have reasonably good classification. Therefore, even if our application falls in the second case, the MAP estimator is still a sensible choice.

To apply this theorem to our model, note that $\boldsymbol{W}$ corresponds to a composition of the edge formation probability $\Phi(\boldsymbol{\gamma}'\boldsymbol{\beta})$, the conditional weight distribution $\boldsymbol{h}$, and the scaling factor $\frac{\log(n)}{n}$. In the Monte Carlo exercises, and the real data application, we will calculate the CH divergence according to Definition 3 and (22)

$$W(k,l;q) = \frac{n}{\log(n)} \Phi(\ \overline{\boldsymbol{\gamma}(k,l)\prime\boldsymbol{\beta}} \ ) h_{kl}(q) \tag{22}$$

where $\overline{\boldsymbol{\gamma}(k,l)\prime\boldsymbol{\beta}}$ is the median of $(\boldsymbol{\gamma}(k,l,z_i,z_j)\prime\boldsymbol{\beta})$ over (i,j) pairs such that $x_i = k, x_j = l$ or $x_i = l, x_j = k$. This accommodates our introduction of covariates in the edge formation process.

4.2 Estimation Algorithm

Apart from the theoretical challenge, the large size of the campaign finance network poses additional computational challenges. The parameter space of $\mathbf{x}$ is m^n. With 3 categories and 5,806 vertices, the parameter space is far larger than the number of atoms in the universe.[23] Therefore, exact solution to MAP in (14) is infeasible, and instead we need an efficient approximation method. In light of these considerations, we propose a Bayesian algorithm to approximate the posterior distribution of the latent ideology as well as other parameters.

In this Bayesian approach, the latent ideology vector $\mathbf{X}$, and the parameters $\epsilon, \boldsymbol{\theta}, \boldsymbol{\beta}$ are treated as random variables with certain prior probability distributions. Adjacency matrix $\mathbf{y}$ and reported affiliation $\hat{\mathbf{x}}$ are treated as one realization of the random variables $\mathbf{Y}$ and $\hat{\mathbf{X}}$. Observable characteristics $\mathbf{z}$ are treated as fixed and exogenous.

[23] According to Jackson (2010), the estimated number of atoms in the universe is on the order of 2^{270}.

The prior distribution of the latent $\mathbf{X}$ is given by (4) in the network formation model. The prior distribution of $\boldsymbol{\theta}$, the parameter governing the unconditional probability distribution of ideology, is assumed to be a Dirichlet distribution:

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha}^{\theta}), \tag{23}$$

where $\boldsymbol{\alpha}^{\theta} \in \mathbb{R}_{+}^{m}$ is a vector of pre-specified concentration parameters. The prior distribution of ϵ, the parameter governing measurement error, is assumed to be a Beta distribution:

$$\epsilon \sim \mathrm{Beta}(\alpha_1^{\epsilon}, \alpha_2^{\epsilon}), \tag{24}$$

where $(\alpha_1^{\epsilon}, \alpha_2^{\epsilon}) \in \mathbb{R}_{+}^{2}$ is a vector of pre-specified concentration parameters. The prior distributions of $\mathbf{h} = \{\mathbf{h}_{kl}\}_{1 \le k \le l \le m}$, the conditional distribution of edge weight (transfer amount), is assumed to be a Dirichlet distribution:

$$\mathbf{h}_{kl} \sim \mathrm{Dir}(\boldsymbol{\alpha}^{h_{kl}}), \tag{25}$$

where $\boldsymbol{\alpha}^{h_{kl}} \in \mathbb{R}_{+}^{Q}$ is a vector of pre-specified concentration parameters. The prior distribution of $\boldsymbol{\beta}$, the parameter governing edge formation probability, is assumed to be a multivariate normal distribution:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}), \tag{26}$$

where $\tau \in \mathbb{R}_{+}$ is pre-specified standard deviation.

Given the prior probability distributions of $\mathbf{X}$, $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, ϵ, and $\mathbf{h}$, the goal of Bayesian estimation is to update the belief on their joint distribution using data $\mathbf{y}, \hat{\mathbf{x}}$ and $\mathbf{z}$, i.e., to compute the posterior distribution $\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h}, \mathbf{X} | \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z})$. Once the posterior distribution is computed, it is straightforward to assess different objects of interest, especially the marginal distributions $\mathbb{P}(X_i | \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z})$. There is neither an analytically nor a numerically convenient form to directly characterize of the joint posterior distribution. Fortunately, calculations of conditional distributions $\mathbb{P}(X_i | \mathbf{x}_{-\mathbf{i}}, \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h},)$, $\mathbb{P}(\boldsymbol{\theta} | \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\beta}, \epsilon, \mathbf{h}, \mathbf{x})$, $\mathbb{P}(\boldsymbol{\beta} | \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \epsilon, \mathbf{h}, \mathbf{x})$, $\mathbb{P}(\epsilon | \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{h}, \mathbf{x})$, and $\mathbb{P}(\mathbf{h} | \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{x})$ are relatively easy. Therefore, a Gibbs sampler algorithm is used to construct the joint posterior distribution. Gibbs sampler is a Markov Chain Monte Carlo (MCMC) algorithm, which repeatedly samples a set of random variables conditional on the values of all other random variables. It is particularly useful when sampling from conditional distributions is convenient.

Computing the Posterior Distribution $\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h}, \mathbf{X}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z})$. By Bayes' rule, the posterior distribution of X_i is given by:

$$\mathbb{P}(X_i = k|\mathbf{x_{-i}}, \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h},) = \frac{\mathbb{P}(\mathbf{y}, x_1, ..x_{i-1}, x_i = k, x_{i+1}, ...x_n, \hat{\mathbf{x}}|\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h})}{\sum_{l=1}^{m} \mathbb{P}(\mathbf{y}, x_1, ..x_{i-1}, x_i = l, x_{i+1}, ...x_n, \hat{\mathbf{x}}|\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h})},$$

which can be reduced to:

$$\mathbb{P}(X_i = k|\mathbf{x_{-i}}, \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h},) \propto \begin{cases} \theta_k (1-\epsilon) \prod_{j \neq i} [\eta_{ij}(k, x_j) h_{kx_j}(y_{ij})]^{\mathbb{1}(y_{ij}>0)} [1 - \eta_{ij}(k, x_j)]^{\mathbb{1}(y_{ij}=0)} & \text{if } k = \hat{x}_i \\ \theta_k \frac{\epsilon}{m-1} \prod_{j \neq i} [\eta_{ij}(k, x_j) h_{kx_j}(y_{ij})]^{\mathbb{1}(y_{ij}>0)} [1 - \eta_{ij}(k, x_j)]^{\mathbb{1}(y_{ij}=0)} & \text{if } k \neq \hat{x}_i \end{cases} \quad \forall i \in \mathcal{V}^o, \tag{27}$$

where θ_k is the ideology prior, $(1 - \epsilon)$ and $\frac{\epsilon}{m-1}$ are measurement accuracy of the report, and $\prod_{j \neq i} [\eta_{ij}(k, x_j) h_{kx_j}(y_{ij})]^{\mathbb{1}(y_{ij}>0)} [1 - \eta_{ij}(k, x_j)]^{\mathbb{1}(y_{ij}=0)}$ is information embedded in network connections. The posterior is an interaction of the three. When $\epsilon > 0$, i.e. allowing for measurement error, if the network data highly favors an ideology different from $\hat{x}_i$, it is possible for the posterior to override the prior, i.e. a posterior mode at $k \neq \hat{x}_i$. This can be viewed as a data-oriented correction of measurement error. Similarly,

$$\mathbb{P}(X_i = k|\mathbf{x_{-i}}, \mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h},) \propto \theta_k \prod_{j \neq i} [\eta_{ij}(k, x_j) h_{kx_j}(y_{ij})]^{\mathbb{1}(y_{ij}>0)} [1 - \eta_{ij}(k, x_j)]^{\mathbb{1}(y_{ij}=0)} \quad \forall i \in \mathcal{V} \backslash \mathcal{V}^o, \tag{28}$$

where θ_k is the ideology prior, and $\prod_{j \neq i} [\eta_{ij}(k, x_j) h_{kx_j}(y_{ij})]^{\mathbb{1}(y_{ij}>0)} [1 - \eta_{ij}(k, x_j)]^{\mathbb{1}(y_{ij}=0)}$ is information embedded in network connections. The posterior is an interaction of the two. Summoning conjugacy, the posterior distribution of $\boldsymbol{\theta}$ is given by:

$$\boldsymbol{\theta}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\beta}, \epsilon, \mathbf{h}, \mathbf{x} \sim \text{Dir}(\boldsymbol{\alpha}^{\theta} + [n_k]_{1 \leq k \leq m}), \tag{29}$$

where $n_k = \sum_{1 \leq i \leq n} \mathbb{1}(x_i = k)$ is the number of vertices with x_i equal to k. The posterior distribution of ϵ is given by:

$$\epsilon|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{h}, \mathbf{x} \sim \text{Beta}(\alpha_1^{\epsilon} + n_e, \alpha_2^{\epsilon} + n_t), \tag{30}$$

where $n_e = \sum_{i \in \mathcal{V}^o} \mathbb{1}(\hat{x}_i \neq x_i)$ is the number of vertices whose self report is different from its ideology, and $n_t = \sum_{i \in \mathcal{V}^o} \mathbb{1}(\hat{x}_i = x_i)$ is the number of vertices whose self report is the same as its ideology. The posterior distribution of $\mathbf{h}$ is given by:

$$\mathbf{h}_{kl}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{x} \sim \text{Dir}(\boldsymbol{\alpha}^{h_{kl}} + [n_{kl,q}]_{1 \leq q \leq Q}), \tag{31}$$

where $n_{kl,q} = \sum_{1\leq i<j\leq n} \max\{\mathbb{1}(x_i = k, x_j = l), \mathbb{1}(x_i = l, x_j = k)\}\mathbb{1}(y_{ij} = q)$ is the number of edges with transfer amount w_q between PCs of ideologies k and l. Constructing the posterior distribution of $\boldsymbol{\beta}$ is more delicate. $\boldsymbol{\beta}$ is essentially the coefficient vector in a Probit regression model, whose posterior distribution does not have an analytically convenient form. Therefore, instead of directly sampling from a closed form distribution, a data augmentation strategy introduced in Albert and Chib (1993) is used. First, sample auxiliary variable $\mathbf{u} = \{u_{ij}\}_{1\leq i<j\leq n}$ from the following truncated normal distributions:

$$u_{ij} \sim \begin{cases} \mathcal{N}(\boldsymbol{\gamma}(x_i, x_j, \mathbf{z}_i, \mathbf{z}_j)'\boldsymbol{\beta}, 1)|u_{ij} > 0 \text{ if } y_{ij} > 0 \\ \mathcal{N}(\boldsymbol{\gamma}(x_i, x_j, \mathbf{z}_i, \mathbf{z}_j)'\boldsymbol{\beta}, 1)|u_{ij} < 0 \text{ if } y_{ij} = 0. \end{cases} \tag{32}$$

Conditional on the auxiliary variable $\mathbf{u}$, the posterior distribution of $\boldsymbol{\beta}$ is given by:

$$\boldsymbol{\beta}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}, \epsilon, \mathbf{h}, \mathbf{x}, \mathbf{u} \sim \mathcal{N}\left((\tau^{-2}\mathbf{I} + \boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{u},\ (\tau^{-2}\mathbf{I} + \boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\right), \tag{33}$$

where $\boldsymbol{\Gamma} = \left[\boldsymbol{\gamma}(x_i, x_j, \mathbf{z}_i, \mathbf{z}_j)'\right]_{1\leq i<j\leq n}$.

The Gibbs sampler algorithm is summarized below:

1. Initialize $\mathbf{x}^0, \boldsymbol{\theta}^0, \boldsymbol{\beta}^0$, ϵ^0, $\mathbf{h}^0$.

2. Iteratively sample from conditional posterior distribution. Specifically, in iteration t, we sample one set of parameters $(\mathbf{x}^t, \boldsymbol{\theta}^t, \epsilon^t, \mathbf{h}^t, \boldsymbol{\beta}^t)$ with the following procedure:

   (a) Sample $\{x_i\}^t_{1\leq i\leq n}$ sequentially from distribution $\mathbb{P}(X_i|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}^{t-1}, \boldsymbol{\beta}^{t-1}, \epsilon^{t-1}, \mathbf{h}^{t-1}, x_1^t, ..x_{i-1}^t, x_{i+1}^{t-1}, ...x_n^{t-1})$ using (27) and (28).

   (b) Sample vector $\boldsymbol{\theta}^t$ from distribution $\mathbb{P}(\boldsymbol{\theta}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\beta}^{t-1}, \epsilon^{t-1}, \mathbf{h}^{t-1}, \mathbf{x}^t)$ using (29).

   (c) Sample ϵ^t from distribution $\mathbb{P}(\epsilon|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\beta}^{t-1}, \mathbf{h}^{t-1}, \mathbf{x}^t, \boldsymbol{\theta}^t)$ using (30).

   (d) Sample vectors $\mathbf{h}^t$ from distribution $\mathbb{P}(\mathbf{h}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \boldsymbol{\beta}^{t-1}, \mathbf{x}^t, \boldsymbol{\theta}^t, \epsilon^t)$ using (31)

   (e) Sample auxiliary vector $\mathbf{u}^t$ from distribution $\mathbb{P}(\mathbf{u}|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}^{t-1}, \mathbf{x}^t)$ using (32).

   (f) Sample vector $\boldsymbol{\beta}^t$ from distribution $\mathbb{P}(\boldsymbol{\beta}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z}, \mathbf{x}^t, \boldsymbol{\theta}^t, \epsilon^t, \mathbf{h}^t, \mathbf{u}^t)$ using (33).

3. Burn in the first T_1 iterations. Use samples from iterations $T_1 + 1$ to $T_1 + T_2$ to construct posterior distribution.

The first step initializes the Markov chain. The initial values should not affect the steady state and can be determined either at random or by other algorithms. We use the former in our application. The second step simulates the Markov chain by repeatedly sampling one parameter conditional on the values of all other parameters. The sampling order of the parameters is arbitrary, and a different order can be used, e.g. one can sample $\boldsymbol{\theta}$ before $\mathbf{x}$. In order to speed up convergence, we use the newly sampled parameter immediately in the following sampling procedures and do not wait until the next iteration, e.g., $\mathbf{x}^t$ sampled from (a) is used in the sampling of $\boldsymbol{\theta}^t$ in (b). The final step discards the initial portion of the Markov chain, namely the first T_1 iterations, where steady state is not reached. Pooling the remainder samples gives an approximate joint posterior distribution $\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{\beta}, \epsilon, \mathbf{h}, \mathbf{X}|\mathbf{y}, \hat{\mathbf{x}}, \mathbf{z})$.

5 Monte Carlo Evidence

In this section, we present Monte Carlo evidence to evaluate the performance of the community recovery algorithm proposed in Section 4. We conduct four sets of Monte Carlo simulations that differ in the specifications of the edge formation process and the network size, both of which affect the Chernoff-Hellinger divergence measure as defined in (18) because they enter the expression of $W(k, l; q)$ (see Eq. (22)).

The first three sets of Monte Carlo simulations share a framework that is similar to the homogeneous labeled stochastic block model as in Jog and Loh (2015), even though the edge distributions are not exactly homogeneous due to the introduction of the covariates in the edge formation probability. Our Monte Carlo results show a strong confirmation of Theorem 1 that CH divergence of 1 is a sharp threshold for exact recovery. In the first specification, the network edge formation patterns and the network size imply a CH divergence of 1.0074, and we find that the misclassification rate is on average 1.10%. The second specification differs from the first one only in the weight distribution, resulting in a smaller CH divergence of 0.4719 and we find a larger average misclassification rate of 5.68%. The third specification differs from the second one only in the network size, which results in a larger CH divergence of 1.7537 and we find a smaller average misclassification rate of 0. In the fourth set of Monte Carlo simulations, the network size is similar to the real data (about 6,000 vertices), and the edge distributions are heterogeneous in a flexible way, resulting in a CH divergence of 6.0110. This is intended to assess the performance of the algorithm in a data set that resembles the real data. The results show that the algorithm performs surprisingly well with an average misclassification rate of 0.0002%, although it is slower due to the scale of the network. We summarize our main simulation results in the text, but many of the less essential details are left in Appendices C-F.

Common Specifications Across All Four Sets of Monte Carlo Simulations. The specifications that are common across all four sets of Monte Carlo studies are listed as follows. The aspects of the specifications that are unique to each set of Monte Carlo studies, as well as their specific results, are presented separately in subsequent subsections.

- The number of ideologies: $m = 3$.
- Marginal distribution of ideology: $\boldsymbol{\theta} = (1/3, 1/3, 1/3)$.
- Probability of measurement error: $\epsilon = 0.05$.
- Fraction of vertices with reported affiliation: 40%.
- Edge formation probability is specified by $y_{ij} > 0$ iff

  $$\boldsymbol{\gamma}(x_i, x_j, \mathbf{z}_i, \mathbf{z}_j)'\boldsymbol{\beta} + e_{ij} > 0,$$

  where $\boldsymbol{\gamma}(x_i, x_j, \mathbf{z}_i, \mathbf{z}_j) = (\boldsymbol{\gamma_x}, \boldsymbol{\gamma_z})$ is a list of 19 variables and $\boldsymbol{\beta} = (\boldsymbol{\beta_x}, \boldsymbol{\beta_z}) \in \mathbb{R}^{19}$. To make the simulated data as close as possible to the real data we use for the empirical analysis, we include in the vector $\mathbf{z}_i$ the PC's budget, state, industry, dummy for House campaign, dummy for Senate campaign, dummy for Presidential campaign, dummy for qualified PAC, dummy for qualified Party, dummy for national committee, dummy for authorized by a candidate, and dummy for joint fund-raiser, and we construct $\boldsymbol{\gamma_z} \in \mathbb{R}^{13}$ based on $(\mathbf{z}_i, \mathbf{z}_j)$; specifically,

  $$\boldsymbol{\gamma_z} = \left(\begin{array}{c} \mathbb{1}_{\text{state}_i=\text{state}_j}, \mathbb{1}_{\text{industry}_i=\text{industry}_j}, \\ \mathbb{1}_{(\text{house}_i=1)\vee(\text{house}_j=1)}, \mathbb{1}_{(\text{senate}_i=1)\vee(\text{senate}_j=1)}, \mathbb{1}_{(\text{president}_i=1)\vee(\text{president}_j=1)}, \\ \mathbb{1}_{(\text{qualified PAC}_i=1)\vee(\text{qualified PAC}_j=1)}, \mathbb{1}_{(\text{qualified Party}_i=1)\vee(\text{qualified Party}_j=1)}, \\ \mathbb{1}_{(\text{national}_i=1)\vee(\text{national}_j=1)}, \mathbb{1}_{(\text{authorized}_i=1)\vee(\text{authorized}_j=1)}, \mathbb{1}_{(\text{fundraiser}_i=1)\vee(\text{fundraiser}_j=1)}, \\ \left[\ln b_i + \ln b_j\right], \left[(\ln b_i)^2 + (\ln b_j)^2\right], \ln b_i \ln b_j \end{array} \right). \tag{34}$$

  We set

  $$\beta_{\mathbf{z}} = \left(\begin{array}{c} 0.3, 0.3, 0.1, 0.1, 0.1, 0.2, 0.2, \\ 0.15, 0.15, 0.15, 0.01, -0.01, 0.001 \end{array} \right)$$

  in all four sets of Monte Carlo simulations.

  $\boldsymbol{\gamma_x} \in \mathbb{R}^6$ includes a constant term and interaction terms of x_i and x_j; specifically,

  $$\boldsymbol{\gamma_x} = \left(\begin{array}{c} 1, \mathbb{1}_{x_i=x_j=\text{Dem}}, \ \mathbb{1}_{x_i=x_j=\text{Rep}}, \ \mathbb{1}_{x_i=x_j=\text{Ind}}, \\ \mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})}, \ \mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})} \end{array} \right). \tag{35}$$

- Transfer amount is discretized into four bins. Therefore, conditional on $y_{ij} > 0$, $y_{ij} \in \{1, 2, 3, 4\}$.

5.1 Monte Carlo I: 500 Networks with $n = 100$, CH Divergence Exceeding 1

In the first set of Monte Carlo simulations, 500 networks are simulated and estimated. They have network size $n = 100$, the coefficients in the network formation probability are given by $\boldsymbol{\beta_x} = (-1.5, 0.5, 0.5, 0.5, 0, 0)$, and the edge's weight distributions are given by $\mathbf{h}_{\text{Dem,Dem}} = \mathbf{h}_{\text{Rep,Rep}} = \mathbf{h}_{\text{Ind,Ind}} = (0.05, 0.1, 0.4, 0.45)$, and $\mathbf{h}_{\text{Dem,Rep}} = \mathbf{h}_{\text{Dem,Ind}} = \mathbf{h}_{\text{Rep,Ind}} = (0.4, 0.3, 0.2, 0.1)$. Note that the specification of $\boldsymbol{\beta_x} = (-1.5, 0.5, 0.5, 0.5, 0, 0)$ implies that the link formation depends on whether the two vertices are of the same ideology or are of differen ideologies, and fits into the homogeneity special case we described in Eq. (20). The implied CH divergence according to (21) and (22) is $1.0074 > 1$.

For these simulations, the total execution time is 235,405 seconds (about 65 hours). The speed of convergence in terms of the number of iterations varies, which is shown in Figure C1 where we plot the histogram of the number of iterations (including burn-in and posterior) across the 500 networks we simulated. The distribution of misclassification rates is summarized in Figure C2 and Table C1.[24] They are small in general, and in 37.4% of the simulations, there is no misclassification.[25] Moreover, 98.8% of the simulations have misclassification rates lower than the measurement error rate 0.05, which implies that in most cases our algorithm successfully corrects some of the misreports. Table C2 provides a detailed tabulation of the estimated vs. true ideologies.

The results above focus only on the posterior mode, and the following analysis further investigates the patterns of the posterior distributions. For the correctly classified vertices, our categorical classification based on posterior mode is rather precise. Figure C3 shows that the differences in posterior probability between the highest posterior probability (i.e., the posterior of the true ideology) and the second highest posterior probability are highly concentrated around 1. This indicates that for most of these vertices, the posterior distribution is strongly informative of the true ideology. For the misclassified vertices, however, the scales of misclassification vary. Figure C4 shows that the differences in the posterior probability between the highest posterior probability and the posterior probability of the true ideology are approximately uniformly distributed between 0 where the classification is only a bit off and 1 where the classification is far off. This suggests that the mis-

[24] Estimated ideology is defined as the posterior mode.

[25] 37.4% exact recovery is lower than the theoretical prediction, for several reasons. First, probability of exact recovery approaches 1 only when network size approaches infinity. In this case, the network size is only 100, which may be too small. Second, our model is not exactly the same as the model in Jog and Loh (2015) because we include covariates in the edge formation probability, which is more complicated. Third, we do not use an exact MAP estimator, which may also introduce approximation error.

classification is likely to be caused by unusual realizations of the networks process rather than the failure of our estimation algorithm. The randomness in the network formation renders the ideology information unclear or even misleading for some vertices, though this occurs rarely. Additional analysis of the posterior mean of other parameters are presented in Tables C3-C6. These tables show that the algorithm recovers the true parameters effectively, except for the weight distribution parameters $\mathbf{h}_{kl}$. We will show that these parameters will be estimated more precisely as the network size gets larger in Monte Carlo simulations III (where $n = 500$) and IV (where $n = 6000$).

5.2 Monte Carlo II: 500 Networks with $n = 100$, CH Divergence Less Than 1

The specifications for the second set of Monte Carlo simulations are the same as those for the first set except for the weight distributions $\mathbf{h}_{\text{Dem,Dem}}$, $\mathbf{h}_{\text{Rep,Rep}}$ and $\mathbf{h}_{\text{Ind,Ind}}$. Specifically, the edge's weight distributions are given by $\mathbf{h}_{\text{Dem,Dem}} = \mathbf{h}_{\text{Rep,Rep}} = \mathbf{h}_{\text{Ind,Ind}} = (0.2, 0.15, 0.35, 0.3)$, and $\mathbf{h}_{\text{Dem,Rep}} = \mathbf{h}_{\text{Dem,Ind}} = \mathbf{h}_{\text{Rep,Ind}} = (0.4, 0.3, 0.2, 0.1)$. Again, 500 networks with size $n = 100$ are simulated and estimated. The implied CH divergence according to (21) and (22) decreases to 0.4719, which is now less than 1.

The total execution time for these simulations is 418,322 seconds (about 5 days). The speed of convergence in terms of the number of iterations is relatively slow; Figure D1 depicts the histogram of the number of iterations (including burn-in and posterior). Figure D2 and Table D1 summarize the distribution of misclassification rates across the 500 simulations. Comparing with Monte Carlo I, the misclassification rates are higher; for the worst case, the misclassification rate is as high as 24%. Only 2.8% of the simulations have 0 misclassification. This is consistent with the theoretical prediction of Theorem 1 that a CH divergence lower than 1 is associated with low probability of exact recovery. Table D2 provides a detailed tabulation of the estimated vs. true ideologies in this set of Monte Carlo simulations. Figure D3 plots, for the correctly classified vertices, the histogram of the difference in posterior probability between the highest posterior probability (i.e., the posterior for the true ideology) and the second highest posterior probability. It is shown to be mostly concentrated at 1, though relative to Figure C3 for Monte Carlo I, the difference is somewhat more likely to be less than 1 and is more spread out. This indicates that, when CH is less than 1, our categorizations are not as informative even though we obtained the correct classification. Similarly, Figure D4 plots, for the misclassified vertices, the histogram of the difference in posterior probability between the highest posterior probability and the posterior probability of the true ideology. The difference is rather evenly distributed, which is similar to Figure C4 in Monte Carlo I. Additional analysis of the posterior mean of other parameters are presented in Tables D3-D6. These tables show that the algorithm recovers the true parameters effectively, except for the weight distribution parameters $\mathbf{h}_{kl}$.

5.3 Monte Carlo III: 500 Networks with $n = 500$, CH Divergence Exceeding 1

The specifications for the third set of Monte Carlo simulations are the same as those of Monte Carlo II except for the network size. Now we simulate 500 networks, each with a network size of $n = 500$. As a result of the increase in the network size, the implied CH divergence according to (21) and (22) is now 1.7537, which is larger than 1.

The total execution time for these simulations is 31,974 seconds (about 9 hours). The speed of convergence is fast in terms of the number of iterations (ranging from 600 to 900); Figure E1 depicts the histogram of the number of iterations (including burn-in and posterior). The misclassification rates in all 500 simulations are 0. Therefore, this set of simulations over-perform the previous two in terms of both convergence speed and accuracy rate. Figure E2 plots the histogram of the difference in posterior probability between the highest posterior probability (i.e., the posterior for the true ideology) and the second highest posterior probability for the correctly classified vertices (which are all vertices because of 0 misclassification rate). The difference is almost completely concentrated at 1, indicating that our categorization based on posterior mode is very informative; in fact, for 99.9976% of the correctly classified vertices, the posterior probability on the true ideology is 1. Additional analysis of the posterior mean of other parameters are presented in Tables E1-E4. These tables show that the algorithm recovers the true parameters effectively, including the weight distribution parameters $\mathbf{h}_{kl}$ (see Table E3).

5.4 Monte Carlo IV: 100 Networks with $n = 6,000$, CH Divergence Exceeding 1

In the fourth set of Monte Carlo simulations, we make several important changes. First, we increase the network size to $n = 6000$, which is comparable to the size of the political contribution network in our data. Also importantly, we deviate from the symmetry in the within-community and between-community link formation probabilities. Both changes are intended to assess the performance of our estimation algorithm in an environment that resembles the actual data. We again simulate and estimate 100 networks. Specifically, the coefficient in the network formation probability is given by $\boldsymbol{\beta_x} = (-3, 1, 1, 0.7, 0.3, 0.3)$, and the edge's weight distributions are given by $\mathbf{h}_{\text{Dem,Dem}} = \mathbf{h}_{\text{Rep,Rep}} = (0.1, 0.2, 0.2, 0.5)$, $\mathbf{h}_{\text{Ind,Ind}} = (0.25, 0.25, 0.25, 0.25)$, $\mathbf{h}_{\text{Dem,Rep}} = (0.5, 0.2, 0.2, 0.1)$, and $\mathbf{h}_{\text{Dem,Ind}} = \mathbf{h}_{\text{Rep,Ind}} = (0.3, 0.3, 0.3, 0.1)$. Using expressions (18) and (22) to evaluate pairwise divergence between ideology communities, we have $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(\text{Dem}), \boldsymbol{W}(\text{Rep})) = 13.1003$, and $D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(\text{Dem}), \boldsymbol{W}(\text{Ind})) = D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(\text{Rep}), \boldsymbol{W}(\text{Ind})) = 6.011$, and thus the CH divergence $6.011 > 1$.

The total execution time for the fourth set of simulations is 930,673 seconds (about 11 days). The speed of convergence is fast in terms of the number of iterations (ranging from 600 to 900); Figure F1 depicts the histogram of the number of iterations. Therefore, the long execution time is a result of heavy computation in each iteration, not the number of iteration. Misclassification rates are 0 for 99 simulations, 0.0167% (i.e., 1 vertex is misclassified) for 1 simulation. For the correctly classified vertices, the numerical posterior distributions of ideology are degenerate in the true ideology. For the only vertex that is incorrectly classified in this set of simulations, the difference between the highest posterior probability and the posterior probability of the true ideology is 0.13. These results suggest that the simulated data exhibit strong information on the community structure, and that our algorithm is efficient in identifying this structure. Additional analysis of the posterior mean of other parameters are presented in Tables F1-F4. These tables show that the algorithm recovers the true parameters effectively, including the weight distribution parameters $\mathbf{h}_{kl}$.

To summarize, our algorithm has excellent performance when the data is generated with CH divergence greater than 1. It has reasonably good performance even when the data is generated with CH divergence lower than 1. The results also suggest that a large network is not necessarily undesirable. On the one hand, it brings in more computational burden and increases the runtime; on the other hand, it also embodies more information and speeds up the convergence.

6 Empirical Implementation and Results

6.1 Empirical Implementation.

We empirically infer the ideologies of 5,806 PCs from the giant component in the campaign finance network depicted in Figure 3. There are 3 categories of ideologies: Democratic, Republican, and Independent. For the small number of PCs whose self-reported affiliations do not belong to these categories, we treat them as if we do not observe their report. The set $\mathcal{V}^o$ contains PCs with self-reported affiliations $\hat{x}_i$'s. We assume the following functional form for the edge formation probability:

$$
\begin{aligned}
\boldsymbol{\gamma}(x_i, x_j, \mathbf{z}_i, \mathbf{z}_j)'\boldsymbol{\beta} \;\; = \;\; & \beta_1 + \beta_2 \mathbb{1}_{x_i=x_j=\text{Dem}} + \beta_3 \mathbb{1}_{x_i=x_j=\text{Rep}} + \beta_4 \mathbb{1}_{x_i=x_j=\text{Ind}} \\
& + \beta_5 \mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})} + \beta_6 \mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})} \\
& + \beta_7 \mathbb{1}_{\text{state}_i=\text{state}_j} + \beta_8 \mathbb{1}_{\text{industry}_i=\text{industry}_j} + \beta_9 \mathbb{1}_{(\text{house}_i=1)\vee(\text{house}_j=1)} \\
& + \beta_{10} \mathbb{1}_{(\text{senate}_i=1)\vee(\text{senate}_j=1)} + \beta_{11} \mathbb{1}_{(\text{president}_i=1)\vee(\text{president}_j=1)} \\
& + \beta_{12} \mathbb{1}_{(\text{qualified PAC}_i=1)\vee(\text{qualified PAC}_j=1)} + \beta_{13} \mathbb{1}_{(\text{qualified Party}_i=1)\vee(\text{qualified Party}_j=1)} \\
& + \beta_{14} \mathbb{1}_{(\text{national}_i=1)\vee(\text{national}_j=1)} + \beta_{15} \mathbb{1}_{(\text{authorized}_i=1)\vee(\text{authorized}_j=1)} \\
& + \beta_{16} \mathbb{1}_{(\text{fundraiser}_i=1)\vee(\text{fundraiser}_j=1)} + \beta_{17} \left[\ln b_i + \ln b_j\right] + \beta_{18} \left[(\ln b_i)^2 + (\ln b_j)^2\right] \\
& + \beta_{19} \ln b_i \ln b_j, \qquad (36)
\end{aligned}
$$

where the first term is a constant characterizing the baseline connection probability between Democratic and Republican PCs; the second to the sixth terms characterize the connection probabilities for other ideology pairs; the seventh and the eighth terms capture the effect of the two PCs belonging to the same state or industry; the ninth to the sixteenth terms capture the effects of PCs' institutional characteristics: whether one of them is a House campaign, a Senate campaign, a Presidential campaign, a qualified PAC, a qualified Party, a national committee, authorized by a candidate, or a joint fundraiser; and the seventeenth to the nineteenth terms capture the effect of both PCs' budgets on link formations probability, which is a restrictive form of that in (8) and assumes that the effect of financial and institutional characteristics are the same across ideological pairs. The main reason for this parsimonious specification is to reduce the computational intensity. The estimation results do not seem to show signs of severe mis-specification. The transfer amount Y_{ij} is discretized into multiples of \$500, and can take values of $\{0, 1, 2, ..., 100\}$ where 100 includes all the transfer higher than \$50,000. The initial values in the Gibbs sampler are randomly generated, and different sets of initial values are used.

6.2 Estimation Results: Posterior Mean and Standard Deviations.

Table 10 presents the posterior mean and standard deviation of $\boldsymbol{\beta}$, the coefficients in edge formation probability. The second to the sixth coefficients are all positive, indicating that Democratic and Republican PCs (the baseline case) have the lowest connection probability. Additionally, the Democratic PCs have stronger within party connection than the Republican PCs. Moreover, Independent PCs have a higher probability of connecting with Democratic or Republican PCs than other Independent PCs. It is also interesting to note that, *everything else equal,* pairs of Republican

PCs are less likely to form a link than Republican/Independent or Democratic/Independent pairs of PCs.

Tables 11 and 12 present the posterior mean and standard deviation of $\boldsymbol{\theta}$, the unconditional probability of ideology; and ϵ, the measurement error. Tables 11 shows that in the population of all PCs, 40.01% are Democratic, 42.74% are Republican, and 17.24% are Independent. The posterior standard deviations of these estimates are small. Table 12 shows that the self-reported ideologies of the PCs are likely to be erroneous with probability 7.37%.

Due to the large number of parameters in the weight distribution function, we present the estimates of $\mathbf{h}$ in Figure 10, which shows the posterior means of all the values of $(h_{kl,1}, ..., h_{kl,100})$ for all $k, l \in$ {Dem, Rep, Ind} pairs graphically.

$\boldsymbol{\beta}$	Posterior Mean	Posterior Standard Deviation
constant	-7.4729	0.0166
$\mathbb{1}_{x_i=x_j=Dem}$	1.3301	0.0099
$\mathbb{1}_{x_i=x_j=Rep}$	0.8210	0.0130
$\mathbb{1}_{x_i=x_j=Ind}$	0.6405	0.0134
$\mathbb{1}_{(x_i=Dem,x_j=Ind)\vee(x_i=Ind,x_j=Dem)}$	1.2909	0.0121
$\mathbb{1}_{(x_i=Rep,x_j=Ind)\vee(x_i=Ind,x_j=Rep)}$	1.5797	0.0119
Same state	0.6399	0.0043
Same industry	0.2185	0.0117
One of them is a House campaign	0.5306	0.0034
One of them is a Senate campaign	0.3783	0.0035
One of them is a Presidential campaign	0.0212	0.0113
One of them is a qualified PAC	0.7006	0.0049
One of them is a qualified Party	-0.5334	0.0066
One of them is a national committee	0.9421	0.0133
One of them is authorized by a candidate	-0.4473	0.0090
One of them is a joint fundraiser	-0.7623	0.0059
$(\ln b_i + \ln b_j)$	-0.0442	0.0023
$((\ln b_i)^2 + (\ln b_j)^2)$	0.0162	0.00004
$\ln b_i \ln b_j$	-0.0010	0.0002

Table 10: Posterior Distribution of β

$\boldsymbol{\theta}$	Posterior Mean	Posterior Standard Deviation
$\mathbb{P}$(Dem)	0.4001	0.0045
$\mathbb{P}$(Rep)	0.4274	0.0046
$\mathbb{P}$(Ind)	0.1724	0.0036

Table 11: Posterior Distribution of θ

$\boldsymbol{\epsilon}$	Posterior Mean	Posterior Standard Deviation
	0.0737	0.0042

Table 12: Posterior Distribution of ϵ

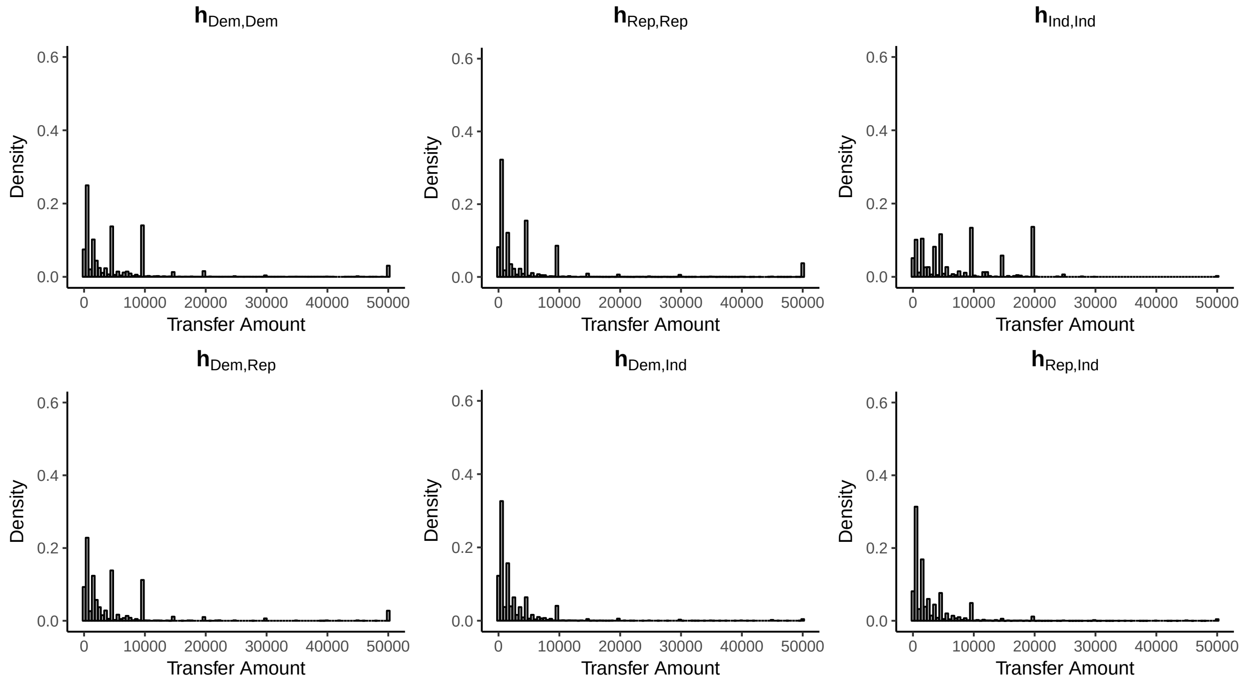


Figure 10: Posterior Mean of h

Note: Bin size is 500. Probability of transfer amount higher than $50,000 is plotted at $50,000.

Chernoff-Hellinger Divergence of the Estimated Model. Using the posterior mode of $\mathbf{x}$, and the posterior means of $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{h}$, we calculate the implied Chernoff-Hellinger divergence:

$$D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(\text{Dem}), \boldsymbol{W}(\text{Rep})) = 14.7760,$$
$$D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(\text{Dem}), \boldsymbol{W}(\text{Ind})) = 26.6116,$$
$$D_{L+}(\boldsymbol{\theta}, \boldsymbol{W}(\text{Rep}), \boldsymbol{W}(\text{Ind})) = 22.4425,$$

and thus the CH divergence is $14.7760 > 1$. Therefore, the data generating process, corresponding to our estimated parameters, satisfies the condition for exact recovery as stated in Theorem 1.

6.3 Comparing Estimated and Self-Reported Ideologies for PCs that Self Report Ideologies

Using the posterior mode as a point estimate of the ideology, Table 13 presents the cross tabulation of all PCs according to self-reported and estimated ideology. Overall, 90.70% of our estimates match the self reports: 94.36% for self-reported Democratic PCs, and 89.49% for self-reported Republican PCs.

	Estimated Dem	Estimated Rep	Estimated Ind
Self-Reported Dem	954 (94.36%)	43 (4.25%)	14 (1.38%)
Self-Reported Rep	46 (4.60%)	894 (89.49%)	59 (5.91%)
Self-Reported Ind	16 (29.09%)	14 (25.45%)	25 (45.45%)
No Reported Affiliation	748 (20.00%)	1,202 (32.13%)	1,791 (47.87%)

Table 13: Tabulation of Estimated vs. Self-Reported Ideology
Note: The percentages are calculated for each row.

6.4 Re-examining the Network Statistics Using Estimated Ideologies

In Section 3, we presented the network statistics based on self-reported ideologies for those PCs that self reported their ideologies. Here we re-examine these statistics based on estimated ideologies of all PCs. Conditional on the estimated ideologies, the empirical distribution of transfer amount is shown in Figure 11, and the mean of each distribution is shown in Table 14. On average, the transfer amount is the highest for Democratic PC pairs and Republican PC pairs, and this is partially due to the heavy tail of the within-party contributions. The transfer amount is smaller between Democratic and Republican PCs, and smallest when it involves Independent PCs. This is consistent with the estimates on connection probability. Independent PCs have overall high connection probability, but the associated transfer amount is small. Democratic and Republican PCs have relatively lower within-party connection probability, but the associated transfer amount is larger.

Dem, Dem	Rep, Rep	Ind, Ind	Dem, Rep	Dem, Ind	Rep, Ind
26,311.03	22,165.53	8,144.17	18,507.13	3,206.26	3,769.65

Table 14: Mean of Transfer Amount Conditional on Estimated Ideology

Next, we compare the contribution patterns of the PCs according to their self-reported vs. estimated ideologies. For each PC i, let numDem_i, numRep_i, and numInd_i denote its numbers of

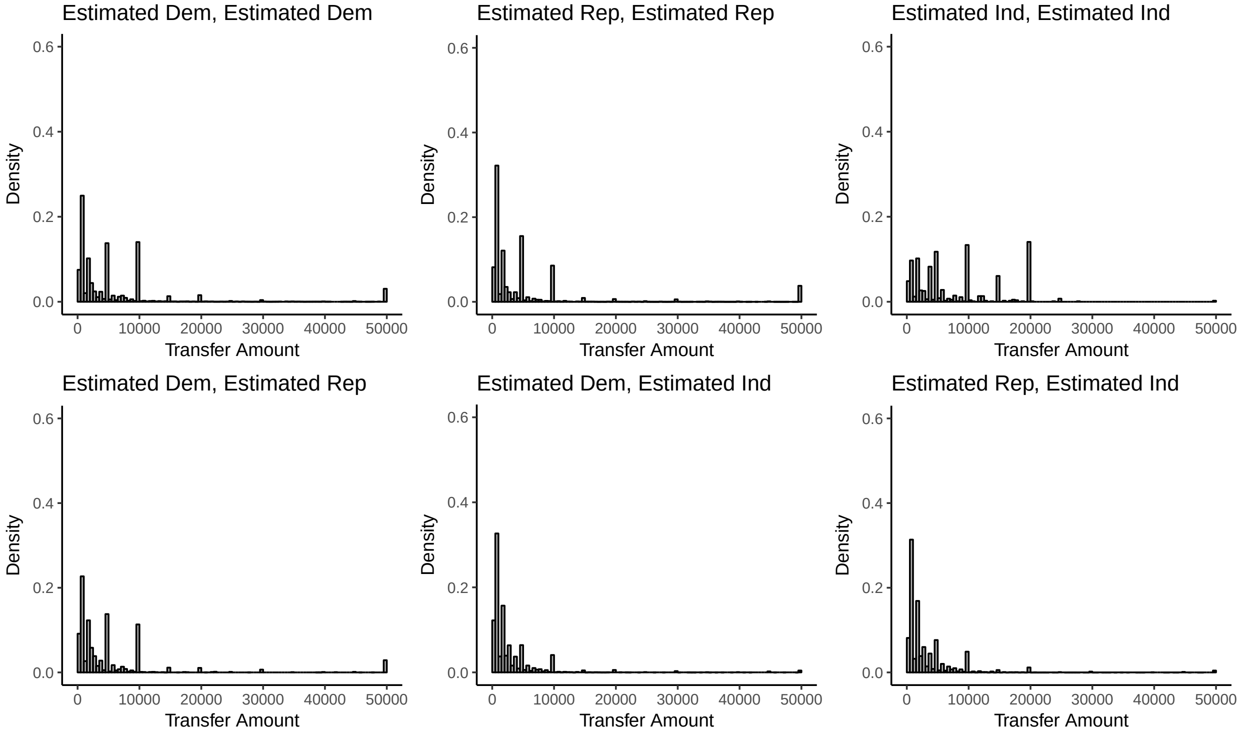


Figure 11: Empirical Distribution of Transfer Amount Conditional on Estimated Ideology
Note: Bin size is 500. Observations with transfer amount higher than $50,000 are plotted at $50,000.

connections with (estimated) Democratic, Republican, and Independent PCs respectively. In Figures 12-14, we plot the distributions of numDem (dark bar), numRep (white bar), and numInd (gray bar) for different groups of PCs, focusing on the difference between self-reported and estimated ideologies. The left panel in Figure 12 presents the distributions for PCs that *self reported* to be Democratic, and the right panel for PCs that did not self report but are estimated to be Democratic PCs. Figures 13 and 14 are similar, but for Republican and Independent PCs respectively. Qualitatively, the degree distributions have similar patterns for self-reported and estimated PCs with the same ideology. As a robustness check, we redo the analysis above, with each connection weighted by transfer amount. Specifically, for each PC, we calculate its total amount of transfer to and from (estimated) Democratic, Republican, and Independent PCs respectively, and then plot the distributions of these numbers for different groups of PCs. The histograms are shown in Figures 15, 16, and 17, respectively for Democratic, Republican and Independent PCs. Again, the distributions are similar for self-reported and estimated PCs with the same ideology. These results demonstrate that PCs without self-report, classified according to our estimation results, behave similarly to PCs with the corresponding self reports of affiliations.

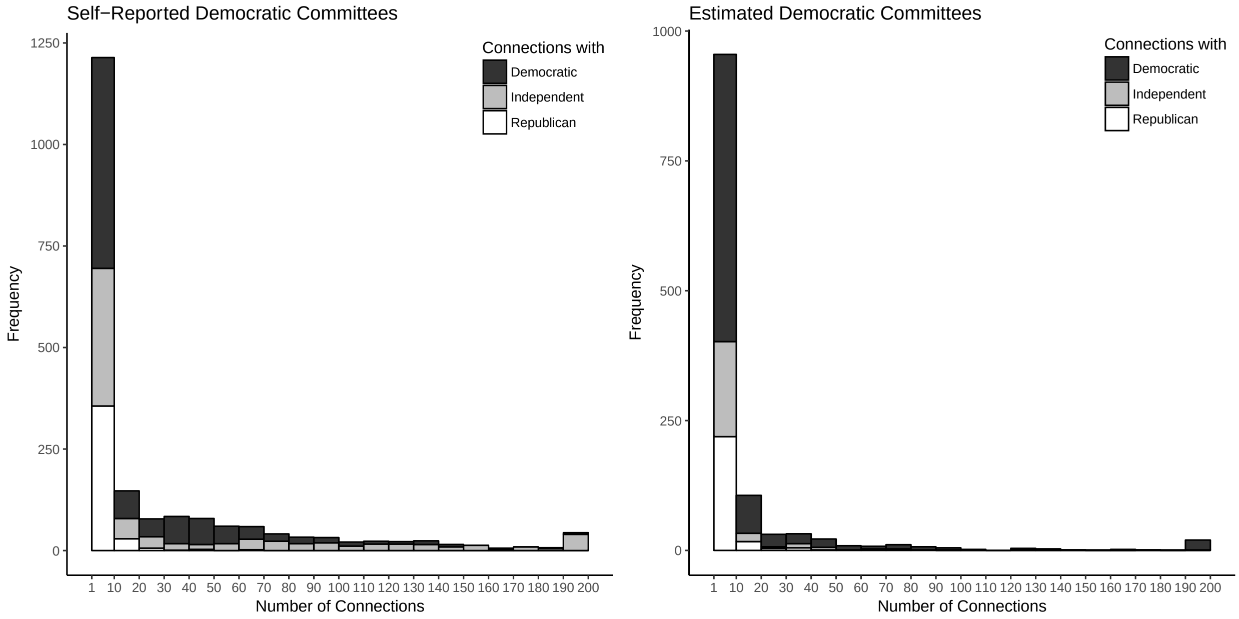


Figure 12: Self-reported vs. Estimated Democratic PCs: Distributions of Number of Connections

Note: These figures only include observations with positive number of connections. Observations with more than 200 connections are plotted at 200.

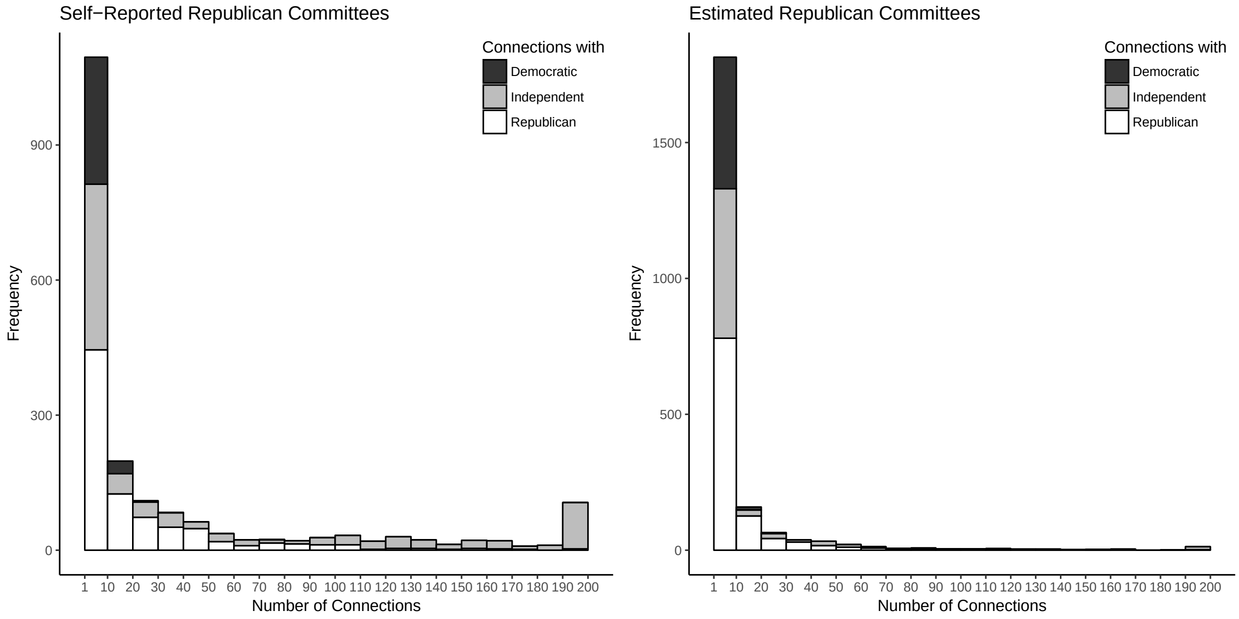


Figure 13: Self-reported vs. Estimated Republican PCs: Distributions of Number of Connections

Note: These figures only include observations with positive number of connections. Observations with more than 200 connections are plotted at 200.

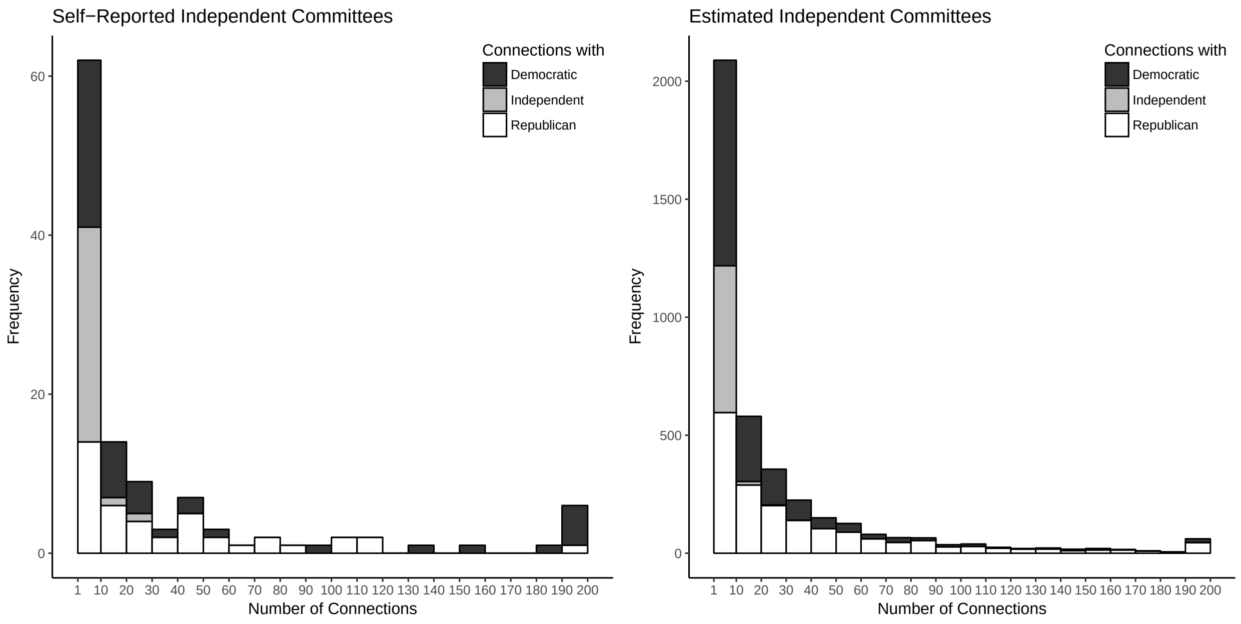


Figure 14: Self-reported vs. Estimated Independent PCs: Distributions of Number of Connections

Note: These figures only include observations with positive number of connections. Observations with more than 200 connections are plotted at 200.

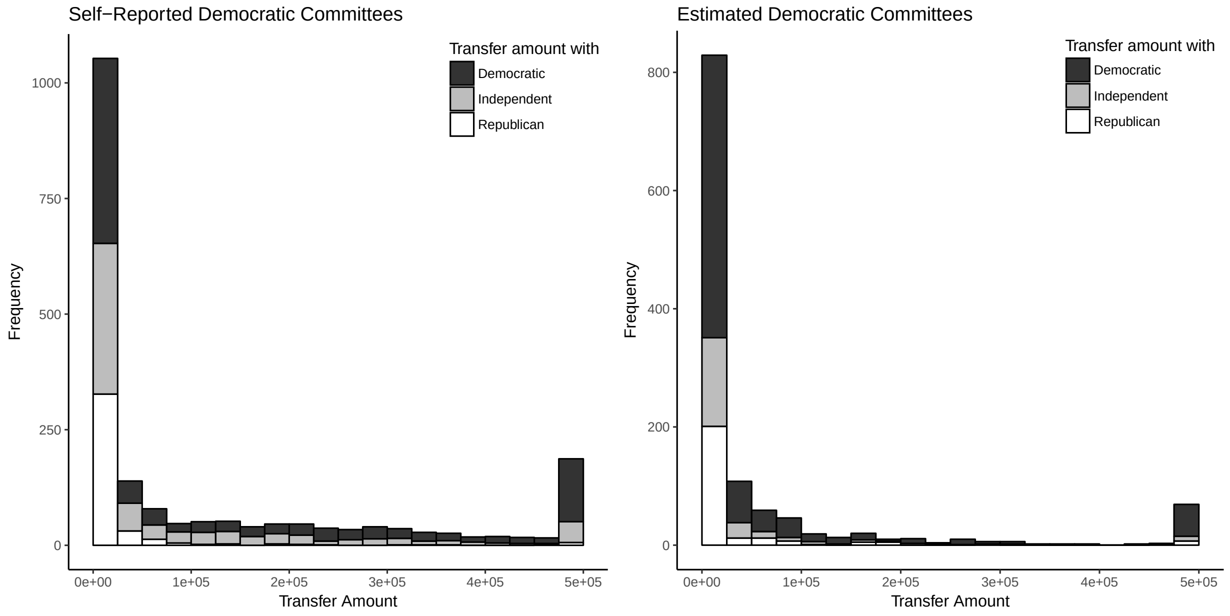


Figure 15: Self-reported vs. Estimated Democratic PCs: Distributions of Transfer Amount

Note: Bin size is 25,000. Observations with more than $500,000 transfer to and from one class of committees are plotted at $500,000.

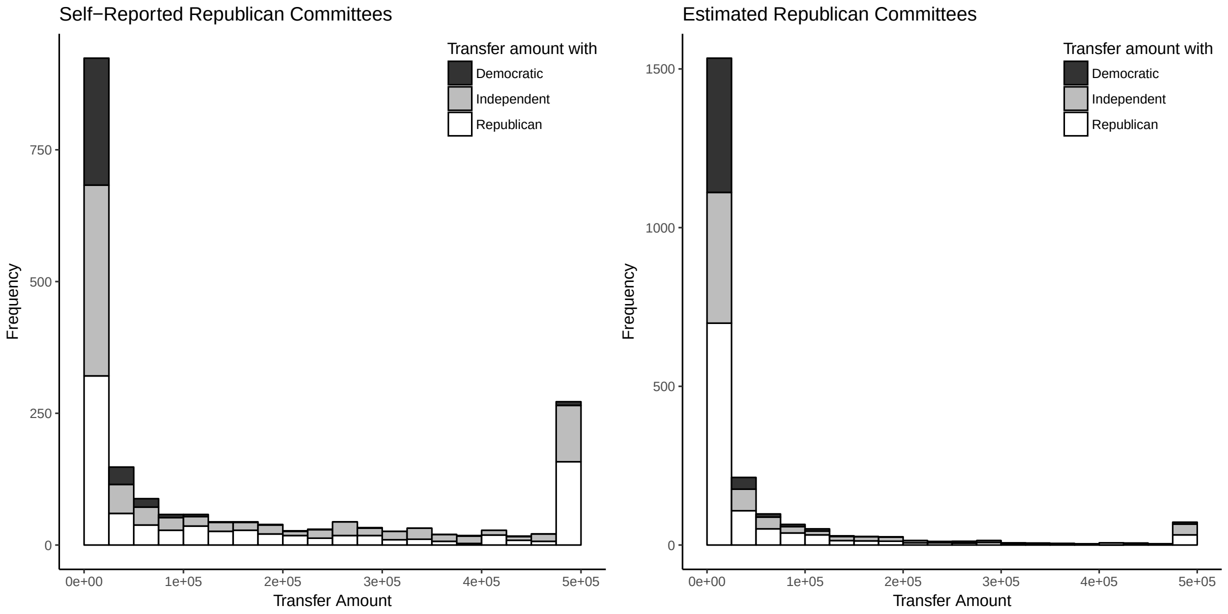


Figure 16: Self-reported vs. Estimated Republican PCs: Distributions of Transfer Amount

Note: Bin size is 25,000. Observations with more than $500,000 transfer to and from one class of committees are plotted at $500,000.

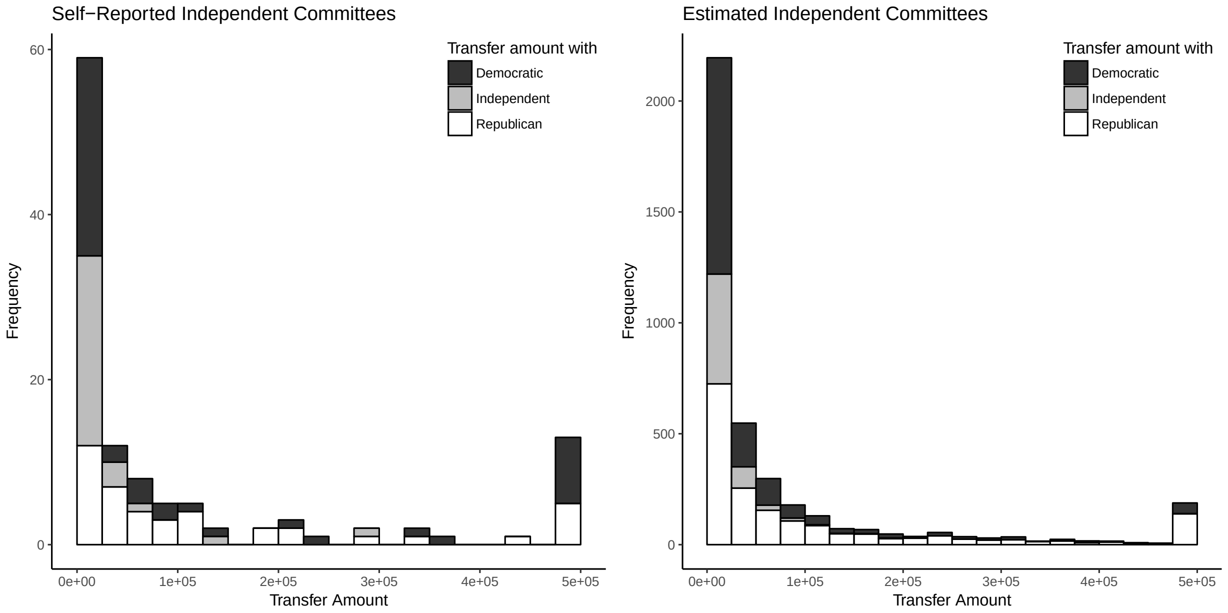


Figure 17: Self-reported vs. Estimated Independent PCs: Distributions of Transfer Amount

Note: Bin size is 25,000. Observations with more than $500,000 transfer to and from one class of committees are plotted at $500,000.

6.5 Estimation Results: Analyzing the Posterior Distribution of PCs' Political Ideologies

Now we analyze the posterior distribution of political ideology. Figure 18 plots, for PCs whose estimates are the same as the self reports (i.e., the reported ideology has the highest posterior probability), the distribution of the differences between the highest and the second highest posterior probability. These differences concentrate around 1, meaning the posterior probabilities concentrate on the self-reported affiliation. This confirms that we do not obtain these classifications by luck. We do a similar analysis in Figure 19, for PCs whose estimates differ from the self reports, by plotting the distributions of the differences between the highest posterior probability and the posterior probability of the reported ideology. These distributions have larger spread, but still a mass around 1. This indicates that these are not "near misses": our estimate strongly favors an ideology different from the PC's self-reported ideology.

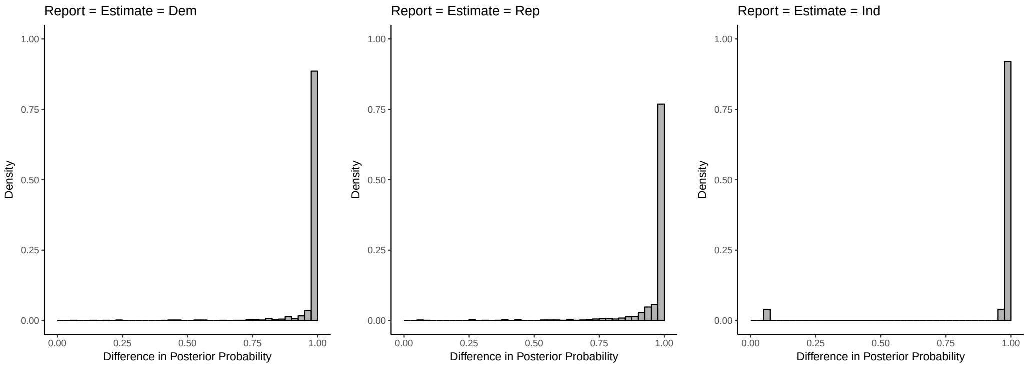


Figure 18: Distribution of Difference in the Posterior Probability of Ideology
Note: Horizontal axis is the difference between the highest posterior probability (i.e., the posterior probability of the self-reported ideology) and the second highest posterior probability. Bin size is 0.025.

Diagnostics of the Discrepancy. We further investigate the reason for the discrepancy between our estimates and the self reports by comparing PCs with different estimated ideology but the same self reported affiliation.

First of all, their financial conditions are different. Table 15 shows that self-reported Democratic PCs that are estimated as Republican are mostly PCs with high budget, and self-reported Democratic PCs that are estimated as Independent are mostly PCs with low budget. Similarly, among self-reported Republican PCs, the ones estimated to be Democratic or Independent are mostly PCs with lower budget.

Second, their contribution patterns are different. For each PC i, we define DemShare_i as its

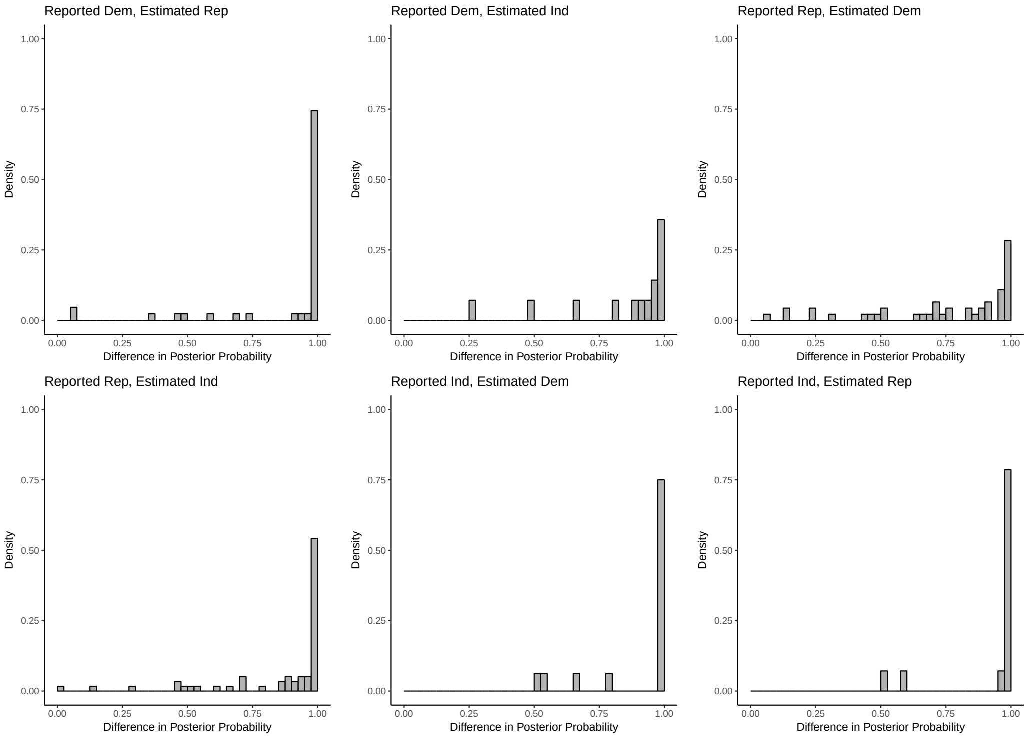


Figure 19: Distribution of Difference in the Posterior Probability of Ideology

Note: Horizontal axis is the difference between the highest posterior probability and the posterior probability of the self-reported ideology. Bin size is 0.025.

	Estimated Dem	Estimated Rep	Estimated Ind
Self-Reported Dem	32.26	206.52	3.50
Self-Reported Rep	10.00	77.96	6.77
Self-Reported Ind	31.00	3.00	46.00

Table 15: Median Budget (in $1,000)

share of connections with (estimated) Democratic PCs $\text{DemShare}_i = \frac{\text{numDem}_i}{\text{numDem}_i + \text{numRep}_i + \text{numInd}_i}$, and similarly for RepShare_i and IndShare_i. In Table 16, we compare the means of DemShare, RepShare, and IndShare for different groups of PCs. As a robustness check, we also calculate the shares in terms of transfer amount and present the results in the same table. The table supports our categorization of some PCs as Republican (Democratic) whose self-reports are Democratic (Republican) because they have significantly higher shares of connections with the Republican (Democratic) PCs, a pattern exhibited by all estimated Republican (Democratic) PCs.

	Number of Connection			Transfer Amount		
	Dem Share	Rep Share	Ind Share	Dem Share	Rep Share	Ind Share
Reported Dem, Estimated Rep	8.71%	19.65%	71.64%	28.36%	20.51%	51.13%
All PCs Estimated as Dem	72.58%	4.40%	23.02%	78.24%	4.15%	17.62%
Reported Rep, Estimated Dem	49.99%	13.44%	36.58%	54.56%	21.20%	24.24%
All PCs Estimated as Rep	6.69%	52.98%	40.33%	6.01%	55.77%	38.22%

Table 16: Mean of DemShare, RepShare, and IndShare

Next, we will discuss the discrepancy in more detail case by case. The discrepancies will fall into one of six cases, as shown in the panel label of Figure 20. Figure 20 depicts the distribution of the number of connections with PCs with different ideologies according to our estimates.

In the first case, some self-reported Democratic PCs are estimated to be Republican. In the data, these PCs have a small number of connections with PCs that self reported affiliations, most of which are Democratic. However, they are mostly connected with PCs without self reported affiliations, most of which are Independent by our estimates. As our estimate $\boldsymbol{\beta}$ suggests, Republican PCs have the highest connection probability with Independent PCs. Therefore, although they self reported to be Democratic and are connected with few self-reported Republican or Independent PCs, their overall contribution patterns are close to that of the Republican PCs. In the degree distribution presented in Figure 20 for this case, it can be seen that the tail distribution is very close to that of the self-reported Republican PCs in Figure 13.

In the second case, some self-reported Democratic PCs are estimated as Independent. Only one of these PCs has one connection with another Independent PC, and the rest have no connection with Independent PCs. According to estimate $\boldsymbol{\beta}$, classifying them as Independent, rather than Democratic, better rationalize this pattern because Independent-Independent connection probability is lower than Democratic-Independent connection probability.

In the third case, some self-reported Republican PCs are estimated as Democratic. These PCs have similar number of connections to all three classes of PCs, with slightly more Democratic connections. Their number of connections to the Republican and the Independent PCs are not jointly high enough to be estimated as Republican.

In the fourth case, some self-reported Republican PCs are estimated as Independent. These PCs have more connections with Republican than Democratic PCs, but not large enough connections with Independent PCs to be estimated as Republican. In other words, they do not exhibit the heavy tail on connection with Independent PCs which is observed for other self-reported Republican PCs.

In the fifth case, some self-reported Independent PCs are estimated as Democratic. These PCs' numbers of connections with Democratic PCs significantly outweigh that with the Republican and Independent PCs. There are not large enough connections with Republican PCs to be estimated as Independent.

In the sixth case, some self-reported Independent PCs are estimated as Republican. These PCs do not have large enough connections with Democratic PCs to be estimated as Independent.

Discussion. Here we present and analyze the discrepancy between our estimate and the self-report, in an attempt to better understand the implications of our model and algorithm. We are not making a claim that these PCs strategically misreported their party affiliations. We show that under our model, their contribution patterns to other PCs are different from the majority of the PCs self reporting the same affiliation as they did. Simplifications in our model are also potential reasons for the discrepancy. For example, our model does not capture all aspects of the incentives in political contribution between PCs. Additionally, we only study the contributions among PCs, and do leave out other campaign activities such as collection of individual contribution and independent expenditures. In Appendix G we list the PCs that self reported to be Democratic (Republican), but are estimated to be Republican (Democratic).

7 Conclusion

About two thirds of the political committees registered with the Federal Election Commission do not self identify their party affiliations. In this paper we propose and implement a novel Bayesian approach to infer about the ideological affiliations of political committees based on the network of the financial contributions among them. In Monte Carlo simulations, we demonstrate that our estimation algorithm achieves very high accuracy in recovering their latent ideological affiliations when the pairwise difference in ideology groups' connection patterns satisfy a condition known as the Chernoff-Hellinger divergence criterion. We illustrate our approach using the campaign finance record in 2003-2004 election cycle. Using the posterior mode to categorize the ideological affiliations of the political committees, our estimates match the self reported ideology for 94.36% of those committees who self reported to be Democratic and 89.49% of those committees who self

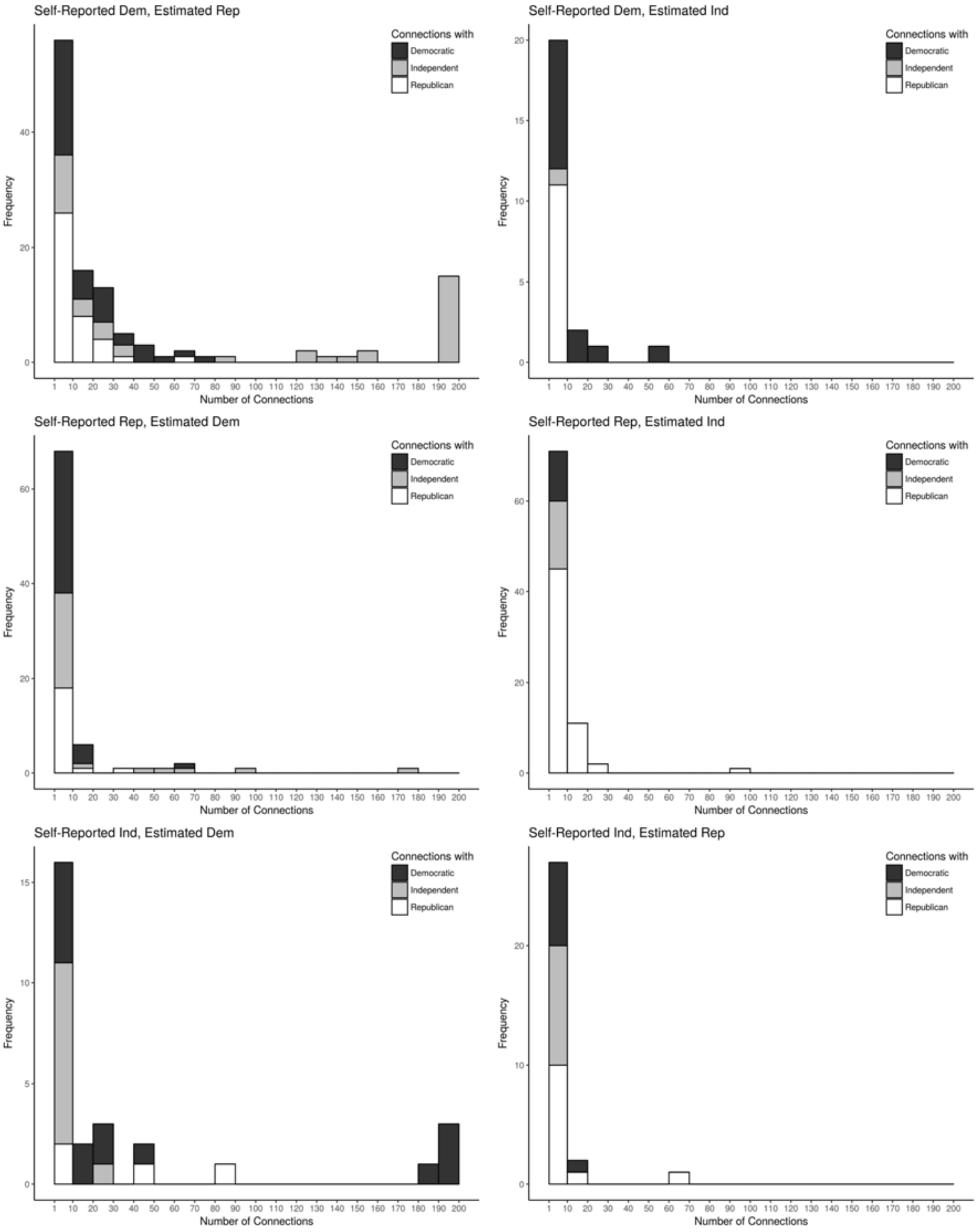


Figure 20: Distributions of Number of Connections with Different Committees

Note: Observations with more than 200 connections are plotted at 200.

reported to be Republican.

Since PCs are required to report to the FEC the financial contributions among each other, our proposed methods to infer the ideological affiliations of political committees via financial contributions network can be implemented readily. To the extent that the estimated ideologies for the PCs are close to their true latent ideologies, our estimated PC ideology can fill the missing "ideologies" problem for researchers who are interested in studying *individuals'* political contribution patterns using FEC's "Contributions by Individuals" data. Moreover, since our estimation methods can be implemented using data from only one election cycle, we can estimate the ideological affiliations of the same PCs using data from different election cycles. This would allow us to study the possible evolutions of ideological affiliations of PCs over time. We can also exploit the permanent presence of the national committees such as Democratic National Committee (DNC) and Republican National Committee (RNC) and use their network links as a vehicle to study the possible changes of party platforms that are not necessarily reflected in official documents. These are exciting areas for future research.

References

Abbe, Emmanuel. 2017. "Community Detection And Stochastic Block Models: Recent Developments." *arXiv preprint arXiv:1703.10146.*

Abbe, Emmanuel, Afonso S Bandeira, and Georgina Hall. 2014. "Exact Recovery In The Stochastic Block Model." *arXiv preprint arXiv:1405.3267.*

Abbe, Emmanuel, and Colin Sandon. 2015. "Community Detection In General Stochastic Block Models: Fundamental Limits And Efficient Algorithms For Recovery." 670–688, IEEE.

Albert, James H, and Siddhartha Chib. 1993. "Bayesian Analysis Of Binary And Polychotomous Response Data." *Journal of the American statistical Association*, 88(422): 669–679.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis*, 23(1): 76–91.

Bonica, Adam. 2013. "Ideology And Interests In The Political Marketplace." *American Journal of Political Science*, 57(2): 294–311.

Breiger, Ronald L, Scott A Boorman, and Phipps Arabie. 1975. "An Algorithm For Clustering Relational Data With Applications To Social Network Analysis And Comparison With Multidimensional Scaling." *Journal of mathematical psychology*, 12(3): 328–383.

Cai, T Tony, Tengyuan Liang, and Alexander Rakhlin. 2017. "Weighted Message Passing and Minimum Energy Flow for Heterogeneous Stochastic Block Models with Side Information." *arXiv preprint arXiv:1709.03907.*

Chandrasekhar, Arun, and Randall Lewis. 2011. "Econometrics Of Sampled Networks." *Unpublished manuscript, MIT.*

Jackson, Matthew O. 2010. "An Overview Of Social Networks And Economic Applications." *The handbook of social economics*, 1: 511–85.

Jog, Varun, and Po-Ling Loh. 2015. "Information-Theoretic Bounds For Exact Recovery In Weighted Stochastic Block Models Using The Renyi Divergence." *arXiv preprint arXiv:1509.06418.*

Kanade, Varun, Elchanan Mossel, and Tselil Schramm. 2016. "Global And Local Information In Clustering Labeled Block Models." *IEEE Transactions on Information Theory*, 62(10): 5906–5917.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles And Techniques.* MIT press.

Leung, Michael. 2016. "A Weak Law for Moments of Pairwise-Stable Networks." *Available at SSRN: https://ssrn.com/abstract=2663685.*

McCarty, Nolan, Keith T Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance Of Ideology And Unequal Riches.* MIT Press.

McCarty, Nolan M, and Keith T Poole. 1998. "An Empirical Spatial Model Of Congressional Campaigns." *Political Analysis*, 7: 1–30.

McKay, Amy. 2008. "A Simple Way Of Estimating Interest Group Ideology." *Public Choice*, 136(1): 69–86.

McKay, Amy. 2010. "The Effects Of Interest Groups' Ideology On Their Pac And Lobbying Expenditures." *Business and Politics*, 12(2): 1–21.

Mossel, Elchanan, Joe Neeman, and Allan Sly. 2014. "Consistency Thresholds For Binary Symmetric Block Models." *arXiv preprint arXiv:1407.1591.*

Newman, Mark EJ. 2006. "Modularity And Community Structure In Networks." *Proceedings of the national academy of sciences*, 103(23): 8577–8582.

Newman, Mark EJ. 2013. "Spectral Methods For Community Detection And Graph Partitioning." *Physical Review E*, 88(4): 042822.

Nowicki, Krzysztof, and Tom A B Snijders. 2001. "Estimation And Prediction For Stochastic Blockstructures." *Journal of the American Statistical Association*, 96(455): 1077–1087.

Peng, Lijun, and Luis Carvalho. 2015*a*. "Bayesian Degree-Corrected Stochastic Block Models for Community Detection." *arXiv preprint arXiv:1309.4796.*

Peng, Lijun, and Luis E Carvalho. 2015*b*. "Group-Corrected Stochastic Blockmodels for Community Detection on Large-scale Networks."

Poole, Keith T, and Howard Rosenthal. 1985. "A Spatial Model For Legislative Roll Call Analysis." *American Journal of Political Science*, 357–384.

Snijders, Tom AB, and Krzysztof Nowicki. 1997. "Estimation And Prediction For Stochastic Blockmodels For Graphs With Latent Block Structure." *Journal of Classification*, 14(1): 75–100.

Stoer, Mechthild, and Frank Wagner. 1997. "A Simple Min-Cut Algorithm." *Journal of the ACM (JACM)*, 44(4): 585–591.

Trebbi, Francesco, and Eric Weese. 2015. "Insurgency and Small Wars: Estimation of Unobserved Coalition Structures." *NBER Working Paper*, , (w21202).

Yun, Se-Young, and Alexandre Proutiere. 2016. "Optimal Cluster Recovery In The Labeled Stochastic Block Model." *in Advances in Neural Information Processing Systems*, 29: 965–973.

Appendix A Construction of Variable *industry**

We construct PCs' characteristic "*industry**" using the industrial breakdown information from OpenSecrets.org. Their breakdown has three nested levels - from coarse to fine - "sector", "industry", and "category". [26] We define variable *industry** to be the sector for PCs in sectors that are relatively homogeneous such as agricultural business. We define *industry** to be a finer level, industry or category, for PCs in sectors that are relatively heterogeneous such as miscellaneous business. The reason is that we use interaction term "same *industry**" in our analysis; and we want, within each *industry**, similar level of heterogeneity. A detailed description of the construction of the variable *industry** is given below, and the corresponding codebook is given in Table A1.

1. *industry**=sector if:

   A PC belongs to one of the following sectors: Agribusiness, Communications/Electronics, Construction, Defense, Energy & Natural Resources, Finance, Insurance & Real Estate, Health, Labor, Lawyers & Lobbyists, Transportation.

2. *industry**=industry if:

   (a) A PC belongs to the Miscellaneous Business sector.

   (b) A PC belongs to the Other sector, but not in the Other industry.

   (c) A PC belongs to the Ideological/Single-Issue sector, but not in one of the following industries: Misc Issues, Republican/Conservative, Democratic/Liberal, Leadership PACs, Candidate Committee.

3. *industry**=category if:

   (a) A PC belongs to the Other industry in the Other sector, but not in the Other category.

   (b) A PC belongs to the Misc Issues industry in the Ideological/Single-Issue sector.

4. *industry**=NA if:[27]

   (a) A PC belongs to one of the following sectors: Unknown, Joint Candidate Committee, Party Committee, Candidate, Non-contribution.

   (b) A PC belongs to one of the following industries in the Ideological/Single-Issue sector: Republican/Conservative, Democratic/Liberal, Leadership PACs, Candidate Committee.

   (c) A PC's sector, industry and category are all Other.

[26] The codebook is available at `https://www.opensecrets.org/downloads/crp/CRP_Categories.txt`.

[27] We define the interaction term "same *industry**" to be 0 if one or both PCs' *industry**s are NAs.

Code	Description
A	Agribusiness
B	Communic/Electronics
C	Construction
D	Defense
E	Energy/Nat Resource
F	Finance/Insur/RealEst
H	Health
H6000	Welfare & Social Work
J1300	Third-party committees
J3000	Consumer groups
J4000	Fiscal & tax policy
J7200	Elderly issues/Social Security
J7600	Animal Rights
J8000	Labor, anti-union
J9000	Other single-issue or ideological groups
K	Lawyers & Lobbyists
M	Transportation
N00	Business Associations
N01	Food & Beverage
N02	Beer, Wine & Liquor
N03	Retail Sales
N04	Misc Services
N05	Business Services
N06	Recreation/Live Entertainment
N07	Casinos/Gambling
N08	Lodging/Tourism
N12	Misc Business
N13	Chemical & Related Manufacturing
N14	Steel Production
N15	Misc Manufacturing & Distributing
N16	Textiles
NA	Not Available
P	Labor
Q04	Foreign & Defense Policy
Q05	Pro-Israel
Q08	Women's Issues
Q09	Human Rights
Q11	Environment
Q12	Gun Control
Q13	Gun Rights
Q14	Abortion Policy/Anti-Abortion
Q15	Abortion Policy/Pro-Abortion Rights
W02	Non-Profit Institutions
W03	Civil Servants/Public Officials
W04	Education
X5000	Military

Table A1: Codebook for Variable *industry**

Appendix B Contribution Limit

Contribution limits for the 2003-2004 Election Cycle are given in Table B1.[28] They differ by the identity of the contributor and that of the recipient. As a contributor, individual, national party committee, non-national party committee, multicandidate PAC, and non-multicandidate PAC have different contribution limits. As a recipient, candidate or candidate committee, national party committee, non-national party committee, and other PC have different contribution limits. As shown in the table, there is no limit for contribution between party committees, national or local. For the rest of the contributions between PCs, the limits range from \$2,000 to \$25,000, most of which are set at \$5,000.

[28] Source: `http://www.fec.gov/pages/brochures/ContributionLimits2003-2004.htm`.

Contributor	Recipient				
	Candidate or candidate committee per election	National party committee per calendar year	State, district and local party committee per calendar year	Any other PC[1] per calendar year	Special Limits
Individual	$2,000	$25,000	$10000 (combined limit[6])	$5,000	$95,000 overall biennial limit: $37,500 to all candidates, $57,500 to all PACs and parties.[2]
National Party Committee	$5,000	No limit	No limit	$5,000	$35,000 to Senate candidate per campaign[3]
State, District & Local Party Committee	$5,000 (combined limit)	No limit	No limit	$5,000 (combined limit)	No limit
PAC (multicandidate[4])	$5,000	$15,000	$5,000 (combined limit)	$5,000	No limit
PAC (not multicandidate)	$2,000[5]	$25,000	$10,000 (combined limit)	$5,000	No limit

[1] A contribution earmarked for a candidate through a political committee counts against the original contributor's limit for that candidate. In certain circumstances, the contribution may also count against the contributors limit to the PAC. 11 CFR 110.6. See also 11 CFR 110.1(h).

[2] No more than $37,500 of this amount may be contributed to state and local party committees and PACs.

[3] This limit is shared by the national committee and the Senate campaign committee.

[4] A multicandidate committee is a political committee with more than 50 contributors which has been registered for at least 6 months and, with the exception of state party committees, has made contributions to 5 or more candidates for federal office. 11 CFR 100.5(e)(3).

[5] A federal candidate's authorized committee(s) may contribute no more than $1,000 per election to another federal candidate's authorized committee(s). 11 CFR 102.12(c)(2).

[6] Combined limits are shared by federal accounts of all other state and local committees of the same party in the same state.

Table B1: Contribution Limits for the 2003-3004 Election Cycle

Appendix C Figures and Tables: Monte Carlo I

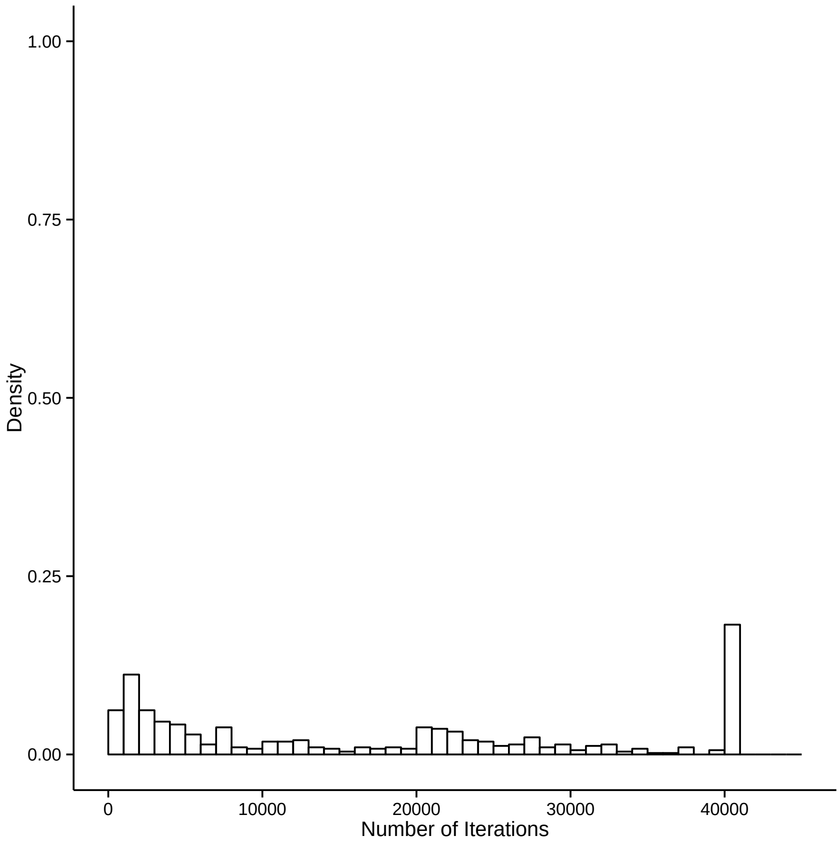


Figure C1: Histogram of Number of Iterations

Note: Bin size is 1000.

Average	Standard Deviation	Minimum	Maximum
0.011020	0.011637	0.000000	0.080000

Table C1: Summary Statistics on Misclassification Rates

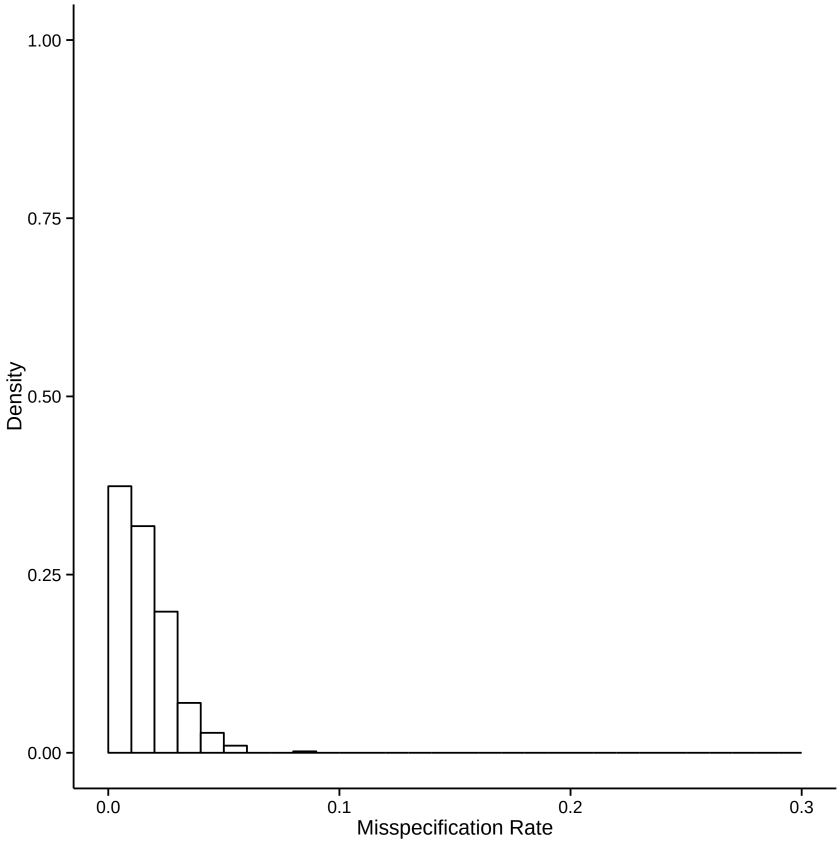


Figure C2: Histogram of Misclassification Rate

Note: Bin size is 0.01.

	Estimated Dem	Estimated Rep	Estimated Ind
True Dem	32.7520%	0.1820%	0.1540%
True Rep	0.1700%	32.7440%	0.2000%
True Ind	0.1840%	0.2120%	33.4020%

Table C2: Tabulation of Estimated vs. True Ideology

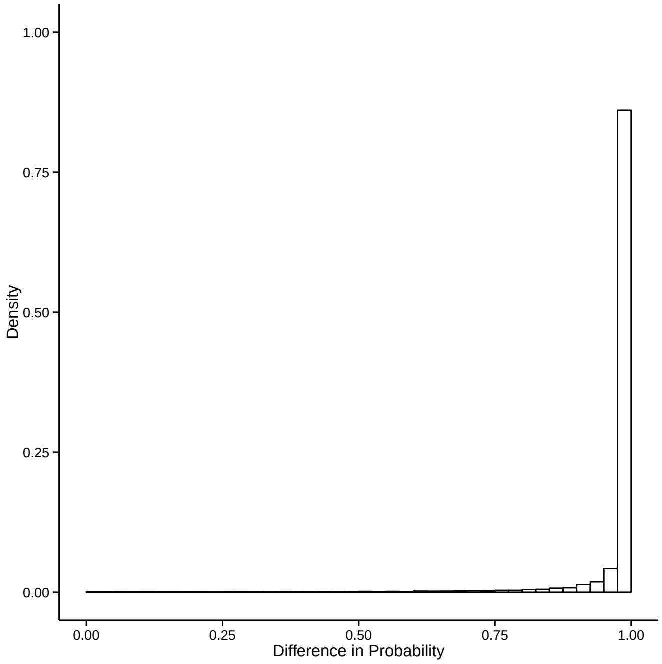


Figure C3: Histogram of Difference in Posterior Probability for Correctly Classified Vertices

Note: Each observation is a vertex. Horizontal axis is the difference between the highest posterior probability(i.e., the posterior probability of the true ideology) and the second highest posterior probability. Bin size is 0.025.

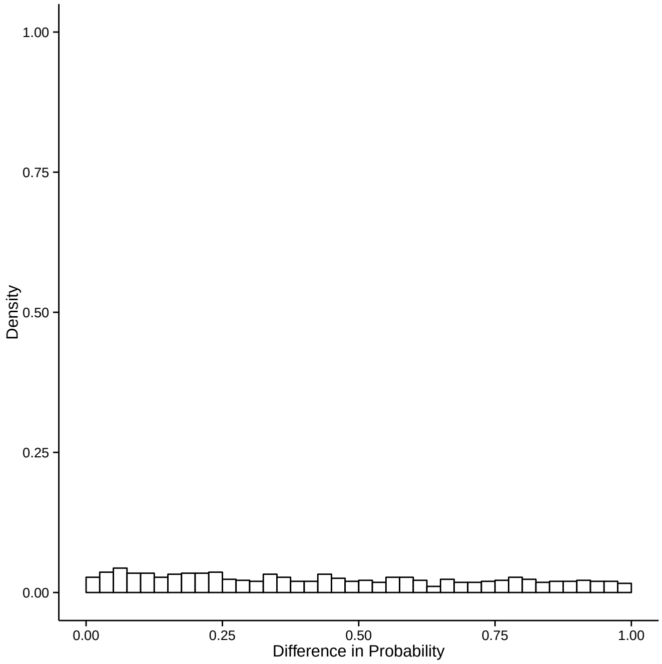


Figure C4: Histogram of Difference in Posterior Probability for Misclassified Vertices

Note: Each observation is a vertex. Horizontal axis is the difference between the highest posterior probability and the posterior probability of the true ideology. Bin size is 0.025.

$\boldsymbol{\beta}$	Bias	Standard Deviation
constant	-0.020032	0.306216
$\mathbb{1}_{x_i=x_j=\text{Dem}}$	0.014907	0.075353
$\mathbb{1}_{x_i=x_j=\text{Rep}}$	0.008465	0.075682
$\mathbb{1}_{x_i=x_j=\text{Ind}}$	0.017410	0.078703
$\mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})}$	0.001838	0.064869
$\mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})}$	-0.001370	0.064845
Same state	0.001803	0.043134
Same industry	0.002705	0.044662
One of them is a House campaign	-0.001016	0.046084
One of them is a Senate campaign	-0.004365	0.046909
One of them is a Presidential campaign	-0.001202	0.046442
One of them is a qualified PAC	0.000268	0.044655
One of them is a qualified Party	0.001389	0.042022
One of them is a national committee	0.001199	0.050153
One of them is authorized by a candidate	0.001038	0.046411
One of them is a joint fundraiser	-0.001212	0.051018
$(\ln b_i + \ln b_j)$	0.004666	0.103812
$((\ln b_i)^2 + (\ln b_j)^2)$	-0.000569	0.013868
$\ln b_i \ln b_j$	-0.000553	0.025041

Table C3: Parameters Governing Edge Formation Probabilities β

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\boldsymbol{\theta}$	Bias	Standard Deviation
$\mathbb{P}(\text{Dem})$	-0.000223	0.004705
$\mathbb{P}(\text{Rep})$	-0.000224	0.004567
$\mathbb{P}(\text{Ind})$	0.000447	0.004620

Table C4: Parameters Governing the Fraction of Ideologies θ

Notes: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\mathbf{h}_{\text{Dem,Dem}}$	Bias	Standard Deviation
	0.136574	0.014174
	0.101386	0.011487
	-0.102415	0.014947
	-0.135545	0.017527

$\mathbf{h}_{\text{Rep,Rep}}$	Bias	Standard Deviation
	0.136643	0.014089
	0.102114	0.012182
	-0.102477	0.014942
	-0.136280	0.017799

$\mathbf{h}_{\text{Ind,Ind}}$	Bias	Standard Deviation
	0.134438	0.014390
	0.100568	0.011735
	-0.101278	0.014835
	-0.133727	0.017879

$\mathbf{h}_{\text{Dem,Rep}}$	Bias	Standard Deviation
	-0.097963	0.013732
	-0.033662	0.011047
	0.032859	0.009625
	0.098766	0.010588

$\mathbf{h}_{\text{Dem, Ind}}$	Bias	Standard Deviation
	-0.096669	0.013237
	-0.032747	0.011223
	0.032331	0.009789
	0.097085	0.009931

$\mathbf{h}_{\text{Rep, Ind}}$	Bias	Standard Deviation
	-0.097534	0.013724
	-0.032846	0.011521
	0.032789	0.009901
	0.097592	0.010524

Table C5: Parameters Governing the Weight Distribution h

Notes: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

ϵ	Bias	Standard Deviation
	-0.002427	0.000660

Table C6: Parameter Governing Measurement Error ϵ

Notes: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

Appendix D   Figures and Tables: Monte Carlo II

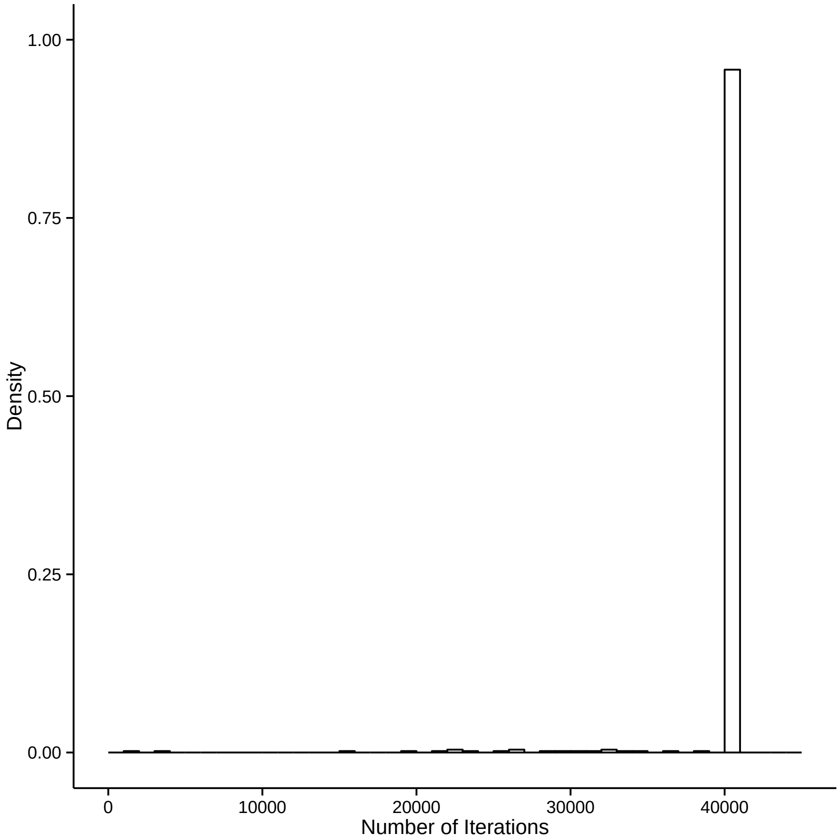


Figure D1: Histogram of Number of Iterations

Note: Bin size is 1000.

Average	Standard Deviation	Minimum	Maximum
0.056800	0.033604	0.000000	0.240000

Table D1: Summary Statistics on Misclassification Rates

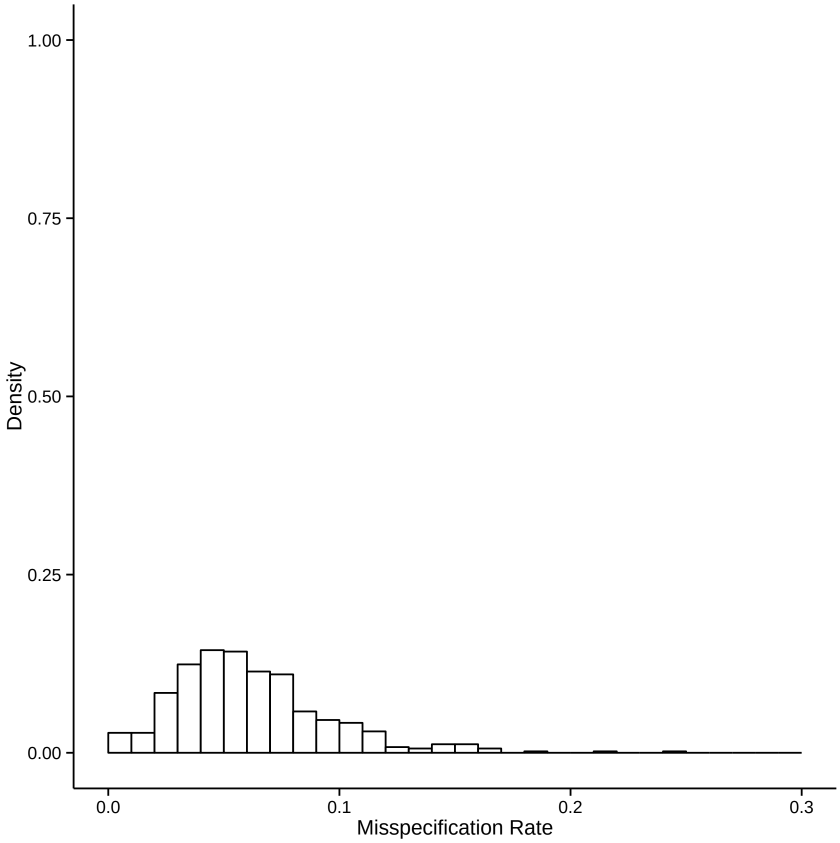


Figure D2: Histogram of Misclassification Rate

Note: Bin size is 0.01.

	Estimated Dem	Estimated Rep	Estimated Ind
True Dem	31.5080%	0.9440%	0.8760%
True Rep	1.0740%	31.3140%	0.8820%
True Ind	0.8940%	1.0100%	31.4980%

Table D2: Tabulation of Estimated vs. True Ideology

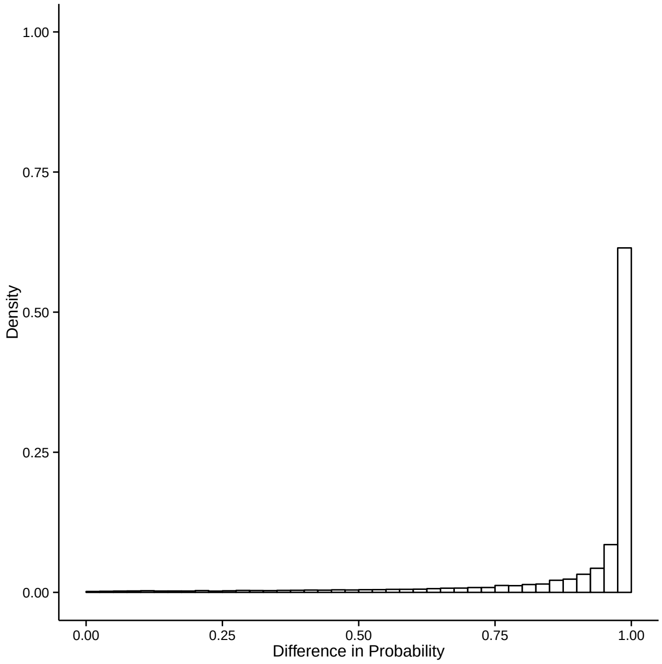


Figure D3: Histogram of Difference in Posterior Probability for Correctly Classified Vertices
Note: Each observation is a vertex. Horizontal axis is the difference between the highest posterior probability(i.e., the posterior probability of the true ideology) and the second highest posterior probability. Bin size is 0.025.

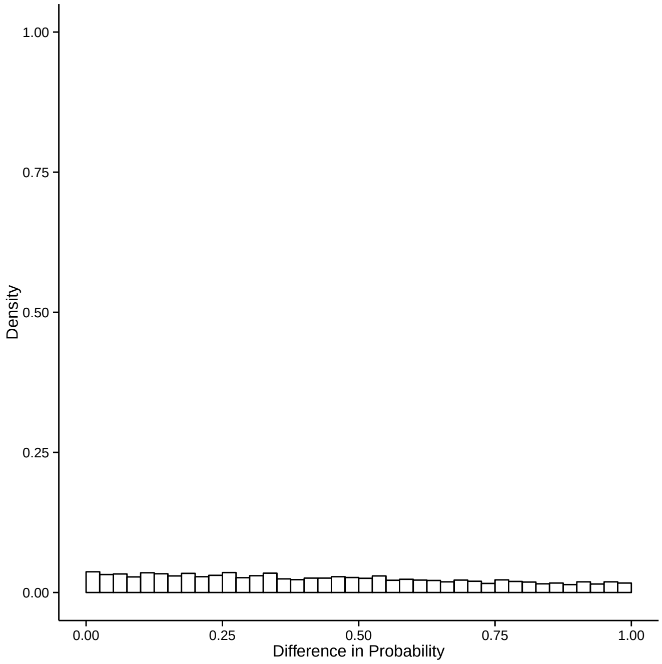


Figure D4: Histogram of Difference in Posterior Probability for Misclassified Vertices
Note: Each observation is a vertex. Horizontal axis is the difference between the highest posterior probability and the posterior probability of the true ideology. Bin size is 0.025.

$\boldsymbol{\beta}$	Bias	Standard Deviation
constant	-0.053234	0.278777
$\mathbb{1}_{x_i=x_j=\text{Dem}}$	0.004676	0.088396
$\mathbb{1}_{x_i=x_j=\text{Rep}}$	0.005889	0.085760
$\mathbb{1}_{x_i=x_j=\text{Ind}}$	0.008628	0.090062
$\mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})}$	-0.005467	0.069264
$\mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})}$	0.000659	0.071066
Same state	0.004227	0.043947
Same industry	0.002177	0.043028
One of them is a House campaign	0.001558	0.046981
One of them is a Senate campaign	0.000138	0.051027
One of them is a Presidential campaign	0.001651	0.046874
One of them is a qualified PAC	0.002929	0.048328
One of them is a qualified Party	0.002520	0.043062
One of them is a national committee	0.001504	0.048683
One of them is authorized by a candidate	0.005082	0.047008
One of them is a joint fundraiser	0.007028	0.049847
$(\ln b_i + \ln b_j)$	0.013339	0.095571
$((\ln b_i)^2 + (\ln b_j)^2)$	-0.001264	0.013323
$\ln b_i \ln b_j$	-0.002407	0.023131

Table D3: Parameters Governing Edge Formation Probabilities β

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\boldsymbol{\theta}$	Bias	Standard Deviation
$\mathbb{P}$(Dem)	0.000043	0.004747
$\mathbb{P}$(Rep)	-0.000114	0.004533
$\mathbb{P}$(Ind)	0.000071	0.004761

Table D4: Parameters Governing the Fraction of Ideologies θ

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\mathbf{h}_{\text{Dem,Dem}}$	Bias	Standard Deviation
	0.039407	0.010862
	0.070890	0.011031
	-0.070779	0.012694
	-0.039519	0.011386

$\mathbf{h}_{\text{Rep,Rep}}$	Bias	Standard Deviation
	0.038909	0.010678
	0.072241	0.011493
	-0.071994	0.012912
	-0.039157	0.012040

$\mathbf{h}_{\text{Ind,Ind}}$	Bias	Standard Deviation
	0.039131	0.011047
	0.071208	0.011610
	-0.070468	0.013130
	-0.039871	0.012104

$\mathbf{h}_{\text{Dem,Rep}}$	Bias	Standard Deviation
	-0.100952	0.014387
	-0.036391	0.011860
	0.036400	0.011109
	0.100943	0.010614

$\mathbf{h}_{\text{Dem,Ind}}$	Bias	Standard Deviation
	-0.100720	0.013911
	-0.035439	0.011370
	0.035755	0.009990
	0.100404	0.010030

$\mathbf{h}_{\text{Rep,Ind}}$	Bias	Standard Deviation
	-0.101129	0.014001
	-0.036941	0.011595
	0.036431	0.010460
	0.101638	0.010507

Table D5: Parameters Governing the Weight Distribution h

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

ϵ	Bias	Standard Deviation
	-0.002484	0.000581

Table D6: Parameter Governing Measurement Error ϵ

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

Appendix E Figures and Tables: Monte Carlo III

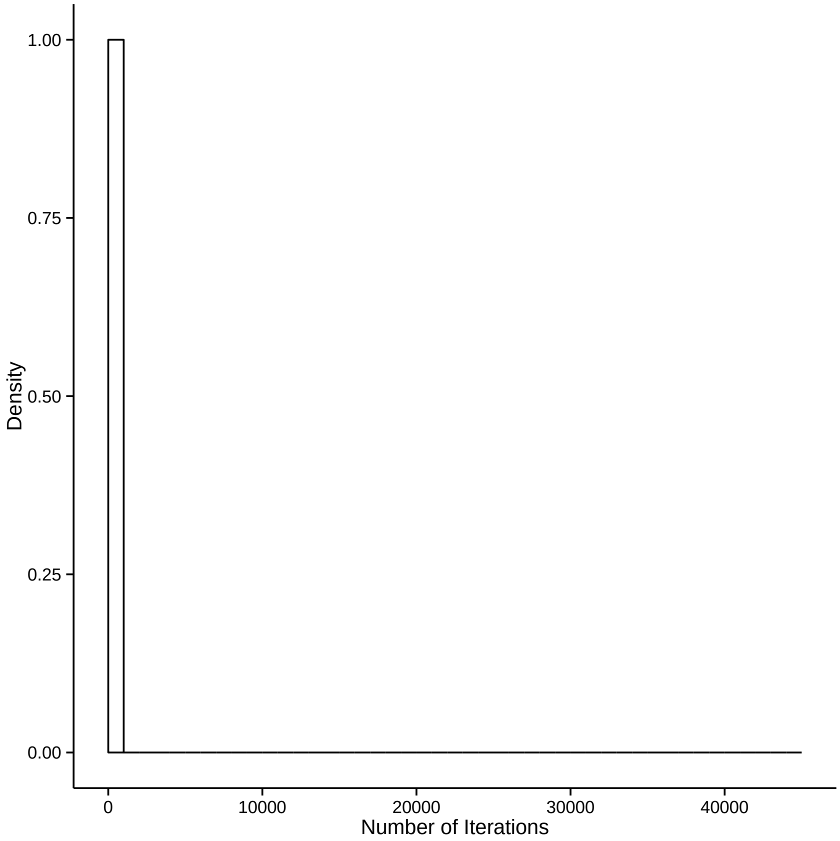


Figure E1: Histogram of Number of Iterations

Note: Bin size is 1000.

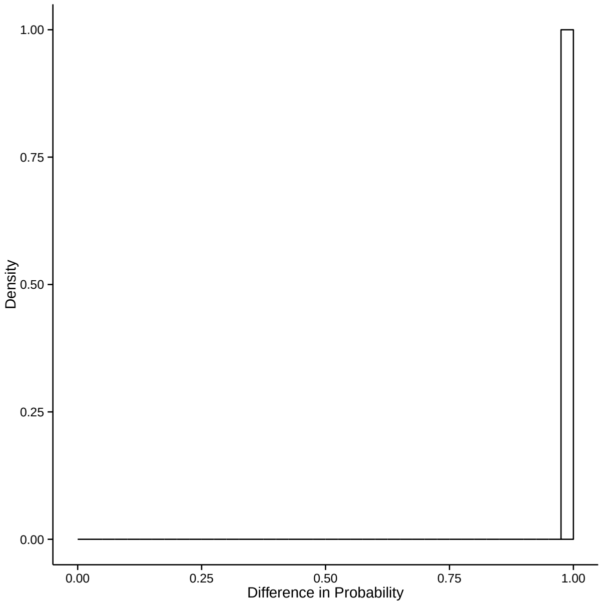


Figure E2: Histogram of Difference in Posterior Probability for Correctly Classified Vertices

Note: Each observation is a vertex. Horizontal axis is the difference between the highest posterior probability(i.e., the posterior probability of the true ideology) and the second highest posterior probability. Bin size is 0.025.

$\boldsymbol{\beta}$	Bias	Standard Deviation
constant	-0.000600	0.045500
$\mathbb{1}_{x_i=x_j=\text{Dem}}$	0.000375	0.013841
$\mathbb{1}_{x_i=x_j=\text{Rep}}$	0.001279	0.014102
$\mathbb{1}_{x_i=x_j=\text{Ind}}$	0.000725	0.014023
$\mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})}$	0.001145	0.012192
$\mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})}$	0.000070	0.011883
Same state	-0.000222	0.008429
Same industry	-0.000045	0.008724
One of them is a House campaign	-0.000023	0.009036
One of them is a Senate campaign	-0.000173	0.009085
One of them is a Presidential campaign	-0.000214	0.009082
One of them is a qualified PAC	-0.000737	0.008557
One of them is a qualified Party	-0.000620	0.008545
One of them is a national committee	0.000057	0.009115
One of them is authorized by a candidate	-0.000122	0.009020
One of them is a joint fundraiser	0.000534	0.009923
$(\ln b_i + \ln b_j)$	0.000375	0.015589
$((\ln b_i)^2 + (\ln b_j)^2)$	-0.000113	0.002002
$\ln b_i \ln b_j$	0.000054	0.004421

Table E1: Parameters Governing Edge Formation Probabilities β

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\boldsymbol{\theta}$	Bias	Standard Deviation
$\mathbb{P}$(Dem)	0.000469	0.007948
$\mathbb{P}$(Rep)	-0.000514	0.007469
$\mathbb{P}$(Ind)	0.000046	0.007534

Table E2: Parameters Governing the Fraction of Ideologies θ

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\mathbf{h}_{\text{Dem,Dem}}$	Bias	Standard Deviation
	0.003591	0.005047
	0.007005	0.004737
	-0.006713	0.006200
	-0.003883	0.006027

$\mathbf{h}_{\text{Rep,Rep}}$	Bias	Standard Deviation
	0.003633	0.005297
	0.007228	0.004696
	-0.007282	0.006413
	-0.003580	0.005937

$\mathbf{h}_{\text{Ind,Ind}}$	Bias	Standard Deviation
	0.003769	0.005023
	0.007100	0.004713
	-0.007326	0.005899
	-0.003543	0.005966

$\mathbf{h}_{\text{Dem, Rep}}$	Bias	Standard Deviation
	-0.009173	0.006111
	-0.003547	0.005576
	0.002779	0.004879
	0.009941	0.003743

$\mathbf{h}_{\text{Dem,Ind}}$	Bias	Standard Deviation
	-0.009563	0.005808
	-0.003075	0.005394
	0.003214	0.004920
	0.009424	0.003579

$\mathbf{h}_{\text{Rep, Ind}}$	Bias	Standard Deviation
	-0.009567	0.005856
	-0.003141	0.005524
	0.003000	0.004760
	0.009708	0.003814

Table E3: Parameters Governing the Weight Distribution $\mathbf{h}$

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

ϵ	Bias	Standard Deviation
	-0.002183	0.001299

Table E4: Parameter Governing Measurement Error ϵ

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

Appendix F Figures and Tables: Monte Carlo IV

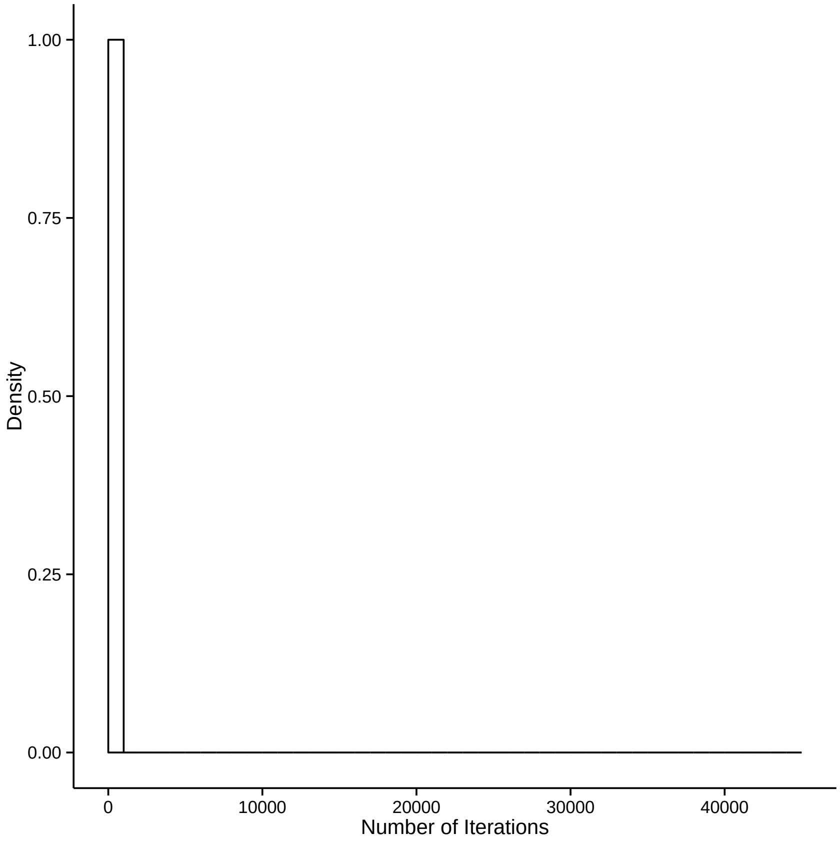


Figure F1: Histogram of Number of Iterations

Note: Bin size is 1000.

$\boldsymbol{\beta}$	Bias	Standard Deviation
constant	-0.000099	0.003286
$\mathbb{1}_{x_i=x_j=\text{Dem}}$	0.000333	0.002225
$\mathbb{1}_{x_i=x_j=\text{Rep}}$	0.000288	0.002065
$\mathbb{1}_{x_i=x_j=\text{Ind}}$	0.000045	0.002128
$\mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})}$	0.000012	0.002087
$\mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})}$	0.000441	0.002084
Same state	0.000135	0.001168
Same industry	-0.000033	0.001090
One of them is a House campaign	-0.000167	0.001289
One of them is a Senate campaign	-0.000063	0.001331
One of them is a Presidential campaign	-0.000026	0.001187
One of them is a qualified PAC	0.000091	0.001104
One of them is a qualified Party	0.000060	0.001147
One of them is a national committee	0.000015	0.001267
One of them is authorized by a candidate	-0.000166	0.001376
One of them is a joint fundraiser	0.000039	0.001281
$(\ln b_i + \ln b_j)$	-0.000084	0.000527
$((\ln b_i)^2 + (\ln b_j)^2)$	0.000005	0.000216
$\ln b_i \ln b_j$	0.000030	0.000583

Table F1: Parameters Governing Edge Formation Probabilities β

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\boldsymbol{\theta}$	Bias	Standard Deviation
$\mathbb{P}(\text{Dem})$	-0.000759	0.006007
$\mathbb{P}(\text{Rep})$	0.000469	0.004874
$\mathbb{P}(\text{Ind})$	0.000289	0.005318

Table F2: Parameters Governing the Fraction of Ideologies θ

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

$\mathbf{h}_{\text{Dem,Dem}}$	Bias	Standard Deviation
	0.000352	0.000592
	0.000110	0.000951
	0.000025	0.000896
	-0.000488	0.001082

$\mathbf{h}_{\text{Rep,Rep}}$	Bias	Standard Deviation
	0.000309	0.000557
	-0.000051	0.000850
	0.000283	0.000831
	-0.000542	0.001025

$\mathbf{h}_{\text{Ind,Ind}}$	Bias	Standard Deviation
	-0.000002	0.001127
	0.000030	0.001127
	0.000158	0.001154
	-0.000186	0.001181

$\mathbf{h}_{\text{Dem,Rep}}$	Bias	Standard Deviation
	-0.001770	0.001944
	0.000260	0.001699
	0.000455	0.001516
	0.001056	0.001108

$\mathbf{h}_{\text{Dem,Ind}}$	Bias	Standard Deviation
	-0.000148	0.001221
	-0.000079	0.001344
	-0.000163	0.001343
	0.000389	0.000966

$\mathbf{h}_{\text{Rep,Ind}}$	Bias	Standard Deviation
	-0.000176	0.001315
	-0.000162	0.001176
	-0.000244	0.001361
	0.000581	0.000852

Table F3: Parameters Governing the Weight Distribution $\mathbf{h}$

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

ϵ	Bias	Standard Deviation
	-0.001132	0.002546

Table F4: Parameter Governing Measurement Error ϵ

Note: Bias is defined as the difference between the average of posterior means across simulations and the true parameter value. Standard Deviation is defined as the standard deviation of posterior means across simulations.

Appendix G List of PCs with Estimated Ideology Different from Self Report

FEC ID	Committee Name
C00113662	NEW MEXICANS FOR BILL RICHARDSON
C00165753	LEADERSHIP '02 (FKA FRIENDS OF ALBERT GORE JR INC)
C00247734	COMMITTEE TO ELECT WILLIAM J JEFFERSON TO THE UNITED STATES CONGRESS
C00401364	FRIENDS OF JOHN SWEENEY
C00178418	BOUCHER FOR CONGRESS COMMITTEE
C00387829	DEMOCRAT GRAYSON FOR THE HOUSE
C00396101	JON PORTER FOR CONGRESS COMMITTEE
C00316596	CHRIS JOHN FOR CONGRESS
C00316141	RE-ELECT HAROLD FORD
C00220145	GENE TAYLOR FOR CONGRESS COMMITTEE
C00223230	FRIENDS OF JOHN TANNER
C00366401	MARK PRYOR FOR US SENATE
C00315176	FEINSTEIN FOR SENATE
C00347310	FRIENDS OF CHRIS DODD 2004
C00386292	NORTH DAKOTA 2004
C00143438	FRIENDS OF BYRON DORGAN
C00305110	A LOT OF PEOPLE WHO SUPPORT JEFF BINGAMAN
C00364364	DANIEL K INOUYE IN 2004
C00280917	DANIEL K INOUYE IN 98
C00396044	JOHN KENNEDY FOR US SENATE INC
C00391110	VICTORY 2004
C00389957	COBLE FOR US SENATE
C00224972	FRIENDS OF SENATOR ROCKEFELLER
C00215830	JOHN BREAUX COMMITTEE
C00391862	LOUISIANA SENATE 2003
C00325126	FRIENDS OF MARY LANDRIEU INC
C00317214	MARY LANDRIEU FOR SENATE COMMITTEE INC
C00202754	FRIENDS OF KENT CONRAD
C00368209	NELSON 2006
C00306712	NELSON 2000
C00385013	NEVADA SENATE 2004
C00204370	FRIENDS FOR HARRY REID
C00387449	MONTANA NEVADA VICTORY FUND
C00308676	WYDEN FOR SENATE
C00201533	TIM JOHNSON FOR SOUTH DAKOTA INC
C00402008	MONTANA ARKANSAS VICTORY FUND
C00255463	FRIENDS OF BLANCHE LINCOLN
C00385633	ARKANSAS SENATE 2004
C00349217	CARPER FOR SENATE
C00344051	BILL NELSON FOR U S SENATE
C00306860	EVAN BAYH COMMITTEE
C00383497	BAUCUS VICTORY FUND
C00328211	FRIENDS OF MAX BAUCUS 2002

Table G1: PCs that Self Reported to be Democratic, but are Estimated to be Republican

FEC ID	Committee Name
C00400598	ILLINOIS US SENATE VICTORY COMMITTEE
C00387001	OHIO VICTORY COMMITTEE
C00406322	REPUBLICAN PARTY OF KENDALL COUNTY
C00188078	SEVENTH CONGRESSIONAL DISTRICT REPUBLICAN PARTY OF WISCONSIN
C00350496	BRADY FOR CONGRESS
C00371443	DANNY DAVIS FOR CONGRESS
C00400531	KY 04 CONGRESSIONAL VICTORY COMMITTEE
C00376749	RODNEY ALEXANDER FOR CONGRESS INC.
C00272211	PETE KING FOR CONGRESS COMMITTEE
C00272153	COMMITTEE TO ELECT MCHUGH
C00378158	ZARELLI FOR CONGRESS
C00396523	ROSELYN FOR CONGRESS
C00401703	FRIENDS OF ALJANICH
C00385542	PHELPS FOR CONGRESS
C00388884	HOOSIERS FOR HARDY
C00400556	LA 03 CONGRESSIONAL VICTORY COMMITTEE
C00400507	LA 07 CONGRESSIONAL VICTORY COMMITTEE
C00190637	TIERNEY FOR CONGRESS COMMITTEE
C00386060	DEROSSETT FOR CONGRESS
C00399675	FRIENDS OF STEVE MORROW
C00392860	BRAUNER FOR CONGRESS
C00398776	HUFFMAN FOR CONGRESS
C00403642	PAUL RODRIGUEZ FOR CONGRESS
C00386078	BELL FOR CONGRESS COMMITTEE
C00388991	RICCARDI FOR CONGRESS
C00399295	MATT MUEDA FOR CONGRESS
C00395061	JANE ESHAGPOOR FOR CONGESS
C00398834	ASAY FOR CONGRESS COMMITTEE
C00331108	REP DON YOUNG CONSTITUTIONAL DEFENSE FUND
C00320168	ASA HUTCHINSON FOR CONGRESS COMMITTEE
C00395731	TERESA DOGGETT TAYLOR FOR CONGRESS
C00333294	DOUG OSE FOR CONGRESS ’98
C00219204	PORTER GOSS RE-ELECTION TEAM
C00335190	CONNELLY FOR CONGRESS
C00300699	NODLER FOR CONGRESS COMMITTEE
C00096412	COMMITTEE TO REELECT CONGRESSMAN CHRIS SMITH
C00091298	THE COMMITTEE TO RE-ELECT CONGRESSWOMAN MARGE ROUKEMA
C00334334	DON SHERWOOD FOR CONGRESS
C00397075	SANTA CRUZ ACTION COMMITTEE
C00394346	BUSH ADMINISTRATION RETIREMENT FUND PAC (BARF PAC)
C00405688	DUMP BUSH MISSOULA
C00374652	DIANE ALLEN FOR US SENATE
C00389692	DR KATHURIA FOR US SENATE
C00349795	GORMLEY FOR SENATE PRIMARY ELECTION FUND
C00366237	CHAFEE FOR SENATE
C00325571	SENATOR JOHN WARNER COMMITTEE

Table G2: PCs that Self Reported to be Republican, but are Estimated to be Democratic

Appendix H Alternative Model: No Measurement Error, but with Hold Out Sample

In this alternative model, we randomly select 200 PCs in $\mathcal{V}^o$ to be the holdout sample. Additionally, we assume no measurement error, i.e., $\epsilon = 0$ and $x_i = \hat{x}_i$ when $\hat{x}_i$ is available. We estimate this model pretending $\hat{x}_i$ is not available for the holdout sample. In order to assess our estimates, we compare the estimated posterior distribution and the self report for the holdout sample.

Posterior distributions of $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{h}$ are summarized in Table H1, H2, and Figure H1.

$\boldsymbol{\beta}$	Posterior Mean	Posterior Standard Deviation
constant	-72.9487	1.1780
$\mathbb{1}_{x_i=x_j=\text{Dem}}$	47.9304	0.4056
$\mathbb{1}_{x_i=x_j=\text{Rep}}$	47.8478	0.4027
$\mathbb{1}_{x_i=x_j=\text{Ind}}$	47.8889	0.3907
$\mathbb{1}_{(x_i=\text{Dem},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Dem})}$	48.3420	0.4055
$\mathbb{1}_{(x_i=\text{Rep},x_j=\text{Ind})\vee(x_i=\text{Ind},x_j=\text{Rep})}$	48.4466	0.4062
Same state	0.7792	0.0114
Same industry	0.5151	0.0272
One of them is a House campaign	0.5315	0.0082
One of them is a Senate campaign	0.2261	0.0044
One of them is a Presidential campaign	-0.2852	0.0169
One of them is a qualified PAC	0.3990	0.0051
One of them is a qualified Party	-0.7151	0.0133
One of them is a national committee	0.7163	0.0116
One of them is authorized by a candidate	-0.4360	0.0071
One of them is a joint fundraiser	-0.8477	0.0102
$(\ln b_i + \ln b_j)$	1.4648	0.0579
$((\ln b_i)^2 + (\ln b_j)^2)$	0.0056	0.0003
$\ln b_i \ln b_j$	-0.1017	0.0038

Table H1: Posterior Distribution of β

$\boldsymbol{\theta}$	Posterior Mean	Posterior Standard Deviation
$\mathbb{P}$(Dem)	0.3839	0.0045
$\mathbb{P}$(Rep)	0.3913	0.0045
$\mathbb{P}$(Ind)	0.2248	0.0038

Table H2: Posterior Distribution of θ

In the following part, we focus on the holdout sample. Table H3 cross tabulates PCs in the holdout sample according to self-reported and estimated ideology. Our estimated ideology matches the self report for 87.76% of those who reported to be Democratic and 79.17% of those who reported

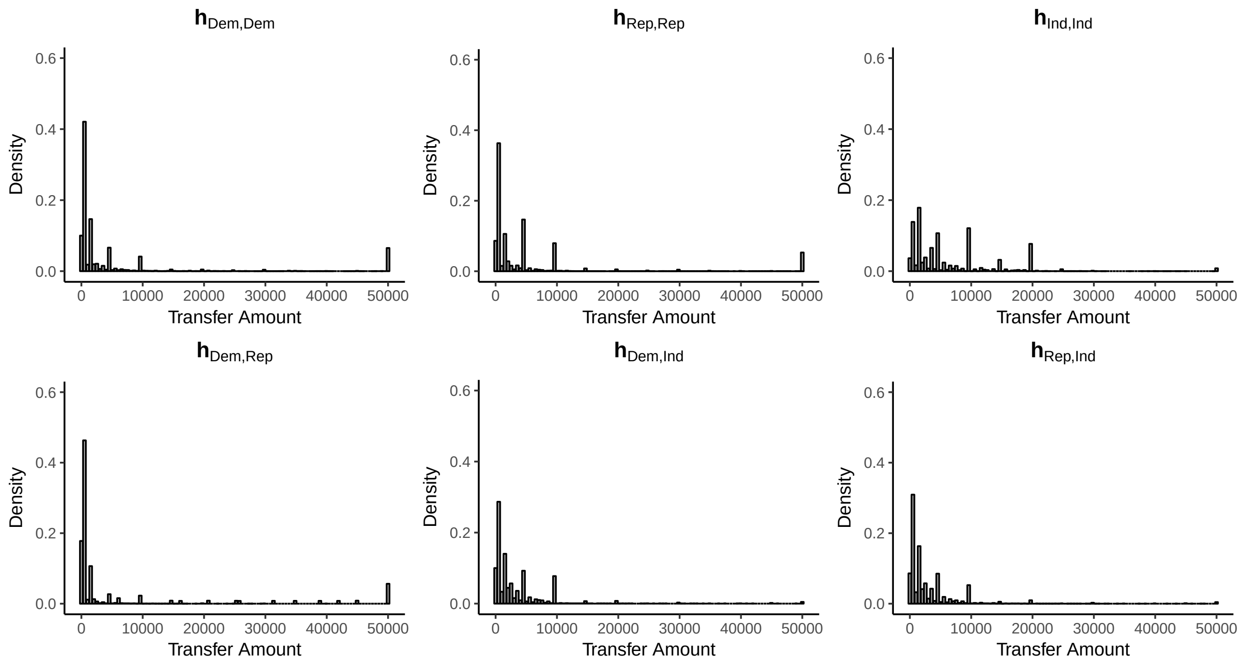


Figure H1: Posterior Mean of h

Note: Probability of transfer amount higher than \$50,000 is plotted at \$50,000.

to be Republican.

In this part, we analyze the posterior distribution of political ideology. Figure H2 plots, for PCs whose estimates are the same as the self-reports (i.e., the reported ideology has the highest posterior probability), the distribution of the differences between the highest and the second highest posterior probability. These differences concentrate around 1, meaning the posterior probabilities concentrate on the self-reported affiliation. This confirms that we do not obtain these classifications by luck.

We do a similar analysis in Figure H3, for PCs whose estimates differ from the self-reports, by plotting the distributions of the differences between the highest posterior probability and the posterior probability of the self-reported ideology. The distributions for PCs which self reported to be Democratic (Republican) but are estimated as Republican (Democratic) are concentrated around 0, indicating that these are "near misses".

Finally, Table H4 and H5 list the PCs that self reported to be Democratic (Republican), but are estimated to be Republican (Democratic).

	Estimated Dem	Estimated Rep	Estimated Ind
Self-Reported Dem	86 (87.76%)	9 (9.18%)	3 (3.06%)
Self-Reported Rep	10 (10.42%)	76 (79.17%)	10 (10.42%)
Self-Reported Ind	1 (16.67%)	3 (50.00%)	2 (33.33%)

Table H3: Tabulation of Estimated vs. Self-Reported Ideology

Note: The percentages are calculated for each row.

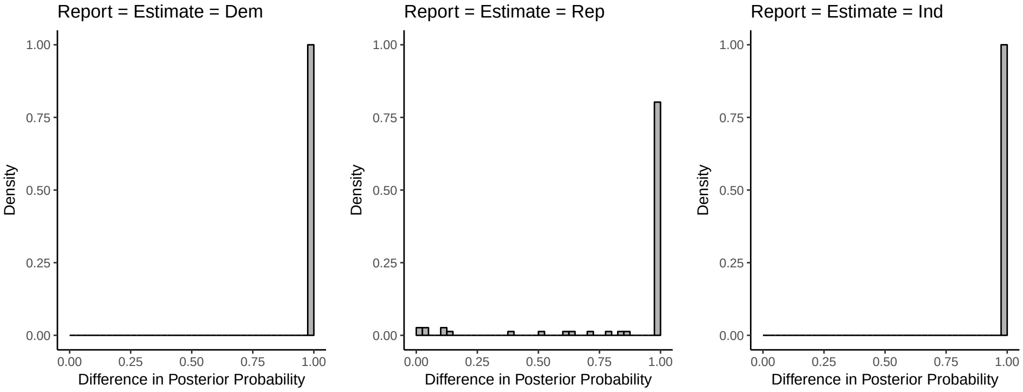


Figure H2: Distribution of Difference in the Posterior Probability of Ideology

Note: Horizontal axis is the difference between the highest posterior probability (i.e., the posterior probability of the self-reported ideology) and the second highest posterior probability. Bin size is 0.025.

FEC ID	Committee Name
C00327403	FRIENDS OF JONATHAN MILLER
C00367060	JOHN MILKOVICH FOR CONGRESS
C00381350	MARK BUDETICH
C00388454	JOHNSON FOR US SENATE
C00390245	REYES FOR CONGRESS
C00394858	VICTORY 04
C00399097	JOHN SALAZAR AND KEN SALAZAR JOINT COMMITTEE
C00399154	BURKS FOR US SENATE CAMPAIGN COMMITTEE
C00402149	FRIENDS TO ELECT JEFF MILLER

Table H4: PCs that Self Reported to be Democratic, but are Estimated to be Republican

FEC ID	Committee Name
C00188078	SEVENTH CONGRESSIONAL DISTRICT REPUBLICAN PARTY OF WISCONSIN
C00349795	GORMLEY FOR SENATE PRIMARY ELECTION FUND
C00367839	SALAZAR FOR CONGRESS
C00371443	DANNY DAVIS FOR CONGRESS
C00375485	RUSTY GLOVER FOR CONGRESS
C00386078	BELL FOR CONGRESS COMMITTEE
C00387571	STARK FOR CONGRESS
C00389130	FRIENDS OF JOE NEGRON
C00390203	RISLEY FOR CONGRESS
C00404772	RANDY EASTWOOD FOR CONGRESS

Table H5: PCs that Self Reported to be Republican, but are Estimated to be Democratic

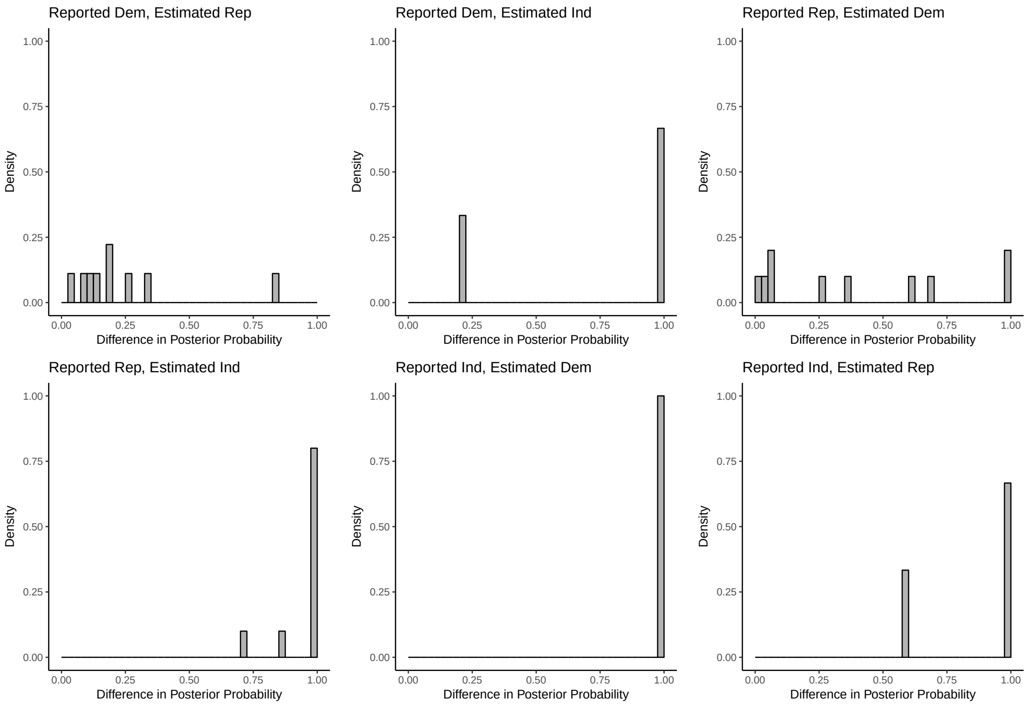


Figure H3: Distribution of Difference in the Posterior Probability of Ideology

Note: Horizontal axis is the difference between the highest posterior probability and the posterior probability of the self-reported ideology. Bin size is 0.025.