國立臺灣大學社會科學院經濟學系
碩士論文
Department of Economics
College of Social Sciences
National Taiwan University
Master Thesis

Facebook 作為社會劫盜地圖：意識形態估計、媒體偏斜、與輿情隔離

Ideology Estimation, Media Slant, and Opinion Segregation: Facebook as a Social Barometer

張耕齊
Keng-Chi Chang

指導教授：
林明仁 博士、江淳芳 博士

Advisors:
Ming-Jen Lin, Ph.D.
Chun-Fang Chiang, Ph.D.

中華民國 106 年 7 月
July, 2017

doi:10.6342/NTU201702253

摘要

利用 Facebook 去識別化的公開資料，我們提出一個廣泛的框架，將美國不同類型的政治參與者 (如政治人物、新聞媒體、利益團體、與社會大眾等) 全部定位在共同的意識形態光譜上。透過辨認潛藏意識形態資訊的粉絲專頁，並選擇可能提供訊息的使用者，我們提供了新的關於政治人物意識形態與媒體偏斜的估計，這些估計也重製了傳統衡量的結果。此外，對一般大眾意識形態的估計結果也較符合全國與各州實際上的分配。與過去研究不同的是：我們的方法並不侷限在政治生活的特定層面；產生的大眾意識形態分配較為平滑合理；估計能隨著時間改變；並且可以依據不同議題做進一步分析。這使得我們的方法能延伸，並且更具使用價值。我們也討論了一些因為這個衡量方式所產生的未來研究方向，例如預測選舉結果，以及衡量在社群媒體上輿情隔離的程度。

關鍵詞：意識形態估計、媒體偏斜、輿情隔離、社群媒體。

JEL 分類代號：D72、L82、D83、C81。

doi:10.6342/NTU201702253

Abstract

We present a general framework to place different political actors including politicians, news outlets, interest groups, and the mass public all on the same ideological spectrum, using only de-identified, publicly available Facebook data. By specifying a potential ideological universe of fan pages and selecting informative users, we are able to give some new evidence and reproduce conventional measures regarding political ideal points and media slants, and also replicate ideology distribution of citizens both at national and at state levels. Unlike previous works, our procedure does not constrain to a specific aspect of political life, can generate a reasonably smooth mass ideology distribution, is time-variant, and is also topic-decomposable. This makes it extensible and useful for future research. Several new avenues of research made possible by our estimates such as election forecasting and measuring opinion segregation on social media are also discussed.

Keywords: ideal point estimation, media slant, segregation, social media.

JEL Classification: D72, L82, D83, C81.

doi:10.6342/NTU201702253

Contents

doi:10.6342/NTU201702253

doi:10.6342/NTU201702253

List of Figures

doi:10.6342/NTU201702253

doi:10.6342/NTU201702253

List of Tables

doi:10.6342/NTU201702253

Chapter 1

Introduction

> *Only connect! That was the whole of her sermon.*
> *Only connect the prose and the passion, and both will be exalted,*
> *And human love will be seen at its height.*
> *Live in fragments no longer.*
>
> — E. M. Forster, *Howards End*

Statistics initiated with a 14th-century Florentine statesman's (or *statista* in Italian) desire to understand his Republic. The purpose of his detailed political arithmetic regarding population, business, and religion is to control the society. Today we are lucky enough to live in an era where people usually want to do the reverse: elites want to know what people are thinking so that they can adjust themselves to suit others.

Estimating people's political preference is vital to understand one's behavior since many important choices are based upon these views. Voters elect political elites, interest groups lobby politicians for changes, and media provide relevant information for citizens, which enhances their decision making. Leaving anyone out is prone to miss something from this interactive and circular process.

However, previous studies focus mostly on linking only two types of players and leaving others out, given the fact that it is usually hard to find a common place to connect different types of actors. For example, building on their seminal works of measuring legislative preference (Poole and Rosenthal 1985, 1997, 2007; Clinton et al. 2004), attempts were made to connect political elites and ordinary citizens (Jessee 2009; Bonica 2014; Bond and Messing 2015; Barberá 2015), political elites and media (Groseclose and Milyo 2005), and also citizens with media (Gentzkow and Shapiro 2011).

Nowadays, social media websites along with their mobile apps make connecting people at

doi:10.6342/NTU201702253

historically low-cost, especially bringing the possibility to connect different types of political actors. On social media, political elites can communicate directly to their voters, news outlets want their stories to get read in exchange for visits and advertisements, and political groups hope their ideas be seen and spread.

Given this, it is natural to study the largest and the most influential social media today, which is Facebook. According to recent surveys (Pew Research Center 2016a,b) , within US internet users, which is 86% of adults, 79% of which uses Facebook (compared to 24% in Twitter); 76% Facebook users use it daily (while 42% in Twitter); also, 44% of US adults get news from Facebook (in contrast to 9% in Twitter); furthermore, Facebook usage and engagement are still on the rise, whereas others stagnated.

Although there are already several studies on social media that try to measure people's ideological positions, each has some rooms for improvement. Barberá (2015) uses Twitter data, which is considered far from being representative. Bond and Messing (2015) uses fan page *following* data, which is not publicly available and thus restricts its potential for general use. Also, nature of the data behind citizen's *following* of politicians makes their estimates harder to become dynamic, given that *unfollowing* afterwards are usually quite rare.

Furthermore, these papers only considers fan pages or accounts of politicians. However, would a moderate choose to follow politicians? If not, is there a way to put these moderates back on the ideological spectrum, if we really want to take Facebook estimates with respect to the mass public seriously, provided that this is indeed a strength of social media data and these moderates often decide many important political outcomes?

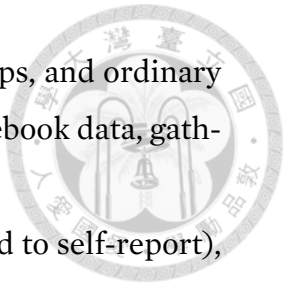
Perhaps an evidence suggesting that only focusing on politicians are not enough is presented in Figure 2 of Bond and Messing (2015), where they plot the densities of ideology estimates for both politicians and individuals. The distribution of individuals is far more polarized than that of politicians. This contradicts to most conventional mass ideology measurements, where moderates should at least occupy a significant proportion of the distribution.

Or, it can also be the opposite. Although there are 37% self-report moderates (National Opinion Research Center 2017), could it be that most of these people are quite extreme to some extent so that it is hard for us to distinguish them from others, at least behaviorally?

These are all important questions, but only looking at behaviors on politician fan pages may not help us address these problems. There are still other aspects of political life.

In this paper, we specify a possible ideological universe that does not depend on a predefined pool of pages, and explore one of the most common actions on Facebook: *likes*. Assume that people are more likely to *like* the posts from fan pages that are closer to their own ideo-

doi:10.6342/NTU201702253



logical position, we are able to place politicians, news outlets, interest groups, and ordinary citizens on the same ideological spectrum, using only publicly available Facebook data, gathered through Facebook's free Graph API.

Also, this measure is based on actions or revealed preference (as opposed to self-report), can be collected at lower cost (compared with surveys), and almost in real time. Furthermore, since we are looking at *liking of posts* not *following of pages*, this adds the whole universe of time and post content dimensions that are worth long-term investigating. Last but not least, compared to methods focusing only on politician pages, our estimates of mass ideology is distributed far more smoothly and seems to replicate the ideological distribution both at national levels and at state levels.

We also provide some interesting applications on the dynamics of voter-politician and media-audience interactions, inspecting echo chambers, forecasting 2016 presidential election, and measuring opinion segregations on social media.

This thesis proceeds as follows. Chapter 2 reviews the literature. Chapter 3 specifies the model and explains the methods adopted. Chapter 4 describes our data and presents the outcome. Chapter 5 compares our results to other related findings. Chapter 6 provides some applications with discussions, and finally concludes.

doi:10.6342/NTU201702253

Chapter 2

Literature Review

2.1 Measuring Ideology of the General Public

Ideology measurements of individuals are generally conducted in surveys. Researchers usually ask respondents to place themselves on a 7-point liberal-conservative scale (see General Social Surveys (National Opinion Research Center 2017) and American National Election Study (American National Election Studies 2017)).

This method, though convenient and straightforward, has some potential problems. A discrete measure makes it hard to transform or combine with other measures. It also does not account for the multidimensional nature of ideology if separate questions for economic, moral, or other social or policy issues were not presented.

Perhaps the most disturbing fact is that respondents may interpret the questions differently (Bauer et al. 2016), or there may be certain social pressure for respondents to respond in a certain way (Schiffer 2000; Gervais and Najle 2017). Though parallel problems may exist in Facebook data, the high dimensional nature compared with surveys may provide chances to overcome or decompose such bias.

2.2 Ideal Point of Political Elites

There is a vast literature on estimating the ideal point of politicians. Most of which involves using roll-call voting records to estimate the ideological positions of the members of Congress. Poole and Rosenthal (1997) builds their foundational work on supposing legislators would vote for roll-calls that are closer to their own ideal point. By further assuming the functional forms of the utility function of the legislators and the error term, they developed the well-

doi:10.6342/NTU201702253

known DW-Nominate method that can estimate the ideal points via maximum likelihood.

Clinton et al. (2004) extends the procedure into a Bayesian setting that is more flexible to incorporate other information (priors) and can be estimated using Markov Chain Monte Carlo (MCMC) simulations through maximizing the posterior distribution. Other than likelihood-based methods, Heckman and Snyder (1997) uses a form of dimension reduction to estimate legislative preference that lowers computational costs and achieves similar result presented in DW-Nominate method.

A serious problem in this line of research is that we cannot apply it to people outside Congress. More broadly speaking, since different political actors make different choices, we cannot estimate their ideal points jointly.

Bonica (2014) is a creative breakthrough to the just-mentioned problem (see also Bonica (2016)). By making use of campaign finance data and assume that people contribute to politicians similar to their own ideological positions, we can jointly estimate the ideological positions of some citizens and politicians outside Congress. To reduce computational cost, they use correspondence analysis (a form of dimension reduction) for estimation. However, since there may be serious self-selection problem in campaign finance data, namely perhaps only politically active people would donate to politicians, it may be hard to generalize its interpretation to the general public.

2.3 Understanding Media Bias

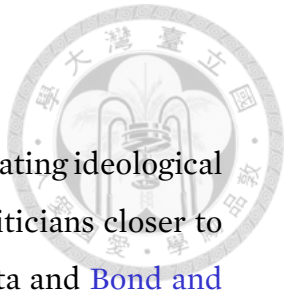
Media also plays an indispensable role in our political life, though there seem to be slightly fewer works on measuring the ideological positions of media, given that we also need to find links to connect media and other political actors in order to get a meaningful result.

Groseclose and Milyo (2005) links media and politicians by counting the times each news outlet cites particular think tank and compare it to the times the members of Congress cite those think tanks. Gentzkow and Shapiro (2011) links media and the public by looking at browsing records of news outlet websites and visitor's self-reported ideology.

Although these are very interesting results, citing may be a rare event and self-reported data may have some above-mentioned problems. More importantly, we cannot place these actors directly on the same scale.

doi:10.6342/NTU201702253

2.4 Ideal Point Estimation Using Social Media



Bond and Messing (2015) and Barberá (2015) are main contributions to estimating ideological scores using social network data. Assuming that people tend to follow politicians closer to their own unobserved ideological position, Barberá (2015) uses Twitter data and Bond and Messing (2015) uses Facebook data to estimate a joint ideology score for politician and mass.

However, Twitter users are less representative, and following data is not publicly available on Facebook. Also, following (or liking fan pages) itself is usually a one-shot action. Using only following data on Facebook is perhaps a waste of information since data on the liking of posts on fan pages is not only publicly available, it also may provide time and post content level dimensions to our estimates that are worth long-term studying.

Lastly, why stop at politicians? [1] Since all users are making the same choice: which post to *like*, and all pages are competing for the same scarce resource: user's attention, Facebook provides historically one of the best environment to jointly estimate ideological positions for different political actors, at least in the eyes of Facebook users.

In this paper, we will try to present an estimation procedure that is based on action-revealed preference, can place different political actors on the same spectrum, is time-variant and topic-decomposable, and requires only publicly available data.

[1] In fact, if we use the method developed in Bond and Messing (2015) on posts of politician fan pages, we will get bad estimates for Democratic politicians as verified by the low correlation between the estimate and DW-Nominate scores. See Section 5.2 for details.

doi:10.6342/NTU201702253

Chapter 3

Model and Method

3.1 Facebook Post Endorsement Model

Similar to Bond and Messing (2015) and Barberá (2015), the fundamental assumption in this paper is that Facebook users tend to like the posts of those fan pages that are closer to their own unobserved ideal point. Below present a modified version of Facebook post endorsement model.

Assume that user i's latent ideological position is θ_i and politician/media/interest group j's position is ϕ_j. User i gains utility from liking page j's post, which is proportional to the negative Euclidean distance between θ_i and ϕ_j. Normalize the event that i not liking j's post to have zero utility. Hence,

$$\begin{aligned} U_{ij}\,(\text{like}) &= -\left\|\theta_i - \phi_j\right\|^2 + \tilde{\alpha}_i + \tilde{\beta}_j - \nu_{ij}, \\ U_{ij}\,(\text{status quo}) &= 0. \end{aligned} \tag{3.1}$$

Note that we account for user and page fixed effects $\tilde{\alpha}_i$ and $\tilde{\beta}_j$ to capture the fact that some users likes more pages (get more utility from liking posts, not so good at distinguishing the latent ideological space, etc.), and that some pages have more likes (more popular, well-known, easier to find, etc.). Also, we preserve a random component ν_{ij} to capture that not all likes yield the same utility.

Thus, user i will like page j's post (denoted by $y_{ij} = 1$) if $U_{ij}\,(\text{like}) > U_{ij}\,(\text{status quo})$. Further assuming that random component $\nu_{ij} \sim \text{logistic}\,(0, 1/\gamma)$, we can derive that the proba-

doi:10.6342/NTU201702253

bility user i likes page j's post to be

$$\begin{aligned}\Pr\left(y_{ij}=1 \mid \tilde{\alpha}_i, \tilde{\beta}_j, \gamma, \theta_i, \phi_j\right) &= \Pr\left(\nu_{ij} < \tilde{\alpha}_i + \tilde{\beta}_j - \left\|\theta_i - \phi_j\right\|^2\right) \\ &= \frac{\exp\left(\gamma\left(\tilde{\alpha}_i + \tilde{\beta}_j - \left\|\theta_i - \phi_j\right\|^2\right)\right)}{1+\exp\left(\gamma\left(\tilde{\alpha}_i + \tilde{\beta}_j - \left\|\theta_i - \phi_j\right\|^2\right)\right)} \\ &= \operatorname{logit}^{-1}\left(\alpha_i + \beta_j - \gamma\left\|\theta_i - \phi_j\right\|^2\right),\end{aligned} \tag{3.2}$$

where we reparameterized $\gamma\tilde{\alpha}_i = \alpha_i$ and $\gamma\tilde{\beta}_j = \beta_j$.

3.2 Identification

The parameters θ_i and ϕ_j in Equation 3.2 are generally not identified. Observe that we can add a constant or scale with nonzero constant (including reflecting with negative constants) to θ_i and ϕ_j without changing model specifications.

There are two ways to address this problem. One is assume two arbitrary candidates to have positions -1 (liberal) and $+1$ (conservative) (Clinton et al. 2004). Another is to shift and scale the estimated positions to have mean zero and standard deviation one (Bond and Messing 2015; Barberá 2015). Although the latter does not solve the reflection problem (that is, the left-right direction can be reversed), one can always flip it back in order to have an ease of interpretation.

3.3 Traditional Estimation Method

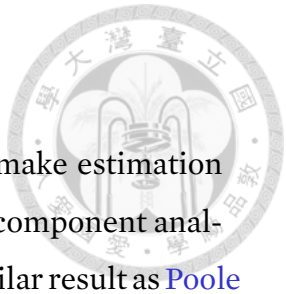
Traditionally Equation 3.2 is solved by Markov-Chain Monte Carlo (MCMC) algorithm through assuming some prior distributions of $\alpha_i, \beta_j, \theta_i$, and ϕ_j to maximize joint posterior density given data via simulation (Clinton et al. 2004; Gelman et al. 2013)

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma \mid \mathbf{y}\right) &= \prod_{i \in \text{user}} \prod_{j \in \text{page}} \operatorname{logit}^{-1}\left(\pi_{ij}\right)^{y_{ij}} \left(1 - \operatorname{logit}^{-1}\left(\pi_{ij}\right)\right)^{1-y_{ij}}, \\ \left\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\right\} &= \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma \mid \mathbf{y}\right),\end{aligned} \tag{3.3}$$

where $\pi_{ij} = \alpha_i + \beta_j - \gamma\left\|\theta_i - \phi_j\right\|^2$. However, this is extremely slow once we have tens of millions of users' ideal points θ_i to estimate.

doi:10.6342/NTU201702253

3.4 Estimation Using Dimension Reduction



To address this problem, a number of papers use dimension reduction to make estimation more computationally efficient. Heckman and Snyder (1997) uses principal component analysis to estimate legislative preference using roll-call voting and generates similar result as Poole and Rosenthal (1997). Barberá et al. (2015) uses Correspondence Analysis (CA) to estimate Twitter ideal points. They also use a sample of their data to verify that estimation using Correspondence Analysis and Bayesian simulation are almost the same ($\rho = 0.98$). Bond and Messing (2015) uses Singular Value Decomposition (SVD) to recover the latent ideological position.

In this paper, we use the two-step procedure suggested by Bond and Messing (2015). We first create a page by page matrix that embedded information from users and employ Principal Component Analysis (PCA) on the corresponding matrix. This is identical to using SVD if one standardizes their data before starting the decomposition. [1] Normalizing each column also makes sense since we want to remove fixed effects, as described in the model. After estimating the positions of pages, we then backward calculate the positions of users.

We also verify that using PCA generates similar results as using Correspondence Analysis (Barberá et al. 2015) while computationally less demanding (see Chapter 4.4). [2] Another advantage of PCA is that it has a more intuitive interpretation. That is, principal axes point out the directions that can explain the largest variation in the original data. Principal components are the projections of the original data on these directions.

Though computationally efficient, there are also two drawbacks of dimension reduction. The first is that we don't really know what each dimension means. One (and possibly only one) way to figure the meaning of the dimensions is to guess and verify using other reliable estimations. The second problem is that we have to subjectively determine the numbers of dimensions that is worth studying. Statisticians generally suggest that one can use scree plot to determine the optimal number of dimensions (see Section 4.5).

[1] If $\mathbf{X}$ is a centered data matrix so it has zero sample mean in each column, the empirical covariance matrix is thus $\mathbf{C} = n^{-1}\mathbf{X}^\intercal\mathbf{X}$. What PCA does is to diagonalize $\mathbf{C}$ such that $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^\intercal$. The principal components are the projection of the data on the eigenvectors, which are columns of $\mathbf{X}\mathbf{V}$. If we employ SVD on $\mathbf{X}$ such that $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\intercal$, then the Eckart-Young Theorem (Eckart and Young 1936) says that the nearest possible matrix of rank k to $\mathbf{X}$ is $\mathbf{U}_k\mathbf{S}_k\mathbf{V}_k^\intercal$, which is basically projecting the first k principal components $\mathbf{U}_k\mathbf{S}_k$ back to the original space. We can also derive that $\mathbf{U}\mathbf{S} = \mathbf{X}\left(\mathbf{V}^\intercal\right)^{-1} = \mathbf{X}\mathbf{V}$, which are the principal components, since $\mathbf{V}\mathbf{V}^\intercal = \mathbf{I}$ holds in spectral decomposition.

[2] To use Correspondence Analysis, one needs to decompose the user by page matrix, which is difficult when users are large.

doi:10.6342/NTU201702253

Chapter 4

Data Processing and Results

4.1 Specify the Ideological Universe

In order to make the first principal components related to ideology, we need to specify a set of politics-related fan pages. There is a trade-off on selecting pages. On one hand, if we include only fan pages of politicians, other political actors will be neglected and we also ignore people's behavior on other pages, especially on media, given that news consumption may also be an important indicator of one's political preference. If we believe that media may not be as polarized as politicians, we will make our estimates of mass ideology biased away from the center. This is perhaps what happened in Bond and Messing (2015) and Barberá (2015). On the other hand, if we include too many unrelated pages, the resulting principal components may not be the underlying political preference we are interested in.

To address this problem, we select two sets of pages into our main sample.

First, we select fan pages that ever mentioned two major presidential candidates: Donald J. Trump and Hillary Clinton, in August 2016. We calculate the total number of likes, comments, and shares of candidate-related posts in these pages, and weight them by factors 1:7:14 (a weight suggested by social media consultant), respectively, to determine which pages to include. [1] Also, changing the weights does not change the pool of pages much. We end up with top 1000 election related pages that include all major news outlets, presidential candidates, and policy interest groups.

Second, we include all fan pages of current national politicians, including members and candidates of the Senate, the House, and the past and present Governors. Many politicians

[1] Since no one knows the exact algorithm Facebook adopted to calculate EdgeRank, the score that determines post visibility, we use the weights proposed by (Calero 2013).

doi:10.6342/NTU201702253

own two pages, one official page, and one personal page. We include all of them and use the page that generates more posts to represent the politician when necessary. We end up with a total of 1475 politician fan pages (with 1225 pages have posted in 2015 or 2016).[2]

Finally, we use Facebook Graph API to get the posts and reactions to these posts of these two sets of political-related pages, from January 2015 to November 2016.

4.2 Select Potential US Users

Another major difficulty regarding using Facebook open data is that we know nothing about user's background. Unlike Twitter API, where user's location and much other information are available (Barberá 2015), Facebook does not provide any information other than user's Facebook id number. On the other hand, Bond and Messing (2015) use Facebook internal data, where user's country is known.

What complicates the situation, even more, is the fact that many US fan pages, especially some news outlets, are also well-known globally. Since what we want to estimate is the ideological positions of these pages, at least as close as possible to those in US citizen's eyes, if we just naïvely use all users that ever reacted to the posts of these pages, we may end up with a messy result.

For example, since both The New York Times and Fox News are quite well known outside the US, we may find these two pages have many shared fans and thus making their ideological position close to each other. What makes these pages share many fans, though, is not because they share similar ideologies, but because they happen to be inside some users' limited information set while other pages don't.

To address this issue, we select all users that *ever* reacted to any national level politicians' posts (Senate, House, and Governors; presidential candidates are not included) in 2015 and 2016 to be our supposed US users. This will end up with a total of 29 million users. We only use data of these users to estimate ideal points. Though these users may not be representative of US population (while we intentionally choose not to put any restrictions on reaction times so that some moderates can be included), this is perhaps the simplest way to our knowledge to select users before Facebook's willing to open their black box. [3]

[2] There are 9 overlaps between these two sets of pages, which are: Tim Kaine, Bernie Sanders, U.S. Senator Bernie Sanders, Elizabeth Warren, U.S. Senator Elizabeth Warren, Rand Paul, Ted Cruz, Governor Jan Brewer, and Al Franken.

[3] We also tried to combine any users that ever reacted to any posts related to Super Bowl in top 1000 pages in that week in order to capture more politically moderate users. The results are generally the same.

doi:10.6342/NTU201702253

Time Period	2015-01-01 to 2016-11-30
Total Reactions	19,085,783,534
US User Likes	16,180,488,916
Total Users	366,840,068
US Users	29,412,610
Total Posts	24,788,093
Total Pages	2132
Politician	1225
News Outlets	560
Political Groups	211
Other Public Figures	93
Others	43

Notes: US users are defined by any user that at least reacted to any national politicians' (Sen, Rep, Gov) post once in 2015 and 2016.

Table 1: Data Summary (Main Sample)

Table 1 gives a brief summary of our main sample. Figure 1 shows the cumulative distribution of the number of pages and likes (of posts) reacted by each US user on these pages. The distribution is quite light-tailed–with 50% of users likes only 16 different pages and 86 different posts, and 10% of users likes more than 68 pages and 1176 posts.

4.3 Build Matrices

We follow the procedure proposed by Bond and Messing (2015). Since what we analyze is reaction on posts, we define *fans* of a page to be US users that ever *likes* at least one post in that page in a given period of time. We do not include other reactions (love, haha, wow, sad, and angry) to have an ease of interpretation. Then we are able to construct an affiliation matrix (see Table 2 for an example). The diagonal elements of this matrix are the numbers of unique fans on each page. The off-diagonal elements are the numbers of shared fans between pages. The time period selected in this example are posts from 2015-01-01 to 2016-11-07. We can observe that there are large differences between shared fans among different pages.

We then transform the affiliation matrix to agreement matrix in order to extract meaningful features from shared fans data (see Table 3 for an example). For each element in affiliation matrix $\mathbf{A}$, we compute $g_{ij} = a_{ij}/a_{ii}$ to get agreement matrix $\mathbf{G}$. For example, 0.48 is the number of shared fans between Trump and Fox News divided by the total number of Trump fans, while 0.44 is the number of shared fans between Trump and Fox News divided by the total number of Fox fans.

doi:10.6342/NTU201702253

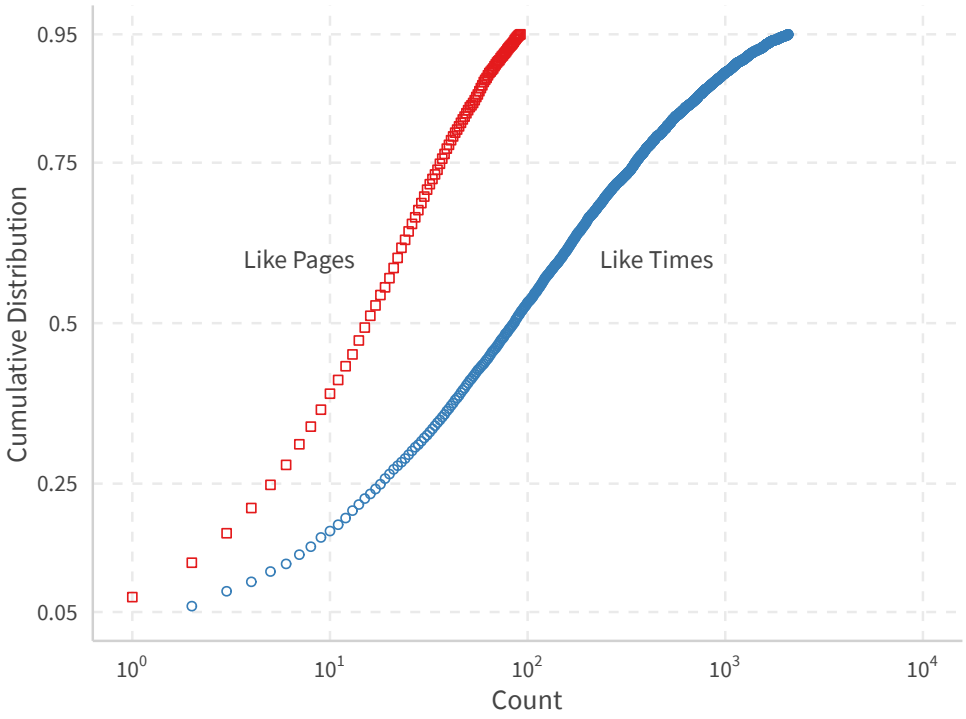


Figure 1: Distribution of Pages and Post per User Likes

Notes: x-axis is log scaled.

	Trump	FoxNews	TeaParty	Clinton	CNN	NYTimes
Trump	2,243,216	1,078,513	128,225	32,731	120,963	25,842
FoxNews	1,078,513	2,449,174	148,016	87,084	186,850	63,401
TeaParty	128,225	148,016	242,089	1528	10,738	2162
Clinton	32,731	87,084	1528	1,768,980	351,210	367,021
CNN	120,963	186,850	10,738	351,210	1,201,156	216,163
NYTimes	25,842	63,401	2162	367,021	216,163	986,613

Notes: Diagonal numbers are unique US users like at least one post of each pages, off-diagonal numbers are shared unique US users at least one post in both pages. Data ranges from 2015-01-01 to 2016-11-07. US users are defined by any user that at least reacted to any national politicians' (Sen, Rep, Gov) post once in 2015 and 2016.

Table 2: Affiliation Matrix (Part)

doi:10.6342/NTU201702253

	Trump	FoxNews	TeaParty	Clinton	CNN	NYTimes
Trump	1.00	0.48	0.06	0.01	0.05	0.01
FoxNews	0.44	1.00	0.06	0.04	0.08	0.03
TeaParty	0.53	0.61	1.00	0.01	0.04	0.01
Clinton	0.02	0.05	0.00	1.00	0.20	0.21
CNN	0.10	0.16	0.01	0.29	1.00	0.18
NYTimes	0.03	0.06	0.00	0.37	0.22	1.00

Notes: For each row in the affiliation matrix, we divide each element by the diagonal element to get agreement matrix. So the numbers in each row are the proportions of shared fans between that page and the pages in each column. Data ranges from 2015-01-01 to 2016-11-07.

Table 3: Agreement Matrix (Part)

This transformation is meaningful so that we can interpret each row as observations and each column as features, with ratios meaning the degree that each observation possess those features. For instance, Trump page is 100% similar to Trump feature, 48% similar to Fox News feature, and 1% similar to Clinton feature, since the denominators are all the number of Trump fans.

4.4 Conduct Principal Component Analysis

After getting the agreement matrix, we run Principal Component Analysis (PCA) on the agreement matrix. The principal axes are linear combinations of the original features. The first principal axis points out the direction that preserves the largest variation in the original data. The first principal component (PC1) projects the original data (agreement matrix) on the first principal axis, which we interpret it as ideology scores of fan pages. This reduces the dimension of the original data from thousands to one.

As discussed in Section 3.2, to partially solve the identification problem, we scale the ideology scores to have mean zero and standard deviation one. We also multiply all scores by -1 when necessary to have an ease of interpretation.

We have also discussed the problems of dimension reduction in Section 3.4. Figure 2 presents the scree plot. We can see that proportion of variation explained for the kth principal component decreases dramatically. This provides evidence that considering the first dimension (the first principal component) may be sufficient for us if we want to focus on the traditional liberal-conservative one-dimensional divide.

Figure 20 in Appendix A shows the scatter plot of pages on the first two dimensions, with PC1 on the x-axis and PC2 on the y-axis. After inspecting the location of pages, we can see

doi:10.6342/NTU201702253

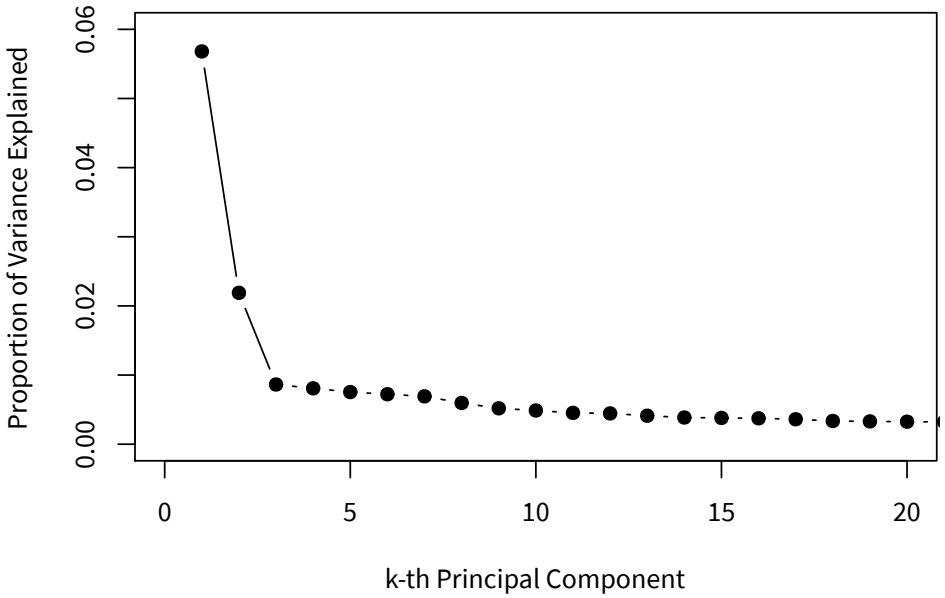


Figure 2: Scree Plot for Principal Component Analysis

that Democratic-related pages are on the left, and Republican-related on the right.

4.5 Results of Fan Pages

We group the pages into three major categories: news outlets, public figures (including politicians and journalists), and political groups (including parties and policy interest groups). Figure 3 gives the distribution of different page types, using data from 2015-01-01 to 2016-11-07. We also annotate some reference points such as Trump, Clinton, Fox News, and The New York Times to give more context to the distribution.

We can observe that news outlets mainly has one mode, public figures and political groups have two modes, while the latter is more dispersed. This is consistent with the roles of these political actors: media serves the general public and interest groups serves politicians. Also, note that we can see most media page are in the center (though slightly left-leaning), there are also a number of pages cluster on the right.

We can also group media pages into categories. Figure 4 shows the result. One can observe that TV, newspapers, and magazines are quite centered (while more left-leaning accordingly), although radio and website news is more dispersed. Appendix A gives other density plots and annotates some notable pages. For example, Figure 26 shows all the major parties in the US, with Democratic, Green, Libertarian, Republican, and Tea Party from left to right. Most media pages replicates recent studies in media bias, such as Groseclose and

doi:10.6342/NTU201702253

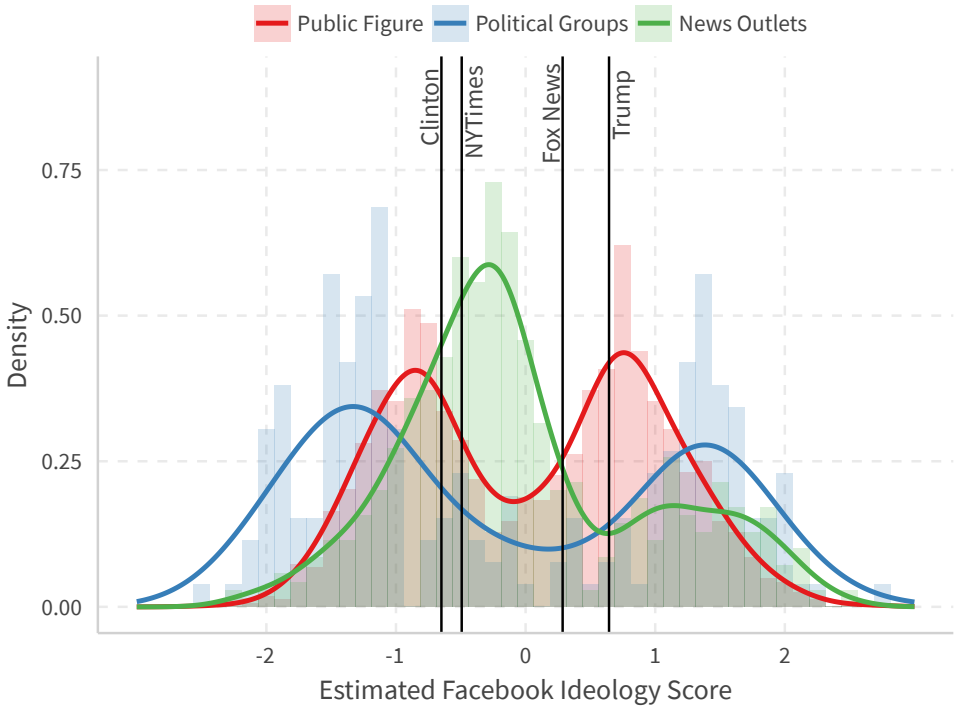


Figure 3: Histogram and Density for Different Page Types

Milyo (2005), Gentzkow and Shapiro (2011), and Pew Research Center (2014).

4.6 Results of Users

Once we estimate the values of ϕ_j as $\hat{\phi}_j$, we can then estimate θ_i by simply calculating the sample means of $\hat{\phi}_j$ that user i likes, since, as described by Equation 3.1, what we want to minimize is $\|\theta_i - \phi_j\|^2$ and sample mean is the minimizer of squared error.

Figure 5 presents the density for all US users. We have made the following adjustments to the data. First, we remove a huge jump created by users only like one and only one page: former California Governor and movie star Arnold Schwarzenegger (1,412,747 users, 5% of the sample). We believe most of these users are fans abroad. We then guess the location of the user by the locations of their maximum likes of national politicians (see Section 5.5 for details) and take random samples of users by comparing to 2016 population in each states (U.S. Census Bureau 2017) if that state is overrepresented in our sample relative to the population.

Still little is known about the ideological compositions of users on Facebook. Bond and Messing (2015) mentions that Facebook users are relatively young, white, educated, female, and liberal. But a caveat is that these are from data in 2012 when the social media giant is still at its early stage. On the other hand, a recent survey (Pew Research Center 2016c) indi-

doi:10.6342/NTU201702253

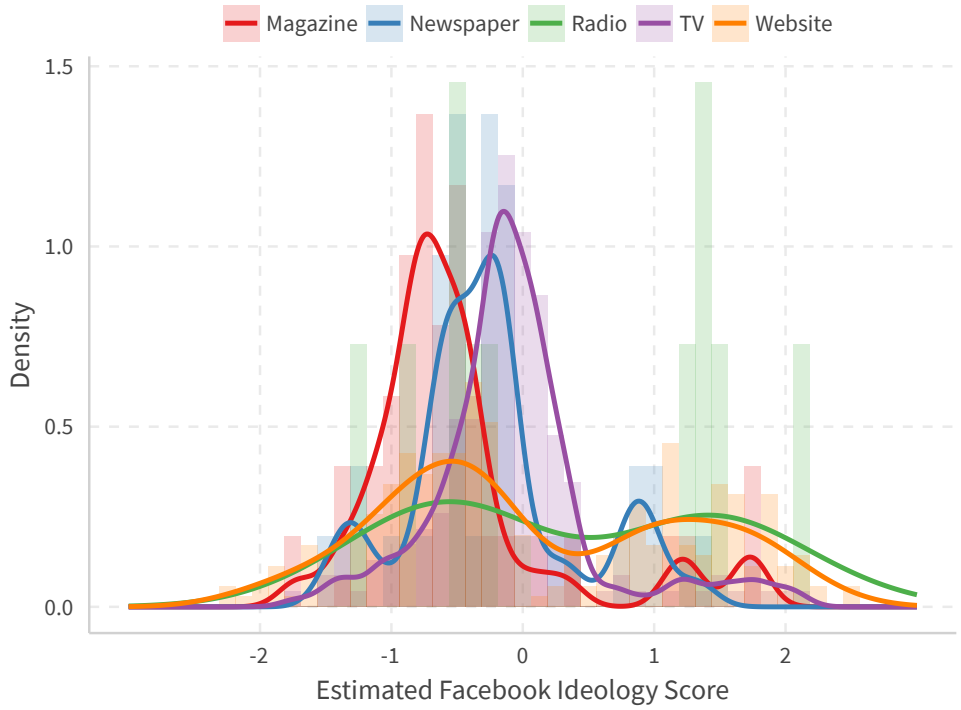


Figure 4: Histogram and Density for Different Media Page Types

cates that 26% of Republicans and 25% of Democrats follow public figures. Among those who follow public figures, 65% follows figures shares their views, only 3% says that they follow figures that are not like-minded. This enforces our confidence of our data, yet more works can be done.

To get what these estimates represent, one can naïvely match these cumulative percentages with self-reported ideology in surveys by assuming that these two represent the same population. Colors in Figure 5 gives the result by matching with General Social Surveys (National Opinion Research Center 2017). [4]

One may also be curious about the usefulness of these self-reported labels. Figures 27 and 27 in Appendix A further shows similar graphs using US users like more than 10 and 20 pages and posts, respectively. The corresponding shapes and quantile values do not change too much as we change the selection of user intensity. This also suggests the potential usefulness of these labels if one wants to interpret the estimates. See Sections 5.4 and 5.5 for further decompositions of our results of users.

[4] In 2016 General Social Surveys, self reported ideologies from extreme liberal to extreme conservatives are: 4.9%, 12.7%, 11.2%, 37.4% (moderate), 13.9%, 15.5%, and 4.4%, respectively. This is quite close to Gallup's 25% liberal, 34% moderate, and 36% conservative estimate (Gallup 2017). The latest numbers in National Election Study we can find is in year 2012 (American National Election Studies 2017).

doi:10.6342/NTU201702253

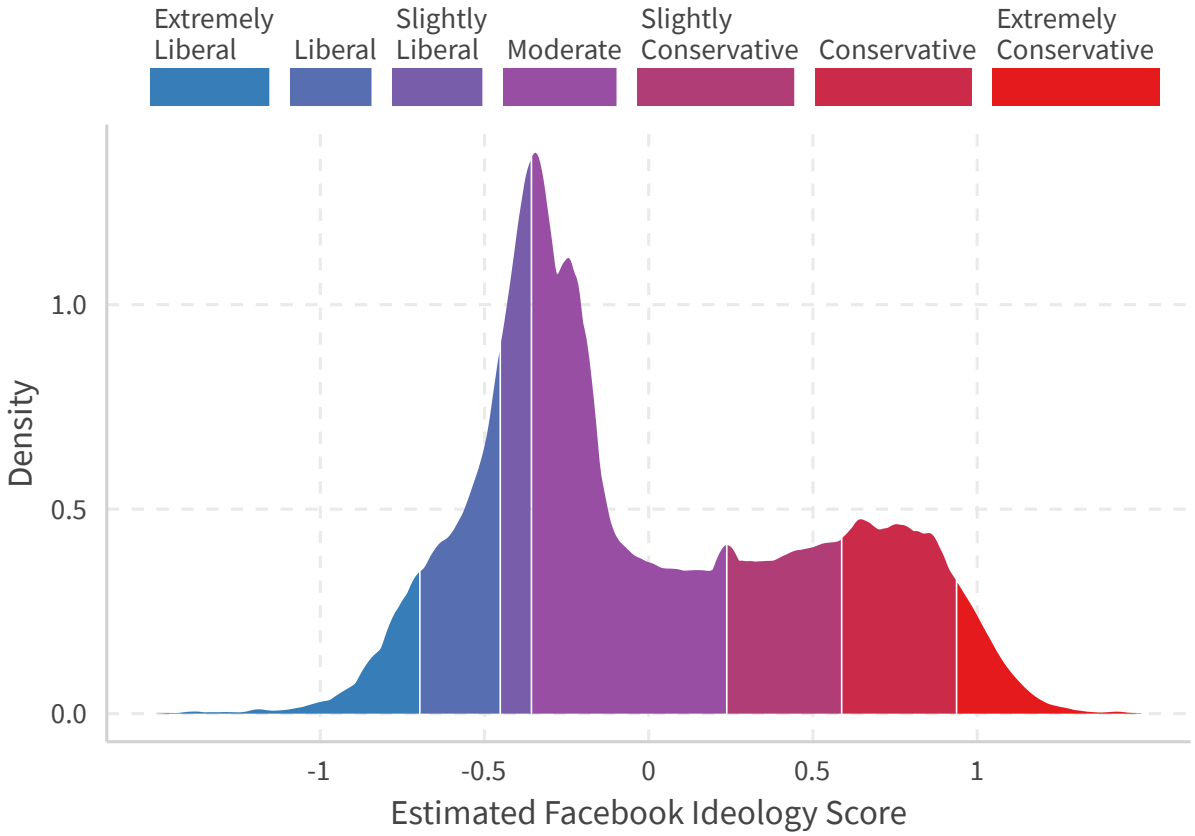


Figure 5: Density for All US Users and Self-Report Ideology Shares in GSS

Notes: Colors represent matching densities with self-reported ideology shares in 2016 General Social Surveys (National Opinion Research Center 2017). US users are defined by any user that at least reacted to any national politicians' (Sen, Rep, Gov) post once in 2015 and 2016, and we guess user's location by the maximum national politician they liked in that state. We remove a huge jump created by users only like one and only one page: Arnold Schwarzenegger. We then sample users by 2016 population in each state (U.S. Census Bureau 2017) if that state is overrepresented in our sample relative to the population.

doi:10.6342/NTU201702253

Chapter 5

Validations

5.1 Methodological Issues

However, how reliable is the first principal component we calculated as a proxy to the positions on the liberal-conservative spectrum?

On a methodological perspective, Barberá (2015) shows that estimating Twitter ideal points using Bayesian simulation and dimension reduction (Correspondence Analysis, CA) are almost the same ($\rho = 0.98$). But procedures proposed by Bond and Messing (2015) have not been verified. Figure 29 in Appendix B shows the comparison between CA and PCA. The results are largely the same, with correlations between pages 0.94 and those between users 0.99. [1] Other than technical limitations such as computer memory, calculation time, and software support, PCA also has strengths in terms of interpretability. These may all facilitate the availability of public use.

5.2 Political Ideal Points

To validate that our measure captures the liberal-conservative divide, one most straightforward approach is to compare our result with the traditional, most wildly-used DW-Nominate Score. Figure 6 shows this scatter plot using data for the 114th Congress (2015–2017). [2] Same

[1] Since CA needs to decompose a user by page matrix, which needs extremely large computer memory, here I conducted CA using users likes more than 70 pages (76,585 users) and pages own more than 10,000 fans (1027 pages).

[2] Many politicians own multiple fan pages. Here we only use the page that produces more post to represent that politician. Some politicians also have pages similar to fans club and not directly-related to the politicians themselves (examples: "Donald Trump, The Political Movement", "Hillary Clinton Supporters"). We count this type of pages as political groups as opposed to politicians. Data for DW-Nominate is retrieved from voteview.com.

doi:10.6342/NTU201702253

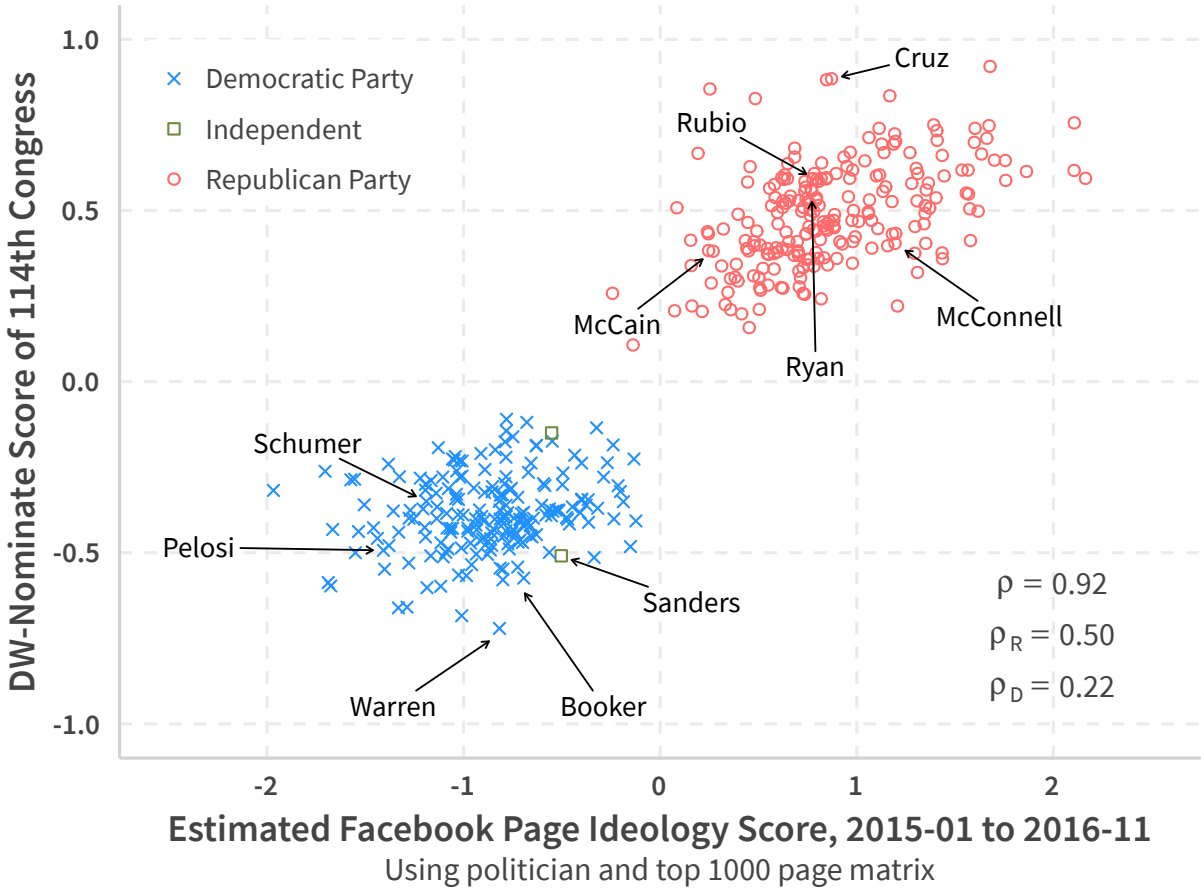


Figure 6: DW-Nominate vs. FB Estimate (114 Congress)

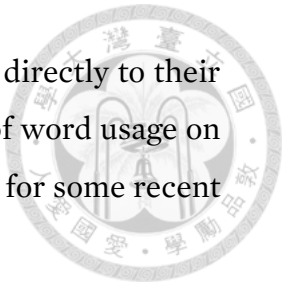
as DW-Nominate scores, our estimate clearly separates politicians into two groups. The overall correlation between two measurements is high (0.92), although correlation inside the Democratic party is relatively low (0.22).

Nevertheless, if we use only politician pages to form a matrix and calculate ideological positions (method in Bond and Messing (2015)), we will get a lower correlation in Democrats, (0.15), as shown in Figure 30 in Appendix B. Also, one will get an even lower correlation if we use Facebook estimates to forecast 115 Congress if one use the procedure in Bond and Messing (2015) (0.15 vs. an almost no correlation 0.09; see Figures 31 and 32 in Appendix B). This suggests that adding other political-related pages does not mess up, instead, it intensifies, our ability to recover people's perceptions of the hidden political spectrum.

The deeper question here is: why do Democrats have in general poorer capability of seeing their political representatives than their Republican counterparts? Could it be that being a minority, as Democratic legislators did, limits their potential to cast votes, based on their beliefs or underlying political liability? On the other hand, perhaps the difference between two measures, one is how voters see them and the other is how they actually act in the Congress, can be interpreted as a measure of how successful political propaganda is. These are of grow-

doi:10.6342/NTU201702253

ing importance given that more and more politicians tend to communicate directly to their supporters. A straightforward route could be to investigate the difference of word usage on Facebook and in Congress (see Gentzkow et al. (2016) and Kim et al. (2017) for some recent attempts to analyze Congressional speech).



5.3 Media Slants

Most findings on media slant or media bias have some potential drawbacks, be it limited sample size, lacking in consistent numeric representation for their estimate, or data-used relatively dated (Groseclose and Milyo 2005; Gentzkow and Shapiro 2011; Pew Research Center 2014).

This makes us hard to perform meaningful comparison, but the big picture they provide for the major news outlets are largely the same: the New York Times and the Washington Post are quite left; ABC News, USA Today, and the Wall Street Journal are considered centrist; while Fox News almost monopolizes the major news market of the right.

Here we present a similar and yet more interesting validation. Define users to be *Republican-affiliated* if their likes in all politicians are more of Republican politicians (compared with other major parties). We can then compute the *share* of Republican-affiliated users on each news outlets. [3]

Figure 7 shows the just mentioned measure against our Facebook Estimate. Our estimate not only replicates both previous studies and the alternative measure, we can see that there are still quite an amount of pages that are on both ends of the spectrum (with almost either only or no Republican users; this also highlights a shortcoming of this alternative straightforward measure), and many of them are still quite popular.

This indicates one of the strengths of our method. Many studies of media slant have to rely on a predefined pool of news outlets or a choice in surveys. This may subject to some sort of bias imposed implicitly by the researchers since most individuals have limited knowledge of what others are seeing. Our evidence, not from a presumed pool, shows that there are still quite a number of sizable right-wing news sources other than Fox News.

This is consistent with findings by Groseclose and Milyo (2005) which suggests a liberal bias for almost all major news outlets. But the demand and supply are still there, on the op-

[3] To remove potential bias created by active users and to be consistent with other papers (such as Gentzkow and Shapiro (2011)), we only count user once per day if they like more than one post of that fan page on that day. We then sum all this kind of so-called *daily* users across day to compute an average share. We use data from 2015-01-01 to 2017-03-31.

doi:10.6342/NTU201702253

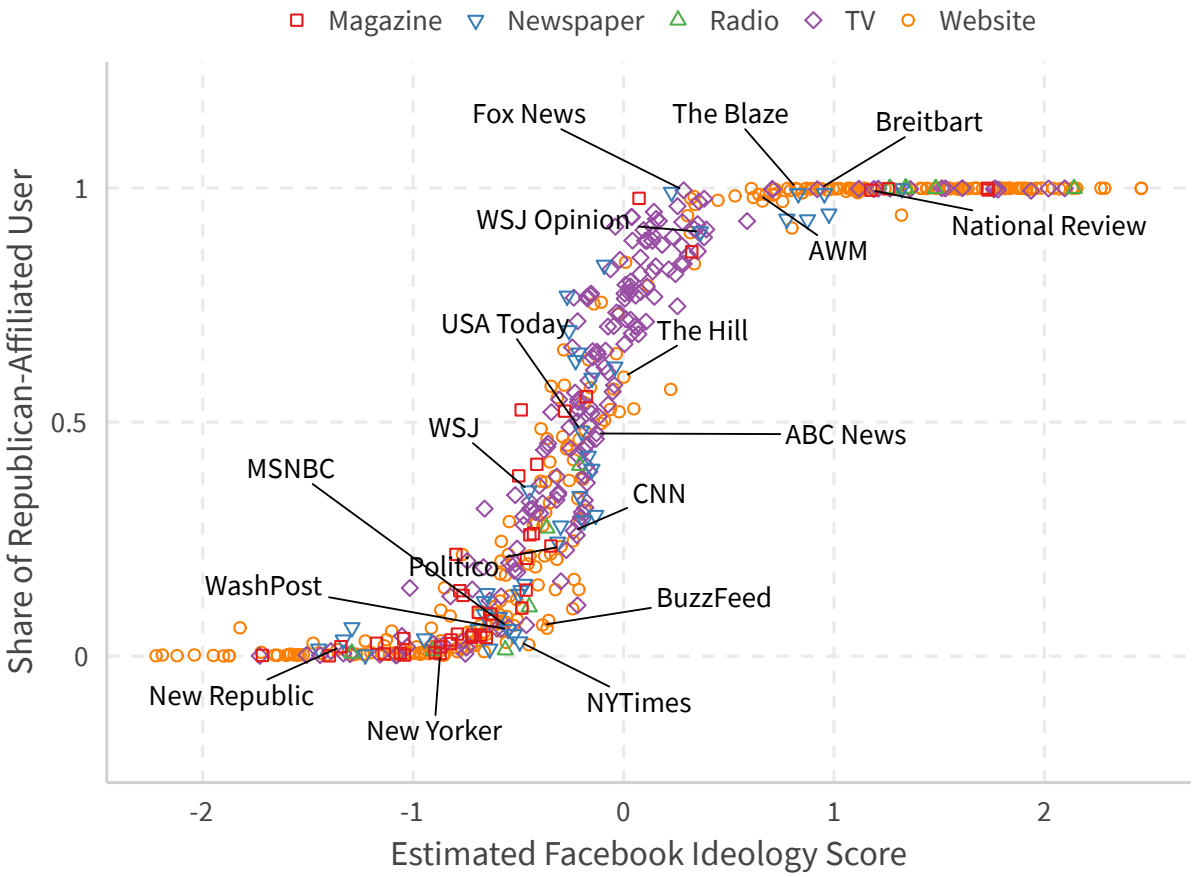


Figure 7: Validation of Media Slant

Notes: A user is *Republican-affiliated* if their likes in all politicians are more of Republicans. We only count user once a day on a page if they like more than one post on that day on that page. We then sum all this kind of daily users up across each day. Data ranges from 2015-01-01 to 2017-03-31.

doi:10.6342/NTU201702253

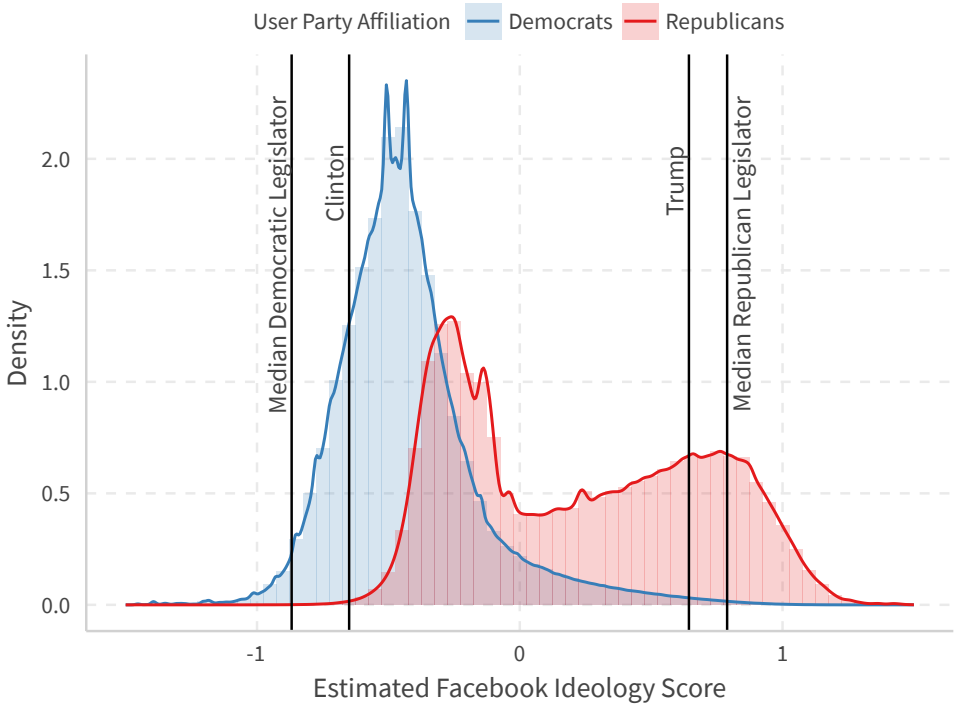


Figure 8: User Party Affiliation vs. FB Estimate

Notes: A user is *Republican-affiliated* if their likes in all politicians are more of Republicans. Data ranges from 2015-01-01 to 2017-03-31. We remove a huge jump created by users only like one and only one page: Arnold Schwarzenegger.

posite end of the world, just not got fully-understood. Our evidence also enables people to notice their own position relative to others while consuming news.

5.4 User Ideologies

We apply the previous-defined user's party affiliation to validate our user ideology estimation.[4] Figure 8 shows our findings. Positions of 2016 Presidential candidates and median legislators are also presented for reference. Extreme regions are dominated by users affiliated with parties, with some cross-affiliation users in the center. One can also see that the jump around zero is possibly caused by Republican-affiliated centrist consuming news around that region.

Furthermore, most users lie between party legislator medians. If we use only pages of politicians to quantify user's position, we may tend to bias user's position from zero, as shown in Bond and Messing (2015) that individuals are more polarized than politicians. This is

[4] Since both Independent legislators and are either previous Democrats (Angus King, Sen-ME) or left-leaning (Bernie Sanders, Sen-VT), we classify them as Democratic politicians.

doi:10.6342/NTU201702253

perhaps due to the fact that endorsing candidates may be only a small part of one's political life. Figure 33 in Appendix B compares our results with the result using only politician pages (Bond and Messing 2015). The politician-only method not only indicates a more heavy-tailed distribution of users, it is also more jumpy and noncontinuous, which seems less consistent with our belief that people could have extremely complicated views on various issues and thus creates a smooth representation, which we call it ideology.

5.5 State Report Cards

Since we have all pages of national politicians (Sen, Rep, and Gov), we can further guess the location of a user by their maximum endorsement of a politician from some state. That is, if one likes more politicians from New York (compared with other states), one should more likely to be from New York. [5]

Figure 9 provides state level densities in six selected states. The top panel is consistently liberal states, the middle states swings from supporting Obama in 2012 to Trump in 2016, and the bottom is conservative states. Colors are using quantiles matched to all US users with national level self-reported ideologies in General Social Surveys, the same as in Figure 5. We can observe the striking disparities among ideology distributions between these states. Also, if we use only, politician pages to calculate user's ideology (Bond and Messing 2015), we will get Figure 10. One can see sharp distinctions between the results of two methods. Plots for all 50 states are presented in Figure 34 of Appendix B.

[5] We treat the event of a user has multiple maxima like states as missing in this series of graphs.

doi:10.6342/NTU201702253

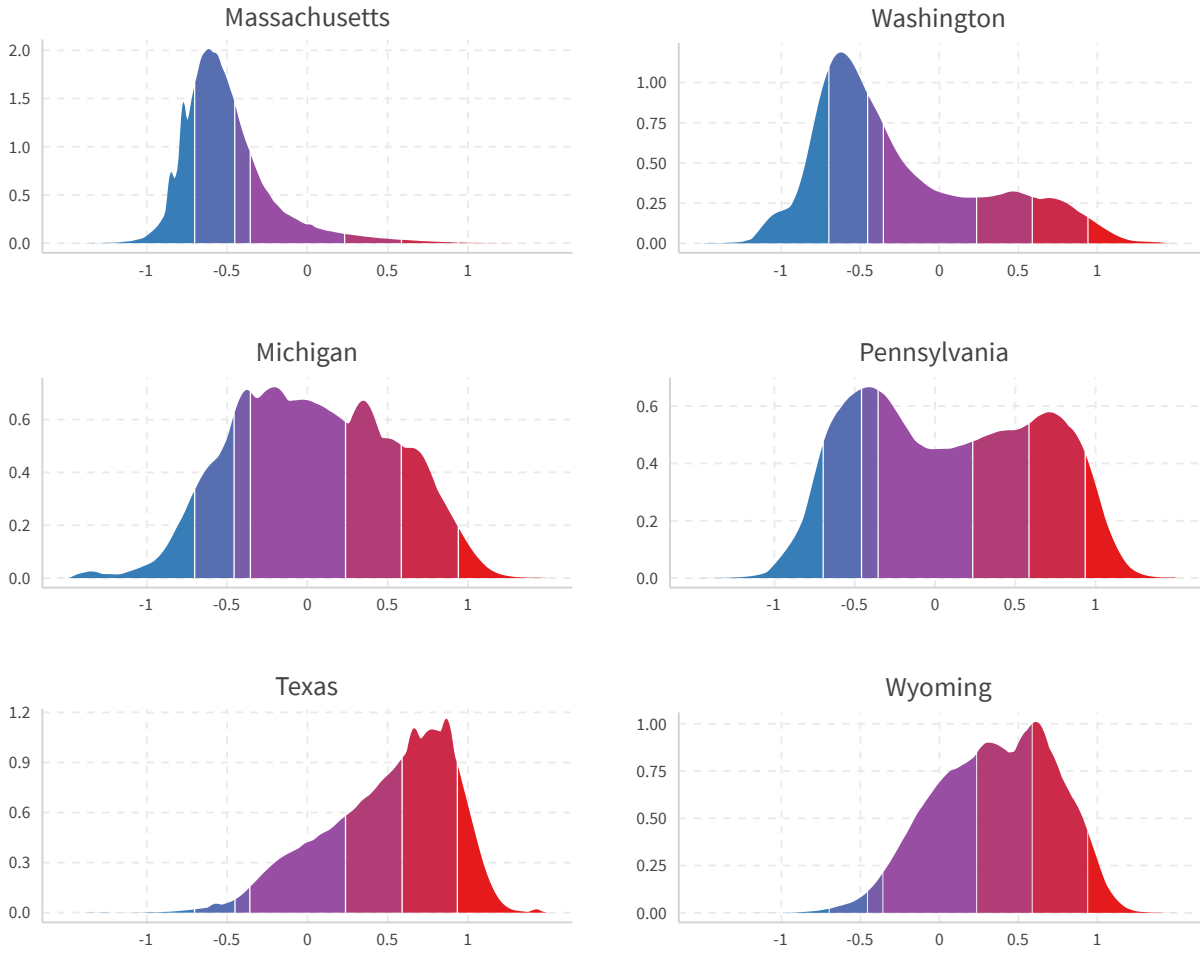


Figure 9: Users in Selected Liberal, Swing, and Conservative States

Notes: States are guessed by the maximum state on likes of national politicians (Sen, Rep, Gov). Colors represent matching densities with self-reported ideology shares in 2016 General Social Surveys (National Opinion Research Center 2017).

doi:10.6342/NTU201702253

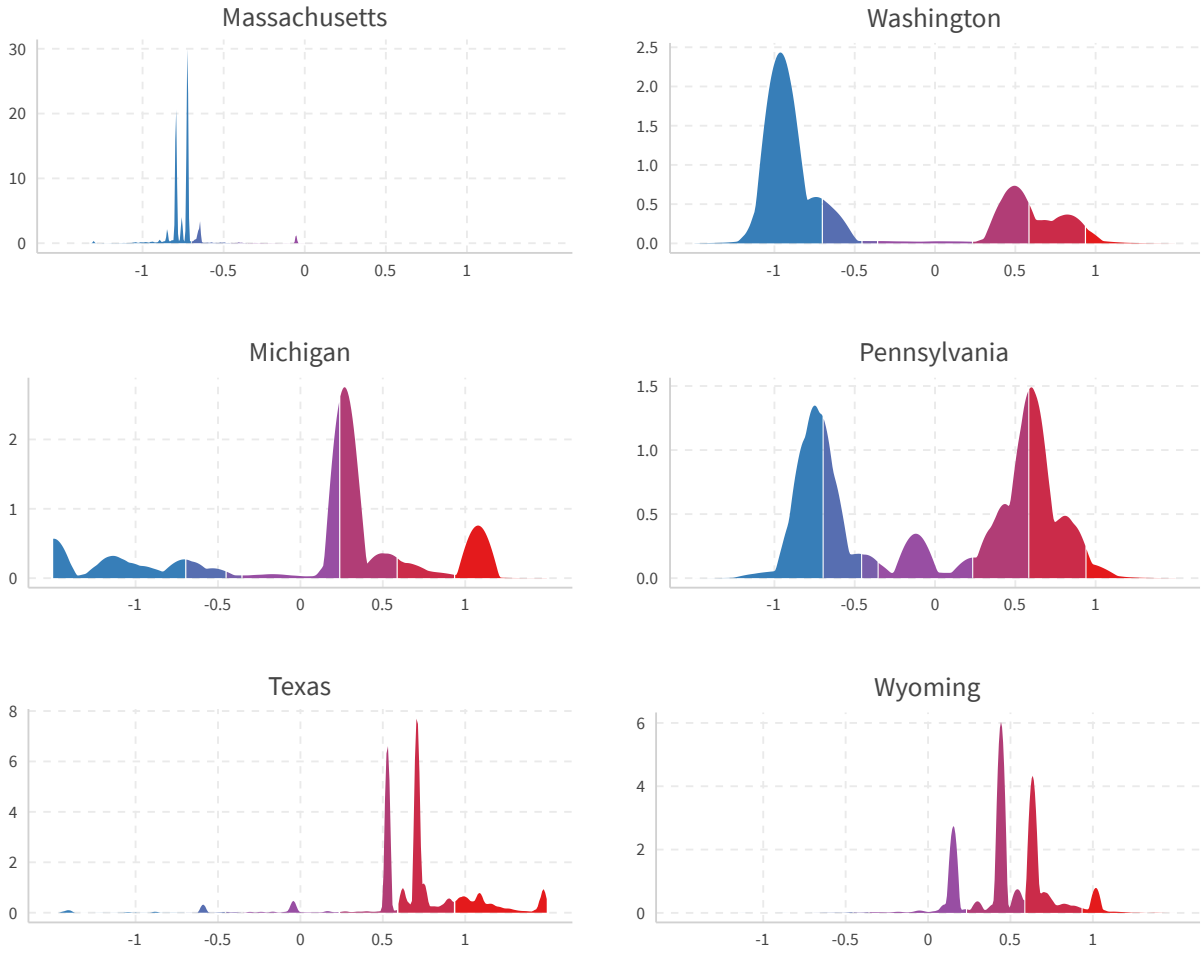


Figure 10: Users in Selected States, Politician-Only Method (Bond and Messing 2015)

Notes: This figure shows user ideology estimate using only politician pages (Bond and Messing 2015). States are guessed by the maximum state on likes of national politicians (Sen, Rep, Gov). Colors represent matching densities with self-reported ideology shares in 2016 General Social Surveys (National Opinion Research Center 2017).

doi:10.6342/NTU201702253

Chapter 6

Applications and Discussions

6.1 Time Dimension: Polarization and Spatial Voting

The first strength of our measure, compared with Bond and Messing (2015) and Barberá (2015), is that since we focus on the liking of posts, it is natural to add on time constraint. The main challenge here is purely technical: in order to produce just *one* point estimate for a specific period of time, one needs to process a large number of likes in that period in order to generate a page by page matrix and then compute principal components. Computation time increases, both as we want to generate a more intensive time series and as we want to include sufficient amount of likes in order to get a less-sparse matrix and thus a more reliable estimate. Here we demonstrate some preliminary results using a 4-month time frame.

Figure 11 plots the time series for some news outlets. One can observe that although their positions are quite stable (partly due to we use a large time frame), pages seem to get more polarized as election approaches. Do they choose to do so, or just responding to their core audiences, and will this trend persist, is quite an interesting question.

A large number of rational choice theories are based on spatial models where political elites move to occupy dense ideological spectrum and voters vote accordingly (Hotelling 1929; Downs 1957). Traditionally we can only test this in elections using vote shares, but since elections are rare, it is hard for us to observe any potential dynamic interactions.

Figure 12 plots a rough outcome for 2016 major Presidential primary candidates. We can see that most candidates move to the center after their announcement. It persists during official primaries from February to June 2016. Some even tend to move back to the extreme after withdrawal.

doi:10.6342/NTU201702253

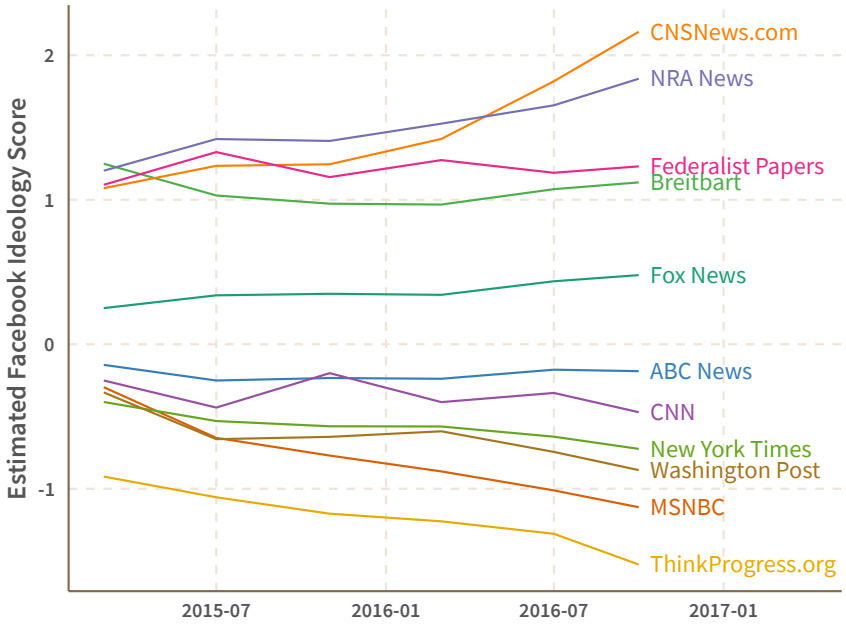


Figure 11: Ideological Time Series for Selected News Outlets

Notes: Preliminary result using 4-month time frame.

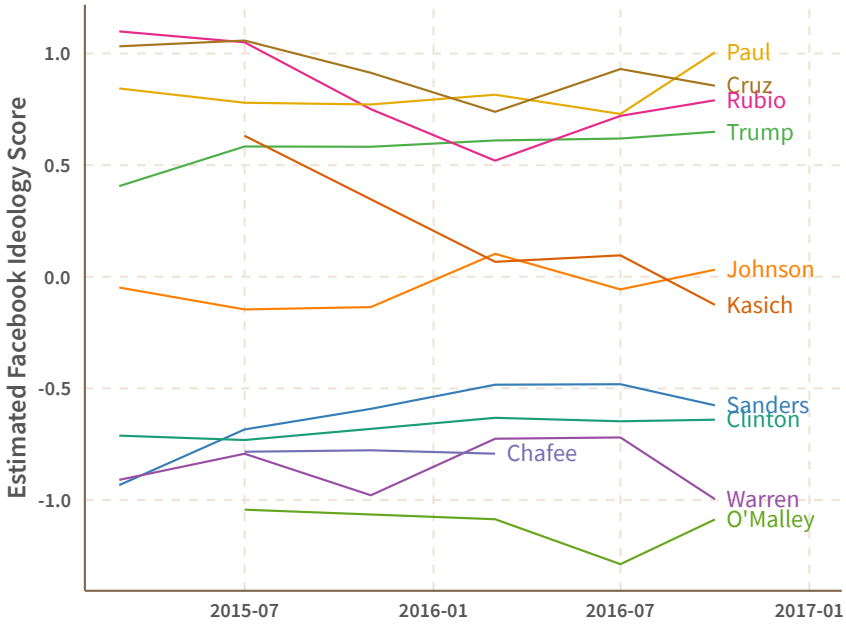


Figure 12: Ideological Time Series for Major Presidential Primary Candidates

Notes: Preliminary result using 4-month time frame. Elizabeth Warren is for reference. Ted Cruz and Hillary Clinton announces in March and April 2015, respectively. Martin O'Malley withdraws in October 2015. The official primaries are held from February to June 2016. Marco Rubio and Ted Cruz drop out in March and May 2016. Bernie Sanders fights until last minute when Democratic National Convention is held in June 2016.

doi:10.6342/NTU201702253

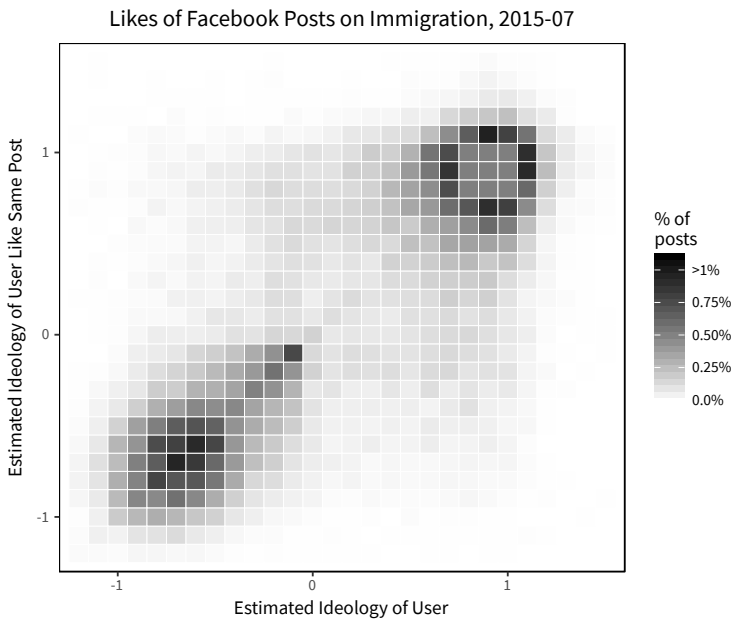


Figure 13: Heatmap of Likes on Posts Related to Immigration

Notes: We randomly select users that likes immigration-related posts, place these people on the x-axis, and then randomly select other users that likes the same post as those on the x-axis. Data used: post created in July 2015. Keywords used to search for posts: immigration, immigrant, immigrants, border. Figures for other months have slight differences but hard to summary and are still quite polarized.

6.2 Post Content Dimension: Echo Chambers

The second strength of our proposed method is that we can dig into the whole universe of post content. Before using some more advanced text analysis techniques, we start by looking at how people across ideologies react to post that conveys different issues.

Figure 13 plots the heat map of likes on immigration-related posts across the political spectrum. We can see the probabilities of two people like the same post are clustered at two like-minded corners. This is perhaps a direct visual evidence of what echo chambers or filter bubbles look like.

Figure 14 plots parallel heat map for Chicago Cubs in October 2016, where they received World Series Champions in Major League Baseball (MLB). Since Illinois is generally a liberal state and MLB fans tend to be more liberal, we can see the likes are clustered around liberal users.

There are still many could be done with respect to text analysis, given the fact that people are of growing interest in how text predicts or determines one's thought (Gentzkow et al. (2016); Kim et al. (2017) are some recent related attempts), and also with the fact that related artistry are closer to their prime time (also see Gentzkow et al. (2017) for an introduction).

doi:10.6342/NTU201702253

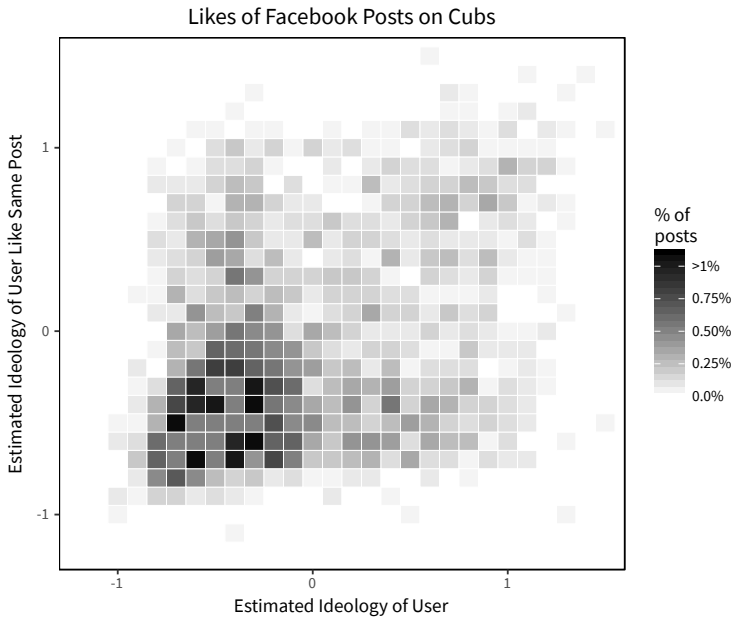


Figure 14: Heatmap of Likes on Posts Related to Chicago Cubs

Notes: We randomly select users that likes related to posts on Chicago Cubs (they received World Series Champions in October 2016), place these people on the x-axis, and then randomly select other users that likes the same post as those on the x-axis. Data used: post created in October 2016. Keywords used to search for posts: Chicago Cubs, Cleveland Indians.

6.3 Forecasting Presidential Election

Another possible use of our data is to forecast elections. As a direct application of spatial model such as Hotelling (1929); Downs (1957), assume that people vote to candidates closer to their own ideological position, given that we can guess the state that user lives in, we can thus calculate the share of users closer to each candidates in each state. Although there may still be bias due to the fact that turnouts would not be uniform across states and some other factors may also affect one's voting decision, this can still be a reasonable forecast for election outcomes.

Figure 15 shows the result using data between 2016-10-01 and 2016-11-07 (the election is held on 2016-11-08). On the x-axis we plot the share of users closer to Hillary Clinton in each state, and on the y-axis we plot the ex-post vote share Hillary Clinton gets. We can see that, rely only on ideology estimates, Facebook data predicts vote share quite well ($\rho = 0.73$).

Also, if we assume that Hillary Clinton wins those states where she is closer to more than 50% of users in term of ideology and Donald Trump wins other states, we can almost get the national election outcome (with some exceptions in some really small states such as Maine, Montana, and Alaska), where Trump gets a total of 292 out of 538 electoral college votes.

Table 4 compares our results with other major election forecasts such as FiveThirtyEight

doi:10.6342/NTU201702253

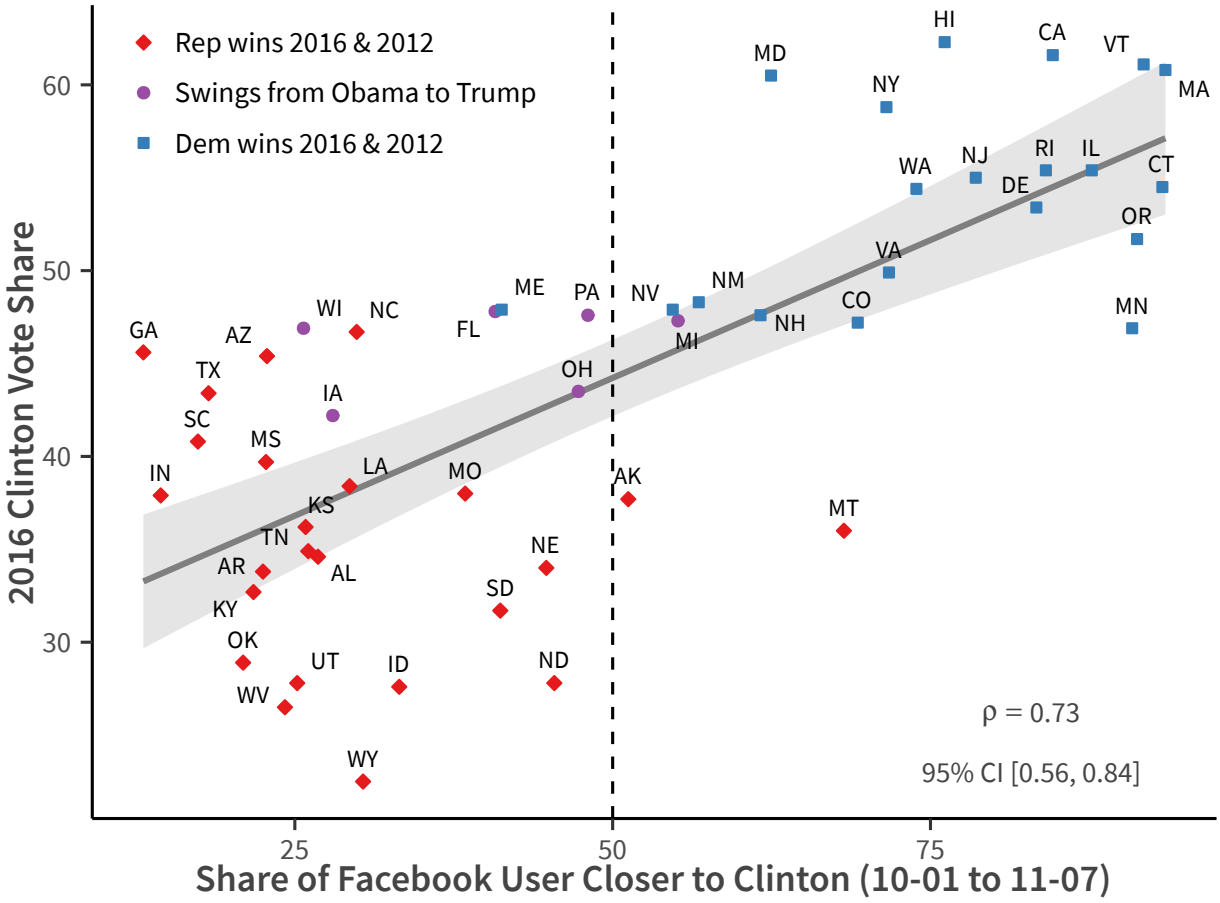


Figure 15: Forecasting 2016 Presidential Election

Notes: States are guessed by the maximum state on likes of national politicians (Sen, Rep, Gov). Data used: 2016-10-01 to 2016-11-07.

doi:10.6342/NTU201702253

State	Electoral Votes	Actual Winner	Facebook	538	NYT	PEC
Wisconsin	10	Trump	○	×	×	×
Iowa	6	Trump	○	○	○	○
Florida	29	Trump	○	×	×	×
Pennsylvania	20	Trump	○	×	×	×
Ohio	18	Trump	○	○	○	○
Michigan	16	Trump	×	×	×	×
Maine	2	Clinton	×	○	○	○
Alaska	3	Clinton	×	○	○	○
Montana	3	Trump	×	○	○	○
Trump's Electoral Vote		306	292	235	216	215

Notes: Here we only list the states where Facebook estimate are wrong or generate different result with other forecasts. Wisconsin, Iowa, Florida, Pennsylvania, Ohio, and Michigan are states swings from Obama to Trump. *Sources*: FiveThirtyEight (2016); The New York Times (2016); Princeton Election Consortium (2016)

Table 4: Election Forecasts Comparison

(2016), The New York Times (2016), and the Princeton Election Consortium (2016). Our result is the most pessimistic for Hillary Clinton and the only one that correctly predicts the rise of President Trump. Furthermore, for most swing states, where the voters swings from Obama to Trump, we correctly predicts the winner, except only Michigan.

6.4 Ideological Segregation at Media Level

The bigger and perhaps the one of the ultimate questions social scientists ought to answer in this decade is, how, indeed, ideologically-segregated platforms like Facebook is, compared with other forms of human interaction?

This is nothing fresh at all. Back when online news starts to dominate the market, people have raised similar concerns (Sunstein 2001). Gentzkow and Shapiro (2011) use web browsing data and self-report ideology to compares ideological segregation on Internet news, TVs, newspapers, and even face-to-face interactions. They suggest people are usually too pessimistic about the online news: The degree of segregation on Internet (0.075) is even smaller than the degree of national newspapers (0.104). They also find no sign of time trend, and on major event dates the index is even relatively lower.

When time goes on, naturally people's concerns have now changed to social media. To have a glance at this issue, we can calculate parallel indexes (see Equation 6.1) as in Gentzkow and Shapiro (2011) for fan pages on Facebook and using GSS-matched self-report ideology

doi:10.6342/NTU201702253

quantiles as proxies for liberals and conservatives.

$$S_m = \sum_{j \in J_m} \left(\frac{\text{cons}_j}{\text{cons}_m} \cdot \frac{\text{cons}_j}{\text{visits}_j} \right) - \sum_{j \in J_m} \left(\frac{\text{lib}_j}{\text{lib}_m} \cdot \frac{\text{cons}_j}{\text{visits}_j} \right) \tag{6.1}$$

That is, for each news outlet j of type m, we can calculate the share of conservative daily visitors (defined by likes) and weight by the relative importance of that page inside the conservative or liberal campaign. Index 0 means that all conservatives and liberals visits the same page, while index 1 means that conservatives only visits all conservative pages, and vise versa.

Results are presented in Figure 16 We find is a high isolation based on likes on Facebook news outlet fan pages. For example, in 2015 newspaper pages, the index is around 0.3, which is close to face-to-face interactions with people you trust, as described by Gentzkow and Shapiro (2011). However, our estimates are based on likes, not views, so we cannot make a direct comparison with their result.[1]

Other than the number difference, we also observe an increasing trend of isolation for most pages, partly due to the fact that election is held in 2016 (but notice that such trend seems to start in 2015). This is not consistent with their findings that isolation even lowers on the day near the election.

6.5 Opinion Segregation at Issue Level

We can also calculate parallel index for different issues by introducing post content dimension. Figures 17 and 18 present the result. We can make the following observations. First, soft issues (kids, pets, sports) generally have lower segregation than hard issues (election, immigration, healthcare). Second, some events may indeed (eg. Russia Hackers intervene the Democratic National Committee email servers in June 2016) trigger higher isolation.

Last but not least, segregation in most hard issues does not change much over time, if not having some slight decreasing trend (eg. healthcare, environment, inequality). This is a striking difference between observations in news outlet level, where we can see a clear rising trend.

An interesting interpretation could be that although people tend to look at contents they like, and media are smart enough to attract their potential supporters, this does not change

[1] We also tried using party affiliation as defined in 5.3. We can see this as a lower bound for index using likes since this almost cuts people into two connected parts so that they have a higher probability to get exposed to each other.

doi:10.6342/NTU201702253

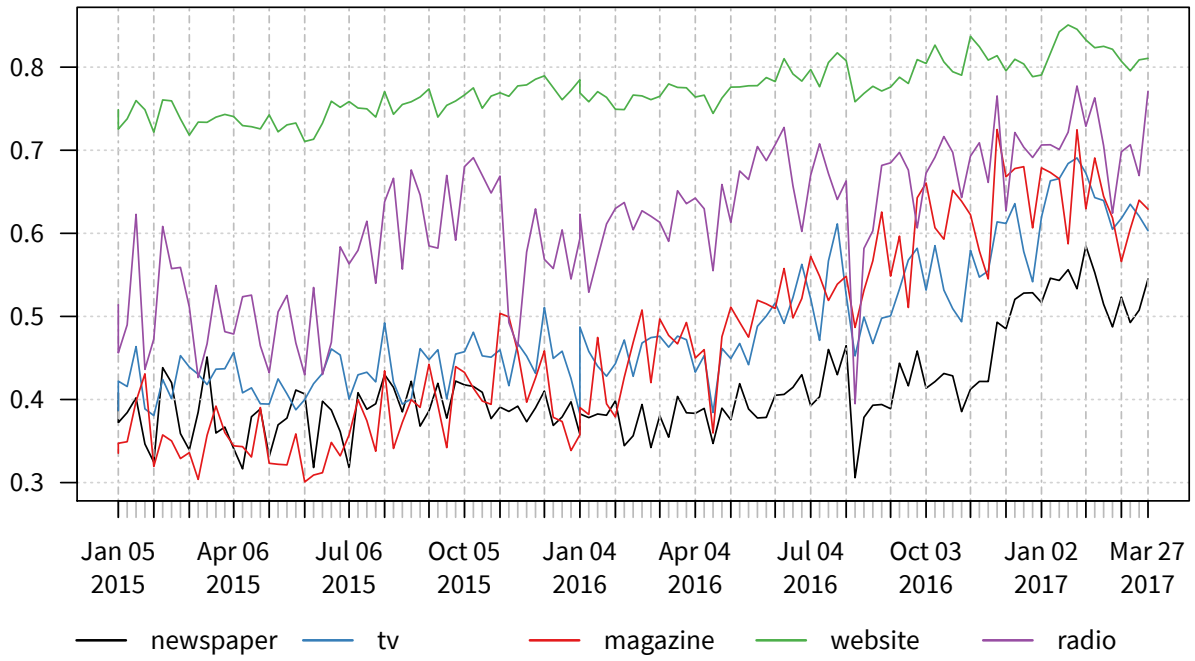


Figure 16: Isolation Index of Likes on Facebook News Outlet Pages

Notes: We use isolation index defined in Gentzkow and Shapiro (2011). Data used: 2015-01-01 to 2017-03-31.

people's opinion on various issues that much. It is worth revisiting these trends on other forms of communications in order to generate more meaningful interpretation of the role of social media on public discussions.

6.6 Potential Causes of Segregation

What causes these possible difference in terms of segregation between online news consumption (as in Gentzkow and Shapiro (2011)) and social media news consumption? A possible reason could be that market structures differ. On Internet news, news websites are more vertically differentiated (remember the days when Yahoo! is still a large portal site and people often read the news there?), where a large amount of traffic goes to few centrist news outlets. As calculated in (Gentzkow and Shapiro 2011, Figure 5), top 1 news sites (Yahoo! News) owns 20% of traffics, and top 20 news sites take over nearly 80% of the total news view traffics.

We calculate parallel figures based on likes, which is shown in Figure 19. If we can take likes as proxies for views, top 1 news fan page (Fox News) only owns 0.5% of likes, and top 20 news fan pages only take over 30%. News views on Facebook are perhaps far more horizontally differentiated than online news views. What even possibly intensifies segregation is the fact that few of these fan pages could be considered as a centrist.

doi:10.6342/NTU201702253

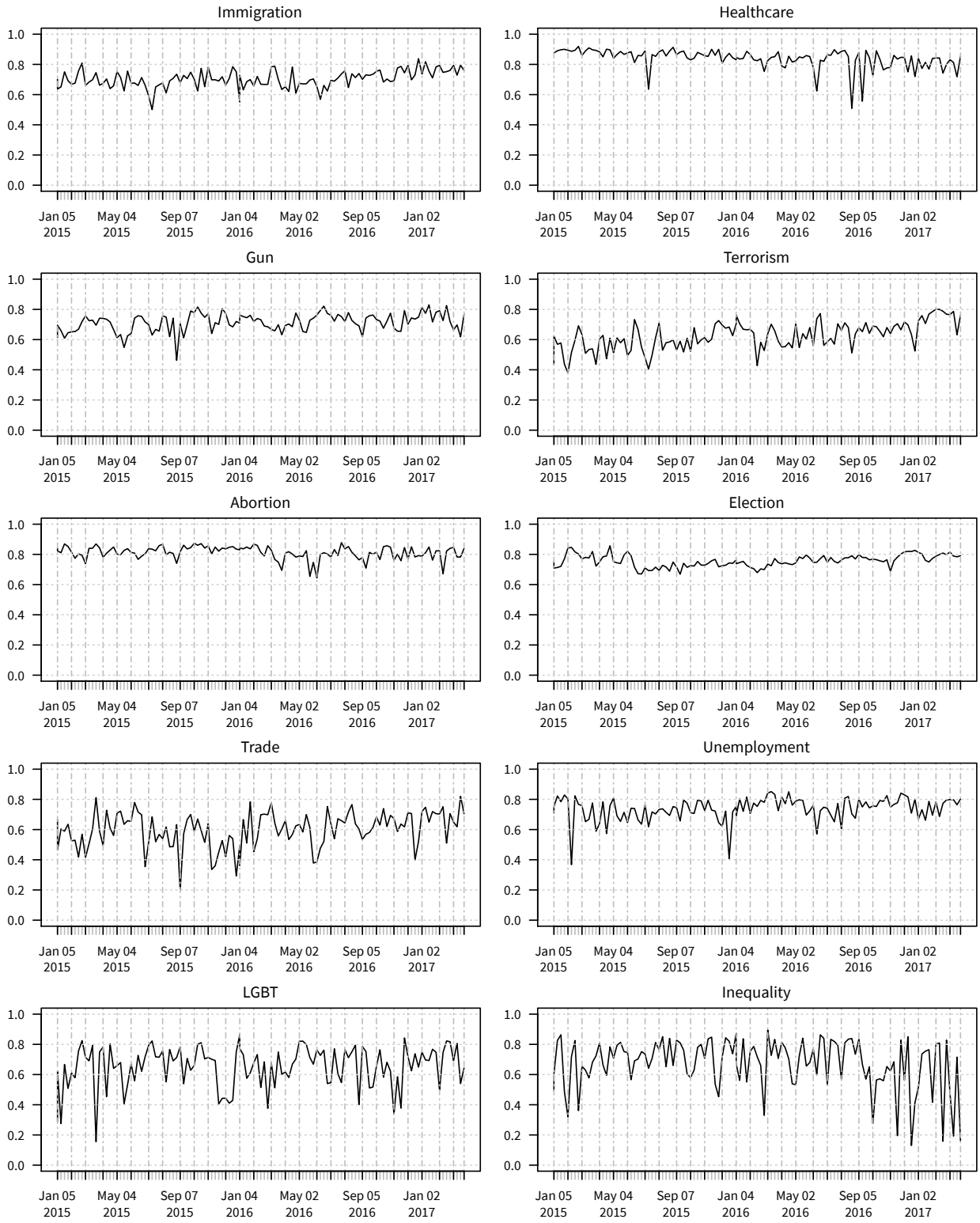


Figure 17: Isolation Index of Likes by Issue, I

Notes: We use isolation index defined in Gentzkow and Shapiro (2011) and replace "media" levels to "issue" levels. Data used: 2015-01-01 to 2017-03-31.

doi:10.6342/NTU201702253

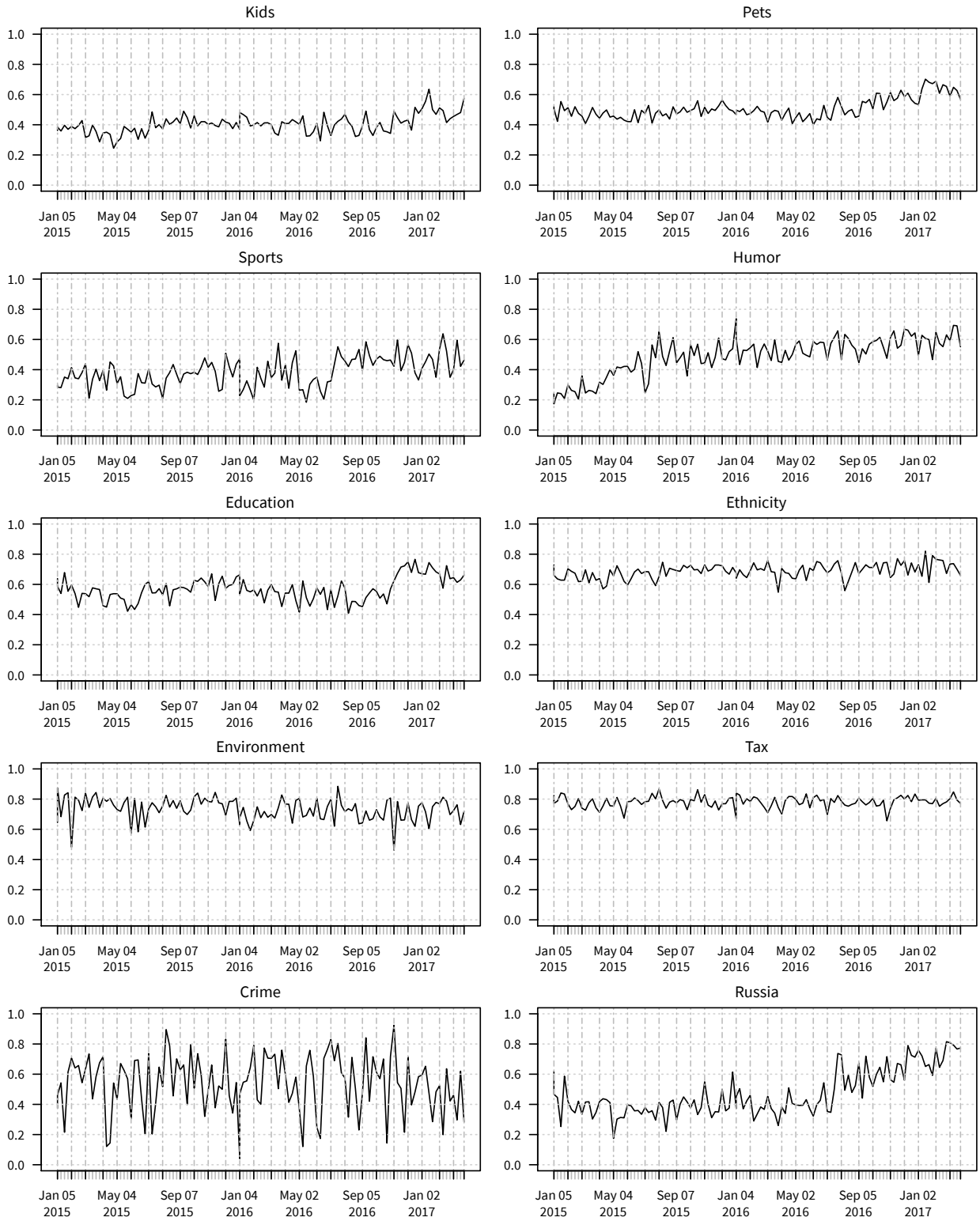


Figure 18: Isolation Index of Likes by Issue, II

Notes: We use isolation index defined in Gentzkow and Shapiro (2011) and replace "media" levels to "issue" levels. Data used: 2015-01-01 to 2017-03-31.

doi:10.6342/NTU201702253

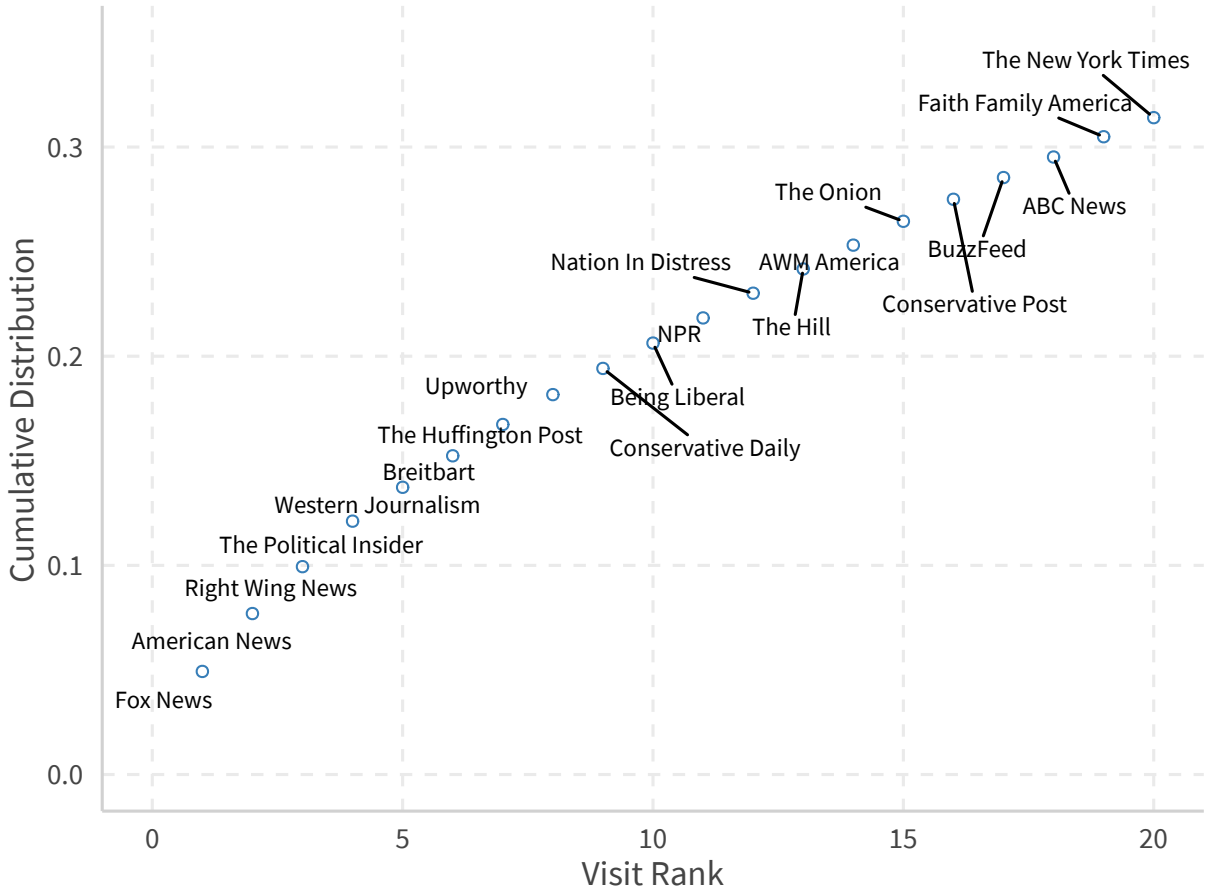


Figure 19: Cumulative Distribution of News Page Likes

Notes: Parallel to (Gentzkow and Shapiro 2011, Figure 5), we only include daily unique visitors. Data used: 2015-01-01 to 2017-03-31.

doi:10.6342/NTU201702253

6.7 Promise and Pitfalls of Social Media

The convenient environment Facebook creates makes connecting easier. This also reduces the cost to find people that are similar to you. The environment is also efficient: users feel happy, and the social media giant gets traffic along with the advertisement. Such trend regarding a more personalized experience to acquire information may be irreversible. The horizontally-differentiated feature, compared with traditional Internet world, also prone to make people stuck inside bubbles echoed with like-minded views. This could be detrimental to a society since people are more likely to form false beliefs, while beliefs shape human behavior. Our paper serves as some attempts to open this black box, using only data that are open to anyone interested in these challenging problems.

doi:10.6342/NTU201702253

References

American National Election Studies. 2017. "American National Election Study." http://www.electionstudies.org/nesguide/toptable/tab3_1.htm.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91. http://dx.doi.org/10.1093/pan/mpu011.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26 (10): 1531–1542. http://dx.doi.org/10.1177/0956797615594620.

Bauer, Paul C., Pablo Barberá, Kathrin Ackermann, and Aaron Venetz. 2016. "Is the Left-Right Scale a Valid Measure of Ideology?" *Political Behavior*: 1–31. http://dx.doi.org/10.1007/s11109-016-9368-2.

Bond, Robert M., and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109 (1): 62–78. http://dx.doi.org/10.1017/s0003055414000525.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58 (2): 367–386. http://dx.doi.org/10.1111/ajps.12062.

Bonica, Adam. 2016. "Inferring Roll Call Scores from Campaign Contributions Using Supervised Machine Learning." *Working Paper*. https://ssrn.com/abstract=2732913.

Calero, Antonio. 2013. "Likes vs. Comments vs. Shares." http://www.antoniocalero.com/2013/05/06/facebook-likes-comments-shares/.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–370. http://dx.doi.org/10.1017/S0003055404001194.

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy* 65 (2): 135–150. http://dx.doi.org/10.1086/257897.

Eckart, Carl, and Gale Young. 1936. "The Approximation of One Matrix by Another of Lower

doi:10.6342/NTU201702253

Rank." *Psychometrika* 1(3): 211–218.
http://dx.doi.org/10.1007/BF02288367.

FiveThirtyEight. 2016. "Who Will Win the Presidency?."
https://projects.fivethirtyeight.com/2016-election-forecast/.

Gallup. 2017. "US Conservatives Outnumber Liberals by Narrowing Margin." January.
http://www.gallup.com/poll/201152/conservative-liberal-gap-continues-narrow-tuesday.aspx.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. London, UK: Chapman & Hall/CRC.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." Working Paper 23276, National Bureau of Economic Research.
http://dx.doi.org/10.3386/w23276.

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. "Ideological Segregation Online and Offline." *Quarterly Journal of Economics* 126(4): 1799–1839.
http://dx.doi.org/10.1093/qje/qjr044.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2016. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." Working Paper 22423, National Bureau of Economic Research.
http://www.nber.org/papers/w22423.

Gervais, Will M., and Maxine B. Najle. 2017. "How Many Atheists Are There?" *Social Psychological and Personality Science* forthcoming.
http://dx.doi.org/10.1177/1948550617707015.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120(4): 1191–1237.
http://dx.doi.org/10.1162/003355305775097542.

Heckman, James J., and James M. Snyder, Jr. 1997. "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators." *RAND Journal of Economics* 28: S142–S189.
http://www.jstor.org/stable/3087459.

Hotelling, Harold. 1929. "Stability in Competition." *Economic Journal* 39(153): 41–57.
http://dx.doi.org/10.2307/2224214.

Jessee, Stephen A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103(1): 59–81.
http://dx.doi.org/10.1017/S000305540909008X.

Kim, In Song, John Londregan, and Marc Ratkovic. 2017. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* forthcoming.
http://web.mit.edu/insong/www/pdf/sfa_pa.pdf.

National Opinion Research Center. 2017. "General Social Surveys."
https://gssdataexplorer.norc.org/variables/178/vshow.

Pew Research Center. 2014. "Political Polarization and Media Habits."

doi:10.6342/NTU201702253

http://www.journalism.org/2014/10/21/political-polarization-media-habits/.

Pew Research Center. 2016a. "News Use Across Social Media Platforms 2016." http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

Pew Research Center. 2016b. "Social Media Update 2016." http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

Pew Research Center. 2016c. "The Political Environment on Social Media." http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/.

Poole, Keith T., and Howard L. Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–384. http://www.jstor.org/stable/2111172.

Poole, Keith T., and Howard L. Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Poole, Keith T., and Howard L. Rosenthal. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.

Princeton Election Consortium. 2016. "Final Projections: Clinton 323 EV, 51 Democratic Senate Seats, GOP House." http://election.princeton.edu/2016/11/08/final-mode-projections-clinton-323-ev-51-di-senate-seats-gop-house/.

Schiffer, Adam J. 2000. "I'm Not That Liberal: Explaining Conservative Democratic Identification." *Political Behavior* 22 (4): 293–310. http://dx.doi.org/10.1023/A:1010626029987.

Sunstein, Cass R. 2001. *Republic.com*. Princeton: NJ: Princeton University Press.

The New York Times. 2016. "Who Will Be President?." https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html.

U.S. Census Bureau. 2017. "State Population Totals Tables: 2010-2016." https://www.census.gov/data/tables/2016/demo/popest/state-total.html.

doi:10.6342/NTU201702253

Appendix A

Further Results

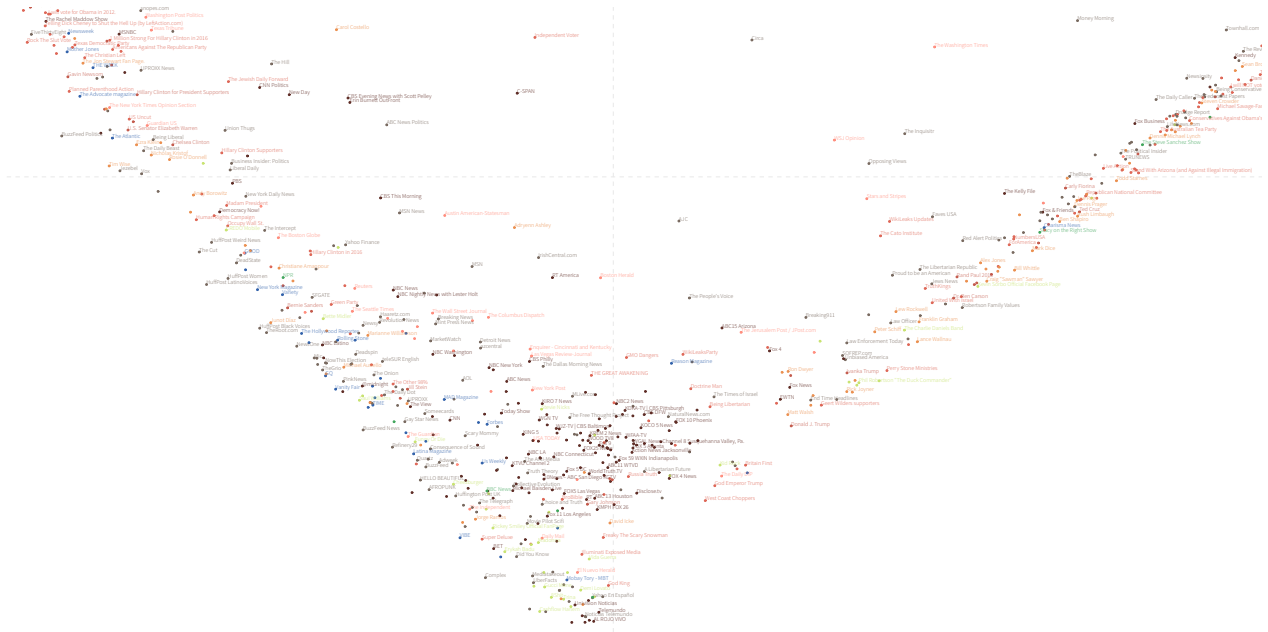


Figure 20: Scatter Plot on the First and Second Dimension (Part)

Notes: First dimension on the x-axis and second dimension on the y-axis. Colors correspond to different page types. Data used: 2015-01-01 to 2016-11-07.

doi:10.6342/NTU201702253

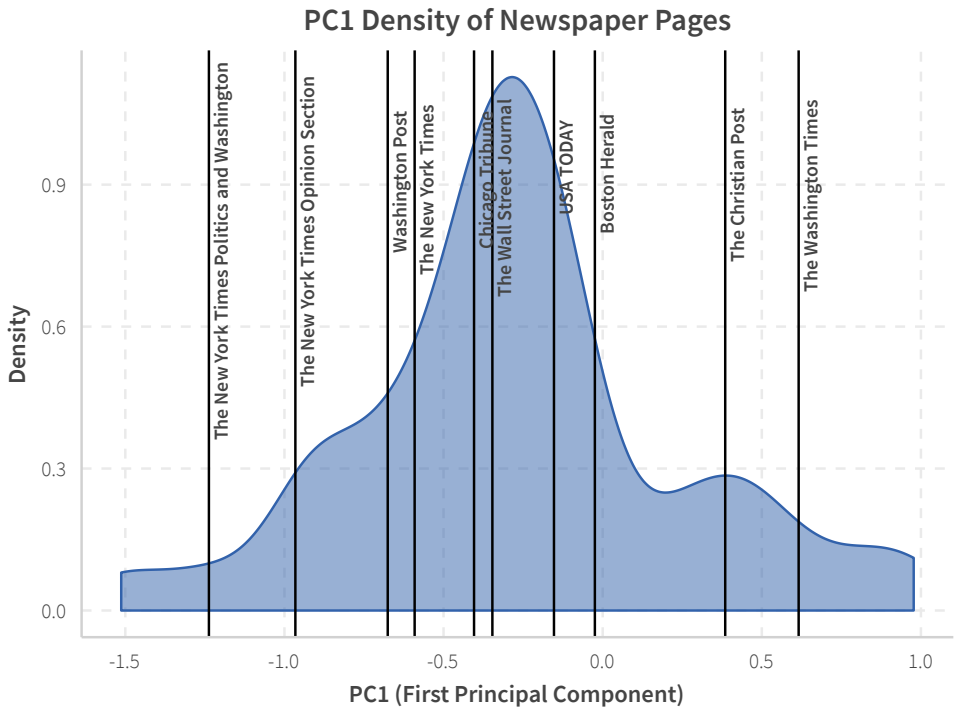


Figure 21: Density for Newspaper Pages

Notes: Data used: 2016-10-01 to 2016-11-24.

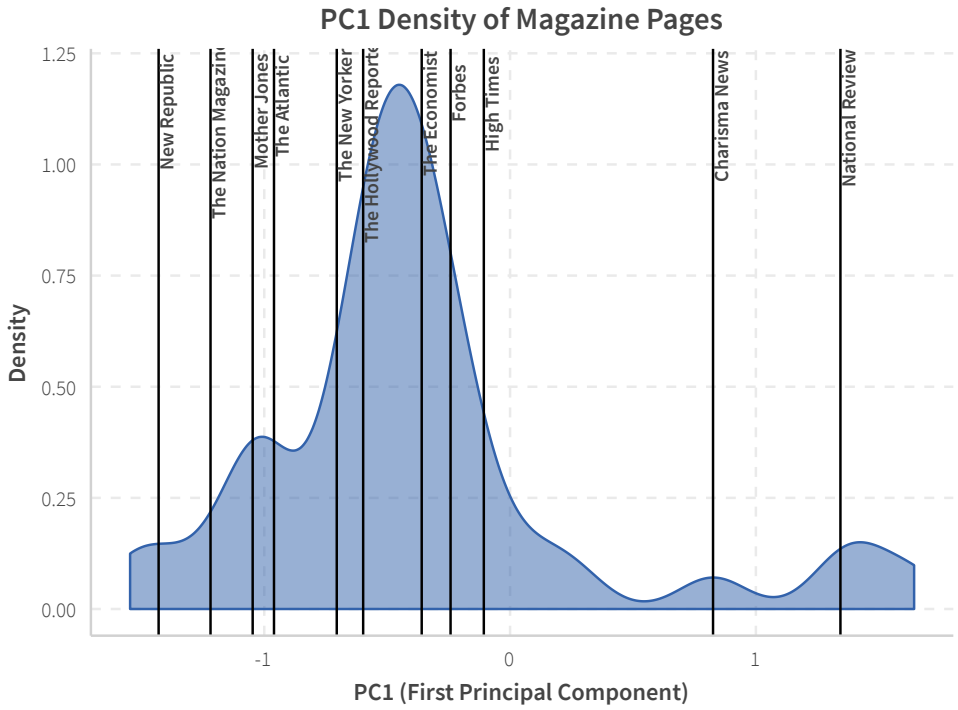


Figure 22: Density for Magazine Pages

Notes: Data used: 2016-10-01 to 2016-11-24.

doi:10.6342/NTU201702253

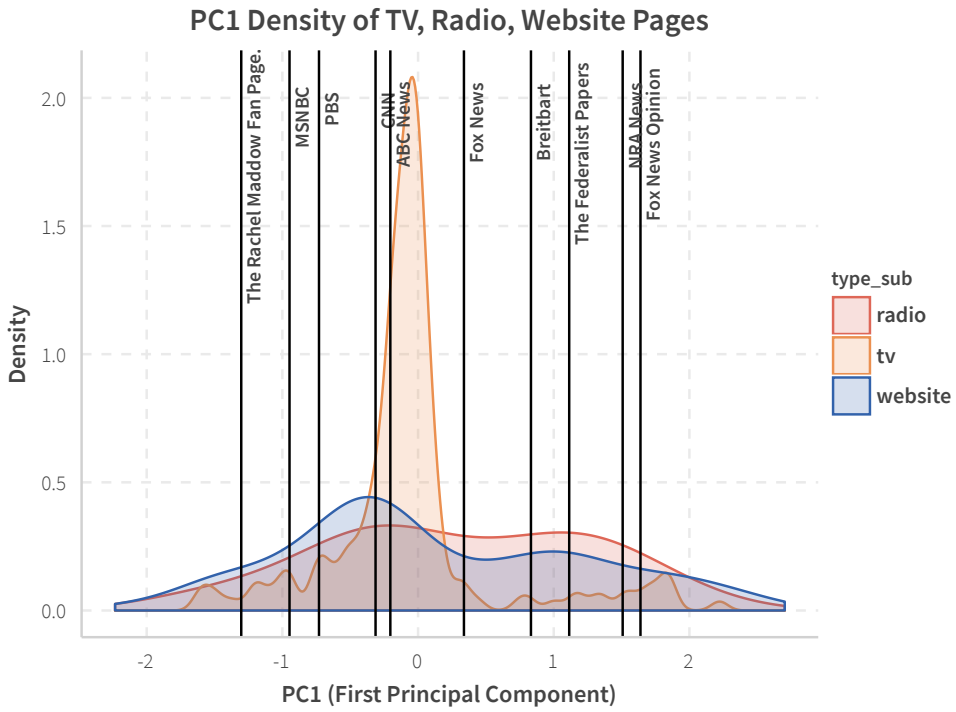


Figure 23: Density for TV, Radio, and Website Pages

Notes: Data used: 2016-10-01 to 2016-11-24.

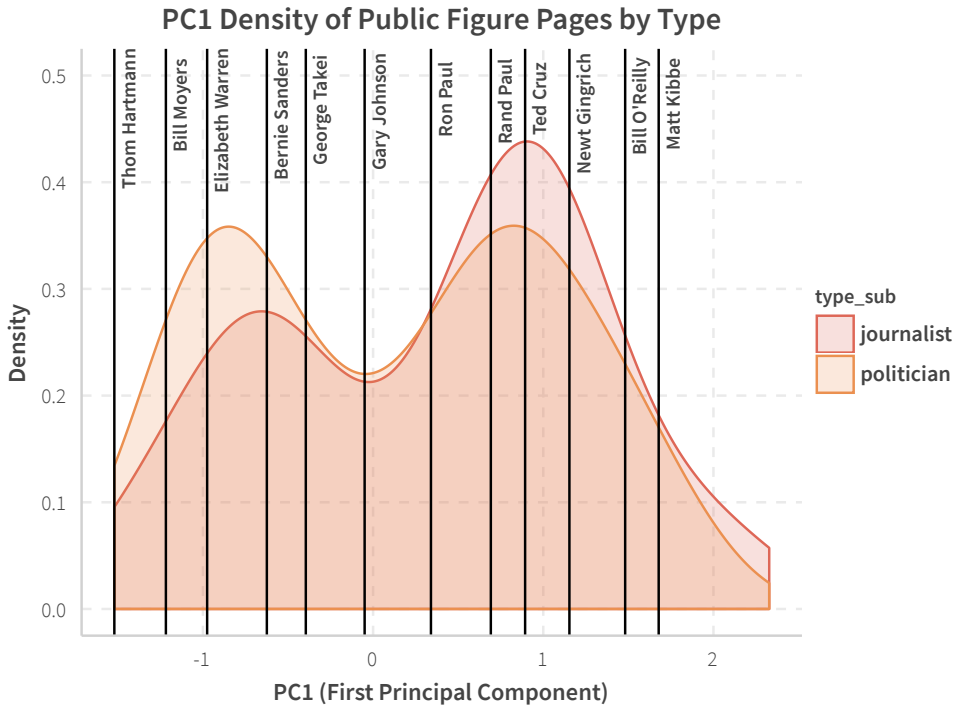


Figure 24: Density for Public Figure Pages

Notes: Data used: 2016-10-01 to 2016-11-24.

doi:10.6342/NTU201702253

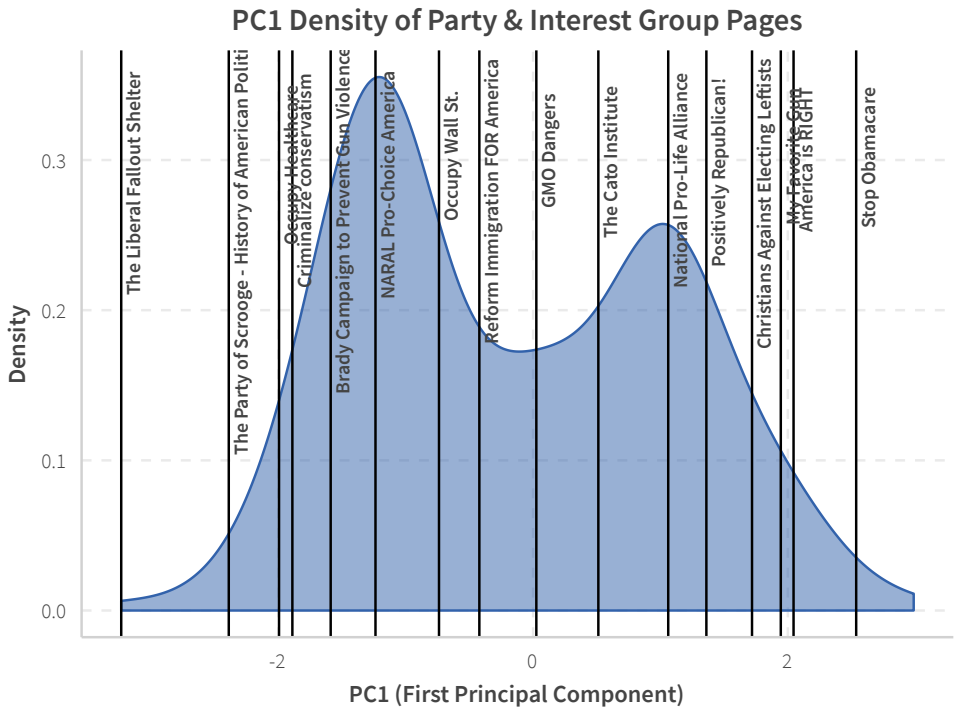


Figure 25: Density for Interest Group Pages

Notes: Data used: 2016-10-01 to 2016-11-24.

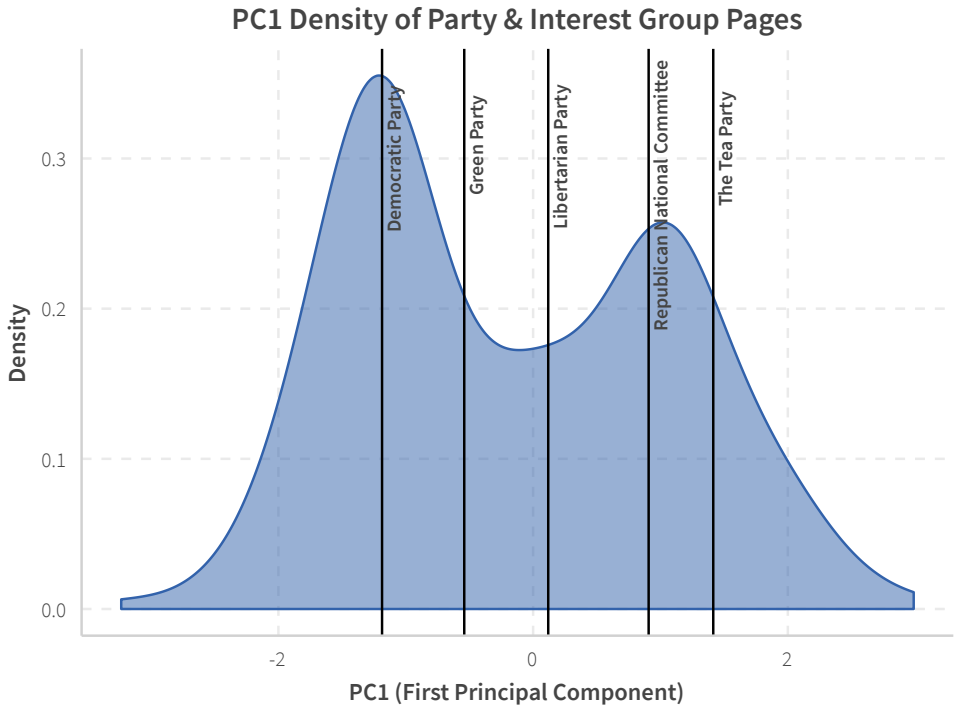


Figure 26: Density for Party Pages

Notes: Data used: 2016-10-01 to 2016-11-24.

doi:10.6342/NTU201702253

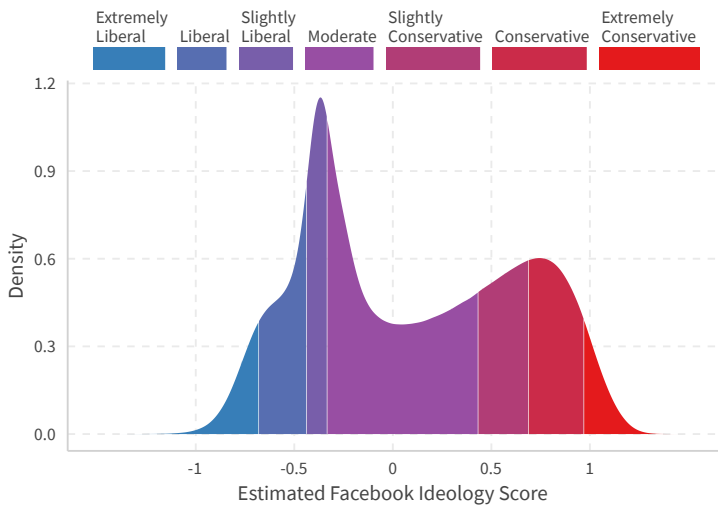


Figure 27: Density for Active US Users (>10) with Self-Report Ideology Shares in GSS

Notes: Colors represent matching densities with self-reported ideology shares in 2016 General Social Surveys (National Opinion Research Center 2017). US users are defined by any user that at least reacted to any national politicians' (Sen, Rep, Gov) post once in 2015 and 2016, and we guess user's location by the maximum national politician they liked in that state. We remove a huge jump created by users only like one and only one page: Arnold Schwarzenegger. We then sample users by 2016 population in each state (U.S. Census Bureau 2017) if that state is overrepresented in our sample relative to the population.

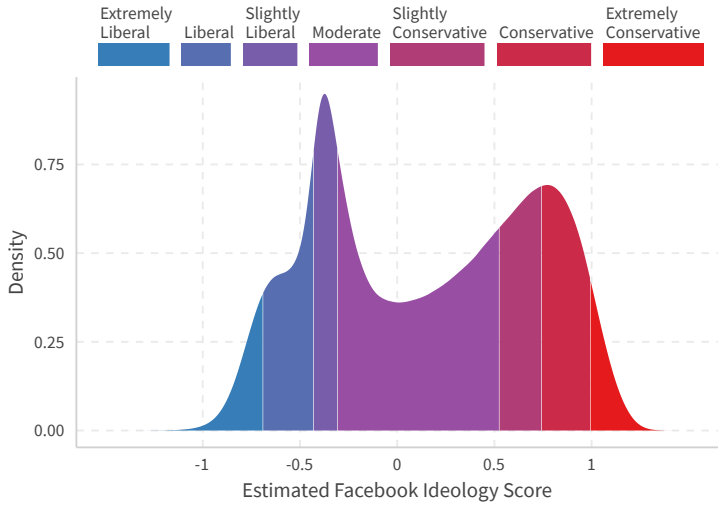


Figure 28: Density for Active US Users (>20) with Self-Report Ideology Shares in GSS

Notes: Colors represent matching densities with self-reported ideology shares in 2016 General Social Surveys (National Opinion Research Center 2017). US users are defined by any user that at least reacted to any national politicians' (Sen, Rep, Gov) post once in 2015 and 2016, and we guess user's location by the maximum national politician they liked in that state. We remove a huge jump created by users only like one and only one page: Arnold Schwarzenegger. We then sample users by 2016 population in each state (U.S. Census Bureau 2017) if that state is overrepresented in our sample relative to the population.

doi:10.6342/NTU201702253

Appendix B

Further Validations

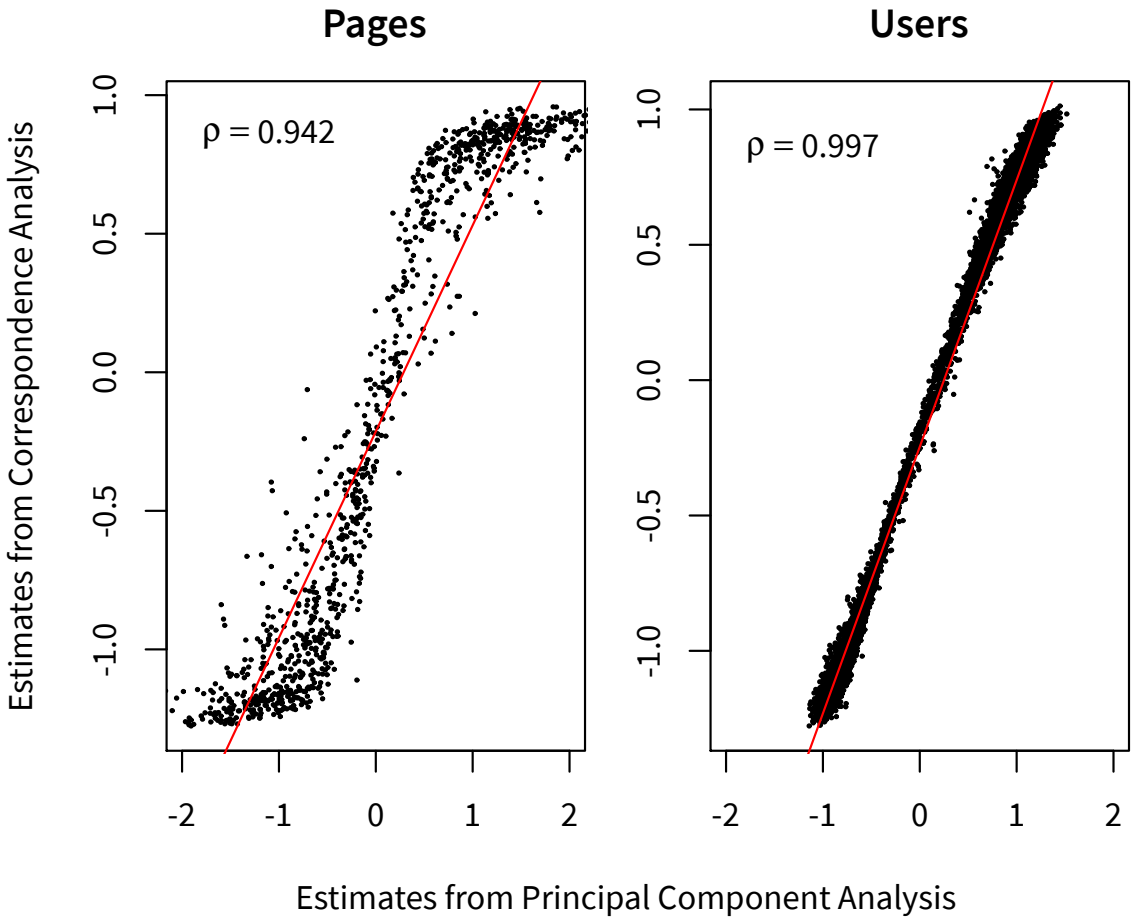


Figure 29: Estimation using PCA vs. CA (Barberá 2015)

Notes: Since CA needs to decompose a user by page matrix, which needs extremely large computer memory, here I conducted CA using users likes more than 70 pages (76,585 users) and pages owns more than 10,000 fans (1027 pages).

doi:10.6342/NTU201702253

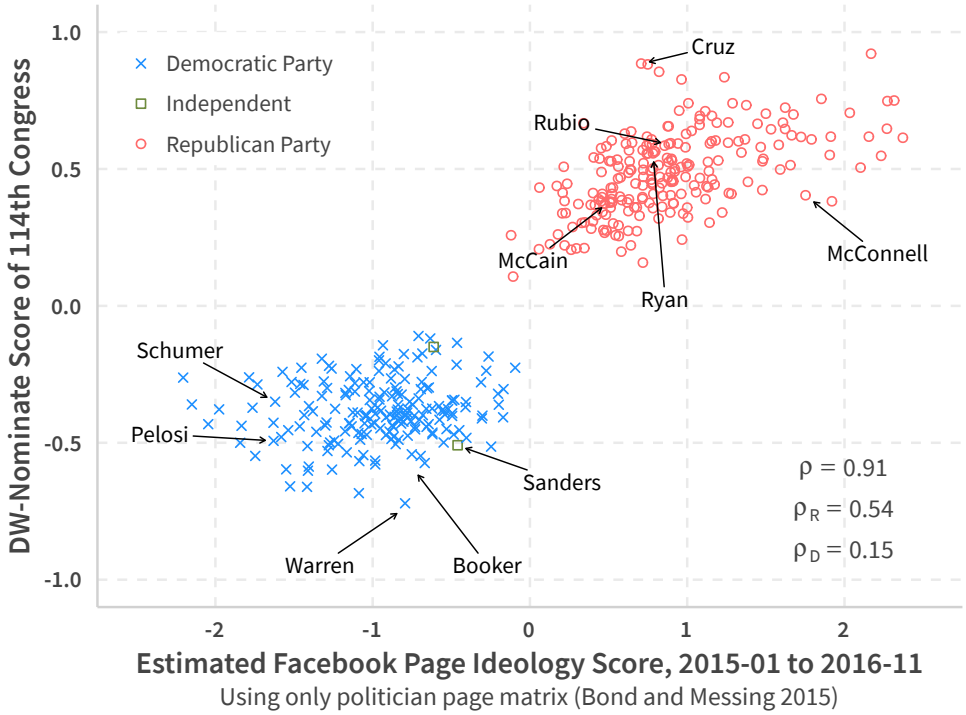


Figure 30: DW-Nominate vs. FB Estimate (Bond and Messing (2015), 114 Congress)

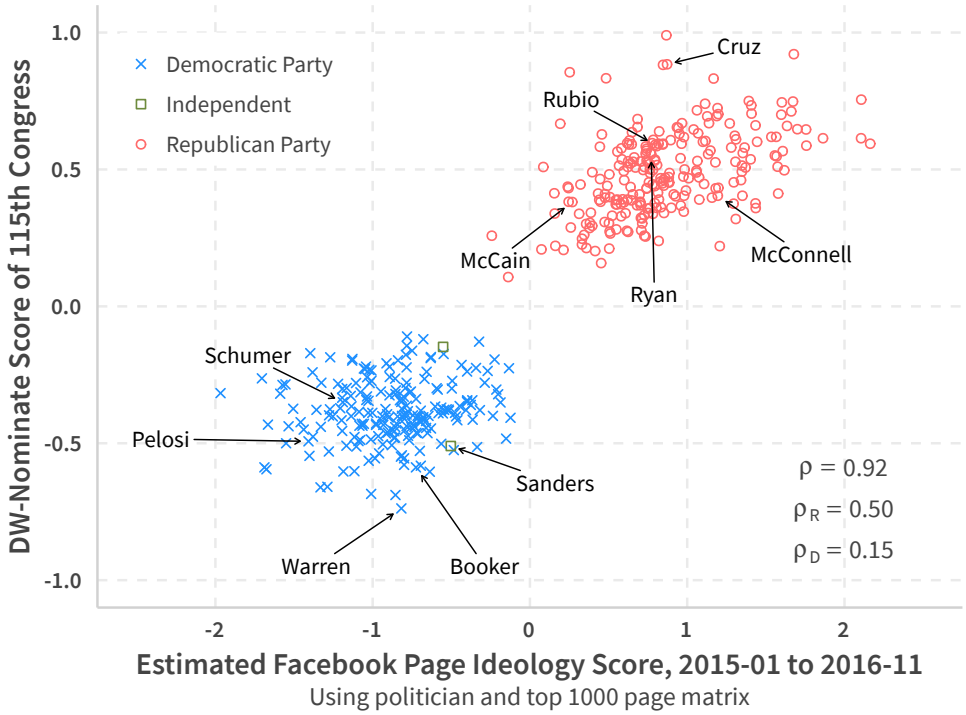


Figure 31: DW-Nominate vs. FB Estimate (115 Congress)

doi:10.6342/NTU201702253

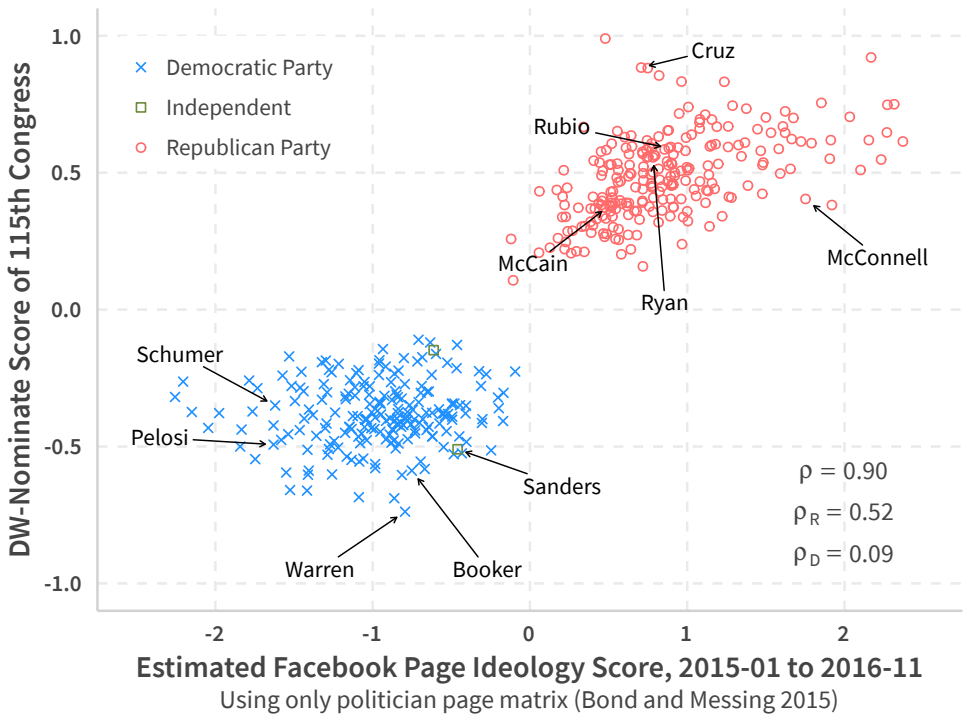


Figure 32: DW-Nominate vs. FB Estimate (Bond and Messing (2015), 115 Congress)

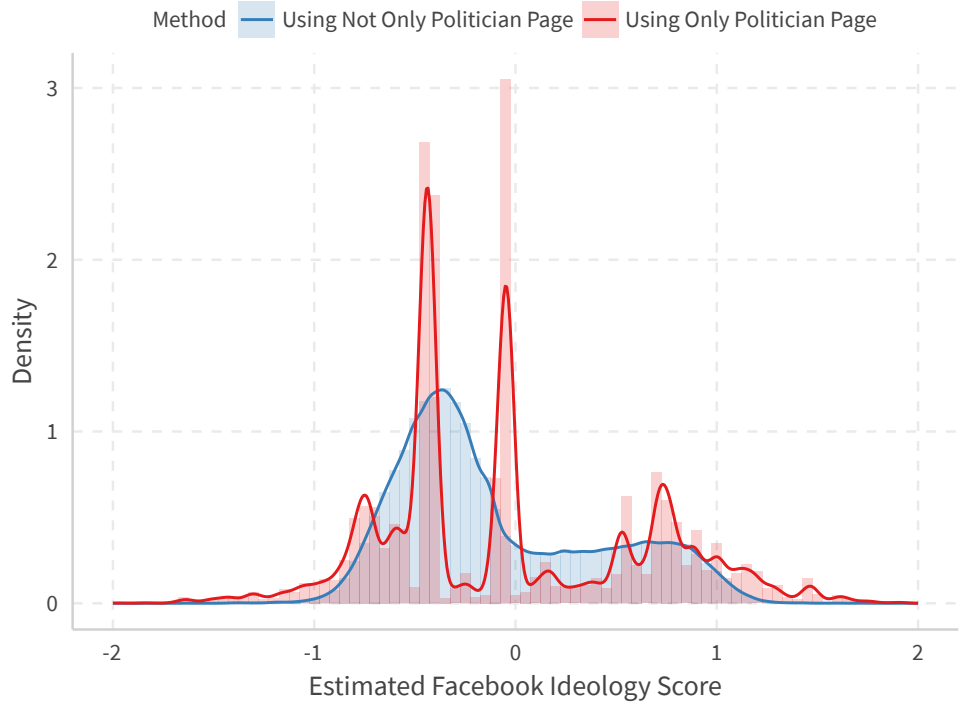


Figure 33: User Density by Politician-Only (Bond and Messing 2015) vs. Our Method

Notes: Blue region represents the method used in our paper. Red region uses the procedure suggested by Bond and Messing (2015) where one only considers politician fan pages and calculate user ideology accordingly. We remove a huge jump created by users only like one and only one page: Arnold Schwarzenegger.

doi:10.6342/NTU201702253

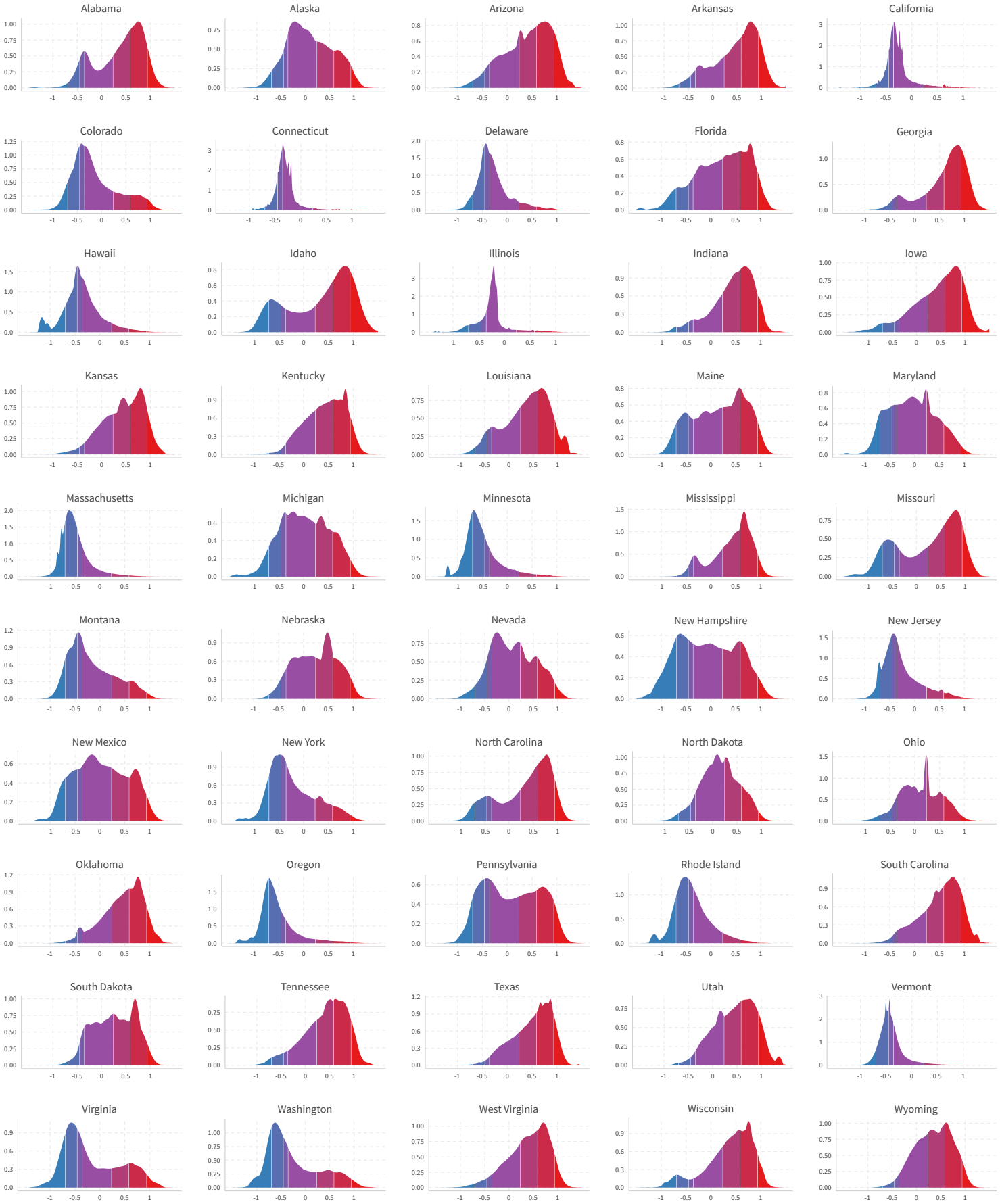


Figure 34: User Densities by 50 States with National Ideology Shares in GSS

Notes: States are guessed by the maximum state on likes of national politicians (Sen, Rep, Gov). Colors represent matching densities with self-reported ideology shares in 2016 General Social Surveys (National Opinion Research Center 2017). We remove a huge jump created by users only like one and only one page: Arnold Schwarzenegger.

doi:10.6342/NTU201702253