DOI: 10.1177/0956797615594620

SUPPLEMENTARY MATERIALS
Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?

Contents

DOI: 10.1177/0956797615594620

1 Real Time Estimation of Social Media Users' Ideological Positions

1.1 The Statistical Model

We consider ideology as a position on a latent multidimensional space that can be inferred by observing Twitter users' following decisions (Barberá, 2015). The key assumption in our approach is that individuals prefer to follow political actors (legislators, candidates, media outlets, think tanks...) that they perceive to be "close" on this latent space. This assumption can be understood as a manifestation of ideological homophily (McPherson et al., 2001) if we treat Twitter as a social networking site, and also as the result of users' selective exposure to ideologically congenial content (Lazarsfeld et al., 1944; Bryant and Miron, 2004), if we consider Twitter as a news media (Kwak et al., 2010). Our approach is also similar in nature to other measurement methods that rely on spatial voting assumptions (Enelow and Hinich, 1984; Poole and Rosenthal, 1997; Jessee, 2009; Bonica, 2013a).

Suppose that each Twitter user $i \in \{1, \ldots, n\}$ is presented with a choice between following or not following another target user $j \in \{1, \ldots, m\}$, where j is a political actor who has a Twitter account in social media network $\mathbf{g}$. Let $Y_{ij} = 1$ if user i decides to follow actor j, and $Y_{ij} = 0$ otherwise. We expect this decision to be a function of d_{ij}, the distance in the latent ideological dimension between user i and political account j. To this core model, we add two additional parameters to account for user- and actor-random effects, α_i and β_j. The former accounts for the different levels of political interest of user i ("out degree"), while the latter measures the popularity of actor j ("in degree").

The probability that user i follows a political account j is then formulated as a logit model:

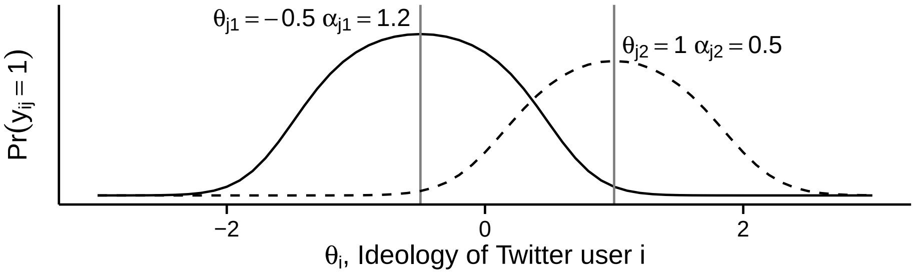
$$\Pr(Y_{ij} = 1 | \alpha_i, \beta_j, d_{ij}) = \text{Logit}(\alpha_i + \beta_j - d_{ij}) \tag{1}$$

where d_{ij} is the Euclidean distance between θ_i, the ideological position of user i, and θ_j, the ideological position of political account j. Figure S1 illustrates how our model works for hypothetical values of our parameters. Assume that we're interested in the probability that a user i follows Barack Obama or Mitt Romney, at different values of θ_i, fixing all other parameters to their means. Liberal Twitter users are more likely to follow Barack Obama, and this probability is maximized when their ideology equals the estimated ideology for Barack Obama ($\theta_i = \theta_{j1}$, and therefore $d_{ij} = 0$)[1]. The same logic applies to the estimated probability of following Mitt Romney, but in this case the predicted probability when $\theta_i = \theta_{j2}$ is lower because the popularity parameter for Romney's Twitter account is likely to be smaller.

[1] Note that, unlike in the standard item-response theory models, the probability of a positive outcome is not monotonically increasing or decreasing in ideology. On the contrary, it is decreasing as the distance between users i and j increases. Continuing with the example, this model is consistent with the intuition that extremely liberal individuals are less likely to follow Barack Obama because they do not view him as "liberal enough".

DOI: 10.1177/0956797615594620

Figure S1: Estimated probability that a given Twitter user i follows Barack Obama (j_1) or Mitt Romney (j_2), as a function of the user's ideal point, for hypothetical values



1.2 Estimation Using Correspondence Analysis

There are two key challenges in the estimation of this model: first, finding a computationally efficient method of fitting the model; second, choosing the set of m political actors that allow us to identify the latent multidimensional space as "ideology." This section describes our approach to deal with these two issues.

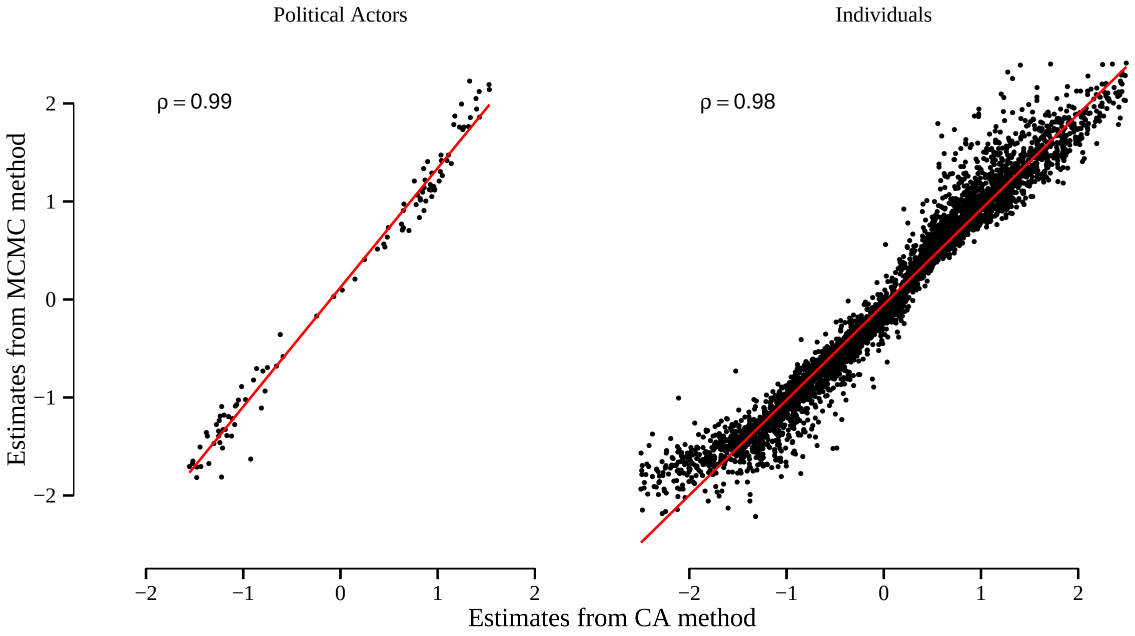
The large number of parameters that need to be estimated (in the order of millions, as we explain below) require developing estimation techniques that can be computationally efficient even at this scale. Latent space models are usually estimated using Markov-Chain Monte-Carlo methods, since standard maximum likelihood methods are intractable for medium to large networks. However, Bayesian methods become computationally inefficient for the type of large-scale networks that we find on social media sites. Instead, we use correspondence analysis (Greenacre, 1984, 2010), adapting its implementation in the `ca` package for `R` (Nenadic and Greenacre, 2007). As Lowe (2008) shows, this method is mathematically close to a log-linear latent space model. Figure S2 demonstrates that this is also the case in our application. Here, we compare estimates using the MCMC method employed by Barberá (2015) and using the method we describe below for a random sample of 500 political actors and 5,000 individuals. Both sets of estimates are very highly correlated.

Correspondence analysis considers $\mathbf{Y}$, the $n \times m$ adjacency matrix indicating whether user i (row) follows user j (column), as a representation of a set of points in a multidimensional space. This matrix is converted into the correspondence matrix $\mathbf{P}$ by dividing by its grand total, $\mathbf{P} = \mathbf{Y}/\sum_{ij} y_{ij}$, and used to compute the matrix of standardized residuals, $\mathbf{S}$, where $\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{P}-\mathbf{rc}^T)\mathbf{D}_c^{1/2}$, where $\mathbf{r}$ and $\mathbf{c}$ are the row and column masses, with $r_i = \sum_j p_{ij}$ and $c_j = \sum_j p_{ij}$, which are then used to construct the diagonal matrices $\mathbf{D}_r = \text{diag}(\mathbf{r})$ and $\mathbf{D}_c = \text{diag}(\mathbf{c})$. As described in Bonica (2013b), this step is equivalent to including the random effects α_i and β_j in the estimation. $\mathbf{S}$ is therefore a matrix of residuals between the observed and expected values based on the marginal distribution of the following matrix $\mathbf{Y}$; and correspondence analysis will scale the rows and columns under the assumption that these deviations respond to the distance between them on a latent multidimensional space.

The main step of the computational algorithm is to calculate the singular value decompo-

DOI: 10.1177/0956797615594620

Figure S2: Comparing Estimates Using MCMC Methods and Correspondence Analysis



sition of $\mathbf{S}$, such that $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathbf{T}$ where $\mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{V}^\mathbf{T}\mathbf{V} = \mathbf{I}$. Once we have identified the low-dimensional solution using SVD, we then project all rows and columns onto that plane by computing the standard coordinates: $\mathbf{\Psi} = \mathbf{D}_r^{1/2}\mathbf{U}$ for rows (ordinary users) and $\mathbf{\Gamma} = \mathbf{D}_c^{1/2}\mathbf{V}$ for columns (political accounts).

One particular advantage of correspondence analysis, which will be particularly useful in our application, is the possibility of projecting supplementary points onto the same subspace. In other words, it is possible to run this method once with a subset of $\mathbf{Y}$, and then estimate the positions of political actors and ordinary users as if they had been included in the original estimation. In order to do so, e.g. for a political actor, we take the column $\mathbf{h}$ of length n indicating what users follow her, standardize it to $\mathbf{h}' = \mathbf{h}/\sum_i h_i$, and then project it to compute $\mathbf{g} = \mathbf{h}'^T\mathbf{\Psi}$, where $\mathbf{g}$ will be the position of that actor on the latent multidimensional space.

The second challenge in implementing this method is the choice of the m target Twitter users: the set of users with “discriminatory” predictive power such that the decision to follow them (or not) provides sufficient information about an individual’s ideology. This decision presents the following trade-off: choosing only political accounts with a clear ideological profile (legislators, candidates, think tanks...) facilitates the identification of the ideological subspace, but limits the number of sample of users for which ideology can be estimated, since many users do not follow politicians on Twitter; but expanding this subset of accounts might give the latent dimensions a different interpretation, as “homophilic” networks emerge based not only on political traits, but also as a result of many other types similarities.

Our approach combines these two options by dividing the estimation in three stages:

DOI: 10.1177/0956797615594620

1. **Identifying the relevant ideological subspace**. First, we narrow down our set of m target accounts to political accounts with high ideological discrimination: legislators, president, candidates, media outlets, interest groups, etc. We compute the model only with users who are highly interested in politics (follow 10 or more of these accounts), since we expect their following decisions to be more informative about the ideological locations of these political accounts. This first step allows us to find the subspace that most closely matches with conventional measures of ideology.

2. **Expanding the number of political accounts,** m. In the second step, we identify the most popular accounts among liberal and conservative users in the first stage (at the top and bottom 20% of the ideological distribution) that were not previously included in the analysis. In particular, the index of popularity we compute is $\text{pop}_{jc} = n_{jc} - n_{jl}$ for conservatives, where n_{jc} is the number of conservative users included in the first stage that follow account j, and n_{jl} is the equivalent measure for liberals. Our assumption here is that the decision to follow one of these accounts also contains information about users' ideology, even if some of these belong to non-political figures. We then add these political accounts to the $\mathbf{Y}$ matrix and estimate their ideological positions by projecting these additional columns onto the ideological subspace.

3. **Expanding the number of users,** n. Finally, in our third stage we project additional users (rows) onto the ideological subspace, but now we include all the accounts, political and non-political (from both the first and second stage), into the analysis. This allows us to estimate the ideological positions of users who do not follow *any* political account, as long as they follow one of the accounts included in the second stage.

As we show in the following sections, this procedure yields valid ideology estimates that replicate conventional measures of political preferences.

One additional minor difficulty in the estimation is reflection invariance: the direction of the latent ideological dimension could be reversed without changing the prediction of the models. This is a problem for interpretation, but not for estimation, and can be easily solved by multiplying all the estimated ideological positions by -1 if the scale is not in the proper liberal-conservative direction. To facilitate the interpretation, we also standardize our ideological estimates so that they have a normal distribution with mean of zero and a standard deviation of one.

2 Data Collection, Curation and Annotation

2.1 Individual-level Data

As described above, the first step in our analysis is to identify a list of political accounts such that the decision to follow them is informative about ideology. Our list includes, among others, the Twitter accounts of all Members of Congress with more than 5,000 followers, the President (`@BarackObama`) and Vice-President (`@JoeBiden`), the Democratic and Republican parties (`@TheDemocrats`, `@GOP`), candidates in the 2012 Republican primary election (`@THEHermanCain`, `@GovernorPerry`, `@MittRomney`, `@newtgingrich`, `@timpawlenty`, `@RonPaul`), relevant political figures not in Congress (@algore, @HillaryClinton, @SarahPalinUSA, @KarlRove, GeorgeHWBush), think tanks and civil society group (@Heritage, @HRC, @OccupyWallSt, @BrookingsInst), and journalists and media outlets that are frequently classified as liberal (@nytimes, @msnbc, @NPR, @KeithOlbermann, @maddow, @MotherJones) or conservative (@limbaugh, @glennbeck, @FoxNews, @drudge_report). Table S1 provides the full list of accounts included in this step.

Table S1: Accounts included in first stage of model

Members of the U.S. Congress					
lisamurkowski	repdonyoung	SenatorBegich	SenatorSessions	SenShelby	RepMarthaRoby
RepMikeRogersAL	Robert_Aderholt	RepMoBrooks	BachusAL06	RepTerriSewell	JohnBoozman
SenMarkPryor	RepRickCrawford	RepTimGriffin	rep_stevewomack	RepTomCotton	JeffFlake
SenJohnMcCain	RepKirkpatrick	RepRonBarber	RepRaulGrijalva	RepGosar	RepMattSalmon
RepDavid	RepTrentFranks	RepSinema	SenFeinstein	SenatorBoxer	RepLaMalfa
RepJeffDenham	askgeorge	NancyPelosi	RepBarbaraLee	RepSpeier	RepSwalwell
RepJimCosta	RepMikeHonda	RepAnnaEshoo	RepZoeLofgren	RepHuffman	RepSamFarr
RepDavidValadao	DevinNunes	GOPWhip	RepLoisCapps	BuckMcKeon	JuliaBrownley26
RepJudyChu	RepAdamSchiff	RepCardenas	RepGaramendi	BradSherman	RepGaryMiller
gracenapolitano	WaxmanClimate	RepBecerra	RepMcLeod	CongressmanRuiz	RepKarenBass
RepLindaSanchez	RepEdRoyce	RepMcClintock	RepRoybalAllard	RepMarkTakano	KenCalvert
MaxineWaters	Rep_JaniceHahn	RepJohnCampbell	LorettaSanchez	RepLowenthal	DanaRohrabacher
DarrellIssa	RepThompson	Rep_Hunter	RepJuanVargas	RepScottPeters	RepSusanDavis
DorisMatsui	RepBera	RepPaulCook	RepMcNerney	MarkUdall	SenBennetCO
RepDianaDeGette	RepJaredPolis	RepTipton	repcorygardner	RepDLamborn	RepMikeCoffman
RepPerlmutter	ChrisMurphyCT	SenBlumenthal	RepJohnLarson	RepJoeCourtney	rosadelauro
jahimes	RepEsty	EleanorNorton	JohnCarneyDE	SenatorCarper	ChrisCoons
SenBillNelson	marcorubio	RepJeffMiller	RepWebster	RepRichNugent	RepGusBilirakis
USRepKCastor	RepDennisRoss	VernBuchanan	TomRooney	RepMurphyFL	treyradel
Rep_Southerland	RepTedDeutch	RepLoisFrankel	DWStweets	RepWilson	MarioDB
JoeGarcia	RosLehtinen	RepTedYoho	AnderCrenshaw	RepCorrineBrown	RepDeSantis
congbillposey	AlanGrayson	SaxbyChambliss	SenatorIsakson	JackKingston	RepPaulBrounMD
RepPhilGingrey	repjohnbarrow	repdavidscott	RepTomGraves	SanfordBishop	RepWestmoreland
RepHankJohnson	repjohnlewis	RepTomPrice	AustinScottGA08	RepDougCollins	maziehirono
SenBrianSchatz	RepHanabusa	TulsiPress	ChuckGrassley	SenatorHarkin	BruceBraley
daveloebsack	TomLatham	SteveKingIA	MikeCrapo	SenatorRisch	Raul_Labrador
CongMikeSimpson	SenatorDurbin	SenatorKirk	RepBobbyRush	RepSchneider	RepBillFoster
RepBillEnyart	RodneyDavis	RepHultgren	RepShimkus	RepKinzinger	RepCheri
RepAaronSchock	RepRobinKelly	RepLipinski	RepGutierrez	RepMikeQuigley	PeterRoskam
RepDannyDavis	RepDuckworth	janschakowsky	SenDanCoats	SenDonnelly	RepVisclosky
RepWalorski	RepStutzman	ToddRokita	SusanWBrooks	RepLukeMesser	RepAndreCarson
RepLarryBucshon	RepToddYoung	JerryMoran	SenPatRoberts	CongHuelskamp	RepLynnJenkins
RepKevinYoder	RepMikePompeo	McConnellPress	SenRandPaul	RepEdWhitfield	RepGuthrie
RepJohnYarmuth	RepThomasMassie	RepHalRogers	RepAndyBarr	DavidVitter	SenLandrieu
SteveScalise	RepRichmond	RepBoustany	RepFleming	RepMcAllister	BillCassidy
MarkeyMemo	SenMoCowan	SenWarren	RepRichardNeal	RepMcGovern	nikiinthehouse
RepJoeKennedy	RepTierney	RepStephenLynch	USRepKeating	SenatorCardin	SenatorBarb
RepAndyHarrisMD	Call_Me_Dutch	RepJohnSarbanes	repdonnaedwards	WhipHoyer	RepJohnDelaney
RepCummings	ChrisVanHollen	SenatorCollins	SenAngusKing	chelliepingree	RepMikeMichaud
stabenow	SenCarlLevin	CongressmanDan	CandiceMiller	RepKerryB	john_dingell
repjohnconyers	RepGaryPeters	RepHuizenga	repjustinamash	RepDaveCamp	RepDanKildee
RepFredUpton	RepWalberg	RepMikeRogers	repsandylevin	alfranken	amyklobuchar
RepTimWalz	repjohnkline	RepErikPaulsen	BettyMcCollum04	keithellison	MicheleBachmann
USRepRickNolan	RoyBlunt	McCaskillOffice	RepAnnWagner	RepHartzler	repcleaver
auctnr1	JoAnnEmerson	RepJasonSmith	SenThadCochran	SenatorWicker	RepAlanNunnelee
BennieGThompson	GreggHarper	CongPalazzo	SteveDaines	MaxBaucus	SenatorBurr
SenatorHagan	GKButterfield	PatrickMcHenry	RepMarkMeadows	RepHolding	RepReneeEllmers
RepWalterJones	RepDavidEPrice	virginiafoxx	HowardCoble	RepMikeMcIntyre	RepRichHudson
reppittenger	SenJohnHoeven	RepKevinCramer	SenatorHeitkamp	Mike_Johanns	SenatorFischer
JeffFortenberry	LEETERRYNE	RepAdrianSmith	KellyAyotte	SenatorShaheen	RepSheaPorter
RepAnnieKuster	CoryBooker	FrankLautenberg	SenatorMenendez	RepDonaldPayne	USRepRodney
RushHolt	RepLoBiondo	RepJonRunyan	RepGarrett	FrankPallone	RepLanceNJ7
RepSires	BillPascrell	MartinHeinrich	SenatorTomUdall	RepLujanGrisham	RepStevePearce

repbenraylujan	SenatorReid	SenDeanHeller	repdinatitus	MarkAmodeiNV2	RepJoeHeck
RepHorsford	chuckschumer	SenGillibrand	TimBishopNY	RepJerryNadler	repmichaelgrimm
RepMaloney	cbrangel	repjoecrowley	RepJoseSerrano	RepEliotEngel	NitaLowey
RepSeanMaloney	RepChrisGibson	RepPeteKing	RepPaulTonko	BillOwensNY	RepRichardHanna
RepTomReed	RepDanMaffei	louiseslaughter	RepBrianHiggins	RepChrisCollins	RepSteveIsrael
RepMcCarthyNY	GregoryMeeks	RepGraceMeng	NydiaVelazquez	RepJeffries	YvetteClarke
SenSherrodBrown	robportman	RepSteveChabot	RepMikeTurner	RepMarciaFudge	RepTimRyan
RepDaveJoyce	RepSteveStivers	RepJimRenacci	RepBradWenstrup	JoyceBeatty	Jim_Jordan
boblatta	RepBillJohnson	RepBobGibbs	SpeakerBoehner	RepMarcyKaptur	TomCoburn
RepJBridenstine	RepMullin	FrankDLucas	tomcoleok04	RepLankford	RonWyden
SenJeffMerkley	RepBonamici	repgregwalden	BlumenauerMedia	RepPeterDeFazio	RepSchrader
SenToomey	SenBobCasey	RepBrady	RepTomMarino	RepLouBarletta	KeithRothfus
USRepMikeDoyle	RepJoePitts	RepCartwright	RepTimMurphy	chakafattah	MikeKellyPA
RepScottPerry	CongressmanGT	JimGerlach	RepMeehan	RepFitzpatrick	RepBillShuster
SenJackReed	SenWhitehouse	RepCicilline	jimlangevin	GrahamBlog	SenatorTimScott
RepSanfordSC	RepJeffDuncan	TGowdySC	RepMickMulvaney	Clyburn	RepTomRice
SenJohnThune	RepKristiNoem	SenJohnsonSD	SenBobCorker	SenAlexander	DrPhilRoe
RepJohnDuncanJr	RepChuck	DesJarlaisTN04	repjimcooper	RepDianeBlack	MarshaBlackburn
RepFincherTN08	RepCohen	JohnCornyn	SenTedCruz	replouiegohmert	McCaulPressShop
ConawayTX11	RepKayGranger	MacTXPress	TXRandy14	USRepRHinojosa	RepBillFlores
JacksonLeeTX18	RandyNeugebauer	JudgeTedPoe	JoaquinCastrotx	LamarSmithTX21	PeteOlson
RepPeteGallego	RepKenMarchant	RepRWilliams	michaelcburgess	farenthold	RepCuellar
RepGeneGreen	SamsPressShop	RepEBJ	JudgeCarter	PeteSessions	RepVeasey
RepFilemonVela	RepLloydDoggett	SteveWorks4You	RalphHallPress	RepHensarling	RepJoeBarton
CongCulberson	RepKevinBrady	RepAlGreen	SenMikeLee	SenOrrinHatch	RepRobBishop
RepChrisStewart	jasoninthehouse	RepJimMatheson	MarkWarner	SenKaineOffice	RobWittman
RepWOLFPress	GerryConnolly	RepScottRigell	repbobbyscott	Randy_Forbes	RepRobertHurt
RepGoodlatte	GOPLeader	Jim_Moran	RepMGriffith	PeterWelch	SenatorLeahy
SenSanders	CantwellPress	PattyMurray	RepDelBene	RepDennyHeck	RepRickLarsen
HerreraBeutler	DocHastings	cathymcmorris	RepDerekKilmer	RepJimMcDermott	davereichert
RepAdamSmith	SenRonJohnson	SenatorBaldwin	RepPaulRyan	repmarkpocan	RepRonKind
RepGwenMoore	JimPressOffice	RepSeanDuffy	RepRibble	SenRockefeller	Sen_JoeManchin
RepMcKinley	RepShelley	HouseTransInf	CynthiaLummis	SenatorEnzi	SenJohnBarrasso

Other political accounts

BarackObama	algore	Schwarzenegger	MittRomney	SarahPalinUSA	KarlRove
JoeBiden	WhiteHouse	GovMikeHuckabee	RickSantorum	RonPaul	newtgingrich
THEHermanCain	GovernorPerry	TheDemocrats	GOP	HillaryClinton	billclinton
GeorgeHWBush	dccc	TimPawlenty	HouseDemocrats	SenateDems	Senate_GOPs
HouseGOP	nytimes	FoxNews	NPR	maddow	glennbeck
KeithOlbermann	limbaugh	DRUDGE_REPORT	Heritage	OccupyWallSt	HRC
RANDCorporation	BrookingsInst	CatoInstitute	AEI		

Next, using the Twitter REST API, we obtained the entire list of followers (as of July, 2014) for all $m = 406$ political accounts, resulting in a entire universe of Twitter users following at least one such account of $n = 60{,}130{,}443$.

As explained above, in the first stage of our estimation approach we only include users who follow 10 or more political accounts, which represents a total of 178,676 users. Then, we took a random sample of 1,000 users in the extremes of the ideology distribution (below the 20th percentile and above the 80th percentile), downloaded the list of accounts they follow, and identified the 400 most popular among conservatives and the 400 most popular among liberals. Table S2 lists some of the 800 accounts that we added in this step. Our final sample size, considering all Twitter users that follow at least one of the $m = 1{,}206$ target accounts and passed the spam and location filter we describe in the following section is $n = 3{,}731{,}957$ active users.

Table S2: Accounts added in second stage of model (top 200 accounts for each group)

Accounts followed by liberals

JenGranholm	stephenfry	latimes	lenadunham	EricBoehlert	MarkRuffalo
politicususa	AlterNet	linnyitssn	glaad	LeftOutLoud	MichaelEDyson
LeoDiCaprio	ProPublica	truthout	davidgregory	KevinSpacey	richardwolffedc
FlaDems	daveweigel	crooksandliars	GottaLaff	chrisrock	jeremyscahill
JohnCleese	AnnCurry	sullydish	cher	ladygaga	amnesty
clairecmc	NatGeo	jonathanalter	NowWithAlex	FareedZakaria	emilyslist
PolitiFact	MikeBloomberg	DavidShuster	NBCNews	BuzzFeed	azizansari
stefcutter	RollingStone	soledadobrien	ElMonte08	katiecouric	hrw
TheTweetOfGod	Newsweek	BashirLive	AFLCIO	kathygriffin	jilevin
anamariecox	BetteMidler	funnyordie	ThePlumLineGS	DemGovs	JohnKerry
mindykaling	guardian	Reuters	PoliticalLine	AriMelber	NOH8Campaign
RichardEngel	TheEconomist	SEIU	JoyVBehar	AJEnglish	rcooley123
SethMacFarlane	pourmecoffee	BoldProgressive	BillMoyersHQ	SteveMartinToGo	CharlesMBlow
PPact	PBS	140elect	StateDept	TeaPartyCat	pattonoswalt
GStephanopoulos	politico	UN	SamuelLJackson	howardfineman	samsteinhp
PaulBegala	ReadyForHillary	dscc	EJDionne	UniteBlue	MHPshow
ACLU	Politics_PR	CapehartJ	mtaibbi	louisck	TheAtlantic
ebertchicago	Upworthy	camanpour	mitchellreports	ABFalecbaldwin	TheScienceGuy
AmbassadorRice	BBCBreaking	alexwagner	TheLastWord	rickygervais	krystalball
finneyk	tomhanks	BBCWorld	Change	GeorgeTakei	sethmeyers
TheNewDeal	democracynow	upwithsteve	NickKristof	ConanOBrien	CornelWest
GabbyGiffords	lizzwinstead	JoyAnnReid	TheRevAl	ClintonFdn	SandraFluke
washingtonpost	AP	JohnFugelsang	chucktodd	TIME	BillGates
Eugene_Robinson	BorowitzReport	OFA	thedailybeast	jimmyfallon	TheDailyEdge
Slate	CNN	ActuallyNPH	TPM	TheOnion	Salon
cnnbrk	KatrinaNation	WeGotEd	HuffPostPol	joanwalsh	VanJones68
PressSec	hardball_chris	markos	WendyDavisTexas	MoveOn	SarahKSilverman
dailykos	elizabethforma	edshow	Oprah	mmfa	DalaiLama
neiltyson	msnbc	nprpolitics	NewYorker	VP	donnabrazile
LOLGOP	ChelseaClinton	RBReich	DavidCornDC	thenation	TheEllenShow
Lawrence	NYTimeskrugman	ariannahuff	andersoncooper	NateSilver538	MHarrisPerry
MaddowBlog	HuffingtonPost	davidaxelrod	FLOTUS	MMFlint	chrislhayes
nprnews	ezraklein	MotherJones	StephenAtHome	thinkprogress	TheDailyShow
MichelleObama	billmaher				

Accounts followed by conservatives

michellemalkin	seanhannity	AllenWest	tedcruz	marklevinshow	IngrahamAngle
megynkelly	AnnCoulter	DennisDMZ	PRyan	krauthammer	DLoesch
theblaze	BretBaier	DanaPerino	FreedomWorks	TwitchyTeam	RealBenCarson
rushlimbaugh	TeaPartyNevada	brithume	Judgenap	greggutfeld	DavidLimbaugh
MonicaCrowley	DRUDGE	PAC43	BreitbartNews	gretawire	chuckwoolery
fredthompson	AndyWendt	BraveLad	RealJamesWoods	KatiePavlich	JimDeMint
Miller51550	ForAmerica	realDonaldTrump	oreillyfactor	LessGovMoreFun	RedState
KatyinIndy	AndreaTantaros	jjauthor	FBNStossel	netanyahu	NatShupe
TIMENOUT	secupp	Reince	BobG231	TuckerCarlson	kimguilfoyle
stephenfhayes	AmbJohnBolton	BobbyJindal	TeamCavuto	Dbargen	JedediahBila
CarrollStandard	PolitixGal	foxnewspolitics	BarracudaMama	Rasmussen_Poll	CharlieDaniels
politichickAM	irritatedwoman	ericbolling	Daggy1	BillyHallowell	MarkRMatthews
mikandynothem	rwhitmmx	PJStrikeForce	Norsu2	DailyCaller	JessicaChasmar
RightWingArt	lr3031	LindaSuhler	JohnFromCranber	MiaBLove	DickMorrisTweet
ArcticFox2016	libertyladyusa	TrucksHorsesDog	DrMartyFox	RightCandidates	WayneDupreeShow
JONWEXFORD	blackrepublican	AnnDRomney	GeneMcVay	AppSame	WashingtonDCTea
JonahNRO	TPPatriots	ConserValidity	PoliticalLaughs	AriDavidUSA	JDCorbinPM
johnboehner	newsbusters	EWErickson	BossHoggUSMC	BlueWaterDays	guntrust
NaughtyBeyotch	DougDauntless	CandiceLanier	KirstenPowers	ChuckNellis	ByronYork
foxandfriends	foxnation	TimTebow	BraveConWarrior	AHMalcolm	Thomasjwhitmore
LibertyBritt	LifeNewsHQ	iowahawkblog	TheDavidMcGuire	TedNugent	DanJoseph78
BluegrassPundit	FishWithDan	benshapiro	saramarietweets	mericanrefugee	HeyTammyBruce
mkhammer	DawnRiseth	mundyspeaks	jmattbarber	NickEgoroff	concreteczar
WarmingWhiners	palmaceiahome1	IndyEnigma	hannityshow	scrowder	forewit

ChasD3
Thomas_More_Law
AnneBayefsky
RightOrgs
KenWahl1
ShannonBream
MattBatzel
Chris_1791
Moonbattery1
ReaganWorld
jeanniemcbride
EmfingerSScout
LeMarquand
PolitixFireball
UniteRight
TheBubbleBubble
occupycorruptDC
3Quarters2Day
TracyAChambers
NolteNC
Reagan_Nation
ATHudd
nf3l
OBAMA_CZAR
famousquotenet
ChristiChat
aaronrobinow
biggovt
mrclean2012
GarySinise
TruthCry
TxRightWing
Stonewall_77
JudicialWatch
EricaRN4USA
Mike_Beacham
joethepatriotic
loudobbsnews
Obama_Clock
TheToady
marthamaccallum
PatDollard
iSheeple1
Herb_Slojewski
AndrewBreitbart
KurtSchlichter
MercuryOneOC
TheTeaParty_net
stephenstephan
WayneBogda
toddstarnes
stephenkruiser
Pudingtane
LibertySurfer
AFPhq
CaptYonah

Our sample selection process required obtaining information about each user, which we extracted from their profiles using Twitter's REST API. In addition, we also parsed the location information into geographic coordinates for a random sample of 300,000 of them using the Data Science Toolkit geocoder, which allowed us to identify the state in which they are located in 71% of the cases.

In addition, we also matched a sample of Twitter users from the states of California, Florida, Pennsylvania, Ohio, and Arkansas with their voter registration records, publicly available through the Secretary of State in each state. In particular, we selected all users that mentioned strings associated to each state in their profiles[2], and then matched them with voters in each state only when their first name, last name, and county had a unique record in both databases. A total of 42,008 Twitter users were matched with their voter registration history.[3] Party affiliation was available for 25,094 of these voters.

2.2 Activity, Location, and Spam Filters

One important obstacle in any analysis of Twitter data is that an extremely high proportion of users of this platform are either inactive, spam bots or located outside of the U.S. This can be particularly problematic in the study of political discussions, since political campaigns can "buy" Twitter followers or create "bots" to promote their platform using spam messages.

To address this concern, in our analysis we limit our sample to only *active* users in the United States, which we define as those who (1) have sent more than 100 tweets, (2) have 25 or more followers, (3) follow 100 or more other accounts, (4) tweet in English, and (5) mention keywords related to two or more of our collections. Conditions 1 to 4 implement a simple activity and location filter, while condition 5 addresses the concern about spam bots. In particular, the reason why we restrict our sample to users tweeting in *two or more* of our collections is that some spam bots are built so that they send automatic tweets that mention trending topics and popular hashtags in order to "hijack" them (Thomas et al., 2012). Filters based on numbers of followers are not effective in these cases, because these bots tend to follow each other (Yardi et al., 2009). Most of these accounts are automatically detected by Twitter and subsequently deleted, but their tweets would remain in our datasets. By focusing on accounts whose tweeting activity is observed at multiple times, we are able to discard spam bots that fall within this category.

[2] For example, in Ohio we searched for users whose location field mentions the strings string "ohio" or "OH", or any of the major cities (Columbus, Cleveland, Cincinnati, Toledo, Lima, etc.).

[3] 22,544 users in California, 3,492 in Florida, 5,370 in Pennsylvania, 8,885 in Ohio, and 1,716 in Arkansas

In all, these may appear to be highly restrictive conditions, but users who meet them generated over 70% of the tweets in all our collections. Given the variety of topics we consider in our collection, our sample thus includes a broad and very large sample of users who were actively tweeting about politics and news events in the United States.[4]

2.3 Twitter Data Collection for Selected Events and Elections

In the text of the article, we analyze the structure and content of the political conversation in Twitter in relation to twelve significant events or issues in the past year, both political and non-political in nature. We collected our datasets from the Twitter Streaming API using the R programming language. Table S3, an expanded version of Table 1 in the article, provides a list of our collections and the keywords we used to filter the tweets. Figure S3 shows the evolution in the daily volume of tweets for each collection. We find significant variation across collections: some events generate many tweets during specific moments in time (e.g. the Oscars or the State of the Union), whereas other issues receive lower levels of attention, but more constant over time (e.g. conversations about the minimum wage or the budget). Furthermore, note that particular spikes in the number of tweets correspond to relevant events, such as the three presidential debates in the 2012 election collection.

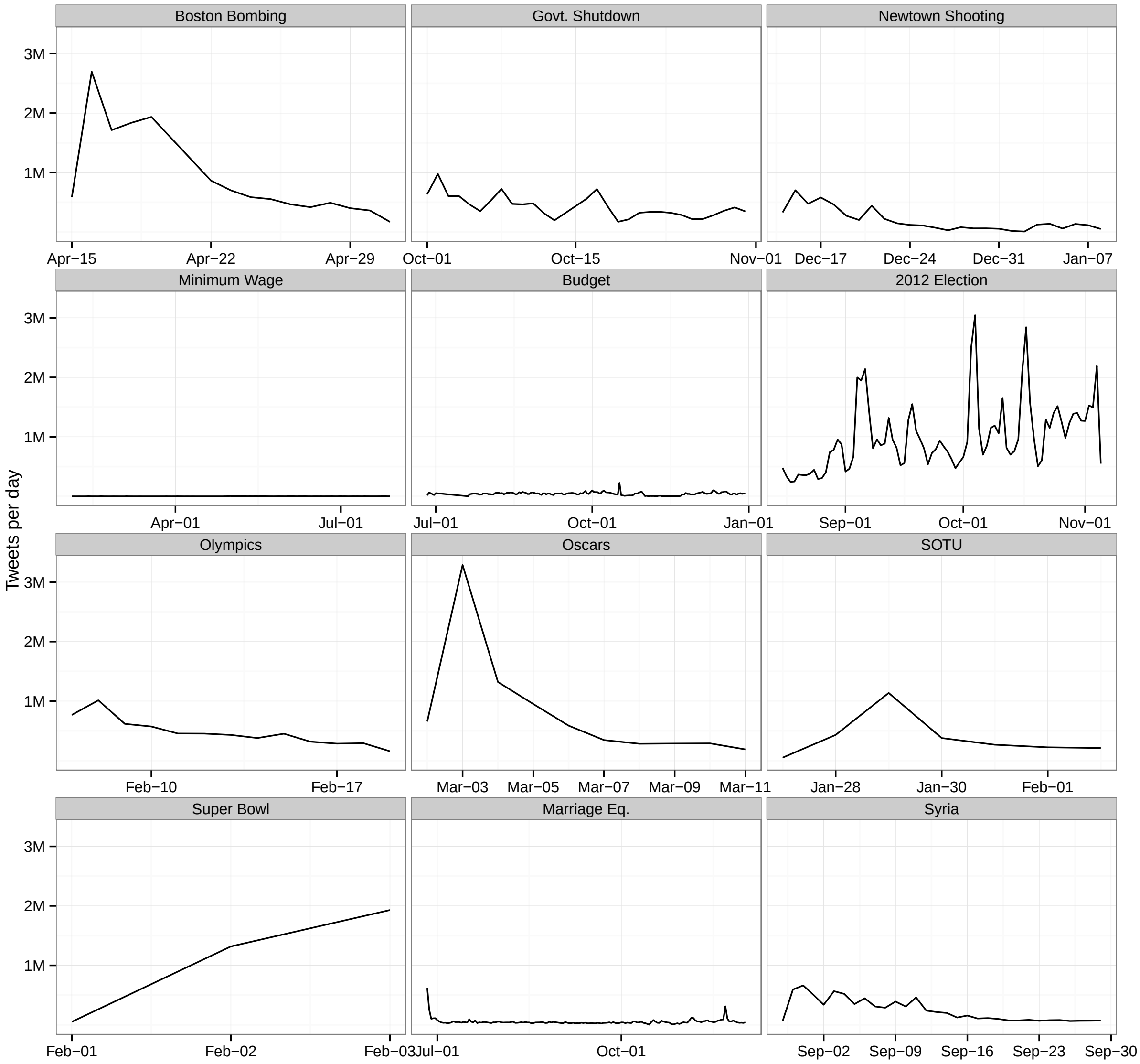
Table S3: Summary of tweets in dataset

Tweet collection	Period	Tweets
2012 Presidential Campaign: obama, romney	2012/08/15 – 2012/11/06	62.3M
Government shutdown: government shutdown, shutdown, shutting down, shut down, furlough, budget, debt, gopshutdown, fairnessforall, enoughalredy, enoughtea, obamacare, boehner, reid, government, slimdown, demandavote, debtlimit, debtceiling, govtshutdown, speakerboehner, senatorreid, harry reid, mcconnellpress, mitch ccconell, tedcruz, ted cruz	2013/10/01 – 2013/11/01	12.4M
Minimum wage: minimumwage, minimum wage, minimum hourly wage, raisethewage, timefor1010, giveamericaaraise, tented, actontenten	2014/02/03 – 2014/04/16	0.2M
Budget: budget, deficit, debt, entitlements, sequester, social security, medicare	2013/06/01 – 2013/12/31	7.7M
Marriage equality: scotus, supreme court, supremecourt, kagan, sotomayor, alito, breyer, ginsburg, justice thomas, justice roberts, justice john, justice clarence, kennedy, scalia, gay marriage, same sex marriage, doma, prop8, prop 8, proposition 8, lgbt, loveislove, marriageequality	2013/06/26 – 2013/12/02	8.2M
State of the Union 2014: situ, sotu14, sotu2014, stateoftheunion, state of the union, obama, barackobama, republican response, gop response, sotugop, mcmorris, mcmorris-rodgets, cathymcmorris, mike lee, senmikelee, rand paul, senrandpaul, opportunityforall, madeinamerica, college opportunity, actonclimate, actonjobs, investinstem, rebuildamerica, actonprek, connected, actonui, actoncir, raisethewage, equal pay, 1010, actontenten	2014/01/27 – 2014/02/02	2.7M
Boston marathon attack: Boston, marathon, explosion, bostonmarathon, attack	2013/04/15 – 2013/04/30	13.3M
Newtown school shooting: prayfornewtown, bra, ctshooting, guncontrol, guns, newtown, gunfights, amendment, gun control, gun rights, sandyhook	2012/12/10 – 2013/01/08	5.1M
Syria: syria, syrian, chemical weapons, saris, assad, bahsar, al-assad, cruise missile, cruise missiles, b2, b117, s300, weapons inspectors, weapons inspector, damascus, jobar, irbin, ghouta, muadhamiya, poison gas, chemical attack, nerve gas	2013/08/28 – 2013/09/30	7.8M
Super Bowl 2014: Super Bowl, broncos, seahawks, touchdown	2014/02/01 – 2014/02/03	5.0M
Oscars 2014: oscars, oscars2014, academy awards	2014/02/19 – 2014/03/11	10.6M
Olympic Games: olympics, sochi, team USA, olympics2014	2014/02/07 – 2014/02/19	7.7M

[4] There may well be networks of single-issue Twitter users, and with our data we wouldn't be able to capture them.

DOI: 10.1177/0956797615594620

Figure S3: Number of Tweets per Day in Each Collection



DOI: 10.1177/0956797615594620

3 Results of Ideology Estimation

3.1 Ideology Estimates

Figure S4 displays the ideology estimates for the most relevant political actors and media outlets. Their positions are consistent with what we would expect: Democrats are on the left and Republicans are on the right; news outlets generally thought to be liberal are on the left, and those generally thought to be conservative are on the right.

Figure S4: Ideology Estimates for Key Political Actors and Media Outlets

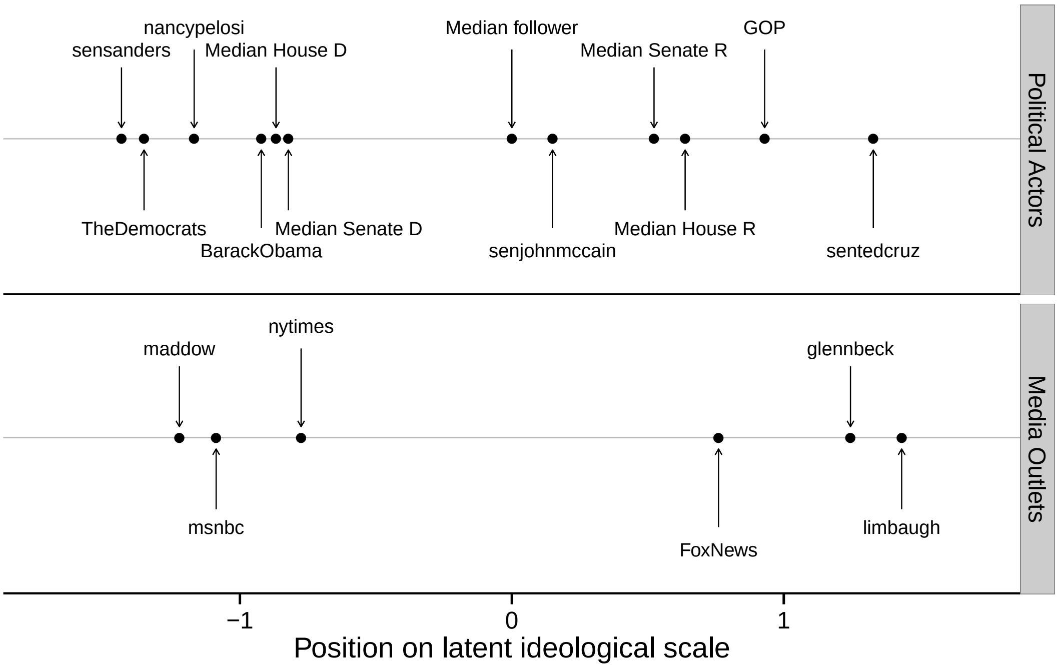


Figure S5 locates the political actors included in the first stage of the model on a two-dimensional space. This figure allows us to confirm that the first dimension clearly overlaps with political ideology. Based on what actors are located on the extremes, we label the second dimension as “congressional politics”: low values correspond to Members of Congress, whereas high values correspond to national-level politicians and media outlets. As shown in Figure S6, which displays the singular values from the SVD step, these two first dimensions are by far the most relevant in explaining users’ following decisions, as evidenced by the fact that the magnitude of the singular values decreases steeply after the second value.

3.2 Model Fit

Before we turn to validate our ideology estimates, we show the results of a battery of predictive checks for binary dependent variables. This allow us to assess whether our estimated parameters from the first stage fit the data; in other words, whether Twitter users’ following decisions are indeed guided by ideological concerns and therefore. As we report in Table S4, all the predictive

DOI: 10.1177/0956797615594620

Figure S5: Distribution of Ideology Estimates, First Two Dimensions

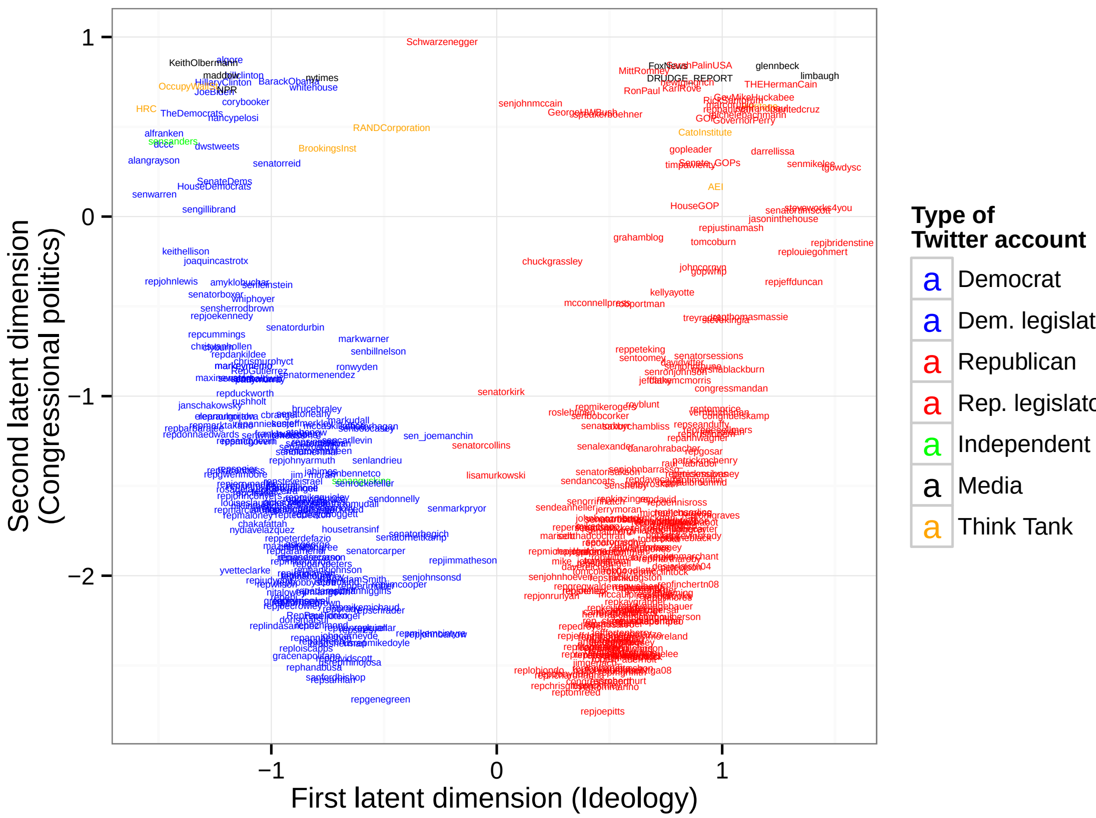
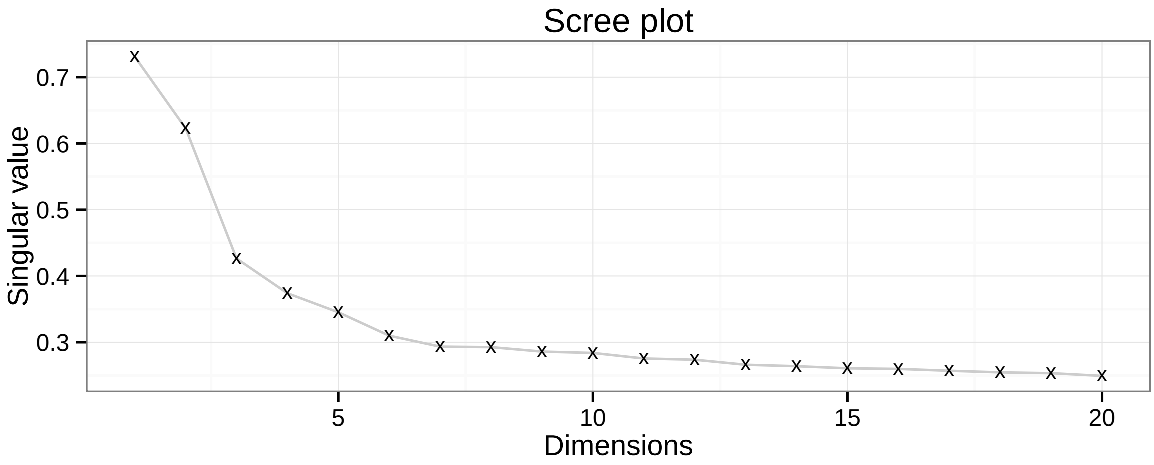


Figure S6: Scree Plot: Importance of Each Dimension



checks suggest that the fit of the model is adequate. Despite the sparsity of the 'following' matrix (less than 8% of values are 1's), the model's predictions improve the baseline (predicting all y_{ij} as zeros), which suggests that Twitter users' following decisions are indeed guided by ideological concerns. In addition to the Pearson's ρ correlation and the proportion of correctly predicted values, Table S4 also shows the AUC and Brier Scores. The former measures the probability that a randomly selected $y_{ij} = 1$ has a higher predicted probability than a randomly selected $y_{ij} = 0$ and ranges from 0.5 to 1, with higher values indicating better predictions (Bradley, 1997). The latter is the mean squared difference between predicted probabilities and actual values of y_{ij} (Brier, 1950).

Table S4: Model Fit Statistics.

Statistic	**Value**
Pearson's ρ Correlation	0.556
Proportion Correctly Predicted	0.946
PCP in Baseline (all $y_{ij} = 0$)	0.944
AUC Score	0.887
Brier Score	0.054
Brier Score in Baseline (all $y_{ij} = 0$)	0.056

3.3 Validation

This section validates our method by comparing our estimates with existing alternative measures of ideology, conducting a similar battery of validation tests as in Barberá (2015). Using different external sources of information, we show that Twitter-based ideology estimates replicate conventional measures of ideology for members of the U.S. Congress, statewide ideology averages, and voters in the state of Ohio. In all cases, we demonstrate that our approach yields estimates of ideology that correctly scale Twitter users on the correct position on a latent ideological dimension.

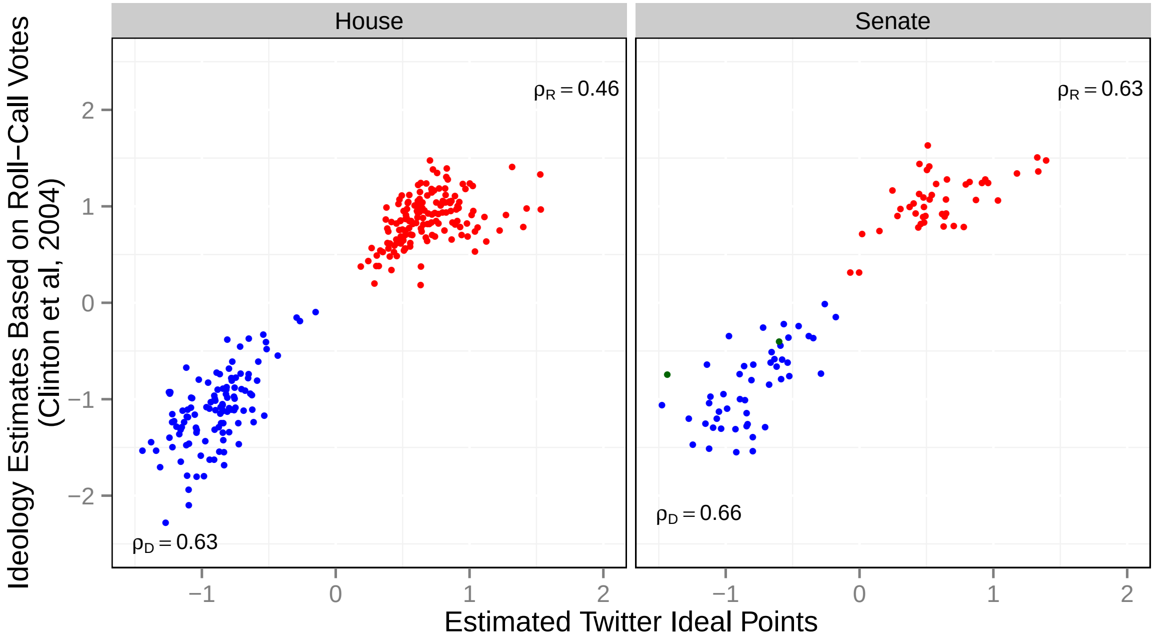
First, we compare our ideology estimates for 365 members of the 113th U.S. Congress with more than 5,000 followers with their "ideal points" based on their roll-call votes in Congress (Jackman, 2014). Figure S7 shows the results of our analysis. Each letter corresponds to a different member of congress, where D stands for democrats and R stands for republicans, and the two panels split the sample according to the chamber of Congress to which they were elected.

The estimated ideal points are clustered in two different groups, that align well with party membership. The correlation between Twitter- and roll-call-based ideal points is $\rho = .956$ in the House and $\rho = .943$ in the Senate. Within-party correlations are also high: $\rho = .442$ for republicans, $\rho = .647$ for democrats. As a comparison, Maestas et al. (2014) found that aggregating individual responses in a survey where voters were asked about the ideological location of Members of Congress led to within-party correlations of $\rho = .71$ for democrats and $\rho = .53$ for republicans. Our result is particularly impressive given that our model estimates legislators' ideology without observing their voting behavior or without aggregating responses from a large and costly survey.

Figure S8 compares the distribution of ideological ideal points for the different types of

DOI: 10.1177/0956797615594620

Figure S7: Comparing Ideal Points Based on Roll-Call Records and Based on Twitter Network of Followers in the U.S. Congress



Twitter users in the sample. The pattern that emerges is consistent with the standard result in the literature (see for example Figure 5 in Bafumi and Herron, 2010), namely, that political actors are much more polarized than mass voters.

As an additional validation test, in Figure S9 we show the ideology of the median Twitter user in each state of the continental U.S., estimated using a random sample of 200,000 users for which we identified their geographic location. Despite Twitter users being a highly self-selected sample of the population, this figure nonetheless presents a close resemblance to ideology estimates based on surveys for each state. As we show in Figure S10, Twitter-based ideal point estimates by state are highly correlated ($\rho = .867$) with the proportion of citizens in each state that hold liberal opinions across different issues, as estimated by Lax and Phillips (2012) combining surveys and socioeconomic indicators. Ideology by state is also a good predictor of the proportion of the two-party vote that went for Obama in 2012, as shown in the right panel of Figure S10, but the correlation coefficient is smaller ($\rho = -.813$).

To further validate the ideal point estimates we introduced in the previous section, now we turn to examine the results from the sample of 42,008 Twitter users in California, Florida, Arkansas, Pennsylvania, and Ohio whose names were matched with the voter files. Figure S11 displays the distribution of our ideology estimates for Twitter users registered as Democrats and Republicans in each state. As we can see, Republican voters are systematically more conservative than Democratic voters.

Figure S12 complements our analysis by taking advantage of the fact that each voter's registration history is available since 2000 in Ohio. Here the horizontal axis groups voters according

DOI: 10.1177/0956797615594620

Figure S8: Distribution of Political Actors and Ordinary Twitter Users' Ideal Points

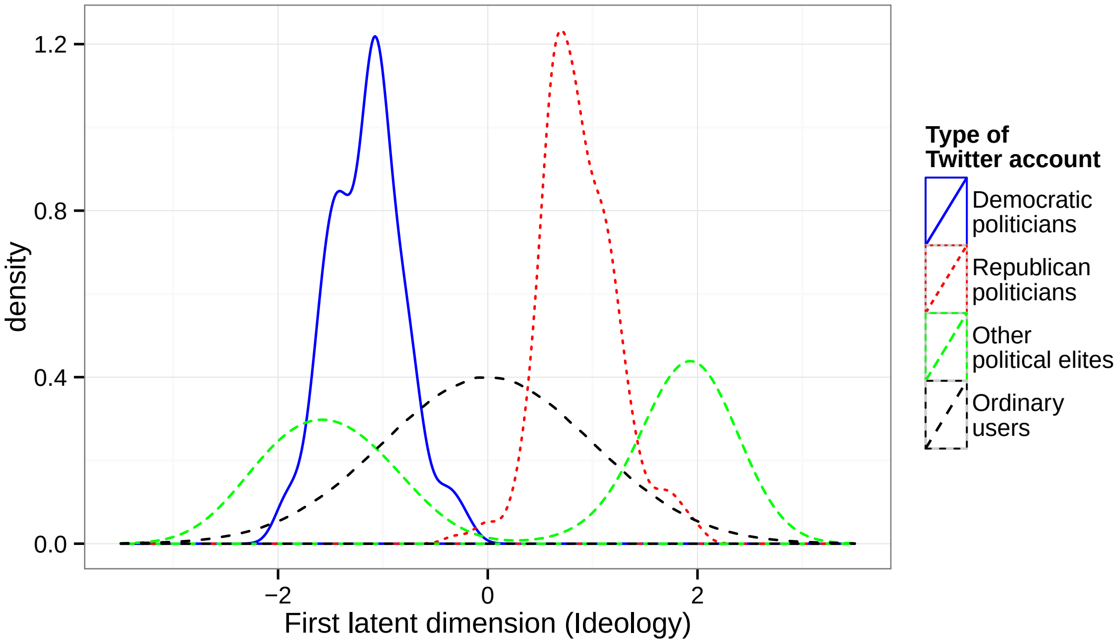
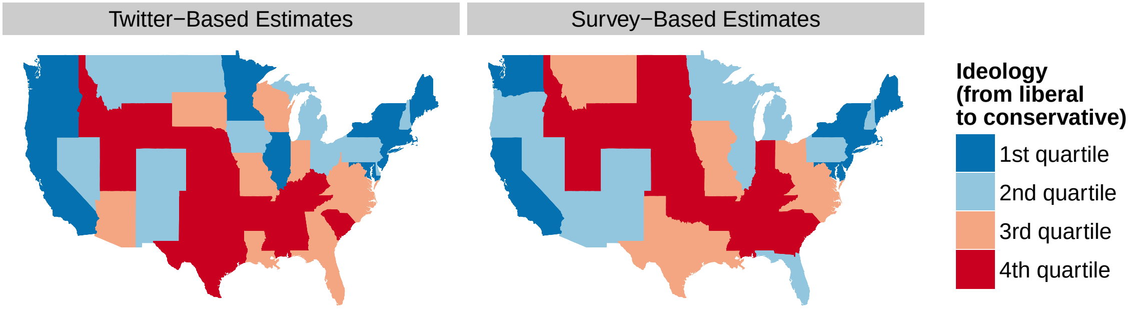


Figure S9: Ideal Point of the Median Twitter User in the Continental US, by State



to how often they have voted in the Democratic primary elections. The vertical axis displays their Twitter-based ideology estimates. We demonstrate that the most conservative (liberal) voters in Ohio tend to consistently register as Republican (Democrat) in the primary elections. This figure confirms that ideology is a very powerful predictor of each voter's registration history.

4 Ideological Polarization: Additional Results.

This section describes the full set of results from our study of political polarization in the United States, measured through social media interactions. As in the main text of the article, we utilize

Figure S10: Twitter-Based Ideal Points, by State

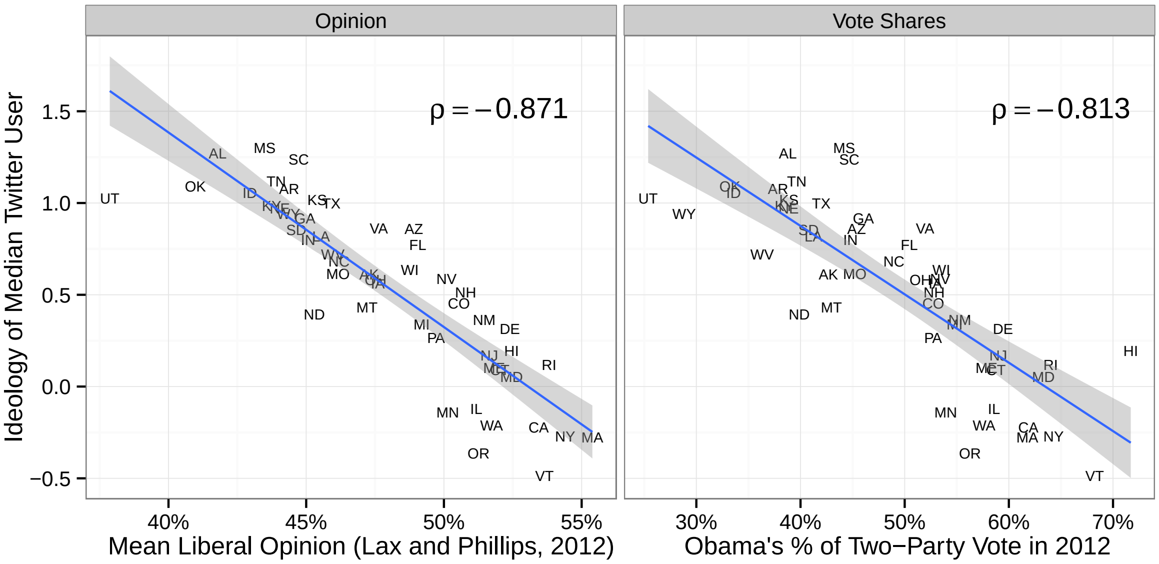
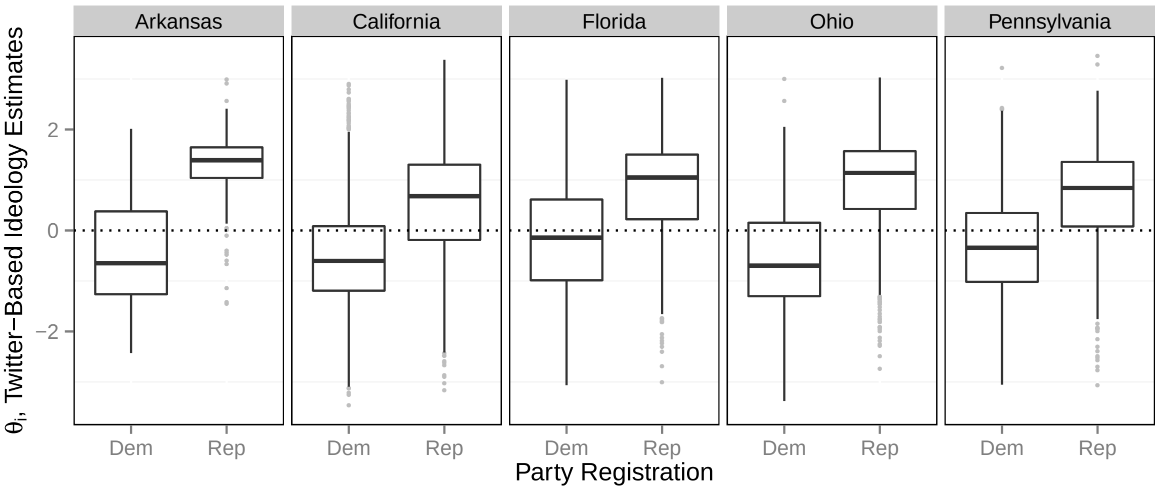


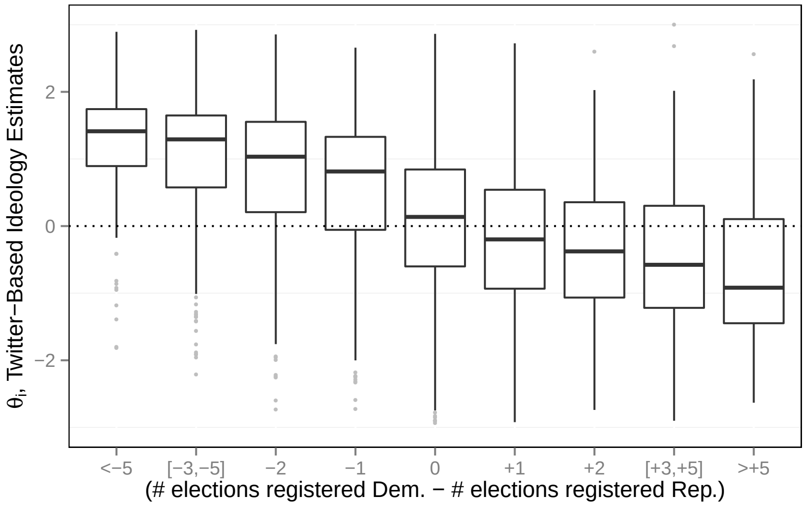
Figure S11: Ideal Point Estimates and Party Registration



three different metrics of polarization, which aim at examining three different aspects of online interactions: ideological homophily in dyadic retweeting interactions, the degree of ideological clustering in the network of information diffusion, and the extent to which ideologically extreme content is shared more often than non-ideological content.

DOI: 10.1177/0956797615594620

Figure S12: Ideal Point Estimates and Party Registration History (Ohio)



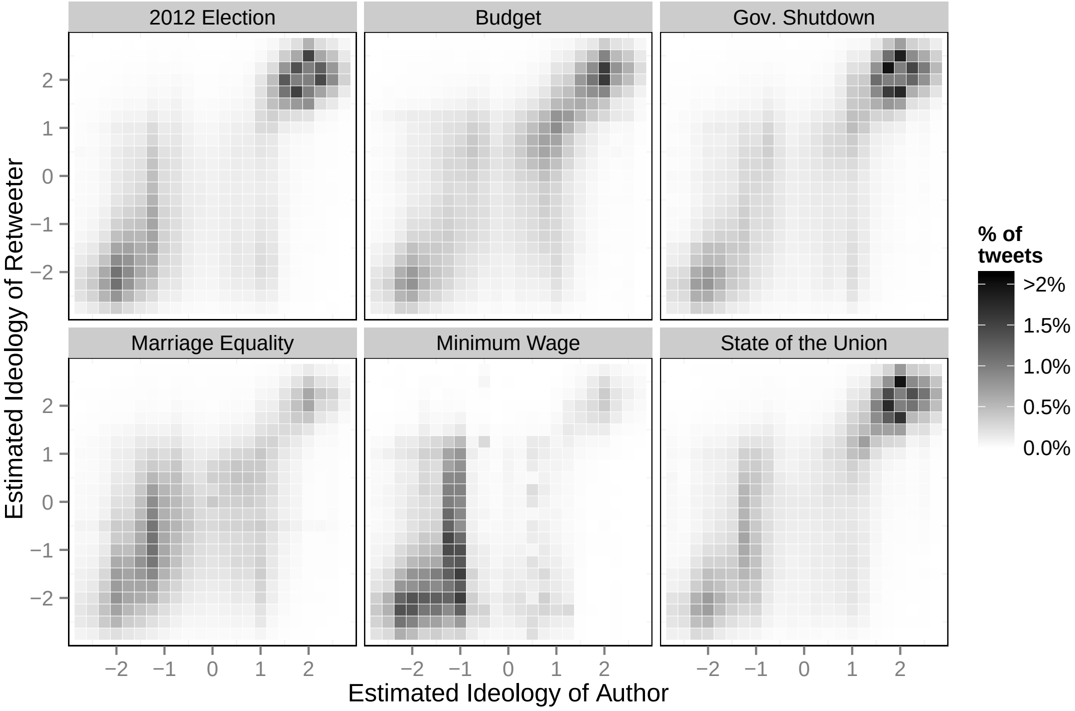
4.1 Ideological Homophily in Dyadic Interactions

Figures S13 and S14 provide additional evidence that the importance of ideological distance in information diffusion varies across different topics. As in the main text of the article, we use heat plots to visualize the most common type of interaction in Twitter: retweets, where one user re-posts another user's content with an indication of its original author. Retweets are used whenever the 'retweeter' wants to publicize the content of the original post, but they are not necessarily a sign of endorsement. In politics, candidates often encourage their followers to retweet their messages. The color of each cell (of size 0.25×0.25) represents the proportion of tweets in the sample that were retweets/mentions of users with ideal point X to users with ideal point Y. For example, in the second panel on the top row of Figure S14 we can see that around 1% of all retweets mentioning Newtown had an original author a Twitter user whose ideal point was in the interval between 2.0 and 2.25, and were retweeted by Twitter users in the same interval.

Polarization in the heat map would be indicated by dark cells along the 45-degree line that slopes up from left to right. Note that the cells could be dark anywhere along the line and it would still indicate polarization: we could see dark corners in the bottom left and top right of the graph; or we could see dark cells along the 45-degree line very close to the center of the graph. The first case would suggest that most discussion is by people at the extremes, but each extreme is retweeting itself. The second case (dark cells along the center) would suggest that most discussion happens among moderates, but that it is people who are moderate-left retweeting the moderate-left, and people on the moderate-right retweeting those on the moderate right.

We find a stark contrast between topics that we labeled as "political" and "not political." On one hand, information about events like the 2012 presidential election or the government shutdown is spread mostly among individuals of similar ideological positions. As we show in Figure S13, most retweets take place in the top-right and bottom-left corner of the plot, where

Figure S13: Political Polarization in Retweets Related to Political Topics

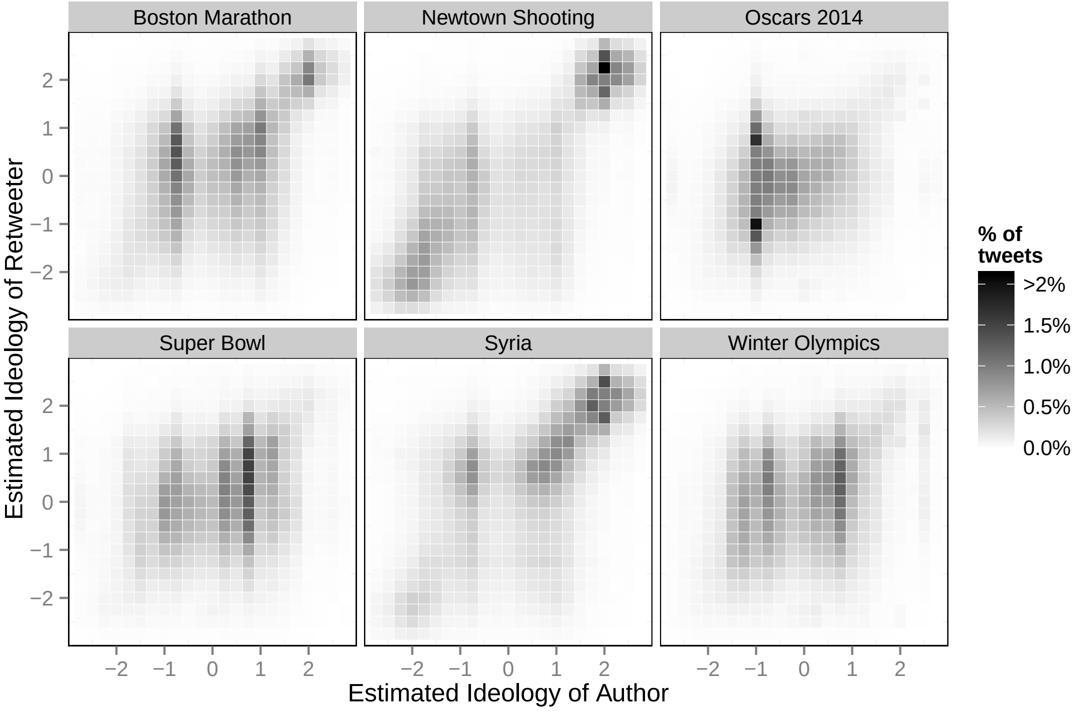


extreme liberals and conservatives (more than one standard deviation away from the center) are located. For example, 38% of all retweets about the 2012 election took place between extreme conservatives (more than one standard deviation above the mean); and 28% between extreme liberals (more than one standard deviation below the mean), even if each of these groups represent only 16% of all users in our sample. We also observe that tweets published by liberal users are generally also retweeted by individuals in the center, but the opposite is usually not the case. This pattern is particularly observable in our collections about the minimum wage and marriage equality, and is due to the success of tweets sent by liberal political accounts, which were retweeted thousands of times. In contrast, tweets originally published by conservative political accounts are hardly retweeted by liberal and moderate users, and at the same time conservative users do not spread tweets posted by liberal and moderate users, as evidenced by the contrast between the relatively lighter cells on the right and top sides of all six heat maps, and the rest of cells.

Patterns of information diffusion in the case of non-political topics, shown in Figure S14, exhibit a completely different pattern. Here, the greatest concentration of retweets occurs in the middle of the ideological spectrum, particularly in the case of non-political events such as the Super Bowl or the Winter Olympics. Furthermore, here we find that conservative users do spread information generated by moderate or liberal sources, as is evident in the case of the 2014 Oscars (note that most actors and directors have liberal ideological positions) and the Boston Marathon Bombing. Note that in these figures we do not see shading clustered around the 45-degree line as we examine the cells in the figure, at least until we reach the far

DOI: 10.1177/0956797615594620

Figure S14: Political Polarization in Retweets Related to Non-Political Topics



right – the exception to this pattern is the set of tweets about the Boston Marathon bombing, where we observe that tweets from the far right were generally only retweeted by other far right users. This result shows that ideological homophily in the propagation of content related to non-political events is lower or non-existent, and therefore is counter to the idea of social media as an "ideological echo chamber."

4.2 Ideological Clustering in Networks of Information Diffusion

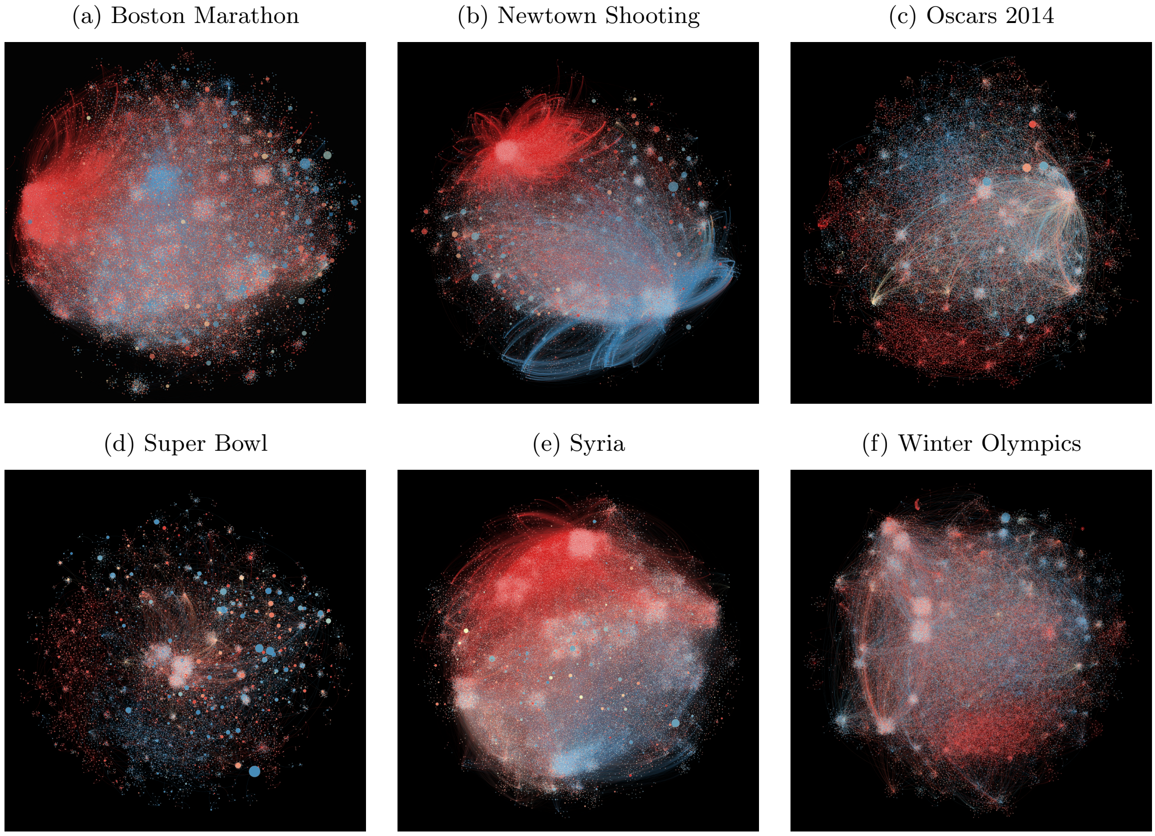
An alternative approach to the analysis of information diffusion on social media is to treat each dyadic interaction as an edge of a network, where each node is a social media user. We can then visualize these large-scale networks using force-directed layout algorithms to observe the existence of different clusters or "cliques" of users, and whether these groups are ideologically homogenous, to ascertain the extent to which similarity in ideological positions determines the structure of communication networks.

We report the results of this analysis in Figures S15 and S16. As in the previous section, we divide our collections in two groups, according to whether they are related to political and non-political events or issues. Given the magnitude of these networks, we report only retweet interactions among a set of 300,000 users, chosen randomly, but giving more weight to users that sent more tweets about different topics and had more followers. In particular, we sampled with the following weights:

$$w_i = 0.5 \times \frac{\log f_i}{\sum_i \log f_i} + 0.5 \times \frac{\log \sum_c \frac{t_{ic}}{n_c}}{\sum_i \log \sum_c \frac{t_{ic}}{n_c}} \tag{2}$$

where f_i is the number of followers of user i, t_{ic} is the number of tweets sent by user i in collection c, and n_c is the total number of tweets in collection c.

Figure S15: Information Diffusion Networks for Non-Political Topics



(a) Boston Marathon

(b) Newtown Shooting

(c) Oscars 2014

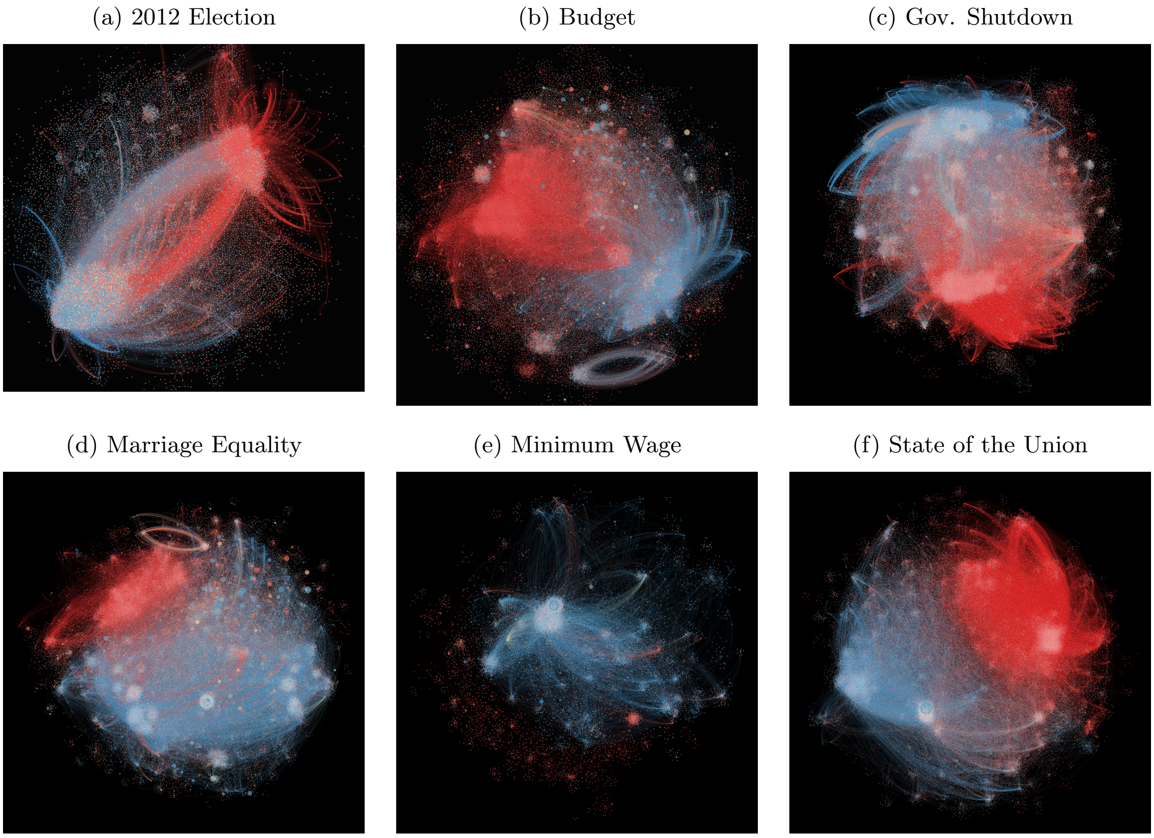
(d) Super Bowl

(e) Syria

(f) Winter Olympics

Then, we extracted the retweets in each of our collections that took place among this set of users, considering each retweet as a directed edge from the author of the tweet to the user that retweets it. After building the network using the open-source software Gephi (Bastian et al., 2009), we identified the largest weakly connected component (Tarjan, 1972), which in all cases contained between 80 and 90% of the nodes and over 90% of the edges, and discarded the rest of the network. Removing users who did not retweet or were retweeted by any other user in the largest subgraph of the network provides a more clear visualization of the different existing clusters without affecting the interpretation of our results. Then, we use the OpenOrd layout algorithm (Martin et al., 2011), a variant of force-directed layout algorithms such as the Fruchterman-Reingold algorithm (Fruchterman and Reingold, 1991) that can scale to a large number of nodes. This type of algorithm positions the nodes of a network in a two-dimensional space so that nodes that are connected by a larger number of edges are located closer to each other and nodes that are not connected are farther apart. By coloring the nodes according to their estimated ideology (from dark red for conservatives to dark blue for liberals), this algorithm

Figure S16: Information Diffusion Networks for Political Topics



(a) 2012 Election

(b) Budget

(c) Gov. Shutdown

(d) Marriage Equality

(e) Minimum Wage

(f) State of the Union

allows to observe the level of ideological clustering in information diffusion via retweets.

As earlier, our results show a gradient of ideological clustering. First, for events like the Super Bowl, the 2014 Oscars, the Winter Olympics, we observe a single large cluster, where liberals and conservatives appear to retweet each other regarding of their ideology. The retweet networks for the Boston Marathon bombing, the Newtown shooting and the military intervention in Syria represent intermediate cases, where we also observe small segregated groups that are ideologically homogenous, but the visualization still indicates that most users are located in a single large cluster where information propagates regarding of ideology.

In contrast, the retweet networks about political topics clearly show the emergence of two different clusters, one composed of conservative users and the other of liberal users. Their size varies across topics in the expected direction: the conservative cluster is larger in the budget network, while the liberal cluster is larger in the marriage equality network. An extreme example is the minimum wage network, in which we observe a single cluster that is overwhelmingly liberal, because conservative Twitter users barely mentioned this issue during our period of analysis. If we examine the separation across these two groups, we find that it is larger for the most political topics, such as the State of the Union address and, in particular, the 2012 election. In these two cases, we observe two very tightly interconnected groups of individuals, which implies that information shared in one of them is unlikely to reach the other group. As we argue in the main text of the article, most existing studies overestimate the degree of ideological clustering

DOI: 10.1177/0956797615594620

because they focus only on highly political issues which, as we show here, are extreme cases when it comes to the importance of ideology in online interactions.

4.3 Variation in Aggregate Levels of Political Polarization

One important limitation in our analysis is that the extent to which information diffusion is guided by ideological distance is dependent on the production of content from ideologically extreme sources. In this section we address this concern by examining political polarization in both information creation and information diffusion.

Figure S17 compares the ideological distribution of tweets and retweets in each of our collections, after assigning to each tweet or retweet the ideology of its author (in the case of retweets, the ideological position of its original author, not who retweeted it).[5] The density plot in black indicates the ideological distribution of all tweets, whereas the red line shows the distribution of retweets.[6] As expected, we find that individuals with non-moderate ideology are more active in collections about political topics, such as the State of the Union address or the 2012 Election. In contrast, tweets about non-political topics tend to be generated by moderate sources. More importantly, this figure also allows us to compare whether ideologically extreme information within each collection is shared more often than would be expected according to the baseline distribution for all the tweets that are being sent.

We provide a metric that summarizes the differences between the ideological distribution of tweets and retweets on the left column of Figure S18. In particular, we computed the ratio of the proportion of retweets generated by individuals on the extremes of the ideological distribution (more than one standard deviation away from the center) and the tweets generated by users in that same interval. High values in this ratio imply that ideologically extreme content is more popular and more likely to be propagated, controlling for the proportion of all generated content that falls within this category. Low values in this ratio, on the other hand, correspond to collections in which content generated by moderate accounts is relatively more popular than that created by individuals with extreme ideological positions.

We find that information created by this set of extreme users is relatively more popular in collections about political topics: on average, for every 100 ideologically extreme tweets, there are 110 extreme retweets in these collections.[7] In contrast, tweets about non-political topics show a more balanced ratio, around one, which implies that the information that is spread via retweets has an ideological distribution approximately equal to the information that is being produced.[8]

[5] Here we display results only for a random sample of 1 million tweets and retweets from each collection.

[6] Note that this second line has more spikes due to particular tweets that were retweeted a large number of times; for example, the clear spike in the Oscars collection corresponds to the "selfie" tweet posted by Ellen Degeneres, which as of July 2014 is the tweet with the highest number of retweets.

[7] One exception to this pattern is the set of tweets about Syria. In this case, tweets with moderate ideological content appear to be relatively more popular. One explanation for this difference could be related to the fact that this is our only foreign policy collection, which may foster the propagation of information vis-à-vis opinion, given that this is a topic about which Americans are less likely to be informed about.

[8] Note that in our analysis of the Oscars collection we exclude retweets of Ellen Degeneres' "selfie" tweets, in order to facilitate the comparison. This single tweet accounts for almost 33% of all retweets in our collection, and given Ellen Degeneres' liberal ideological position, represents an extreme outlier. If it were included, the popularity ratio for this collection would be 1.88.

DOI: 10.1177/0956797615594620

Figure S17: Ideological Distribution of Content Shared on Twitter

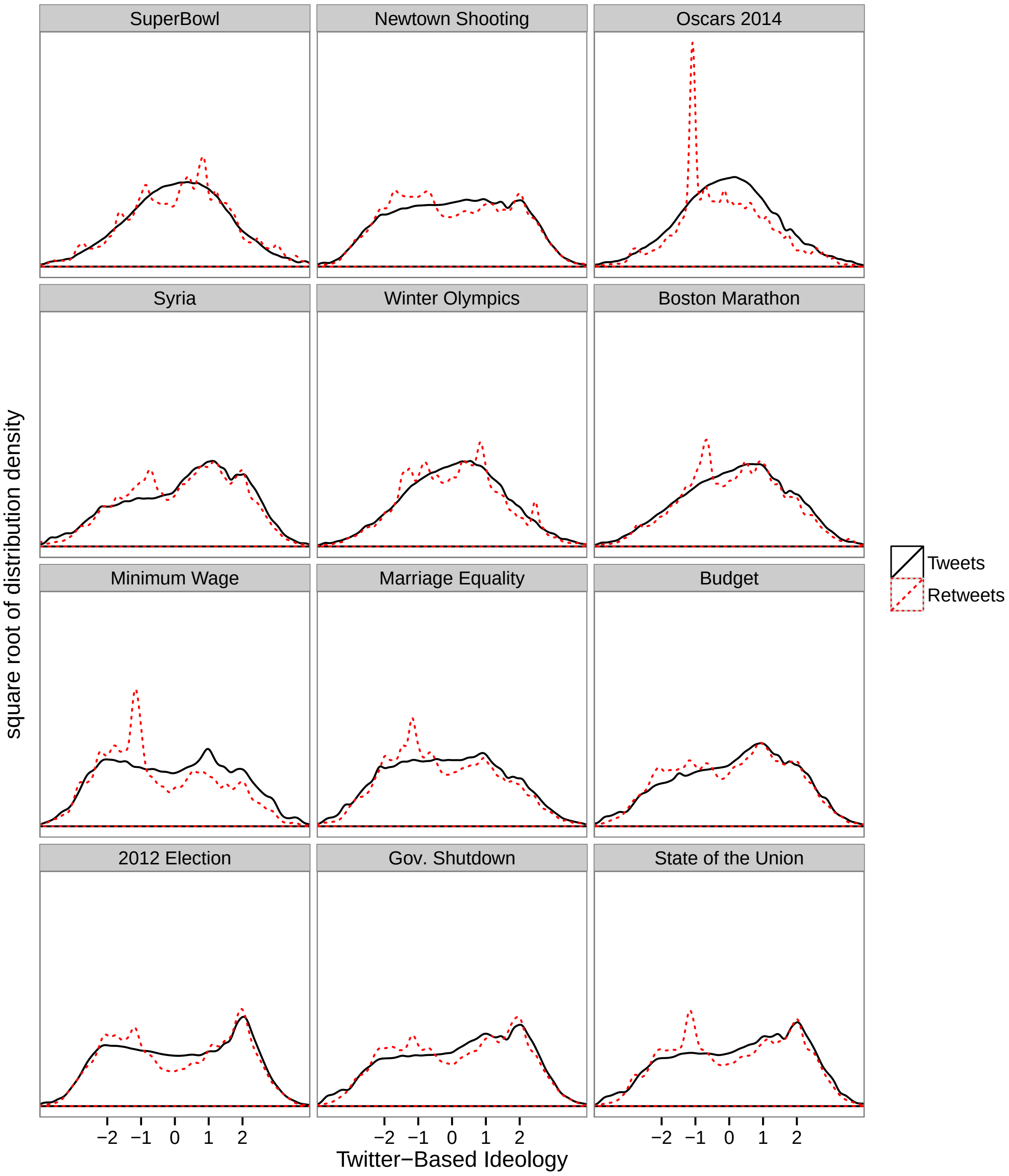
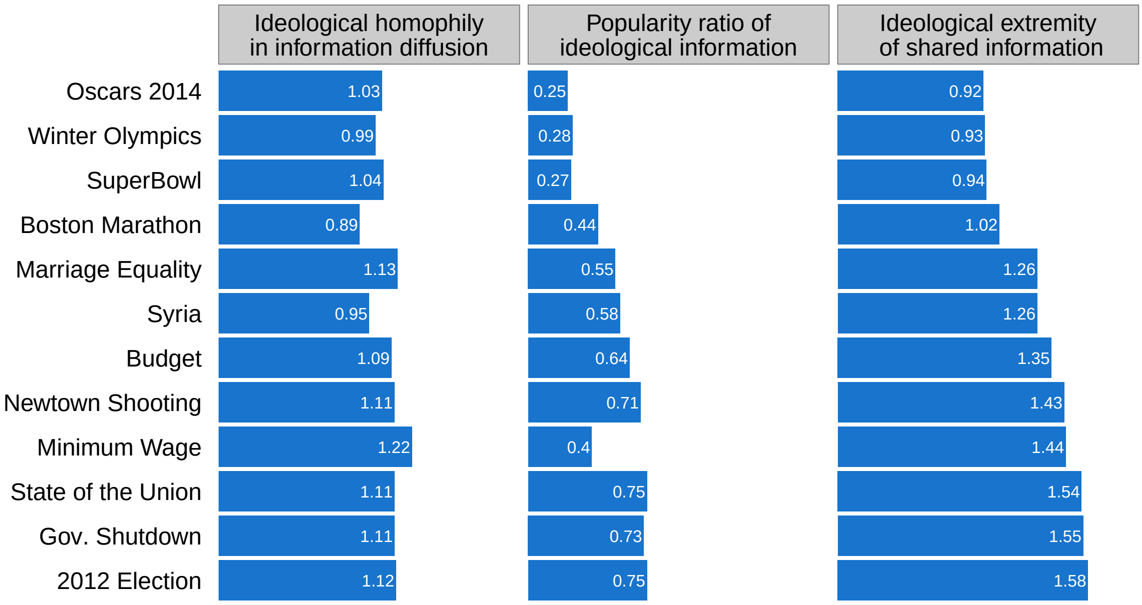


Figure S18 also includes other polarization indices for each of our collections. The center column displays the slope coefficient of an ordinary least squares regression of the ideology of the "retweeted" on the ideology of the "retweeter", where the retweet is the unit of analysis. If individuals were retweeting only content shared by other users with their exact same ideological position, then this coefficient would be equal to one. If ideological proximity was unrelated to the probability of retweeting, then the slope would be zero. Finally, the right column replicates

DOI: 10.1177/0956797615594620

Figure S18: Alternative Measures of Political Polarization in Information Diffusion



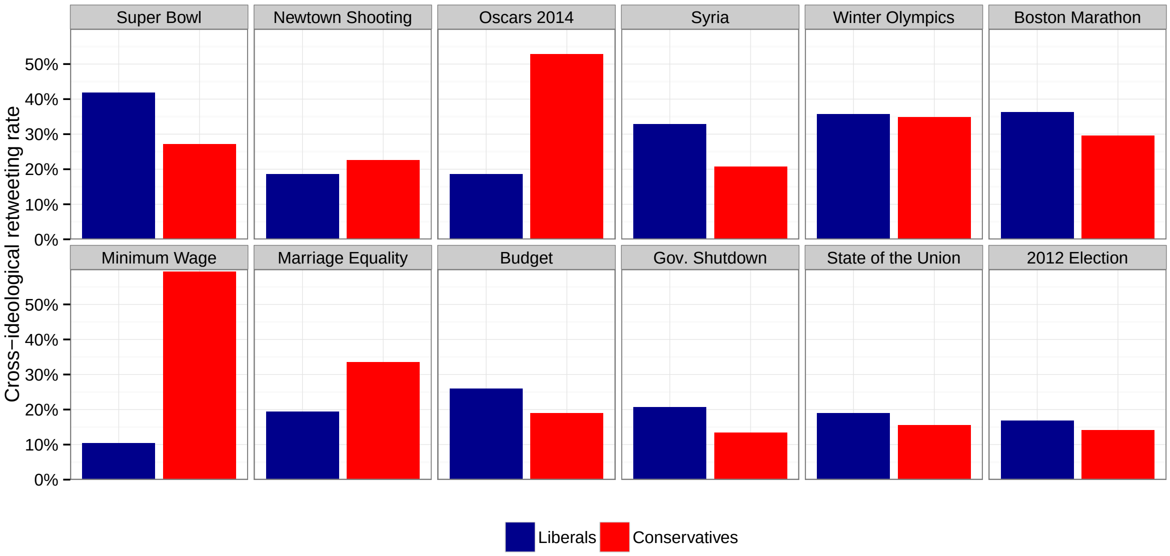
the polarization index in the main text of the article: the average absolute distance between the author of a tweet and the ideological center, only for tweets that were retweeted. Higher values in this indicator imply that the information shared via retweets features content that is more ideologically extreme. All three indicators reinforce our previous conclusions: (1) there is substantive variation in the extent to which social interactions on Twitter are ideologically polarized, and (2) this variation is related to the type of issue or event that sparks these interactions in systematic ways.

5 Ideological Asymmetries in Information Diffusion

5.1 Descriptive Analysis

We are interested in determining whether the rates of cross-ideological retweeting are higher among liberals than among conservatives. In order to do so, we compared liberal and conservative rates of cross-ideological retweeting: the proportion of messages written by conservatives that were retweeted by liberals and vice versa. Figure S19 displays the cross-ideological retweeting rates for each topic and ideological group. Consistent with the foregoing, we see that these rates are higher in general for non-political than political topics. At the same time, even for clearly non-political topics such as the Winter Olympics and the Super Bowl, cross-ideological retweeting rates are lower than 50%, which is the proportion one would expect if ideology was entirely irrelevant to retweeting decisions. This is consistent with the existence of ideological homophily in communication networks, insofar as liberals are more likely to retweet messages of liberals and conservatives are more likely to retweet messages of conservatives.

Figure S19: Observed Rates of Cross-ideological Retweeting



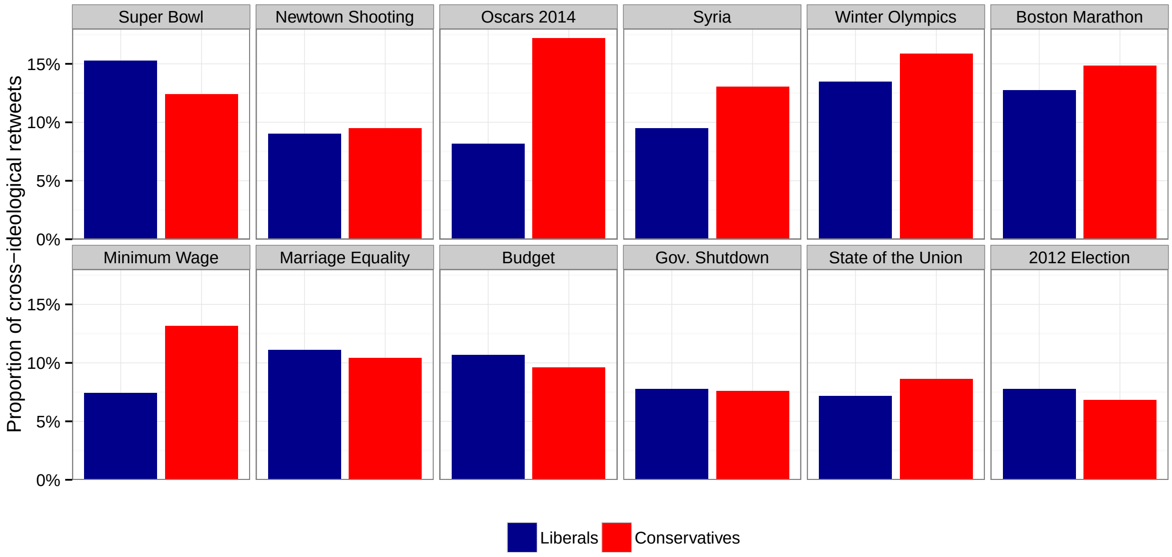
We also observe that the cross-ideological retweeting rate is higher for liberals than for conservatives with regard to most topics, although some of the estimates are biased by the tendency for one ideological group to tweet more often than the other about a given topic. For instance, liberals generated most of the tweets about the minimum wage debate, marriage equality, and the 2014 Oscars, whereas conservatives generated most of the tweets about the Super Bowl (see Supplementary Information). These ideological disparities in tweet production contribute to seemingly anomalous results, such as the observation that 60% of conservatives' retweets about minimum wage were originally posted by liberals. This percentage seems high, but it might mask a relatively high degree of selective retweeting nonetheless.

To address this problem, we computed the observed proportion of cross-ideological retweets, which refers to the number of cross-ideological retweets divided by the total number of retweets for that topic (see Figure S20). Using this measure, we observe that 6.8% of all retweets pertaining to the 2012 election involved conservatives retweeting liberals, and 7.7% involved liberals retweeting conservatives. Although this measure takes into account the unequal production of tweets, it is affected by the opposite problem: ideological disparities in the propensity to retweet. Thus, in tweets pertaining to the State of the Union, it appears as if conservatives were more likely to retweet liberals than vice versa, but this masks the fact that conservatives are more likely than liberals to retweet all types of messages.

5.2 Unbiased Estimates of Cross-Ideological Retweeting

To overcome these two issues and thus obtain the unbiased estimates of cross-ideological retweeting behavior for the results displayed in Figure 3 in our article, we split the ideological space

Figure S20: Observed Proportions of Cross-ideological Retweets



into 20 different bins of size 0.5 (e.g., ideology from -1 to -0.5) and counted the number of retweets in each of the 400 (20x20) possible combinations of bins (e.g., how many retweets have a retweeter in the interval -1 to -0.5 and a retweeted author in the interval 0.5 to 1). These counts served as the dependent variable in a poisson regression where the main independent variables were dummy variables for each possible bin for retweeter and retweeted users. To this baseline model we added two additional dummy variables: one signifying that a bin combination corresponds to a liberal retweeting a conservative, and another one signifying a conservative retweeting a liberal. The coefficients for each of these dummy variables can thus be interpreted as the difference in the propensity to engage in each type of cross-ideological (versus within-ideological) retweeting, adjusting for the baseline probability that a retweet would occur for that bin combination.

5.3 Exploring Possible Effects of Age and Personality

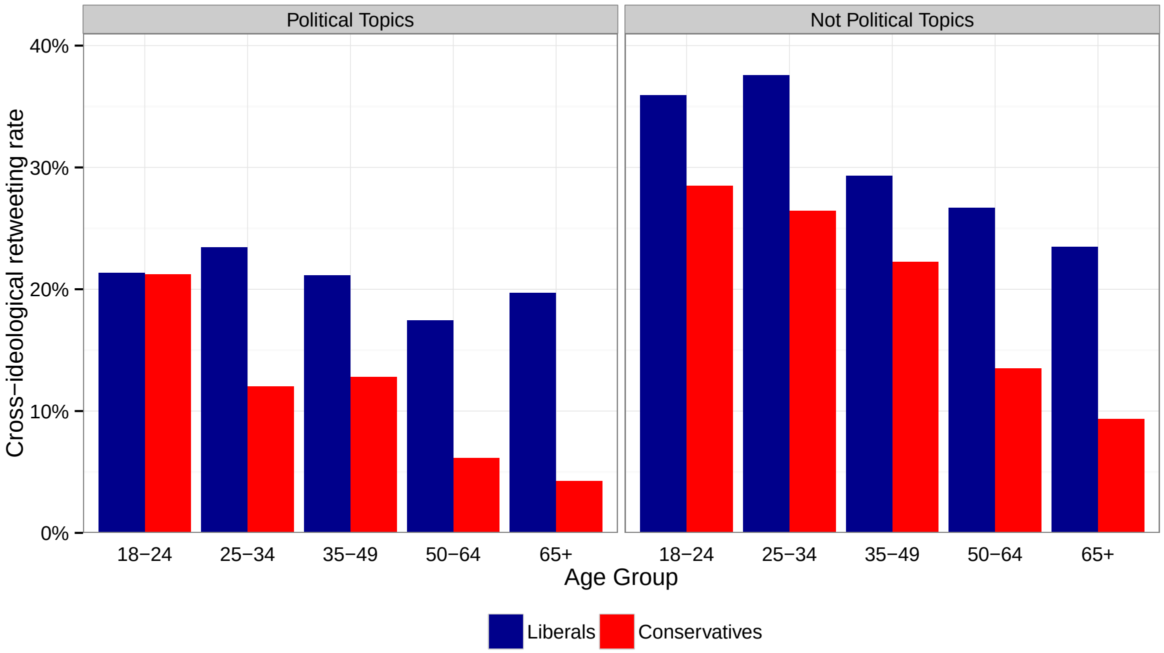
There may be several social and psychological mechanisms that help to explain the ideological asymmetry in retweeting behavior that we have observed. Here we consider two factors in particular, namely age and personality characteristics. With regard to the first, there is considerable evidence that individuals become more conservative as they grow older (e.g., Sears and Funk, 1999). It also seems probable that younger people would make more sophisticated use of Twitter as a social media platform by following more political accounts, being more active, and retweeting more often. If so, it is conceivable that some of the differences in cross-ideological interaction rates between liberals and conservatives could be attributable to age differences.

Unfortunately, we did not have access to the ages of individuals in our main sample, because Twitter does not allow users to include this information in their profiles. However, by matching

DOI: 10.1177/0956797615594620

Twitter accounts with voting registration records (described in Section 2.1), we were able to obtain the year of birth for a subset of users (27,613 voters from five different states, resulting in a data set of 166,795 retweets), and to investigate cross-ideological retweeting behavior in different age groups. This analysis demonstrates that the general pattern of results holds even after adjusting for age: liberals are more likely to retweet across ideological lines than conservatives, and cross-ideological retweets are more likely to take place in non-political than political domains (see Figure S19). The only exception is for the youngest age group (18-24): young conservatives are just as likely as young liberals to engage in cross-ideological retweeting of political messages. Results also suggest that cross-ideological retweeting is less frequent with increasing age, an observation that is consistent with theories pertaining to the crystallization of political attitudes. At the same time, these results should be interpreted with caution, given the relatively small sample size in the older age groups (e.g. only 1,250 individuals in our sample were 65 or older).

Figure S21: Observed Rates of Cross-ideological Retweeting, by Age Group



Given prior research (e.g., Carney et al., 2008; Gerber et al., 2010), we also considered the possibility that liberals, who are more likely than conservatives to engage in cross-ideological dissemination of information, would score higher on the "Big Five" personality trait of Openness to New Experiences (and lower on Conscientiousness) and that these traits might be related to retweeting behavior. Table S6 provides evidence that such a correlation may exist in our dataset, too. Although we did not have access to personality scores for our sample of Twitter accounts, we were able to conduct exploratory analyses at an aggregate level of analysis, comparing our statewide averages of ideology (see Figure S10) with statewide measures of personality computed by Rentfrow et al. (2008). As expected, we observed a significant negative correlation between openness and conservatism and a significant positive correlation between conscientiousness and

DOI: 10.1177/0956797615594620

Table S5: Correlation Coefficients Between Statewide Averages of Political Ideology and Personality Dimensions

	E	A	C	N	O
Correlation between ideology and z-scores (Pearson)	0.21 (p=0.15)	0.14 (p=0.32)	0.36 (p=0.01)	-0.07 (p=0.61)	-0.41 (p<0.01)
Rank correlation between ideology and z-scores (Spearman)	0.22 (p=0.13)	0.24 (p=0.09)	0.44 (p<0.01)	-0.06 (p=0.66)	-0.43 (p<0.01)

E=Extraversion; A=Agreeableness; C=Conscientiousness; N=Neuroticism; O=Openness

Source of Personality Data: Rentfrow et al. (2008)

Table S6: Correlation Coefficients Between Statewide Rates of Cross-Ideological Retweeting and Personality Dimensions

	E	A	C	N	O
Correlation between cross-ideological retweeting and z-scores (Pearson)	-0.10 (p=0.33)	-0.02 (p=0.83)	-0.19 (p=0.05)	0.17 (p=0.09)	0.24 (p=0.02)
Rank correlation between cross-ideological retweeting and z-scores (Spearman)	-0.08 (p=0.42)	-0.09 (p=0.35)	-0.21 (p=0.03)	0.18 (p=0.08)	0.25 (p=0.01)

E=Extraversion; A=Agreeableness; C=Conscientiousness; N=Neuroticism; O=Openness

Source of Personality Data: Rentfrow et al. (2008)

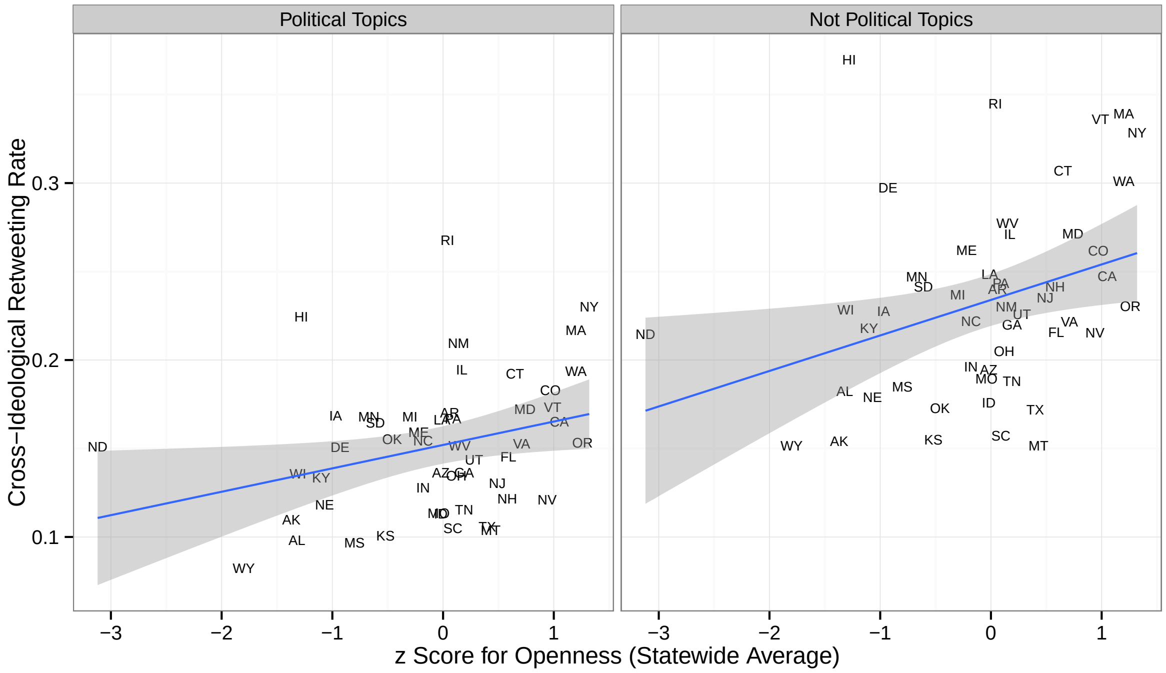
conservatism. In other words, states in which the average voter is more conservative tend to exhibit lower averages for the first trait, and higher averages for the second trait.

We also extended this analysis by comparing statewide rates of cross-ideological retweeting and averages of the Big Five personality traits across states. We did so by focusing on the random sample of 300,000 Twitter users in our dataset whose location we identified using the method described in Section 2.1. We split this sample by state, extracted the retweets in our dataset whenever the "retweeter" was one of these individuals, and then computed the proportion of retweets that took place between individuals of different ideology across various topics. In Table S6 we list the correlations between statewide personality scores and cross-ideological retweeting behavior. Results are congenial to theoretical expectations: at the aggregate level of analysis, cross-ideological retweeting is positively correlated with openness and negatively correlated with conscientiousness. Figure S22 illustrates this relationship by comparing cross-ideological retweeting rates with average openness scores for each state.

It is important to note that, in our view, personality should not be considered as a statistical confound but rather a plausible mechanism by which political ideology affects behavior (see also Carney et al., 2008). While it is difficult to extrapolate to the individual level of analysis on the basis of aggregate (statewide) correlations, recent work designed to estimate personality traits by analyzing the content of Twitter messages (Quercia et al., 2011; Golbeck et al., 2011) suggests an avenue for future research on this topic. An individual-level analysis that combines estimates of political ideology and personality traits as well as behavioral measures regarding social interaction and communication could prove especially useful in strengthening our understanding of the ways in which ideology affects political behavior.

DOI: 10.1177/0956797615594620

Figure S22: Average Values of Openness and Cross-Ideological Retweeting Rates, by State



References

Bafumi, J. and Herron, M. C. (2010). Leapfrog representation and extremism: A study of american voters and their members in congress. *American Political Science Review*, 104(03):519–542.

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

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bonica, A. (2013a). Database on ideology, money in politics, and elections: Public version 1.0 [computer file]. Stanford, CA: Stanford University Libraries. http://data.stanford.edu/dime.

Bonica, A. (2013b). Mapping the ideological marketplace. *American Journal of Political Science (forthcoming)*.

Bradley, A. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Bryant, J. and Miron, D. (2004). Theory and research in mass communication. *Journal of communication*, 54(4):662–704.

Carney, D. R., Jost, J. T., Gosling, S. D., and Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, 29(6):807–840.

DOI: 10.1177/0956797615594620

Enelow, J. and Hinich, M. (1984). *The spatial theory of voting: An introduction.* Cambridge Univ Pr.

Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.

Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M., and Ha, S. E. (2010). Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review*, 104(01):111–133.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE.

Greenacre, M. (2010). *Correspondence analysis in practice.* CRC Press.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.*

Jackman, S. (2014). Estimates of members' preferences, 113th u.s. house and senate. Retrieved on June 1st, 2014.

Jessee, S. A. (2009). Spatial voting in the 2004 presidential election. *American Political Science Review*, 103(01):59–81.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lax, J. and Phillips, J. (2012). The democratic deficit in the states. *American Journal of Political Science*, 56(1):148–166.

Lazarsfeld, P., Berelson, B., and Gaudet, H. (1944). *The People's Choice: How the Voter Makes up his Mind in a Presidential Election.* New York: Duell, Sloan and Pearce.

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, 16(4):356–371.

Maestas, C. D., Buttice, M. K., and Stone, W. J. (2014). Extracting wisdom from experts and small crowds: Strategies for improving informant-based measures of political concepts. *Political Analysis, forthcoming.*

Martin, S., Brown, W. M., Klavans, R., and Boyack, K. W. (2011). Openord: An open-source toolbox for large graph layout. In *IS&amp;T/SPIE Electronic Imaging*, pages 786806–786806. International Society for Optics and Photonics.

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

Nenadic, O. and Greenacre, M. (2007). Correspondence analysis in r, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3):1–13.

Poole, K. and Rosenthal, H. (1997). *Congress: A political-economic history of roll call voting.* Oxford University Press, USA.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE.

Rentfrow, P. J., Gosling, S. D., and Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, 3(5):339–369.

DOI: 10.1177/0956797615594620

Sears, D. O. and Funk, C. L. (1999). Evidence of the long-term persistence of adults' political predispositions. *The Journal of Politics*, 61(01):1–28.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160.

Thomas, K., Grier, C., and Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats*, pages 13–13. USENIX Association.

Yardi, S., Romero, D., Schoenebeck, G., et al. (2009). Detecting spam in a twitter network. *First Monday*, 15(1).