Revealed Public Opinion on Twitter: The Supreme Court of the United States Same-Sex Marriage Decisions

Tom S. Clark  
Department of Political Science  
Emory University  
tom.clark@emory.edu

Jeffrey K. Staton  
Department of Political Science  
Emory University  
jeffrey.staton@emory.edu

Eugene Agichtein  
Department of Math and Computer Science  
Emory University  
eugene@mathcs.emory.edu

Yu Wang  
Department of Math and Computer Science  
Emory University  
yu.wang@emory.edu

March 31, 2014

Abstract
1 Introduction

Modern political science has typically sought to reveal mass preferences via public opinion polling. Although polls are ideal for many tasks, they are less so for others. Naturally, polls are not well-designed to capture revealed opinions as they manifest naturally in the world, where a person chooses to express a position without necessarily being prompted. It is also hard, if not impossible, to field public opinion polls on many topics at regular intervals. And since some political events emerge or change suddenly, a polling approach may be impossible for some studies of the dynamic relationship between opinion and politics. This is particularly true in the context of political institutions and policies that are more unpredictable and of lower salience to the popular media. Yet, we should be measuring features in these political contexts and we should be measuring what people say publicly as much as what they reveal when prompted by a survey researcher. Good answers to important questions in many research areas in political science depend on it.

This paper introduces a near real time approach for tracking revealed features of public opinion regarding the Supreme Court of the United States and the political issues that it tackles each year. We generate measures of concepts regarding the Supreme Court itself, a variety of features of the opinions individuals express about the Court and its decisions, as well as features of the individuals whose opinions we track. To do so, we combine machine learning techniques for natural language processing of microblog text with scaling techniques developed for measuring latent ideology of discussants in a legal network (Clark and Lauderdale 2010). Pairing social media with developments in the quantitative analysis of text provide a potential source of detailed, accessible, and extensive information about public attitudes about political issues. People increasingly engage in political discourse on microblogs, such as Twitter (e.g. Vergeer, Hermans and Sams 2013, Guo and Saxton 2012, Khondker 2011). What they say can be revealing about the political issues they care about and what they think about those issues. This is particularly promising in the context of political institutions like the Supreme Court, which are less frequently the subject of political polling. As our study suggests, the approach we take is useful for the evaluation of core questions in the study of public opinion and the Supreme Court, but the advantages are broader. Microblog text can be used to generate daily measures of issue salience. We also suggest how we can develop ideological measures of the Twitter users, as well as a variety of demographic features of Twitter users.
themselves. And because microblogs like Twitter are used in multiple countries and quantitative text analysis can be done in multiple languages, our approach is extensible to a global level. For these reasons, a careful approach to the measurement of features of public discourse regarding the Supreme Court can influence a wide variety of research questions, both in American politics and comparatively.

Although many applications are possible, we focus here on a particular study, in which we consider the effect of the Supreme Court’s 2013 same sex marriage decisions in United States v. Windsor and Hollingsworth v. Perry, on revealed opinion regarding same-sex marriage.¹ We track a panel of frequent Twitter users, who commonly commented on the issue; and, we compare the findings in the panel to a very large daily random sample of Twitter entries on the issue. Consistent with Franklin and Kosaki (1989) and Johnson and Martin (1998) we find that the Court’s opinion was polarizing. We also find considerable evidence suggesting that the decision influenced how people expressed their policy opinions. We uncover a pronounced emotional effect concerning both the intensity and anger of expressed opinions in ways related to revealed opinion expressed before the decision. That Supreme Court decisions in salient cases impact the emotionality of policy discourse is not surprising in light of research on emotional responses to major political and economic events (e.g. Bollen, Mao and Pepe 2011) but it has yet to be incorporated in research on the role of the Supreme Court in American policy making. The finding has important implications for a wide variety of subjects, ranging from work on how political entrepreneurs might use emotion to advance their policy goals (Lupia and Menning 2009) in the wake of significant political events to more particular concerns regarding how peak courts judges are incentivized to manage their images (Baum 2006, Davis 1994, Staton 2010).

In what follows we first review briefly familiar empirical implications derived from mass opinion research conducted on the Supreme Court of the United States. We then introduce our approach for tracking revealed public opinion on Twitter and present our findings, derived from both a daily random sample of Tweets and a panel of frequent Tweeters who we tracked for six months. We conclude by discussing the implications of our approach for other research questions in judicial

politics as well as a web-based teaching tool that we have constructed to encourage use of the new data we have generated.

2 Theoretical Motivation

American democracy involves great tensions between distinct visions of democratic governance (Powell 2000). At its core the American system is committed to majoritarian outcomes, as evidenced by its classically majoritarian electoral system. Yet the American constitution also fragments the decisional processes of governance across multiple institutions in search of multiple checks on authority. It admits that multiple governing pluralities are possible and in incentivizing cooperation among these pluralities, we advance a key goal of the proportional vision. Similarly, its liberal constitution implies limits on the power of the American state, limits beyond which no governing coalition should be able to go.

Research on mass opinion and the Supreme Court speaks to fundamental and related normative concerns about the role of constitutional review in American governance, which emerge in a political system possessing democratic institutions at cross-purposes. Perhaps most obviously, scholars have asked whether it is legitimate in a democracy to allow unelected judges the ability to constrain the will of a majority through constitutional review (Bickel 1962). This concern is heightened by the general finding that the decisions of U.S. Supreme Court justices are powerfully related to the judges’ personal, ideological policy preferences (e.g., Segal and Spaeth 2002). On the other hand, a broad literature suggests that the Court’s decisions, especially over the long-run, reflect well mass preferences over policy outcomes (e.g. Ura 2013, Mishler and Sheehan 1996). Likewise, a similarly broad literature suggests that the Court is influenced by outside political actors, including most naturally the U.S. Congress and President (e.g. Carrubba and Zorn 2010, Martin 2006), but also by interest groups and lawyers (Spriggs and Wahlbeck 1997). These strands of research on the Supreme Court suggest alternative normative concerns. If the Supreme Court is in fact capable of producing counter-majoritarian outcomes, then the majoritarian element of American democracy is threatened. Yet, of course, American constitutional design has a powerful liberal element, and so a counter-majoritarian strain is to be welcomed on that account. Yet again, if the Court is
constrained by external, majoritarian influences, then the liberal vision of the Court’s role may be undermined, as well.

It is in this context that research on the capacity of the Supreme Court to influence public opinion becomes of great normative importance. One line of thought, beginning with Dahl (1957), suggests that the Supreme Court of the United States has a unique capacity among major institutions of American government to leverage its legitimacy in order to change mass opinion regarding salient policies (Generally, see Gibson and Caldeira 2009, Caldeira and Gibson 1992, Gibson 1989). If this is possible, then counter-majoritarian concerns with constitutional review are misguided, since majoritarian views on issues are endogenous to the Court’s opinions. Likewise, concerns about the inability of the Court to hold majorities to constitutional limitations are similarly misguided, because the Court may be able to influence its power and corresponding ability to influence policy by influencing the incentives of elected officials through changing the preferences of their constituents. If the Dahlian hypothesis is correct, then the Supreme Court’s same-sex marriage decisions should have resulted in a measurable change in opinion.

Research on this implication of Dahl’s hypothesis has been extremely mixed (e.g. Mondak 1994, Hoekstra 2003). A primary finding in survey research is that the Court is polarizing, creating more supportive opinions of the policies it reviews among those who supported the policy before the decision and more negative opinions among those who opposed the policy prior to the decision (Franklin and Kosaki 1989, Johnson and Martin 1998). This is especially true in landmark decisions, and so the Court’s opinions in Windsor and Hollingsworth present excellent opportunities to evaluate the polarization hypothesis. Experimental research has found that the Court is only likely to be able to change opinion among those who do not already possess strong views or who have not been exposed to much information about the policy. And recent research by Ura (2014) suggests that the short run effect of patterns of Supreme Court decisions is to undermine very general support for these patterns, while the long-run effect may be to enhance support. In short, as Supreme Court policy becomes increasingly liberal (conservative) public support for liberal (conservative) policy declines in the short run but increases in the long run. Ura’s research is conducted at the aggregate level, but the theoretical argument motivating it, the theory of “thermostatic” response, has clear micro-level implications. In the context of the same-sex marriage cases, we ought not to observe strong effects among liberal and conservative discussants since the decisions would have
been pleasing to liberals and since conservatives would have very little room for change. On the other hand, a thermostatic response at the micro-level ought to be observable in moderates, who would have been expected to reflect more conservative opinions regarding same-sex marriage in the wake of the decision.\textsuperscript{2}

In addition to these well-known effects on public opinion we focus on new features of opinion that have yet to be evaluated in Supreme Court research. Specifically, we will consider the extant to which the Court’s decisions produced a change in the intensity and anger with which Twitter users express their opinions. Although not the subject of prior research, existing theory does speak to these opinion features. If Dahl’s legitimation hypothesis is correct, then we should observe that individuals who were opposed to same-sex marriage prior to the decision were less angry and intense in their opposition after the decision. The Court should have been able to produce increased tolerance among the initially opposed (\textbf{Tom: Cite to Gibson?}). On the other hand, if the Court is polarizing in these types of cases, then we might expect that those individuals who were initially opposed to the decision became more intense and angry in their responses to the decisions. Likewise, those who were initially in favor of the policy should have reduced the intensity and anger of their expressions, having received a significant policy concession from the peak court of the United States.

To summarize then, our primary research goal will be to evaluate three simple predictions about the effect of Supreme Court decisions on public opinion: (1) The Supreme Court pulls opinion in the direction of the policy outcome favored by its decision (Dahl), (2) The Supreme Court polarized public opinion, especially in landmark cases, and (3) The Supreme Court pushes opinion away from the policy outcome favored by its decision, especially among moderates. Our secondary goal will be to track the intensity and anger of revealed opinion, evaluating two predictions: (4) the Supreme Court’s decision reduced the intensity and anger of all expressions regarding same-sex marriage, and (5) the decision had polarizing effects on anger and intensity, making the initially opposed more intense and angry while making the initially supportive less intense and angry.

\textsuperscript{2}It may be argued that the thermostatic response argument operates only in the aggregate, once multiple decisions are accumulated. But micro-processes must underlie the argument. And some decisions should weigh more than others. Indeed, if that is true then some decisions may be the equivalent of groups of other decisions. Our view is that the same-sex marriage decisions represented a massive policy statement for the Court in a highly salient way. In 2013, there was no other decision like them in this context. If there were decisions in 2013 capable of moving opinion, these would have been they.
3 Data and Measures

Our approach began with the development of keywords to identify Tweets relating to the Supreme Court’s involvement in two same-sex marriage cases—*Hollingsworth v. Perry* and *US v. Windsor*. In the former case, the Supreme Court considered the constitutionality of California’s Proposition 8, a constitutional amendment enacted via ballot referendum that made same-sex marriage illegal. In the latter case, the Supreme Court considered the constitutionality of Section 5 of the Defense of Marriage Act, which prohibited the United States Government from recognizing otherwise legal same-sex marriages. The Court decided both of these cases (essentially) in favor of same-sex marriage recognition on June 26, 2013.

The architecture of data collection and analysis is outlined in Figure 1. First, the Twitter stream is filtered by using topically relevant keywords, language filters, and desired author characteristics. The resulting set of relevant Tweets are automatically annotated with a pipeline of topic-specific classifiers, annotating each tweet with estimated author demographic information, topic salience, support, emotional intensity, and other quantities of interest, which are then used for opinion estimation. The Tweet text, as well as the original and inferred metadata, are stored in a database for later search, analysis, and visualization. A sample of the tweets is also manually coded for validation of the automated classification, and to provide continuous stream of up-to-date training data for the classifiers. Finally, all the summary and estimated statistics, as well as tools for developing new topics of interest, are made available through a web server, on which we publicly release our data. We describe first the collection of Tweets and section the construction of relevant measures from those data.

3.1 The data collection process

Twitter provides streaming API to deliver Tweets in real-time. There are several types of filters can be applied to the API. A Language filter selects only Tweets written in that particular language will be delivered. The language of a Tweet is determined by Twitter’s language classifier. A Keyword / phrase filter can represent topics and help select on-topic Tweets. If specified, Tweets containing at least one of the keywords or phrases will be delivered. A User filter can help tracking user groups of interest. If specified, all Tweets authored or retweeted by the particular users will be delivered. If
Figure 1: Architecture of the Twitter-based opinion estimation system.

none of the filters is specified, Twitter streaming API will constantly produce 1% random sample of all Tweets. When the filters are applied, the API will send all the Tweets that match the filtering parameters as long as the volume is lower than 1% of all Tweets. Once the percentage of matched Tweets is higher than 1%, the API will only return random sampled matched ones up to 1%.\(^3\)

To obtain “gay marriage” Tweets, we set the API language filter to be English and developed “gay marriage” keywords. The “gay marriage” keywords are “ssm”, “same sex marriage”, “DOMA”, “Prop8”, and “gay marriage”. Our procedure produced over 2,500,000 Tweets between March 26, 2013 and August 10, 2013, with 87,575 daily Tweets on average. On June 26, 2013, the day the Supreme Court decided its two gay marriage cases, we collected 335,399 Tweets on the gay marriage topic. We developed keywords and requested a sample of Tweets that matched the keywords. The keywords used are “ssm”, “same sex marriage”, “DOMA”, “Prop8”, and “gay marriage”. We also specified the language of sampled Tweet to be English only. Our procedure

\(^3\) Additionally, we plan to automatically discover additional topical keywords and hashtags by mining the text of the Tweets, in conjunction with coder decisions about the relevance of Tweets. That is, we plan to automatically identify misleading keywords (e.g., ambiguous terms), and words that are highly predictive of relevance but are not included in our set of keywords. For this, we would use well established computer science techniques for topic detection and tracking (e.g., ??). Note that these techniques can also be applied for cross-language topic tracking, i.e., to also identify relevant Tweets on the same topic in other languages, notably Spanish (??) – necessary for the subsequent stages of our proposed work (We will further elaborate on the issues of cross-language sentiment analysis below). To evaluate the performance of these discovered keywords, we would use our ongoing manual labeling efforts as well as the feedback from the users on the courtometer website, to evaluate whether those terms should be included in the set of keywords for each topic. We are aware of the potential issue of Twitter manipulation through bots, crowdsourcing, organizations, or colluding individuals. This is an active area of research in computer science, with existing techniques that would allow us to at least detect manipulation retroactively (e.g., ??) (and remove such Tweets from our data), and potentially extend this to near-real time detection of rumours and mis-information (e.g., ?), to reduce the effect on the resulting opinion estimates.
produced over 2,500,000 Tweets between March 26, 2013 and August 10, 2013, with 87,575 daily Tweets on average. On June 26, 2013, the day the Supreme Court decided its two gay marriage cases, we collected 335,399 Tweets on the gay marriage topic. The precision (i.e., fraction of retrieved Tweets manually verified to be on-topic) was 92.3%. This provides strong support for the accuracy of our topic filtering techniques based on designing a set of precise keywords for each topic.

In addition to the daily sample of Tweets, in early May 2013 we identified about 700 frequent Tweeters regarding the “gay marriage” topic. Thereafter, we tracked every Tweet those individuals posted, thereby constructing a panel of Twitter accounts for which we had the universe of Tweets over a 3-month period. Critically, the individuals in our panel do not know they are in a panel; we are simply collecting all of their public utterances on Twitter. We refer to the daily sample of Tweets as our “random sample” of Tweets, recognizing, though, they do not represent a national probability sample, or anything approaching one. We refer to the set of Tweets from our panel as our “panel sample.”

3.2 Measures

3.2.1 Tweet features: a classifier approach

To measure the content of the Tweets we collected, we adopted classification algorithms specifically designed to detect the three features of Tweet content in which we are interested: supportiveness, intensity, and anger. Each of the measures is handled by one classifier. For example, to estimate supportiveness of Tweets, we classify every Tweet to one of the following classes: supportive, neutral, and opposing. For intensity, the classes are intense vs. non-intense. And for anger, the classes are angry, neutral, and happy.

We begin by manually labeling a set of “training” Tweets. We developed detailed labeling instructions and hired Political Science graduate students as labelers. We trained them in person. For training, we labeled 1,400 Tweets for the “gay marriage” topic, sampled at the rate of 100 per day, over the period of two weeks, immediately prior and subsequent to the DOMA decision. Coding rules for the research assistant tasks can be found in the appendix. Inter-coder reliability, as measured by the Fleiss’ κ statistic is highest on the “relevance” and “support” items and lowest
on “anger”. With the human-labeled Tweets in hand, we developed our classification algorithms to automatically label Tweets.

To classify Tweets, we developed several groups of features to represent them in feature space:

- Popularity: Number of times the message has been posted or favored by users. As for a Tweet, this feature means number of Retweets and favorites.

- Capitalization and Punctuation: It has been shown that capitalization and punctuation carry valuable signals of emotional intensity in sentiment analysis.

- The text: Unigram, bigram, and trigram of the Tweet text.

- Character N-gram in text: Trigram and four-gram of the characters in the Tweet text.

- Sentiment score: The score is computed via a comprehensive sentiment dictionary, SentiWordNet\(^4\), as well as stylistic features described in ?.

We experimented with a variety of automated classification algorithms, and for this experiment report the performance of Naive Bayes algorithm (simple, fast, and shown to be surprisingly robust to classification tasks with sparse and noisy training data). The classification performance, using 10-fold cross validation, for supportiveness, intensity, and anger of the tweet text is reported in Table 1. Even with the small amount of the available training data, the classifier is able to accurately identify supportive and neutral Tweets, with precision of 73% and 76% respectively, but not the more rare occurrences of opposing Tweets. Classifier performance is also acceptable for distinguishing emotionally intense from dispassionate (factual) Tweets, of which there was a 69% majority in our sample.

Indeed, we find that our classifier reliably recovers the levels of support, intensity, and anger among on a daily basis. Figure 2 shows a comparison of the classifier’s estimates of each of these features and the “ground truth”—the labels our research assistants assigned. FINISH!!!

The results of our classification are reported in Figure 3. The top row reports the results from our random sample of Tweets, and the bottom row reports the results from the relevant Tweets among our panelists. The three columns show the daily average value for each of the dimensions we label, along with a smoothed trend. FINISH!!

\(^4\)http://sentiwordnet.isti.cnr.it/
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Value</th>
<th>Precision (%)</th>
<th>Coverage (%)</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>Supportive (48%)</td>
<td>73</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neutral (45%)</td>
<td>76</td>
<td>67</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Opposed (7%)</td>
<td>17</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Emotional</td>
<td>Intense (31%)</td>
<td>56</td>
<td>60</td>
<td>73</td>
</tr>
<tr>
<td>Intensity</td>
<td>Dispassionate (69%)</td>
<td>81</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>Sentiment</td>
<td>Pleased (10%)</td>
<td>48</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td>Neutral (79%)</td>
<td>84</td>
<td>78</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Angry (11%)</td>
<td>24</td>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Preliminary sentiment classifier performance for Tweets on “Gay Marriage” topic, 10-fold CV, 1,400 tweets.

Figure 2: Comparison of human-assigned labels and classification algorithm-assigned labels. The open points show daily averages of human-labeled Tweets. The solid points show daily averages of classification algorithm-labeled Tweets.
Trends on three sentiment features among the random sample and panel Tweets. The first column shows supportiveness for same-sex marriage; the second column shows Tweet intensity; and the third column shows Tweet anger. The top row shows the random sample, and the bottom row show the panelists’ Tweets. The points show daily averages and are sized proportional to the number of Tweets collected that day. The lines are fit lines from a GAM smoother; shaded areas show 95% confidence bands.
3.3 Tweeter ideology: a scaling model

The final feature we measure using Twitter data is the political ideology of our panel members. To do so, ideally, we would administer a survey to each Tweeter in our data to develop an index of conservatism/liberalism for each individual. This is not possible, for several reasons. Twitter’s Terms of Service prohibit us from directly contacting the individuals whose Tweets we monitor and collect. However, some users voluntarily provide some indicators of their ideology through their public profiles. For example, some Tweeters self identify as “Republican” or as “conservative”. Indeed, in our sample, 176 of the 522 Tweeters indicate their political ideology in some form in their profile. Unfortunately, that percentage is still relatively low, and one might worry that relying on self-reported ideology in a public profile induces substantial selection bias in terms of which Tweeters for whom we have a measure of ideology. To overcome these limitations, we develop a latent variable model to estimate latent ideology for the Tweeters. Our model rests on the assumption that who one “follows” on Twitter is a manifestation of latent ideology. In other words, the social network to which one belongs on Twitter is informative about one’s underlying political leanings.

The behavioral model of “following” on Twitter that underlies our model is intuitive—we assume that all individuals on Twitter have a location in a latent dimension. The closer two individuals are on that dimension, the more likely one is to follow the other. The intuition is similar to that of Clark and Lauderdale (2010), who assume that Supreme Court opinions have latent locations in an underlying dimension and that relatively closer opinions are more likely to cite each other favorably. Of course, we recognize that not all subjects on Twitter may not be equally discriminating across the latent dimension; some individuals may be followed by Tweeters across the latent dimension, whereas others may be followed only by people very close to them. Our model allows the effect of distance between a Tweeter and an individual to be followed to vary across “followed” individuals.

Consider a set of \(N\) Tweeters and \(J\) followed individuals. Let \(F\) be an \(N \times J\) matrix, where \(f_{ij} = 1\) if Tweeter in row \(i\) follows individual in column \(j\) and 0 otherwise. We assume each Tweeter and followed individual has a location \(\theta\) in a latent 1-D space. Formally, our model is given by

\[
Pr(f_{ij} = 1|\alpha, \beta_j, \theta_i, \theta_j) = \Phi \left( \alpha + \beta_j \cdot (\theta_i - \theta_j)^2 \right)
\]

(1)
where \( \Phi \) is the cumulative normal distribution. Of course, the model given by equation (1) is not identified without further restrictions. We identify the scale by assigning a prior distribution to the unobserved ideal points, such that

\[
\theta \sim N(0, 1).
\]

Even with that constraint, the model is still only identified up to a polar rotation. As we show below, self-declared political ideology is well-correlated with the latent dimension we recover. We therefore select the polarity of the model that makes self-declared conservatives more likely to be at the right end of the dimension and self-declared liberals more likely to be at the left end of the dimension. Finally, we assign diffuse normal priors to the intercept and slope parameters. Specifically, we assume that the intercept has an improper uniform prior and that the slope parameter has a strictly negative uniform prior, enforcing the assumption that increasing distance decreases the propensity to follow someone else: \( \alpha \sim U(-\infty, \infty) \) and \( \beta_j \sim U(-\infty, 0) \) for \( j = 1, \ldots, J \). We program and estimate our model in \textit{R} and \textit{JAGS} (R Development Core Team 2011, Plummer 2003).

### 3.3.1 The data

To estimate individual Tweeters’ latent ideology, we collect data from their Twitter profiles. Specifically, we identify the universe of Tweeters that each individual Tweeter “follows.” We also collect each Tweeter’s public profile and identify whether the Tweeter self-identifies as “liberal”, “Democrat”, “conservative”, or “Republican”. We refer to the Tweeters in our panel as “Followers” and the individuals they follow as “Followees.” There are \( N \) Followers and \( J \) Followees. We construct a \( N \times J \) matrix \( A \), where \( a_{ij} = 1 \) if Follower \( i \) follows Followee \( j \). This matrix is incredibly sparse. There are 653 Followers and 162,509 Followees. The average number of Followees for each Follower is 1168 (median is 706), and the average number of Followers for each Followee is 5 (median is 2). This is because most Followees are not the type of salient individuals likely to have many Followers; as a consequence, they are also not likely to be useful in discriminating individuals into latent ideology. Thus, we subset the data and retain only those Followees in the top 1% of the distribution of the number of Followers one has. This results in 1551 Followees remaining in our data. Similarly, we subset the data and retain only those Followers in the top 80% of the distribution of the number of Followees for each Tweeter. This results in 522 Followers remaining in our
data. (The distribution of Followees per Follower is extremely right skewed.) We then estimate the latent ideal points, $\theta$ for the Followers in our data (as well as the Followees, of course). Of the 2052 individuals in the matrix, 21 appear as both Followers and Followees.\(^5\) This suggests that rather than collecting data on Tweets among an insular network of individuals talking to each other, we have data on Tweets by people who are part of broader Twitter networks.

### 3.3.2 Estimates

The estimates we report below are based on a 10,000-iteration simulation (thinned by 10), after a discarded 10,000-iteration burn-in period. Standard diagnostic tests suggest the model converges and mixes within the burn-in period.\(^6\) Figure 4 reports the distribution of posterior mean estimates of $\theta_i$ for each of the 522 Twitter Followers in our data. We divide the data into three groups—those self-identifying as conservative, those self-identifying as liberal, and those who do not declare a political preference. Conservatives are individuals who self-identify as conservative or Republican in their profile; Liberals are individuals who self-identify as liberal or Democrat in their profile. This plot provides strong evidence of the validity of our measurement model. Self-identified liberals are distributed around a mode to one end of our scale, whereas self-identified conservatives are distributed around a model to the other end of our scale. The average ideal point for self-identified liberals is $-0.14$, whereas the average ideal point for self-identified conservatives is $0.22$, and the $t$-statistic for the difference between these two groups is $12.3$, (147 df).

What is more, and importantly, Tweeters who do not self-declare as either liberals or conservatives are distributed bimodally. There is a mode right around the self-identified liberals, and another mode right around the self-identified conservatives. This finding is consistent with the interpretation of our estimates as a measure of ideology, in a world in which self-identification as liberal or conservative on Twitter is not necessarily associated with being an ideological extremist; rather whether one chooses to self-identify as liberal or conservative may be a function of features other than their agent ideological predispositions, such as interest in politics, profession, or general public profile. In either event, these data provide strong evidence of the validity of our measurement

---

\(^5\)We do not currently constrain those individuals to have just a single ideal point. In future iterations, we will.

\(^6\)For example, for a single chain, none of the $z$-scores for the individual parameters is able to reject a null hypothesis that the mean of the first 10% of the chain and the last 50% of the chain are equal. In addition, when simulating multiple chains, the Gelman-Rubin statistics for the myriad parameters are all nearly 1.
Figure 4: Estimated ideal points of Tweeters in the panel study. Figure shows posterior mean estimates of latent ideology for individuals in our Twitter panel. Those labeled Liberals are individuals who identify themselves as either liberals or Democrats in their Twitter profiles. Those labeled Conservatives are individual who identify themselves as either conservatives or Republicans in their Twitter profiles. Those labeled Undeclared did not indicate a political preference in their Twitter profiles. Estimates based on a 10,000-iteration simulation after a discarded 10,000-iteration burn-in period. Plot shows distribution of posterior means.
strategy and allow us to examine in more fine-grained detail how different individuals react to the Supreme Court’s SSM decisions.

Finally, it bears mentioning who some of the most liberal and conservative Tweeters in our sample are. We focus on the Followees, as they tend to be high-profile institutions (they have many followers, by definition), whereas the Followers (members of our panel) tend to be private individuals. However, because all of these individuals are scaled in a common dimension, the Followees are useful for interpreting what it means to be at one end of the dimension. Also, recall that this panel is constructed of Tweeters who often Tweet about same-sex marriage. Among the individuals at the far-right end of the dimension—those we interpret as “most conservative”—are “Team Santorum KY”, a Rick Santorum-led political group; “Roaring Republican”, a conservative group whose profile says that “#Liberalism is a disease”; “Patriot Airborne”, a self-described “Proud #NRA member”; and “Andrea Silver”, a Tweeter whose profile says “Christian, Pro-Israel, Pro-constitution.” Among the most liberal members of our group are “Big Gay News”, a gay media account; “SEIU”, the official Twitter account of the Service Employees International Union; and “SenateDems”, the Twitter account of the Senate Democratic Policy and Communications Committee; and “GOP Unplugged”, a Twitter account that mocks Republican politicians.

4 Analysis

4.1 The effect of the Court’s decision on supportiveness

The various theoretical hypotheses outlined above have direct implications for the patterns we should see in the supportiveness of Tweets in our sample. Specifically, the Dahlian hypothesis is that the Court’s decision will shift the public towards supporting same-sex marriage. By contrast, the thermostatic model of opinion predicts that at least initially opinion will move away from supportiveness. According to this model, the liberalism of the Court’s same-sex marriage decisions should induce a reactive conservatism among the public. Finally, the polarization hypotheses predict opinion will change heterogenously—supporters will become increasingly supportive, whereas opponents will become increasingly opposed to same-sex marriage.

Consider first the patterns we saw above in the left-most column of Figure 3. In both the random sample and the panel, there is a noticeable uptick in support in the immediate wake of the
Court’s decision, though it is short-lived in the random sample and appears to be so in the panel. (Unfortunately, our panel ends sooner after the decision than does our random sample.) While on its face this may seem to support the Dahlian perspective, there are important caveats to note.

First, we cannot document evidence of any long-run effect. The limited period we study precludes inference about long-lasting effects. Second, while there is a brief uptick, the trend, especially in the panel, is a negative one over time. In fact, when we consider the distribution of supportiveness before and after the Supreme Court’s decision, we find that the average level of support is lower after the decision than before. In the random sample, the average level of support before the decision is 1.71, compared with 1.68 after the decision. The $t$-statistic for this difference is 43.0. This extremely large test statistic, however, is not surprising, given we have roughly 2.5 million Tweets. In the panel data, the average level of support is 1.18 before, compared with 1.12 after. The $t$-statistic is 12.5. Again, the size of the test statistic is not surprising, as we have over 15,000 Tweets before the decision and nearly 9000 after. However, even if we limit our inquiry to Tweets during the week before and the week after the decision, we find differences of comparable magnitude. Among the random sample, the means before and after are 1.68 and 1.56 ($t = 57.2$), respectively; among the panel the means before and after are 1.14 and 1.10 ($t = 3.3$), respectively. What is more, these effects, which statistically meaningful, may not be of substantive consequence. Recall that a score of 1 on our scale is the “neutral” category. What we observe here is a change of about 0.1 standard deviations away from Support towards neutrality.

Turning from the macro-level predictions from the Dahlian and thermostatic models, we consider the micro-level predictions from the thermostatic and polarization arguments. Both models predict (different) heterogeneous patterns in opinion change. The thermostatic model predicts that the most movement should be among political moderates. By contrast, the polarization model predicts that changes in supportiveness will be positive among “early supporters” and negative among “early opponents.” To investigate whether these predictions, we turn to our panel. Among our 522 panelists, 396 Tweeted about SSM before and after the SSM decisions. With our estimates of Tweeters’ conservatism in hand, we can investigate how these individuals changed, in terms of the features of their Tweets, after the SSM decisions.

In the left-hand panel of Figure 5, we report the signed change in average supportiveness among for each of our panelists (only those Tweeting both before and after the decision) against the
Figure 5: Change in panelists’ supportiveness after the Supreme Court’s decisions, as a function of conservatism and early support. Left-hand panel shows signed change in panelists’ average supportiveness as a function of their estimated latent conservatism. Right-hand panel shows signed change in panelists’ average supportiveness as a function of their average support before the Supreme Court decisions. All change measures exclude Tweets on the day of the Supreme Court’s decisions. Lines are locally-weighted GAM regressions with 95% confidence intervals. Points are sized proportional to the number of Tweets collected from each panelist.

Tweeter’s estimated conservatism. This is simply each Tweeter’s average Tweet label (0=opposed, 1=neutral, 2=supportive) after June 26, 2013, minus the average Tweet label before June 26. The thermostatic model predicts a U-shaped pattern, where the most moderate Tweeters would exhibit the greatest shifts towards opposition. These data do not support the thermostatic prediction. There appears to be no relationship between the latent conservatism of a Tweeter and the direction or magnitude of her change in supportiveness. The line we see in the left-hand panel of Figure 5 (the fit from a generalized additive model (GAM)) is essentially flat, and if we regress change in support on conservatism, we find a negative but substantively small and statistically insignificant correlation.

Turning to the polarization model, the right-hand panel of Figure 5 reports the signed change in supportiveness as a function of average support before the decision. The polarization model predicts a positive slope that crosses 0. This would indicate that the most supportive Tweeters increase their supportiveness, whereas the least supportive Tweeters decrease their supportiveness. Indeed, this is precisely what we find. Tweeters whose support levels were below average (below 1.17 on our scale) generally have a negative change in supportiveness: $-0.17$ on average. Tweeters whose support levels were above average generally have a positive change in supportiveness: $0.07$ on average. The $t$-statistic for this difference is $6.9$. Indeed, if we regress the signed change in support on a Tweeter’s
early support level, we find a substantively large and statistically significant correlation ($\hat{\beta} = 0.64$, $se = 0.06$). This finding is consistent with a polarizing effect of the Supreme Court’s decisions in the same-sex marriages cases on supportiveness for same-sex marriage among Tweeters in our panel.

4.2 Emotional effects of the Court’s decision

The trends in supportiveness for same-sex marriage both within our random sample and among our panelists suggest a polarizing effect of the Supreme Court’s decisions. Our data, though, allow us to go further than simply examining the supportiveness embodied in their Tweets. Specifically, we can examine how Tweeters talk about same-sex marriage. The measures of anger and intensity described above, in particular, provide insight into how the emotionality of Twitter discourse about same-sex marriage shifted in the wake of the Court decisions.

Figure 3 above documents the aggregate trends in intensity and anger within the random sample and among our panelists. With respect to intensity, we saw a potential, though minor, uptick in the wake of the Court’s decisions, though, at least in the random sample, it is short-lived. With respect to anger, we see evidence of a considerable increase in anger after the decision. On it’s face, though, this seems strange, given the overwhelming supportiveness of our Tweeters. Why would supportive Tweeters become angrier after the Court’s decision?

In Figure 6, we document the trends in intensity and anger, distinguishing among supportive and opposed Tweets. What we see here is remarkable. In both samples—the random sample and the panel—there are significant increases in intensity and anger in the aftermath of the SSM decisions, but only among opposed Tweets. Supportive Tweets do not change in these two emotional aspects.
Intensity among supportive and opposed Tweets

Anger among supportive and opposed Tweets

Random Sample

Panel

Date

Figure 6: Estimated intensity and anger among Tweets by estimated supportiveness over time, January 31, 2013-August 11, 2013. Left column shows estimated intensity among Tweets by day, divided into Tweets estimated to be supportive and opposed to same-sex marriage, and those before and after the Supreme Court’s decisions in same-sex marriage cases. Right column shows estimated anger among Tweets by day, divided into Tweets estimated to be supportive and opposed to same-sex marriage, and those before and after the Supreme Court’s decisions in same-sex marriage cases. Top row shows the random sample of Tweets; the bottom row shows the panel of Tweeters.
5 Additional Uses of the Data

In a future version of this paper, we will present two other uses of the data.

Salience  The most common indicator of the salience of policies reviewed by the Supreme Court was produced by Epstein and Segal (2000). The Epstein and Segal measure is a binary indicator of whether the NY Times runs a story about a Supreme Court’s decision dealing with that policy on its front page. Our project permits the measurement of policy salience, as well as the salience of the Court, in a variety of ways. At the level of the policy-day, the simple volume of Tweets is an appropriate indicator, but so would be an measure of daily volume relative to some baseline, e.g, the average level of Tweeting about the policy. At the level of the policy-Court term, average volume over time would serve, as well. The point is that our approach permits a variety of options. Figure 7, which comes from our pilot study, shows the (logged) number of Tweets on three policy specific topics—gay marriage, the Voting Rights Act, and Corporate Taxation—plus a fourth topic which is about the Supreme Court generally. Three important features emerge. First, we see considerable cross-sectional variation—gay marriage was the most Tweeted of these three topics, whereas the tax case was the least-Tweeted. Second, we see over-time variation—there are notable spikes for each topic on important dates, such as oral arguments and decision dates. Third, we see important differences in the trends. Following the Supreme Court’s decision in the gay marriage cases, there was not much of a shift in the amount of Tweeting about gay marriage. However, following its decision in the Voting Rights Act case, we see a marked shift upwards in the amount of Tweeting, likely due in part to both the content of the decision and the Obama Administration’s (continued) reaction to the case.

These results suggest that in addition to considerable flexibility we can offer, our approach can capture the same kind of variation that the well-used and validated Epstein and Segal salience measure identifies. Specifically, the voting rights case (Shelby County v. Holder) and the gay marriage cases (US v. Windsor and Hollingsworth v. Perry) were all covered on the front page of the NY Times, while the corporate taxation case (PPL v. Commissioner) was not. We are satisfied that we can capture salience with these data and our approach.

Teaching Uses  This section will show the Court-o-Meter website.
Figure 7: *Plot of count of Tweets for each of four topics, March 2013 through August 2013.* The figure shows the raw number of Tweets collected using four sets of keywords from Twitter's free streaming API service. Counts are plotted on the log scale; vertical grey lines show important dates for each topic.

6 Conclusion

There will be a conclusion.
References


