A method for computing political preference among Twitter followers

Jennifer Golbeck [a,*], Derek Hansen [b]

[a] Human-Computer Interaction Lab, College of Information Studies, University of Maryland, College Park, MD, United States
[b] Information Technology, Brigham Young University, Provo, UT, United States

A B S T R A C T

There is great interest in understanding media bias and political information seeking preferences. As many media outlets create online personas, we seek to automatically estimate the political preferences of their audience, rather than examining the bias of the media source. In this paper, we present a novel method for computing the political preferences of an organization's Twitter followers. We present an application of this technique to estimate the political preferences of the audiences of U.S. media outlets, government agencies, and interest groups and think tanks. We also discuss how these results may be used and extended.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction and related work

As government agencies, media outlets, and other organizations establish online presences in social media, understanding the characteristics of their audiences is an important task. It has implications for how information is presented in social environments where personalization is often expected. Furthermore, it can provide valuable information to marketers, policy analysts, and social science researchers.

Fortunately, users of these social media services leave a public trail of their activities, and that behavior can be used to infer information about their preferences. In this paper, we present a technique for estimating audience preferences in a given domain on the microblogging service Twitter. We use U.S. politics as our domain and motivating example. We present a method for computing the political preference of an audience by analyzing their "following" behavior and illustrate this approach with media outlets, government, and interest groups and think tanks. This method was first outlined in our previous work (removed for anonymity), and we extend the method and analysis here.

Although our study is about the political preferences of audiences—not media bias, or the biases of other organizations—it is worth briefly discussing the extensive research on analyzing media bias as it is strongly correlated with audience preferences. A subset of this work uses automated methods to infer liberal/conservative bias of news stories and outlets. These automated methods do not depend on subjective measurements of bias,

although the specific techniques used to infer bias can be problematic and are highly contested.

For example, one approach is to compute a media bias score based on citations in news stories. News outlets that cite "think tanks" which are also cited by Representatives with a known liberal bias are assumed to be more liberal (Groseclose and Milyo, 2005). Another approach is to compare keywords and phrases used by Representatives of known political persuasions with those used in news articles; news outlets that use terms like "death tax" and "illegal immigration" are more likely to be conservative (Gentzkow and Shapiro, 2010). A final approach assigns a liberal/conservative score to web documents based on the number of times they are co-cited with other web documents that have a known political bias (Efron, 2004).

In contrast to these approaches, we estimate the political preferences of news outlet audiences, not the news outlet content itself. Our strategy is similar to Groseclose and Milyo (2005) in that we use Representatives' liberal/conservative ratings as a starting point for our scoring. However, we use Twitter Follow relationships rather than article citations. Using Follow relationships avoids a key concern with Groseclose and Milyo (2005)—that results rely too much on the nuances of journalist and Representative citation practices. Our approach does not require a researcher to read stories and find citations (as in Groseclose and Milyo, 2005), or access to large corpuses of news stories and congressional speeches. Instead, it relies instead on freely available and open access data from Twitter.

2. Method and sampling

We compute the political preference of an audience by using sites like Twitter that embed users' information-seeking choices into social media ties. Our examples and applications are in the political domain, but the technique is generalizable when the right

* Corresponding author. Tel.: +1 3014057185.
E-mail addresses: jgolbeck@umd.edu (J. Golbeck),
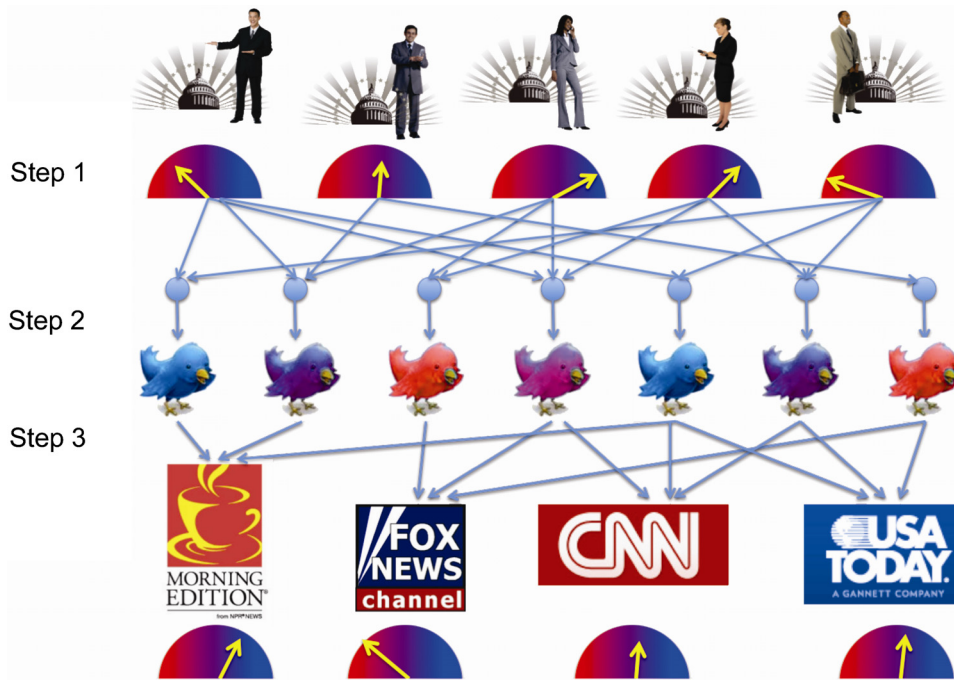dlhansen@byu.edu (D. Hansen).

Fig. 1. An illustration of our process for scoring users and organizations.

background information is available. Fig. 1 illustrates our approach, which comprises the following steps.

Step 1: Apply known scores to a seed group. (In our case, these are Representatives using Twitter.) We then need a source for liberal/conservative ratings of Representatives. Two sources were available to us:

- The first set was obtained from Americans for Democratic Action (ADA), who puts out an annual report that considers the voting record of members of Congress (ADA, 2009). ADA defines a key set of votes that indicate liberal and conservative positions, and uses the Representative's voting record to assign each a score. The most liberal score is a 1.0, and the most conservative is 0.0. This is a widely accepted measure of political position. We apply the 2009 ADA ratings to our main dataset of 111th Congress members.
- In addition, we used DW-nominate scores (Clinton, 2004). The DW-Nominate scores have two dimensions, and the first most closely represents traditional liberal/conservative positions. These scores range from −1 to +1, with lower values representing more liberal positions and higher values representing conservative positions. These values could be used in place of ADA scores anywhere in our analysis, and we test them in the next section where we validate our method.

Step 2: Map the scores of the seed group onto their followers to create P-Scores. We collect the list of followers for each member of Congress on Twitter (i.e., Congress Followers). An inferred political preference score (*P-Score*) for each Congress Follower is computed as the simple average of the liberal/conservative scores for all Representatives he or she follows. Our approach relies on the assumption that people's political preferences will, on average, reflect those of the Representatives they follow. Prior literature on "selective exposure" to political information suggests this assumption is reasonable since people seek after information from those with similar political views (Frey, 1986; Garret, 2009). We verify this assumption through surveys and quantitative measures, discussed below in the next section on "Validating Follower Scores".

Step 3: Map the inferred scores of the seed group followers (Congress Followers) onto the target of the investigation. In this case, that includes the Twitter accounts of media outlets, government agencies, and interest groups and think tanks. A simple approach is to assign the average of the liberal/conservative scores of all Congress Followers who also follow the target media outlet. However, this approach raises a problem: Twitter users may not represent the overall population well. Therefore, we use a sampling method to more closely approximate the distribution of liberal and conservative tendencies as described below.

The main dataset we used was collected during the 111th Congress. Although Democrats outnumbered Republicans in Congress at the time, Republicans significantly outnumbered Democrats on Twitter (127 to 103). Furthermore, Republican Twitter users tended to have disproportionately more followers than Democrats. Even excluding John McCain, who had over 1.7 million followers as a result of the 2008 Presidential election (more than 30 times the next most-followed member of Congress, and nearly twice the total number of followers of all other Representatives), the total number of Republican followers was 581,997, compared to 291,050 for Democrats. (We duly note that some of these users likely follow both Republicans and Democrats.)

Without any adjustment, news outlet audiences would appear more conservative than they would if Twitter users' politics mirrored those of the more general U.S. population.

Since Republicans are overrepresented on Twitter when compared to their representation on Congress, we sampled members of congress to more closely match the roughly equal ratio of Republicans and Democrats in the general population. John McCain was excluded as an outlier, because the followers gained during his 2008 Presidential run makes him a particularly abnormal and overly influential data point.

Representatives were broken into groups by the number of followers they had: *over 10,000; 1000–5000; 500–1000; 100–500; 10–100*. Within each group, we randomly selected equal numbers of Republican and Democratic representatives until we had the maximum number for the least represented group. For example,