Computing Political Preference among Twitter Followers

Jennifer Golbeck
Human-Computer Interaction Lab
College of Information Studies
University of Maryland
College Park, MD 20742
jgolbeck@umd.edu

Derek L. Hansen
CASCI, Human-Computer Interaction Lab
College of Information Studies
University of Maryland
College Park, MD 20742
dlhansen@umd.edu

ABSTRACT
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 of the outlet itself. In this paper, we present a novel method for computing preference among an organization's Twitter followers. We present an application of this technique to estimate political preference of the audiences of U.S. media outlets. We also discuss how these results may be used and extended.

Author Keywords
Twitter, politics, liberal, conservative, news, journalism

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors

INTRODUCTION
As major media outlets establish online presences in social media, understanding the characteristics of those audiences is an important task. It has implications for how information is presented in this environment where personalization is expected. Furthermore, it can provide valuable information to marketers and social media analysts.

As a first step toward understanding audiences, we present a technique for estimating audience preferences in a given domain on the microblogging service Twitter. We use U.S. politics as our motivating example by estimating the political preferences of media outlets' audiences.

BACKGROUND
While we are not studying media bias, but rather the political preferences of audiences, it is worth briefly discussing the extensive research on analyzing media bias.

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. One approach is to compute a media bias score based on citations in the news story – news outlets that cite "think tanks" that are also cited by Congressperson's with a known liberal bias are assumed to be more liberal [8]. Another approach is to compare keywords and phrases used by Congresspeople 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 [7]. 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 [3].

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 [8] in that we use Congresspeople's American for Democratic Action's (ADA) scores as a starting point for our scoring; however, we use Twitter Follow relationships rather than article citations. Using Follow relationships avoids the concern with [8] that results rely too much on the citation practices of journalists and Congresspeople. Our approach does not require coding of data (as in [8]) or access to large corpuses of news stories and congressional speeches; it relies instead on freely available and open access data from Twitter.

METHOD AND SAMPLING
Unlike [8], we are not interested in predicting media bias. Instead, we are interested in predicting the political preference of the *audience* of different media outlets and organizations by using sites like Twitter that embed social ties. Our examples and applications are in the political domain, but the technique is generalizable when the right background information is available. Our approach includes the following steps:

Step 1: Apply known scores to a seed group, in this case Congresspeople using Twitter. The base data of liberal/conservative scores are obtained from Americans for Democratic Action (ADA), who puts out an annual report that considers the voting record of members of Congress [1]. ADA defines a key set of votes that indicate liberal and conservative positions, and uses the Congressperson'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 the 111th Congress members.

Step 2: Map the scores of the seed group onto their followers. 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 ADA scores for all Congresspeople he or she follows. Our approach relies on the assumption that people's political preferences will, on average, reflect those of the Congresspeople 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 [4, 5]. We tested this assumption in the Twitter context by surveying a convenience sample (recruited via Twitter and email by the authors and friends) of 47 subjects who follow politicians on Twitter. Of those, 66% stated that the politicians they follow "mostly share [their] political views", while only 4% follow politicians who "mostly hold political views that oppose [their] own." The remaining 30% reported following a mix of both, with an average of 56% of the politicians they follow sharing their own views. Thus, overall, users tend to follow politicians with similar views; even when there is a mix of political views, those that match the user's views tend to dominate. These preliminary results suggest that our assignment of a P Score to congressional followers is a reasonable approximation of real political preference. An adjusted P Score could be computed based on data from more comprehensive surveys that compare self-reported liberal/conservative scores with inferred P Scores.

Step 3: Map the inferred scores of the seed group followers (i.e., Congress Followers) onto the target of the investigation, in this case the Twitter accounts of media outlets. 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 population well.

Indeed, on Twitter Republican members of Congress significantly outnumber members who are Democrats (127 to 103), though Democrats outnumber Republicans in the 111th Congress. Furthermore, Republican Twitter users tend to have disproportionately more followers than Democrats. Even excluding John McCain, who has 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 double the total number of followers of all other Congresspeople), the sum of Republican followers is 581,997, compared to 291,050 for Democrats. As a result, the pool of Congress Followers is significantly biased toward conservatives. Without any adjustment, the audiences of news outlets will incorrectly appear more conservative. Our solution is to selectively sample Congress Followers so they more closely match the roughly equal ratio of Republicans and Democrats in the general population, minimizing the effect of the initial selection bias. John McCain was excluded as an outlier because his Presidential run makes him a particularly abnormal and overly influential data point.

Congresspeople were broken into groups by the number of followers they had: over 10,000, 5,000, 1,000, 500, 100, 50, and 10. Within each group, we randomly selected equal numbers of Republican and Democratic representatives until we had the maximum number for the least represented group. We chose equal numbers from each party since in recent elections there are roughly equal numbers in Congress. For each selected Congressperson, we randomly selected the number of followers equal to the group in which he was a member (e.g. for a Congressperson from the "over 1,000" group, we selected 1,000 random followers).

To compute the final media audience P Scores, we used a 10-fold validation; we drew 10 samples using this technique, computed the P Score for each media audience using each sample, and averaged the scores across samples.

CONGRESS FOLLOWER SCORES

Figure 1 shows a distribution of scores for all Congress Followers, as well as the distribution of ADA scores for all Congresspeople (not just those on Twitter).

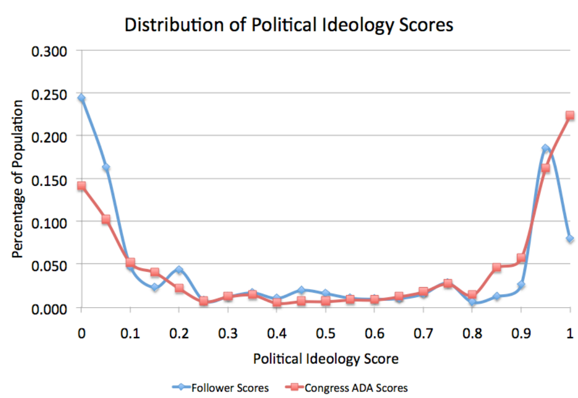

Figure 1. The distribution of P Scores among all Congress Followers and ADA scores for Congresspeople. A score of 0.5 is moderate, 0.0 is very conservative and 1.0 is very liberal.

The distribution of scores among the Congress Followers and the sampled follower population is strongly bimodal, since it is based upon the congressional ADA scores which are bimodal. This is in contrast to other evidence that suggests the distribution of political ideologies among the general public is a more normal distribution, with most of the population as moderate [6]. This discrepancy may arise from two sources. First, followers of Congresspeople may not share the views of those they follow, although our survey and prior "selective exposure" literature makes this unlikely.

Media Outlet	Twitter User ID	Total Followers	Avg. % of Total Followers sampled	Avg. P Score
Fox News	foxnews	266,121	7.33%	0.26
The Drudge Report	Drudge_Report	102,981	20.52%	0.27
Washington Times	washtimes	13,545	33.75%	0.29
Wall Street Journal	WSJ	392,332	5.75%	0.41
US News & World Report	usnews	7,836	28.60%	0.49
The L.A. Times	latimes	72,296	9.91%	0.51
USA Today	USATODAY	63,714	9.90%	0.51
Good Morning America	gma	1,698,875	0.72%	0.51
The News Hour	NewsHour	54,787	12.79%	0.52
CBS News	CBSNews	1,578,599	0.94%	0.53
Newsweek	Newsweek	1,256,536	1.25%	0.54
Washington Post	washingtonpost	154,400	9.51%	0.54
The Today Show	todayshow	629,088	1.89%	0.55
The Early Show	theearlyshow	12,628	11.16%	0.55
Time Magazine	TIME	2,134,411	0.99%	0.57
ABC World News	abcworldnews	12,973	14.93%	0.57
CNN Breaking News	Cnnbrk	3,314,716	1.40%	0.58
NBC Nightly News	nbcnightlynews	27,137	16.85%	0.58
The New York Times	nytimes	2,553,291	1.40%	0.60
Morning Edition (NPR)	MorningEdition	6,146	17.22%	0.66

Table 1. The average audience P Scores for the twenty popular media outlets studied in [8]. Averages are computed over 10 randomly drawn samples using the sampling method described above. Results are sorted from most conservative to most liberal.

Second, those who follow Congresspeople on Twitter may have more polarized political tendencies than the overall US population. This is likely the case, as those who actively decide to follow a Congressperson are likely more politically aware and active, characteristics that are often accompanied by more extreme political tendencies. If this is true, our estimated audience political preference scores for media outlets will rely on the Twitter follow practices of "politically savvy users" rather than the general population. This is not necessarily problematic, although it does impact the meaning of the P Scores. In our political example with the bi-modal distribution, the Avg. P Score for a media outlet is roughly equivalent to the ratio of "political savvy" liberals and conservatives that follow that news outlet.

MEDIA AUDIENCES' POLITICAL PREFERENCE

Our first application of this method was to estimate the audiences' political preferences, through their P Scores, of the same popular media outlets evaluated in [8]. Table 1 shows the audience bias scores along with the percentage of each outlet's followers who were considered when computing the value.

For traditionally conservative outlets [9], such as Fox News [2], we found audiences with correspondingly conservative P Scores: Fox News (0.256), the Drudge Report (0.265), and the Washington Times (0.290). There are no outlets with audiences that have P Scores that are liberal to the same extent that these are conservative, but some liberal preference is visible in the audiences of outlets like the New York Times (0.604) (considered to be liberal leaning [9, 11]) and NPR's Morning Edition (0.659).

The vast majority of these media outlets' audiences – 15 out of 20 – fall between the moderate scores of 0.4 and 0.6. Half are even closer to the midpoint, falling within 0.05 of the perfect moderate 0.5 value. This suggests that most media outlets audiences have roughly equal numbers of liberal and conservative Congress Followers.

Our findings of audience preference are similar to the estimates of political orientation of media outlets found in [3], which used co-citation of hyperlinks to infer political orientation of web documents and their associated news outlets. Although, the scales differ, like [3] we found that the Wall Street Journal and The New York Times equally deviated from the middle in opposite directions (conservative and liberal respectively). This similarity may come from the fact that [3] relies partially on news outlet audiences by using website linking behavior. Our method provides a more direct estimate of audience political preferences by focusing on follower relationships.

Note that these scores do not imply that the outlets themselves present news in a way that reflects their audience's political preferences.

DISCUSSION

There are a number of implications and areas for future work that follow from these results. Many users expect personalized web-based content. Some personalization comes from the structure of the services themselves (e.g. who users choose to follow on Twitter affects which tweets they see). However, there are many opportunities to further personalize and enhance the way information is presented.

Understanding the political preference of an audience can be important for presenting tailored information (including or excluding information according to the user's tastes) and personalizing the user's experience (e.g. through recommender systems). For example, an audience's political preference can be used as input into recommender systems. In collaborative filtering systems, items are recommended by finding people with tastes similar to the user and recommending things those people like. In this context, if we know a user's political preferences, we can find media outlets that have audiences with a similar preference, mimicking the basic idea behind collaborative filtering. Tweets (or information provided on other social media sites) can be highlighted, filtered out, or sorted based on the similarity of their audiences' political preferences to those of the user. Alternatively, our method could be used to help recommend tweets commonly read by people on both sides of the political spectrum reducing homophily [10]. Finally, marketers and analysts can use our method to measure their Twitter reach within different political markets to see if they are reaching their intended audience.

Outside of personalization, this technique may have applications for studying media bias in social media. While we have set out to estimate audience political preference, not media bias, previous work has shown that news consumers have a significant preference for like-minded media outlets [4, 5] and use new social media tools to actively seek out those with similar views. This implies that people may choose a media outlet because its presentation of the news reflects their own political beliefs, and thus the preference of an audience may generally reflect the bias of the outlet. While we do not have evidence to support this connection, predicting media bias based on audience preferences is an area for future research.

Finally, while we have used political preference and media outlets as our example case, we believe this technique is applicable in other domains. We have begun to use this technique to estimage the political preferences of Twitter audiences of government agencies, think tanks, political organizations, and individuals. It could also be used for non-political analysis. For example, a similar analysis could be done using the Green Scores that rate how environmentally responsible companies are. These could be used to create an environmental score for Twitter followers of organizations, politicians, and other Twitter accounts.

CONCLUSION

In this paper, we have presented a technique for estimating the political preferences of Twitter account followers. Using the political domain as our motivating example, we present a method that uses the follower connections in the Twitter network to propagate liberal/conservative scores from members of Congress to Congress Followers and then to audiences of media outlets. Our results show that the estimated political preferences of media outlets' audiences reflect the liberal/conservative leanings of the media outlets as presented in prior literature.

The results have potential applications for motivating new interface personalization techniques, understanding media bias, and for detecting other types of audience preferences in different domains. There is much future work to be done in this space and we hope this initial work serves as motivation to pursue those issues.

ACKNOWLEDGMENTS

Thanks to members of the University of Maryland's Human Computer Interaction Lab (HCIL) for their helpful comments on an earlier draft.

REFERENCES

1. Americans for Democratic Action. Annual Voting Records, 2009. Available at http://www.adaction.org/
2. DellaVigna, S. and Kaplan, E. The Fox News Effect: Media Bias and Voting. *Quarterly Journal of Economics, 122*, 3 (2007), 1187-1234.
3. Efron, M. 2004. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. In *Proceedings of CIKM'04.* (Washington, D.C., November 2004), ACM Press, 390-398.
4. Frey, D. Recent research on selective exposure to information. *Advances in experimental social psychology*, 19, 1986.
5. Garret, K. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communications, 59*, 4 (December 2009), 676-699.
6. Gelman, A. *Red state, blue state, rich state, poor state: why Americans vote the way they do*. Princeton Univ Press, Princeton NJ, 2008.
7. Gentzkow, M. and Shapiro, J. What drives media slant? Evidence from US daily newspapers. *Econometrica, 78*, 1 (2010), 35–71.
8. Groseclose, T. and Milyo, J. A Measure of Media Bias. *The Quarterly Journal of Economics, 120*, 4 (2005), 1191–1237.
9. Mondo Times: The worldwide news media directory. Available at http://www.mondotimes.com
10. Munson, S. A. and Resnick, P. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of CHI'10*. (Atlanta, GA, April 2010). ACM Press, 1457-1466.
11. Puglisi, R. Being the New York Times: The Political Behaviour of a Newspaper. Available at SSRN: http://ssrn.com/abstract=573801