Poster: Information Flow Experiments to study News Personalization

Amit Datta
Carnegie Mellon University
amitdatta@cmu.edu

Anupam Datta
Carnegie Mellon University
danupam@cmu.edu

Suman Jana
Stanford University
suman@cs.stanford.edu

Michael Carl Tschantz
Int'l Computer Science Institute
mct@icsi.berkeley.edu

I. Introduction

With the advancement of tracking technologies and the growth of online data aggregators, data collection on the Internet has become a serious privacy concern. Data aggregating companies track users across many websites to obtain a global view of each user's interests. These insights are then used to serve personalized content to users on websites (e.g., [1]).

There are concerns over online personalization. Through content surveys, Thurman et al. [2] uncover increasing degrees of personalization at top national news websites over a period of three years. Pariser [3] points out that such personalization may trap users in their own customized 'filter bubble'. Such concerns have led to much interest in determining whether such filter bubbles exist. Englehardt et al. study several news platforms, but do not find high levels of personalization [4]. To detect personalization, we look at the methodology due to Tschantz et al. [5], which demonstrates how to conduct experiments in a systematic way to detect instances of information flow in web systems. Datta et al. [6] extend the methodology due to Tschantz et al. and study properties of transparency, choice, and discrimination on Google Ads and Ad Settings. They developed AdFisher, an automated tool to run browser-based experiments and analyze data using machine learning and significance tests.

This poster presents an extension of AdFisher to study news personalization. News articles are a gateway for users to the happenings around the globe. Personalizing news articles affects users' perceptions of the world. We study articles served on Google News[1], which is among the top news websites on the Internet[2] and gets millions of unique visitors in a span of a few days [7]. We run several experiments to study different forms of personalization on Google News. We find evidence for active personalization, where users' explicit actions affect the news feed, and passive personalization in sections where it is expected (like 'Suggested for you'). In particular, we find that editing news personalization settings has immediate and significant effects, and reading and searching for sports-related articles lead to relevant suggested articles. However, we do not find unexpected passive personalization.

II. News Personalization

We study two forms of personalization as identified by Thurman et al. [2]: active and passive. In the former, the user is actively involved in customizing content on the page, for example via news personalization settings. Passive personalization involves inferring interests and demographics from user profiles or behaviors to customize content. This form of personalization may or may not occur in sections where users expect it. For example, on Google News, passive personalization can be observed in the category 'Suggested for you' [8] is expected. Concerns of filter bubbles arise in passive personalization, especially when unexpected.

Google News customizes content differently depending on whether a user is signed in or signed out [9]. We study news personalization in both these settings. When signed in, personalization is based on the user account and follows the user on different computers and browsers. When signed out, articles are personalized based on past news browsing activities linked to a cookie on the browser [9].

In this poster, we examine how online actions and behaviors affect the news articles served in both signed in and signed out settings. We observe the effects of visiting websites related to a certain topic, editing interests on the Google Ad Settings page, signing in to Google accounts, reading certain Google News articles, editing news personalization settings, and performing Google searches. For most experiments, we aim to detect unexpected passive personalization. In the next Section, we discuss the different experiments we perform and our findings from them.

III. Experiments and Results

We extend the AdFisher tool to perform our experiments. AdFisher launches several fresh browser instances, randomly assigns them to the control and experimental groups, and drives them to perform actions and collect measurements. Then, it automatically detects differences in the collected measurements using machine learning and executes a test of significance specialized for the difference it found. The experiments are performed in a blocked fashion, each block comprising of ten browsers. Every experiment has 100 blocks, thus providing measurements from a total of 1000 browser instances. We analyze the collected data using the classifier-based analysis built into AdFisher. 90% of the blocks are randomly selected for training a classifier, while the remaining 10% are used to carry out significance testing. We test for significance at the 5% level and apply the Bonferroni correction to adjust the p-values.

We run one experiment to study active personalization, one for expected passive personalization, and eleven for unexpected passive personalization. Table I lists all the experiments we perform.

[1]news.google.com
[2]www.alexa.com/topsites/category/Top/News

Common actions	Experimental actions	Length	# articles
Sign in	Edit news personalization settings	14 hrs	145,542
-	Read news and perform search	164 hrs	-
-	Visit websites	28 hrs	147,296
Sign in	Visit websites	28 hrs	148,861
-	Edit Ad Settings	22 hrs	145,055
Sign in	Edit Ad Settings	30 hrs	110,704
-	Sign in to fresh account	21 hrs	146,534
-	Sign in to existing account	55 hrs	141,539
-	Sign in to separate existing accounts	15 hrs	146,224
-	Read news from an agency	28 hrs	29,261
Sign in	Read news from an agency	26 hrs	135,308
-	Read news from a category	34 hrs	131,404
Sign in	Read news from a category	31 hrs	129,784

TABLE I. TABLE LISTING THE DIFFERENT EXPERIMENTS WE PERFORM. THE COMMON ACTIONS ARE PERFORMED BY ALL BROWSERS. THE EXPERIMENTAL ACTIONS ARE PERFORMED BY BROWSERS IN THE EXPERIMENTAL GROUP. THE LENGTH OF EACH EXPERIMENT AND THE TOTAL NUMBER OF ARTICLES COLLECTED ARE ALSO SHOWN.

1) Active Personalization: When signed in, Google News provides the option to edit news personalization settings (the analogue of Ad Settings for news), wherein users can adjust the frequency of articles served from specific news sources. When we initially collect news articles for two fresh Google accounts, we did not find any significant difference. Then, in the experimental account, the news settings were adjusted so as to receive fewer articles from the Wall Street Journal, while the control account was left as it was. Articles collected after this modification had significant differences (adjusted p-value: 0.0022). The explanations provided by AdFisher's analysis showed all the top five distinguishing news articles were from the Wall Street Journal served to the control group. This finding demonstrates the presence of active personalization on Google News.

2) Expected Passive Personalization: We performed a week long experiment to observe effects on the 'Suggested for you' section. Using a fresh Google account, we simulated an interest in 'Sports' by reading all Google News articles from the Sports category and then searching for sports related queries and visiting the top three search results. Then, we collected the suggested stories. We drove the browser to perform this sequence of actions over and over. We observed that the first suggested stories appear after about five days from the start of the experiment. All the suggested stories were related to sports, with topics ranging from College Basketball, NBC Sports, and Major League Soccer. Our studies into expected passive personalization have been exploratory, but they suggest that this form of personalization can be observed in the suggested stories.

3) Unexpected Passive Personalization: We focus the bulk of our experiments on detecting unexpected passive personalization. We carry out seven experiments to test the effects of actions outside of Google News on the news articles. We study visiting sports related websites, adding several sports related interests on Ad Settings, and signing in to a fresh as well as existing Google account, but do not find significant effects. We also use two existing Google accounts, whose suggested stories were visibly different, indicating expected passive personalization. But, there was no evidence of unexpected passive personalization. We perform four experiments to study whether reading specific news articles from the Google News page leads to personalization. To simulate reading an article, we had a browser instance click on the article and spend 20 seconds on the destination page. We studied two kinds of reading behavior: (1) reading articles from a specific set of news agencies, and (2) reading articles from a specific category. For the former, we tested with 'USA Today' while signed out, and with 'The Economist', 'Reuters', and 'The Wall Street Journal' while signed in. For the latter, we experimented with the 'Sports' category, while both signed in and signed out. We could not find unexpected passive personalization for any of the above reading behaviors.

IV. DISCUSSION AND FUTURE WORK

We run several experiments to study news personalization on Google News. We detect active and expected passive personalization, but do not find evidence of the more concerning unexpected passive personalization. We observe expected passive personalization in the suggested stories after about five days. It is clear that relevant interests are inferred, but there is no transparency into these interests before they are displayed along with targeted articles. As future work, we would like to investigate how information flows into the suggested stories, thereby improving transparency into the system. Our current findings indicate that longer experiments are necessary to study suggested stories.

Our findings indicate that there may not be unexpected passive personalization on Google News. However, it is also possible that our experiments were not long enough or that our analysis was not powerful enough to detect to find this form of personalization. News stories rapidly change with time, and our experimental design and analysis may not be well-equipped to detect differences varying over time. In future work, we would like to improve our techniques to detect information flow in time-varying measurements and perform longer running experiments.

REFERENCES

[1] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *IEEE Symposium on Security and Privacy*, 2012, pp. 413–427.

[2] N. Thurman and S. Schifferes, "The future of personalization at news websites: Lessons from a longitudinal study," *Journalism Studies*, 2012.

[3] E. Pariser, *The filter bubble: How the new personalized Web is changing what we read and how we think*. Penguin, 2011.

[4] S. Englehardt, C. Eubank, P. Zimmerman, D. Reisman, and A. Narayanan, "Web privacy measurement: Scientific principles, engineering platform, and new results," Manuscript at randomwalker.info/publications/WebPrivacyMeasurement.pdf, 2014, accessed May 6, 2015.

[5] M. C. Tschantz, A. Datta, A. Datta, and J. M. Wing, "A methodology for information flow experiments," in *Proceedings of the 28th IEEE Computer Security Foundations Symposium (CSF)*, 2015.

[6] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," in *Proceedings on Privacy Enhancing Technologies (PoPETs)*. De Gruyter Open, 2015.

[7] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: Scalable online collaborative filtering," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 271–280. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242610

[8] Google, "Suggested stories - news help," http://support.google.com/news/answer/6029465, accessed April. 3, 2015.

[9] ——, "News history personalization - news help," http://support.google.com/news/answer/3010317, accessed April. 3, 2015.