DATACTIVE

Does Facebook's NewsFeed Algorithm Reinforce Filter Bubbles?

An Experimental Methodology Using *fbtrex*

ALEX Working Paper Series
No 1/2019

Davide Beraldo
University of Amsterdam

Address
Stefania Milan, DATACTIVE/ALEX PI
s.milan@uva.nl
University of Amsterdam
Department of Media Studies (2.23)
Turfdraagsterpad 9
1012 XT Amsterdam
The Netherlands

Corresponding author
Davide Beraldo
d.beraldo@uva.nl

Algorithms Exposed: Investigating Automated Personalization and Filtering for Research and Activism is funded with a Proof of Concept grant (grant agreement no. 825974).

Data activism: The politics of big data according to civil society is funded by a Starting Grant of the European Research Council (grant agreement no. 639379).

Both projects are based at the Department of Media Studies of the University of Amsterdam awarded to Dr. Stefania Milan as Principal Investigator.

The ALEX Working Paper Series presents results of the ALEX research project and/or showcases research from invited authors. An editorial committee consisting of the DATACTIVE PI and Postdoctoral fellows reviews the quality of the Working Papers. The Series aims to disseminate research results in an accessible manner to a wider audience. All Working Papers are available for download in PDF format at https://data-activism.net.

Readers are encouraged to provide the authors with feedback and/or questions.

Does Facebook's NewsFeed Algorithm Reinforce Filter Bubbles?
An Experimental Methodology Using *fbtrex*

v 0.1

March 13, 2019

Abstract

This working paper presents a preliminary research design developed to empirically assess the contribution of Facebook's NewsFeed algorithm to the phenomenon of so-called filter bubbles. The browser extension *fbtrex* will be used to collect snapshots of timelines related to *ad hoc* users. Those users will be following the same pages, but will signal a specific orientation towards a political issue with a selective liking activity. By comparing the posts effectively appeared on the users' timelines with the overall set of posts produced by the pages, the analysis will test a set of converging hypothesis related to whether and to what extent the coherence between a user's and a post's orientation towards the issue affect the probability of a post to be served to the user, under a variety of structural scenarios. Whereas the research cannot account for all the socio-technical factors behind the phenomenon of filter bubbles, nor for the overall logic of the algorithm, this experimental setting permits to isolate the effect of the algorithm from other controlled factors, such as users' individual choices and different friendship network configurations.

Introduction

Fueled by continuous streams of data primarily extracted by proprietary platforms, algorithms increasingly mediate social processes in a variety of ways (Gillespie 2014). They categorize, filter, recommend, predict our lives, affecting virtually any political, economic and social phenomena. Algorithms exert a new form of social power (Beer 2017) and often result in (re)producing systematic discrimination (O'Neil 2016). Nonetheless, they are fundamentally unaccountable in their effects, since they are generally obscure, shielded behind industrial secret and sometimes even their technical undecipherability (Pasquale 2015). This has given rise to a particular form of data activism (Milan 2016), articulated around the several individuals, organizations and coalitions advocating and practicing forms of algorithmic accountability (Diakopoulos 2015; Sandvig et al. 2014).

One of the most discussed issues related to the effects of algorithms on society is that of so-called filter bubbles (Pariser 2011) and the alleged intensification of political polarization that they might cause. Whereas in recent years the issue has generated a lot of buzz, there is a lack of empirical knowledge on whether and to what extent proprietary algorithms contribute to diminishing the variety of a netizen's information diet. Some authors warn against the pernicious effects of personalized algorithms on the exercise of democratic citizenship (Tewksbury, Hals, and Bibart

2008); however, a review of the topic argues that there is need for more (independent) empirical evidence in order to sustain this claim (Borgesius et al. 2016).

The present study aims at filling this gap, by testing hypothesis and providing measures on the effects of Facebook News Feed algorithm on the diversity of a user's timeline over a certain issue. This will we be done by trying to isolate the outcomes of the algorithm from other concomitant variables with an experimental setting. This working paper devises an initial set of converging hypotheses, to be possibly expanded and refined at a later stage.

Research Goal

Facebook is nowadays an important source of information: according to the Pew Research Center, 43% of the American adult population use the platform to consume news.[1] Previous studies have shown how at least a portion of users is unaware of the existence of any filtering activity selecting the posts they happen to see on their timelines (Eslami et al. 2015; Rader and Gray 2015).

Existing research on Facebook's algorithm is largely conducted by Facebook-embedded researchers (Bakshy, Messing, and Adamic 2015; Kramer, Guillory, and Hancock 2014), the only having access to the granularity of data required to control for the many variables involved. A controversial publication appeared on Science, based on a detailed study conducted by Facebook researchers on a huge sample of users (Bakshy et al. 2015), argues that 'individual choices more than algorithms limit exposure to attitude-challenging content in the context of Facebook' (p. 1130). The outcome of the study, and this bold claim in particular, have however been at the center of fierce criticism (Lumb 2015), with some arguing that results have been (mis)interpreted in a way convenient to Facebook itself (Sandvig 2015; Tufekci 2015). Conversely, an independent study[2] indicates that News Feed algorithm amplifies user's signaled political orientation, by over-representing Facebook pages they have liked as compared to those they have simply followed (Hargreaves et al. 2018). The present research aims at further developing the last research direction, by more specifically focusing on the role of the algorithm in generating a 'filter bubble' around a specific political issue.

Whereas the social process generating filter bubbles emerge from the (complex) interplay between algorithmic 'pre-selection' and users' own 'self-selection' (Borgesius et al. 2016), the present research design aims to isolate News Feed algorithm's effects from those of user's behavior (liking, commenting, sharing and friendship-making activity) and different structural configurations (e.g. liking patterns / friendship networks that proxy different scenarios of political polarization). Consequently, the goal is not to understand the generative mechanisms behind the emergence of filter bubbles, nor the importance of filter bubbles in real-world scenarios; this would indeed require a naturalistic setting, observing spontaneous user activity on a large scale. Despite its attempt to isolate

[1] Source: http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.

[2] To which the present research designed is inspired.

the algorithm itself, the goal of this project is also not much to reverse engineering the News Feed algorithm properly, as this would require a more granular understanding of the many intervening variables at play. Both goals could most likely be achieved by Facebook-embedded research only which, however, can be quite obviously expected to accommodate Facebook's marketing purposes more than the cause of algorithmic accountability.

The specific goal of this project is to *measure the empirical outcomes of News Feed's content selection on the diversity of political news presented to users*. Whereas this could affect actual news consumption patterns and other aspects of users' behavior in different ways, the attempt to assess algorithmic effects in isolation from other variables is a fundamental step in order to articulate a public critique in support of algorithmic accountability. If specific effects are identified, these cannot be ignored when discussing the issue of filter bubble proper -and can further ground claims for accountability.

Data Collection

This study will make use of samples of timelines generated for *ad hoc* users, collected through the browser extension *fbtrex* (facebook.tracking.exposed). It thus represents a form of sock puppet audit study (Sandvig et al. 2014). Research based on Application Programming Interfaces (APIs) provides limited resources to investigate the inner functioning of algorithms, considering the aggregated and curated nature of the data they produce. Moreover, APIs methods are susceptible to the changing policies of platform corporations, resulting in discontinuities and, especially in the aftermath of the Cambridge Analytica scandal, growing restrictions.

Fbtrex, instead, relies on data collected from the point of view of the user. It creates a (privacy-aware) copy of a user's timelines, and aggregates the data enabling to consider the user's own point of view (i.e. which posts have effectively appeared as an outcome of algorithmic selection). Whereas the 'real-world' adoption of the tool, currently installed by more than 3,000 users, has very promising applications for both awareness-raising and research, for this specific test we follow previous studies (Hargreaves et al. 2018) that have designed a controlled research setting making use of *ad hoc*, partially automated Facebook users.[3] The main problematic issue of the real-world users dataset at this stage is that of self-selection sampling, which would prevent to make generalizations considering its most likely biased composition.[4] The main shortcoming of controlled users profiles relate to difficulty of managing an excessive number of profiles; however, this choice allows to observe the responses of NewsFeed outcomes on predefined user's behavior. The possibility of accurately

[3] Creating *ad hoc* accounts introduces an issue concerning the possible violation of Facebook's Terms of Service. This aspect will be duly covered in a another paper. It suffices to remind here that this methodology is the only viable, scientifically sound way to independently assess claims related to Facebook's algorithms effects.

[4] This limitation is expected to diminish when reaching a critical mass of adoption.

controlling intervening variables different than algorithmic selection largely compensate the lower number of observations that the burden of managing a number of accounts entails.

Considering the interest of this project towards so-called filter bubbles, a relevant, highly divisive political issue will be selected (e.g. migrants; European Union; climate change), so as to test the outcome of the interaction between the (unknown) News Feed behavior and the (controlled) user behavior. More in particular, we will look at this interaction in terms of the relative probability of posts appearing on a user timeline to be coherent, rather than cross-cutting, with respect to the user's orientation towards the issue.

The data collection procedure can be summarized as follows. The collection of data will produce a dataset of Facebook posts observed by a number of freshly created, *ad hoc* Facebook users ('impressions') over 30 days. Users will initially follow (without liking) the same sources (approx. 30 Facebook pages related to news and 10 politicians known for their orientation towards the issue). Users will be then preliminary 'polarized' by selectively liking, according to a controlled procedure, posts and pages coherent with a specific orientation of the issue. We expect that, by tracking users behavior, the algorithm will implicitly categorize users as 'pro' or 'against' the issue, based on the preferences that they revealed with their liking patterns. This way, we end up with three groups: 10 users 'in favor' of the issue ('pro' group); 10 users 'against' the issue ('anti' group); 10 users will not like any post related to the issue ('ctrl' group).

Gradually, again according to a planned procedure, users will start making friendship connections among each other and liking pages (among the initial ones and new ones). These choices are designed so as to introduce different scenarios in terms of individual behavior (liking pages and content of coherent and cross-cutting orientation) and of network structure (friendship connections intra- and cross-group; different network structures), proxying varying configurations of self-selected political polarization.

Another data set will be collected including all the posts produced by each of the followed pages ('posts'). This will allow us to measure the probability of a certain post to appear on each user's timeline (p_{post}). The content of the 'posts' dataset will be selected for those related to the chosen issue; the selected posts will be manually classified along their orientation towards the issue. In this way we can also classify every impression as either 'coherent' or 'cross-cutting', according to the relative orientation between user and impression.[5] By comparing the effects of various factors and scenarios on the share of impressions coherent with a user's orientation, we aim at isolating the effects of NewsFeed's selection from those of individual behavior and network structure.

Research Questions and Hypothesis

[5] Obviously, this will include a residual category for posts of uncertain orientation. Possibly, each post 'orientation' will consider different degrees of strength (very in favor, in favor, moderately in favor, etc.), so that the variable indicating coherence between user's orientation and post will be continuous instead of dichotomous.

The general question guiding this research has to do with how Facebook's algorithm treats content a user is supposed to be following, based on whether that content is coherent or not with a user's previously signaled orientation towards an issue. Although not considering the whole complex interacting social phenomena related to filter bubbles, answering this question is a crucial premise for further investigation. Hence the main research question:

RQ: *Does (and under which conditions) the NewsFeed algorithm privilege content coherent with a user's orientation towards a political issue?*
The main research question is operationalized in the following three sub-questions, each contributing from different angles to give an empirical answer to the same question.

RQ1: *Is the probability of cross-cutting posts to appear on a user's timeline lower than that of coherent posts?*
Considering the polarized groups only ('pro' and 'anti'), if NewsFeed algorithm contributes to create filter bubbles, then the probability of a coherent post to appear on a timeline (p_{coherent}) will be higher than that of a cross-cutting post ($p_{\text{cross-cutting}}$). From this follows our first hypothesis:

HP1: The probability of encountering cross-cutting posts is lower than the probability of encountering coherent posts.

$$p_{\text{coherent}} > p_{\text{cross-cutting}}$$

RQ2: *Does exhibiting a polarized orientation increase the chances of being served by the NewsFeed algorithm more posts coherent with a user's orientation?*
The control group's collected impressions will be used to account for random effects of the algorithm. We can calculate the odds between cross-cutting and coherent impressions ($o_{\text{cross-cutting}} = \mathrm{p}_{\text{cross-cutting}} / \mathrm{p}_{\text{coherent}}$) per each group of users ($o_{\text{cross-cutting}}^{pro}$; $o_{cross\text{-}cutting}^{anti}$; $o_{cross\text{-}cutting}^{ctrl}$). Subsequently, we can calculate the odds ratio ($o_{\text{cross-cutting}} / o_{coherent}$) of polarized (both pro- and anti-issue) vs. non-polarized users. We hypothesize that the odds ratio of both polarized groups to be lower than 1, signaling an association between a user's polarization and the likelihood to be exposed to more issue-coherent content. Hence the second hypothesis:

HP2: The probability of encountering cross-cutting posts diminishes if a user signals a specific orientation.

$$o_{\text{cross-cutting}}^{pro} \,/\, o_{cross\text{-}cutting}^{ctrl} < 1, \quad o_{\text{cross-cutting}}^{anti} \,/\, o_{cross\text{-}cutting}^{ctrl} < 1$$

RQ3: *What effect does the fact of a post being coherent with a user's orientation have on the probability of that post to appear on a user's timeline, controlling for intervening variables?*

We can model the probability of a post to appear on a user's timeline (p_{post}) as a function of whether the post is coherent with the user's orientation ($X_{\text{dummy\_coher}}$). Using a logistic regression model, we can control for the effects of intervening variables such as different levels of cross-cutting friendship ties ($X_{\text{nfriends\_cross}}$), different user liking behavior ($X_{\text{nlikes\_cross}}$), test group to which a user belongs ($X_{\text{dummy\_pro}}$) and other post-level confondents such as the engagement of a post ($X_{\text{post\_engage}}$), the type of post ($X_{\text{post\_type}}$) and the relative time of posting (X_{recency}).[6] Looking at the regression coefficient related to $X_{\text{dummy\_coher}}$, we can estimate the effect of a post's coherence with a user's orientation, *ceteris paribus*, on the likelihood of a post to be selected by the algorithm. In accordance with the filter bubble hypothesis, we would expect this coefficient to be positive. We can thus formulate a third hypothesis:

HP3: The probability of a post to be served to a user's timeline increases when the post is coherent with a user orientation, controlling for intervening variables.

$$logit\ (p_{\text{post}}) = \alpha_0 + \beta_0 X_{\text{dummy\_coher}} + \beta_1 X_{\text{nfriends\_cross}} + \beta_2 X_{\text{nlikes\_cross}} + \beta_3 X_{\text{dummy\_pro}} +$$
$$+ \beta_4 X_{\text{post\_engage}} + \beta_4 X_{\text{post\_type}} + \beta_5 X_{\text{recency}} : \beta_0 > 0$$

Implications, Limitations and Extensions

The question being posed in the present working paper has pretty straightforward implications: if the News Feed algorithm favors content coherent with a user's signaled orientation, as suggested by initial findings (Hargreaves et al. 2018), Facebook can be legitimately supposed to contribute to exacerbate so-called filter bubbles. Moreover, estimating the distinct effects of concomitant factors, such as user-post relative orientation or the number of cross-cutting friends, on the likelihood of a post to be selected contribute to introducing baseline measurements for further auditing research.

The present research design has a number of limitations. In particular, the main shortcoming is in term of external validity, considering that what we are observing is an artificial environment, whereas the actual phenomenon of filter bubbles depends both on algorithmic curation and user spontaneous behavior -as well as their complex entanglement. However, estimating algorithmic-specific effects is a necessary step for an informed debate on the overall social issue to develop.

Another limitation that is necessary to mention is that, despite the stated attempt to isolate algorithmic behavior, the number and inter-relatedness of confounding factors is too high to produce exact claims on the algorithm's logic. The News Feed algorithm most likely has a stochastic component, evolves

[6] Compared to the known factors affecting News Feed choices (see https://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed/?guccounter=1), this set of control variables does not include the interest shown by users towards the page, which will be controlled with a uniform user-page interaction activity.

in time, and depends on a complex set of interacting variables. Again, the ambition is not to understand how the News Feed algorithm works, but rather to observe its systematic effects given certain conditions, in order to provide with empirical grounds the hypothesis of algorithmically-generated filter bubbles.

This working paper represents a work-in-progress. We can envision at least two interesting possibilities of expansion: the first related to network effects, the second related to the relative importance of the algorithm and user's behavior.

We could focus on different network configurations more in particular, by replicating the same analysis on different network scenarios. We can measure political polarization in terms of network structure by looking at the proportion of cross-cutting ties or at other network parameters, such as network density and clustering. This might allow us to test the effects of different fine-grained network configurations, based on friendship patterns among users, on a measure of political polarization based on the diversity of impressions on each user's timeline.

A further direction that the research could take relates to the relative importance of algorithmic versus users' choices in determining the level of cross-cutting content shown. If it is possible to appropriately operationalize the abstract variable 'user behavior' (e.g. looking at liking patterns and friendship making), it would possible to estimate its contribution to the probability of a post to appear (e.g. with a logistic regression model similar to the one described HP3), and compare the resulting coefficients with that of the dummy 'the post is coherent with the user'. Whereas Facebook's own research claims that users' choices are more determinant than News Feed selection (Bakshy et al. 2015), the validity of this claim has been criticized on grounds on sampling and data interpretation issues (Sandvig 2015; Tufekci 2015). An experimental setting based on the data produced by *fbtrex*, however, could possibly give us an opportunity to test this (suspicious) claim even without the possibility to run large-scale natural experiments that Facebook embedded researchers have, representing an achievement also in terms of coping with today's growing 'data divide'.

Sample Research Procedure

The following guidelines are still general and tentative, as they will depend on more fine-grained operative choices, expectations on further hypothesis to test and the availability of partners interested in co-managing a number of *ad hoc* accounts.

1. Set up of approx. 30 Facebook users, using different IP addresses on different virgin browsers.
2. Each of these users will follow (without liking) 40 newspapers and politicians of different political views.
3. 10 of those users ('pro' group) are gonna like exclusively pro-issue; 10 ('anti' group) are gonna like exclusively anti-issue; 10 ('ctrl' group) are gonna like exclusively unrelated posts (approx. 5 articles per day).

4. On regular intervals of time (approx. between 5 and 10 days), friendship connections are going to be added and removed, following a dynamic adjacency matrix to be compiled in order to simulate different scenarios (i.e. 'highly pre-polarized' scenario: more intra-group connections; 'lowly pre-polarized' scenario: more cross-group connections).
5. Similarly, on regular intervals of times, users are gonna start liking specific pages, simulating again different scenarios of self-selected polarization.

References

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348(6239):1130–32.

Beer, David. 2017. "The Social Power of Algorithms." *Information, Communication & Society* 20(1):1–13.

Borgesius, Frederik J.Zuiderveen, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. de Vreese, and Natali Helberger. 2016. "Should We Worry about Filter Bubbles?" *Internet Policy Review*.

Diakopoulos, Nicholas. 2015. "Algorithmic Accountability." *Digital Journalism* 3(3):398–415.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "'I Always Assumed That I Wasn'T Really That Close to [Her]': Reasoning About Invisible Algorithms in News Feeds." Pp. 153–162 in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, *CHI '15*. New York, NY, USA: ACM.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." Pp. 167–94 in *Media technologies: Essays on communication, materiality, and society*, edited by T. Gillespie, P. Boczkowski, and K. Foot. Cambridge, MA: MIT Press.

Hargreaves, Eduardo, Claudio Agosti, Daniel Menasché, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. 2018. "Biases in the Facebook News Feed: A Case Study on the Italian Elections." *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 806–12.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." 111(24):8788–90.

Lumb, David, David Lumb, and David Lumb. 2015. "Why Scientists Are Upset About The Facebook Filter Bubble Study." *Fast Company*. Retrieved March 11, 2019 (https://www.fastcompany.com/3046111/why-scientists-are-upset-over-the-facebook-filter-bubble-study).

Milan, Stefania. 2016. *Data Activism as the New Frontier of Media Activism*. *SSRN Scholarly Paper*. ID 2882030. Rochester, NY: Social Science Research Network.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1 edition. New York: Crown.

Pariser, E. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Rader, Emilee and Rebecca Gray. 2015. “Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed.” Pp. 173–182 in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, *CHI ’15*. New York, NY, USA: ACM.

Sandvig, Christian. 2015. “The Facebook ‘It’s Not Our Fault’ Study.” *Social Media Collective*. Retrieved February 23, 2019 (https://socialmediacollective.org/2015/05/07/the-facebook-its-not-our-fault-study/).

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. “Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms.”

Tewksbury, David, Michelle L. Hals, and Allyson Bibart. 2008. “The Efficacy of News Browsing: The Relationship of News Consumption Style to Social and Political Efficacy.” *Journalism & Mass Communication Quarterly* 85(2):257–72.

Tufekci, Zeynep. 2015. “Facebook Said Its Algorithms Do Help Form Echo Chambers, and the Tech Press Missed It.” *New Perspectives Quarterly* 32(3):9–12

ISSN coming soon