Social Media Bubbles Reinforce Negative Behaviour

Sebastian E. Lamerichs

Curtin University

Abstract

This paper explores the use of “filter bubbles” (or just “bubbles”) on the Internet, referring to the practice of using metadata for each individual user of an online service such as a social media platform or search engine, and constructing personalised results for that user that align with their existing interests, intended to increase engagement on those platforms. While the concept of filter bubbles culturally is not a new phenomenon, this paper draws on the popularised idea of online filter bubbles as defined and used by Pariser (2011) and posits that these bubbles directly result in users of services utilising them experiencing a feedback loop of like-minded people and pages, reinforcing their own beliefs and behaviours. Specifically, we look at filter bubbles created by Facebook friend groupings and Reddit communities (“subreddits”), and posit that negative behaviours are heavily susceptible to the effects of these bubbles, to a greater extent than positive or neutral ones.

Social Media Bubbles Reinforce Negative Behaviour

With the advent of the Internet in its use as a social platform for personal use, rather than its original use by academics or the military, many websites have begun creating personalised content for each individual user of that platform. This is not a particularly new phenomenon, as dynamic content predicated on each user's metadata was one of the founding principles of "Web 2.0" content (Best, 2006); a service such as Twitter would not be particularly useful or engaging if you were shown the same content as everyone else on the website regardless of who you had "followed". The term "filter bubble" (or just "bubble"), however, refers to a more secretive filtering process that isn't based on a user's explicit choices in the content that they'd prefer to see, but rather the algorithmic process of a piece of code analysing a user's metadata and making guesses on what type of content that user is going to be more engaged with in the future (Pariser, 2011).

Nearly everyone who uses the Internet regularly will experience this bubbling in some form or another. The most immediately apparent example is in search engines. Someone in South Australia using Google to search "football results" will usually see pages about the Australian Football League, while someone in the United Kingdom is likely to see results about the English Premier League, and someone in the United States may see the National Football League. While this specific example may seem fairly harmless, the result of this bubbling can be far more insidious: in an interview with *Salon.com* (Parramore, 2010), Eli Pariser recounted an example where two of his friends Googled the acronym "BP", referring to the oil and gas company British Petroleum. One of them received news articles about 2010 Deepwater Horizon oil spill, while the other received "a set of links that was about investment opportunities in BP". Because Google's algorithm had determined that this second searcher was more likely to be interested in investing in British Petroleum, it had the end result of actively hiding a large news story from a certain subset of users at the behest of a faceless machine's assumptions.

This content filtering has a profound impact on social media, where it results in users creating social circles around common interests, particularly political beliefs. A study by Facebook data analysts revealed that for every four Facebook friends an average user has that share the same side of the left-right political spectrum, that user is friends with only one person who is on the opposite side of the political spectrum (Bakshy, Messing, & Adamic, 2015). While it was not made clear whether or not this disparity in user connections was a result of the algorithm preferentially recommending other politically like-minded users as friends or rather the result of user choice in who they associate with, the study also determined that the use of an algorithmic ranking process

on the Facebook News feed further reduced the amount of content an average user would experience opposed to their existing political views by anywhere from 5% to 8% depending on that user's political stance. When this effect is compounded with the existing 4-to-1 ratio of like-minded versus opposing users available in that feed to begin with, and the fact that users are up to 17% less likely to click on links that conflict with their pre-held beliefs even when they do see it (Bakshy et al., 2015), the end result is a system in which reaffirming content has an advantage in every step of the process from content creation to a user viewing it, entrenching the user in their bubble.

This is by no means a phenomenon isolated to Facebook, either. Popular online content aggregator site *reddit.com* has also been criticised for its prevalence of "echo chambers", communities where the overwhelming discourse is one-directional and dissenting opinions are rarely seen or even outright banned (LaViolette, 2017). Reddit's default content display algorithm is based on a user voting process where content that accumulates a lot of "upvotes" very quickly is displayed at the top of the page (Munroe, 2009), and while this helps combat some of the problems with algorithmically biased content as it shows what's popular regardless of whether or not Reddit thinks that it aligns with your political beliefs, it doesn't combat any of the problems caused by the creation of insular communities where everyone voting has similar beliefs. Reddit allows for any user to create their own community on the site – called a "subreddit" – and the creator(s) of that subreddit are free to moderate it at their own discretion, removing any content that they feel does not fit the intended theme of the community. For example, the subreddit "/r/The_Donald", a community for the U.S. President Donald Trump with over 600,000 registered users, states in its rules that the subreddit "is for Trump supporters only" and explicitly bans dissent ("The_Donald", 2018).

While the Reddit vote system may seem like it prevents communities like this from becoming a filter bubble, it ends up being almost completely ineffectual in extreme cases like /r/The_Donald. As part of a feature that is almost completely undocumented, users who are banned from a subreddit (preventing them from being able to post submissions or comments) are also not allowed to vote on posts in that subreddit (Harvey, 2011). This means that moderators of a subreddit can not only remove content that they personally dislike, but they can also artificially inflate the seeming popularity of content that they do like by banning anyone who would otherwise have voted it down.

This has the end result of a post's "ranking" (and thus visibility) determined almost exclusively by its engagement, rather than its quality. A post that is universally beloved but only seen by 1,000 people will have 1,000 "points", whereas a post that is seen by 200,000 people with only 1% of those people liking the post enough to upvote it can still be ranked higher if the other 99% of people are banned from the community and cannot vote on it. Just like how Facebook preferentially shows people posts that they already agree with, creating an illusion of consensus where it may not exist, Reddit's handling of voting algorithms and subreddit autonomy results in communities where content that fits the moderators' existing beliefs can seem well-received and overwhelmingly supported even if the opposite is true.

This pattern is evident across almost all social media platforms. Facebook has a "Like" button that will increase a post's chance of being selected to appear on a user's News feed due to the increased engagement, but no equivalent button to "Dislike" a post that will achieve the opposite. Twitter allows you to "Like" someone's tweet, or even "Retweet" it out to all of your followers to extend its audience, but there's no way to give negative feedback. Google+ has "+1s", but no "-1s". Reddit is one of the few large social networks that gives users any ability to give a post or comment negative feedback, but even there this feedback can be stifled. When the success and visibility of content is determined almost entirely by the number of people who like it with no regard for the number of people who dislike it, content creators (which in the context of social media covers almost anyone using the platforms) are incentivised to create content that reaches the maximum number of eyes, with minimal regard for what that content actually entails. As Bakshy et al. (2015) showed, people who already agree content are both more likely to see it and more likely to click on it, so on any even remotely divisive topic, heavily-viewed content will seem popular regardless of its stance.

With such a strong impetus to publish content that is as widely-seen as possible with minimal regard for the quality or accuracy of the content, the end result is publishing negativity. People are more likely to read negative news stories than positive ones, even if they explicitly voice a preference for positive ones (Trussler & Soroka, 2014). Even on a physical level, humans not only react faster to negative words like "cancer" and "war" than positive ones like "smile" and "fun", but are also more likely to react to them at all (Dijksterhuis & Aarts, 2003). This seemingly-innate predisposition towards negative news has a profound impact on what people read online; when Russian news website "City Reporter" performed an experiment where they only published positive news for a day, their readership dropped to a third of its usual figures (Epstein, 2014).

Negative content having such an advantage over positive content in terms of viewership (and thus success) encourages an all-out assault of negativity. Marketing research into online reactions show that social media is far more prone to backlash and negativity than traditional communication mediums, and that this negativity propagates and grows much faster than positive sentiments (Pfeffer, Zorbach, & Carley, 2014), in a phenomenon known as "online firestorms".

Compounding this effect further is the strong biases humans display when evaluating information. In addition to being significantly more likely to believe or trust evidence that conforms with a pre-existing opinion than evidence that doesn't (Stanovich, West, & Toplak, 2013), humans are also biased towards believing negative information over positive information (Morewedge, 2009), and are more likely to recall information that was associated with a negative stimulus when they were first presented with it (Costantini & Hoving, 1973).

All of these factors together combine into a deeply negative vicious cycle of social media reinforcement. From the very start of a piece of content's life cycle to the very end, positive content and positive behaviour – particularly when it challenges existing beliefs – is at a disadvantage. Negative content is more popular and more

successful, and as such more of it is produced by those who wish for their content to be popular and successful (Pffefer et al., 2014). This content is then classified with various metadata – notably political stance – and then algorithmically distributed on social media to those most like to already agree with it (Pariser, 2011). These users are then more likely to click on this negative content because it is both in line with their pre-held beliefs (Bakshy et al., 2015) and because it is negative (Trussler & Soroka, 2014). On further examination of the content, they will pay more attention to the content because it's negative (Dijksterhuis & Aarts, 2003). When actually processing the content, bubbled users are more likely to believe it because it's negative (Morewedge, 2009), and because they already held the same belief (Stanovich et al., 2013). These users are also more likely to remember it long-term because it's negative (Costantini & Hoving, 1973), and if they share the content, they're typically sharing it to others in the same filter bubble, such as on Facebook where 80% of a user's friends share broadly similar political views (Basky et al., 2015).

Many of these causal links for the advantage of negativity and proliferation of negative content existed long before the Internet, but the existence of filter bubbles that shut out any conflicting views allowed for an unprecedented flow of constant reinforcement for negativity, causing the creation of communities, groups, and friend circles that form not over a common interest, but a common disinterest: a hatred for someone or something. As filter bubbles are a relatively new phenomenon more research needs to be done into the long-term impacts of such heavy exposure to a bubbled environment of negative reinforcement, but it is clear that in the short term these bubbles have contributed to making social media a deeply negative place.

References

Bakshy, E., Messing, S., & Adamic, L. (2015), Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. doi: 10.1126/science.aaa1160

Best, D. (2006), *Web 2.0 Next Big Thing or Next Big Internet Bubble?* Lecture Web Information Systems: Techni sche Universiteit Eindhoven.

Costantini, A.F., & Hoving, K.L. (1973), The effectiveness of reward and punishment contingencies on response inhibition. *Journal of Experimental Child Psychology*, 16(3), 484–494. doi:10.1016/0022-0965(73)90009-X.

Dijksterhuis, A., & Aarts, H. (2003), On Wildebeests and Humans: The Preferential Detection of Negative Stimuli. *Psychological Science*, 14(1), 14–18. doi:10.1111/1467-9280.t01-1-01412

"The_Donald" (2018). Retrieved April 22, 2018 from https://www.reddit.com/r/The_Donald

Epstein, A. (2014), *Here's what happened when a news site only reported good news for a day*. Retrieved April 22, 2018 from https://qz.com/307214/heres-what-happened-when-a-news-site-only-reported-good-news-for-a-day/

Hamlin, J.K., & Baron, A.S. (2014), Agency Attribution in Infancy: Evidence for a Negativity Bias". *PLoS ONE*, 9(5). doi:10.1371/journal.pone.0096112.

Harvey, J. (2011), *How about we stop allowing banned users to vote?* Retrieved April 22, 2018 from https://www.reddit.com/k1cgz

LaViolette, J. (2017), *Cyber-pragmatics and alterity on reddit.com*. Retrieved April 22, 2018 from https://www.tilburguniversity.edu/upload/6614d6f8-3b03-4c8a-8ac9-b56ecf4b9cb1_TPCS_196_LaViolette.pdf

Morewedge, C.K. (2009), Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, 138(4), 535–545. doi:10.1037/a0016796

Munroe, R. (2009), *reddit's new comment sorting system*. Retrieved April 22, 2018 from https://redditblog.com/2009/10/15/reddits-new-comment-sorting-system/

Palminteri S., Lefebvre G., & Kilford E.J., Blakemore S-J. (2017), Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback

processing. *PLoS Computational Biology*, 13(8). doi:10.1371/journal.pcbi.1005684

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

Parramore, L. (2011), *Eli Pariser on the future of the Internet.* Retrieved April 22, 2018 from https://www.salon.com/2010/10/08/lynn_parramore_eli_pariser/

Pfeffer, J., Zorbach, T., & Carley, K.M. (2014), Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20(1–2), 117–128. doi:10.1080/13527266.2013.797778

Stanovich, K.E., West, R.F., & Toplak M.E. (2013), Myside Bias, Rational Thinking, and Intelligence. *Current directions in Psychological Science*, 22(4), 259–264. doi:10.1177/0963721413480174

Trussler, M., & Soroka, S. (2014), Consumer Demand for Cynical and Negative News Frames. *The International Journal of Press/Politics*, 19(3), 360–379. doi:10.1177/1940161214524832