Selected Papers of #AoIR2023:
The 24th Annual Conference of the
Association of Internet Researchers
Philadelphia, PA, USA / 18-21 Oct 2023

AFTER DEPLATFORMING: RETRACING CONTENT MODERATION EFFECTS ACROSS PLATFORMS

Emillie de Keulenaar
University of Groningen

Marcelo Alves dos Santos Junior
Pontifical Catholic University of Rio de Janeiro

João Carlos Magalhães
University of Groningen

Richard Rogers
University of Amsterdam

Bharath Ganesh
University of Amsterdam

Abstract

Half a decade ago, social media platforms were widely perceived as revolutionary devices for maximizing political expression around the world. By opening the floodgates to expression, however, the same platforms were also accused of opening the floodgates of hate – allowing, for example, the self-claimed "revolutionary" return of ideas, speech and actors long thought to be relegated to the dustbins of history. This panel examines a three-fold revolution, namely: populist revolutions (on the right) facilitated by agnostic content moderation philosophies; the internal revolutions that platform content moderation underwent to address the political violence of the former; and the adjustments that digital methods research needs to adopt to facilitate content moderation research in a "post-API" environment. The first paper of this panel examines how Twitter's content moderation has undergone several arbitrary changes before reaching a form of "normative plasticity", with reinforcement techniques such as demotion and other forms of conditional content obfuscation. The second paper looks at how, despite making profound changes to prevent furthering political violence during

Suggested citation (APA): de Keulenaar, E., Alves dos Santos Junior, M., Magalhães, J. C., Rogers, R., & Ganesh, B. (2023, October). *After deplatforming: Retracing content moderation effects across platforms*. Panel presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from: http://spir.aoir.org

elections, Twitter, Facebook, YouTube and Instagram have tended to moderate the Brazilian elections in a dislocated fashion, turning a blind eye to Brazilian militaristic content and focusing instead on what it primarily moderates in a US context. Finally, the third paper offers a set of methods for empirical researchers to capture and study content moderation metadata over time. All three papers aim to contribute to attempts at archiving and studying speech moderation as a public good, in an international context.

Introductory statement

Half a decade ago, social media platforms were widely perceived as revolutionary devices for maximizing political expression around the world (Klonick, 2018). At the heart of this view was a newfound opportunity to open the "the floodgates to online expression" (Keipi et al., 2016, p. 114) with affordances that circumvented 'legacy' media networks strongly gatekept by corporate or governmental interests. One of the key affordances to free speech was a largely non-interventionist philosophy of content moderation, according to which platforms, as "tech companies" (Castillo, 2018), did not bear the right to adjudicate the moral, political or epistemic validity of user-generated content. Besides being a convenient philosophy for platforms to function within free and "multi-sided markets" (Rieder and Sire, 2014), it was also a founding characteristic of platforms as public forums intended to consolidate "big data" as the most diverse possible conglomeration of information, ideas and user cultures.

By 2021, this positive take on content moderation had largely faded. By opening the floodgates to expression, the same platforms were also accused of opening the floodgates of hate – allowing, for example, the return of ideas, speech and actors long thought to be relegated to the dustbins of history (Fœssel, 2021). News media and scholarly literature have tended to focus on the "return" of extreme right-movements in Europe and the U.S., with antiquated ideas of race and identity making their ways into the mainstream of online public spheres. Oft-forgotten cases also touch upon Latin America, where a once widely condemned, authoritarian brand of Brazilian militarism has been normalized by a string of pro-military hashtag campaigns, Facebook and Instagram influencers, Telegram groups, YouTube and other content throughout Bolsonaro's presidency (2018-2022). Ironically, these movements have not called for a return of authoritarianism per se (at least not explicitly), but demanded their own, competing version of an open democracy whose anti-establishment values merited revolutionary means of realization, as seen in Washington D.C. on January 6, 2021, or in Brasília on January 8, 2022.

In response, social media platforms have sought to return to some of the historically constituted speech norms they had initially destabilized. Against the re-emergence of antisemitic content online, for example, YouTube, Twitter and Facebook introduced specific clauses against Holocaust denialism in 2018 (YouTube, 2019); against biological brands of racism and white nationalism, they have introduced new clauses in their "hateful conduct" policies throughout 2017 (see **Paper 1**); and against militaristic content in Brazil, Facebook had made a specific policy forbidding such content through arguably imperfect means (Catucci, 2022). These changes partook in a longer revolution in the field of content moderation, which, since 2017, saw platforms zig-zagging through largely arbitrary decisions as to what users can and cannot say in the public spheres they maintain. One of the latest u-turns in this field, for example, has been Elon Musk’s claim to a return to ideological “neutrality” in Twitter’s content moderation – which has, since his arrival, favored content on the right of the American, Brazilian and European political spectra (Getahun and Tangalakis-Lippert, 2022).

This panel examines a three-fold revolution, namely: popular revolutions (on the right) facilitated by agnostic content moderation philosophies; the internal revolutions that platform content moderation underwent to address the political violence of the former; and the adjustments that digital methods research needs to adopt to facilitate content moderation research in a “post-API” environment (Perriam, Birkbak and Freeman, 2020). The first paper of this panel examines how Twitter’s content moderation has undergone several arbitrary changes before reaching a form of “normative plasticity”, with reinforcement techniques such as demotion and other forms of conditional content obfuscation. The second paper looks at how, despite making profound changes to prevent furthering political violence during elections, Twitter, Facebook, YouTube and Instagram have tended to moderate the Brazilian elections in a dislocated fashion, turning a blind eye to Brazilian militaristic content and focusing on what it primarily moderates in US elections: electoral disinformation. In doing so, it offers a meta-analysis of content moderation practices (deletion, labeling and demotion) in five platforms, both “fringe” and “mainstream” (De Zeeuw and Tuters, 2020): Telegram, Facebook, Instagram, YouTube, and Twitter. Finally, the third paper offers a set of methods for empirical researchers to capture and study content moderation metadata – which is increasingly prone to the volatility of a “post-API” research environment.

All three papers aim to contribute to digital methods (and related) scholarship that attempt to archive and study speech moderation as an international public good. While paper 3 offers a set of methods to allow scholars to gain systematic access to content moderation metadata, paper 1 offers a wider historical outlook on how the recently acquired Twitter reflects on larger societal transformations in corporate and public speech norms. Paper 2 highlights how a “dislocated” form of US-based content moderation complicates difficult attempts at finding truth and reconciliation in fragile democracies, and outlines a few guidelines for policy makers and platform trust and safety teams to focus on preventing contention around issues that threaten deep ruptures of consensus.

Paper 1: Modulating moderation: a history of objectionability in Twitter moderation practices

Half a decade ago, social media platforms were widely seen as online models of nominally liberal, democratic societies (Klonick, 2018). At the heart of this view was their newfound opportunity to open the “the floodgates to online expression” (Keipi et al., 2016, p. 114) with affordances that circumvented strongly gatekept “legacy” media networks. One of these affordances was a largely non-interventionist philosophy of content moderation, according to which platforms do not bear the right to adjudicate the moral, political or epistemic validity of user-generated content, in order to host the most diverse possible conglomeration of ideas, information and user cultures. But when the same platforms were also accused of opening the floodgates of hate – allowing, for example, the return of ideas, speech and actors long thought to be relegated to the dustbins of history (Fœssel, 2021) – they were pressured to consider the many ways in which the link between “online speech” and “offline harm” could become “demonstrably real” (Dorsey, 2021).

This question alone invites many different answers, encapsulated within different ideals of social media platforms as “public spheres” and speech as a public good. Twitter, in particular, went through several approaches to moderate what it called “hateful conduct”, “abusive behavior”, “violent threats” and other more granular forms of verbal abuse. It invested in more human moderators, expanded systems of automated speech control, and had legal experts contribute to content moderation policies for an increasingly complex array of problematic situations. At the core of this process were considerations as to what amounted to objectionable language, a question reflected in ongoing battles of ideas as to what behavior, words and historical periods constitute falsehoods and offenses to religion, race, gender, and other forms of identity. By adopting some definitions over others, and thereby removing certain content over others, Twitter became an intermediary to comprehensive but highly contentious transformations of public speech norms. In the process, its boundaries as an aspiring digital “public sphere” have broadened, retracted and become altogether more fluid.

The general contours of the modulations of Twitter as an online public sphere are relatively clear. But partly because they involve an opaque and private organization, little scholarship has systematically documented, examined, and theorized the evolution of Twitter’s definitions of objectionable speech – particularly after its acquisition by Elon Musk. This article contributes to this research by analyzing how Twitter’s conceptions and approach to objectionable speech changed between 2006 and 2022. Methodologically, we look at both *concepts* and *techniques* that aim to define and discipline objectionable language linked to what Twitter calls “hateful conduct” and “abusive behavior”. This implies a systematic web history (Brügger, 2013) of several content moderation policies published by the platform, taking note of what is deemed problematic, specific examples offered for each type of content, and the techniques used against them. We then use a combination of digital methods (Rogers, 2013) that rely on both scraping and Twitter’s Academic API to trace the moderation of objectionable speech in practice. We examine how a dataset of Tweets linked to the U.S. elections of 2020 and online slurs were demoted, flagged, removed and (some) eventually “redeemed” after a period of temporary suspension. This means tracing instances of updated, overwritten, removed and “replatformed” content as various “platform effects” (Malik & Pfeffer, 2016).

Paper 2: Three dislocated content moderation enforcements of political militarism during the Brazilian elections of 2022

This paper examines traces of content moderation by Twitter, YouTube, Facebook, Instagram and Telegram during the Brazilian electoral year of 2022 to 2023, specifically between August 15, 2022 (the official start of political campaign) and January 15, 2023 (one week after the January 8 depredation). It argues that moderation suffers from three “dislocations”. First, US-based platforms lack a regional comprehension of the national political history and speech legislation of Brazil, which causes them to moderate incriminating content with a primarily US understanding of local political violence. Second, content moderation enforcement is dislocated *in time*, in the sense that it is applied primarily during the election cycle, missing out of violence incited *as a result* of electoral results. And third, content moderation is dislocated by users, who devise strategies to circumvent moderation with language obfuscation, VPNs, and a broader network of fringe-to-mainstream platforms. Empirically, this paper builds on digital methods and “dynamic archiving” of content moderation traces (see **Paper 3**) to recreate enforcement mechanisms and moderation decisions over time.

Since the United States Capitol Storm on January 6, 2022, there has been a growing concern over the propagation of electoral disinformation on social media (Munn, 2021). The #StoptheSteal campaign was largely spread on social media challenging the timely removal of illegal or otherwise unsafe content from multiple platforms (see **Paper 1**). One year later, the far-right president Jair Bolsonaro was defeated in a runoff on October 30, 2022, after the center-left won the Brazilian presidential election by a slim margin of 2 million votes. Even though Brazilian events suggest a similar scenario of the delegitimization of elections resulting from social media and violent acts of invasion and depredation of public buildings, local specificities add another layer of complexity to content moderation policies in politically unstable democracies.

On YouTube, we sampled a list of 849 public channels that produce content about the Brazilian elections. This rendered a unique dataset of 193,482 videos. Next, we created an R script using the tubeR package that connects to Youtube Data API v3 and collects the last 10 videos of each channel every six hours to record video metadata, including rankings for further analyses of demotion. At the end of each week, we also used the scraping library youtube-dl (Gonzalez et al. 2023) to extract transcripts of the content. Then, we called the Youtube Data API v3 to determine which videos were unavailable and scraped the platform's reasons for removal. We complemented this dataset by using youtube-dl to search videos mentioning any of 365 queries susceptible to returning problematic (and thus likely moderated) results about the Brazilian elections. Some examples include hashtags supporting a military coup (#SOSFFAA), denying the integrity of the electoral process (#BrazilWasStolen), or attacking Brazilian institutions responsible for overseeing the elections and a peaceful transfer of power (#FIMdasUrnasEletrônicasJá, #ForaSTF).

On Telegram, content moderation is seldom enacted by the platform itself. Users can decide when and how to delete their content with features optimized for the kind of “private sociality” the platform promotes (Rogers, 2020), such as message auto-deletion. We also found that a lot of messages had been promptly deleted after the January 8 riots for fear of recrimination. To collect these content moderation traces, we first collected 24,905 Telegram messages from an expert list of 250 Telegram groups active between January 1, 2022 and January 15, 2023. We then compared the availability of each group and their posts using Selenium, a Web scraper, on January 10, 2023, limiting our results to the top 500 most engaged posts from August 15, 2022 to January 10.

Unlike Twitter, Instagram and Facebook are notoriously wary of scrapers, and have routinely banned researchers for using them on their sites (Bond, 2021). For this reason, we have opted to manually copy the statuses of the 500 most engaged posts from August 15, 2022 to January 10, 2023. The data we have copied moderation statuses from are 678,029 Facebook and Instagram posts collected with the above-mentioned queries on Crowdtangle. On Twitter, we again used Selenium to scrape the statuses of all Tweets dating from January 1, 2022 to January 10, 2023. This was done on a dataset of 3,486,622 Tweets obtained with the Academic API, using the same queries as above. For consistency, we filtered all results from all platforms to posts dating from August 15, 2022 to January 10, 2023.

The first problem revealed by our data is a general "dislocation" in US-based content moderation policies. By this, we mean that YouTube, Twitter, Facebook and Instagram unilaterally conceived their content moderation policies based on their experience of the US elections of 2020, overlooking the local institutional, political and historical factors in Brazil that contributed to the violence seen on January 8. YouTube's Electoral Misinformation Policies list multiple rules that penalize false information on electoral procedures, such as voter suppression, obstruction or claiming widespread fraud. Nonetheless, Brazil's electoral issues are derived from a longstanding history of institutional ruptures caused by military and political coups d'État, the latest of which resulted in a military dictatorship that ended in 1985. Brazilian law includes safeguards against defenses of the military dictatorship, for encouraging a military intervention, or for inciting a democratic breakdown (de Albuquerque, 2020). We find that a majority of moderated videos and posts are moderated not for these reasons but for spreading electoral disinformation; content that called for a military coup, even by militaries themselves, was left mostly unmoderated.

The second issue is a *temporal* dislocation of content moderation. By this, we mean that platforms tended to relax their enforcement policies after the election ended on October 31, disregarding how violence was fostered following the electoral results. Moderation after the fact, i.e., after October 31st, was insufficient to mitigate the propagation of electoral disinformation or convocations to violent protests. This dislocation in time is caused by two major factors. One concerns platforms' period of attention, monitoring and coverage of the electoral cycle. On YouTube, for example, we found that the majority of offline videos were created prior to October. The second factor is strike waivers granted by YouTube. The enforcement of new policy clauses targeting electoral disinformation always gave the platform a 30-day buffer time, during which content could be removed without penalizing entire channels.

The last shortcoming is *user* dislocation, i.e., ever-evolving content moderation evasion strategies. From August 2022 to January 2023, users engaged in calls for military coups or disseminating electoral misinformation routinely exchanged tips for evading moderation from either platforms or from the Supreme Federal Court. Some of these strategies included using VPNs to access content exclusively banned in Brazil, including detailed plannings of the January 8 attacks on Twitter. Another was language obfuscation, which consisted in slightly changing terms that are likely to be blacklisted in content moderation word lists ("fraud", for example, becomes "fr4ud" or "f.r.4.u.d."). A more complex strategy consisted in exchanging incriminating information in an increasingly complex network of more or less moderated platforms. Users of Facebook, Instagram or YouTube have also relied on Telegram, GETTR, BitChute, Rumble and file transfer websites to exchange plans for the January 8 attacks, or documents purportedly showing electoral fraud. This "fringe-to-mainstream" platform ecology makes platform-specific content moderation ineffective on a larger scale.

These three dislocations contribute to a larger issue stemming from US-based content moderation of Brazilian political content, which is that it complicates a difficult process of truth and reconciliation of a fairly new and fragile democracy. Since the military regime ended in 1985, tortures and other human rights abuses by the military regime had never been prosecuted. The return of militarism to the mainstream of Brazilian public debate has also contributed to a serious institutional crisis after January 8, where the judiciary branch pained to keep military branches of power in check. We make recommendations for platform trust and safety teams to collaborate with local authorities to broaden the nature of their work outside of a punitive logic, and consider moderating – in the sense of *balancing* – long-term dialogue between opposing political voices.

Paper 3: After Deplatforming: the Return of Trace Research for the Study of Platform Effects

When platforms were considered as “intermediaries”, or conduits through which content would flow unfettered (Gillespie, 2018), they could serve as sites for the study of user traces: online actions and behaviors interactively registered by the platforms and made available to researchers in the form of hit logs, links, likes, retweets, shares and so forth. One could observe user behavior and attitude as if ‘in the wild’. These ‘unobtrusive measures’ would provide insights in collective mood and sentiment, give indicators of opinion and perhaps even capture animal spirits behind stock price or cryptocurrency movements (Watts and Dodds, 2007; Lazer *et al.*, 2008).

But there is a certain artificiality to platform data in at least four senses. On the platform side, there are attempts to make content “stick” in order to increase the time users spend consuming it on their services through the optimization of “watch time” or even the introduction of “dark patterns” underlying user interfaces that steer users towards forms of conversion (such as purchases). There are also content moderation techniques, such as deplatforming, labeling, or demoting objectionable content. On the user side, one finds the amplification or manipulation of user content, such as inflating likes and views, acquiring fake followers, gaming recommendation algorithms in order to up-rank certain content on top of their peer’s lists of visible content. There is optimisation from SEO (search engine optimisation) to analytics-driven content production. Users also protect themselves by employing code words while linking to extreme material, introducing (mirrored) images with explicit text, or setting accounts or posts to private to dodge the oversight of public or platform content moderation.

Though there have been more or less isolated studies about the function, roles and construction of each of these effects, there is no clear idea about their overall significance. There are of course platform transparency reports and blog posts detailing the percentage of content taken down by hate speech, misinformation or state legislations. This kind of documentation is most often framed as “PR exercises”, that is, diplomatic statements that one is required to read between the lines — or, in any case, secondary sources of information with no strong objective value.

Ironically, a solution to the “untraceability” of platform transparency reports has been to do trace research. That is, one needs trace research to answer the question of platform effects. In the field of content moderation studies, some examples are recovering traces of content removal by continually archiving posts, Tweets or videos (de Keulenaar *et al.*, Forthcoming); following patterns of demotion by scraping the rankings of search results on YouTube or Twitter (Keulenaar, Burton and Kisjes, 2021); or scraping labels, “context” flags and other post statuses on Twitter over time.

On this matter, this chapter discusses a new kind of “digital forensics”, or set of methods one can use to reconstruct platform and user traces in content moderation. In other words, it proposes methods to reconstruct the scene after or on which platform or user data has disappeared in the context of one specific platform effect: content moderation. We discuss five examples: contextualizing the disappearance of user data by doing a web history of content moderation policies with the Wayback Machine; reverse-engineering content moderation with “dynamic archiving” of data susceptible to being removed, demoted and otherwise moderated; and using platform metadata or scraping moderation traces, such as flags, prompts and labels. We also include one example of user effects, namely, capturing neologisms designed to obfuscate platform moderation, and tracing outlinks to alternative platforms to redirect audiences toward content that the host platform does not allow.

Each of these research techniques rely on the traces remaining after content has been re-versioned, updated, overwritten, suppressed or removed. As such, it puts trace research to a new use. Rather than studying user behavior by treating the platform as an intermediary, it studies platform behavior by treating users as actively moderated. It thereby constitutes a different strategy for deploying trace research. In the earlier version of trace research, one would collect platform data about users and consider the extent to which it can be cleaned to remove platform effects and other content befouling causes (such as bots, optimized content or artificial amplification). Now the trace research takes as its point of departure platform efforts to cleanse and police the site of rule-breaking or offending content, artfully capturing those acts. Trace research thereby gains a new purpose. It seeks the answer, if only in part, to the question of the effects of content moderation on the content under study. In a sense, it could be seen as a necessary step prior to content analysis or trace research in its original form. Or it could be considered a research practice in itself that has as its main aim content reconstruction, which has a series of uses.

We first outline the methods and use cases linked to each of our examples — first, for platform effects, and then, for user effects. We then conclude with a discussion on the larger significance of empirical content moderation research, be it for scholars invested in digital methods, platform governance, or historians tracing the modulation of speech norms in complex media environments. We also cover important limitations — some of which are proper to any kind of historiographical (Web) research (Arora et al., 2016), and others due to the vagaries of “post-API” research (AoIR, 2018).

References

AoIR staff (2018) *Facebook Shuts the Gate after the Horse Has Bolted, and Hurts Real Research in the Process*. Available at: https://aoir.org/facebook-shuts-the-gate-after-the-horse-has-bolted/ (Accessed: 1 March 2023).

Arora, S.K. *et al.* (2016) ‘Using the wayback machine to mine websites in the social sciences: A methodological resource’, *Journal of the Association for Information*

Science and Technology, 67(8), pp. 1904–1915. Available at: https://doi.org/10.1002/asi.23503.

Bond, S. (2021) ‘NYU Researchers Were Studying Disinformation On Facebook. The Company Cut Them Off’, *NPR*, 4 August. Available at: https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f (Accessed: 1 March 2023).

Brügger, N. (2013) ‘Web historiography and Internet Studies: Challenges and perspectives’, *New Media & Society*, 15(5), pp. 752–764. Available at: https://doi.org/10.1177/1461444812462852.

Castillo, M. (2018) ‘Zuckerberg tells Congress Facebook is not a media company: “I consider us to be a technology company”’, *CNBC*, 11 April. Available at: https://www.cnbc.com/2018/04/11/mark-zuckerberg-facebook-is-a-technology-company-not-media-company.html (Accessed: 1 March 2023).

Catucci, A. (2022) ‘Facebook e Instagram começam a remover publicações com pedidos de intervenção militar no Brasil, diz Meta’, *G1*, 4 November. Available at: https://g1.globo.com/tecnologia/noticia/2022/11/04/facebook-e-instagram-vao-remover-publicacoes-com-pedidos-para-intervencao-militar-no-brasil-diz-meta.ghtml (Accessed: 1 March 2023).

de Albuquerque, A. (2020) ‘The Two Sources of the Illiberal Turn in Brazil’, *Brown Journal of World Affairs*, 27, p. 127. Available at: https://heinonline.org/HOL/Page?handle=hein.journals/brownjwa27&id=407&div=&collection=.

de Keulenaar E, Kisjes I, Smith R, et al. (2023) Twitter as accidental authority: how a platform assumed an adjucative role during the COVID-19 pandemic. In: Rogers R (ed.) *The Propagation of Misinformation in Social Media: A Cross-Platform Analysis*. Amsterdam: Amsterdam University Press, pp. 109–138. Available at: https://www.aup.nl/en/book/9789048554249.

de Keulenaar, E., Burton, A.G. and Kisjes, I. (2021) ‘Deplatforming, demotion and folk theories of Big Tech persecution’, *Fronteiras - estudos midiáticos*, 23(2), pp. 118–139. Available at: https://doi.org/10.4013/fem.2021.232.09.

de Zeeuw, D. and Tuters, M. (2020) ‘Teh Internet Is Serious Business: On the Deep Vernacular Web Imaginary’, *Cultural Politics*, 16(2).

Fœssel, M. (2021) *Récidive 1938*. Quadrige. Paris: Presses Universitaires de France.

Garcia Gonzalez, R., Amine, R. and M., S. (2022) ‘youtube-dl’. youtube-dl. Available at: http://ytdl-org.github.io/youtube-dl/about.html (Accessed: 22 February 2021).

Getahun, K.T.-L., Hannah and Tangalakis-Lippert, K. (2022) ‘Elon Musk says his politics are in the center but extremism experts say he’s using Twitter to increasingly empower right-wing viewpoints’, *Business Insider*, 11 December. Available at: https://www.businessinsider.com/elon-musk-right-wing-extremism-twitter-mythology-of-the-center-2022-12 (Accessed: 1 March 2023).

Keipi, T. *et al.* (2016) *Online Hate and Harmful Content: Cross-National Perspectives*. Taylor & Francis.

Klonick, K. (2017) ‘The New Governors: The People, Rules, and Processes Governing Online Speech’, *Harvard Law Review*, 131, p. 1598. Available at: https://heinonline.org/HOL/Page?handle=hein.journals/hlr131&id=1626&div=&collection=.

Jack Dorsey (2021) ‘I do not celebrate or feel pride in our having to ban @realDonaldTrump from Twitter, or how we got here. After a clear warning we’d take this action, we made a decision with the best information we had based on threats to physical safety both on and off Twitter. Was this correct?’, *Twitter*. Available at: https://twitter.com/jack/status/1349510769268850690 (Accessed: 15 July 2022).

Lazer, D. *et al.* (2008) ‘Networks and Political Attitudes: Structure, Influence, and Co-Evolution’. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.1280328.

Malik, M. and Pfeffer, J. (2016) ‘Identifying Platform Effects in Social Media Data’, *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), pp. 241–249. Available at: https://doi.org/10.1609/icwsm.v10i1.14756.

Munn, L. (2021) ‘More than a mob: Parler as preparatory media for the U.S. Capitol storming’, *First Monday* [Preprint]. Available at: https://doi.org/10.5210/fm.v26i3.11574.

Perriam, J., Birkbak, A. and Freeman, A. (2020) ‘Digital methods in a post-API environment’, *International Journal of Social Research Methodology*, 23(3), pp. 277–290. Available at: https://doi.org/10.1080/13645579.2019.1682840.

Rieder, B. and Sire, G. (2014) ‘Conflicts of interest and incentives to bias: A microeconomic critique of Google’s tangled position on the Web’, *New Media & Society*, 16(2), pp. 195–211. Available at: https://doi.org/10.1177/1461444813481195.

Rogers, R. (2020) ‘Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media’, *European Journal of Communication*, p. 0267323120922066. Available at: https://doi.org/10.1177/0267323120922066.

Watts, D.J. and Dodds, P.S. (2007) ‘Influentials, Networks, and Public Opinion Formation’, *Journal of Consumer Research*, 34(4), pp. 441–458. Available at: https://doi.org/10.1086/518527.