British Journal of Political Science (2022), **52**, 1482–1489
doi:10.1017/S000712342100017X

British Journal of Political Science

COMMENT

The Perils of Estimating Disengagement Effects of Deadly Terrorist Attacks Utilizing Social Media Data

Tanja Marie Hansen

Aarhus University, Aarhus, Denmark
E-mail: tanja@ps.au.dk

(Received 15 April 2020; revised 13 December 2020; accepted 27 March 2021; first published online 31 January 2022)

Abstract

This comment discusses the impact of social media rule enforcement protocols on research on online data sources. It argues that the conclusions of the article 'Do Islamic State's Deadly Attacks Disengage, Deter, or Mobilize Supporters?' concerning the recruitment effects of deadly attacks cannot be assumed to hold when considering the timing of Twitter account suspensions. It highlights four ways in which suspensions can confound evidence of demobilization despite the introduction of control variables and fixed-effects model specifications. All change the composition of the sample in four non-random ways. First, suspending connected Islamic State accounts may result in follower loss. Secondly, Twitter suspension procedures may be tied to account characteristics, such as follower accrual rates. Thirdly, suspended accounts that re-emerge introduce replication bias. Fourthly, account closure may reflect user movement to other platforms in response to changing security environments following deadly attacks. In conclusion, caution is advised when platform-introduced variation risks altering the sample composition in non-random ways.

Keywords: political violence; terrorism; signaling; social media data

The study of patterns of online behavior has gained ground in recent years. The topics studied have ranged from online election tampering and disinformation campaigns to censorship, foreign influence operations and patterns of online radicalization. However, one important aspect is frequently overlooked within this up-and-coming research tradition – researchers' limited control of the data-generating process, and the bias this risks introducing to analyses of social media data.

The opaque process that underlies rule development and enforcement introduces unknown and potentially systematic variation to analyses of social media data. Social media platforms are essentially exempt from legal liability for the content posted by their users in both Europe and the United States (Nouri, Lorenzo-Dus and Watkin 2019, 4–5), and platform policy statements like 'Each situation is evaluated on a case by case basis and ultimately decided upon by a cross-functional team, [and is influenced by] the changing nature of online speech, including how our rules are applied and interpreted in different cultural and social contexts' (Twitter: Our approach to policy development and enforcement philosophy) indicate large-scale variation among and within platforms. As such, patterns not only differ from country to country, as rules are enforced differently depending on cultural and social contexts, but also over time as social media platforms are free to change the rules and procedures that guide their rule enforcement. For instance, such platforms can alter the number of employees tasked with ensuring rule enforcement, the technological capabilities that bring potential rule breakers to its attention, or the type of content deemed to violate rules as political and social climates change in surrounding societies. Social media platforms are not required to disclose or elaborate on these changes to the public or researchers.

https://doi.org/10.1017/S000712342100017X Published online by Cambridge University Press

In this article, I delve into a recent study by Barceló and Labzina (2020) on terrorist (de)mobilization to demonstrate the undue influence that social media platform enforcement decisions can have on otherwise strong research designs. I further discuss the strong demands these problems place on researchers, who need to invest additional effort into guarding their models against such empirical challenges. In the study by Barceló and Labzina, platform interference takes the form of account suspensions, as platforms (in this case Twitter) weed out accounts with terrorist connections and those glorifying violence (Twitter: Terrorism and violent extremism policy; Glorification of violence policy).

I present four ways in which account suspensions on social media platforms can impact measures of online mobilization, all of which should be taken into account in future research in this area. First, platforms' suspension of accounts might be related to the events being studied. For instance, accounts that accrue followers at a higher rate are also more likely to face suspension as they draw more attention to themselves. Secondly, account suspension can lead to follower loss in other related accounts. Thirdly, suspended accounts that resurge under new but similar aliases risk introducing replication error to analyses, as (near) identical accounts are treated as unique accounts (identical in all but user ID) – essentially leading the researcher to count the same account multiple times. Fourthly, voluntary account closure and user movement to other platforms potentially coincide with perceived security risks influenced by the timing of deadly attacks. For example, a drop in followers in the aftermath of deadly attacks could be a result of deterrence, as Islamic State (IS) members move to other, safer, online platforms rather than demobilization and disaffiliation with IS online activities.

The article proceeds as follows. I first present the theory and findings of the original article on online demobilization published by Joan Barceló and Elena Labzina. I then discuss in detail the four ways account suspensions can affect measures of online mobilization, how the authors guarded their analysis against these challenges, and why there is nevertheless still cause for concern about the robustness of the results. I conclude by broadening the discussion to the general topic of biases inherent to the study of social media data.

An Overview of the Study By Barceló and Labzina

In their article 'Do Islamic State's Deadly Attacks Disengage, Deter, or Mobilize Supporters?', Barceló and Labzina discuss three separate and competing hypotheses regarding the expected effect of deadly terrorist attacks on Twitter follower behavior: attacks may (1) mobilize explicit online support by displaying organizational capacity of deadly violence and increased attention in the news, etc., (2) disengage online supporters who become morally repulsed by excessive use of violence or (3) deter online supporters who are fearful of facing personal ramifications of affiliation with IS. A mobilization effect should be visible as an *increase* in the number of followers of IS accounts in the aftermath of IS terrorist attacks, whereas both disengagement and deterrence effects are expected to show as reductions in Twitter follower counts.

Barceló and Labzina analyzed 127 days of Twitter data from the Spring of 2016 on the number of followers of Twitter accounts associated with IS, and found evidence that deadly terrorist attacks have a demobilizing effect. They test their hypotheses using a two-step approach that combines (1) an interrupted time-series design centering on the large-scale IS attacks in Brussels and Nice[1] and (2) a panel regression design incorporating a total of 70 IS-attributed attacks during the study period. The panel data approach incorporates a smoothing temporal approach, allowing deaths to impact (de)mobilization for an extended period of time beyond the day of the attack.

[1] Table 1 in the Appendix provides recalculated p-values reflecting a two-sided significance test, indicating that the Brussels attack was less statistically significant than first estimated.

https://doi.org/10.1017/S000712342100017X Published online by Cambridge University Press

Taking Suspensions into Account

Social media account closure – whether voluntary closure or forced suspension – is a widely discussed phenomenon in the debate and literature on terrorist presence on social media platforms (Berger and Perez 2016; Conway et al. 2019; Huey 2015; Klausen, Marks and Zaman 2018; Mitts 2020; Nouri, Lorenzo-Dus and Watkin 2019; Pearson 2018; Shehabat and Mitew 2018; Weimann 2019; Wright et al. 2016). To combat the spread of terrorist content, many social media platforms have engaged heavily in efforts to close accounts with terrorist affiliations (Abutaleb 2016; Alba, Edmondson and Isaac 2019; Twitter: 'An Update on Our Efforts to Combat Violent Extremism' 2016), especially after the surge of IS on Twitter during the early 2010s (Pearson 2018; Shehabat and Mitew 2018).

Figure 1 charts the overarching causal chains of the study by Barceló and Labzina and emphasizes why social media data is a difficult data source for the exploration of what motivates human behavior – especially under conditions of platform-introduced variation. While their study attempts to estimate whether deadly terrorist attacks cause mobilization, disengagement or deterrence among online sympathizers, its dependent variable measures changes in Twitter followers of IS-related accounts. They interpret an increase in the number of followers as a sign of increased mobilization. They treat a loss of followers as either sympathizers' loss of belief in the cause (disengagement) or sympathizers' strategic choice to cut ties with the organization out of fear of repercussions (deterrence). Either of the second two motivations should be visible as a decision to unfollow.

From the outset, this logic entails a causal leap of faith, as it carries the strong assumption that individual psychological effects can be gauged from patterns of online activity. In addition to this, the suspension related activity of Twitter raises questions about whether reductions in the number of Twitter followers always reflect individuals' decisions and motivations to unfollow accounts. What can be inferred about the mental states of suspended account holders who were involuntarily shut down? What if an overall loss of followers reflects spill-over effects from suspended accounts that used to follow other IS accounts? What if suspended accounts quickly reappear under new unique account IDs, but struggle to regain their previous followings? What if a loss of followers on Twitter reflects moves to other, safer, platforms where sympathizers are perhaps even encouraged to mobilize further, and expand their range of engagement with IS activities? And most importantly, what if the very initiation of suspension activities is tied to the main independent variable of interest: the timing of deadly terrorist attacks?

A visualization of the impact of Twitter account suspensions on the total number of IS followers in the dataset collected by Barceló and Labzina, presented in Figure 2, elicits concerns that suspensions are indeed a force to be reckoned with when utilizing social media data. The timing of suspensions appears to be non-randomly tied to the timing of major terrorist events,[2] and suspension activity seems to strongly restrict the total number of followers of IS accounts.

Additionally, when the impact of major deadly terrorist attacks is gauged by the mean number of Twitter followers – a daily number indicating the average number of followers per account – the demobilizing effect of deadly terrorism reported by Barceló and Labzina becomes harder to spot with the naked eye. As their study utilized a fixed-effects within-unit research design, the findings of Figure 2, with its focus on across-account variation, do not undermine the finding of a demobilizing effect of deadly attacks. The figure merely highlights that suspensions are worthy of researcher attention when estimating social media followings, as they introduce a source of non-random variation.

[2] In 2016 Twitter also directly mentioned 'spikes in suspensions immediately following terrorist attacks' (Twitter: 'An Update on Our Efforts to Combat Violent Extremism').

https://doi.org/10.1017/S000712342100017X Published online by Cambridge University Press

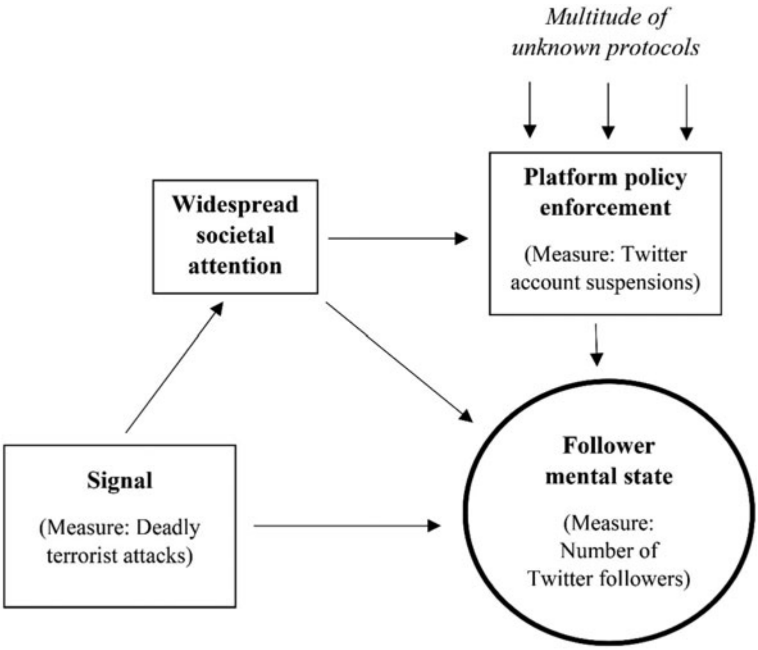
Multitude of unknown protocols
Widespread societal attention
Platform policy enforcement
(Measure: Twitter account suspensions)
Signal
(Measure: Deadly terrorist attacks)
Follower mental state
(Measure: Number of Twitter followers)

Figure 1. Causal chart of the study by Barceló and Labzina

Account Suspension, Resurgence and Voluntary Closure as Sources of Bias

Suspension, resurgence and voluntary account closure can cause concerns when utilizing social media data for analytical purposes, as they alter the composition of the sample in non-random ways. Below I demonstrate four ways in which this may have occurred in the panel data analysis of IS (de)mobilization by Barceló and Labzina, all of which should be considered more generally in this type of research.

First, Twitter's account suspensions are potentially related to the event being studied. For instance, accounts that accrue followers at a faster pace than less active accounts can draw attention to themselves and thus face a greater risk of suspension. This process is likely intensified following violent attacks that highlight the importance of terrorists' online presence. In 2016, Twitter hinted as much, writing that 'Daily suspensions are up over 80 percent since last year, with *spikes in suspensions immediately following terrorist attacks*' (Twitter: 'An Update on Our Efforts to Combat Violent Extremism' 2016, emphasis added). As suspension removes these accounts from the sample, the accounts left for estimation are biased towards less active accounts, which themselves can lose followers during such periods as a general trend due to lower activity and posting levels. This would skew estimations in the direction of lower or even negative follower growth, akin to that found in the analyses by Barceló and Labzina (2020).

Secondly, suspending one account can lead to follower loss in other related accounts. The speedy follower-accrual rates upon resurgence found by Wright et al. (2016), and the existence and maintenance of lists of accounts that are supportive of IS intended to guide IS supporters to follow each other (Berger and Perez 2016),[3] indicate the tight-knit community characteristic of IS' online presence. IS accounts appear to follow each other. Given that the owners of IS accounts deliberately follow other IS-related accounts that share similar information and hold

[3] A famous example was the 'Baqiya Shoutout' list on Twitter in 2015, which was 'compiled manually by the Baqiya Shoutout user, a prolific and highly motivated ISIS social media activist whose online activity was devoted primarily to network-building' (Berger and Perez 2016, 5).

https://doi.org/10.1017/S000712342100017X Published online by Cambridge University Press

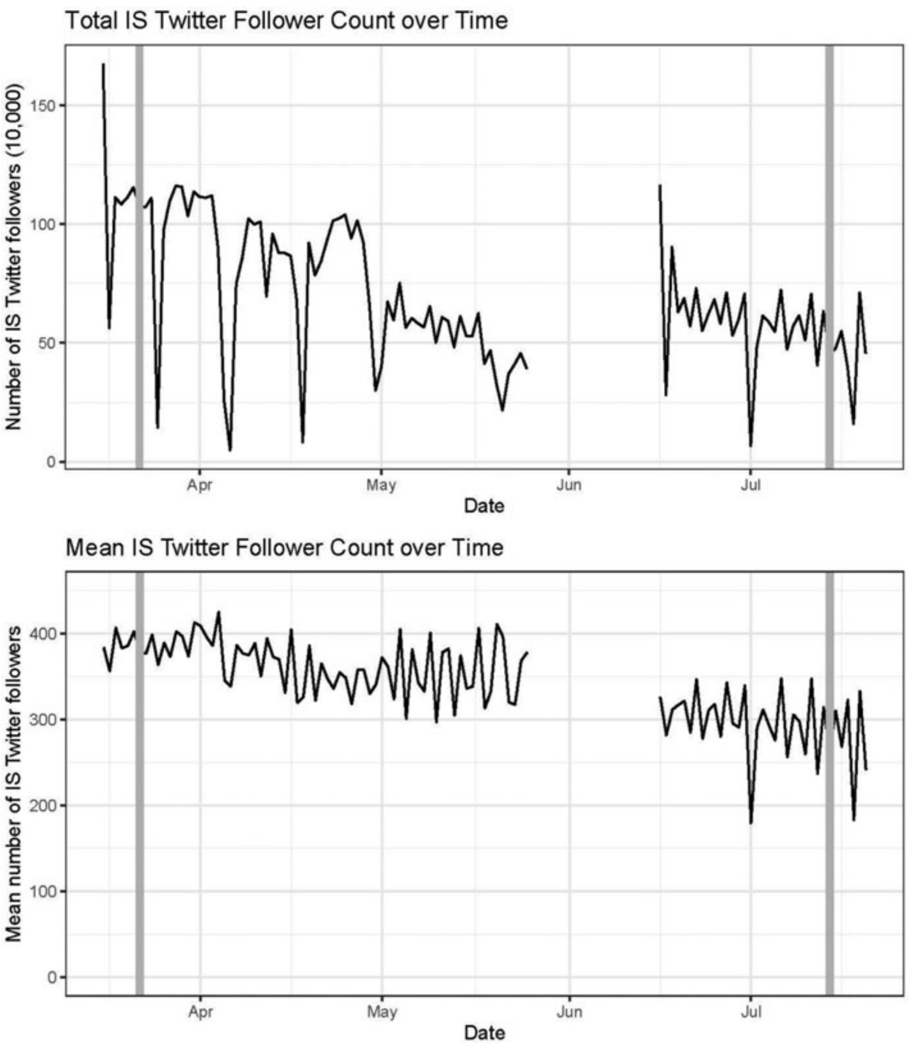

Figure 2. Developments in IS Twitter followers over time

comparable world views, suspension of one account can cause follower loss not only in the suspended account, but also in the remaining affiliated IS accounts that lose followers.

Thirdly, suspended accounts that resurge under new user IDs risk introducing bias as they re-enter the data. Terrorists often react to suspension by promptly returning to the platforms under new aliases, and quickly regaining followers through previously established networks (Wright et al. 2016). Some researchers even argue that terrorists see suspension as akin to an online battle wound, something that can be survived with glory by re-entering the platform, and they describe community building effects of the process of resurgence (Pearson 2018). Wright et al. (2016) highlight the risk of replication error resulting from resurgent accounts being treated as unique accounts – essentially counting them multiple times, thus reinforcing their weight in the analysis. The analysis by Barceló and Labzina could suffer from this bias, as their dataset includes new user IDs as late as day 127 – the last day of data collection. If more accounts are suspended following terrorist attacks, more accounts can also resurge in the aftermath of attacks. Although resurgent accounts regain some of their previous followers, this

https://doi.org/10.1017/S000712342100017X Published online by Cambridge University Press

is a lengthy process: Berger and Perez (2016, 4) found that, 'Over time, individual users who repeatedly created new accounts after being suspended suffered devastating reductions in their follower counts'. If these follower-amputated resurgent accounts re-enter the analysis at higher levels after terrorist attacks, it could bias estimations in the direction of lower follower growth, akin to that found in the analysis by Barceló and Labzina (2020). It is unclear whether this reflects follower disenchantment with IS.

Fourthly, attack-timed drops in the number of active accounts may have less to do with suspensions and more to do with voluntary account closures. By quoting FBI Director James Comey, the authors themselves point out the tendency for recruits to be 'move[d] […] to an encrypted mobile-messaging app so they go dark to us' (Barceló and Labzina 2020, 1,543). Within the field of terrorism studies, several scholars have discussed this tendency, showing how terrorist online activity in the mid-2010s partially moved to other, more secure platforms like the encrypted platform Telegram and the anonymous file-sharing site Justpaste.it following the surge in suspensions on mainstream platforms (Berger and Perez 2016; Bloom, Tiflati, and Horgan 2019; Mitts 2020; Nouri, Lorenzo-Dus, and Watkin 2019; Shehabat and Mitew 2018). If IS recruiters time such moves based on the present security environment, which should be especially hostile directly following deadly terrorist attacks, large-scale account closures and inactivity – reflecting moves to more secure platforms – may co-occur with deadly terrorist attacks. Again, the timing of large-scale drops in the number of active IS-related accounts is potentially non-random and directly tied to the timing of terrorist attacks. This effect thus supports deterrence and mobilization effects rather than a disengagement effect, as sympathizers move to other online locations that enable deeper integration with the illegal terrorist network.

Guarding Analyses Against Platform Introduced Variation

Barceló and Labzina readily present their concern that 'the effects we observe may not be driven by changes in the behavior of ISIS's online audience, but instead they might be a product of the time-variant aggressive account suspension efforts by Twitter and the Anonymous's hacktivist group in the aftermath of terror attacks' (Barceló and Labzina 2020, 1,548). To account for this potential bias, they include two post-treatment measures to control for the confounding effects of an increased number of suspensions by Twitter and reports by Anonymous following terrorist attacks.[4] Though they find a highly statistically significant negative effect of their measure 'number of suspensions', and a moderately significant positive effect of 'number of reports', their overall results remain largely unchanged following their inclusion. From this they conclude that the demobilizing finding of their analysis is robust to the unwanted influence of Twitter suspensions.

There are, however, several ways in which this approach falls short of guarding the analysis against unwanted bias introduced by account suspensions, resurgence and voluntary closure, which I suggest needs to be considered in future research in this field. As I will argue in the following section, the influence of platform-introduced variation in the form of account suspensions (and subsequent resurgence) cannot be controlled away in their given research design. While the introduction of control variables goes some way towards guarding the analysis against the unwanted influence of account suspensions, it remains a lot to ask from count-based control variables, and researchers should be acutely aware of at least three problems associated with using social media data.

First, controlling for the number of suspended accounts per day does not capture spill-over effects from a loss of followers of the remaining accounts. While the number of suspended accounts may be correlated with the accompanying fall in followers, it is unclear to what degree

[4] This is only a viable option for their panel regression, as interrupted time-series analyses are run without control variables.

https://doi.org/10.1017/S000712342100017X Published online by Cambridge University Press

this is the case. The level of interconnectedness among suspended and remaining accounts can be expected to vary wildly, with some suspended accounts following many other IS accounts, and others following only a few.

Secondly, controlling for the number of suspended accounts also fails to protect the analysis against resurgent accounts re-entering the analysis under new unique account IDs. As resurgent accounts must gradually rebuild their online networks, they re-enter at reduced follower rates, biasing results in favor of a pattern of disengagement.[5]

Thirdly, the variable 'number of suspensions' does not measure voluntary account closure or account inactivity – reflecting moves to safer online platforms. Follower loss on Twitter as accounts disappear from the analysis is likely to at least partially reflect follower gains on other platforms. This is especially problematic for the analysis if such moves are deliberately timed to coincide with major terrorist attacks.

Conclusion

Social media data present new and intriguing ways to study human behavior, whether it be our fascination with videos of cats, elections or decapitations. Yet researchers must be acutely aware of the complexities associated with this type of data. Relying on social media companies to oversee the data-generating process risks introducing unintended biases that can be hard to spot and even harder to control for. In this article, I have demonstrated these biases by discussing how Twitter suspensions potentially skewed the findings of a 2020 article by Barceló and Labzina on terrorist (de)mobilization, despite their attempts to clear this hurdle by utilizing control variables. Though this example focused on the consequences of social media platform suspensions on research on terrorist (de)mobilization, it has implications for a variety of fields that study social media data, including analyses of online disinformation, hate speech, censorship, online election campaigns and foreign influence operations. When the subject matter introduces a risk of platform-introduced interference,[6] researchers must be acutely aware of the non-random ways this alters the composition of the sample.

Supplementary Material. Online appendices are available at https://doi.org/10.1017/S000712342100017X.

Data Availability Statement. Data replication sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/UY3HTY

References

Abutaleb Y (2016) Twitter Suspended 360,000 Accounts for 'Promotion of Terrorism.' *Reuters*, 18 August. Available from https://www.reuters.com/article/us-twitter-terrorism-%20idUSKCN10T1ST (accessed 29 October 2020).

Alba D, Edmondson C and Isaac M (2019) Facebook Expands Definition of Terrorist Organizations to Limit Extremism. *New York Times*, 17 September 2019. Available from https://www.nytimes.com/2019/09/17/technology/facebook-hate-speech-extremism.html (accessed 29 October 2020).

Barceló J and Labzina E (2020) Do Islamic State's deadly attacks disengage, deter, or mobilize supporters? *British Journal of Political Science* **50**(4), 1539–1559.

Berger JM and Perez H (2016) The Islamic State's Diminishing Returns on Twitter: How Suspensions Are Limiting the Social Networks of English-Speaking ISIS Supporters. Program on Extremism, George Washington University Occasional Paper (February).

Bloom M, Tiflati H and Horgan J (2019) Navigating ISIS's preferred platform: telegram1. *Terrorism and Political Violence* **31**(6), 1242–1254.

Conway M et al. (2019) Disrupting Daesh: measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism* **42**(1–2), 141–160.

Hansen TM (2021) "Replication Data for: The Perils of Estimating Disengagement Effects of Deadly Terrorist Attacks Utilizing Social Media Data", https://doi.org/10.7910/DVN/UY3HTY, Harvard Dataverse, V1.

[5] This is only problematic in random-effects models, as fixed-effects models measure effects at a within-unit level.

[6] In addition to suspension this can include things like viewer warnings, content deletion and limited distribution.

Huey L (2015) This is not your mother's terrorism: social media, online radicalization and the practice of political jamming. *Journal of Terrorism Research* **6**(2), 1–16.

Klausen J, Marks CE and Zaman T (2018) Finding extremists in online social networks. *Operations Research* **66**(4), 957–976.

Mitts T (2020) Countering Violent Extremism and Radical Rhetoric. Working Paper, Columbia University. Available from http://tamarmitts.com/research/ (accessed 29 October 2020).

Nouri L, Lorenzo-Dus N and Watkin A-L (2019) *Following the Whack-a-Mole: Britain First's Visual Strategy from Facebook to Gab*. London: Royal United Services Institute.

Pearson E (2018) Online as the new frontline: affect, gender, and ISIS-take-down on social media. *Studies in Conflict & Terrorism* **41**(11), 850–874.

Shehabat A and Mitew T (2018) Black-boxing the black flag: anonymous sharing platforms and ISIS content distribution tactics. *Perspectives on Terrorism* **12**(1), 81–99.

Twitter (2016) An update on our efforts to combat violent extremism. Available from https://blog.twitter.com/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html (accessed 20 October 2020).

Twitter (n.d.) Glorification of violence policy. Available from https://help.twitter.com/en/rules-and-policies/glorification-of-violence (accessed 20 October 2020).

Twitter (n.d.) Our approach to policy development and enforcement philosophy. Available from https://help.twitter.com/en/rules-and-policies/enforcement-philosophy (accessed 20 October 2020).

Twitter (n.d.) Terrorism and violent extremism policy. Available from https://help.twitter.com/en/rules-and-policies/violent-groups (accessed 20 October 2020).

Weimann GJ (2019) Competition and innovation in a hostile environment: how Jabhat Al-Nusra and Islamic State moved to Twitter in 2013–2014. *Studies in Conflict & Terrorism* **42**(1–2), 25–42.

Wright S et al. (2016) Resurgent insurgents: quantitative research into jihadists who get suspended but return on twitter. *Contemporary Voices: St Andrews Journal of International Relations* **7**(2), 1–13.

Cite this article: Hansen TM (2022). The Perils of Estimating Disengagement Effects of Deadly Terrorist Attacks Utilizing Social Media Data. *British Journal of Political Science* **52**, 1482–1489. https://doi.org/10.1017/S000712342100017X