Farkas, J., & Neumayer, C. (2017). 'Stop Fake Hate Profiles on Facebook': Challenges for crowdsourced activism on social media. *First Monday*, *22*(9).

'Stop Fake Hate Profiles on Facebook': Challenges for crowdsourced activism on social media

Johan Farkas* and Christina Neumayer**

* IT University of Copenhagen, direct comments to jjfs@itu.dk.
** IT University of Copenhagen, chne@itu.dk.

Abstract

This research examines how activists mobilise against fake hate profiles on Facebook. Based on six months of participant observation, the article demonstrates how Danish Facebook users organised to combat fictitious Muslim profiles that spurred hatred against ethnic minorities. The article concludes that crowdsourced action by Facebook users is insufficient as a form of sustainable resistance against fake hate profiles. A viable solution would require social media companies such as Facebook to take responsibility in the struggle against fake content used for political manipulation.

Keywords: online activism; crowdsourced activism; fake profiles; hate profiles; Facebook

Introduction

> This is how we shut his page down. We're nearly 1300 members and if we each spend 5 seconds reporting his page, it'll be removed in no time.
>
> (Stop Fake Hate Profiles on Facebook, post by admin, 1 July 2015)

In June 2015, a closed Facebook group named *Stop Fake Hate Profiles on Facebook* (*STOP falske HAD-PROFILER på FACEBOOK*) was created to combat fake profiles spurring anti-Muslim discourses in Denmark. Within 24 hours, the group attracted over 1000 members engaging in several forms of cooperative contestation. Most notably, the group used collective reporting of content for violations of Facebook's community standards (Facebook, 2016).

Stop Fake Hate Profiles on Facebook was created in reaction to several Facebook pages that sparked hundreds of hateful comments and shares from Danish Facebook users in spring 2015. These pages were all constructed around fictitious Muslim identities, claiming to represent a wider Muslim community in Denmark. Their consistent message was that Danish Muslims were conspiring to take over the country, rape Danish (white) women, and kill all non-Muslims (Farkas et al., 2017). Most users who reacted to this hateful content did not realise the identities were fake and expressed aggression as well as xenophobic sentiments in comments. Furthermore, users who contested the pages' authorship in comments were systematically removed and blocked by the anonymous page administrator(s).

Journalists from the Danish public service broadcaster (*Danmarks Radio*) eventually reported on the phenomenon, highlighting that the Facebook pages were fake and likely constructed by far-right activists to smear Muslims (Nielsen, 2015). The latter finding, however, could not be positively confirmed, as Facebook enables page administrators to remain invisible, challenging any

legal action against them. *Stop Fake Hate Profiles on Facebook* thus represented the only systematic attempt to resist and combat the fake Muslim Facebook pages. This occurred through crowdsourced reporting of the pages to Facebook in order to get the company to close them down.

The power of crowdsourced online activism as a form of collective resistance has long been heralded, though particularly in the early days of social media (Benkler, 2006; Jenkins, 2006; Shirky, 2009). Scholars have argued that datafication of personal information and the rise of many-to-many communication enables new forms of political mobilisation based on a politics of numbers (Loader and Mercea, 2011). A core aspect of such political mobilisation is crowdsourced collective action (in the streets and online), often through social media platforms that enable large-scale coordination and organisation (Lotan et al., 2011). There are, however, limitations to this form of action. Given the increasing range of opportunities for engagement in the digital era, it has become common to lament that online participation is no more than feel-good 'slacktivism' (Morozov, 2011), 'clicktivism' (White, 2010), or altogether lacking a collective altruistic component (Bauman, 2001). While this criticism might ring partially true in the case of *Stop Fake Hate Profiles on Facebook*, this article argues that participation and activism organised in the group was conditioned and limited by Facebook's digital architecture. Based on participant-observational findings, the article explores the challenges that *Stop Fake Hate Profiles on Facebook* faced in its struggle. Drawing upon these findings, the article suggests that crowdsourcing user action can only make a marginal contribution to sustainably preventing fake hate profiles on social

media under current conditions. A sustainable solution would require that Facebook takes on greater responsibility as a company and provide more than its currently limited and opaque user support.

The crowdsourcing ideology on social media

Jeff Howe coined the term 'crowdsourcing' in 2006 in a *Wired* article. The idea of crowds acting and creating together was present in early discourses about social media. Tim O'Reilly's (2005) concept of 'Web 2.0' had the "wisdom of the crowds" as a key component. These ideas mainly included crowdsourcing in a business context, focusing on bottom-up creative processes in which companies adopt ideas from crowds, fans, and amateurs. In a discourse analysis of popular press articles concerning crowdsourcing, Brabham (2012, p. 407) concludes that the concept was also promoted as "a potentially powerful tool to spur public participation and transparency in government affairs." Brabham argues, however, that the 'amateur' label in this context delegitimises otherwise-worthy agents by devaluing their roles as participants and citizens in democratic society. Liberatory technological discourses – a powerful part of the corporate identities of social media companies such as Google and Facebook (Turner, 2006) – have thus been adopted in both contemporary business cultures and democratic discourses and processes.

Based on an analysis of the political economy of the digital media industry, Sandoval (2014, p. 252) argues that, rather than being social (as asserted in corporate social responsibility statements), social media companies exploit labour and "are feeding on the commons of society." Social media and other

tech companies co-opt ideas of the radical left, such as participation, decentralisation, spontaneous interaction, and lack of discipline and hierarchy (Žižek, 2009), in concepts such as crowdsourcing. These discourses of empowerment, however, shift the obligation for action on social media to the users. This creates potentials for user action as well as disempowerment since social media companies can disown corporate responsibility for phenomena on their platforms such as fake hate profiles.

Facebook's community standards state that the company strives "to welcome people to an environment that is free from abusive content. To do this, we rely on people like you" (Facebook, 2016). The company's model for handling abusive content is thus built around free user labour. This is economically beneficial for Facebook, as it only employs commercial content moderators to review content reported by cost-free users (Fuchs, 2015; Roberts, 2016). It also enables the company to distance itself, both legally and communicatively, from abusive material on its platform by granting users primary responsibility. This evasion strategy is central to Facebook, which is currently seeking to increase this delegation of responsibility: "The idea is to give everyone in the community options for how they would like to set the content policy for themselves" (Zuckerberg, 2017). As we show in this article, Facebook's user-centred approach is problematic, as the company circumvents responsibility for countering abuse while providing inadequate and opaque tools for user action. This disempowers users and limits the potential for counteracting phenomena such as fake hate profiles.

Activism and social media logics

Many challenges confront activists using corporate social media platforms to counter-act dominant discourses, including racism. Poell and Borra (2011, p. 695) note that for "crowd-sourcing alternative [news] reporting," the content of tweets is framed by mainstream news to produce visibility. Leistert (2015) argues that corporate social media have become algorithmic mass media, using algorithms to censor, normalise, and standardise activist communications. The silencing of critical voices reinforces neoliberal values in which corporate social media platforms are embedded (Couldry, 2010). In order to successfully achieve political goals, activists in social media environments must thus adapt their political strategies to corporate social media logics such as connectivity, popularity, and datafication (van Dijck and Poell, 2013). Through this adaptation, activists risk being co-opted by the social media logics that they attempt to use against the system (Galis and Neumayer, 2016). In other words, instead of empowering activists, "power has partly shifted to the technological mechanisms and algorithmic selections operated by large social media corporations" (Poell and van Dijck, 2015, p. 534).

In his philosophy of technology, Feenberg (2002) focuses on human agency, arguing that technology reinforces prevailing political hierarchies and power relations. Feenberg suggests, however, that technological invention also provides new opportunities for subversive actors to challenge political systems by appropriating new media technologies for their cause. A critical analysis of technology must consequently be "balanced by description of what people actually do in practice" (Mackenzie, 2006, p. 458). This requires us to open the

black box of social media materiality "as active agents shaping the symbolic and organizational processes of social actors" (Milan, 2015, p. 897). In the following, we seek to unpack this black box by analysing the social media practices of *Stop Fake Hate Profiles on Facebook*. In so doing, we explore how the group navigates social media logics in its struggle against fake hate profiles.

A participant-observational inquiry

This article builds upon data collected during six months of participant observation within the closed Facebook group *Stop Fake Hate Profiles on Facebook*. The fieldwork commenced in late June 2015, shortly after the creation of the group, and ended in early January 2016. Levels of activity within the group varied over the course of the six months, with concentrations around occurrences of fake hate profiles. During the research period, *Stop Fake Hate Profiles on Facebook* contested eight fake hate profiles, which attracted a total of over 14,000 comments and 6000 shares from Danish Facebook users. Prior to the group's creation, data from five fake Muslim Facebook pages had already been collected in April and May 2015 (Farkas et al., 2017). When *Stop Fake Hate Profiles on Facebook* was created in response to fake Muslim Facebook pages, it was thus possible to initiate research within the group shortly thereafter. Data on fake hate profiles collected prior to the existence of the group enables a comparative perspective on fictitious profiles before and after initiation of the group's collective contestation.

The dataset of 13 fake hate profiles – eight of which were contested by *Stop Fake Hate Profiles on Facebook* – derives from our qualitative approach. Based

on online participant observations (Hine, 2015), our research objective is to explore and investigate the people, objects, controversies, conflicts, and negotiations surrounding fake hate profiles and the struggle against them. Throughout the six months of research, we continuously observed and participated in the activities of *Stop Fake Hate Profiles on Facebook*. This involved a high degree of engagement. We supported the group's cause and interacted regularly with group members, particularly the group administrator. The primary purpose of these interactions (which can best be described as informal dialogues) was to understand the ways in which the group was organised and operated. Based on these observations, this article seeks to unravel the delicate practices and tactics of *Stop Fake Hate Profiles on Facebook* as well as the challenges facing the group's crowdsourced user action. In future work, quantitative measures could advantageously be included to examine the scale and proliferation of fake hates profiles such as those contested by this group.

Informed consent was secured from members of *Stop Fake Hate Profiles on Facebook* by first contacting the page administrator and receiving permission from him. We thereafter asked the administrator to post a statement in the group for all members to see, disclosing our research agenda and requesting permission to do fieldwork. In this statement, we assured group members that we would protect everyone's anonymity. The group responded positively to our request. User activity within the group was archived through screenshots and 'print page' functionalities to ensure the existence of data in case the group or its content were deleted. In total, we collected 38 posts (all made by the

group administrator) and 943 comments. Subsequent to our fieldwork, all names of group members have been anonymised, and the act of translation from Danish to English renders the content unsearchable.

Stop Fake Hate Profiles on Facebook

As its name suggests, *Stop Fake Hate Profiles on Facebook* was created with the purpose of finding and combatting what it terms 'fake hate profiles' on Facebook. In the group's mission statement, this term encompasses "fake profiles […] groups, or pages created to incite fear and hatred towards specific groups in Danish society" (Post by group administrator, 21 June 2015). The group's objective was to expose and combat such profiles through collective efforts using Facebook's digital architecture and community standards. All group members could invite new users to the group, though they had to be approved by the administrator. The group was explicitly non-partisan, and political discussions were not allowed.

The fake hate profiles combatted by the group were identified on the basis of a number of characteristics, most prominently: use of stolen profile pictures, falsely proclaimed affiliations with existing organisations; deletion of user comments questioning the profiles' authorship, lack of response when contacting the profiles, and rhetoric similar to that of previous profiles identified as fake. The fake hate profiles used fictitious Muslim identities to construct a narrative of Muslims plotting to overrun Danish society, killing and raping ethnic (white) Danes in the process:

> Islam is NOT about peace but subjugation to Allah. Once we get sharia law in Denmark, all you infidel pigs will have to submit to Islam… It's okay to kill, as long as the victims are infidels. Allahu Akhbar!
>
> (Facebook post, Mohammed El-Sayed, 30 June 2015)

> You Danes can laugh at me now, but just wait until we get sharia law in Denmark, then all non-Muslims will be 'removed' (if you know what I mean) ☺. Allahu Akhbar! You should by the way know that I take your money, I have sex with your cheap women, and I make them pregnant.
>
> (Facebook post, Mehmet Dawah Aydemir [1], 9 September 2015)

Most posts from these fake hate profiles contained direct threats to oppress, rape, and kill (non-Muslim) Danes. Others provoked by rejoicing in the September 11 terrorist attacks or stating that all Danes are stupid pigs and dogs. On all profiles, the aggressive statements were presented as originating from young, Danish-speaking Muslims living in Denmark (Farkas et al., 2017). These fictitious identities were all constructed around existing xenophobic stereotypes of Muslims as violent, hypersexual, and alien threats to the Danish welfare state (Hervik, 2011). Stolen images, text, and hyperlinks were thus all deployed to personify these stereotypes as credible and authentic individuals. On each profile, images of Arab-looking people were presented alongside links to existing Muslim organisations, posts about Muslims destroying Denmark from within, and images of burning Danish flags or the flag of ISIS. The fake profiles all claimed to speak on behalf of a wider Muslim community in Denmark, all participating in a large-scale conspiracy: "We Muslims have come

to stay. We haven't come in peace, but to take over your shitty country" (Facebook posts, Zahra Al-Sayed, 2 July 2015). Rhetoric and wording were highly similar across the pages, indicating that their creators were likely connected or identical. As Facebook enables page administrators to remain completely anonymous, however, the actual identities and motives of these authors cannot be established. Consequently, in terms of motive, we can only conclude that all fake hate profiles deliberately sought to provoke and spark anti-Muslim aggression from Danish Facebook users – an agenda in which they largely succeeded.

Across the various fake hate profiles, the violent rhetoric prompted thousands of user comments from Danes believing in the stated authorship and responding with hatred towards the fictitious identities as well as Muslims and immigrants in general:

> Go home to your own country! We didn't ask you to come here to our country"
>
> (User comment, Mehmet Dawah Aydemir [1], 9 September 2015)
>
> What the fuck is this, you fucking pig!!! We help you come to Denmark and this is how you thank us!
>
> (User comment, Mehmet Dawah Aydemir [1], 9 September 2015)

Not all users reacted with aggression towards the fictitious Muslim identities. Numerous users actively tried to dismantle the hatred and warn others that

the profiles were fake. The anonymous page administrators, however, systematically obstructed such attempts, as we show below.

Stop Fake Hate Profiles on Facebook was formed in June 2015 to organise and increase contestation of fake hate profiles. This contestation involved four distinct, concurrent processes: (1) finding and reporting pages, (2) alerting users, (3) alerting journalists and authorities, and (4) speculating about culprits. These processes were continuously negotiated and iteratively developed by group members in order to increase the effectiveness of their efforts. In the following sections, we explore the group's crowdsourced contestation, focusing on the socio-technical tactics deployed in their struggle. Based on this examination, we discuss the limitations and opportunities for crowdsourced user action on social media and their implications for the prevention of fake hate profiles.

(1) Finding and reporting fake hate profiles on Facebook

The first step in the group's contestation was to search for Facebook profiles, groups, or pages using fictitious identities to disseminate hate speech. When members located such content, they would contact the group administrator and get him to share a link within the group alongside a short statement, for example:

> We've received a tip from a member and it seems this profile is fake. The rhetoric is similar to previous profiles, and I will therefore encourage you all to report the page, so we can shut it down.

(Post by group administrator, 5 January 2016).

Users would follow the link and report the profile to Facebook for violations of the company's community standards, which prohibit both fake identities and hate speech (Facebook, 2016). Key to this operation was Facebook's 'report' button, which can be found on all profiles and pages as well as posts, pictures, and videos. When reporting violations to Facebook, group members would subsequently post comments within the group, often simply writing: 'Reported'. Members would thereby continuously make their (otherwise-invisible) actions visible to each other. Some users deliberately reported the same profile for numerous violations (e.g. fake identity, hate speech, harassment) and also reported its individual posts. This was done in the hope that larger quantities of reports would cause Facebook to pay more attention.

Facebook's processing of filed user reports is a highly opaque process (Roberts, 2016), making it difficult to discern how the company operates. Consequently, group members would iteratively exchange personal experiences and hypotheses in an attempt to maximise the effectiveness of their crowdsourced contestation. A recurrent finding by group members was that the quantity of reports played a major role in Facebook's response, although the company officially denies this (Facebook, 2016). Often, when filing reports, group members would initially receive a standard response from Facebook, stating that the reported profile(s) did *not* violate Facebook's community standards. Group members would take screenshots of these replies and post them within the group accompanied with statements of disbelief:

Really!? They've checked the page and won't shut it down… !!!"

(Comment by group member, 22 June 2015)

I can't believe Facebook claims this isn't violating their community standards? A fake profile spreading hate speech, this must be a violation of the rules?

(Comment by group member, 1 July 2015)

After numerous additional reports, Facebook's verdict would often be reversed, causing users to post new screenshots accompanied with statements of celebration: "Together WE ARE STRONG… evil will be conquered in this way! <3" (Comment by group member, 2 July 2015). The pattern of reversed verdicts from Facebook caused members to speculate that the company at first responds algorithmically to filed reports and only later involves actual human staff: "Keep reporting the profiles. Facebook uses robots to go through the complaints. Real humans will only look into it if there are lots of reports" (Comment by group member, 22 June 2015).

As exemplified by the above quotes, group members felt empowered through their collective contestation, as it enabled them to influence (what were otherwise felt to be) unwavering decisions made by Facebook. Simultaneously, however, group members also felt disempowered by Facebook's secrecy and lack of collaboration, with no apparent interest in the group's crowdsourced activism. The group's power seemed to lie solely in its size. Group members and the group administrator would therefore repeatedly emphasise the

importance of *all* members filing as many reports as possible and complaining if Facebook did not respond positively to their request(s):

> We need to keep reporting his [the anonymous administrator's] page. At some point, Facebook will get tired and look at what he's actually written. This is how we shut his page down. We're almost 1300 members, and if we all spend 5 seconds reporting his page, it'll be removed in no time.
>
> (Post by group administrator, 1 July 2015).

The contestation surrounding Facebook's 'report' button shows how *Stop Fake Hate Profiles on Facebook* engaged in tactical socio-technical negotiations, continuously attempting to unlock Facebook's secretive digital architecture and use it strategically to further its cause. These strategies proved largely successful, as contested hate profiles often only existed for a few days before Facebook removed them (see **Table 1**).

< Insert **Table 1 -** Overview of fake hate profiles and their durations of existence. >

(2) Alerting users

On several occasions, hate profiles contested by *Stop Fake Hate Profiles on Facebook* received hundreds or even thousands of comments from Danish Facebook users. Most users accepted the proclaimed Muslim identities and expressed anger, hostility, and even racism:

Fuck you, you fucking monkey

(User comment, Mohammed El-Sayed, 1 July 2015)

Disgusting animal! Get the fuck out of my country… you don't belong here!

(User comment, Mehmet Dawah Aydemir [1], 11 September 2015)

It's because of people like you that more and more people turn racist

(User comments, Mehmet Dawah Aydemir [1], 11 September 2015).

Stop Fake Hate Profiles on Facebook sought to dismantle this hatred towards Muslims and immigrants by alerting users that the profiles were fake and deliberately created to incite aggression. Group members would post comments on the profiles, warning users not to believe in the proclaimed identities and political manipulation. After making such comments, members would notify each other of their actions within the closed group: "Wrote on his page, a warning and a link to this group" (Comment by group member, 1 July 2015).

The anonymous page administrator(s) running the fake hate profiles, however, continuously sabotaged these efforts. On all Facebook profiles and pages, administrators can remove any content without notifying its author and can block any user from making (additional) comments. The administrator(s) of the fake hate profiles systematically used this technological feature to their advantage by deleting all comments and blocking all users who contested their proclaimed authorship. New users encountering the hate profiles would thus be exposed exclusively to user comments affirming the legitimacy of the

sources. Group members and their warnings were continuously deleted and blocked even though they still attempted to alert users:

> You get blocked so fast in there, but at least I got to post 20 times that the page was fake before it was over.
>
> (Comment by group member, 1 July 2015);
>
> I was removed right away!! The person behind must know that we work together and are on his trail!!
>
> (Comment by group member, 1 July 2015).

Due to the systematic moderation performed by the anonymous page administrator(s), the effectiveness of the group's efforts to alert users as to the existence of fake Muslim hate profiles seems to have been limited. The administrators of the hate profiles tactically exploited Facebook's digital architecture to silence any contestation. Nevertheless, a few group members reported that they had in fact first believed in the fake authorship and only later became aware of its deceptive nature due to comments made by group members: "Yesterday, I really thought that someone was being this hostile and I jumped in feet first and cursed him back. I'm glad someone told me it was fake." (Comment by group member, 12 September 2015). This highlights how the struggle between *Stop Fake Hate Profiles on Facebook* and various fake hate profiles fundamentally concerned visibility and awareness. The hate profiles sought to render all contestation invisible, leaving only comments accepting the proclaimed authorship. The group's goal, in contrast, was to make its

contestation as visible as possible to warn users while simultaneously making the pages invisible (through deletion by Facebook).

In several respects, Facebook's digital architecture seems to have supported the hate profiles' agenda by providing unlimited anonymity to their administrator(s) as well as asymmetrical power relations between administrator(s) and users (e.g. through the ability to remove any comment). The counter group's efforts to alert users regarding fake hate profiles might have furthermore had the unforeseen consequence of contributing to their proliferation. Facebook's algorithms continuously evaluate content and 'decide' how far it should spread based on a number of parameters. A central parameter in this process is the number of likes, comments, and shares received by the content in question (Bucher, 2012). Comments posted by members of *Stop Fake Hate Profiles on Facebook* to warn users might thus have indirectly increased the fake hate profiles' reach, potentially deceiving additional Facebook users. Thus, despite the group's collective efforts, fake hate profiles continued to pose a complex challenge. As we discuss below, however, the group also pursued the goal of making their contestation visible outside of Facebook.

(3) Alerting journalists and authorities

Although *Stop Fake Hate Profiles on Facebook* primarily operated within the boundaries of social media, the group also sought to reach out and involve journalists and authorities in their struggle. The group managed to attract the attention of several major Danish media institutions, including the national

tabloid *Ekstra Bladet* (Ryde, 2015), the newspaper *Information* (Skovhus, 2015), and the TV broadcaster *TV2*. These media outlets all reported on the phenomenon of fake hate profiles on Facebook, the latter two interviewing the group's administrator as part of their coverage. The public outreach agenda pursued by the group was primarily undertaken to warn the Danish public about potential democratic dangers posed by fake hate profiles. Simultaneously, it enabled the group to attract more members to participate in their struggle. These efforts largely proved successful. Yet as with the group's efforts to warn users on Facebook, the increased attention to fake hate profiles achieved through mass media could potentially also have led more users to engage with the profiles, indirectly increasing their proliferation on Facebook.

In parallel with the group's efforts to reach journalists, group members also contacted the Danish police and the intelligence agency (*PET*) in order to instigate investigations into the originators of the fake hate profiles. The ephemerality of the contested content, however, presented an obstacle to this agenda. The short time periods in which the fake hate profiles existed meant that archived material was necessary in order to file police reports. The group addressed this challenge by working collectively to compile such material:

> REQUEST: A member is asking for screenshots from the hate profiles that have been shut down since the police want to look into the case… please send them to me in a private message or post them below, so they're visible.
>
> (Post by group administrator, 2 July 2015)

In addition to the challenge of piecing together deleted evidence, the ephemerality and anonymity of fake hate profiles proved problematic. Ephemerality of content meant that authorities could never observe the consequences of fake hate profiles as they unfolded. Furthermore, the complete anonymity of fake hate profile creators, enabled by Facebook's design, meant that no charges could be filed directly against anyone. Doing so would first require a thorough investigation and close contact with Facebook. This caused frustration and feelings of disempowerment for members of *Stop Fake Hate Profiles on Facebook*, as the social media company showed no apparent interest in collaborating with them. The group was thus totally reliant on Danish authorities for conducting investigations, yet the group also experienced a lack of support from authorities in identifying and investigating the creators of fake hate profiles. This caused distress:

> I don't understand why IT specialists in the police can't find their [the administrator's] IP address… These fake profiles are so far out…
>
> (Comment by group member, 13 September 2015)

> I don't think we can achieve anything through police reports.
>
> (Comment by group member, 1 July 2015)

The most powerful means available to members of *Stop Fake Hate Profiles on Facebook* thus continued to be their collective efforts to report fake hate profiles to Facebook and get them deleted. Yet this strategy had severe limitations, as the group could never get to the root of the problem due to Facebook's digital architecture and (apparent) lack of interest in collaboration.

The anonymous creators of fake hate profiles could continuously (re-)create new fictitious identities each time old ones were removed. For members of *Stop Fake Hate Profiles on Facebook*, this caused frustration, even in situations in which Facebook deleted fake hate profiles: "Yes :)! Finally, they [the fake hate profiles] are removed.. but he [the anonymous administrator] will just create a new one :(:("(Comment by group member, 1 July 2015). The lack of collaboration from authorities and Facebook led to investigations by group members to identify the anonymous content creators.

(4) Speculating about culprits

A recurring theme within *Stop Fake Hate Profiles on Facebook* was speculations as to who were behind the fake hate profiles combatted by the group. On Facebook, all page and profile administrators can remain completely anonymous. Even if a page or profile is removed, no information is provided as to who created it. Due to numerous similarities across different fake hate profiles, group members became convinced that several profiles were created by the same administrator(s): "This is exactly the same rhetoric as on the last one. It's the same person who's behind it, fucking coward" (Comment by group member, 24 October 2015); "You just know it's a 20-year-old kid with no friends and Nazi tendencies who's behind the keyboard." (Comment by group member, 1 July 2015). Several members expressed frustration at the ability of the anonymous administrator(s) to continually construct new fake hate profiles and spark aggression, even though Facebook continually deleted the pages. Members also expressed hope that authorities would react and investigate the culprits: "I really hope he [the administrator] will be punished

for the hatred he creates" (Comment by group member, 12 September 2015). Others conducted their own detective work and formulated hypotheses about specific people who could be behind the profiles, including far-right activists. Such speculations were, however, criticised by other members, who argued that *Stop Fake Hate Profiles on Facebook* should not become a vigilante group: "This is exactly what I mean. A suspicion isn't enough to accuse people" (Comment by group member, 12 September 2015). The group never successfully identified any hate content creators, though there were strong indications that several of the fake hate profiles combatted by the group had the same administrator(s).

Crowdsourced social media activism

Having explored the activist practices of *Stop Fake Hate Profiles on Facebook*, we now address how group members navigated Facebook's social media logics in their struggle against fake hate profiles. We furthermore discuss whether the group's crowdsourced activism proved successful in terms of its overarching goal of stopping fake hate profiles. The group's crowdsourced contestation and reporting does indeed seem to have succeed in shortening the lifespans of fake hate profiles (see **Figure 1**). Fake hate profiles studied in our research that existed prior to the group's formation existed significantly longer than did profiles that were created after the group's formation. Facebook, however, maintains in its community standards that quantity of reports does not influence the evaluation of flagged content (Facebook, 2016). Facebook's opaque evaluation procedures mean that different factors could have had an impact.

< Insert **Figure 1** - Lifespans of fake hate profiles before and after the formation of *Stop Fake Hate Profiles on Facebook*. >

Temporality is not the only available parameter for evaluating the groups' struggle against fake hate profiles. Some profiles that existed for the shortest periods of time (and were created after the group's formation) were also those that received the most comments and shares from Danish Facebook users (see **Figure 2**). For example, a profile named Mehmet Dawah Aydemir [1], which existed for less than two days in September 2015, managed to attract 10,426 comments and 4954 shares within this period. Most commenting users did not recognise the page's deceptive character. Its rapid proliferation can partially be explained by Facebook's algorithmic prioritisation of short time decays when assessing the importance of content and 'deciding' its reach (Bucher, 2012). Posts, images, and videos can in other words reach thousands of users within hours if they spark a large number of interactions. This seems to have been the case with Mehmet Dawah Aydemir [1].

< Insert **Figure 2** - Times of existence of fake hate profiles and numbers of comments and shares. >

Although *Stop Fake Hate Profiles on Facebook* successfully reduced the lifespans of fake hate profiles, the profiles could still deceive and provoke thousands of users within this period. Lifespan, then, does not seem to be the best indicator of the group's success in dismantling fake hate profiles. This

raises the question of whether this form of crowdsourced activism represents a viable trajectory for stopping hatred and manipulation on social media. Should it be up to users to stop phenomena such as fake hate profiles on Facebook? Or should social media companies take greater responsibility?

Based on the challenges and limitations facing the crowdsourced activism of *Stop Fake Hate Profiles on Facebook*, we argue that Facebook's delegation to users of responsibility for reporting violations is problematic. Unless social media corporations take greater responsibility in combatting faceless hatred and racism produced by anonymous administrators, no action can go beyond solely closing down such hate profiles. Reaching out to and collaborating with authorities – finding content and identifying its creators – could be part of a solution. Removing unlimited anonymity for page and profile administrators could be another. Such efforts, however, would require Facebook to change its self-image, which is currently that of a tech company and *not* a media company (Seetharaman, 2016). If Facebook is to stop fake hate profiles on its platform, the company must acknowledge that problems associated with fake identities and hatred are partially its responsibility and not only that of users. This argument has recently been raised in debates concerning 'fake news' (Stromer-Galley, 2016). Hopefully, future research can help address these issues by expanding current knowledge on the extensiveness of fake hate profiles on social media as well as related phenomena such as fake news spread by social bots (see Shao et al., 2017). Such efforts could advantageously draw upon both big data analysis and machine learning (see Ferrara et al., 2016).

On the basis of the present article's findings, Facebook's limited response to the phenomenon of fake hate profiles highlights a discrepancy between the company's business model and its corporate ideals. Facebook's business model is built around commodification of user-generated data and user attention, making the quality of content economically secondary to the attention it receives. At the same time, Facebook's corporate identity, which hijacks left-wing ideas of participation and decentralisation (Žižek, 2009), burdens users with responsibility for tackling problems such as fake hate profiles. In this process, the company provides only limited opportunities for users to engage in crowdsourced activism. Even though Facebook refers to its platform as a "global community" (Facebook, 2016), the company seems to prioritise commodification of user attention over the empowerment of users and quality of content.

In the current state of affairs, crowdsourcing of responsibility leaves users with a 'report' button as their only weapon. Even if fake hate profiles only exist for short periods of time, their visibility can still be great due to social media logics that algorithmically privilege content that quickly attracts comments, likes, and shares – even if these reactions express hatred and racism. Although *Stop Fake Hate Profiles on Facebook* continuously performed crowdsourced resistance, Facebook's architecture disempowered the group by limiting its possibilities for action, while fake hate profiles could continue to spur hatred, aggression, hostility, and racism. If this is to change, social media companies must reduce hierarchical power relations, increase the potential for user action, and take responsibility for hatred and racism on their platforms.

About the authors

Johan Farkas is Assistant Lecturer at the IT University of Copenhagen. His research interests include political participation and disguised propaganda in digital media.

Christina Neumayer (PhD, IT University of Copenhagen/MA, University of Salzburg) is Associate Professor of digital media and communication in the Digital Design department at the IT University of Copenhagen. Her research interests include digital media and radical politics, social media and activism, social movements, civic engagement, publics and counterpublics, surveillance and monitoring, and big data and citizenship.

Acknowledgements

The authors would like to thank *Stop Fake Hate Profiles on Facebook*, particularly its administrator, for making this work possible. The authors would also like to thank the anonymous reviewers of this journal.

References

Zygmund Bauman, 2001. *The individualized society*. Cambridge, UK: Polity Press.

Yochai Benkler, 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Vol. 7). New Haven: Yale Press.

Daren C. Brabham, 2011. "The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage". *Information, Communication & Society*, volume 15, number 3, pp. 394-410.

Taina Bucher, 2012. "Want to be on the top? Algorithmic power and the threat of invisibility on Facebook". *New Media & Society*, volume 14, number 7, pp. 1164–1180.

Nick Couldry, 2010. *Why voice matters: Culture and politics after neoliberalism*. London: SAGE Publications.

Facebook, 2016. "Facebook Community Standards" at https://www.facebook.com/communitystandards, accessed 13 January 2016.

Johan Farkas, Jannick Schou, and Christina Neumayer, 2017. "Cloaked Facebook Pages: Exploring Fake Islamist Propaganda in Social Media". *New Media & Society*. https://doi.org/10.1177/1461444817707759

Andrew Feenberg, 2002. *Transforming technology: A critical theory revisited*. Oxford, UK: Oxford University Press.

Emilio Ferrara , Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan, 2016. "Predicting online extremism, content adopters, and interaction reciprocity". In: E. Spiro and YY. Ahn (editors), *Social Informatics.*

SocInfo 2016. Lecture Notes in Computer Science, Springer. https://doi.org/10.1007/978-3-319-47874-6_3

Christian Fuchs, 2015. *Culture and economy in the age of social media*. London: Routledge.

Vasilis Galis and Christina Neumayer, 2016. "Laying claim to social media by activists: a cyber-material détournement". *Social Media+ Society*, volume 2, number 3, pp. 1-14.

Peter Hervik, 2011. *The Annoying Difference: The Emergence of Danish Neonationalism, Neoracism, and Populism in the Post-1989 World*. New York: Berghahn Books.

Christina Hine, 2015. *Ethnography for the Internet: Embedded, Embodied, and Everyday*. London: Bloomsbury.

Jeff Howe, 2006. "The Rise of Crowdsourcing". *Wired*, Issue 14, number 6. https://www.wired.com/2006/06/crowds/, accessed 19 July 2017.

Henry Jenkins, 2006. *Convergence culture: Where old and new media collide*. New York: New York University.

Oliver Leistert, 2015. "The revolution will not be liked: On the systematic constrains of corporate social media platforms for protest". In: L. Dencik and O.

Leistert (editors), *Critical perspectives on social media and protest: Between control and emancipation*. London: Rowman & Littlefield Publishers Inc, pp. 5–52.

Brian D. Loader and Dan Mercea, 2011. "Networking Democracy?". *Information, Communication & Society*, volume 14, number 6, pp. 757–769. http://doi.org/10.1080/1369118X.2011.592648

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, danah boyd, 2011. "The Arab Spring| The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions". *International Journal of Communication*, volume 5, pp. 1375–1405. http://doi.org/1932–8036/2011FEA1375

Adrian Mackenzie, 2006. "Innumerable transmissions: Wi-Fi® from spectacle to movement". *Information, Communication & Society*, volume 9, number 6, pp. 781–802. http://dx.doi.org/10.1080/13691180601064139

Stefania Milan, 2015. "From social movements to cloud protesting: The evolution of collective identity". *Information, Communication & Society*, volume 18, number 8, pp. 887–900. http://dx.doi.org/10.1080/1369118X.2015.1043135

Evgeny Morozov, 2011. *The net delusion: How not to liberate the world*. London: Penguin Books.

Stine Bødker Nielsen, 2015, 19 May. "Vi overtager Danmark: Falske facebook-sider sætter muslimer i dårligt lys". *DR Nyheder*. http://www.dr.dk/Nyheder/Indland/2015/05/18/110828.htm, accessed 19 July 2017.

Tim O'Reilly, 2005. "What is web 2.0" at http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html, , accessed 19 July 2017.

Thomas Poell and Erik Borra, 2011. "Twitter, YouTube, and Flickr as platforms of alternative journalism: The social media account of the 2010 Toronto G20 protests". *Journalism*, volume 13, number 6, pp. 695–713. http://doi.org/10.1177/1464884911431533

Thomas Poell and Jose van Dijck, 2015. "Social Media and Activist Communication," In: C. Atton (editor). *The Routledge Companion to Alternative and Community Media*. London: Routledge, pp. 527-537.

Sarah T. Roberts, 2016. "Commercial Content Moderation: Digital Laborers' Dirty Work," In: S. U. Noble & B. Tynes (editors), *The Intersectional Internet: Race, Sex, Class and Culture Online*. New York: Peter Lang, pp. 147–160. http://doi.org/10.1007/s13398-014-0173-7.2

Ronja Ryde, 2015. 12 September. Ekspert om Mehmet-profil: Forargende Dannebrogs-pisser er formentlig falsk. *Ekstra Bladet*. http://ekstrabladet.dk/nyheder/samfund/ekspert-om-mehmet-profil-forargende-dannebrogs-pisser-er-formentlig-falsk/5729027, accessed 19 July 2017.

Marisol Sandoval, 2014. *From corporate to social media: Critical perspectives on corporate social responsibility in media and communication industries*. Abingdon, UK: Routledge.

Deepa Seetharaman, 2016. 25 October. "Facebook Leaders Call It a Tech Company, Not Media Company". *The Wall Street Journal*. https://www.wsj.com/articles/facebook-leaders-call-it-a-tech-company-not-media-company-1477432140, accessed 19 July 2017.

Clay Shirky, 2009. 11 December. The net advantage. *Prospect Magazine*. http://www.prospectmagazine.co.uk/2009/12/the-net-advantage, accessed 19 July 2017.

Chengcheng Shao, Giovanni L. Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer, 2017. "The spread of fake news by social bots". Eprint arXiv:1707.07592.

Phillip R. Skovhus, 2015. 22 September. "Ægte had fra falske Facebook-profiler". *Information*. http://www.information.dk/546054, accessed 19 July 2017.

Jennifer Stromer-Galley, 2016. 2 December. "Three ways Facebook could reduce fake news without resorting to censorship". *The Conversation*. http://theconversation.com/three-ways-facebook-could-reduce-fake-news-without-resorting-to-censorship-69033, accessed 19 July 2017.

Fred Turner, 2010. *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. Chicago: University of Chicago Press.

José Van Dijck and Thomas Poell, 2013. "Understanding social media logic." *Media and Communication*, volume 1, number 1, pp.2-14. http://www.cogitatiopress.com/mediaandcommunication/article/view/70

Micah White, 2010, 12 August. "Clicktivism is ruining leftist activism". *The Guardian*. http://www.guardian.co.uk/commentisfree/2010/aug/12/clicktivism-ruining-leftist-activism, accessed 19 July 2017.

Slavoj Žižek, 2009. *Violence: Six sideways reflections.* London: Profile Books.

Mark Zuckerberg. 2017, “Building a Global Community” at https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/, accessed 7 august 2017.

Name	Time period	Number of days
Ali El-Yussuf [3]	16/06-22/06 2015	7
Mohammed El-Sayed	30/06-02/07 2015	3
Fatimah El-Sayed	01/07-02/07 2015	2
Zarah Al-Sayed	02/07-02/07 2015	1
Mehmet Dawah Aydemir [1]	09/09-12/09 2015	4
Mehmet Dawah Aydemir [2]	13/09-15/09 2015	2
Ebrahim Said	24/10-25/10 2015	2
Mohammed Al-Dawah	05/01-07/01 2016	3

Table 1: Overview of fake hate profiles and their durations of existence.

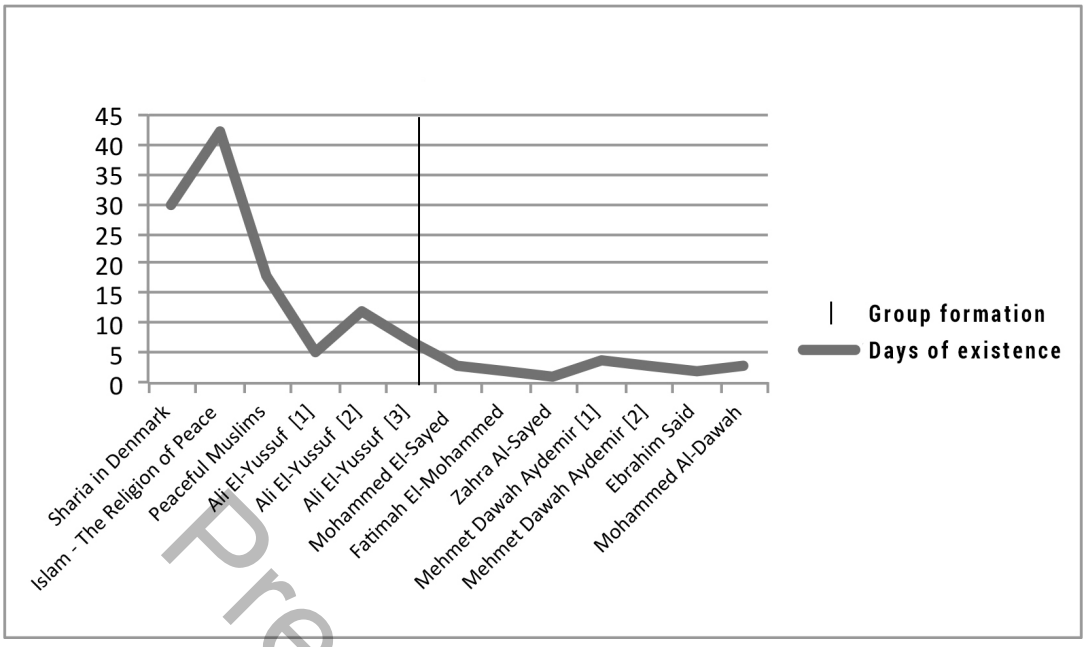

Figure 1: Lifespans of fake hate profiles before and after the formation of *Stop Fake Hate Profiles on Facebook.*

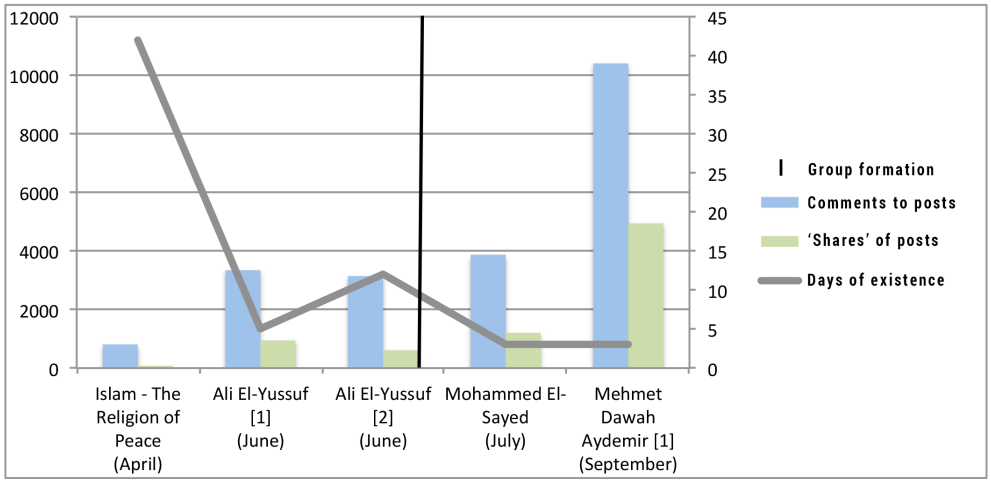

Figure 2: Times of existence of fake hate profiles and numbers of comments and shares.