STS Encounters

Accessibility statement

This is an accessibility statement for the journal: Encounters.

Conformance status

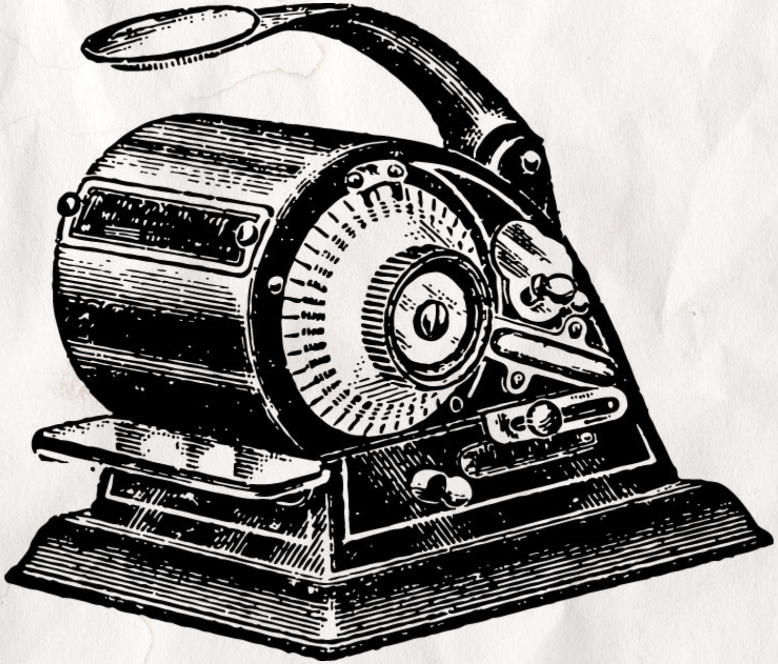
The Web Content Accessibility Guidelines (WCAG) defines requirements for designers and developers to improve accessibility for people with disabilities. It defines three levels of conformance: Level A, Level AA, and Level AAA. This statement is relevant for volume 10, number 5, 2018 through volume 12, number 1, 2021. Encounters is partially conformant with WCAG 2.1 level AA. Partially conformant means that some parts of the content do not fully conform to the accessibility standard. Despite our best efforts to ensure accessibility, footnotes and graphs may not be accessible for screen readers at this point in time.

Feedback

We welcome your feedback on the accessibility of the journal. Please let us know if you encounter accessibility barriers. You can reach us at:

E-mail: imvko@cc.au.dk
Address: STS Center, Helsingforsgade 14, 8200 Aarhus N

ENGAGING THE DATA MOMENT



Special issue
Volume 11• Number 1 • 2020

STS Encounters

SPECIAL ISSUE

Volume 11• Number 1• 2020

Beyond issue publics? Curating a corpus of generic Danish debate in the dying days of the Facebook API

Anders Kristian Munk
TANTLab, Department of Culture and Learning, Aalborg University, Denmark

Asger Gehrt Olesen
TANTLab, Department of Culture and Learning, Aalborg University, Denmark

DASTS is the primary academic association for STS in Denmark. Its purpose is to develop the quality and breadth of STS research within Denmark, while generating and developing national and international collaboration.

Abstract

This article recounts and reflects on our experience of interacting with Facebook's data infrastructure during some pivotal months of change in early 2018. We show how the technical affordances of the Application Programming Interface (API) have critical consequences for the practice of digital controversy mapping and hence argue for the necessity of engaging with changes to these affordances: a consequential data moment for digital STS. The tools that controversy mappers have developed over the past 20 years have focused predominantly on the construction and curation of issue-specific datasets. This is partly justified in the theoretical positions underpinning actor-network theoretical controversy analysis, but it is also technically more convenient than demo- or geographical delimitations. Through the example of mapping the Danish HPV debate, we demonstrate the necessity of being able to challenge the issue-specific approach, and we show how this involves direct engagement with the API. We thus provide an inside perspective from a research practice that relies heavily on data from digital platforms and discuss how the closure of public access to API endpoints severely limits this kind of critical engagement.

Keywords: Digital methods, Issue publics, Controversy mapping, Facebook, API-based research.

We work in a digital methods laboratory where data are perhaps not so much a moment as they are a permanent condition or an ongoing event. Yet, data certainly have their moments and if there ever was one, January 2018 was it. We had spent the previous year trying to get a sense of what the upcoming European General Data Protection Regulation (GDPR) was going to look like and how it would impact our work. In late November 2017 we received news that Facebook was going to change its public data access in a series of radical steps leading up to GDPR taking effect in May 2018. In the middle of all of it we had to complete a project on the Danish HPV vaccine in a way that was not supported by our standard tools for doing Facebook research. As a consequence, we found ourselves experimenting with a changing and in many ways dying data infrastructure. We did not come to the data moment having to find out how to engage with it. On the contrary, we were already deeply engaged in trying to solve some fundamental methodological issues when data became a moment. This is an account from within that situation.

In digital STS, the practice of mapping 'issues' (Marres & Rogers 2005) or 'controversies' (Venturini 2012) online typically entails the construction of datasets in which specific types of digital entities are taken to represent engaged actors. On the open web, such a dataset would comprise websites that take a stance in a debate. For example, in the case of the Narmada Dam network in Uzbekistan, Noortje Marres and Richard Rogers built a dataset around the websites of local and international NGOs that articulated different issues in relation to the construction project (Marres & Rogers 2008): on Twitter, it could be user handles tweeting around certain hashtags; on Facebook, groups or pages dedicated to certain topics. Rather than random samples of activity on specific media platforms, much less in national publics or demographic groups – *within which* issues can subsequently be traced and actors identified – datasets in digital controversy mapping are curated and delimited from their inception as the issues and actors of a debate.

The reason for this is at least twofold. *First*, a theoretical emphasis on the 'generative force' of controversies (Whatmore 2009) and their ability to 'spark' new publics 'into being' (Marres 2005) means that issue publics are understood as emergent communities brought into existence by shared stakes in a problem. They can, therefore, not be captured as a random subset of an already known population or electorate, much less of all users in some geographical area. *Second*, the fact that users on most social media do not natively organize according to socio-economic factors but around shared interests, which is why Rogers calls these media 'post-demographic machines' (2009), makes it necessary to think differently about data curation. It is simply impractical to craft representative samples when the full population is not known, filter on the basis of demographic criteria when these are not available as metadata, or make unambiguous geographical delimitations, although it is possible to study how various digital devices perform differently in a national web (Rogers et al. 2012).

What is both more practical, and seemingly more aligned with the understanding that controversies are generative events, is to let a seed of known actors point the researcher to other actors in the idiosyncratic ways of a specific medium. On Twitter, for example, user handles of known actors can point to other relevant handles through follows, mentions, retweets, or replies. On the open web it can happen through hyperlinks; on Facebook through likes, shares, or comments. In this way, controversy mapping solicits the actors, in their media specific guises, to decide who and what should be included, and the result is a dataset which is an analytical result in its own right. If issue publics are emergent, then the first task for digital controversy analysis is to locate them and describe what has emerged in specific situations. The method for generating digital datasets outlined above can be said to accomplish this task, since we presume that the entities comprising the dataset have pointed each other out as a consequence of the controversy.

The role of tools in the curation of datasets for controversy mapping

The practice of letting actors in a controversy deploy themselves digitally, however, is not only contingent on the idiosyncratic ways in which this can happen on different media, but also on the mechanics of the tools we have at our disposal to do so (Rieder 2020). Early versions of web crawlers like the Issue Crawler (Marres & Rogers 2005) or the Navicrawler (Jacomy et al. 2007), for example, allowed you to input a seed of webpages. From there, the tool would mine all the hyperlinks at a set distance (number of link steps) from those pages and thus collect a corpus of linked web entities. While the Issue Crawler did so automatically, the Navicrawler prompted the researcher to curate which of the discovered pages to include in the corpus manually. This difference in tool design clearly also entails an analytical difference in how actors are allowed to deploy themselves. Furthermore, neither the Navicrawler nor the Issue Crawler allowed you to discriminate between different sections of a website (such as different national versions of Greenpeace like www.greenpeace.org/africa/ or https://www.greenpeace.org/usa/). That feature has now become available in a crawler like Hyphe (Jacomy et al. 2016), which introduces a difference to what can be delimited as an actor. When we say that an issue public has deployed itself on the web through hyperlinks we have therefore not only followed a specific medium but also a specific 're-tooling' of that medium (Elmer 2006).

Other popular data collection tools in digital STS, such as the Twitter Capture and Analysis Toolkit (Bruns et al. 2014) or the now defunct Netvizz for Facebook (Rieder 2013), are of necessity re-tooling their respective media, with similar consequences. The design of a graphical user interface and a functional backend makes it necessary to choose what kind of digital traces can be followed and how. Without such trade-offs, where user-friendliness is gained at the expense of complexity and choice, there would be little point in building tools in the first place. In this paper we address the consequences of these trade-offs in relation to a particular research problem involving the curation of a generic corpus of public Danish Facebook debate, and in response to a particular data moment prompted by the closure of data access in the aftermath of the Cambridge Analytica scandal and the introduction of the European General Data Protection Regulations (GDPR).

In 2017, while Netvizz was still in function as the preferred tool for doing Facebook research in digital STS, we began a collaboration with a group of doctors and anthropologists to map the controversy around the HPV vaccine in Denmark. Facebook was, at the time, considered to be one of the main catalysts for opposition to the vaccine, especially in the wake of a critical documentary that had been aired by one of the national broadcasters in 2015. To build a dataset with Netvizz, you first had to input the ID of the pages or groups of interest. The tool would then retrieve all posts, comments, and reactions from those groups and pages. If you were interested in a particular controversy, such as that surrounding the HPV vaccine, you would therefore have to construct the dataset around groups or pages where you knew that the debate was taking place, since there was no way to simply ask for all HPV related posts and comments independently of the groups or pages where they had been posted. This was, as we shall see below, a direct consequence of the way Facebook made (and to some extent still makes) data retrieval possible through its so-called Application Programming Interface (API), the data architecture on top of which Netvizz and all other Facebook applications are built.

Netvizz provided two methods for constructing an issue-specific set of pages and groups within the framework afforded by the API. The first was a search engine that identified pages or groups containing certain search terms in their names or 'about' sections. The second was the production of a so-called 'page like network' which allowed you to input the ID of a page and retrieve a dataset of other pages liked by that page. As an example, in a recent analysis of wind energy controversies we used these methods to identify 73 groups and pages protesting wind turbine projects in Denmark and retrieve their posts and comments. This dataset was then treated as the issue public emerging around Danish wind turbines on Facebook (Borch et al. 2020).

As became clear rather quickly in the HPV project, however, the controversy was not only taking place in groups and on pages that were set up specifically to discuss the vaccine. This had, of course, always been true of most debates on Facebook, but it became particularly acute and hard to ignore in a case like the Danish HPV debate where a TV documentary was widely assumed to have sparked much of the controversy. The Facebook page of the broadcaster would, for instance, be an obvious place for people to discuss the documentary. But what about the pages of other news networks or media outlets? Or pages dedicated to tangential issues like alternative medicine, parenting, teenagers, diets, or healthy lifestyles where the documentary could have been shared and debated? None (or at least very few) of these pages would be found by following likes from vaccine-related pages or querying their 'about' sections for mentions of vaccines.

As a consequence of the way both Netvizz and the API were re-tooling the medium, there was no way of discovering individual posts about vaccines without first having collected all posts from a set of pages and then querying their text. We therefore decided to take a radically different approach. Rather than following the medium to build an issue-specific dataset, we would attempt to build a dataset of public Danish Facebook conversation and subsequently locate traces of the HPV controversy within it. The Atlas of Danish Facebook Culture, as we began calling the project, ultimately covered 24,272,461 posts and 703,693 events from 68,825 pages located in Denmark. These posts and events had been engaged by 19,851,399 users who had reacted 740,635,475 times with a 'like' or an emoji, made 134,381,871 comments, and declared their interest in an event 87,358,664 times (see Appendix A).

From capturing issue publics to capturing media publics

As we argued in the introduction, attempting to build a corpus that does not trace the contours of some hotly contested topic but claims to reflect a national public conversation as enacted by a platform, sits uneasily with both the theoretical premise of digital controversy mapping and the affordances of online media. Facebook is no exception in this respect. Indeed, you could say that the very idea of Danish Facebook is nonsensical given that users are not restricted by geography in their interactions. As we shall see below, repurposing the API to construct such a dataset has tangible consequences for the way in which a conversation can be said to be 'Danish' or 'public'.

The problem with mapping controversies in topically delimited datasets, however, is that we risk naturalizing any pattern we find as indicative of said controversy. Developments in activity over time in a set of tweets with the same hashtag are easily construed as having something to do with that hashtag (i.e. the dynamics of the controversy) but there is no way of knowing if the changes are actually reflective of some larger trend on the platform. Furthermore, as became clear in our mapping of the HPV debate, the issue-specific datasets that become available through particular re-toolings of a medium like Facebook can be dramatically skewed towards certain types of actors, since those who take the trouble of setting up dedicated groups and pages to discuss vaccines are typically committed to that debate in very particular ways.

A central finding in the project was that Facebook conversations about HPV from the period prior to the airing of the documentary in March 2015 tended to be engaged by two separate groups of users, namely, a group assembling around vaccine-skeptical and another around vaccine-positive posts. As shown in Figure 1 below, these two groups rarely came into contact with each other before the documentary. The networks on the left and right are identical and comprise posts about the HPV vaccine connected by the degree to which they are commented on or reacted to by the same users. On the left, posts from 2012-2013 are highlighted, on the right, posts from 2016-2017. The visual layout is produced with a force vector algorithm, which means that nodes that are visually close can be understood as a cluster of posts engaged by the same group of users. The effect of the documentary, then, was that two isolated groups of users, each either promoting or objecting to the vaccine, became one group of users discussing the vaccine with each other.

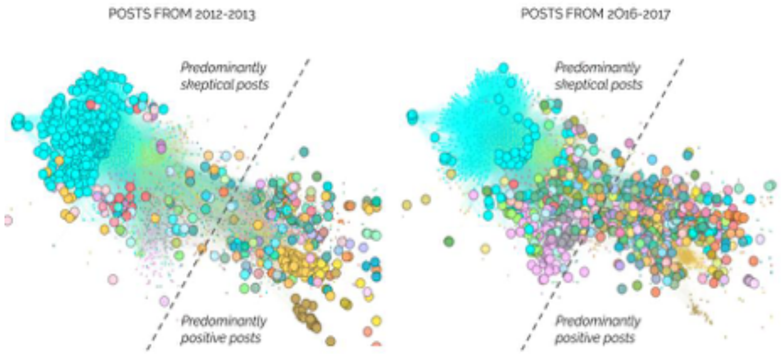


Figure 1: Network of HPV posts connected by user similarity (the degree to which two posts are commented on or reacted to by the same users). Nodes on the left are sized to show posts from 2012-2013; nodes on the right are sized to show posts from 2016-2017. Nodes are colored by the page or group from which the post was harvested.

If the dataset had been topically delimited to groups and pages that were dedicated to vaccine debate, this change in user behavior would have gone unnoticed; so would the scale and perhaps even the existence of the positively inclined user group. As shown in Figure 2 below, the posts that bring the skeptical and the positive user groups into conversation with one another in the years following the documentary are predominantly found on pages that are not dedicated to vaccine issues.

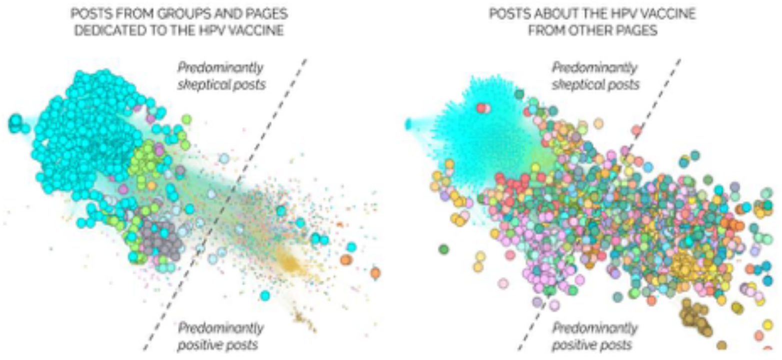

POSTS FROM GROUPS AND PAGES DEDICATED TO THE HPV VACCINE
Predominantly skeptical posts
Predominantly positive posts
POSTS ABOUT THE HPV VACCINE FROM OTHER PAGES
Predominantly skeptical posts
Predominantly positive posts

Figure 2: Network of HPV posts connected by user similarity (the degree to which two posts are commented on or reacted to by the same users). Nodes on the left are sized to show posts from issue-specific groups and pages; nodes on the right are sized to show posts about the issue from other pages. Nodes are colored by the page or group from which the post was harvested.

While it is possible on a medium like Twitter to request a random sample of total activity through the API as a baseline for comparison (Gerlitz & Rieder 2013), this is not an option on Facebook. And even when the possibility exists, any platform-wide sample would be unlikely to capture the patterns that are characteristic of a national discourse in a small country like Denmark. The corpus we collected for the atlas project, however, shows clear annual rhythms of precisely such a national character in the way users post and comment (see Figure 3). The holidays in summer, over the new year, and to some extent Easter, are associated with significant dips in monthly post and comment activity. Christmas is associated with an even more marked spike in comment activity. And if we visualize the daily post activity as a ratio of the monthly activity, it is even possible to reproduce the national calendar of public holidays and weekends for each year (see Figure 4). Some of these public holidays, such as Constitution Day on 5 June, are uniquely Danish phenomena. The same is true for days like 29 November, 2015 when Storm Gorm and its ensuing floods created a national emergency, or 27 November 2016 when a new government was announced following a parliamentary crisis. Such days stand out as unusually active.

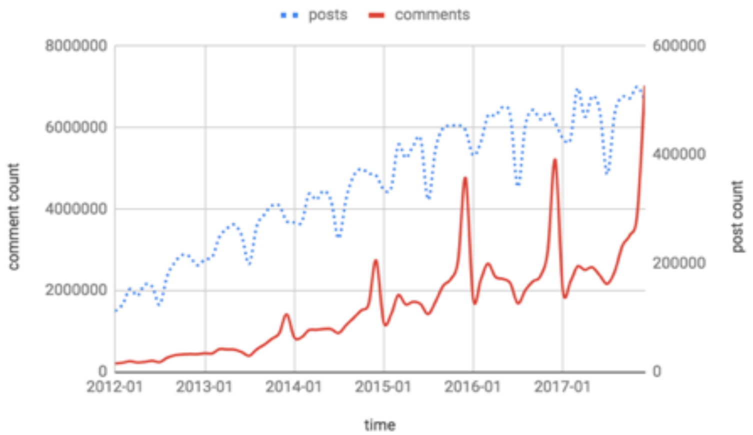


Figure 3: Number of comments (left vertical axis) versus number of posts (right vertical axis) month by month.

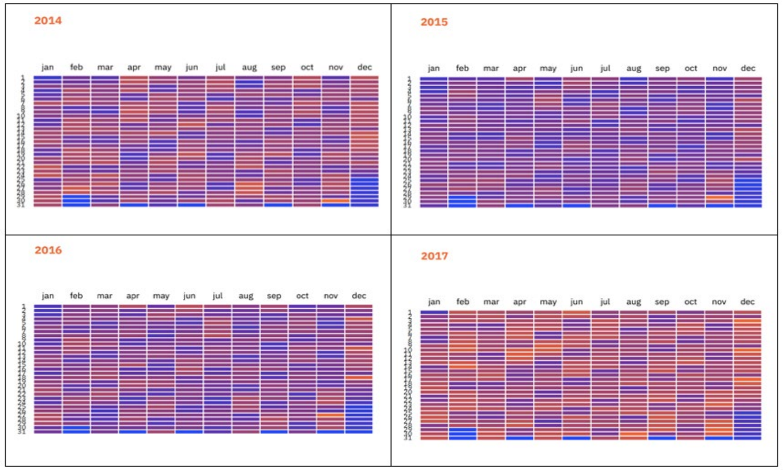


Figure 4: Daily post activity as a ratio of monthly post activity over four years. Blue indicates less activity; orange more activity.

Although social media platforms are post-demographic in the sense that they convene communities of interest rather than representative samples of a population, demography and in this case geography as well, leave tangible imprints on the ways we interact with these platforms. Importantly, however, this does not happen in a correspondence-like fashion where every major event in the 'real' world is straightforwardly reflected in the signal from social media. The national election on 18 June, 2015, for example, is not particularly visible in the post activity. We may not be looking at a specific issue public, but we are certainly not looking at some imprint of 'the general public' either. Media publics, which is what we should assume this to be, co-exist as ongoing results of the shifting ways in which platforms like Facebook, Twitter, or Instagram perform publicity, as shown, for instance, by Andreas Birkbak (2016), Jean Burgess and Ariadna Matamoros-Fernández (2016), or Noortje Marres and David Moats (2015).

Nevertheless, whereas digital STS has devoted considerable attention to such performative media effects in the context of issues (Marres 2015) or controversies (Venturini et al. 2015) – that is, situations where a public is also (and perhaps foremost) brought into being by its stakes in a problem – less consideration has been given to the ongoingness and rhythmicity of media publics themselves. Besides the fact that a comprehensive national mapping of public discourse on a specific medium would be useful for testing claims about political 'echo chambers' (Sunstein 2001) or 'filter bubbles' (Pariser 2011, Hendricks & Hansen 2014), it would also help us situate more case-oriented controversy mapping projects like the analysis of the HPV debate, which was the impetus for building the atlas in the first place. A shift in Facebook activity around a controversial new vaccination program is normally taken as an indication that the issue is heating up or cooling down. The atlas allows you to gauge if such a shift in activity should instead be taken as an expression of wider demographic or media-related rhythms.

One of the clearer indications that the rhythms we observe are closely linked to the intricacies of the medium comes when we track the development in user reactions to content over time (see Figure 5). Up until 2016 we observe a steady year-by-year increase in the number of 'likes' that resembles the increase we see in post and comment activity (Figure 3). In 2016 and 2017, however, the 'like' count slightly decreases, before it picks up again towards the end of 2017. Facebook users will know that in early 2016 the platform introduced a series of alternative reactions to the conventional 'like'. These emoji-based reactions ('love', 'wow', 'haha', 'sad', 'angry', and for a while also 'pride' and 'thankful', the latter for Mother's Day) offered a wider range of options that could be expected to take attention away from the 'like'. What becomes clear in Figure 5 is how the 'love' reaction specifically filled this role.

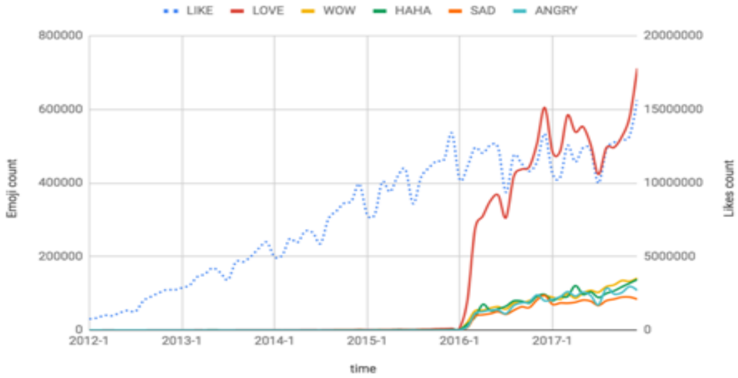


Figure 5: Number of emoji reactions, i.e. 'love', 'wow', 'haha', 'sad' or 'angry' (left vertical axis), versus number of likes (right vertical axis) month by month.

The atlas thus provides a rhythmic backdrop against which we can begin to ground claims about the 'liveness' (Marres & Weltervrede 2013) of issues. The process of constructing the atlas also constituted a good opportunity to consider the grounds on which we can actually talk about and measure features of a particularly Danish discourse or sphere of activity on Facebook. As we will see in the next section, none of this liveliness is available to researchers anymore. We can no longer scrutinize how content or activity gets to be counted as 'public' or 'Danish' in different ways, nor what kind of consequences such constructions have for the analysis. Since none of this was part of the documentation provided in the API reference, the only way to find out was to attempt to produce such a data corpus ourselves, experimentally.

Engaging the data moment through the dying endpoints of an API

On 7 November, 2017 Facebook announced its intention to 'deprecate' (i.e. discontinue) a number of 'endpoints' for its API.[1] These endpoints allowed third parties to retrieve user-related information from public Facebook pages. Three months later, on 30 January, 2018, similar deprecations were announced for endpoints relating to open Facebook groups and events.[2] Since the changes would effectively break existing third-party applications (hence categorized as so-called 'breaking changes' by Facebook itself), they were announced with 90 days prior warning in order to give developers of these applications a chance to come up with new designs and revise their code. In many cases, however, the announced deprecations jeopardized the relationship between the platform and its third-party stakeholders. This was not least the case in digital methods research where apps like Netvizz, which had served the research community as a tried and tested tool for API-based data retrieval and platform scrutiny (e.g. van Es et al. 2014, Munk 2014, Rieder et al. 2015, Lev-On et al. 2015, Poell et al. 2016, Larsson 2016, Ben-David & Matamoros-Fernández 2016, Farkas et al. 2018, Madsen & Munk 2019), were suddenly existentially threatened.[3]

It was not the first time that breaking changes to the API had been announced, nor was it the first time that they would impact the data tools available to digital methods research. Indeed, when the abuse by political consultancies like Cambridge Analytica first came to the platform's attention in 2015, Facebook deprecated all API endpoints that, at the time, gave third parties access to private profile information, such as the friends network of any member of an open group. The reason for these deprecations only came to the public's attention much later in spring 2018, but they were clearly noticed by the research community as they were put into effect (Rider 2015). Generally speaking, these 2015 API changes were seen as a long-needed move to shore up some of the blatant privacy problems in the way Facebook shared data with its third parties. The changes that were announced to take effect in early 2018, however, prompted a much more complex set of questions and concerns.

On the one hand, the user-related information that Facebook wanted to prevent third parties from harvesting could, if treated in sufficient volume, be misused to profile individuals and thus target political content. A like or a comment on a public page may not be private or sensitive information, but it is certainly personally identifiable information, which meant that it would fall within the remit of the upcoming European General Data Protection Regulation (GDPR). It is also the kind of information that machine-learning algorithms can use to guess otherwise undisclosed personal characteristics of the user, such as gender, political orientation, or level of education (e.g. Kristensen et al. 2017). In the age of big data analysis this is in itself an argument in favor of limiting access to data.

On the other hand, we were talking about information that had been deliberately self-published by users, often in an effort to influence a debate, advocate a point, and thus sway public opinion. It had been deposited as posts, comments, and 'likes' on the pages of political parties, companies, and interest organizations as part of a public conversation. Relevant questions about the spread of misinformation, the polarization of online conversation, the role of bots and fake profiles in political debate, or the ability of citizens to organize and mobilize around their matters of concern would become much harder to answer after the API

[1] Changes to the Facebook Graph API announced as v2.11 on 7 November, 2017: https://developers.facebook.com/docs/graph-api/changelog/version2.11#gapi-90 (last accessed 30 April, 2019).

[2] Changes to the Facebook Graph API announced as v2.12 on 30 January, 2018: https://developers.facebook.com/docs/graph-api/changelog/version2.12 (last accessed 30 April, 2019).

[3] From August 2019 Netvizz was no longer publicly available.

changes took effect. From a digital methods perspective, and arguably also from the point of view of a democratic interest in the way social media platforms and the multinational corporations behind them have become part of public life, limited data access posed a serious challenge. Indeed, Facebook would still be selling the ability to tailor campaigns and target users with specific interests or demographic characteristics (this was still the case in November 2019). The announced deprecations would prevent third parties from doing so, but not Facebook itself, the argument being that the platform could then control how content was being targeted and, as they have to some extent started doing, make it transparent who was buying. The capacity to target content, however, would still be available to political operatives and commercial actors alike.

In late 2017, Facebook was by far the preferred social media platform and thereby also the dominant arena for public debate and news dissemination in Denmark and other Western countries. In the wake of Brexit and the Trump election it had become commonplace to question the democratic consequences of this dominance critically, questions that could only be answered if there was a public record documenting which stories were being shared and circulated by whom. There was a schism, then, between Facebook taking back control of its publicly available data and the platform closing itself off from public scrutiny (e.g. Perriam & Birkbak 2019, Venturini & Rogers 2019, Ben-David 2020). The potential clash between the need for democratic society to conduct inquiries on the state of its own public sphere, concerns about privacy and personal information in the age of algorithms, the role and power of multinational media corporations, and attempts to make such corporations accountable through regulation, landed us squarely in the intricacies of what the editors of this special issue call 'the data moment'. The question was: how to engage it?

Engaging the dying endpoints of the Graph API to construct a corpus of the magnitude of the atlas in a relatively short time frame (we had 4 weeks at our disposal from the decision was made until the API changes kicked in) turned out to be a productive empirical situation in the

Deweyan sense that the framing of the problem had to be negotiated in an ongoing process of inquiry (Dewey 1938). It was one that involved, among other things, the API environment, the technical means at our disposal for interacting with this environment, the changing privacy policies and regulations, our own research interests, and the need for robust protocols that would support these interests.

As pointed out by Mirko Tobias Schäfer and Karin van Es in the introduction to their edited volume, *The Datafied Society*, "the translation of the social into data involves a process of abstraction that compels certain compromises to be made as the data are generated, collected, selected and analysed" (Schäfer & Es 2017:13). Negotiating the endpoints of the API to construct a representation of public life on Danish Facebook required a series of such translations, each of which constituted its own potential occasion for learning and critical reflection. We thus took the construction of the atlas as an occasion to move into 'critical proximity' (Latour 2005, Birkbak et al. 2015) with Facebook as a research infrastructure. Coming back to *The Datafied Society*, José van Dijck notes in his foreword that:

> In a society where many aspects of language, discourse and culture have been datafied, it is imperative to scrutinize the conditions and contexts from which they emanate. Researchers from the humanities and social sciences increasingly realize they have to valorise data originating from Web platforms, devices and repositories as significant cultural research objects. Data have become ontological and epistemological objects of research – manifestations of social interaction and cultural production. (Schäfer & Es 2017: 11)

When we are doing research on and with Facebook, such 'conditions and contexts' (that we must imperatively scrutinize) are often features of the API environment. Some of the most commonly used data capture tools for Facebook have been built as applications on top of publicly accessible API endpoints and it has been suggested that we sometimes need to move beyond such tools because they are easily conflated with the platform itself; we can tend to naturalize the data-world offered to us by the tool as if this was the data-world of the platform (Skeggs & Yuill 2016). As we will see in the following sections, the consequences of this conflation become extremely tangible and anything but trivial when you have to decide what to count as *public* and as *Danish* in the construction of a dataset.

Negotiating 'publicness' between the API and the GDPR

What is private and public is not an easy distinction to make online (Birkbak 2013). This is also the case on Facebook where different levels and versions of publicness coexist and intersect. *Pages* are certainly public. They cannot host a closed forum or be kept secret. The administrators of a page can decide not to let visitors author their own posts, but whatever happens on a page remains visible to everyone. This visibility even extends to people who are not on Facebook. Similarly, it is not necessary to like a page to be able to comment on or react to posts on the page. *Groups* can be public as well, although in a slightly different way. If a group is set to 'public', non-members can openly follow the discussion, read the comments, and see who is reacting, although you have to be a member to comment or react yourself. If a group is set to 'closed' or 'secret' you have to be approved as a member by the group's administrators in order to follow the discussion. Members of such groups can thus have a reasonable expectation of privacy, although arguments could be made that if a group has enough members it should no longer be characterized as a private forum. Something similar is the case for personal profiles where users have a reasonable expectation of privacy, except in some cases where users have so many friends and/or have loosened privacy settings to such an extent that their personal profile pages effectively become public forums.

When we tried to determine these questions in conversation with the API, however, the outcome was quite different. Without a special access token (specific permission from a user or an admin of a group), it was impossible to retrieve information from personal profiles or closed/secret groups. From the point of view of the API as it looked in January 2018, there was a clear distinction between the kind of publicness you could argue for open groups and pages and the kind of publicness you could argue for very large closed groups or private profiles with public settings. Prior to the 2018 changes there was arguably some alignment between endpoints that the API allowed you to call without a specially obtained access token, and material that was already publicly available to anyone on or off Facebook. Posts, comments, and reactions from pages and public groups could thus be requested from the API with a generic access token, whereas the same material from a closed group or a personal profile could only be obtained with the express permission (in the form of a temporary access token) from the user or group administrator in question. As we have already discussed, this alignment was only partial since it could be argued in certain circumstances that closed groups are indeed public forums. However, even this partial alignment was temporary. The announced API deprecations would effectively make most of the self-published material from pages and public groups unavailable. At the moment of harvest, then, it was possible to establish a definition of publicness that could be aligned with the API, but part of what made this moment so momentary was the fact that this alignment was not going to last.

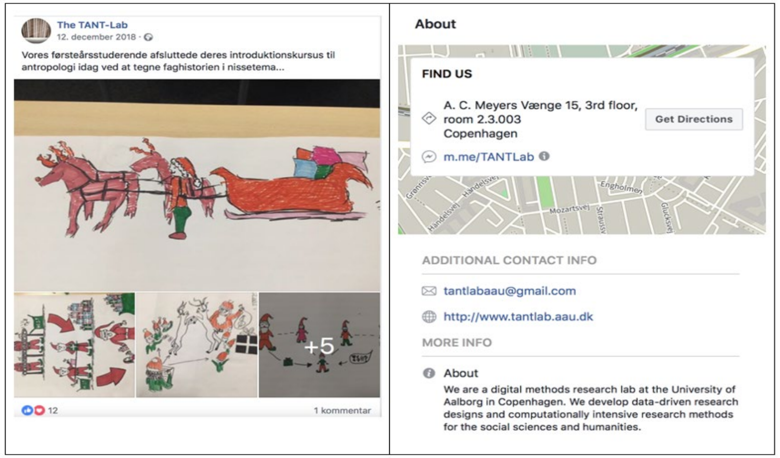
One of the reasons why the API endpoints were being deprecated was very clearly GDPR-related. Or rather, GDPR was not yet phased in but the prospect of it being so was certainly on everybody's mind in 2018, and Facebook used it as part of its justification for the announced changes. Even though the material available through the page endpoints had been self-published, it was nonetheless personally identifiable information. The fact that the information is public makes anonymization almost impossible since a simple Google search for the full text of any comment or a post will immediately reveal the name of its author. Notably, the latter does not even require that the person who is performing the search is logged into Facebook and this is still the case after the deprecation of the endpoints. The situation is somewhat paradoxical: the same publicness that seems to make page interactions on Facebook legitimate objects of research, in the sense that they are already in the public domain, also makes them harder to treat in a GDPR-compliant way. The first question that the construction of the corpus prompted, then, was whether there was a genuine research purpose that justified treatment of the data.

Since it is not possible to obtain informed consent from millions of users, the only way to treat personally identifiable social media data in a GDPR-compliant way is by justifying a research interest. And since it is not possible to use this justification without making the data subjects aware that the treatment is taking place, which is equally unfeasible with millions of users unless the data is already available through a third party, it becomes even more important to argue how the API makes what kind of data available without further consent. The data registration procedures of our university thus became a key factor in determining how we defined 'publicness'.

As we have laid out above, being able to ground the apparent liveness of issues against national or media-related rhythms is certainly of interest to digital STS research. That, and the fact that we could document that the data we were harvesting were both self-published and already being made available to third parties through the API, allowed us to complete the necessary data registration. Doing so, however, also meant that the data for the atlas would have a limited life. Once treated for the purposes laid out in the original registration the corpus must either be deleted or anonymized. As we have already discussed, the latter is near impossible, and we have therefore committed to deleting parts of the corpus when the treatment is over.

Negotiating 'Danishness' with a post-demographic machine

A more difficult problem arose when we had to decide from which pages or groups to harvest data. While the API of the day could help us argue for a version of publicness that was aligned with the availability of data, it was less clear how it could help us define Danishness in a way that could be put to practical use. As described earlier, Netvizz comprised a search module that allowed you to discover pages or groups based on different query terms. This was well adapted to an issue-based approach to building corpora but not of much use for building a national corpus. In this respect, Netvizz seemed to reflect the possibilities of the API, which offered no way of searching for all pages and groups from Denmark. There was, however, another option that was not built into the tool but could be accessed by calling the API directly. Facebook places – locations where users have the ability to check in when they post – could be discovered through a geographical search module. We could thus draw circles around a series of geographical points with a 5km radius, covering the territory of Denmark, and then call the API to retrieve more than 70,000 places where users can check in. As it happens, some of these places are also pages: for example, when a restaurant, a theatre, or a university offers the possibility to check in. In our case we ended up with 2,454 places that we could verify as being pages. Figure 6 shows how a page (in this case our own) contains location information that can be used to check in. These pages became the seed from which to begin the construction of the corpus.



The TANT-Lab
12. december 2018 ·

Vores førsteårsstuderende afsluttede deres introduktionskursus til antropologi idag ved at tegne faghistorien i nissetema...

About

FIND US

A. C. Meyers Vænge 15, 3rd floor,
room 2.3.003
Copenhagen

m.me/TANTLab

ADDITIONAL CONTACT INFO

tantlabaau@gmail.com

http://www.tantlab.aau.dk

MORE INFO

About
We are a digital methods research lab at the University of Aalborg in Copenhagen. We develop data-driven research designs and computationally intensive research methods for the social sciences and humanities.

Figure 6: Example of page post with comments and reactions (left) and an 'about' section with location info (right).

We then decided on a strategy of snowballing. We would start with the seed list of known Danish Facebook pages (pages with a geographical location inside the territory of Denmark) and ask the API which other pages they 'liked'. In the same API call we would specify that we were not only interested in the names and IDs of these 'liked' pages, but also in their location info if available. This allowed us to filter the results returned by the API so that we were left with a new list of pages that all had the country 'Denmark' in their location info and were not already present in the seed list. The process could in principle be repeated until no new pages were found. In practice we proceeded through 15 iterations to find a total of 68,825 pages that we could claim to be public and Danish at the same time. The fact, however, that this combination of page location info and page likes, and the associated API endpoints, became the way in which we could operationalize the construction of the corpus, also meant that public groups could not become part of the corpus. It is not possible for groups on Facebook to like each other and the API offered no other possibility for snowballing more groups from a seed list of groups, except to search through the actual post activity of the group for links to other groups, which we assessed to be unrealistic under the given time constraints. Groups are also not allowed to have a location with country info. It would therefore have been hard to determine which groups were Danish and which were not, based on criteria comparable to those used for pages.

We considered alternatives to the location-based criteria for Danishness, the most obvious one being a linguistic criterion. Implementing such a criterion would first of all have required an additional step for each new page or group we had to evaluate in the snowball where we would perform language recognition on the description of the page. This was also a challenge within our limited time frame and it obviously assumes that Danish pages speak Danish. Pages that are geographically located in Denmark but communicate in English (as is the case for certain restaurants or bars, for instance) would thus not be recognized. As shown in Figure 7 below, 17,537 of the geographically defined Danish pages in the atlas have non-Danish 'about' sections

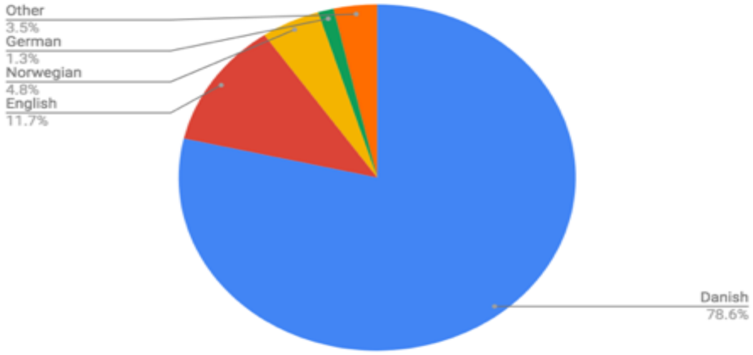


Figure 7: Distribution of detected languages in the 'about' sections from 62,067 Facebook pages geo-located in Denmark. Language detection failed on 6,758 pages which are not included in the diagram.

Even with a geographical criterion there were multiple ways to proceed. Our initial (and eventual) inclination was to go with the self-declared country stated by a page in its location info. There are pages, however, that do not declare a country in this section even though they clearly have an address with a Danish city, postal code, street name, and sometimes even geographical coordinate. As an experiment we scraped a list of 4,092 Danish place names from the geography section of the online version of the Great Danish Encyclopedia.[4] We then asked the API to return pages with a city matching one of the places on the list. This produced an additional 2,454 pages that were not already part of the corpus. None of these pages had 'Denmark' as their country (this is to be expected as they would otherwise likely have been found in our first snowball). Some of the pages state other countries while some of them do not state a country at all. The former come in different categories. A page like *Events Bornholm*, for example, which advertises events on the Baltic Island of Bornholm and was found through the search for city names, erroneously has 'Australia' as its country. Sometimes the Danish place names are ambiguous. This is the case for pages from the city of Greve which happens to be both one of Copenhagen's southern suburbs and a market town in Tuscany (Greve in Chianti). Then there are formerly Danish places, notably in Northern Germany and Southern Sweden, which emerge as an effect of using an encyclopedia with a historic perspective as the ground truth for what counts as places in Denmark!

Conclusion

We have discussed some of the consequences involved in re-tooling a post-demographic machine like Facebook to construct a generic corpus of public Danish debate. The construction of such a behemoth data body involved non-trivial choices about what should count as 'Danish' or 'public'; how such notions could be operationalized within the technical constraints of the Facebook API; the tools available for interacting with it; the mechanisms for storing and accessing data; the time and resource constraints imposed by the reality of API changes (announced and unannounced); how the construction could be justified in relation to GDPR, which in itself turned out to have far reaching technical ramifications; and how to square this with various platform policies. We had to decide, for example, whether 'Danish Facebook' should be defined as the parts of Facebook that speak Danish. Such a translation would exclude non-Danish speaking pages but also necessitate the use of a language detection algorithm, which requires a fairly substantial input of text in order to work and therefore takes time. Even though this extra time is negligible for one page, over the course of thousands of pages it would jeopardize the ability to get data before the closure of the API. In the end, curating a corpus of generic Danish Facebook debate is a matter of negotiating a host of situations that all, in their own ways, embed the complexities of the data moment. Learning to talk to the API through a process of 'explorative programming' (Montfort 2016), that is, scripting API commands and experimenting with the returned results to piece together a strategy in the absence of proper documentation, made it possible to construct a version of what 'public' and 'Danish' could realistically mean in a conversation with the medium.

We have also argued that the construction of such datasets is of critical importance to the practice of controversy mapping in digital STS. We have showed that a strategy of data collection based around issues risks missing important parts of a debate. It can lead us to mistake the rhythms of a medium or a national context for signals in a controversy. As an example, we showed how a conventional approach to capturing the issue public around the HPV debate on Facebook would have left us with a dataset that did not include the pivotal moment when skeptics and supporters of the vaccine began debating each other in the wake of a critical documentary. Although most digital media are more amenable to an issue-oriented data generation strategy, following from the tendency of users to organize in communities of interest rather than along demographic or geographic distinctions, as well as the tendency of both media and data collection tools to support such organization, the case of the Danish Facebook atlas demonstrates the importance of comparing the consequences of this issue orientation

[4] http://denstoredanske.dk/Danmarks_geografi_og_historie/Danmarks_geografi, accessed January 2018.

against other ways of curating datasets. The ongoing closure of API endpoints makes such comparisons increasingly unfeasible and digital STS should therefore consider them as urgent data moments to be explored and exploited as occasions for critical proximity with the media infrastructures on which we rely.

References

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. International Journal of Communication, 10, 1167-1193.

Ben-David, A. (2020). Counter-archiving Facebook. European Journal of Communication, 0267323120922069.

Birkbak, A. (2016). Caring for publics: How media contribute to issue politics (Doctoral dissertation, Aalborg Universitetsforlag).

Birkbak, A. (2013). What is public and private Anyway? A pragmatic take on privacy and Democracy. XRDS: Crossroads, The ACM Magazine for Students, 20(1), 18-21.

Birkbak, A., Petersen, M. K., & Elgaard Jensen, T. (2015). Critical proxim-ity as a methodological move in techno-anthropology. Techné: Research in Philosophy and Technology, 19(2), 266-290.

Borch, K., Munk, A. K., & Dahlgaard, V. (2020). Mapping wind-power controversies on social media: Facebook as a powerful mobilizer of local resistance. Energy Policy, 138, 111223.

Bruns, A., Weller, K., Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. Aslib Journal of Information Management.

Burgess, J., & Matamoros-Fernández, A. (2016). Mapping sociocultural controversies across digital media platforms: One week of# gamer-gate on Twitter, YouTube, and Tumblr. Communication Research and Practice, 2(1), 79-96.

Dewey, J. (1938). The Theory of Inquiry. New York: Holt, Rinehart & Wiston, USA.

Elmer, G. (2006). Re-tooling the network: Parsing the links and codes of the web world. Convergence, 12(1), 9-19.

Farkas, J., Schou, J., & Neumayer, C. (2018). Platformed antagonism: racist discourses on fake Muslim Facebook pages. Critical Discourse Studies, 15(5), 463-480.

Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. M/C Journal, 16(2).

Hendricks, V. F., & Hansen, P. G. (2014). Infostorms: how to take informa-tion punches and save democracy. Springer Science+ Business Media BV.

Jacomy, M., Ghitalla, F., & Diminescu, D. (2007). Méthodologies d'analyse de corpus en sciences humaines à l'aide du Navicrawler. Programme, TIC-Migrations. Paris: Fondation de la Maison des Sciences de l'Homme.

Jacomy, M., Girard, P., Ooghe-Tabanou, B., & Venturini, T. (2016, March). Hyphe, a curation-oriented approach to web crawling for the social sciences. In Tenth International AAAI Conference on Web and Social Media.

Kristensen, J. B., Albrechtsen, T., Dahl-Nielsen, E., Jensen, M., Skovrind, M., & Bornakke, T. (2017). Parsimonious data: How a single Facebook like predicts voting behavior in multiparty systems. PloS one, 12(9).

Larsson, A. O. (2016). Online, all the time? A quantitative assessment of the permanent campaign on Facebook. New media & society, 18(2), 274-292.

Latour, B. (2005). Critical Distance or Critical Proximity. Unpublished manuscript. Available at http://www. bruno-latour. fr/sites/default/files/P-113-HARAWAY. pdf. Accessed March, 31, 2014.

Latour, B. (1999). Pandora's hope: essays on the reality of science studies. Harvard university press.

Lev-On, A., & Steinfeld, N. (2015). Local engagement online: Municipal Facebook pages as hubs of interaction. Government information quarterly, 32(3), 299-307.

Madsen, A. K., & Munk, A. K. (2019). Experiments with a data-public: Moving digital methods into critical proximity with political practice. Big Data & Society, 6(1), 2053951718825357.

Marres, N. (2005). Issues spark a public into being: A key but often forgotten point of the Lippmann-Dewey debate. In B. Latour and P. Weibel (eds.) Making things public: Atmospheres of democracy, 208-217, MIT Press.

Marres, N. (2015). Why map issues? On controversy analysis as a digital method. Science, Technology, & Human Values, 40(5), 655-686.

Marres, N., & Moats, D. (2015). Mapping controversies with social media: The case for symmetry. Social Media+ Society, 1(2), 2056305115604176.

Marres, N., & Rogers, R. (2005). Recipe for Tracing the Fate of Issues and their Publics on the Web.

Marres, N., & Rogers, R. (2008). Subsuming the ground: how local realities of the Fergana Valley, the Narmada Dams and the BTC pipeline are put to use on the Web. Economy and Society, 37(2), 251-281.

Montfort, N. (2016). Exploratory programming for the arts and humanities. MIT Press.

Munk, A. (2014). Mapping wind energy controversies online: introduc-tion to methods and datasets. Available at SSRN 2595287.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Perriam, J., Birkbak, A., & Freeman, A. (2019). Digital methods in a post-API environment. International Journal of Social Research Methodology, 1-14.

Poell, T., Abdulla, R., Rieder, B., Woltering, R., & Zack, L. (2016). Protest leadership in the age of social media. Information, Communication & Society, 19(7), 994-1014.

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. In Proceedings of the 5th annual ACM web science con-ference (pp. 346-355). ACM.

Rieder, B. (2015). the end of Netvizz (?). The Politics of Systems.

Rieder, B. (2020). Engines of order: A mechanology of algorithmic techniques (p. 353). Amsterdam University Press.

Rieder, B., Abdulla, R., Poell, T., Woltering, R., & Zack, L. (2015). Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said". Big Data & Society, 2(2), 2053951715614980.

Rogers, R. (2009). Post-democraphic machines. In A. Dekker, & A. Wolfsberger (Eds.), Walled garden, pp. 29-39. Amsterdam: Virtueel Platform.

Rogers, R., & Ben-David, A. (2008). The Palestinian—Israeli peace process and transnational issue networks: the complicated place of the Israeli NGO. New Media & Society, 10(3), 497-528.

Rogers, R., Weltevrede, E., Niederer, S., & Borra, E. (2012). National Web Studies: Mapping Iran Online. Iran Media Program.

Skeggs, B., & Yuill, S. (2016). The methodology of a multi-model project examining how Facebook infrastructures social relations. Information, Communication & Society, 19(10), 1356-1372.

Schäfer, M. T., & Van Es, K. (2017). The datafied society: Studying culture through data. Amsterdam University Press.

Sunstein, C. R. (2001). Echo chambers: Bush v. Gore, impeachment, and beyond. Princeton, NJ: Princeton University Press.

van Es, K., Van Geenen, D., & Boeschoten, T. (2014). Mediating the Black Pete discussion on Facebook: Slacktivism, flaming wars, and deliberation. First Monday, 19(12).

Venturini, T. (2012). Building on faults: how to represent controversies with digital methods. Public understanding of science, 21(7), 796-812.

Venturini, T., Ricci, D., Mauri, M., Kimbell, L., & Meunier, A. (2015). Designing controversies and their publics. Design Issues, 31(3), 74-87.

Venturini, T., & Rogers, R. (2019). "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. Digital Journalism, 1-9.

Whatmore, S. J. (2009). Mapping knowledge controversies: science, democracy and the redistribution of expertise. Progress in Human Geography, 33(5), 587-598.

Author Bios

Anders Kristian Munk is associate professor and director of the TANTLab, Aalborg University (Copenhagen). He is the co-author of Danish textbook on digital methods and a forthcoming field guide to controversy mapping. Anders holds a D.Phil. from the University of Oxford and has been a senior visiting researcher at the SciencesPo médialab.

Asger Gehrt Olesen is a PhD student in Techno-Anthropology at TANTLab, Aalborg University (Copenhagen). His current research focuses on the use of social media data in urban planning, as well as the consequences of the introduction of social media data in demographic and geospatial methods.

Appendix A

Interaction type	Unique users	Occurrences	Average occurrences per user associated with this interaction type	Average occurrences per user (all users)	Percentage of all reactions	Unique users as a percentage of all users
LIKE	17.390.933	700.124.571	40,3	35,3886	96,7573	87,9045
LOVE	1.592.781	13.547.241	8,5	0,6848	1,8722	8,0509
WOW	634.291	2.559.839	4,0	0,1294	0,3538	3,2061
HAHA	778.776	3.037.078	3,9	0,1535	0,4197	3,9364
SAD	550.409	1.966.947	3,6	0,0994	0,2718	2,7821
ANGRY	500.008	2.334.429	4,7	0,1180	0,3226	2,5273
THANKFUL	9.147	10.899	1,2	0,0006	0,0015	0,0462
PRIDE	4.222	7.184	1,7	0,0004	0,0010	0,0213
POSTS	784.819	2.337.439	3,0	0,1181	N/A	3,9670
COMMENTS	4.108.573	127.223.812	31,0	6,4307	N/A	20,7673
PAGES	18.163.312	192.799.568	10,6	9,7453	N/A	91,8086
EVENTS	4.540.469	87.358.664	19,2	4,4156	N/A	22,9503
ALL INTERACTIONS	19.593.459	940.470.710	48,0	47,5372	N/A	99,0375