

Tweets Are Not Created Equal: Investigating Twitter’s Client Ecosystem

CAROLIN GERLITZ
Siegen University, Germany

BERNHARD RIEDER
University of Amsterdam, The Netherlands

This article offers an investigation into the developer ecosystem of platforms drawing on the specific case of Twitter and explores how third-party clients enable different “ways of being” on Twitter. It suggests that researchers need to consider digital data as traces of distributed accomplishments between platforms, users, interfaces, and developers. The argument follows three main steps: We discuss how Twitter’s bounded openness enables and structures distributed data production through grammatization of action. We then suggest ways to explore and qualify sources by drawing on a weeklong data set of nearly 32 million tweets, retrieved from Twitter’s 1% random sample. We explore how clients show considerable differences in tweet characteristics and degrees of automation, and outline methodological steps to deploy the source variable to further investigate the heterogeneous practices common metrics risk flattening into singular counts. We conclude by returning to the question about the measures of the medium, suggesting how they might be revisited in the context of increasingly distributed platform ecosystems, and how platform data challenge key ideas of digital methods research.

Keywords: platform studies, Twitter, digital methods, developer ecosystem, social media metrics, application programming interfaces, grammatization

Contemporary social media platforms have succeeded in attracting very large numbers of users and, more crucially, have come to serve as communicational infrastructure in broad areas of public and private life. One of the remarkable features of services such as Facebook or Twitter is how they have inserted themselves into the fabric of a diversifying digital media landscape (Bodle, 2011): Not only can we use (subsets of) the functionalities they afford on devices ranging from desktop computers to smartwatches, their widgets populate websites, their messages flicker across television screens, and what people post online increasingly amounts to a news genre in itself. The emerging fields of platform studies (Gawer, 2014; Gillespie, 2010) and, more recently, app studies (Miller & Matviyenko, 2014) have begun to investigate the multifarious entanglements between interfaces, algorithms, data streams, business

Carolin Gerlitz: carolin.gerlitz@uni-siegen.de

Bernhard Rieder: rieder@uva.nl

Date submitted: 2016–06–17

Copyright © 2018 (Carolin Gerlitz and Bernhard Rieder). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at <http://ijoc.org>.

models, stakeholders, use practices, and the wider social and cultural implications these elements connect to. Although the social sciences—and various kinds of marketing research—have shown overwhelming interest in using tweets, posts, images, hashtags, likes, or shares as material to be analyzed, the connection between the designed software and hardware layers making social media possible in the first place and the concrete practices they elicit remains insufficiently understood, both conceptually and empirically. And yet, the analysis of data taken from online platforms generally builds on technologically defined units, such as likes, tweets, and hashtags; The common use of metrics and indicators based on counting these elements suggests that these are indeed similar if not equivalent entities. But are activist tweets about the Ukraine conflict comparable to promotional tweets posted by automated accounts? Is a tweet sent by YouTube whenever a user likes a video the same entity as a tweet the user posted “manually” after watching the clip? When repurposing social media data for analytical ends, we are confronted with the following question: To what extent do “the methods of the medium” (Rogers, 2013, p. 1) imply processes of formalization and production of equivalence that hide important differences, differences that affect the study of online practices—as well as the practices themselves—in significant ways?

In this article, we approach this issue by investigating one side of the software ecosystem that has developed around Twitter, namely, the source of a tweet, that is, the client, service, or script it is sent from. Previous inquiries into Twitter’s software ecosystem have focused mainly on how the company manages relations with developers and third parties (Bucher, 2013). Our approach necessarily includes these platform politics, yet introduces a series of fundamental questions concerning the “utility used to post the Tweet.”¹ Besides engaging in a more conceptual discussion, we explore the client ecosystem empirically, making use of a particularity of data collected from Twitter, namely, the presence of a data field that indicates the source for every tweet sent. This allows us to situate data defined by the medium not only in the context of the platform but also with regard to how client developers build on top of that platform, linking insights from platform studies about the reappropriation of platform data by developers to the empirical work with social media data. In an initial study of a 1% random sample provided by Twitter’s streaming application programming interface (API),² Gerlitz and Rieder (2013) discovered that tweets in fact originate from a diverse set of sources, with only 22.6% sent through the Web interface. The observed variety shows the plurality of what we consider to be different modes of “being on Twitter” and justifies closer inspection. A recent study of Donald Trump’s tweets by Robinson (2016) demonstrates how sources can be of analytical value. When exploring the politician’s tweets, Robinson noted distinct use patterns related to the two different devices used: Tweets sent from the iPhone were issued at different times, included more links, images, and official campaign hashtags, and were more neutral in tone. Tweets sent from Android, however, contained more emotionally charged words and negative sentiment. Taking sources as a starting point led Robinson to the conclusion that Trump’s account may be run by different people, the iPhone tweets being written by staffers and the Android tweets by the politician himself.

¹ This is how Twitter’s Application Programming Interface (API) documentation describes the “source” field. See <https://dev.twitter.com/overview/api/tweets>

² An API is a technical interface that allows one program to request data or functionality from another.

As our focus, here, lies on gaining a better understanding of the client landscape and platform–stakeholder relations, we want to keep an eye on repercussions for researchers working with Twitter data. By putting the spotlight on the production side of tweets, we hope to problematize what digital methods research calls “natively digital objects” (Rogers, 2013, p. 1)—units such as likes, hashtags, tweets, and their associated counts—and ask how we need to understand them as analytical material in increasingly complex and distributed software environments enabled by platforms. Our objective is to situate the data provided by platform media in their wider stakeholder ecosystem and to develop an idea of medium specificity that is not reduced to a single, monolithic understanding of what a medium feature stands for, but considers digital data as outcomes and traces of distributed accomplishments. We do so by exploring how to use the source variable as a means to study this accomplishment and the potentially heterogeneous contexts, practices, and cultures that feed into platform data. If we contemplate Twitter as a stream of messages that flow from specific instances of production to specific instances of engagement, sources play a crucial role in framing and defining this distributed infrastructure. Do different Twitter clients afford different practices and, consequently, do they impact metrics in specific ways? How does the interplay of sources inform the dynamics of the medium, and how can we account for medium specificity in distributed platform infrastructures?

We develop our argument in three main steps: We first discuss how Twitter’s bounded openness enables and structures distributed data production through grammatization of action. We then suggest ways to explore and qualify sources by drawing on a weeklong data set of nearly 32 million tweets, retrieved from Twitter’s 1% random sample. We explore how clients show considerable differences in tweet characteristics and degrees of automation, and we outline methodological steps to deploy the source variable to further investigate the heterogeneous practices common metrics risk flattening into singular counts. The article concludes by returning to the question of the measures of the medium, suggesting how they might be revisited in the context of increasingly distributed platform ecosystems, and how platform data challenge key ideas of digital methods research.

Grammatization

When looking at social media from the perspective of data-driven research, we notice a proliferation of metrics that consist of counting units defined by platforms in technical or functional terms. But, as Driscoll and Walker (2014) point out, “to ascribe a single meaning to any of these behaviors masks the complexities of users’ actual intentions and experiences” (p. 1747). At the same time, however, there is a moment of *commensuration*, a “transformation of different qualities into a common metric” (Espeland & Stevens, 1998, p. 314) that is not simply a question of methodological choice or interpretation on the researcher’s side, but something that happens earlier. One could argue that commensuration, seen as formalization and discretization, on social media does not happen *ex post* but *ex ante*, at the moment users encounter technical interfaces that channel their activities into predefined forms and functions. Commensuration thus affects practice on a primary level in the sense that users inscribe themselves into the spaces of possibility produced and delineated by software. In this section, we approach such delineation as *grammatization* (Agre, 1994) and develop an overview of Twitter’s platform ecosystem to outline the specific role software clients have come to play in what needs to be considered a distributed accomplishment.

Twitter's Software Ecosystem

Social media services have been described as platforms that bring together distinct constituencies who follow distinct interests, from users to advertisers and lawmakers (Gillespie, 2010). Economists (Rochet & Tirole, 2006) have used the term *platforms* in conjunction with so-called multi-sided markets that enable and orchestrate complex economic relationships between different sets of actors, whereas innovation scholars approach platforms as "innovation ecosystems" focusing on engineering networked innovation (Adner & Kapoor, 2010). In this article, we emphasize the infrastructural aspect of the term (Plantin, Lagoze, Edwards, & Sandvig, 2016) and ask, in more detail, how a service such as Twitter provides a set of core functionalities that are not only implemented through different end-user clients but also via connections that insert tweeting into all kinds of configurations and practices.

As Figure 1 indicates, Twitter can be seen as a central database that defines a number of entities (users, tweets, hashtags, etc.), their properties (a tweet has an ID, some text, a post date, etc.), certain relationships between them (users post tweets, hashtags appear in tweets, etc.), and a set of possible actions (writing tweets, following accounts, etc.). All interactions with the database are enabled and governed by middleware providing a set of APIs that define modalities for both input and output. Users interact with the platform through a variety of interfaces that make use of these APIs to read tweets from the backend, write tweets to the backend, or do both. One can think about Twitter as a message delivery infrastructure surrounded by various software devices that "interpret" (Bijker, Hughes, & Pinch, 1987) or "translate" (in the sense of actor-network theory) the basic entities and functions in specific ways. Referencing digital media content, Manovich (2013) argues that "all the new qualities of 'digital media' are not situated 'inside' the media objects . . . they all exist 'outside'—as commands and techniques of media viewers, authoring software, animation, compositing, and editing software" (p. 32). This applies, to a degree, to Twitter: How tweets are produced, displayed, and contextualized depends largely on the interface of the client used.

Within this ecosystem, the possibilities for users and third parties are organized through sets of predefined options that we consider, with Agre (1994), as "grammars of action." Such grammars rely on the capacity of software to structure activity by providing specific "unitary actions" as well as "certain means by which actions might be compounded" (Agre, 1994, p. 746). The grammatization of action into predefined forms such as tweets, retweets, replies, mentions, or hashtags allows platforms to collapse rather than sequence action, grammar, and data capture, inscribing user activities directly into highly formalized units. Grammars thus come with a certain normative force as they demarcate horizons of possible (and impossible) action. On Facebook, users can like, but not dislike, even if they would want to. This is how activities are channeled into technically defined forms that instantly produce equally structured data. However, as we will see, the restrictive aspects of grammatization should not be overemphasized: Both platforms and client software enable bounded reinterpretation, often in ways that "augment" user capacities, for example, through automation.

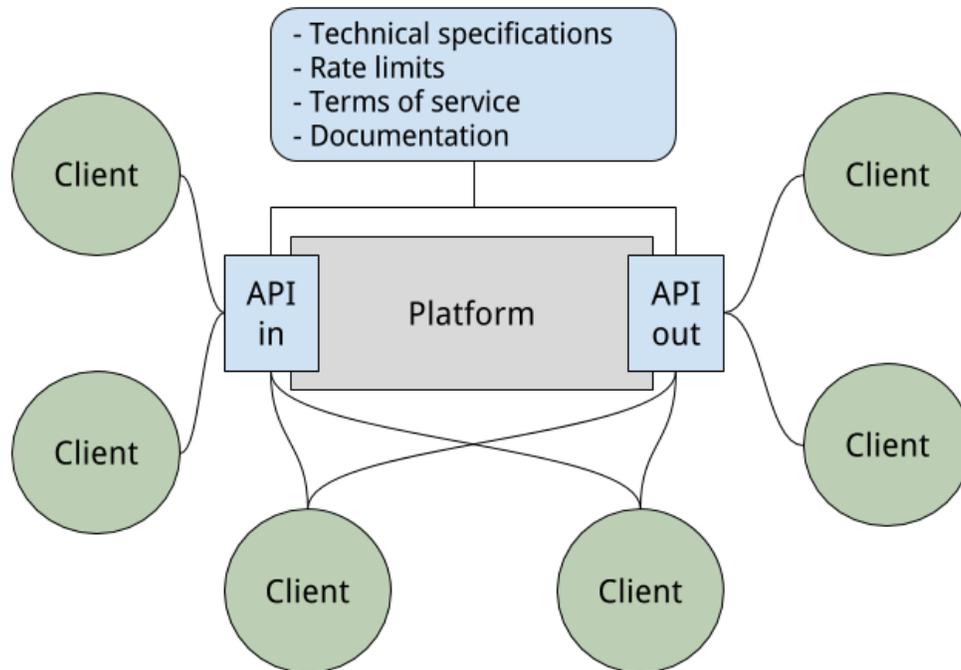


Figure 1. A schematic overview of the Twitter platform as a technical system; it differentiates software clients into three groups, according to whether they are able to send tweets, display tweets, or both. API = Application Programming Interface.

In the context of platform infrastructures, we can distinguish four moments of technical grammatization that echo the schema introduced above: First, a database specifies basic entities and relationships; second, backend and middleware define possible actions that establish the central functionalities of the platform; third, APIs govern the inputs and outputs to and from the system; fourth, sources or clients display and post tweets via end-user interfaces or various forms of automation. Given that the basic forms and functionalities of Twitter have been widely discussed, we limit the following discussion to the API and client levels to focus on grammatization in the context of tweet sources.

API Grammatization

Just a few months after its launch, Twitter released its first API.³ Triggered in large parts by users creating projects with Twitter data, the first Representational State Transfer (REST) API was developed to support structured access for third-party developers wanting to design their own devices to post or retrieve tweets. Today, Twitter offers two key modalities for machine interaction, the REST API and the

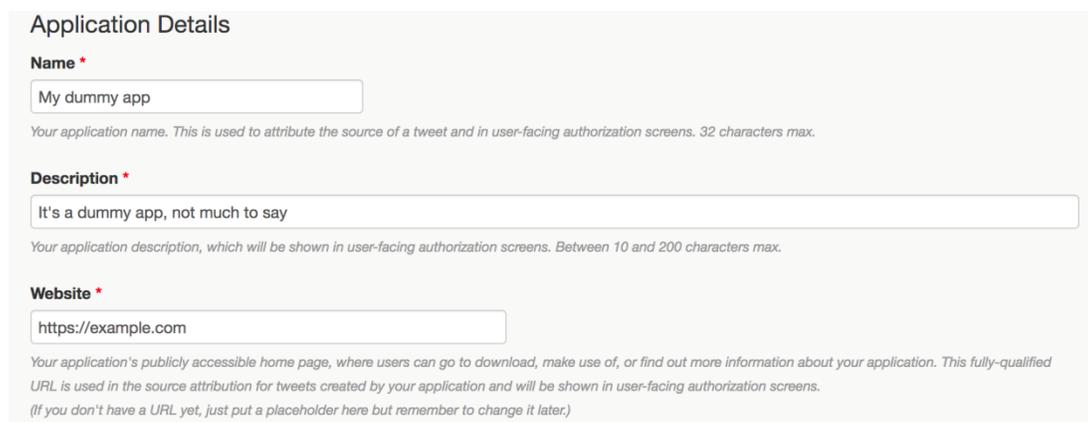
³ <https://blog.twitter.com/2006/introducing-twitter-api>

Streaming API.⁴ Clients connect to the latter via continuous HTTP connections and tweets “stream in” as they are posted, providing near-real-time access to data. The REST API is used for noncontinuous retrieval and, crucially, for sending tweets. To constitute the random sample used for the empirical part of this article, we connected to the statuses/sample end point⁵ provided by the Streaming API and collected the incoming data for a week. However, because the focus of our investigation concerns the sources from which tweets are sent, the following discussion concentrates on the REST API, which is the only means for clients to post tweets.

To interact with Twitter’s APIs, developers need to register their applications online, providing an application name, an application description, and a website (see Figure 2). When a registered app sends a tweet, the values showing up in the metadata of that tweet are presented as an HTML link composed from the given app name and the website. A tweet sent from the official iPhone client, for example, will identify as:

```
<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
```

This signup process implies fairly open grammatization, as Twitter neither verifies the provided URL nor obliges developers to follow restrictive naming conventions or to submit to a review process. This keeps the system open for developers who do not have clear-cut app ideas or associated websites. Twitter facilitates this freedom to experiment and may itself only have relatively little information about all third parties involved. As a consequence, we indeed found 75,096 distinct values for sources in our sample, some of them clearly using automatically generated names.⁶



The image shows a screenshot of the Twitter application registration interface. It is titled "Application Details" and contains three main sections, each with a required field (indicated by an asterisk):

- Name ***: A text input field containing "My dummy app". Below it is a note: "Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max."
- Description ***: A text input field containing "It's a dummy app, not much to say". Below it is a note: "Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max."
- Website ***: A text input field containing "https://example.com". Below it is a note: "Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)"

Figure 2. Twitter’s application registration interface.

⁴ For Twitter’s API history, see Burgess and Puschmann (2013).

⁵ <https://dev.twitter.com/streaming/reference/get/statuses/sample>

⁶ When examining source names, we stumbled, for example, over random variations of the letters *a*, *d*, *g*, and *n*. This is a testament to the open character of the API on this level and raises the question for researchers about how sources that are not immediately comprehensible can be identified.

Actual input to the API requires that users authenticate through the app, which means that each connection to the platform is signed by both an app and a user. All machine interaction is further prestructured through specific developer-facing grammars, divided into search-focused GET grammars and activity-focused POST grammars. To post a tweet, for example, a program needs to connect to the statuses/update end point⁷ and provide at least a not-empty status parameter (i.e., some text to tweet). These technical forms are further constrained by various rate limits for both users and apps. In the case of the REST API, these limits are organized in 15-minute intervals: Search is limited to 180 queries per interval, list, follower, or following retrieval to 15 queries. Restrictions for posting are handled on a per account basis, which means that users are limited to 2,400 tweets across all of the clients they use on a given day.⁸

Beyond API grammatization and rate limiting, client activity is regulated through the general terms of service⁹ and Twitter rules,¹⁰ which refer to content, account usage, copyright, and abuse, as well as via additional terms of use for developers,¹¹ which emphasize rate limits and data use. Twitter is continuously negotiating relations with third-party stakeholders. Initially, Twitter encouraged diverse reinterpretations and repurposing of its data, but since 2012, the platform gradually started to exercise control over (a) how content can be shown on external sources, (b) who can access (Puschmann & Burgess, 2014) and valorize data, and (c) how users can engage with grammars. To “to deliver a consistent Twitter experience” (Sippey, 2012, [blog] n.p.), the platform renamed its “display guidelines” to “display requirements”¹² and curtailed the development of clients by increasingly requiring them to maintain the forms and functions of the platform. If developers do not adhere, access may be revoked. Furthermore, preinstalled clients have to obtain Twitter certification, and any source that requests more than 100,000 user tokens must seek special permission.

As Figure 3 shows, Twitter has been monitoring its developer ecosystem closely. In 2012, the platform identified four strategic areas for developers and decided to support all quadrants except the upper right one, namely, building clients “that mimic or reproduce the mainstream Twitter consumer client experience” (Sippey, 2012, [blog] n.p.). Twitter also stopped showing sources in most of its own interfaces,¹³ whereas selected clients such as Hootsuite continue to do so, opening up questions of governing innovation by design (Gawer, 2014). This politics of (in)visibility is part of Twitter’s attempt to regulate its developer ecology (Paßmann & Gerlitz, 2014) and aims to redirect user activities back to its own interface, limiting the valuation potentials for other stakeholders. Burgess (2016) recently demonstrated the effects of these policy changes: Although 2010–11 marked the heydays of client development, a significant slowdown and muting of clients followed.

⁷ <https://dev.twitter.com/rest/reference/post/statuses/update>

⁸ <https://support.twitter.com/articles/15364>

⁹ <https://twitter.com/tos>

¹⁰ <https://support.twitter.com/articles/18311>

¹¹ <https://dev.twitter.com/overview/terms>

¹² <https://dev.twitter.com/overview/terms/display-requirements>

¹³ <http://thenextweb.com/twitter/2012/08/28/twitter-longer-displays-client-tweet-posted-web-emphasizing-first-party-reading-experience/>

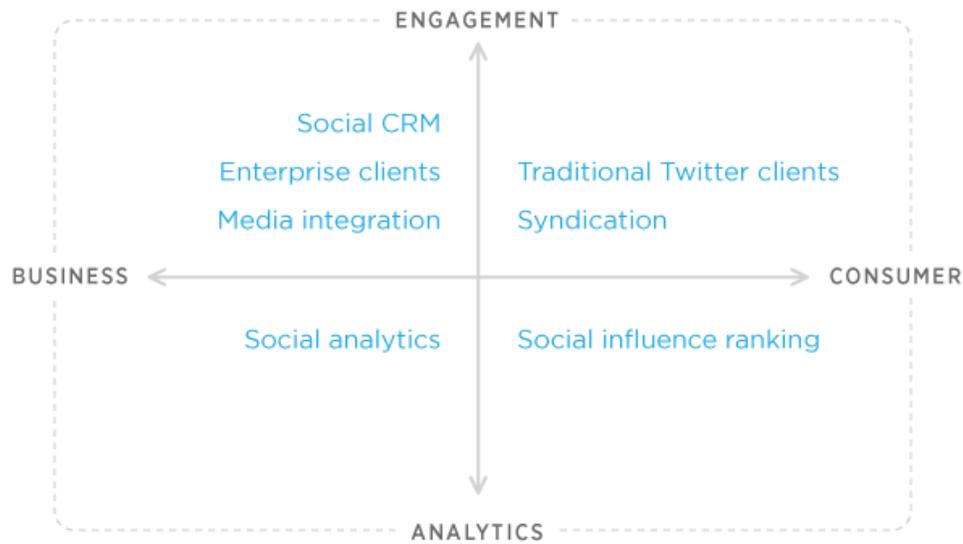


Figure 3. Twitter's vision of its application registration interface ecosystem.
CRM = Customer Relationship Management.¹⁴

Beyond API regulations, Twitter controls its ecosystem through legal means. Again in 2012, Twitter sued five clients for breaking its rules by enabling spam and automation, which eventually led to the shutdown of TweetAdder in 2015.¹⁵ However, “interpretative flexibility” (Bijker et al., 1987) remains, in particular, when it comes to software that does not directly replicate Twitter’s canonical client functionalities. As van Dijck (2011) argues, platforms keep their grammars deliberately open to enable ever more users and third parties to get involved and gain a stake in the platform. And appropriation by users, in particular, still reveals important margins of freedom. To give an example, Twitter’s *favorite* (now *like*) button has been interpreted variably as an internal bookmarking feature or as a gesture of social appreciation (Paßmann & Gerlitz, 2014). Both perspectives were accompanied by respective third-party services that allowed users to further act on their specific interpretation by enabling either the extraction of “favorited” tweets into bookmarking software or the reaggregation of favorites into popularity rankings, as in the case of Favstar—even though the new *like* button is considered to partly disambiguate this freedom. What data stand for is not just subject to platform grammatization, but involves the entire sociotechnical ecosystem of the platform. The role sources play in these assemblages is the focus of the following section.

¹⁴ <https://blog.twitter.com/2012/changes-coming-in-version-11-of-the-twitter-api>

¹⁵ <http://mashable.com/2012/04/05/twitter-sues-spammers/>, see lawsuit at <http://www.scribd.com/doc/88195333/Twitter-Spam-Filing>

Qualifying Sources

To gain a better understanding of Twitter's software ecosystem, we turn to the already mentioned statuses/sample API end point, which offers a stream of randomly selected tweets covering 1% of the total Twitter volume at any point in time, to explore how the source variable can shed light on data production in platform ecosystems as a distributed accomplishment that brings together use practices, client software, and the corporate platform itself. As an empirical example, we draw on a week-long sample retrieved with the Digital Methods Initiative Twitter Capture and Analysis Toolset (Borra & Rieder, 2014) between June 14 and 20, 2014. The data set comprises 31,707,162 tweets sent from 14,313,384 distinct accounts. Although doubts about representativeness remain, a study comparing the 1% sample to the full stream concluded that "the tweets that come through the Sample API are a representative sample of the true activity on Twitter" (Morstatter, Pfeffer, & Liu, 2014, p. 6). Working with randomly selected data allowed us to move beyond the more common keyword-, location-, or user-based samples to engage with a cross-section of tweet practices (Gerlitz & Rieder, 2013). In what follows, we explore and outline methodological tactics to retrace how different sources inform the creation of content, and how this may introduce a platform perspective on social media data and their medium specificity.

To establish an overview, we first identified the 20 most common sources (see Figure 4) out of a total 75,096 unique entries; 86.51% of tweets in our sample were sent from sources other than the Web client, with the official Twitter apps for iPhone (30.27%) and Android (22.01%) coming in first, a finding that still applies to Twitter in 2017. The most used software is thus owned and designed by Twitter itself, which should not come as a surprise given the restrictions discussed. Other top sources are clients that provide users with enhanced software interfaces to one or more social media platforms. Among these are TweetDeck (Twitter only, owned by the social media company itself) and Hootsuite (multiplatform), both of which address advanced or professional users. We also find the official Twitter apps for other devices among the top 20, including Windows Phone, iPad, or Android tablets. Many of these apps have been developed by third-party companies later acquired by Twitter. The list also contains a series of social media platforms, including Facebook and Instagram, which enable cross-platform syndication of content. Finally, the overview reveals what we call Twitter automatons, applications that organize content production and engagement around rule-based triggers. *Twitbot.net*, *TweetAdder*, and *TweetCaster* are specific to Twitter, but more generic Web automation services such as *dlvr.it* and *IFTTT* (If This Then That) appear as well. These services point toward two important observations: First, Twitter content is not necessarily produced within or for Twitter itself; second, their presence in the top 20 highlights the significance of forms of content creation that hover at various stages between manual and automated.

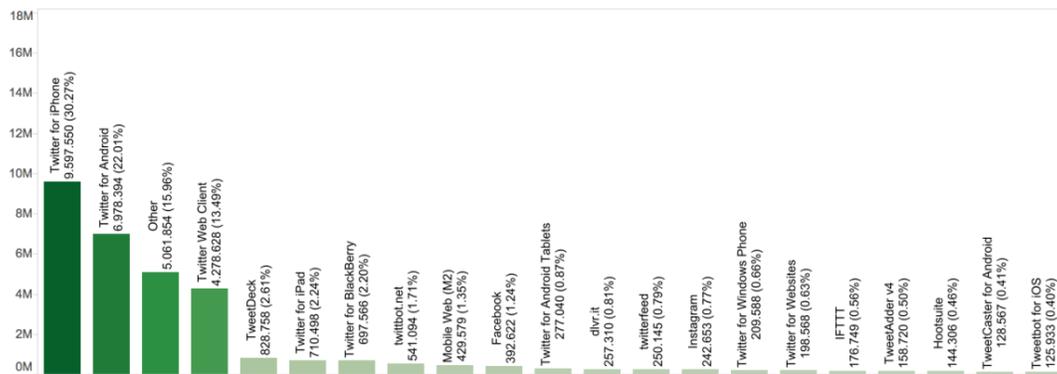


Figure 4. An overview of the top sources appearing in our data set.

Sources and Their Features

The diversity of sources raises the question of how they enact Twitter's basic grammars. To gain a sense of the ways "being on Twitter" is implied by specific clients, it is worthwhile to investigate their interfaces, functionality, and objectives. Take the case of Instagram, for example, or, more precisely, of the possibility to cross-syndicate content from Instagram to Twitter. The feature is built into most Instagram clients; when preparing a post, the final screen offers a variety of sharing options for Facebook, Tumblr, Flickr, and Twitter. Cross-posting cannot be made a default and has to be selected for each post individually. Just like Twitter, Instagram offers hashtags and @username handles as part of its grammaticalized actions, but especially the former takes on a fairly different function, given that hashtags and location features are the only ways to search for content on Instagram. Users often employ long chains of hashtags to make their posts searchable. Instagram allows image captions up to 2,200 characters,¹⁶ including up to 30 hashtags, and this text is cut off automatically to fit Twitter's 140-character limit when cross-posting. Cross-syndication allows for hashtags and handles to be transposed from one platform to the other in a haphazard way that complicates interpretation if the source of a tweet is not taken into account.

Further aspects of both automation and cross-syndication come into play when examining IFTTT, a service for software and Web automation that is often used for content production, sharing, and archiving. Automation on IFTTT is based on so-called *recipes*, short chains of actions that perform a specified task. "DO recipes" enable quick execution of actions, and "IF recipes" are based on triggers that run in the background. In June 2016, IFTTT offered 4,100 (user-generated) recipes involving Twitter focusing mainly on cross-platform syndication, such as sending a tweet whenever a user posts a new blog entry, writes a Facebook status update, or posts to Instagram. Many recipes also enable automated or scheduled content generation, ranging from New Year's or happy weekend wishes to automatically posted scores of sports games. By breaking down automation into channels (platforms) and actions, IFTTT boosts the exploration of interpretive flexibilities, given that recipes make it easy to move data into new contexts

¹⁶ <http://www.iennstrends.com/limits-on-instagram/>

and to enhance manual tweeting with various degrees of automation. Similar to Instagram's cross-syndication, IFTTT invites its users to be on Twitter in ways that differ substantially from the common gesture of writing a message into a form field.

Hootsuite, a commercial social media client, represents yet another interpretation of Twitter's grammars. It offers a social media management dashboard for multiple platforms, accounts, and users, addressing professionals and teams working in social media management, marketing, and customer service. The client provides various features for filtering and aggregating tweets, and adds grammars not available on other clients: Users can alter privacy settings per tweet and target audiences by country. Tweets can be assigned to team members and bulk uploads; scheduling and text templates support long-term strategic tweeting. These features are complemented with additional analytics, such as influence or Klout scores for users, advanced follower/followee ratios, and customized reports, providing calculative capacities for strategic follower building and engagement with influential users. Hootsuite offers a way of being on Twitter that is focused on accountable and effective team tweeting via orchestrated interaction, using the available grammars for strategic promotion or customer management.

These three cases illustrate how different renditions or enactments of Twitter grammars provide functionalities for engaging the platform that differ substantially from the outline of a basic messaging client. They draw attention to the distribution and blending of grammars between platforms and sources, and show how automation is not a binary opposition between "human" and "bot," but a fine-grained and nuanced continuum that is organized around the automation of specific functions. Although custom-made software can broaden the spectrum even further, established clients already provide capabilities that enhance users' capacities in significant ways.

Source Profiling

The differences in how sources orchestrate tweeting bring us back to the question of whether and how actual outputs are affected. What traces do sources leave in the metrics provided by the medium? A first step to approach this question is to pattern how users engage with Twitter-specific grammars across key sources by identifying the percentages of tweets sent from these devices containing links, hashtags, mentions, retweets, and replies (see Table 1). The results for the three largest sources—iPhone, Android, and Web—largely mirror those of the overall data set. A key difference in the case of mobile clients is the lower use of links in tweets at about 5%, and also hashtags uptake is lower. Interestingly, Twitter for Web clients engages slightly less on mentions and retweets while drawing somewhat more on hashtags and links, suggesting that mobile clients enable more user interaction. Most other sources show more distinctive patterns. TweetDeck, for instance, stands out through a high uptake of both hashtags (31.8%) and retweets (38.1%) while featuring fewer replies than the average. This suggests a nuanced shift in practices toward less interactive modes of sharing and broadcasting content through links and retweets. The profile of the tweet automator TweetAdder diverts more drastically from the average practices in the data set: 64% of all TweetAdder tweets contain links, 54% contain hashtags, and 43% constitute retweets. Replies barely matter at less than 1%. This profile suggests that TweetAdder is not used for social interaction, but for broadcasting Twitter-external contents. Together with the lack of direct interaction, this profile may be an indicator for mainly automated activity. An even less interactive pattern

of activity can be found in tweets sent from the automator dlvr.it, whose tweets mainly feature links and hashtags. Tweets sent from the source Tribez, a series of Facebook games, constitute a specific form of automation, as they are triggered by in-game achievements. All Tribez tweets contain URLs redirecting back to the game and feature hashtags referring to the specific title. Thirty-nine percent of all tweets from Instagram feature hashtags, as these are also part of Instagram's grammars, although their usage varies, as we discussed earlier.

Table 1. Percentage of Tweets Containing Platform Elements.

Source	Hashtags	Mentions	Links	Retweets	Replies
Overall	15.9	58.1	16.8	33.0	18.3
Twitter for iPhone	11.0	67.1	4.9	37.4	25.3
Twitter for Android	12.1	64.7	5.0	37.6	19.6
Twitter Web client	17.2	59.3	10.3	30.0	21.4
TweetDeck	31.8	64.7	15.1	38.1	12.8
Twitter for BlackBerry	13.3	63.0	3.8	30.5	15.6
Facebook	8.5	0.8	82.9	0.0	0.0
dlvr.it	27.6	6.7	94.3	4.5	0.0
Instagram	39.1	6.7	100.0	0.0	0.0
IFTTT	40.8	10.5	70.3	0.0	0.0
TweetAdder v4	54.1	49.3	64.6	43.3	0.0
Hootsuite	35.1	42.9	61.5	21.3	5.2
Tribez	100.0	0.0	100.0	0.0	0.0

This exercise provided us with indications of grammar use, but we were also interested in source-specific content and used the top hashtags for selected sources to approximate their topical focus (see Figure 5). There are significant limitations to hashtag research, for instance, their relatively low importance within samples, but they are still considered relevant indicators of content dynamics (Bruns & Stieglitz, 2013; Gerlitz & Rieder, 2013). The top hashtags tweeted from the iPhone again largely—but not entirely—align with the overall data set, containing multiple celebrity- and event-based hashtags, such as #MilenaforMMVA or #SelenaforMMVA referring to the MTV Video Awards, and #WorldCup or #WorldCup2014 relating to the Soccer World Cup. Hashtags coming from TweetAdder show a very different profile, as they predominantly focus on *porn*, *gossip*, and Arabic news. Tweets from Instagram, however, revolve around the topics *love*, *selfie*, *me*, *summer*, *happy*, and *cute*, not to forget *worldcup*. Instagram thus emerges as an island of bliss.

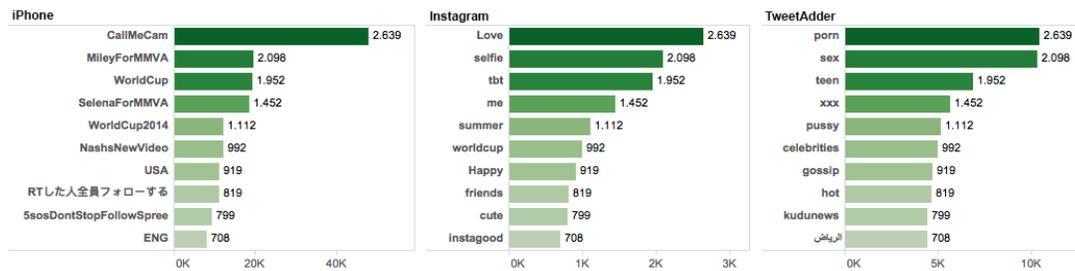


Figure 5. Top hashtags per source.

Given that there are more than 75,000 distinct sources in our sample, a major question is how to move from exploring individual cases to classifying clients in general. Looking for a flexible yet scalable method for classification, we turned to clustering based on similarities between “source profiles,” that is, the uptake of platform grammars per client. Treating sources as vectors and the variables shown in Table 1 as their dimensions, we calculated cosine similarities between clients and used the Gephi graph analysis toolkit (Bastian, Heymann, & Jacomy, 2009) to visualize the similarity network (see Figure 6).

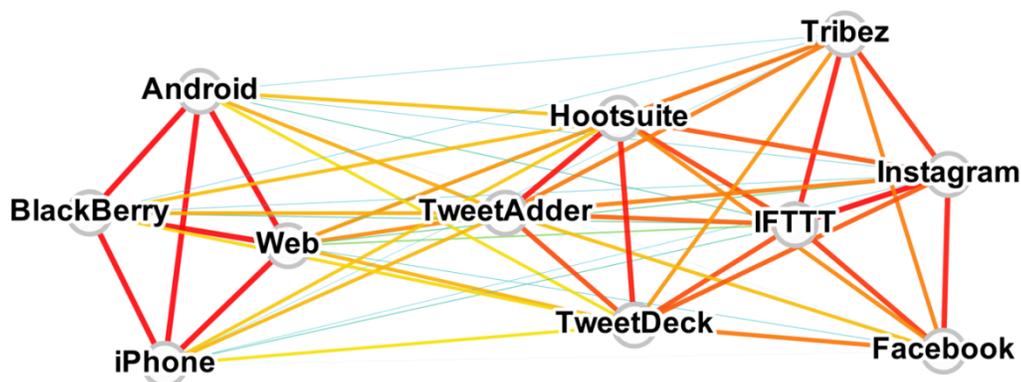


Figure 6. Similarity network of sources. The color and weight of each link show the similarity between sources.

This experimental approach delineates three distinct groups. The cluster on the left in Figure 6 includes the most widespread, general public-oriented personal clients for mobile and Web, including Twitter for iPhone and Android. The three more professional clients Hootsuite, TweetAdder, and TweetDeck form a second cluster that is, however, quite closely related to automators such as IFTTT, which also share similarities with cross-syndicators largely because of increased uptake of links and hashtags. Although this method is clearly underdeveloped, it allows for a large-scale investigation of Twitter sources based on similarities in the tweet practices they enable.

Sources, this section has shown, come with their own interpretations and uptake patterns of platform grammars. They enable distinct modes of being on Twitter, which may enhance, reinterpret, or recombine its grammars, and even feed the grammars of one platform into another.

Disassembling Metrics and Tracing Practices

After investigating clients by describing their features and certain characteristics of the contents they are associated with, we asked how the source variable could be used as an indicator to enhance our understanding of the practices and interpretations that are often equated into seemingly straightforward counts of natively digital objects. Similar questions have been asked by Bruns and Stieglitz (2013), who focused on qualifying hashtag-based samples by identifying a series of key intersecting metrics that provide further insight into the internal composition of samples in terms of user activity, activity over time, network relations, or sentiment. In this article, we work the other way around, asking how the source variable can be connected to dynamics within but also outside the platform. In what way can source variables, applied to specific samples, figure as traces of the heterogeneity of technicity, practice, and meaning, and thereby challenge overly monolithic accounts of medium specificity? We begin to answer these questions by intersecting sources with other metrics with the help of alluvial diagrams produced by Twitter Capture and Analysis Toolset's Sankey Maker. These diagrams facilitate the analysis of relationships between categorical variables of tweets such as languages, time zones, hashtags, hosts, and sources.

The exploratory visualizations provide a series of contextual nuances worth investigating when qualifying data sets. First, investigating the sources making up a data set—whether it is close to the overall distribution, a unique composition of sources, or driven by a few specific sources—can provide traces for the composition of practices and interpretations feeding into the data. In the case of the hashtag #CallMeCam (see Figure 7), a "shoutout" hashtag initiated by the teenage YouTube celebrity Cameron Dallas who promised to call one of his fans using the hashtag #CallMeCam, most tweets come from one particular source, namely, Twitter for iPhone. Given that the iPhone is the preferred device of U.S. teenagers,¹⁷ this intersection adds to the not implausible suspicion that #CallMeCam is driven by adolescent fans and points to an intricate relation between YouTube, Twitter, and the iPhone. In a setting where sociodemographic data are sparse, sources can provide a certain degree of context.

¹⁷ <https://www.cnet.com/news/iphone-dominating-android-more-than-ever-among-teens/>

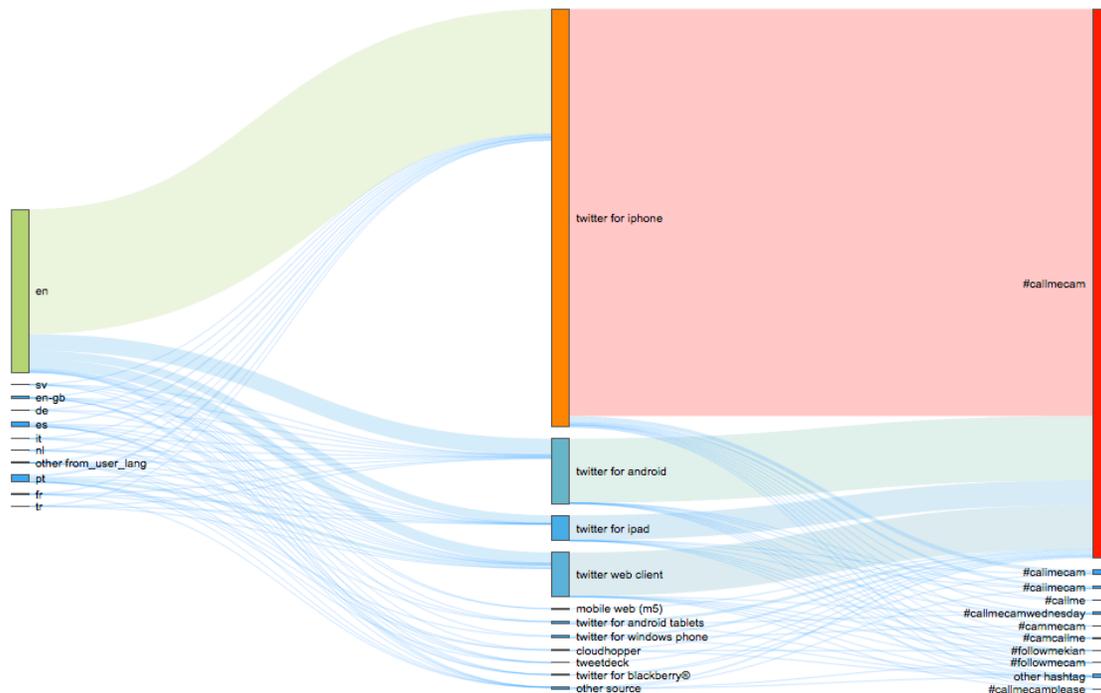


Figure 7. Profile for #CallMeCam with language (left), source (middle), and hashtags (right); Twitter for iPhone dominates the hashtag, and the thicker flow connectors on the right side indicate that users employ the hashtag several times in a tweet, sometimes with variations or spelling errors.

In the case of the hashtag #love, two sources stand out, namely, Instagram and dlvr.it. Although these are relatively rare in the overall data set (0.77% and 0.81%, respectively), they are much more common in the set of tweets containing #love (8.6% and 11.8%, respectively). This indicates that one of the most common hashtags on Twitter is, to a significant degree, animated by two distinct sets of practices, namely, cross-syndication of tweets from and written for another platform, as well as largely automated content production. It also shows that the distribution of sources may deviate strongly from the average when looking at specific subsamples, providing an interesting indicator to researchers. A second methodological exercise in retracing consists of identifying the content, media or websites tweets within a sample link to. In the case of #love (see Figure 8), the links sent from tweets via dlvr.it point to topically nonrelated and promotional content; here, domain names function as traces of hashtag hijacking by spammers. Interestingly, the references to bakmoda.com, a now defunct shopping site, are not coming exclusively from dlvr.it but also from Twitterfeed, a plugin for cross-syndicating blog content. This shows that in our investigation of Twitter, clients open up on a larger software ecosystem that draws on a wide array of tools, services, and platforms that are connected in different ways.

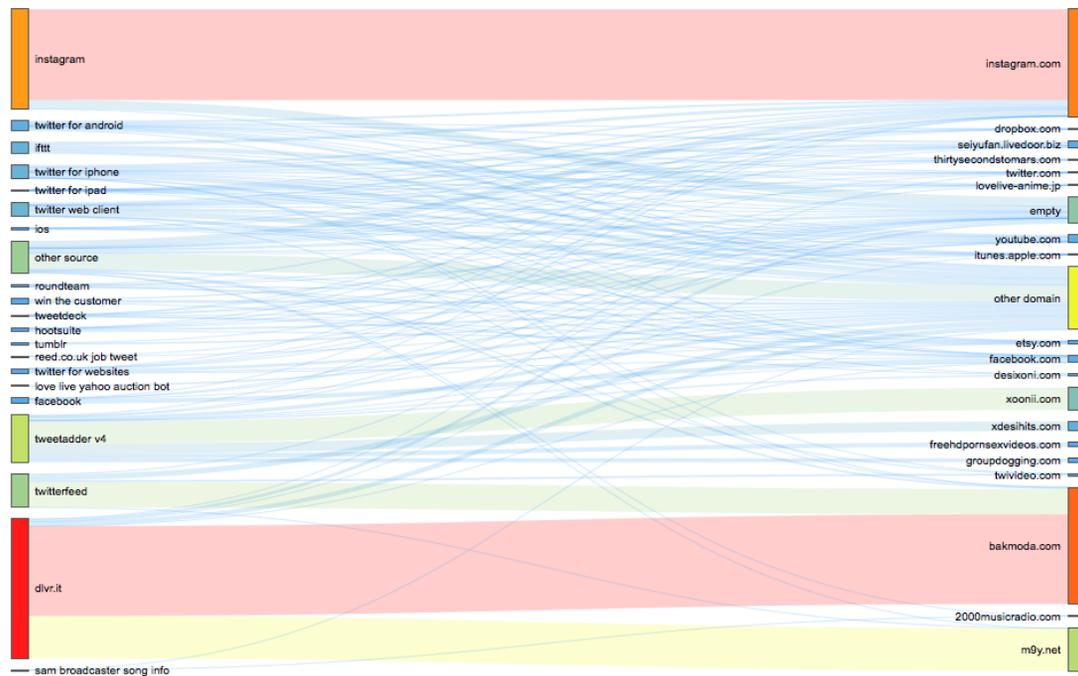


Figure 8. Profile for tweets containing the hashtag #love and a URL, with source (left) and hosts (right).

In both cases, attention to the client software can shed light on the specificities of content production as a distributed accomplishment. Instagram shows how tweeting can be software supported beyond full automation: Users produce the Instagram content and the app automatically readies it for Twitter; yet, users again have to confirm cross-syndication for every single tweet. However, dlvr.it adds timed, triggered, and high-volume tweeting to the repertoire, rendering practices such as hashtag hijacking logistically feasible in the first place.

Although many sources with specific profiles constitute relatively small portions of the overall volume of tweets, they can be key drivers in specific samples and offer traces of particular practices or user groups, adding to our interpretative arsenal. Intersecting sources with other variables allows for a better understanding of what entities such as hashtag or domain counts are animated by. Even straightforwardly countable units such as tweets per time frame remain problematic if we gloss over internal differentiations in terms of tweet production. Including the source in the analysis provides a methodological pathway toward accounting for this differentiation.

Conclusion: Lively Metrics and Platform-Specific Methods

This article started with the question of how researchers can reimagine “methods of the medium” in the context of distributed platform-based media. Starting from an science and technology perspective,

the main premise of digital methods has been that digital data are structured and ordered by the medium itself. In the context of platforms, as we have shown, this needs to include (third-party) clients and practices such as cross-syndication. We used Twitter's source variable to explore the software ecosystem forming around platforms and revisited the question of how social media data are marked by this ecosystem from a platform perspective. The bounded openness of platform grammars invites a variety of stakeholders to realize distinct practices, interpretations, and objectives, and the commensurating effects of platform metrics, we conclude, cannot be reduced to a mere media effect, but need to be regarded as "happening," as Lury and Wakeford (2012) write, as situated accomplishment. Therefore, we propose treating social media metrics as "lively" metrics that are assembled in specific and situated ways. In the sense of Marres and Weltevrede (2013), *lively* means internally dynamic and subject to variation and happening. Social media metrics such as counts of (re)tweets, hashtags, favorites, or links are lively, as they can be enacted through automated accounts, issue campaigning, spam, or manual tweeting, via automator software, Web interfaces, or cross-syndication. They are animated by dynamic practices and platform cultures, and the source variable—which is unfortunately unique to Twitter and cannot be retrieved for other platforms—has allowed us to start retracing this liveliness. For a more complete picture, one would have to address further dimensions of the client ecosystem that inform the liveliness of metrics, for instance, the partially overlapping ecosystem of specialized reading devices that range from embedded tweets on websites to full-fledged analytics software, all of which enable distinct ways of organizing, interpreting, and engaging tweets.

Throughout the article, we further found that liveliness is subject to varying degrees of automation enabled by the wider platform ecosystem. So far, automation has mainly been conceptualized as "bot activity" in social media research, enabled by scripted robots (Geiger, 2014). Bots are often associated with spam and promotional practices, and are considered malicious by users and platforms themselves. Readers may have noted that in this article, we have abstained from talking about bots and instead referred to automation as it opens up a wider spectrum that spans the whole range between manual and fully software-generated content. Professional clients such as Hootsuite allow users to schedule and autopost tweets, perform bulk uploads and random postings, turn RSS feeds into posts, and organize collective writing. These features introduce a degree of automation in which clients do not simply produce contents for users but also assist content composition in various ways. Another level of automation is attained by what we have called automator sources, including IFTTT, dlvr.it, and TweetAdder, which make internal and external content Twitter-ready through syndication and transposition, and which automate selected interactions, list-making, or archiving. In the case of cross-syndication, we encounter a specific form of automation, as the respective content could have been produced manually, just not for the Twitter platform, even if certain grammars seem to translate more or less directly. Limiting automation to the binary categories of bot and human fails to account for what needs to be considered as a distributed accomplishment that includes users, platforms, software, and content. Due to its entanglement with meaningful practices, automation asks for a situated view on how connected software ecosystems enable and favor distinct modes of being on Twitter.

This has a number of implications for doing digital research that move far beyond the mere issue of data cleaning. First, what appears as straightforwardly countable metric is assembled and lively, bringing together potentially heterogeneous practices and data. Frequency counts are not comparable

from the outset, but need to be made comparable through additional interpretation. The source metric offers initial cues to understand what a metric actually counts. Second, the perspective developed in this article requires digital research to engage with automation with great attentiveness and conceptual depth. Instead of trying to eliminate all content and activity that is nonhuman, digital research should attend to the mutual constitution of manual and software-supported practices. Third, thinking platform metrics as lively thus requires us to refine our notion of medium specificity. In the context of platforms, one cannot (always) focus on the single medium, its grammars, and data formats, but needs to take the involvement of third parties, their sources, and interpretations into account. Although platforms partially stabilize the ongoing process of commensuration through rules and regulations, they also make it increasingly opaque, as sources remain largely invisible in frontends and metrics are presented as straightforwardly countable. Throughout this article, we have attempted to render this liveliness accountable again. When developing platform analytics tools, developers could attend to the same challenge and ask whether their analytics render the liveliness of metrics more or less opaque. Fourth, from such a perspective, not only platforms, users, and third parties are involved in the liveliness of metrics but also digital researchers who recombine and intersect them, imposing yet another set of objectives and interpretations on platform data.

References

- Adner, R., & Kapoor, R. (2010). Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. *Strategic Management Journal*, 31(3), 306–333.
- Agre, P. E. (1994). Surveillance and capture: Two models of privacy. *The Information Society*, 10(2), 101–127.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov, & B. Tseng. *Proceedings of the Third International Conference on Weblogs and Social Media* (pp. 361–362). Menlo Park: AAAI Press,
- Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. Cambridge, MA: MIT Press.
- Bodle, R. (2011). Regimes of sharing. *Information, Communication & Society*, 14(3), 320–337.
- Borra, E., & Rieder, B. (2014). Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66(3), 262–278.
- Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, 16(2), 91–108.

- Bucher, T. (2013). Objects of intense feeling: The case of the Twitter API. *Computational Culture*, 3. Retrieved from <http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api>
- Burgess, J. (2016). Twitter's changing device ecology. Retrieved from <https://medium.com/dmrc-at-large/twitters-changing-device-ecology-179dbec5d4dc>
- Dijk, J. van. (2011). Tracing Twitter: The rise of a microblogging platform. *International Journal of Media & Cultural Politics*, 7(3), 333–348.
- Driscoll, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, 8, 1745–1764.
- Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24(1), 313–343.
- Gawer, A. (2014). Bridging differing perspectives on technological platforms: Toward an integrative framework. *Research Policy*, 43(7), 1239–1249.
- Geiger, R. S. (2014). Bots, bespoke code, and the materiality of software platforms. *Information, Communication & Society*, 17(3), 342–356.
- Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal*, 16(2), 1–18.
- Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, 12(3), 347–364.
- Lury, C., & Wakeford, N. (2012). *Inventive methods: The happening of the social*. New York, NY: Routledge.
- Manovich, L. (2013). Media after software. *Journal of Visual Culture*, 12(1), 30–37.
- Marres, N., & Weltevrede, E. (2013). Scraping the social? *Journal of Cultural Economy*, 6(3), 313–335.
- Miller, P., & Matviyenko, S. (2014). *The imaginary app*. Cambridge, MA: MIT Press.
- Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased? Assessing the representativeness of Twitter's streaming API. Retrieved from <https://arxiv.org/pdf/1401.7909>
- Paßmann, J., & Gerlitz, C. (2014). Zur sozio-technischen Geschichte der Plattformaktivitäten Fav, Retweet und Like ["Good" platform-political reasons for "bad" platform-data. On the socio-technical history of the platform-activities fav, retweet, and like]. *Mediale Kontrolle unter Beobachtung*, 3(1), 1–40.

- Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*. Advance online publication. doi:10.1177/1461444816661553
- Puschmann, C., & Burgess, J. (2014). The politics of Twitter data. In K. Weller, A. Bruns, J. Burgees, M. Mahrt, & C. Pushmann (Eds.), *Twitter and society* (pp. 43–54). New York, NY: Peter Lang.
- Robinson, D. (2016). Text analysis of Trump’s tweets confirms he writes only the (angrier) Android half [Blog post]. Retrieved from <http://varianceexplained.org/r/trump-tweets/>
- Rochet, J.-C., & Tirole, J. (2006). Two-sided markets: A progress report. *The RAND Journal of Economics*, 37(3), 645–667.
- Rogers, R. (2013). *Digital methods*. Cambridge, MA: MIT Press.
- Sippey, M. (2012, June 28). Delivering a consistent Twitter experience [Blog post]. Retrieved from <https://blog.twitter.com/2012/delivering-consistent-twitter-experience>