Social Media + Society
January-March 2021: 1–10
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2056305120984472
journals.sagepub.com/home/sms
SAGE

Invisible Data: A Framework for Understanding Visibility Processes in Social Media Data

Christina Neumayer[1], Luca Rossi[2], and David M. Struthers[3]

Abstract

Social media data are increasingly used to study a variety of social phenomena. This development is based on the assumption that digital traces left on social media can provide insights into the nature of human interaction. In this research, we turn our attention to what remains invisible in research based on social media data. Using Andrea Brighenti's work on "social visibility" as a point of departure, we unpack data invisibilities, as they are created within four dimensions: people and intentionality, technologies and tools, accessibility and form, and meaning and imaginaries. We introduce the notion of quasi-visible data as an intermediary between visible and invisible data highlighting the processual character of data invisibilities. With this conceptual framework, we contribute to developing a more reflective and ethical field of research into the study of social phenomena based on social media data. We conclude by arguing that distancing ourselves from the assumption that all social media data are visible and focusing on the invisible will enhance our understanding of digital data.

Keywords

visibilities, social media data, epistemology, computational methods, digital methods

Introduction

People live-tweet, share their stories on Facebook, use YouTube videos to present themselves and their narratives, appear in selfies on Instagram, and retweet, share, like, and comment on posts and images of others. The flood of images, posts, clicks, stories, and shares is turned into data by social media corporations. Despite critical voices (see, for example, Baeza-Yates, 2018), social media data have grown to become a large and rich data source for studying a variety of social phenomena. Yet, despite much of our daily interactions moving online and leaving digital traces, a large amount of data remains entirely invisible or only visible to social media corporations. These data invisibilities can result from the non-coordinated actions of three main actors: platforms, users, and analysts. Platforms can act through design decisions like making digital traces ephemeral, or with the platform's data-policies like restricting access to data or commercializing data access. People can intentionally employ strategies to creatively tamper with data to achieve various goals, from increased to reduced visibility (particularly within censored or highly surveilled platforms) to selective visibility (limited to certain groups of people). Finally, researchers and analysts with tools and methods further curate and analyze the data and produce a different set of visibilities and invisibilities. This article will turn our attention to these (in)visibility processes that occur when social media data are used to represent a social phenomenon.

Several fields and disciplines have discussed the characteristics of social media data being employed for drawing conclusions about social phenomena. Critical software and data studies provide valuable insights into the mystification of data and the imaginaries that surround such mystification processes (Andrejevic, 2013; boyd & Crawford, 2012; Kitchin, 2015). Computational social science has primarily focused on retrieving and analyzing social media data to discern results about human behavior as well as to predict

[1] University of Copenhagen, Denmark
[2] IT University of Copenhagen, Denmark
[3] Independent Researcher, Denmark

Corresponding Author:
Christina Neumayer, Department of Communication, Faculty of Humanities, University of Copenhagen, Karen Blixens Plads 8, 2300 Copenhagen S, Denmark.
Email: christina.neumayer@hum.ku.dk; @nechri

outcomes of social interaction (e.g., Giglietto et al., 2012; Margetts et al., 2015). Social scientists have critiqued computational methods, techniques of automation, and machine learning approaches for their opacity and their inefficacy when it comes to such predictions (see Baym, 2018; boyd & Crawford, 2012; Hargittai, 2018). While social media data have entered a variety of fields in the humanities, social and political sciences and produced valuable results, there is a lack of knowledge about these data, their political, cultural and scientific meaning and the practices that produce, curate, and maintain them (see Earl, 2018).

In this article, we turn our attention to data invisibilities based on an understanding of the process in which data becomes visible. To overcome the dichotomy of visible and invisible, we introduce the notion of "quasi-visible" as an intermediary state of social media data. As we do not consider data visibility as a stable attribute of the data itself, we develop a conceptual framework that helps us understand how data moves through stages of being visible, quasi-visible and invisible. We first introduce our understanding of visibility and invisibility as processes in the context of social media data. We then introduce epistemological claims underlying visibility processes and examine the dimensions within which data move from visible to quasi-visible to invisible in such processes. The article concludes by discussing the consequences of using such a concept for understanding data invisibility processes in studies based on social media data.

Visibility as a Process

We begin by arguing that visibility and invisibility in the context of social media data are not dichotomic constants. Yet, to understand the connection between visibility and invisibilities in social media data, we first need to attend to the process of becoming visible. "Looking at someone who looks back" is the beginning of all social interaction from a societal perspective, however, as Andrea Brighenti (2010a) contends, there is an asymmetry between being looked at and looking, which one cannot simply correct or entirely avoid. If we are to understand social visibility and move beyond a dichotomy of visible and invisible, we need to "complexify our understanding of visibility as not simply a monodimensional or dichotomic, on/off phenomenon" (Brighenti, 2010a, p. 1). Visibility is, from this perspective, context dependent as it is set in a particular technical, political, and social arrangement. Brighenti (2010a) suggests an exploration of visibility unrestricted to the visual and as a dimension of society and the social as it is constituted by a material and immaterial layer. It is these material and immaterial layers that we need to understand to move further in our conceptualization of data invisibilities. Visibility is thus inherently unstable, as narratives, symbols, and images can move from the visible into the invisible and back again. Understanding visibility as deeply social breaks with the assumption that data are simply there to be observed for drawing conclusion about any given social phenomena. We need to trace the visibility processes underlying such data with its material and immaterial layers.

Visibility has been discussed in many different fields and contexts in the social sciences, but there are many unanswered questions about how to conceptualize and measure visibility as well as how to reflect upon our own positioning as researchers (Blaagaard et al., 2017, p. 1115). Brighenti (2010b) further contends that visibility regimes are fundamentally interwoven with technologies of power and constitutive of political regimes. As such, visibility is not just what is visible and visual (such as images) but "meaning inscribed in material processes and constraints" (Brighenti, 2010b). The material is made visible but also makes visible in a similar vein. Rather than free-floating meaning, the visible according to Brighenti (2010b) is the material and strategic connecting to the technical and the social. Techno-social visibility is then dependent on social relationships and norms, and often the normal fades into the background, while the unusual (such as, for example, political protest) alarms us and becomes noticed and visible (Brighenti, 2007, p. 326). We account for these dynamics by constructing a relational understanding of visibility that provides guidance on how such relationships are shaped by imaginaries about classifications, but the imaginaries are also shaped by invisibility processes.

Brighenti (2007, p. 330) uses the example of media representations of migrant criminals to argue that the overemphasis and "supra-visibility" creates a selective focus that becomes representative of moral minorities. Following a similar line, Chouliaraki and Stolic (2017) discuss the regimes of visibility under which refugees' only agentic capacity is that of criminals and terrorists, never their humanity and their presence as equals. Brighenti (2007) adds that processes of visibility are always relational. Media institutions, for example, can confer visibility to groups and people (Brighenti, 2007, p. 332), but we need to understand how this quality of conferring visibility is created. This means conceptualizing that which remains invisible; the invisible supports the stabilization of such supra-visible representations. To understand such ontologies, we need to unpack the processes and relationship of visibility and invisibility, including an understanding of how classifications (such as criminals) contribute to creating visibilities and invisibilities. The normalizations of classifications push any other capacity, except one, into the background. In Bowker and Star's (1999) conceptualization, classification tools become "naturalised" and by doing so, become invisible. The connection between classification and visibility can also be analyzed in the context of research methods and application programming interfaces (APIs) as classifiers rendering data visible or invisible, but also imaginaries that we have about the meaning such data provide.

Returning to Brighenti, "[v]isibility is a double-edged sword: it can be empowering as well as disempowering" (Brighenti, 2007, p. 334). This opens the question of who

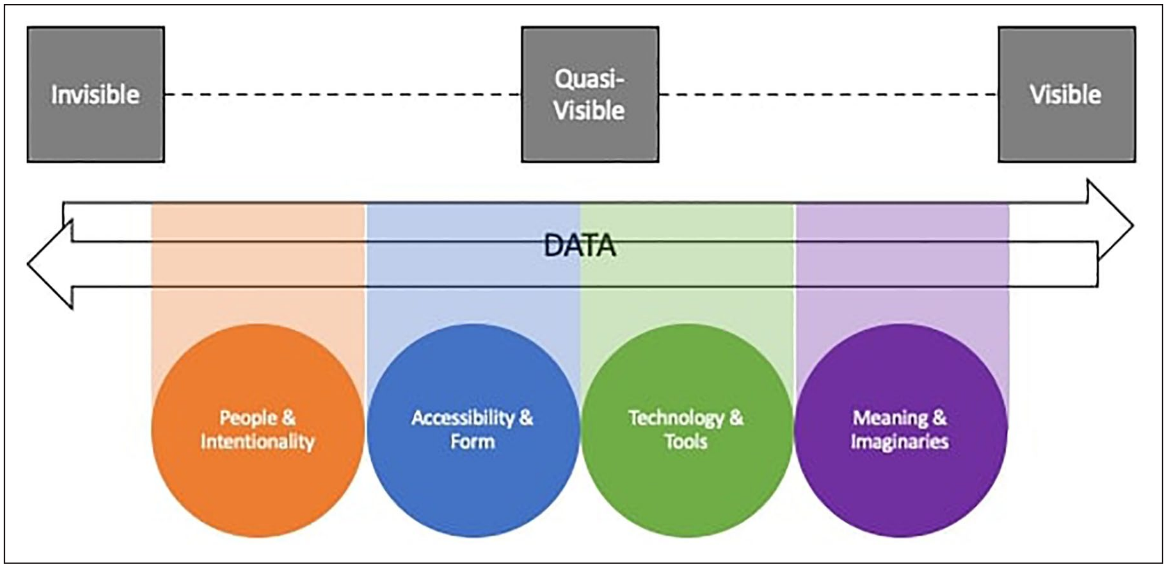

Figure 1. A conceptual model of data (in)visibility. Data move from visible to invisible within 4 socio-technical dimensions.

gets to decide what becomes visible and what can or must remain invisible. This question gets more complex when social media come into play as information and data flows remain largely invisible. Flyverbom et al. (2016) are concerned with how organizations can manage visibility and remain in control of what is disclosed and what should remain invisible. They use the term "visibility management" (Flyverbom et al., 2016 99) to discuss the increasing complexity of managing visibility, as what can be seen is not always clear with the involvement of social media corporations. "Visibility affordances" (Flyverbom et al., 2016) are concerned with the use of technologies by individuals that makes information viewable to others. On social networking sites, this means that interactions within the confines of two people can potentially be seen by third parties (Flyverbom et al., 2016). While these conversations might not be visible per se (except the two people involved in the interaction), they still might be visible to social media corporations. This then can be used to accumulate future economic surplus and capital (Fuchs, 2015). To complicate this further, while people might not actively communicate their preferences in social media, it might be possible to make their preferences visible through the traces that they leave (in searches performed or on their social media profiles). Much attention has recently been paid to visibility processes though social media corporations giving access to this data or restricting access to researchers (see Walker et al., 2019). The intentionality of the people who initially produce content in social media is often mentioned as a concern when social media corporations or authorities use these data (see, for example, Uldam, 2016), but we also have to take this into account when drawing conclusions based on social media data as researchers and analysts (see, for example, Neumayer & Rossi, 2016).

Based on Brighenti's conceptualization of visibility, we introduce four dimensions within which visibility processes take place in social media data: (a) people and intentionality, (b) accessibility and form, (c) technology and tools, and (d) meaning and imaginaries. While data can move between visible to invisible and within these material and immaterial dimensions, we suggest that we also have to conceptually introduce a space in-between which we call "quasi-visible" (see Figure 1). Quasi-visible social media data comprises data that are not visible per se but can be made visible within the dimensions introduced above. Visibility processes can take place from quasi-visible to visible (e.g., by making data visible by using certain methods; social media corporations granting access to data) or from quasi-visible to invisible (e.g., when social media corporations permanently delete data). These processes do not take place in a vacuum but within the four dimensions, which we expand upon below, but first we need to briefly discuss the epistemological assumptions underlying this conceptual understanding.

Chasing the Data: A Note on Epistemology

As researchers, we often treat social media data as if they have always been there waiting for us to be collected and analyzed, if only researchers had the access, tools, and methods to do so. The large volume of social media data generated by users combined with the automated analysis of such data with major advances in digital methods and data science have rendered these data seemingly objective, quantifiable, and generalizable (see Gandomi & Haider, 2015; Gayo-Avello, 2013). This understanding of social phenomena being objectively observable through such data is particularly prevalent in computational social sciences (Lazer et al., 2009), while recently the field has moved toward a more nuanced and theory-driven understanding of data, methods, and unavoidable bias (Lerman, 2018). Objective truth claims are often based on the assumption that social media data can potentially provide an accurate representation of human activity, as they are simply there to be found and collected. Media studies have differentiated between found and made data (Jensen, 2012). While made data are created by the researcher (e.g., by questionnaires or interviews), found data would be, for example, media content. To understand data visibilities and invisibilities, we have to move away from this notion of found data and start from the assumption that all data are *made*.

Garforth (2012) references Latour's early work concerned with the invisible in science. Latour and Woolgar (1986, p. 32) unpack scientific research in the laboratory and study the construction of knowledge through "the process by which scientists make their observations." The construction and stabilization of fact is thus a process that is carried out through practices and discourse. Once facts are stabilized, they appear as if they were always there until their discovery by scientists. According to Latour and Woolgar (1986), however, this discovery is a result of an ongoing process of discursive persuasion across different groups and with different social forces at play. Audiences, they argue, accept these facts as reality because of the mystifying manner in which they are constructed. When working with social media data, we must also examine how such processes take place to understand the underlying meaning of the results produced.

Social scientists have a long tradition of seeking to understand data, where they come from, and what they represent (Mattoni & Pavan, 2018). In this context, critical software and data studies provide insights into algorithmic bias, the mystification of data, and the power imbalances within data-regimes (Andrejevic, 2013; boyd & Crawford, 2012; Kitchin, 2015). Computational methods have been critiqued for their techniques of automation, and machine learning approaches for their opacity and lack of possibilities for human intervention (see Baym, 2018; Hargittai, 2018). This also includes the notion of data as a form of power (see Iliadis & Russo, 2016, for an overview). There is an increasing understanding of these data always being "cooked" and "raw data" being described as "an oxymoron" (Gitelman, 2013). However, this is often considered a result of social media corporations not providing complete data access or datasets being incomplete, biased, or not representative in the form that they can be accessed. Practices and processes underlying the data are usually not considered in research relying on social media data. With this in mind, and returning to visibility, we argue that these data are never just there or simply visible. There are sociotechnical practices underlying these data at play while they are *made* and *made visible*.

From this understanding, data are never visible or invisible per se. As many have argued, especially after the lockdown of social media platforms for researchers, social media corporations do play an important role in *making data visible* and accessible (Bruns et al., 2018). Yet, we argue that scholars also have to appreciate that their research practices make some data visible while rendering other data invisible. To better situate data invisibilities and construct, a more reflexive approach to research based on social media data, we turn our attention to a field that has always worked with "found data." Historians working with archived sources (data) are accustomed to contemplating their role in knowledge creation when writing history based on their sources as well as the way archives are *made*; power relationships are imbedded throughout this process (Foucault, 1969/2003).

For our purposes here, the data process we are interested in begins with historical sources (personal letters, organizational and state documents, etc.) at the point of their creation and continues through to the decisions made about which sources merit preservation, cataloging, and continued storage in formal and informal archives. Historical production—the writing of history by historians from found sources—brings with it power relationships distinct from those present in the data process. This acknowledgment brings two additional inflection points into consideration in the making of historical writing. First, historians make decisions at the moment of "fact retrieval" about which data to collect (Trouillot, 1995, p. 26). Second, when they construct narratives or other forms of presentation, often with epistemological claims rooted in social sciences, historians participate in a broader societal process of valorizing the significance of a particular period or understanding of the past during the period in which they write; "presentism" is as unescapable as it is criticized in the field. Trouillot argues that "silences" are created at each of the four points from the data process to historical production (Trouillot, 1995, p. 26). By using the term "silencing," Trouillot draws out the myriad actions in the process of creating silences.

This article exchanges silencing for invisibility because we want to distinguish the print and other written media of traditional historical sources from contemporary digital data. Digitized historical sources align more closely with the material logics of traditional historical sources than they do with contemporary digital data. Yet, here a comparison is helpful: "Historical power is not a direct reflection of a past occurrence, or a simple sum of past inequalities measured from an actor's perspective or from the standpoint of any 'objective' standard, even at the first moment" (Trouillot, 1995, p. 47). Historical power and silences seep through the four key points discussed in similar ways to the *socio-technical processes* at play that *make* contemporary digital data *visible* as well as *invisible*.

Data Invisibilities as a Four-Dimensional Process

To moving forward with an understanding of the visibility processes of social media data, we first need to take a more careful look at the four dimensions in which these processes take place. Mattoni and Pavan (2018) argue that we have to understand big data as a complex set of political, cultural, and scientific practices including technologies from social media platforms as data infrastructures to mobile phone applications, APIs and analysis software; imaginaries and (academic) discourses around social media data and its consequences; and people, such as data scientists, software developers, social scientists, policy makers, and platform owners. Their focus on practices challenges the assumption of social media data simply *being* there and traces the processes underlying invisibilities while these data are *made*. Taking this further, we transcend the one-dimensional perspectives on data (such as informational, computational, or epistemological data) and understand visibility-processes in social media data as taking place within the four abovementioned socio-technical dimensions: (a) people and intentionality (people leave traces of data with the aim of making visible but also to obfuscate or hide), (b) technologies and tools (as collections of "*fact*" are observed and stored), (c) accessibility and form (data are made accessible in various forms), and (d) meaning and imaginaries (data are believed to have the capacity to measure, represent or unveil social phenomena). Within these dimensions, we can trace how invisibility- and visibility processes take place when social media data are made. In the following, we introduce the four dimensions using illustrative examples from research based on social media data. While these examples are not exhaustive, they contribute to understanding the visibility processes of social media data within these dimensions.

People and Intentionality

When we study social media data to understand human behavior, researchers make assumptions about people who produce such content. Based on social media data, we can quantify users as followers or produce networks connecting different user accounts. However, the numbers used to understand user behavior are often more complex. The multiplicity of tactics and actors leads to a contested and complex flow of digital data. One example is the number of followers of retweets on Twitter which has often been used as a proxy to measure the visibility of tweets (Harada et al., 2017; Suh et al., 2010; Yang & Counts, 2010). Despite the highly quantifiable nature of social media data, knowing the exact number of users who have potential access or actually seen a certain image or message circulating in social media is an impossible undertaking. Even by focusing on a relatively simple context such as Twitter, available metrics such as number of followers, number of retweets or any combination of these two are affected by well-known problems (e.g., nonhuman actors, dead or inactive accounts) thus the numbers should be understood more as potential viewers rather than actual viewers (Davis et al., 2016). More precisely, the number of followers is deeply affected by the high number of inactive users as well as by the large number of bots and fake accounts that exist on Twitter (Davis et al., 2016).

At the same time, more activity-based metrics, such as the number of retweets or interactions, will most probably fail to include the less active segment of users as well as lurkers or lightly engaged users (Bernstein et al., 2013). Moreover, structural network elements (Petrovic et al., 2011) such as the number of friends and followers of the sender can strongly impact a message's retweets. While these metrics can be a useful way to understand social media data and are already quite complex, we are not taking the intentionality of people into account. Yet, it is an important dimension in visibility processes. While some people (such as influencers, see Abidin in this special issue) employ creative strategies to be highly visible, others are doing their best to remain invisible (such as marginalized or vulnerable groups, see, for example, Triggs et al., 2019) and use strategies such as obfuscation (Brunton & Nissenbaum, 2015) or controlling their visibility by adopting creative and unorthodox solutions based on the platform's affordances (Lange, 2007). People can also move to platforms that allow for encrypted communication such as WhatsApp or Telegram (Belair-Gagnon et al., 2018, see Semenzin & Bainotti in this special issue).

Taking the dimension of people and intentionality into account when thinking about visibility processes in studies based on social media data, has two major consequences. The first is that social media data have limitations, as people only become visible in such data through modes of reception such as likes, shares, posts, retweets, while not having an active profile and only viewing might remain invisible. The second is that taking intentionality into account means that we often need to combine computational methods with other social science methods (such as interviews) to move toward more conclusive results (see, for example, Jensen et al., 2020). This also comes with responsibility as computational methods can give an overview of social media activity of a group of people, but this visibility might collide with the intentionality of people (Croeser & Highfield, 2015). Conversely, including the dimension of people and intentionality into our conceptual understanding of visibility processes in social media data, urges social media corporations, but also researchers and analysts to reflect upon their own role in making data visible.

Technology and Tools

Social media platforms, as well as other types of digital media, organize traces left intentionally or unintentionally by human interactions into specific data form. These traces of communicative practices are, for example, organized and, eventually, returned as a list of tweets; social relations of various kinds are organized as a list of friends by Facebook. Users' practices are transformed in "relevant" facts or entities that can be stored and analyzed in the most appropriate data format. This is often what has been called a "social graph," a computer-readable representation of observed social relations and interaction. In 2012, Mejias described the reduction of all human interactions into network dynamics as "nodecentrism" (Mejias, 2012). While this is far from being a new process, the consequences that this has for research are becoming more apparent today. Social graph data are essentially a selection of what social media platforms understand as interesting or useful in the behavior of the users and what can be made available through the platforms' APIs. For example, Twitter APIs describe a single tweet with approximately 31 attributes, describing both elements that have been directly produced by the user (e.g., the text of the tweet) and information that has been added at a later stage, yet before the tweets are made visible through the APIs (e.g., the detected language of the tweet, or the presence of links to potentially sensitive content). The data *directly* available from the APIs show traces of substantial pre-process as well as traces of the suggested use. What is encoded as a data attribute suggests and supports specific types of research and analysis. Tweets' attributes describing the messages in terms of original tweets, retweets, or replies suggest a more informational type of communication rather than the ephemeral or intimate use of the platform (Burgess & Bruns, 2012).

Returning to the work of historians, the idea that the way data are stored also shapes its understanding, is not new. Foucault (1969/2003) in *The Archeology of Knowledge* contends that archives consist not only of shelves and artifacts that historians can investigate, but include the larger apparatus and set of rules and power relationships that allow archives to exist, including the institution and even the building in which it is located. Similarly, social media data are stored and archived within their own logics, algorithms, policies and business models that make them profitable (Gillespie et al., 2014; Van Dijck et al., 2018). The data are arranged by the techno-commercial infrastructure organized by algorithmic classification and sorting based on likes, hashtags, mentions, comments, and favors (Neumayer & Struthers, 2019). Paul Dourish (2017) argues that the materiality of digital objects (including data) is not only the archive itself, but also the technologies we can use to produce the material that can be archived. How we create data in an interactive process with media technologies is an essential part of the materiality of information, which is concerned with the material forms of representation of digital data leading to particular interpretations and actions.

Including the dimension of technology and tools into our conceptual understanding of visibility processes in social media data has two major consequences. The first is that social media platforms already provide a specific technological framework for interaction, and human interaction can be made visible and rendered invisible by such tools and technologies. The second is that social media platforms also already use tools and technologies that "precook" (to use Gitelman's words, 2013) the data as they are processed, archived and stored, before analysts and researchers can retrieve them. Both of these processes can render certain aspects of human interaction highly visible, quasi-visible, and others invisible. Taking these visibility-processes into account, can give us a better idea of what the data we retrieve from social media platforms represent. In short, social media data, even when collected directly from the "source" (platform APIs), are a tailored selection of facts, content, and activities organized according to the platforms' relevance criteria in a way that facilitates specific analysis (deemed to be more interesting).

Accessibility and Form

Once users' behavior has been coded into a specific data structure its accessibility through an API is often part of platforms' business models (Langlois & Elmer, 2013). The "regimes of access" (Burgess & Bruns, 2012) of social media companies regulate the data researchers can collect and use. Such conditions direct scholarly attention toward relatively open platforms (e.g., Twitter) and collection strategies that follow their functionalities (e.g., hashtags). This leads to methodological development in some research areas (e.g., focus on text-based analysis), while computational analysis of visual content lags behind (see Neumayer & Rossi, 2018). Studying social phenomena solely based on Twitter data, for example, can only give us a limited representation of a social phenomenon. At the same time, it makes visible certain aspects that are not visible per se (such as retweet networks, communities based on shared hashtags, etc.). While these social constellations are not naturally there, social media analysists make such data visible through their methods. At the same time (and similar to the historian) turning our attention to the invisible might inform our understanding of power relations as well as the socio-technical materiality of social media data. While social media platforms have an interest in allowing external developers to contribute to their ecosystem, access to users' data from marketing and advertising companies has increasingly become a revenue stream for platforms (Van Dijck et al., 2018).

From a research perspective, by commercializing data access, social media companies create the condition for dividing researchers between "data haves" and "data have-nots" (Weller, 2015; boyd & Crawford, 2012). Since the resources available to acquire data access are unevenly distributed this will likely produce a prioritization of those research topics that are mainly of interest to the "data haves" (Bruns et al., 2018). Yet, despite being largely used by researchers and scholars, APIs are not designed for scientific research, thus many traditional scientific practices (e.g., sampling methods, data sharing, replicability etc.) will not be applicable in this context. While the practical consequences of this varies largely from platform to platform, Burgess and Bruns (2012) quote directly from Twitter's FAQs and show that the platform's APIs are "not meant to be an exhaustive archive of public tweets and not all the tweets are indexed or returned"; and that "some results are refined to better combat spam and increase relevance." One concern of social media platforms is users' experience (or computational scalability) rather than accuracy, completeness, or representativeness of returned results.

The problematic relation between the commercial nature of APIs and research practices partially results from the unstable and unreliable nature of APIs. What is accessible through the APIs changes over time to maximize social media companies' profit opportunities or to minimize public relations risks. An illustrative example is the historical evolution of the free API on Twitter, which over time became less usable for researchers in need of large quantities of data. Another example is the almost full closure of any form of access to Facebook data following the Cambridge Analytica scandal (Bruns et al., 2018, Tromble in this special issue). Returning to our notion of visibility as a socio-technical process that is not stable but in a constant relational making, this also means that new invisibilities and visibilities are constantly introduced. Changes in APIs and company policies can make data accessible in a different form (that requires a new type of analysis), incomplete or restrict access entirely.

Including the dimension of accessibility and form into our conceptual understanding of visibility processes in social media data has two major consequences. First, it describes a dependency of the researchers on access to data in a particular form given by social media corporations. Second, understanding the visibility and invisibilities introduced by changes and the instabilities in terms of access might urge researchers and analysists to move toward methods that go beyond very specific forms of data and access to data from a specific platform.

Meaning and Imaginaries

Once data have been collected, organized, and made accessible by the platform, researchers can analyze them. The dimension of meaning and imaginaries is concerned with what ultimately constitutes research data. Previous studies in the field of digital media (Burgess & Bruns, 2012; Rettberg, 2008) show that there is a direct connection between how the data are made available from the social media platform and how they are classified and analyzed by researchers. To name two examples of such classifications: Returning users' social relations as some sort of connected graph will lead the researchers toward network analysis methods, while content returned as a list of well-organized and itemized textual corpora will steer the researchers toward methods based on natural language processing. This resonates with trends in research methods and platforms of interest (Neumayer & Rossi, 2016). In parallel, imaginaries about social phenomena are introduced alongside technical infrastructures such as Manuel Castells' famous notion of the "network society" (Castells, 2011).

Widely adopted methods in conjunction with the data structure offered by the APIs make visible what otherwise would remain invisible. A prominent example of this is the growing use of network methods to study digital data (P. Jensen et al., 2016). Network analysis is often used as a method to identify communities (or clusters) based on social media data (Bruns et al., 2014; Effing et al., 2011). Using this method to analyze social media data by identifying communities is based on the assumptions that the digital traces left by the users provide information about their individual characteristics (e.g., on which side of an online controversy they position themselves), and that they provide information about their role in more global dynamics (e.g., who acted as an information gatekeeper). With very few exceptions, contemporary community detection methods try to identify clusters based on the assumption that connections between users of the same group will be more common than connections between users belonging to different groups (Capocci et al., 2005). While the algorithmic means to measure this can vary, the underlying concept is stable (Javed et al., 2018) and strongly resonates with our common sense and everyday experiences. However, the "detection" of these structures based on digital data can produce information that is largely unknown to the user, who is supposedly a member of such communities. The structure exists outside of his or her personal experience or control.

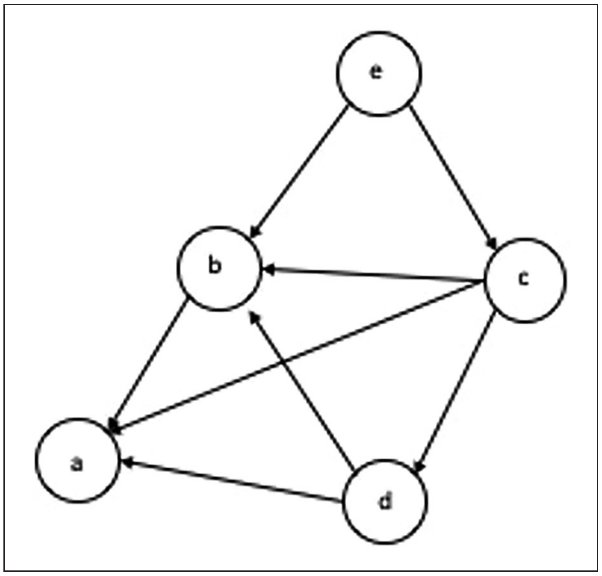

Figure 2. An example of a retweet network. All users will be identified as belonging to the same cluster even if they might be unaware of the existence of a and d.

An illustrative example: A network of retweets (see Figure 2) is commonly used in computational social media research. In such a network users *a, b, c, d*, and *e* are connected with a directed edge if they retweet each other. In our example, *c* has retweeted *b* and *d* and *a*, while *e* only retweeted *b* and *c*. Within this approach, modularity optimization community detection algorithms are widely used to identify a single cluster (or community) assigning *e* to the same community as *a* and *d*. This will happen even if *e* never directly interacted with them and (to the best of our knowledge) user *e* has never seen their tweets or is aware of their existence. Nevertheless, users will become members of a clusters in the analysis, and one or more clusters will be detected. While these are well-known problems in social network analysis (Lancichinetti & Fortunato, 2011), this showcases the limits and raises questions about the effectiveness of such methods. Moreover, the example shows how many of the computational tools we use for processing and analyzing social digital data make structures visible that might not be visible otherwise.

Including the dimension of meaning and imaginaries into our conceptual understanding of visibility processes in social media data has two major consequences. Research methods are developed based on the assumption that meaning ascribed to such data makes social interaction visible, and that we only need access to all the data and employ appropriate analysis methods to understand the social meaning underlying such data. Yet, there is a process in place that makes data visible and invisible, and the researchers and analysts take an active part in such processes. They do so by ascribing meaning to such data and by introducing or stabilizing imaginaries that reinforce such visibility and invisibility processes.

Connecting the Dots: Discussing Data Invisibilities

To summarize, we argued that as data are made, visibilities and invisibilities are created within four dimensions: people and intentionality, technologies and tools, accessibility and form, and meaning and imaginaries. We also argued that there is a liminal space between visible and invisible (these are not stable dichotomies) that we describe with the notion of quasi-visible data. To understand how invisibilities are introduced into studies relying on social media data, we traced the process starting when people with particular intentionality in a specific social context create data, which are then traced and archived by social media platforms, made accessible in particular forms, and are analyzed or ascribed meaning to by researchers and analysts based on particular imaginaries. Visible and invisible are not dichotomous, and we find it useful to add the notion of *quasi-visible data*. Quasi-invisible data are only visible to some, that is, those who have access (such as social media corporations or some researchers) or can be made visible by performing a particular type of (computational) analysis.

The notion that data move between visible, quasi-visible, and invisible within the four socio-technical dimensions of people and intentionality, technologies and tools, accessibility and people, and meaning and imaginaries has three major consequences for research based on social media data. First, it leads us to asking ethical questions about what we make visible as researchers when working with such data. Do we need consent if we make personal information, hidden strategies or data visible outside of the control of the user? To what extent are we able to take into account (over)visibilities as well as invisibilities intentionally created by the users? And under which circumstances can and should we prioritize such intentionality? How do we comply with the recently implemented European General Data Protection Regulation (GDPR) that assumes a fully informed data subject (Kotsios et al., 2019)? How do we live up to the expectation that users need to be aware of what data they share, how their data are processed and with what purpose? And how do we do so when research includes data intentionally kept invisible? Second, it leads to practical and methodological questions. Can we take into account (in)visibilities with the methods and tools that we have? How can we develop methods and tools that are better at considering such (in)visibilities? If we are uncertain about (in)visibilities, how should we frame our results? Are results that cannot account for (in)visibilities reliable? How can we practically work with data (in)visibilities? And when do we need to combine computational methods with qualitative work (such as ethnography) to take into account (in)visibilities? What are the limitations of social media data if we acknowledge invisibilities? And third, conceptualizing (in)visibilities opens up for questions about the politics of research. How do we present research that we know is based on partial data? Can we combine visible and invisible data? What does it mean to acknowledge that we always only work with a small part of what are possible lines of inquiry? To what extent should we let social media corporations and the way platforms provide data guide our research agenda and the questions we ask?

While this article is not an attempt to answer these questions, it rather invites us to reflect upon them. Understanding data invisibilities in the research process when working with social media data might help researchers develop a more responsible approach by reflecting upon the (in)visibilities created, especially if the analysis makes visible what people intentionally tried to hide. Addressing the question of data invisibilities might also reveal power relations within which such processes take place. In other words, by understanding what remains and what must remain invisible, we can better understand what type of analysis we can perform with social media data but also the limitations of our results. While the dimensions discussed in this article are not exhaustive, they are an invitation for using data invisibilities as a sensitizing concept to reflect upon our research and analysis process when relying on social media data and to eventually move toward a more reflective field of research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Velux Foundation [grant number 12823], *Data as Relation: Governance in the Age of Big Data*.

ORCID iD

Christina Neumayer https://orcid.org/0000-0003-0450-2983

References

Andrejevic, M. (2013). *Infoglut: How too much information is changing the way we think and know*. Routledge.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61.

Baym, N. K. (2018). Data not seen: The uses and shortcomings of social media metrics. *First Monday*, *18*(10), 1–15. https://doi.org/10.5210/fm.v18i10.4873

Belair-Gagnon, V., Agur, C., & Frisch, N. (2018). Mobile sourcing: A case study of journalistic norms and usage of chat apps. *Mobile Media & Communication*, *6*(1), 53–70.

Bernstein, M. S., Bakshy, E., Burke, M., & Karrer, B. (2013, April). Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 21–30). ACM.

Blaagaard, B., Mortensen, M., & Neumayer, C. (2017). Digital images and globalized conflict. *Media, Culture & Society*, *39*(8), 1111–1121.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Brighenti, A. M. (2007). Visibility: A category for the social sciences. *Current Sociology*, *55*(3), 323–342.

Brighenti, A. M. (2010a). *Visibility in social theory and social research*. Palgrave Macmillan.

Brighenti, A. M. (2010b). Democracy and its visibilities. In: K. D. Haggerty & M. M. Samatas. (Eds,), *Surveillance and democracy* (pp. 67–84). Routledge.

Bruns, A., Bechmann, A., Burgess, J., Chadwick, A., Clark, L. S., Dutton, W. H., & Howard, P. (2018). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*. https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

Bruns, A., Highfield, T., & Burgess, J. (2014). The Arab Spring and its social media audiences: English and Arabic Twitter users and their networks. In M. McCaughey (Ed.), *Cyberactivism on the participatory web* (pp. 96–128). Routledge.

Brunton, F., & Nissenbaum, H. (2015). *Obfuscation: A user's guide for privacy and protest*. MIT Press.

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of "Big Social Data" for media and communication research. *M/C Journal*, *15*(5), 1–7. https://doi.org/10.5204/mcj.561

Capocci, A., Servedio, V. D., Caldarelli, G., & Colaiori, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and Its Applications*, *352*(2–4), 669–676.

Castells, M. (2011). *The rise of the network society* (Vol. 12). Wiley-Blackwell.

Chouliaraki, L., & Stolic, T. (2017). Rethinking media responsibility in the refugee 'crisis': A visual typology of European news. *Media, Culture & Society*, *39*(8), 1162–1177.

Croeser, S., & Highfield, T. (2015). Mapping movements—Social movement research and big data: Critiques and alternatives. In G. Langlois, J. Redden, & G. Elmer (Eds.), *Compromised data: From social media to big data* (pp. 173–201). Bloomsbury.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016, April 11–15). BotOrNot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273–274). International World Wide Web Conferences Steering Committee.

Dourish, P. (2017). *The stuff of bits: An essay on the materialities of information*. MIT Press.

Earl, J. (2018). The promise and pitfalls of big data and computational studies of politics. *Partecipazione E Conflitto*, *11*(2), 484–496.

Effing, R., Van Hillegersberg, J., & Huibers, T. (2011, August). Social media and political participation: Are Facebook, Twitter and YouTube democratizing our political systems? In *International Conference on Electronic Participation* (pp. 25–35). Springer.

Flyverbom, M., Leonardi, P., Stohl, C., & Stohl, M. (2016). The management of visibilities in the digital age—Introduction. *International Journal of Communication*, *10*(12), 98–109.

Foucault, M. (2003). *The archaeology of knowledge*. Routledge. (Original work published 1969)

Fuchs, C. (2015). *Culture and economy in the age of social media*. Routledge.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.

Garforth, L. (2012). In/visibilities of research: Seeing and knowing in STS. *Science, Technology, & Human Values*, *37*(2), 264–285.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, 31(6), 649–679.

Giglietto, F., Rossi, L., & Bennato, D. (2012). The open laboratory: Limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. *Journal of Technology in Human Services*, 30(3–4), 145–159.

Gillespie, T., Boczkowski, P. J., & Foot, K. A. (Eds.). (2014). *Media technologies: Essays on communication, materiality, and society*. MIT Press.

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. MIT press.

Harada, J., Darmon, D., Girvan, M., & Rand, W. (2017). Prediction of elevated activity in online social media using aggregated and individualized models. In R. Missaoui, T. Abdessalem, & M. Latapy (Eds.). *Trends in social network analysis* (pp. 169–187). Springer.

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*, 10–24. https://doi.org/10.1177/0894439318788322

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 1–7. https://doi.org/10.1177/2053951716674238

Javed, M. A., Younis, M. S., Latif, S., Qadir, J., & Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, *108*, 87–111.

Jensen, K. B. (2012). Lost, found, and made: Qualitative data in the study of three-step flows of communication. In I. Volkmer (Ed.), *The handbook of global media research* (pp. 435–450). Wiley-Blackwell.

Jensen, M. S., Neumayer, C., & Rossi, L. (2020). 'Brussels will land on its feet like a cat': Motivations for memefying# Brusselslockdown. *Information, Communication & Society*, *23*(1), 59–75.

Jensen, P., Morini, M., Karsai, M., Venturini, T., Vespignani, A., Jacomy, M., Cointet, J. P., Mercklé, P., & Fleury, E. (2016). Detecting global bridges in networks. *Journal of Complex Networks*, *4*(3), 319–329.

Kitchin, R. (2015). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE.

Kotsios, A., Magnani, M., Vega, D., Rossi, L., & Shklovski, I. (2019). An analysis of the consequences of the general data protection regulation on social network research. *ACM Transactions on Social Computing*, *2*(3), 1–22.

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E*, *84*(6), Article 066122.

Lange, P. G. (2007). Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication*, *13*(1), 361–380.

Langlois, G., & Elmer, G. (2013). The research politics of social media platforms. *Culture Machine*, *14*, 1–17. http://svr91.edns1.com/~culturem/index.php/cm/article/download/505/531

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., & Jebara, T. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science*, *1*(1), 49–58.

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: How social media shape collective action*. Princeton University Press.

Mattoni, A., & Pavan, E. (2018). Politics, participation and big data. Introductory reflections on the ontological, epistemological, and methodological aspects of a complex relationship. *Partecipazione e Conflitto*, *11*(2), 313–331.

Mejias, U. A. (2012). FCJ-147 liberation technology and the Arab Spring: From utopia to atopia and beyond. *The Fibreculture Journal*, *20*, 204–217.

Neumayer, C., & Rossi, L. (2016). 15 years of protest and media technologies scholarship: A sociotechnical timeline. *Social Media + Society*, *2*(3), 1–13. https://doi.org/10.1177/2056305116662180

Neumayer, C., & Rossi, L. (2018). Images of protest in social media: Struggle over visibility and visual narratives. *New Media & Society*, *20*(11), 4293–4310.

Neumayer, C., & Struthers, D. M. (2019). Social media as activist archives. In: M. Mortensen, C. Neumayer, & T. Poell (Eds.), *Social media materialities and protest: Critical Reflections* (pp. 86–98). Routledge.

Petrovic, S., Osborne, M., & Lavrenko, V. (2011, July). Rt to win! predicting message propagation in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (pp. 586–589). The AAAI Press.

Rettberg, J. W. (2008). *Blogging*. Polity.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August 20–22). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of the 2010 IEEE 2nd International Conference on Social Computing* (pp. 177–184). IEEE Computer Society.

Triggs, A. H., Møller, K., & Neumayer, C. (2019). Context collapse and anonymity among queer Reddit users. *New Media & Society*. Advance online publication. https://doi.org/10.1177/1461444819890353

Trouillot, M. R. (1995). *Silencing the past: Power and the production of history*. Beacon Press.

Uldam, J. (2016). Corporate management of visibility and the fantasy of the post-political: Social media and surveillance. *New Media & Society*, *18*(2), 201–219.

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Societ*, *22*, 1531–1543. https://doi.org/10.1080/1369118X.2019.1648536

Weller, K. (2015). Accepting the challenges of social media research. *Online Information Review*, *39*(3), 281–289.

Yang, J., & Counts, S. (2010, May 23–26). Predicting the speed, scale, and range of information diffusion in Twitter. In *Proceedings of the 4th International Conference on Weblogs and Social Media* (pp. 355–358). Association for the Advancement of Artificial Intelligence.

Author Biographies

Christina Neumayer is associate professor of Media Studies at the Department of Communication at the University of Copenhagen. Her research focuses on the role of digital media technologies, platforms and data for political contention, protest, activism, racism, social movements, and more broadly political communication. Her most recent publications include the volume *Social Media Materialities and Protest: Critical Reflections* (Routledge, 2019) co-edited with Mette Mortensen and Thomas Poell.

Luca Rossi is associate professor of Digital Media and Networks at the Department of Digital Design of IT University of Copenhagen. He is member of the Networks Data and Society (NERDS) research group and of the Digital Platforms and Data research group. He also coordinates the Data Science & Society research lab. His interdisciplinary research tries to connect traditional sociological theory with computational approaches. Within this line of research, he has worked and published in the context of online participation, online activism, political campaign, and election studies.

David M. Struthers is a historian with interdisciplinary pursuits. His monograph *The World in a City: Multiethnic Radicalism in Early Twentieth Century Los Angeles* (University of Illinois Press) explores the dynamics of a city and region with a rapidly developing economy and large-scale immigration. His volume *Wobblies of the World: A Global History of the IWW* (Pluto Press, 2017), edited with Peter Cole and Kenyon Zimmer, is the first global history of the Industrial Workers of the World. He is currently working on a monograph on the history of activist social media before the internet.