Making Sense of Small and Big Data as Onlife Traces

Anja Bechmann, Kjetil Sandvik† & Karin Zelano

Two main societal tendencies have dominated the field of media and communication over the last ten years in different ways. First, ubiquitous internet services and devices have led to more seamless media use in everyday life and subsequently to an increase in human communication, behavioural and interaction data. Additionally, this data increase is accelerating the developments in applying mathematical models such as machine learning and artificial intelligence (AI) models to process, cluster, correlate and make sense of such data, for instance, with the purpose of predicting future usage patterns (Russell & Norvig, 2009). Second, there is a concentration of power in increasingly global, distributed and commercial digital intermediaries such as social media and other app services that hold and gatekeep user data in private companies (Bruns et al., 2018; Langlois, 2015).

These tendencies have created associated challenges when it comes to understanding how humans live, work and are represented through data. Another challenge arises in the way massive power concentration is governed when commercial interests collide with human and societal values at a global scale with regards to issues such as transparency, privacy, antidiscrimination, freedom of expression and democracy (see e.g., Bechmann & Bowker, 2019; Howard, 2005; Nissenbaum, 2009; Zuboff, 2019). The ways in which the processes of decision-making are accounted for and governed when using seemingly unstable and autonomous mathematical models as the backbone of the data-driven society presents a further challenge.

Media and communication research have a long history and interest in understanding user behaviour, incentives, beliefs and opinions through surveys, observations and interviews. However, the character of the existing communication and media landscape provides new possibilities for analysing a larger amount of data than ever before over time, across different spaces and places and in larger user (social and collective) networks. However, such data still create a need for a holistic approach to usage as a contextual challenge in anchoring big data (wide and deep log data) both in a physical situation and in a situation that involves other platforms, services and/or users, just to name a few contextual challenges (Leckner & Severson, 2019; Lai et al., 2019). This is especially important when big data are used to predict behaviour such

as fertility prevention (Karlsson, 2019) or to profile people through correlated behaviours and attempts to intervene in election results based on carefully tailored content placed through social media micro-targeting (Eg & Krumsvik, 2019). Meaningful data in such scenarios are both important and potentially a threat to users and society. Nonetheless, the classical question of meaning and meaningful data deserves renewed attention in our research field in the light of societal and technological changes. When media usage becomes more seamless integrating various forms of physical and digital interwoven lives in big data studies becomes central to create meaningful data (boyd & Crawford, 2012; Mayer-Schönberger & Cukier, 2013). At the same time new types of small n-studies arise along with new theoretical and methodological considerations in classical qualitative studies.

This special issue gathers a number of articles that study onlife empirically, as an integration of the physical and digital life (Floridi, 2015; Simon & Ess, 2015) and will discuss the methods used to infer meaning from data traces to usage and users. In doing so, the issue will build on classical battles and paradigm shifts from the idea of separated digital and physical lives to integrated spheres (Baym, 1995; Markham, 1998; Turkle, 1995) and from single methods to the introduction of triangulation of data points and multi-sited ethnography (Marcus, 1995). Onlife designates the transformational reality that, in contemporary developed societies, our offline and online experiences and lives are inextricably interwoven. Our onlives produce digital traces or footprints, some of which are even produced before birth by our parents and continue to exist after our death (in the shape of registers, bank accounts, social media profiles, etc.). This special issue examines and discusses how we create meaning in and make sense of small and big data as onlife traces.

Anchored in media and communication research, the contributors use and discuss methods to extract meaningful sociological onlife findings from small and big data and discuss the methodological challenges in doing so. The notions "small" and "big" data are problematic in many ways, but in this volume "big data" connotes bigger than usual data (Boellstorff, 2013; Kitchin, 2014; Mayer-Schönberger & Cukier, 2013) and "small data" refers to in-depth and thick accounts of usage or particular contextualized users (Geertz, 1973; Kitchin & Lauriault, 2015). One of the problems of the small and big data discussion is that it replicates a classical distinction between researchers being primarily either quantitative or qualitative in their methodological approach (Leckner & Severson, 2019; Munk, 2019).

The fields of qualitative and quantitative research are often depicted as each other's counterpart even though the mixed-method design has existed for a very long time. Nevertheless, when researchers study onlife as a way of conceptualizing the digital (and physical) layer of our lives, the notions of quantitative and qualitative methods are challenged. Many of the contributions in this volume show how combinations of quantitative and qualitative methods are applied to fertilize the knowledge of usage or users. The quantitative approaches are not merely isolated to large n survey studies and statistical analysis, and the qualitative approaches are not restricted to small n and ethnographically inspired collection methods, such as field diaries, interviews, focus groups and analytical methods such as discourse and narrative analysis. Instead, the special issue illustrates reinterpreted *hybrid method* approaches. How large n log data samples can be used to narrow down and focus on carefully selected smaller n samples.

This sample method creates new opportunities for justifying the specific selection. At the same time, it exemplifies how classical qualitative studies can benefit from the distributed network structure of the online (Lai et al., 2019). The issue further exemplifies how designated apps demands new ephemeral methods (Møller & Robards, 2019). In the context of big data the contributions in this issue show how subjective selections, data cleaning and interpretative labelling of machine learning cluster outputs take place in statistical studies. This results in pinpointing the "qualitative"/interpretative elements in otherwise classical big data studies that can have profound consequences for the outcome of the data processing (e.g. Bruns & Moon, 2019; Elgesam, 2019; Bechmann, 2019; Munk, 2019; Rosales & Fernandez-Ardévol, 2019). In this way, onlife not only adds new methodological combinations to the media and communication field, but in doing so, also highlights how old distinctions and practices of methodological combinations are reinterpreted. One of the largest contributions in this methodological development is an invigorating interest in how meaning can be inferred from data in new triangulated interpretations and/or close data observations that draw on classical discussions yet contribute something new because it is applied to a completely new field, namely that of the onlife.

In the onlife domain, one of the major questions (again) is how can we infer meaning from the digital traces made by the user to the actual use or the human(s) behind (incentives, motives and needs)? This is a classical methodological question within literary studies, media communication and behavioural research about what texts/signs refer to and how the meaning and sense making are generated in a combination between text/signs, users and the surrounding cultural context in, for instance, linguistics (Jakobson, 1995), cultural studies (Hall, 1980) and semiotics (Peirce, 1934). This sense making has a renewed interest in the digital social sciences as we explore different methodological trajectories into the onlife, such as studies of the Internet of Things, apps (e.g. social media, games and self-trackers) or other forms of digital communication and behaviours as traces of digital sociology (Marres, 2017).

Both small and big data study approaches have tried to "solve" this problem of inference by suggesting triangulation or a similar methodological approach. Within big data studies, more data on more users, multiple data points or data on the same user over a longer time span are used to create "clearer signals" and to strengthen the predictions of specific user behaviours, motives, ideologies, incentives and needs. Qualitative studies use, for instance, multi-sited ethnography and methodological triangulation to heighten the validity of findings when it comes to understanding clearly the user and the use behind onlife traces.

User data have become sharable and tradable commodities and governments, media, health and financial sectors build actions and decisions on top of predictions inferred from data traces of human communication and behaviour in digital spaces. Therefore, it is perhaps more important than ever to understand and discuss not whether we can make a solid one-to-one interpretation but how we can advance our inferences. At least explicitly discuss in various studies how we have solved or coped with the issue of inference by advancing our research questions, methodological approaches, philosophical background and conceptual understanding.

The articles in this issue

This special issue compiles 11 articles under the theme of *Making sense of small and big data as onlife traces*, as introduced here. The first group of articles uses large data sets of blog and social media communication and log data to infer meaning from and critically discuss how meaning can be inferred from big data.

Bruns and Moon (2019) use an extensive data set to address the imbalance in most studies of social media platforms, namely that of studying only hashtagged content and not mundane everyday use. Specifically, the study draws on a comprehensive data set that tracks the public activities of all Australian Twitter accounts during a 24-hour period in March 2017.

Bechmann (2019) investigates the Facebook posting behaviour of 922 posting users over a time span of seven years (from 2007 to 2014), using an innovative combination of survey data and private profile feed post counts obtained through the Facebook API to understand the differences in posting behaviour of demographic groups.

Rosales and Fernandez-Ardévol (2019) undertake a critical review of the academic literature on big data and highlight how structural ageism can be found in such studies.

Elgesem (2019) explore the potential and challenges of using hyperlinks as data through a study of polarization in blogs about climate change. The article finds that the bloggers in the sample predominantly link to sources that they agree with and, if they link to a source with different opinions, the link is part of negative criticism of the targeted source.

The second group of articles attempts to make sense of data by using ethnographic methods to infer meaning from onlife traces from ICTs and associated usage.

Leckner and Severson (2019) propose to use digital method triangulation and exemplify the benefits of two cases: digital focus groups and a combination of internet traffic measurements, surveys and diaries.

Møller and Robards (2019) propose to study "ways of being with media", which they call an ethnography of ephemeral mobilities that analytically integrate four dimensions: bodies and affect, media objects and environments, memory and narrative, and the overall research encounter.

Karlsson (2019) uses in-depth individual interviews to find out more about how the use of a specific type of self-tracking apps (period trackers) confers meaning on the everyday lives of users and reproduces traditional feminine cultures such as privacy and shaming.

Saariketo (2019) also studies the use of self-monitoring, but more broadly, and applies a combination of interviews and media diaries with log data prompts to make the participants reflect on their use and the meaning of their digital bodies.

Lai, Pagh and Zeng (2019) argue in favour of a comparative ethnography of communication that emphasizes the study of intermediality by taking a people-centred approach. More specifically, the research design combines network sampling and maximum variation sampling with communication diaries and elicitation interviews.

Munk (2019) explores the meaning problem in data studies by using a web corpus (a collection of websites connected by their hyperlinks) that mapped the so-called New Nordic Food Movement in Scandinavia. The collection of hyperlinks and word clusters, he argues, are onlife traces that tell us something about who these actors are associating with, what they are talking about and when they are doing so. However, they do not

reveal the meaning of those actions to us. In response, Munk suggests four qualitative methods to make sense of these onlife traces by adding a qualitative component to the existing (quantitative) data.

Eg and Krumsvik (2019) in turn apply an experimental set-up to test whether personality traits influence news engagement online.

Beyond this straightforward methodological division of articles into big data and small data studies, the articles are interlinked thematically. Hence, two or more articles address one of the following four dimensions: mundane usage, smartphone and app usage, civic engagement and democracy, and demographic biases.

Bruns and Moon (2019) note that much of the existing research into the use of social media platforms focuses on the exceptional, while everyday social media practices remain comparatively underexamined. Instead of examining hashtags on Twitter, they investigate the entire Australian Twittersphere. Similarly focusing on the mundane, Saariketo (2019) investigates the experiences gained from self-monitoring and the encounters with tracked data by self-identified avid ICT users, and Karlsson (2019) maps the reasons for using period trackers among a group of Danish women.

The smartphone is perhaps the prime representative of onlife. In 2018, the British Office of Communications reported that the average person in the UK spends more than a day a week online and checks his or her smartphone every 12 minutes (Ofcom, 2018). As a consequence, the study of onlife implies the study of smartphone use in one way or another, and three of the articles explicitly deal with smartphone app usage in everyday life. The women interviewed in Karlsson's (2019) study experience period tracker apps as private, shame-free rooms for exploratory engagement with the menstruating body, and the risk of embodied data potentially becoming shareable commodities does not affect the everyday self-tracking practice of these women. Møller and Robards' (2019) article has a purely methodological focus on app usage and considers three methods in qualitative "small data" social media research: the walkthrough, the go-along and the scroll back methods – all related to the use of smartphone app use. In a sense, Saariketo (2019) takes the inquiry of app use even further by elaborating on the ways in which research interventions may repurpose the means of datafication and create possibilities for people to reflect on what it means in their daily lives. Although people and not apps are the centre of Lai, Pagh and Zeng (2019) work, the article illustrates a people's perspective in which apps such as Facetime, WeChat and Youtube appear when investigating communication from a holistic point of view.

Three of the articles explicitly explore the onlife from a civic engagement or democracy perspective, investigating how studying the onlife shows ways of deliberation and participation in modern society. Elgesem's (2019) analysis of blogs is an example of such civic engagement analysis. By considering the functions of the links in the blog posts, we gain a more nuanced understanding of the nature of online discussions. Against the background of recent revelations about how individual user data are used to target political information, Eg and Krumvik (2019) examine the connection between personality and news reading. Despite small effects, the study finds that informative stories engage more extroverted and rational participants while participants with experiential behaviour and information processing are less likely to engage. Munk's (2019) article deals with another kind of civic engagement, namely that of social movements. Munk investigates Nordic Food networks with a focus on local produce and communities around a specific

manifesto encouraging Nordic food. Using the methods of network and cluster analysis as a form of algorithmic sense making, Munk shows how Nordic food communities map out both at scale and in depth.

The final thematic dimension is demographic biases in big data studies, addressed explicitly in the two articles by Rosales & Fernandez-Ardevól (2019) and Bechmann (2019). Rosales & Fernandez-Ardevól (2019) reveal how data that are treated as neutral/unbiased in studies in fact enforce and perpetuate discrimination of older individuals. Specifically, they show how biased samples and biased tools tend to exclude data of older people from the studies and their interests or values from the algorithms, which contributes to reinforcing structural ageism. Bechmann (2019) emphasizes the need to count in time when analysing behaviour online and shows how posting behaviour on Facebook varies between socio-demographic groups and over time.

Looking ahead, moving forward

In conclusion, the special issue makes valuable empirical and methodological contributions to the discussion on sense making in the framework of understanding digital traces from a sociological perspective. The special issue does not claim to *solve* the "sense-making problem", but we hope that the special issue and the many methodological discussions and empirical explorations point out further directions for research in the field of digital sociology.

• • •

This special issue is dedicated to the memory of Kjetil Sandvik and his work in the field of media and communication research.

References

Baym, N. K. (1995). The emergence of community in computer-mediated communication. In S. Jones (eds.) *CyberSociety: Computer-mediated communication and community* (pp. 138-163). Thousand Oaks, CA: Sage Publications, Inc.

Bechmann, A. (2019). Inequality in posting behaviour over time: A study of Danish Facebook users. *Nordicom Review*, 40 (Special Issue 1): 31-49.

Bechmann, A. & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, 6(1): 1-11. doi: https://doi.org/10.1177/2053951718819569

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, 18(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4869, accessed [2019, May 23].

Boyd, D. & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5): 662-679. doi: https://doi.org/10.1080/1369118X.2012.678878

Bruns, A. & Moon, B. (2019). One day in the life of a national Twittersphere. *Nordicom Review*, 40 (Special Issue 1): 11-30.

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

Eg, R. & Krumsvik, A. (2019). Personality filters for online news interest and engagement. *Nordicom Review*, 40 (Special Issue1): 177-194.

Elgesem, D. (2019). The meaning of links: On the interpretation of hyperlinks in the study of polarization in blogging about climate change. *Nordicom Review*, 40 (Special Issue 1): 65-78.

Floridi, L. (2015). *The online manifesto: Being human in a hyperconnected era*. Cham: Springer.

Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books, HarperCollins Publishers.

Hall, S. (1980). Encoding/decoding. In S. Hall (ed.), *Culture, media, language* (pp. 128-138). London: Hutchinson.

Howard, P. N. (2005). *New media campaigns and the managed citizen*. Cambridge: Cambridge University Press.

Jakobson, R. (1995). *On language*. Cambridge: Harvard University Press.

Karlsson, A. (2019). A room of one's own? Using period trackers to escape menstrual stigma. *Nordicom Review*, 40 (Special Issue 1): 111-123.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Kitchin, R. & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, 80(4): 463-475. doi: https://doi.org/10.1007/s10708-014-9601-7

Lai, S. S., Pagh, J. & Zeng, F.H. (2019). Tracing communicative patterns: A comparative ethnography across platforms, media and contexts. *Nordicom Review*, 40 (Special Issue 1): 141-157.

Langlois, G. (2015). What are the stakes in doing critical research on social media platforms? *Social Media + Society*, 1(1), https://doi.org/10.1177/2056305115591178.

Leckner, S. & Severson, P. (2019). Exploring the meaning problem of big and small data through digital method triangulation. *Nordicom Review*, 40 (Special Issue 1): 79-94.

Marcus, G. E. (1995). Ethnography in/of the world system: The emergence of multi-sited ethnography. *Annual Review of Anthropology*, 24: 95-117.

Markham, A. N. (1998). *Life online: Researching real experience in virtual space* (Vol. 6). Plymouth: Altamira.

Marres, N. (2017). *Digital sociology: The reinvention of social research*. Cambridge: Polity Press.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Munk, A.K. (2019). Four styles of quali-quantitative analysis: Making sense of the new Nordic food movement on the web. *Nordicom Review*, 40 (Special Issue 1): 159-176.

Møller, K & Robards, B. (2019). Walking through, going along and scrolling back: Ephemeral mobilities in digital ethnography. *Nordicom Review*, 40 (Special Issue1): 95-109.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Palo Alto, CA: Stanford University Press.

Ofcom. (2018). *Communications market report 2018*. London: Ofcom. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf

Peirce, C. S. (1934). *Collected papers* (Vol. IV). Cambridge: Harvard University Press.

Rosales, A. & Fernandez-Ardevól (2019). Structural ageism in big data approaches. *Nordicom Review*, 40 (Special Issue 1): 51-64.

Russell, S. & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River, CA: Pearson.

Saariketo, M (2019). Encounters with self-monitoring data on ICT use. *Nordicom Review*, 40 (Special Issue 1): 125-140.

Simon, J. & Ess, C. (2015). The ONLIFE initiative – A concept reengineering exercise. *Philosophy and Technology*, 28(1): 157-162.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon & Schuster.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Hachette Book Group.