Broken data and unexpected research questions

Minna Ruckenstein, University of Helsinki

Recent research introduces the concept-metaphor of "broken data", suggesting that digital data might be broken and fail to perform, or be in need of repair (Pink et al 2018). Concept-metaphors, anthropologist Henrietta Moore (1999, 16; see also Moore 2004) argues, are domain terms that "open up spaces in which their meanings – in daily practice, in local discourses and in academic theorizing – can be interrogated." By doing so, concept-metaphors become defined contextually in practice; they are not meant to be foundational concepts, but work as partial and perspectival framing devices.

In this presentation, the concept-metaphor of broken data is discussed in relation to the open data initiative, *Citizen Mindscapes*, an interdisciplinary project that contextualizes and explores a Finnish-language social media data set ('*Suomi24*', or Finland24 in English), consisting of tens of millions of messages and covering social media over a time span of 15 years (see, Lagus et al 2016). The aim of taking advantage of a concept-metaphor in a data-related study is to arrange and provoke ideas and open a conceptual domain within which facts, connections and relationships are identified and imagined. The role of the broken data metaphor in this discussion is to examine the implications of breakages and consequent repair work in data-driven initiatives that take advantage of secondary data. Moreover, the concept-metaphor can sensitize us to consider the less secure and ambivalent aspects of data work. By focusing on how data might be broken, we can highlight misalignments between people, devices and data infrastructures, or bring to the fore the failures to align data sources or data uses with the everyday.

As Pink et al (2018) suggest the metaphorical understanding of digital data, aiming to underline aspects of data brokenness, brings together various strands of scholarly work, highlighting important continuities with earlier research. Studies of material culture explore practices of breakage and repair in relation to the materiality of objects, for instance by focusing on art restoration (Dominguez Rubio 2016), or car repair (Dant 2010). Drawing attention to the fragility of objects and temporal decay, these studies underline that objects break and have to be mended and restored. When these insights are brought into the field of data studies, the materiality of databases,

platforms and software become a concern (Tanweer et al 2016), emphasizing aspects of brokenness and following repair work in relation to digital data (Pink et al 2018).

In the science and technology studies (STS), on the other hand, the focus on 'breakages' has been studied in relation to infrastructures, demonstrating that it is through instances of breakdown that structures and objects, which have become invisible to us in the everyday, gain a new kind of visibility. The STS scholar Stephen Jackson expands the notion of brokenness to more everyday situations and asks 'what happens when we take erosion, breakdown, and decay, rather than novelty, growth, and progress, as our starting points in thinking through the nature, use, and effects of information technology and new media?' (2014: 174). Instances of data breakages can be seen in light of mundane data arrangements, as a recurring feature of data work rather than an exceptional event (Pink et al 2018; Tanweer et al 2016).

In order to concretize further the usefulness of the concept-metaphor of broken data, I will describe how identifying instances of breakage in the data set and repair in the data work can generate new and unanticipated research questions. In particular, I will highlight the role of spam bots, computer programs specifically designed for generating spam messages, in digital work. In the collaborative Citizen Mindscapes initiative, discussing the gaps, or possible anomalies in the data led to conversations concerning the production of data, deepening our understanding of the human and material factors at play in processes of data generation. As described below, the waste of the data world, spam messages, could also be seen as a resource in terms of everyday digital innovation (Tanweer et al 2016).

Working with spam

The Suomi24 data was generated by the media company, Aller; the data silently resided on the servers until the company decided to open the proprietary data for research purposes (see Lagus et al 2016). In the past two years, the *Citizen Mindscapes* -initiative, particularly researchers experienced in working with large data sets, have been cleaning the data in order to make it ready for computational work. The aim is to build a methodological toolbox that researchers, who do not possess computational skills, but are interested in using digital methods in the social scientific inquiry, can benefit from. This entails, for instance, developing user interfaces that narrow down the huge data set and allow to access the data with topic-led perspectives.

The ongoing work has alerted the research collective to breakages of data, raising more general questions about the origins and nature of data (Pink et al 2018). The research report that details and contextualizes the Suomi24 data pays attention to the writers of the social media community as producers of the data; the moderation practices of the company are described to demonstrate how they shape the data set by filtering words and terms, or certain kinds of messages, for instance, advertisement or messages containing sensitive personal information (Lagus et al 2016). When the data work identifies gaps, errors and anomalies in the data, it reveals that data might be broken and discontinuous due to human or technological forces: infrastructure failures, trolling, or automated spam bots.

We have repeatedly used the visual information of gaps in the data (see Figure 1) as a conversation opener with the social media company's employees. We learned that the 2004-2005 is probably a technical error in the database retrieval. The anomaly in the data volume in July 2009 was first defined as a spam bot by the employees (Pink at al, 2018). Later, however, one of the moderators of the company suspected that it could not have been a spam bot after all. The data set was not supposed to contain spam in such quantities, because the data was already cleaned by the programmers.

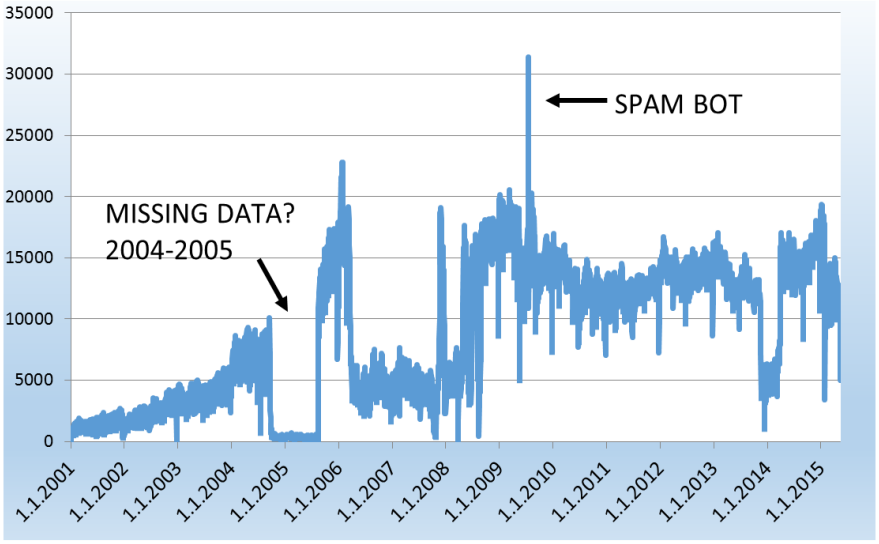

Figure 1: Identified gaps and breakages in the Suomi24-data

In January 2018, we started exploring the July 2009 peak in a more consistent manner. A whole new area of research questions started to emerge about automation and spam bots. Importantly, the

spam bots have an online agency of their own; bots are searching for wikis, blogs and forums that they can use to submit spam. In the Suomi24 forum, the spam messages involve techniques of targeted advertisement, and at times it might be hard to tell human-generated posts apart from automated ones. Other spam messages are not meant to be read by humans at all, but they are posted to increase the number of links to a particular website in order to boost its search engine ranking.

One of the programmers of the company describes the "cat and mouse chase" with the spammers and spam bots; in the past ten years spam has been for him one of the biggest problems of the discussion forum. He feels that only recently they have finally started to master the spam by using machine-enabled filtering. One of the latest incidents of automated posting of links was to boost the search engine ranking of a sport-related event. Occasionally the link posting is also done by humans. Based on IP addresses, the human-generated spamming is mainly produced from India, frequently referred to in the press as "the spam capital of the world".

From the perspective of the broken data metaphor, spam bots raise further questions about brokenness and repair work by paying attention to how the discussion forum, and the data that it generates, is kept clean by filtering it manually and automatically. We now know for sure that the peak in the data on the 17th of July 2009 was a spam bot: the amount of messages on that day was 32 033. Of these messages 17 850 contained the following text: "This message has been removed by admin." From the perspective of data work, analyzing the removal messages, in the context of the discussion forum, might, for instance, tell us about the temporal rhythms of spammers or which conversation threads are more likely to be infected with unwanted content. From the programmer's perspective, on the other hand, the cleaning work is active development of new filters and tools that search the discussion forum in order to identify harmful or rubbish content. For a programmer this work can be quite exciting and enjoyable, trying to get on top of the spammers and being one step ahead. Spam bots call for improvisation and creativity, highlighting the role of repair work as an important resource for knowledge production and innovation (Tanweer et al 2016).

Concluding remarks

The broken data concept metaphor calls for paying more attention to the incomplete, fractured and changing character of digital data. The particular example used to highlight the shifting character of data focused on a peak in a data visualization identified as a spam bot. Acknowledging the incomplete nature of digital data in itself is of course nothing new, researchers are well aware of

their data lacking perfection. With growing uses of secondary data, however, the ways in which data is broken and incomplete might not be known beforehand, underlining the need to explore brokenness and the consequent work of repair. In the case of Suomi24, the data breakages suggest that we need to actively question data production and the diverse ways in which data are adapted for different ends by practitioners. As our *Citizen Mindscapes* collaboration suggests, the production of data is permeated by moments of breakdown and repair that call for a richer understanding of everyday data work and data practices. The intent of this paper has been to suggest that a focus on data breakages is an opportunity to learn about digital work, and to account for how data breakages and related uncertainties challenge linear and too confident stories about data work.

References

Bell, G. (2015). 'The secret life of big data'. In Data, now bigger and better! Eds. T. Boellstorf and B. Maurer. Publisher: Prickly Paradigm Press,7-26

Dant, T., 2010. The work of repair: Gesture, emotion and sensual knowledge. *Sociological Research Online*, 15(3), p.7.

Domínguez Rubio, F. (2016) 'On the discrepancy between objects and things: An ecological approach' *Journal of Material Culture* 21(1): 59–86

Jackson, S.J. (2014) 'Rethinking repair' in T. Gillespie, P. Boczkowski, and K. Foot, eds. *Media Technologies: Essays on Communication, Materiality and Society*. MIT Press: Cambridge MA

Lagus, K. Pantzar, M. Ruckenstein, M. and Ylisiurua M. (2016) Suomi24: Muodonantoa aineistolle. The Consumer Society Research Centre. Helsinki: Faculty of Social Sciences, University of Helsinki.

Moore, H (1999) Anthropological theory at the turn of the century in H. Moore (ed) *Anthropological Theory Today*. Cambridge: Polity Press, pp. 1-23.

Moore, H (2004) Global anxieties: concept-metaphors and pre-theoretical commitments in anthropology. *Anthropological theory*, 4(1), 71-88.

Pink, S., Ruckenstein, M., Willim, R., & Duque, M. (2018). Broken data: Conceptualising data in an emerging world. *Big Data & Society*, *5*(1), 2053951717753228.

Tanweer, A., Fiore-Gartland, B., and Aragon, C. (2016). Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. *Information, Communication & Society*, 19(6), 736-752.