Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–10

DOI: 10.1177/2053951715616649
bds.sagepub.com

SAGE

Hacking the social life of Big Data

Jennifer Pybus[1], Mark Coté[2] and Tobias Blanke[3]

Abstract

This paper builds off the *Our Data Ourselves* research project, which examined ways of understanding and reclaiming the data that young people produce on smartphone devices. Here we explore the growing usage and centrality of mobiles in the lives of young people, questioning what data-making possibilities exist if users can either uncover and/or capture what data controllers such as Facebook monetize and share about themselves with third-parties. We outline the MobileMiner, an app we created to consider how gaining access to one's own data not only augments the agency of the individual but of the collective user. Finally, we discuss the data making that transpired during our hackathon. Such interventions in the enclosed processes of datafication are meant as a preliminary investigation into the possibilities that arise when young people are given back the data which they are normally structurally precluded from accessing.

Keywords

Big Data, data making, datafication, hacking, mobiles, youth

Introduction

Ever since the Snowden revelations in June 2013 there has been a growing awareness of the depth and breadth of the data we generate and how it renders us into ever more traceable objects of surveillance. The profoundly asymmetrical, political economic dimensions of the production and circulations of data have led to deeply problematic power relations wherein every keystroke, website visited or application downloaded are now rich sites of potential surplus value. With the proliferation of mobile platforms, digital footprints are expanding rapidly, especially those of young people.

Teenagers between the ages of 13 and 17 are spending more time on their mobiles and less time accessing social media platforms on their Internet browsers. Nielsen studies show that mobile phone data usage of young people tripled in 2011 (Nielsen, 2011; Osborn, 2012). Platforms such as Facebook can now regularly gather data from over 500 million active users via their Messenger app, in addition to the 30 billion messages that pass daily through their recently acquired WhatsApp (Weisenthal, 2014). In the United Kingdom, 81 percent of teenagers have access to a mobile (Spence, 2013), alongside 88 percent in the United States (Lenhart, 2015). And yet, little is known about the different ways in which apps generate and share data.

When using an Internet browser on a laptop, desktop, mobile or otherwise, there are a number of available plugins such as DisconnectMe or Lightbeam which, to varying degrees, reveal and in some instances block third-party marketing and analytic companies from gathering personal social data.[1] Browsers such as Chrome, Safari or Firefox – the primary windows onto the Internet from a desktop or laptop computer – also offer methods in their preferences or plugins for blocking or clearing some of the unwanted cookies. By contrast, while an app such as DisconnetMe can be downloaded for users to retain some data control when browsing on their mobiles, for the most part, individual apps offer their users even less control over the data being gathered and/or shared (Han et al., 2012). Given the growing usage and centrality of the smartphone in the lives of young people, we see an urgent need to understand and unpack questions around a) the extent to which data is produced

[1]London College of Communication, University of the Arts London, UK
[2]Digital Culture and Society, King's College London, UK
[3]Digital Humanities, King's College London, UK

Corresponding author:
Jennifer Pybus, London College of Communication, University of the Arts London, Elephant & Castle, London SE1 6SB, UK.
Email: jennifer.pybus@lcc.arts.ac.uk

and shared by the apps that teenagers routinely download; b) how this routine extraction of data impacts how young people negotiate and conceptualize digital privacy; and c) what agentic 'data-making' possibilities exist if users can either uncover and/or capture what data controllers monetize and share about themselves with third-parties. Thus our research focuses on what we call big social data (Coté, 2014; Manovich, 2011); that is, the data we produce through our mediated cultural and communicative practices both on the Internet and now on mobile devices.

The leaky mobile ecosystem

According to Han et al. (2012), mobile applications or apps are an inherently 'leakier' medium; that is, they extract more personally identifiable information compared to platforms that run off a desktop browser. The configuration of the mobile's intensive data flow is in part rooted in the infrastructure of applications that do not necessarily distinguish between first and third-parties (Egele et al., 2011; Enck et al., 2010). To clarify, the first party represents the app proprietor such as Facebook or Google or Candy Crush. Conversely, the third-party represents a tracking body that has been granted access to a user's data via the first party, typically for commercial purposes (although not exclusively). Web browsers, unlike apps, entail protocological restrictions between first and third-parties, which have been established by the World Wide Web Consortium (W3C). However, Han et al.'s research suggests that both Android and iPhone applications can and do share the device and SIM identifiers – such as the user's phone number – thereby providing personally identifiable information about the user to third-parties. The problem, they argue, is that "this means that mobile third-party tracking may be done with identifiers that allowed activity to be linked to people over long periods of time" (2012: 3). In addition, the mobile ecosystem offers what Han et al. refer to as sensor data, that is, sensitive data such as location, images and audio from smartphones. The combination of both real world names and sensor data that can now be easily gathered are allowing for even richer and more hyper individuated data profiles.

Coming to terms with the leakiness of the app ecosystem is critical given that our time spent within these siloed environments is only increasing. In the US, a 2014 study performed by ComCast revealed that 60 percent of people's time spent online transpires on smartphones, compared to on browsers (Perez, 2014). Young people are in the vanguard of this trend, with 91 percent of teenagers in the US regularly reaching for smartphones and the ubiquitous connectivity that they offer (Lenhart, 2015). The increasing use of applications on mobile devices does not, however, mean that they are unaware of the compromises required to manage their digital privacy. Raines-Goldie (2010) and Marwick and Boyd (2014) have demonstrated how young people deal with such challenges, and draw on highly developed strategies to obfuscate aspects of themselves within the public platforms they are using. In short, they have learned to hide in the open.

Despite looking for ways in which they can control their online identity, young people often feel caught in what has been referred to as a 'privacy paradox' (Acquisti and Grossklags, 2005; Shklovski et al., 2014). In other words, there is a strong sentiment that levels of data surveillance are simply 'creepy' or rather too intimate. Yet this sentiment notwithstanding, most appear resigned to pressing 'agree' to the countless number of 'terms and conditions' agreements, which function largely to safeguard and legitimize the extraction and monetization of data (Kim, 2014; Shklovski et al., 2014).

Privacy policies and terms and conditions, virtually non-existent prior to 1998, have since become: "the primary legal instruments conveying representations to consumers regarding third-party data sharing" (Kim, 2014: 328). Scholars such as Shklovski et al. (2014) have been trying to come to terms with this imbalanced relationship wherein the access users are granted to mobile apps ultimately surrenders their social and cultural data to private interests. And yet, while users may feel "outraged" or "express dismay", they proceed with "business as usual", going back to the same routine habits that only intensify those processes of extraction (2014: 8).

As Crawford and Boyd argue, it is not simply the amount or size of the data being extracted that matters; rather, what requires our attention is the inherent potential value derived from its relationality with other data points (2012). Here, value is not simply being derived from the petabytes of information being extracted, instead it comes out of a newfound capacity to process large volumes of data from a multitude of different structured and unstructured points, at increasingly higher velocities (Kitchin, 2014). As such, the data that is generated via our social relations online has gained a newfound depth and breadth of potential due to its varied form, routine generation and new modes of algorithmic processing.

Datafication to data making

Mayer-Schoünberger and Cukier (2013) refer to datafication as a process of the quantification of the world into data that is in turn reconstituted into new forms of value. Their emphasis is on its potential for analysis, predictive and otherwise, that they believe provides previously unavailable kinds of insights. We find datafication useful for our research in identifying an emergent

object of study: the quotidian data we generate. This 'born digital' material is distinct from digitisation, wherein analogue cultural expression is converted into digital form, the traditional focus of digital humanities.

For Mayer-Schoünberger and Cukier, datafication is presented primarily via the data-driven economy which creates the conditions for mediated social and cultural practices and expression to be repurposed and transformed into a quantified format – unlocking the potential to create something new that lies dormant within the data. Google Books is an exemplar. Scanning millions of books, hence datafying hundreds of millions of data points on every page that entered their vast digital archive, ultimately led to the creation of a new asset – Google Translate. Value for Google was not a simple effect of the data they extracted from all of the books they scanned; instead, it came through realising the relational potential within those datasets to create what is now the most used translation platform in the world. Such a process is not necessarily forthcoming; for example, there is no Amazon Translate, despite their extensive digitisation of books for Kindle. In both instances, Google and Amazon had access to a massive archive of digitized and datafied books and yet only Google sought to leverage the potential in their data sets into a new service that would generate even more data.

Datafication is therefore not a straightforward process that transforms analogue or digital material into quantifiable data for the production of surplus value. It is equally about recombination; therein finding expressions of relationality that are inherently possible among mutable data points. The multivalent variability of data that allows it to be constantly reimagined, signals that value is not necessarily forthcoming through mere access to a large data set. Instead, value is produced in the algorithms that are programmed to ask different questions, across different data sets, allowing for new forms of recombination and reuse.

Datafication, however, as presented by Mayer-Schönberger and Cukier suffers from what we call the resolution of verisimilitude. We intend a double articulation of resolution denoting i) the will and determination of those who assert that with enough data we can know the world, and ii) the ever more finely granulated data points that are meant to render it transparent and true. We question such faith in data plenitude, an $n = \text{all}$ reality where empirical analysis is transformed into an automated algorithmic effect as subsequent predictive analytics become fact.

Other interlocutors share such reservations. Van Dijck, for instance, notes that the concept of datafication is ''rooted in problematic ontological and epistemological claims'' (2014: 198) and should not be regarded as a neutral process. Instead, it is predicated on both access and the capacity to generate new forms of value from preexisting and newly acquired data. Elsewhere Coté has referred to this asymmetrical process as data motility (2014). This denotes the way in which from the moment we tap send, the data we generate richly circulates almost wholly outside of our control, yet remains profoundly tethered to us, enabling and constraining our conditions of possibility.

The relevance of datafication to our work is two-fold. First, it demonstrates a new materiality of data wherein cheap memory, powerful processors, algorithms and machine learning quantify our world and selves. Secondly, it raises questions of agency within this process. Mayer-Schoünberger and Cukier's framing of datafication fails to move beyond the uncritical capture and transformation of data for economic gain. However, according to O'Neil (2013), this increasingly pervasive process needs to account for *who* is actually controlling and framing the data and for *what* purpose. In short, important questions which appear to be absent in Mayer-Schoünberger and Cukier's account of datafication.

Vis (2013) furthers this critique through a nuanced reckoning of the ways data is shaped by everything from application program interfaces (APIs) to data mining methods to researcher motivation. She notes how these processes of 'data making' create myriad challenges for researchers both in terms of data access and quality. We concur that there is a need for further critical examination of the methods and tools, especially as a means for rebalancing agency within data-intensive ecosystems. Foregrounding 'making data' in processes of datafication can highlight the asymmetry that demarcates this socio-technical assemblage of access and capacities, especially in terms of the ability to activate the transformational potentialities located in the data itself. Whether it be a researcher accessing a Twitter feed, or someone using their smartphone, identifying the 'whom' and 'how' of data making is not so simple.

Datafication frames us as primarily passive generators in the social life of Big Data. Conversely, valorising 'making data' opens the possibility of becoming more active. In the context of datafication, the drive towards 'making data' can thus be seen as a strategic mode of agency that can arise if the subjects of datafication are given tools to both understand and work with the data that they produce. Innovative methods can provide insight into that which is regularly being captured about users. Our research approach is but one of many possible ways of understanding the social life of the data we generate, wherein we might be able to critically leverage the spirit of Jenkins' 'participatory culture' (2006) into a realm of Big Data. However, we don't simply want to celebrate our capacity to produce and widely share content, we need to gain clearer

insight into the extensive ecosystem in which our data is already participating. The approach we took sought to enable the YRSers to become data-makers. Similar to Nick Couldry (2014), we see a need for new kinds of innovative research that examines how we can take up and use analytics more autonomously.

Data literacy can act as an extension and updating of traditional discourses around media literacy[2] by refocusing our attention to the material conditions that surround a user's data within highly proprietary digitised environments. Given the growing imbalance between those who produce data and those who produce value from that data, there is a need to open up new forms of digital literacies, such as privacy literacies, information literacies, code literacies, algorithmic literacies, database literacies and so forth.

Puschmann and Burgess (2013) reach a similar end point in their work on the politics of Twitter's ecosystem. If a user does not understand how they can leverage their API, nor understand its technical constraints, then they are unable to effectively interact with the platform. As a result, mainly corporate actors or those who are technologically adept and/or possess the resources can meaningfully engage with the data that is being collectively produced. Similarly, we are interested in what happens when closed data becomes open, to reveal what young people regularly generate but typically cannot access. In trying to unpack the processes by which data is collected, brokered, aggregated, analysed, monetized and acted upon, we conceived of our project as an initial step toward the development of a holistic approach to data literacy. One that is able to both question how meaning is constructed and (re)presented from the data but equally seeks to unpack those opaque material processes that enable this capture and (re)presentation alongside the active (re)shaping of data infrastructures.

Mobile methodologies: *Our Data Ourselves*

We have taken an interdisciplinary approach with a research team comprised of media and cultural theorists, computer scientists, programmers and youth. The gambit has been that such a grouping would enable a rigorous exploration of our data to facilitate a critical intervention into the datafication of youth cultures via the mobile phone. Our focus has been on the asymmetrical power relations that are deeply imbricated in the structural ways in which data is produced by and yet flows away from the user.

To begin, we set out to work with a very specific demographic of young people between the ages of 14 and 18 years old. The participants were all affiliated with Young Rewired State (YRS). This non-profit, UK based organization brings together communities of youth who have an interest in coding. By using project-focused learning, young people work in groups to improve their coding, alongside both peers and mentors. The members of YRS are also given opportunities to experiment with new technologies, software and computer languages while being actively encouraged to turn their ideas into real prototypes at different hackathon events. The pre-existing rich computer literacies that these young people possessed were seen as an asset to our project, enabling us to establish both a co-learning and research relationship.

There is a wide range of participatory research methods (cf. Reason and Bradbury, 2008) yet we use co-research in explicit reference to the politicised methods of *conricerca*, evoking the Italian *operaismo*, or 'workerist' movement dating back to the 1960s. The subsequent *conricerca* or co-research method developed as both the production of knowledge and organisation. Our method is also situated within the expansive field of action research given the opacity of the material infrastructures that predominantly govern our data beyond our understanding or control. Our partnership with YRS marked an opportunity to examine the data produced by mobile apps without reproducing those conditions which render individuals powerless against the ubiquitous data sharing that transpires with economic and political third-parties. In short, we did not simply want to use our subjects as data producers, instead we wanted to imagine what a data literate subject might look like.

Collaboration with our participants was therefore critical. From the beginning we expected the YRSers to help us access the data they were generating within their mobile environments and then analyse what had been collected through the creation of different applications. The endpoint, however, was not to realize a completed prototype but rather to observe their heuristic process and approach, which could potentially facilitate different pedagogical practices for those who do not have the same technological expertise. We thus found resonance with action based research methodologies, such as those described by Baym (2013) and Kennedy et al. (2014) who specifically argue in favour of "small-scale, qualitative studies [which] open up a space for reflecting on the affordances of digital methods" (p. 174). Similarly, Coleman's study of the myriad technologically savvy direct actions taken by Lulz and Anonymous (2011) points to the creative possibilities of hacking.

As a first step, we provided 20 young participants with smartphones, a six-month data plan and created an app called the MobileMiner. The app was designed to learn about the data being generated by both the devices and applications that our participants were using. While we were aware that most of our participants

were also using their pre-existing mobiles, we hoped that the free data plan we provided would entice them to use our phones. Ultimately, given the exploratory nature of our project, we were more focused on the quality, as opposed to the quantity, of what could be gathered and analysed. That being said, to account for this discrepancy of mobile use, we asked our participants to keep tumblr media diaries to investigate the apps they were regularly using (regardless of the device) and to ascertain their perception of the amount of time they thought they were spending on their devices. In addition, we held two hackathons; the first explored ways of improving the MobileMiner app, and the second allowed our participants to work with the data that our application had collected from their devices.

In addition, we divided all our research participants into groups and held three, two hour focus groups at the beginning of the project. The aim of these sessions was to a) establish relations of trust and b) gain a broad sense of how our participants view, understand and negotiate their privacy online. Additionally, we conducted interviews during and after the second hackathon to observe the individual's reaction once presented with their personal data cache amassed by the MobileMiner app and then later to gain a clearer understanding of the prototypes they constructed.

MobileMiner: Collective construction of research devices

There were two primary deliverables from our co-research: a data-gathering app and the data itself. The MobileMiner application we created was initially downloaded onto the Android smartphones we gave to our participants. It was designed to gather the data which is most typically harvested by apps. The application was always conceptualized as a data-making tool that could eventually be used outside of our research project and can be found on GitHub for free download. To ward off 'creepy' surveillance and enable greater privacy for our co-researchers, they had the autonomy to turn on and off the app as they wished and thus were never forced to have their data mined.

We envisioned our application as a specific contribution to the growing number of open source tools that are making visible the dynamic ways in which a user's cultural practices and sociality become datafied. It's function was to collect in- and outgoing communications from the smartphones, including information on the total amount of data sent and app network activity, alongside location information and identifiers. We wanted the application to function as a kind *data spy*, thereby examining the frequency at which apps routinely harvest data. The approach we undertook is used by commercial providers and has the advantage that it does not require permissions granted by the user when the app is installed. We decided early to not be too aggressive about tracking the app behaviour and excluded, for instance, direct access to GPS information on the phone. Instead, we used opencellid.org to link our app to the cell tower location database. Building trust with our co-researchers factored more strongly into our research design then maximising the harvesting capacity of the MobileMiner. Furthermore, in contrast to surveillant-modes of maximum data scooping, we wanted to find out what non-aggressive modalities of datafication could yield.

The second research deliverable co-developed with the YRSers was the actual data that was collected via the MobileMiner app. We envisioned this as part of what a big social data commons could look like, insofar as it was anonymised but open to exploration and creative use by our co-researchers. Data harvested from our mobile devices almost invariably ends up in proprietary datasets over which we have no access. In contrast, we wanted to explore an alternative infrastructure. To do so, we used a Comprehensive Knowledge Archive Network (CKAN) platform, an open source repository developed by the Open Knowledge Foundation, as the core element of an open data ecosystem.

The CKAN instance was updated at regular intervals from a local database on the mobile phone via the MobileMiner app. Finally, we took the innovative approach of packaging the CKAN data together with a standard toolbox so that this prototype of a big social data commons could be worked over in a virtual machine that is free to download from our website. Once both the MobileMiner and CKAN were in place, we started a controlled experiment to receive data from our participants over a six-month period. We returned this to our co-researchers during the second hackathon for creative inquiry.

Focus groups: From individual agency and control to group privacy

The focus groups revealed that most of our co-researchers are deeply aware of the compromises involved when it comes to managing their digital privacy. All of the teenagers we spoke with know that their data is being routinely mined. At issue, however, is not the knowledge of data extraction but rather what happens to this information once it moves beyond their reach. Reactions among our participants were varied. Kylie, for example, spoke out at length about her frustrations around the monetization of data:

> The problem with this generation is that we are far more scared of the individual, you know the pedophiles and the online bullies rather than corporations. What

> we are not taught about is that fact we shouldn't be allowing big corporations like Google and Facebook to have access to our personal details.

In fact, she was so bothered by the privatization of data that she and her friends had refused having a Facebook, Instagram or Twitter account. Conversely, most of the other participants showed far less concern. Jacob put his thoughts the most succinctly: "Perhaps the most surprising thing is just how little we care about that!" (in relation to the growing number of third-parties mining personal data), followed by a very animated discussion whereby others in the group confirmed and reiterated his opinion.

Jacob's comments, however, need to be taken into context. As one of the stronger coders in the room, he also expressed how important it was to have control over the technologies he was using; a sentiment that was taken up by almost everyone we spoke to. According to Anna (which garnered a number of nods in focus group A):

> Being of this generation and being tech savvy we have some control because we know how to have control, whereas I know that my Mum doesn't have any idea...We know we can control our privacy and a lot of people do but then a lot of people also go and they are just using the technology but don't actually understand how it works.

Control, within this context, links directly with Anna's sense of agency that is bolstered by her intimate knowledge of how to use and manipulate the technologies that she is engaged with. Overall, we would summarise our findings on agency and control over privacy as follows: i) a desire for control over messages being sent about oneself; ii) control over a message in relation to peer groups –thus it mattered more if a close friend uncovered a bad photograph posted on Instagram then a complete stranger; and, iii) a desire for better understanding of the technological affordances of a medium.

We observed the relationship between the desire for agency and control and technological literacies playing out in a number of different ways for our participants. For example, several of our co-researchers actively obfuscated the data being captured about themselves, most notably through the routine removal of Google's geolocation data from their mobile browsers. The tactics used to protect their privacy – 'unmaking data' – ranged from deleting metadata that reveal geolocation in photographs, deleting accounts, withholding personal information such as addresses, setting up proxies (one participant claimed he had set up seven proxies on his mobile phone) and using alternative open sourced platforms that run on Linux. What these strategies unambiguously revealed for this group of young people is the inherent confidence that comes out of their various technological literacies.

Maintaining digital privacy might be somewhat easier for our participants, but if their peer groups want to use a leaky app, we discovered that most of our participants will use it too. Phoebe laughed, recalling when she had tried to get her friends to take up an open source alternative to Skype. She quickly abandoned these efforts because her friends found this platform too difficult and cumbersome to use. Despite how effective this other platform was at preventing data from being tracked, she went back to Skype. She went on to recall why she never stayed on 'Diaspora', an open-source, decentralized social network that has tried to provide users with control over their data. When her friends again refused to migrate, she remained on Facebook. So while she would have preferred to use this alternative network, she elected to remain where her friends were. For her, privacy is arguable not inherently individual, but, rather, far more collectively understood. A similar sentiment was expressed by David when he was talking about using the privacy settings on Facebook: "it's kind of herd thing, you've all got to do it otherwise, one person is in trouble." Again, privacy is not simply about him but about collective choices that are made by his peer group.

Often, when it comes to online privacy, the focus has remained on what the individual does or should do, such as in Livingstone's study on teenagers in the age of social networking (2008) that assesses privacy through the lense of risk assessment. Later, she expands this work with another study performed with Livingstone et al. (2014), wherein their findings focus on the identity management techniques of EU children. Conversely, Boyd and Marwick's (2011) work on privacy explores the various strategies that young people have invoked to maintain control within a networked environment. Their findings suggest that on-line privacy does matter significantly and is observed through the paucity of ways in which their participants had learned to conceal information from undesirable interlocutors while remaining visible to their peer groups. What Boyd and Marwick do not explicitly mention, but what our research suggests, is that there are also inherent group dynamics making it possible for these young people to hide in the open and maintain some degree of control over the content they have chosen to circulate within the wider domain. Seeing the dynamics of privacy as being rooted in a collective, based on the cultural norms that have been established by a peer group, is an interesting finding with practical and theoretical implications, particularly when it comes to pedagogical approaches to data literacy.

Hackathons: Making tools

Our hackathons draw on the digital humanities tradition of 'critical making' as complementary to 'critical thinking'; that is, as a material form of critique which is both 'external' and 'community-oriented' (Svensson, 2012: 53). While it is beyond the scope of this paper to discuss all of the tools and prototypes that were created during our second hackathon, it is instructive to offer a brief overview of what some of our coders were able to produce when given access to their data on the CKAN instance. The challenge with our methodological approach lies in its demand for specific technological competencies to manipulate the data that we were able to provide. Indeed, how can a user participate in building an open and transparent digital commons if they cannot confidently engage with platforms at the level of code, algorithms and databases? There is no comprehensive or immediate way to meet this challenge. Nonetheless, we anticipated that there was much to learn from the technologically savvy young people we worked with.

Here we can present some relevant examples. Phoebe created what she referred to as 'sous-sousveillance' of the apps she was using. This approach was inspired by the initial observations that revealed how some of the applications being used by our participants were extracting significantly more data than others. For example, the MobileMiner registered that three of our YRSers (Kylie, Edward and Julian) were playing a game entitled: 'Don't Tap the White Tile,' which was accessing the Internet on 46, 53 and 42 occasions over periods of 21, two and three days, respectively. According to our analysis, the frequency at which data was exchanged between this app and the server is supported by commercial studies, which report that up to 95% of mobile apps see similar patterns of use (Boris, 2013). However, these data patterns are in stark contrast to Kylie's interaction with another game: 'The Line-Keep In,' a simple app involving the navigation of a dot through a vertically scrolling maze.

On closer inspection of the code, the game uses various tools to gather deep statistics and push user messages. Permission to access player GPS location is requested, even though the Mobile Country Code of any device's last connected cell tower would be sufficient to localize advertisements to the player's country. Despite the game's simplicity, her device called home 1760 times over a period of 27 days, with significant activity registered every day (Blanke et al., 2014). This number is notably higher than the other application we observed and raises questions in relation to the structural differences between the apps users regularly download.

Equipped with the data from MobileMiner, Phoebe wanted to consider what constituted a leaky app. She therefore tried to create an application that could sonify what she called the "attention grabbing-ness of apps . . . so you could tell immediately which were worse for calling home." To clarify, the term calling home refers to when an application accesses its server on the Internet, presumably to relay personal and or GPS data accumulated from the user's smartphone and/or other applications. The MobileMiner app was designed to track the frequency by which data moves between the user and any given downloaded app. Our challenge, however, arose in determining what, specifically, passes between the mobile and the app's server. Nevertheless, the data that we did procure was enough to facilitate Phoebe's hack on the "noisiness of apps."

Ultimately, she wanted to create an application that could *listen* to the data as it moves from an individual phone to a server. She therefore assigned tones and colours to correlate with the frequency of data requests that are routinely made by apps. Had she completed this hack, her app would have generated brighter and more grating sounds to coincide with the leakier and more invasive apps found more broadly on a user's device. When asked why she wanted to turn the data into sounds and colours, Phoebe told us that these kinds of visualizations were in many respects more accurate:

> Most people would prefer this to numbers on a screen or paper, as it's a lot more jarring for the non-savvy, as some could say that higher, louder tones are uncomfortable, whereas seeing numbers is relatively meaningless unless you know the context . . . This could break down the complicatedness for the end user . . . It would be great if you could listen to a list of apps, to find the tones, to find the ones that are potential problems.

The tool she set out to create was meant to easily draw a user's attention towards the leakiness of different apps, regardless of their technological expertise. This is but a small example of the kinds of creative and critical possibilities that might emerge when users have access to their data, in this case via a CKAN instance. Subsequently, Phoebe's hack reflects both her strong political engagement and a desire to create tools that might engage others:

> Far too many people don't understand quite how much they are giving to companies and how much this data is worth to them especially when the privacy policies are shady at best. And when you can't have members of the public check what the Facebooks and Googles are doing inside of these apps and with the data behind closed doors then it becomes very easy for them to exploit the user. After all, data is the new currency and with the amounts these companies have they could buy anyone.

Agency, here, lies in having more transparency in the ability to work on, process and transform the materiality of datafication, elements which typically move outside the user's knowledge or control. The sonification tool represents a creative possibility in a data literacy toolkit, fostering resistance by illuminating those covert 'sous-surveillance', data mining practices that sustain new economic and algorithmic relations. Tom's hack also examined the value of collective surveillance/sous-surveillance practices. However, his interest lay in the frequency of app activity, and whether there was a correlation between app usage and the amount of information being relayed to a server.

Tom used the CKAN instance data to develop a graph generating tool to demonstrate when and for how long our participants were using the Twitter, Facebook and Facebook Messenger app. More specifically, he wanted to know what days of the week these apps were being used, alongside the frequency and times in which they were accessed. He then cross referenced these with the number of notifications that the apps sent back to their home servers. By so doing, he asked three questions of the data: 1) Does using an application result in an increased number of notifications? 2) Is there a day of the week that appears to make the user more vulnerable to the app leaking? 3) Is there a time a day that makes the user more vulnerable to the app leaking? His end point was to visualize and see how this might change over time.

While Tom was less concerned about the appropriation of his data for commercial use in the focus group, it remained important to quantify and qualify his social media usage, again echoing those discourses discussed earlier around control. Yet, instead of tracking the material body's movement through space and time, Tom wanted to grasp his movements within those digitally networked social media environments. He imagines:

> Maybe in the future this could work on a daily basis over months and years rather than just a week. We could see how social media is used more on important days such as during big events. For others it could also be used for statistical analysis along with other social media datasets to create a wider picture of what we do online.

Similar to Phoebe, Tom wants to have access and control over his data to better understand his data profile. However, this is not simply about having more open data[3] but instead, to develop more nuanced ways of understanding his own personal data production in relation to his peers. Here he states:

> While I agree with open datasets, I don't think that data should be taken without permission – people should have control over their data, anonymous or not, and if anything is to be done with it it should be open to all rather than kept by companies.

When we think about how data literacy might be developed through further research, we are drawn to the empowerment that comes from gaining access to what we collectively generate. For these coders, agency began with their ability to manipulate and make something from the entire set of data that the MobileMiner app gathered from *all* of our participants' smartphones in the CKAN instance.

Conclusion

Throughout the project, the active participation of the YRSers has always been paramount. Echoing Kelty (2008), we tried to facilitate the development of a recursive data public by practically modifying the material infrastructure of datafication to critically engage power–knowledge relations. Our research shows that there is tremendous untapped potential in the general intellect and technical practice, not just among our teen coders but in the figure of the data generator who wants to be in control and, more importantly, seeks to understand the data they collectively generate.

For us, a recursive data public is one with augmented critical data making capacity. We therefore see our contribution in two ways: The first lies in creating open social and cultural data sets that are accessible for critical and creative use by both researchers and the general public – an area of research that we are currently pursuing, as are others. Some, for example, are exploring the potential of blockchain technology, the 'distributed ledger' underpinning BitCoin that may facilitate new forms of digital commons (see Bollier, 2015; O'Dwyer, 2015). Others, such as Salvatore Iaconesi and Oriana Persico, are building a ubiquitous commons;[4] that is, a space calibrated by new technological, legal and social protocols which assure greater user control – a sentiment that strongly resonates with our empirical research.

Our second contribution is to further develop non-prescriptive modes of interdisciplinary research, including but not limited to hackathons and workshops. Our project demonstrated the enriching and critical value forthcoming from working with and learning from young coders. More research is required to explore a collective understanding of digital privacy, in addition to innovative digital methods that can further unpack the mobile ecosystem. The realm of datafication is opaque; metaphorically we would like to imagine attaching radio-frequency identification (RFID) tags to granular data points and tracking their flow through the social life of data which drives predictive analytics,

circuits of consumption, business intelligence and state surveillance. More collective interdisciplinary methods that engage our contemporary technological condition beyond the enclosures of platforms, apps and user interfaces are required. Moving forward, if we want to develop new tools and methods to enhance the active participation of users, then we must carefully consider how we can both make and unmake data. The sous-surveillance and social media quantification tools are but two examples meant to highlight the possibilities that can exist if users have access to their own data.

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: Arts and Humanities Research Council, 10.13039/501100000267 (Grant number: AH/L007770/1).

Notes

1. The term is used loosely here, as more often blocking third-parties or cookies ultimately means that the user will then not have access to the website.
2. Media literacy here refers to a large body of work in Cultural and Media Studies that has been aimed at building various tools to empower individuals to think critically about how meaning gets sedimented and in turn how the world around them is constructed. The end point is to empower individual subjects to understand this process so that they can shape their own identity and, as Kellner and Share argue, ''transform the material and social conditions of their culture and society'' (2005: 369).
3. For a more comprehensive critique of open data initiatives please refer to Rob Kitchin's work (Kitchin, 2013).
4. cf. http://www.ubiquitouscommons.org/

References

Acquisti A and Grossklags J (2005) Privacy and rationality in individual decision making. *IEEE Security and Privacy* 3: 26–33.

Baym N (2013) Data not seen: The uses and shortcomings of social media metrics. *First Monday*. Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4873/3752 (accessed 15 June 2015).

Blanke T, Greenway G, Pybus J, et al. (2014) Mining mobile youth cultures. *IEEE* 14–17.

Bollier D (2015) The blockchain: A promising new infrastructure for online commons, news and perspectives on the commons. In: *David Bollier,* 3 April. Available at: http://bollier.org/blog/blockchain-promising-new-infrastructure-online-commons (accessed 15 May 2015).

Boris C (2013) Twenty-two percent of mobile apps are only used once. *Marketing Pilgrim,* 6 September. Available at: http://www.marketingpilgrim.com/2013/09/twenty-two-percent-of-mobile-apps-are-only-used-once.html (accessed 28 January 2015).

boyd D and Marwick AE (2011) Social privacy in networked publics: Teens' attitudes, practices, and strategies. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Available at SSRN: http://ssrn.com/abstract = 1925128 (accessed 10 November 2015).

Coleman G (2011) Anonymous: From the Lulz to collective action. *The New Everyday: A Media Commons Project*. Available at: http://mediacommons.futureofthebook.org/tne/pieces/anonymous-lulz-collective-action (accessed 15 June 2015).

Coté M (2014) Data motility: The materiality of big social data. *Cultural Studies Review* 20(1): 121–149.

Couldry N (2014) Inaugural: A necessary disenchantment: Myth agency and injustice in a digital world. *The Sociological Review* 62: 880–897.

Crawford K and Boyd D (2012) Six provocations for big data. *Information, Communication & Society* 15(5): 662–679.

Egele M, Kruegel C, Kirda E, et al. (2011) PiOS: Detecting privacy leaks in iOS applications. *NDSS*. Available at: https://www.iseclab.org/papers/egele-ndss11.pdf (accessed 16 June 2015).

Enck W, Gilbert P, Chun B-G, et al. (2010) TaintDroid: An information-flow tracking system for realtime privacy monitoring on smartphones. *OSDI*. Available at: http://static.usenix.org/event/osdi10/tech/full_papers/Enck.pdf (accessed 16 June 2015).

Han S, Jung J and Wetherall D (2012) *A study of third-party tracking by mobile apps in the wild*. Report, University of Washington, US, 1 March. Available at: ftp://ftp.cs.washington.edu/tr/2012/03/UW-CSE-12-03-01.PDF (accessed 15 June 2015).

Jenkins H (2006) *Convergence Culture: Where Old and New Media Collide*. New York: NYU Press.

Kellner D and Share J (2005) Toward critical media literacy: Core concepts, debates, organizations, and policy. *Discourse: Studies in the Cultural Politics of Education* 26(3): 369–386.

Kelty CM (2008) *Two Bits: The Cultural Significance of Free Software*. Durham: Duke University Press.

Kennedy H, Moss G, Birchall K, et al. (2014) Balancing the potential and problems of digital methods through action research: Methodological reflections. *Information, Communication and Society* 18(2): 172–186.

Kim N (2014) Three's a crowd: Towards contextual integrity in third party data sharing. *Harvard Journal of Law and Technology* 28(1): 325–347.

Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data and Society* April–June: 1–12.

Kitchin R (2013) Four critiques of open data initiatives. In: LSE: The impact blog, 27 November. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2013/11/27/four-critiques-of-open-data-initiatives/ (accessed 5 October 2015).

Lenhart A (2015) Teens, social media and technology overview 2015. *The Pew Research Centre*, 9 April. Available at: http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/ (accessed 15 June 2015).

Livingstone S (2008) Taking risky opportunities in youthful content creation: Teenagers' use of social networking sites for intimacy, privacy and self-expression. *New Media & Society* 10(3): 393–411.

Livingstone S, Mascheroni G and Murru MF (2014) Social networking among European children: New findings on privacy, identity and connection. *Hermès*. CNRS Editions.

Manovich L (2011) Trending: The promises and the challenges of big social data. In: Gold MK (ed.) *Debates in the Digital Humanities*. Minneapolis: The University of Minnesota Press. Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (accessed 28 January 2015).

Marwick A and Boyd D (2014) Networked privacy: How teenagers negotiate context in social media. *New Media and Society* 16(7): 1051–1067.

Mayer-Schoünberger V and Cukier K (2013) *Big Data. A Revolution that will Transform How we Live, Work, and Think*. London: John Murray Publishers.

Nielsen (2011) New mobile obsession: U.S. teens triple data usage. *Nielsen*, 15 December. Available at: http://www.nielsen.com/us/en/insights/news/2011/new-mobile-obsession-u-s-teens-triple-data-usage.html (accessed 28 January 2015).

O'Dwyer R (2015) The revolution will (not) be decentralised: Blockchains, common transitions. *Common Transition,* 11 June. Available at: http://commonstransition.org/the-revolution-will-not-be-decentralised-blockchains/ (accessed 15 June 2015).

O'Neil C (2013) The rise of big data, big brother. *Mathbabe,* 2 May. Available at: http://mathbabe.org/2013/05/02/the-rise-of-big-data-big-brother/ (accessed 28 January 2015).

Osborn C (2012) Teenagers: Mobile gaming, apps and data greed. *Znet*, 26 March. Available at: http://www.zdnet.com/article/teenagers-mobile-gaming-apps-and-data-greed-infographic/ (accessed 28 January 15).

Perez S (2014) Majority of digital media consumption now takes place in mobile apps. *TechCrunch* 21 August. Available at: http://techcrunch.com/2014/08/21/majority-of-digital-media-consumption-now-takes-place-in-mobile-apps/ (accessed 15 June 2015).

Puschmann C and Burgess J (2013) The politics of Twitter data. *SSRN Electronic Journal*.

Raines-Goldie K (2010) Aliases, creeping and wall cleaning: Understanding privacy in the age of Facebook. *First Monday,* 15(1–4). Available at: http://firstmonday.org/article/view/2775/2432 (accessed 15 June 15).

Reason P and Bradbury H (eds) (2008) *The Sage Handbook of Action Research. Participative Inquiry and Practice*. London: Sage.

Shklovski I, Mainwaring SD, Skúladóttir HH, et al. (2014) Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In: *Proceedings of the 32nd annual ACM conference on human factors in computing systems - CHI '14*, Toronto, Canada, 26 April–1 May, pp. 2347–2356.

Spence E (2013) Smartphones kicks in the UK (with some help from Windows phones). *Forbes*, 4 April. Available at: http://www.forbes.com/sites/ewanspence/2013/04/30/teenage-smartphone-kicks-in-the-uk-with-some-help-from-windows-phone/ (accessed 28 January 2015).

Svensson P (2012) The digital humanities as a humanities project. *Arts and Humanities in Higher Education* 11(1–2): 42–60.

van Dijck J (2014) Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197–208.

Vis F (2013) A critical reflection of big data: Considering APAs, researchers and tools as data makers. *First Monday*, 10(2). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4878/3755 (accessed 28 January 2015).

Weisenthal J (2014) Why Facebook bought WhatsApp in one chart. *Business Insider*, 19 February. Available at: http://www.businessinsider.com/whatsapp-growth-2014-2?IR=T (accessed 15 June 2015).

This article is part of a special theme on *Data and Agency*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/data-agency.