MPC MAJOR RESEARCH PAPER

BOUGHT AND SOLD: EXPLORING THE EFFECTS OF BIG DATA ON USER AGENCY AND COMMODIFICATION

Kristia M. Pavlakos

Dr. Frauke Zeller

The Major Research Paper is submitted
in partial fulfillment of the requirements for the degree of
Master of Professional Communication

Ryerson University
Toronto, Ontario, Canada

Thursday, September 8th, 2016

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PAPER

I hereby declare that I am the sole author of this Major Research Paper and the accompanying Research Poster. This is a true copy of the MRP and the research poster, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this major research paper and/or poster to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP and/or poster by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP and/or my MRP research poster may be made electronically available to the public.

Abstract

Big Data[1] is a phenomenon that has been increasingly studied in the academy in recent years, especially in technological and scientific contexts. However, it is still a relatively new field of academic study; because it has been previously considered in mainly technological contexts, more attention needs to be drawn to the contributions made in Big Data scholarship in the social sciences by scholars like Omar Tene and Jules Polonetsky, Bart Custers, Kate Crawford, Nick Couldry, and Jose van Dijk. The purpose of this Major Research Paper is to gain insight into the issues surrounding privacy and user rights, roles, and commodification in relation to Big Data in a social sciences context.

The term "Big Data" describes the collection, aggregation, and analysis of large data sets. While corporations are usually responsible for the analysis and dissemination of the data, most of this data is user generated, and there must be considerations regarding the user's rights and roles. In this paper, I raise three main issues that shape the discussion: how users can be more active agents in data ownership, how consent measures can be made to actively reflect user interests instead of focusing on benefitting corporations, and how user agency can be preserved. Through an analysis of social sciences scholarly literature on Big Data, privacy, and user commodification, I wish to determine how these concepts are being discussed, where there have been advancements in privacy regulation and the prevention of user commodification, and where there is a need to improve these measures. In doing this, I hope to discover a way to better facilitate the relationship between data collectors and analysts, and user-generators.

[1] While there is no definitive resolution as to whether or not to capitalize the term "Big Data", in capitalizing it I chose to conform with such authors as boyd and Crawford (2012), Couldry and Turow (2014), and Dalton and Thatcher (2015), who do so in the scholarly literature.

Acknowledgements

I would like to extend my utmost thanks to my supervisor, Dr. Frauke Zeller, for her insightful comments, excellent advice, and continued guidance. She was an invaluable resource during this process, and I am thankful for the chance to have worked with her.

I would also like to thank my second reader, Dr. John Shiga, who was wonderful to work with, and whose insight was greatly appreciated.

Table of Contents

Abstract……...iii

Acknowledgements…….iv

Introduction……..1

Literature Review…….3

Research Questions……17

Data Collection Method…….19

Method of Analysis……22

Findings…….23

Discussion……..50

Conclusion…….53

Appendix……56

References……..60

Introduction

Technology is constantly evolving, changing the way that we observe the world and make assumptions about things on a daily basis. Technological innovation has had a major impact on data collection and analysis, and with data sets that are bigger and cover a broader scope of information than ever before, a new way of analyzing and interpreting data is needed. This new way of interpreting data comes in the form of the concept of Big Data analysis. Big Data is often conceptualized as massive amounts of raw numerical information, ready to be appropriated, digested, and analyzed. However, the term itself is somewhat of a misnomer, and as boyd and Crawford (2012) point out, Big Data is "less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets" (p. 663). Big Data sets differ from traditional data sets in the size of the data being analyzed, the rapidity of its aggregation and analysis, and the immense scope of data that it makes accessible. The field of Big Data is a relatively new area of academic study, but it is rapidly expanding as scholars and researchers discover its capacity to revolutionize data collection and analysis. While the analysis of Big Data sets can prove valuable in completing and enhancing research, it can also expose a number of ethical issues regarding user agency, including who owns the data and how to analyze it, as well as how much control users have in the data collection and analysis process. The purpose of analyzing Big Data sets is to draw generalizations about populations and make predictions about future behaviours based on the patterns uncovered. This means that this process is structured to be inherently commodifying, as it seems largely to benefit corporations who have a vested interest in using this data to target marketing materials to consumers. It tends to overlook users as individuals, and instead defines them based on their potential marketing profitability. It also

makes it easier to perpetuate stereotyping, as individuals are viewed *en masse* as an algorithm of aggregate behaviours and patterns.

Despite these ethical issues, Big Data makes analyzing large sets of data easier and more efficient than it ever was before; in today's technologically advanced society, it seems as if Big Data is becoming entwined with our knowledge and understanding of data analysis. As such, issues of user commodification and data collection and analysis in relation to Big Data must be studied in order to discover more innovative and ethical ways of collecting, analyzing, and interpreting this data.

Throughout this paper, I refer to three key terms that shape my research and discussion. Those key terms, and how I apply them in the context of this paper, are as follows:

Commodification: I use the term "commodification" to describe the process of attributing users with value based on the data they provide. Kennedy and Moss (2015) describe the way that "individuals' marketing value is calculated and each individual is categorised as target or waste" (p. 3). Users are commodified when their individuality as people is compromised by the monetary value placed on their data.

Agency: Agency refers to the ability to think and act independently. Having agency, or playing an agentic role in the data collection process, involves users' providing consent for corporations to use their data, or being able to make their own informed decisions and act for themselves instead of being acted upon.

Consent: The term consent implies the act of authorization. In the context of Big Data, consent describes users' act of affording a researcher or a corporation permission to collect and analyze their data.

Literature Review

The Internet is a technological environment where internet users (hereby referred to as "users") become data generators, leaving a trace of "markers" with every word they type and click they make. These markers track things like the terms they search and the websites they visit. All of these aggregate data, known as Big Data, can then be collected and analyzed to draw conclusions or make predictions. However, many ethical issues can arise in this process. Users are at risk of becoming commodified subjects, reduced to the data they generate, which results in a lack of user agency. This can be classified into two distinct categories: data mining and data ownership, and user commodification.

Defining Big Data

First, the distinction needs to be made between Big Data and traditional data. Zeller (2014) distinguishes Big Data from traditional data by attributing it with four key characteristics: "volume, variety, velocity and veracity" (p. 263). Whereas traditional data sets are usually smaller in size, Big Data is large-scale data analysis. These data are diverse in nature, and can be used to predict behaviours and draw conclusions about populations (Zeller, 2014).

Borlin and Schwarz (2015) state that "the debate around the societal and cultural effects of Big Data is extensive, and one of the pressing questions has been whether to see the turn to Big Data as merely a shift in scale, reach, and intensity (a quantitative shift) or as a more profound, truly qualitative shift – implying both a shift in being (ontology) and meaning (epistemology)" (p. 2). In this paper, my research questions argue the latter, in that the "shift in being" and in "meaning" alludes to Big Data's potential to commodify users.

Big Data sets are compiled by collecting and aggregating data from users' online behaviours; while the internet is, arguably, an essential component of daily life to many in

today's society, it is not yet universal, and as such not every member of every population can be represented by the data that comprises these sets. The issue then arises that the processes involved in the collection, aggregation, and analysis of large-scale data sets are essentially commodifying, from the way that corporations "mine" for data, to the way these processes restrict the user's voice and agency and assume ownership of and sell user-generated data for profit.

Data Mining

Data mining involves the excavation of data that can be analyzed to draw conclusions about certain questions and theories, or gain information about populations. Data mining can also be understood as "life mining", or "extracting useful knowledge from the combined digital trails left behind by people who live" (van Dijk, 2014, p. 200). However, the practice of data mining can uncover some ethical issues about users' control over the data they generate. Data mining processes can result in users' inability to control who collects their information and how it is used (Reyman, 2013). User-generated data is attributed as a "creation" of the user, implying that users should be considered the owners of their own data.

This lack of control is furthered by the use of automated data-mining machines, as "platforms that count and sort online data, such as Google and Facebook, work automatically via algorithms, often allowing users only limited degrees of manual adjustment" (Couldry & Powell, 2014, p. 2). What's more, sites like Google create an illusion of agency in users by stating that gathering their information allows them to improve their services, thus improving user experience. This allows users to believe that they are helping themselves by contributing to the eventual improvements of the platform that will facilitate their own use, when their consent is actually benefitting the corporation (Reyman, 2013).

Data mining can also be used for marketing purposes: the data generated from social media platforms is mainly used to determine the online habits of consumers for targeted advertisements, which makes the relation between data mining and user commodification explicit (Reyman, 2013). In this way, corporations can extract users' personal information based on internet use patterns, and can use this information to target their products to consumers.

However, many users are not conscious of the fact that they are contributing this data, and do not realize that it can be used to create a consumer profile for marketing purposes (Reyman, 2013). As users are unaware that they are generating this data, it makes it difficult for them to claim ownership or even realize that it may be being used at all; essentially, users "participate in the metrification of their habits either wittingly or largely unwittingly" (Borlin & Schwarz, 2015, p. 4). The idea of "not knowing" creates an issue of consent; if users are not aware that they are generating data, then they cannot be aware that it is being used. As such, they are unable to claim ownership of this data or consent to its use, which risks making its use a violation of their privacy.

Data Ownership

Many ethical issues surrounding Big Data often relate to the complex question of data ownership. When data is collected, there are generally three roles that are associated with the process: who generates the data, who collects and analyzes it, and who it is being analyzed for.

With the Internet as an integral component of daily life in today's technologically advanced society, the generation of user data occurs daily. This results in a plethora of data sets that can be collected and analyzed in order to draw any number of conclusions about any number of possible theories or questions. While these large amounts of data are available, who owns this

data is a difficult and convoluted question. Because the data is accessible, those collecting and analyzing the data, or "collectors", often assume that it is available to be taken and analyzed.

Data ownership is a widely contested subject, and there is much debate in both the technological field and in academe regarding whether users or analysts own the rights to the data. Some scholars in the field of Big Data, as well as most of the population of users who generate the data, believe that they should have more control of data ownership; they are responsible for generating it, and as such, it is their data. However, those who collect and analyze the data believe that they have the legal right to assume ownership of this data simply because they are the ones who are using it. In this way, the data is not owned by those that generate it, then, as corporations control how this data is used and collected for "commercial ends" (Reyman, 2013, p. 516). Emphasizing the "commercial" aspect of data "transactions" alludes to users' potential for corporate profitability; user data can be sold to corporations like "property", which causes corporations to place their focus not on how to improve user agency and experience, but how to benefit from the appropriation of their data.

Privacy and Consent

Many websites often assume ownership of user data simply because they ask for consent to use this data. Custers (2016) uses the example of social networks like Facebook, Twitter, and LinkedIn to illustrate that many social networking sites imply consent simply by agreeing to a website's Terms of Service. He furthers that:

> By asking their users for broad consent, they create opportunities to collect large datasets for all kinds of business opportunities. The business models are usually of the type in which users do not have to pay for their accounts…usually consent is asked for when registering at the website. (p. 1-2)

The phrases “business opportunities” and “business models” suggest that data collection is a corporate venture, and identifies the language of online consent as part of the network’s business plan in order to appropriate user data in a way that might not be ethical, but is technically legal. Facilitating consent procedures ensures that consent does not hinder the availability of user data by deterring users from providing consent. Websites, and social media sites especially, create a sense of dependency in the user as they offer the use of their platform for free (Reyman, 2013). The user’s reliance on the service provided outweighs the caveat of data appropriation, and in the user’s mind, the reward becomes greater than the risk.

Even when requiring consent is made explicit, initially, consent is usually implied for an indefinite period of time (Custers, 2016). However, this does not take into account the fact that users may change their minds regarding consent as time progresses, and that level of technological understanding may also change, which could potentially alter a previous decision to consent (Custers, 2016). Reyman (2013) relates this to “the asserted distinction between content and data” (p. 525), and as such:

> Users maintain the rights to content shared on participatory Web platforms, but site owners unquestioningly assume control over the data attached to it. Such a distinction fails to recognize the interactive nature of production in social media spaces, in which user contributions of content and data are inextricable. (p. 525)

This “distinction” is implied, which results in a hierarchically skewed structure dividing users and corporations. Corporations then “assume” that they have control over the data simply because this implied role division decreases the likelihood that users will contest.

The idea of control can be further illustrated using Reyman’s (2013) example of geotagging, a process where “a date, time, and location are automatically attached to [a post],

and when generating data about your networks and activities, status updates and comments are composed" (p. 525). Geotagging represents the muddled boundary between "human experience" and "data"; technically the information produced is numeric in nature, but it is produced as a result of human experience. While translating experiences into data can make experiences more valuable by increasing their ability to be translated into different platforms, the issue lies in the way these same corporations appropriate the data and use it to facilitate marketing and advertising efforts.

The process of geotagging can be classified as "spatial Big Data" (Dalton & Thatcher, 2015, p. 4). However, this process is inherently commodifying as spatial Big Data is used for advertising by tracking users' locations on phones or their recently visited websites. Spatial Big Data allows corporations to engage in "personalized targeting" (Dalton & Thatcher, 2015, p. 4), which results in a "quantified individual" (Dalton & Thatcher, 2015, p. 7). The user's data becomes more valuable than the personal entity behind the data, reducing the user to a "number" that represents their online behaviours. These data are used for predictive analytics to calculate what behaviours users are most likely to engage in based on their previous behaviours, or the data that they generate. Dalton and Thatcher (2015) note:

> Both geodemographics and spatial Big Data assume that social identity can be reduced to measurable characteristics that can be algorithmically classified. Furthermore, as with social physics, this assemblage of personal data is predictive, or can be made to be so, commodifying it as valuable in marketing and ultimately making a sale" (p. 4).

Once corporations can predict human behaviour and quantify it into profitable information, they can sell this data to those who engage in targeted marketing.

Just because behaviour is quantifiable and "predictable", though, it does not mean that these predictions are precise, as "the accuracy of a company's output data is verified in competitive marketplaces, rather than more formal scientific or similar scholarly verification processes" (Dalton & Thatcher, 2015, p. 6). As long as data serves its market purpose, it does not necessarily matter whether it is accurate.

Tene and Polonetsky (2013) assert that "predictive analysis is particularly problematic when based on sensitive categories of data, such as health, race, or sexuality" (p. 253), and suggest that "even with non-sensitive data categories, predictive analysis may have a stifling effect on individuals and society, perpetuating old prejudices" (p. 254). Predictive analytics bring the past into the present space, and essentially make a judgement based on users' past behaviours. This does not allow for change or growth, which can propagate prejudiced attitudes so that they are ingrained in society, making this not only an issue of privacy, but that of identity.

While many websites and social networks address the issues of user privacy, information storage and dissemination, and consent in their privacy policies (Custers, 2016), this is not an adequate form of explanation and often seems to serve purposes of legality rather than act in users' best interests. Often, the language used in these policies is much too convoluted for users to understand, marking a division between those who generate the data and those who analyze it, as data analysts often have the ability to write and understand the terms presented in these policies (Custers, 2016). Many times, users provide consent without reading through the actual policy, and they abide by the request because they are asked to do so (Custers, 2016). This is concerning because users inadvertently relinquish their rights to their data in favour of expediting the process of agreeing to the terms of the document. In many cases, these documents are so tedious to read and understand that users bypass them completely. One potential way to

remedy this is by "strengthening consent mechanisms", but this may be counterintuitive as users are more likely to simply consent when confronted with convoluted consent procedures (Schermer, Custers & van der Hof, 2014, p.172). For true consent to be given, users must be provided with all of the necessary information as to what they are consenting to, they must consent freely based on their own decisions or values, and they must have the intention to consent (Schermer, Custers & van der Hof, 2014). It is difficult to prove that all three of these requirements have been met when requesting consent, and therefore it is difficult to obtain the truest form of consent that most accurately reflects users' interests and well-being.

How, then, can consent be requested and obtained so as to make the process more easily understood by the user, while still maintaining or increasing efficiency? Custers (2016) discusses the potential of data reuse to protect users' privacy while still benefitting society as a whole. Data reuse implies using data after the initial process of collecting, analyzing, and storing it (Custers, 2016). This, though, is different from data recycling, which is repeatedly using the same data for the same purposes, or data repurposing, which is when data collected by one company is sold to another for a profit. In the latter example, this is most likely done to serve marketing purposes. Data repurposing can have severe legal consequences, as the data is being evaluated by those who may be unauthorized to use it for purposes other than what it may have been originally collected for. This implies a blurring of the lines of consent.

The main idea of Custers' theory of data reuse is to seek consent initially, and then act on this implied consent for subsequent actions using this data. While this will expedite the process of analyzing the data, and will be less likely to deter users from consenting because of the lack of convoluted consent procedures, it still seems to embody these ethical issues that were previously discussed as to whether implied consent is an adequate form of consent. Because it is not

updated regularly, and because it is not made explicit each time the data is used, there is an increased possibility of the data being used without the user's consent.

In order to prevent user identification in data collection, some scholars have proposed the adoption of the "Do Not Track mechanism", which allows users the option to conduct online behaviours without providing access to the data that they generate (Reyman, 2013, p. 528). However, this "may prevent users from being able to use certain social media features and services" and "fails to offer users an opportunity to shape the development of technologies and policies governing use" (Reyman, 2013, p. 528). Instead, "users themselves need to advocate for a future Internet based on increased user participation in the development of data-mining practices and policies, and for a more balanced ownership structure to manage user data" (Reyman, 2013, p. 529).

Privacy preferences represent ethical issues in and of themselves, "as privacy preferences can be used for personalization or profiling", which can "[yield] less privacy rather than more privacy" (Custers, 2016, p. 4). The question then arises of how to remedy this situation so that user consent always reflects users' most current beliefs. It can be suggested that consent to analyze user-generated data should have an expiry date, but even then, "when consent for processing personal data expires, the anonymized data may still be used for profiling and statistics" (Custers, 2016, p. 4). If the data still exists, the shield of anonymity can often allow analysts to assume that data is "fair game", and appropriate data for purposes other than those originally specified.

The idea of anonymized data alludes to the issue of "dark data", or "non-used raw data" (Zeller, 2014, p. 265). This data, although it has not been used, is still accessible, and reflects certain information about users that may have been generated unwittingly. This raises the

question of what happens to this data; if it has not been used then it is technically not violating any privacy standards, but the fact that it is still accessible means that it has the ability to do so.

User Commodification

User commodification is described as "datafication", or "the process of rendering into data aspects of the world not previously quantified" (Kennedy, Poell & van Dijk, 2015, p. 1). Datafication can also be seen as the "ubiquitous quantification of social life" (Couldry & van Dijk, 2015, p. 4). Datafication encompasses an area of life where the "social" seeps into the technological and the material, making access to information gathered from tracking and aggregating online behaviours a "good" that can be purchased. However, this raises the issues of user agency and quantifying the social sphere, or using information for a purpose other than that for which it was generated (Pybus, Cote & Blanke, 2015). By combining the words "data" and "commodification", the term "datafication" implies a concept that uses user-generated data as a form of information gathering to make a profit. Couldry and van Dijk (2015) assert that "manipulation and commodification are not the *consequences* of datafication; all three mechanisms are intrinsically intertwined in the configuration of the platform ecosystems preconditions of its use and exploitation…" (p. 3). Essentially, datafication does not result from commodification, but is rather a form of commodification adapted to the new world of the online space.

Lewis (2015) explains that Big Data sets "are typically collected by people other than researchers with incentives other than the advancement of science. As such, not only are participants' natural behaviors constrained—they are constrained in ways that are as commonly distortive as illuminating and frequently altogether hidden from the researcher" (p. 2). By describing "illumination" as "distortive", Lewis alludes to the common misconception that Big

Data sets are inherently accurate and representative of all members of a population (boyd and Crawford, 2012). Because the data sets are so large, it is easy to assume that they are representative. However, there are many flaws embodied by the collection and analysis process, and this is not necessarily the case. For example, Big Data sets are compiled by collecting and aggregating data from users' online behaviours; while the Internet is, arguably, an essential component of daily life to many in today's society it is not yet universal, and as such not every member of every population can be represented by the data that comprise Big Data sets.

User commodification not only relates to the loss of the user's individuality, but the collection of individual user data for profit, as "harvesting these data are often directly connected with profit generation, with little or no attention to the difficulties that are connected with connecting, processing, and finally using these data" (Zeller, 2014, p. 156-7). These companies responsible for "profit generation" can be classified as "social actors", or "actors with social ends over and above the basic aim of generating and analysing data (usually for profit)" (Couldry & Powell, 2014, p. 2).

The data users generate are often understood as and referred to as a product. While the information is produced, the language of the "product" implies a corporate nature in which it can be bought and sold; essentially, it is commodified. As such, "the information industry continues to treat personal data as goods. As it stands, anonymous Internet users are not lucrative and personal data have a unit price" (Peacock, 2014, p. 4). Reyman (2013) states that "the type of property that is produced—user data—is not understood as holding intellectual and creative value, but only as a commodity to be bought and sold" (p. 526). This implies that the data sets lose the individuality that provides the data with contextual richness and instead serve as purely a means to be profited from. Further, "data is treated as unclaimed property free for the taking, or

as by-products of technology use to be collected and used by technology providers toward largely self-serving ends" (p. 528). In this way, Reyman illustrates that not only is data seen as "property" and not as creative "voice" (Couldry & Powell, 2014, p. 4), but it is assumed to be the "property" of collectors and not of users. While most data are user-generated, data collectors and analysts see data as just that—numbers that do not have a human "voice" or element, but are purely technologic and therefore available to be appropriated.

Many privacy policies and "terms-of-use statements for social and participatory Web technologies" take the role of data generator away from the user and instead ascribe it to technology. However, "separating an author-user from the productive act of generating user data…privileges a structure in which the technology provider automatically assumes ownership and control over user information" (Reyman, 2013, p. 526). This prevents users from taking control of their data or exhibiting any rights of authorial ownership, and instead gives the agentic role to the corporations that have access to these data analysis technologies (Reyman, 2013). Essentially, "data becomes, at its inception, free to be appropriated and controlled by those responsible for the technology and not by users" (Reyman, 2013, p. 526).

It is difficult for users to change social media platforms, and even so it is unlikely that the privacy policies of these different platforms would be much different (Peacock, 2014). It seems as if users lack choice; if they choose to use a social media platform, then they must consent to the appropriation of their data in order to have the opportunity to do so (Peacock, 2014). This encapsulates the necessity of changing and improving these measures in order to offer users more control over their data.

Loss of Individuality

The idea of datafication results in a loss of user individuality, reducing users to the aggregate data that they generate. Borlin and Schwarz (2015) further that "the techniques for aggregating user data…build on large, algorithmically produced aggregates of information that are the basis for the construction of audience commodity" (p. 1). In amalgamating the data generated by each individual user, the singular "user" and the corresponding originality of the data is transformed into the plural "users", consequently negating the user's status as an individual and instead redefining the user within the broader context of the data itself. This undermines the user's individuality, and instead defines the user's data as a means to an end of collecting a large amount of information. The assertion of "voice", or the user's ability to tell their own stories and give life to their own accounts, is taken away when individuality is subsumed by the whole (Couldry & Powell, 2014, p. 4).

This lack of voice is echoed in the way that Big Data "document[s]" users' behavioural patterns rather than "report[ing]" about them (Lewis, 2015, p. 2). It records things in real-time, giving the data an almost human aspect. In this way, the user is rendered obsolete; even though it is the user who is generating the data, new technologies that allow data to be recorded as it happens appropriates the human experience, resulting in a lack of clarity as to user rights and roles.

Summary

The emergence of the Internet, along with social networks, has "redefined 'the social'" (Couldry & van Dijk, 2015, p. 2) and made it quantifiable by offering numerous ways of gathering and analyzing data. This results in a "measurable kind of "'social'" (Couldry & van Dijk, 2015, p. 5). The inverted commas around the word "social' suggest that what is being

referred to here is the concept of sociality; essentially, the social sphere is being redefined in order to make it more computable. Embodied in the phenomenon of Big Data is this new understanding of sociality, which can prove a valuable resource in understanding how the social is redefined and reinterpreted. However, this data must be protected. Reyman (2013) asserts the importance of user data, and understands it as "not merely a technology by-product to be bought and sold; rather, it forms a dynamic, discursive narrative about the paths we have taken as users, the technologies we have used, how we have composed in such spaces, and with whom we have participated" (p. 516). It is important to develop new ways to understand and define Big Data to continue to expand the scope and definition of knowledge, but to do so in a way that protects users' privacy, offers them a claim on the ownership of the data that they generate, and resists their commodification.

Research Questions

One of the issues that arises in the process of data collection and analysis is who owns the data, and who has control over this data. It is realistic to acknowledge that Big Data will most likely continue to be used for marketing or promotional purposes in the future; as long as corporations continue to have access to these large data sets, as well as the ability to aggregate and analyze them, this practice will continue until a newer and more efficient process comes to fruition. It is necessary, then, to determine how data can be used in a way that affords users a more active role in ownership, and allows them to have more control over how the data they generate are used.

Big Data has proved beneficial in many ways; while it has revolutionized the scope of information and knowledge that is available to us today, the processes of data aggregation and appropriation embody some major ethical concerns, as scholars like Couldry and van Dijk (2015) indicate. My goal is to determine the context in which Big Data is discussed by scholars in the social sciences, and what they are saying about how society can continue to benefit from Big Data to gain knowledge, obtain information, and answer questions while keeping these ethical issues in mind and allowing users more options for individuality, privacy, and agency. In an effort to gain insight into the various issues that I have outlined, I pose three main Research Questions that will help to shape the direction of my research in my MRP:

1. RQ 1: How can users play a more active role in data ownership when it comes to data appropriation?
2. RQ 2: Does a more active role in data ownership include a more explicit way to require consent, and what role do more explicit consent measures play with regards to data ownership?

3. RQ 3: To what extent can the individuality of users who generate data be preserved in the process of Big Data collection and analysis, and how can users be protected from commodification?

Data Collection Method

Sources of Information

Big Data is a relatively new field in academic study and is constantly influenced by the changes and innovations in data collection, interpretation, and analysis that occur on a daily basis. I have drawn my research from journal articles[2]; while there are many insightful book chapters dedicated to exploring user rights and roles in the context of Big Data, the accelerated process for publishing an online journal article ensures that the information obtained from these sources is the most up-to-date. While blog entries can also be a good source of the most recently published information, journal articles are more likely to be peer reviewed, ensuring their reliability. I have found Sage Publications' online journal, *Big Data and Society*, an invaluable resource as it shows both past and current (2016) research being done on Big Data. I have also used the online journals *Social Media and Society*, *Information, Communication, and Society*, *Communication and Mass Media Complete*, *JSTOR*, and *Open Doar*.

Another method that I found helpful was taking note of some of the authors that I cited most frequently in my Literature Review, and those whose names I recognized from my research as being prominent and frequently published scholars in the field of Big Data, like Claudia Aradau, Bart Custers, and Nick Couldry. I looked at their list of publications, and input the titles of the articles that I thought would be most useful into the RULA search catalogue. After reading the article's abstract and determining whether it discussed issues that were pertinent to my research questions, my decision to include the article in my sample group depended on its relevancy to my research questions.

[2] See Appendix (p. 56), which offers a complete list of all 20 journal articles referenced in the Findings section.

Sampling Strategy

To formulate my sample group, I used a total of 20 online journal articles published between 2013-2016. I initially proposed that the sample group would not make reference to any of the articles that I had previously cited in my literature review, but as my research progressed, I discovered that since Big Data is still a new area of scholarly analysis, especially in the social sciences, there were relatively few articles published within the field of Big Data scholarship that did not focus on its scientific or technologic applications. Within the general scope of published articles that focus on Big Data, most discuss its effects as a general phenomenon and not its specific implications on privacy or user rights and roles. I originally planned to search a combination of the terms "Big Data" with "user agency", "commodification", "privacy", and "consent", for a total of 5 search terms in order to generate a group of potential results. However, in order to find articles that specifically addressed my research questions, I found that I had to both broaden the scope of my search and word my search terms differently, as some of the search terms that I had originally proposed to use, like "agency" and "commodification", did not return many (if any) search results. This may be because although these terms are prevalent in the social sciences, the study of Big Data in this context is so new that these are not search terms commonly associated with Big Data yet.

In order to determine which articles to select for the sample group, I searched Sage Publications' online journal, *Big Data and Society*, as well as *Open Doar*, *JSTOR*, and the RULA search catalogue. I then read article titles and abstracts in order to determine which articles were most relevant to what I wished to find out through my research questions. This allowed me to include the articles that were most pertinent to my research questions in my

sample group, and also helped me to reduce the possibility of a skewed data result by eliminating those articles that were not as relevant.

Method of Analysis

I conducted my research using qualitative content analysis. My main goal was to determine in what context these search terms were discussed in article titles, abstracts, discussion sections, and conclusions. I wanted to see what was being said about privacy, consent, and user rights and roles in the context of Big Data in the social sciences to determine what themes were common in the literature, how scholars addressed each other's work, and where there were potential ethical issues that needed to be further expanded upon. Specifically, I wished to determine what the scholarly literature said about Big Data and user commodification and how users can develop a more agentic role in the data appropriation process, how consent measures can better protect user privacy, and how user individuality can be preserved in a process that seeks to predict general patterns through the analysis of aggregate data.

To answer RQ 1, I will be looking at how users can claim ownership of their data; corporations simply appropriate this data, and it is important to discern different methods of reclaiming user agency, and for users to gain a better understanding of how they generate this data and in what way it is being used. To answer RQ 2, I will consider whether more explicit consent measures will better protect user privacy, how improved privacy practices empower users, and whether consent measures positively affect user roles in the data generation process by allowing users to claim ownership of their data. To answer RQ 3, I will examine various ways corporations can protect users from commodification, if there are any methods that allow users to protect themselves, and how users' individuality can be preserved within a system that seeks to identify general patterns and is predisposed to reducing people to the data that they generate.

Findings

With any new technological phenomenon comes a certain level of confusion or lack of understanding—many times the techniques and processes surrounding these phenomena are described using terms and language and that are not universally understood. However, when the phenomenon is Big Data, the level of uncertainty rises due to its multifaceted portrayal and understanding across disciplines.

While Big Data itself is a relatively new phenomenon, and has usually been discussed in scientific and technological contexts, in recent years it has infiltrated the academy and has begun to be discussed in the social sciences, not just for its technological innovation, but for how it is redefining the way we understand our social sphere.

Central to the study of Big Data in the social sciences are the issues of user rights and roles, privacy, and commodification. This phenomenon embodies a dichotomy of "human" and "machine", and as its presence in society increases it is important to understand how and why these issues occur, and discover how they can be remedied.

RQ 1: How can users play a more active role in data ownership when it comes to data appropriation?

In answering RQ 1, one of the most discussed methods to create a more active user was to increase transparency measures, which would allow users to become more aware of the processes involved in data aggregation and analysis. Some scholars (Kennedy and Moss, 2015; Baack, 2015) believed that if users were more informed about these processes, they would be more likely to make informed choices about whether or not to provide their data. They would also feel more like *a part* of the process instead of *apart* from it, allowing them to ascertain more of an agentic role in data ownership. Interestingly, these same scholars (Kennedy and Moss,

2015; Baack, 2015) also acknowledged the limitations of increased transparency measures, saying that that there is no verifiable way to guarantee universal understanding among different levels of technical literacy. Overall, transparency was not touted as a universal solution, but was described as effective when implemented in conjunction with other measures that benefitted users.

The process of data mining contributes to the debate surrounding data ownership. These "automated methods of data extraction" (Alim, 2014, p. 4) do not allow users to have a great deal of control over their own data. Because these methods are mechanized, the "traces of data people leave behind are often unconscious and not meaningful to them" (Baack, 2015, p. 2); users may not even know that they are generating data, let alone that it is being appropriated by the corporations who collect it. This makes it difficult for users to claim ownership. Within the literature, many authors agree that "transparency is the most widely discussed way in which the public can have more control over data mining" (Kennedy and Moss, 2015, p. 5), which implies decreasing the obscurity surrounding data mining practices. Advocates for increasing transparency believe that "making data-mining algorithms and processes public in this way would help to facilitate public understanding, scrutiny and debate about the political effects of data mining, and allow the public and groups acting on the public's behalf to examine and contest data-mining practices" (Kennedy and Moss, 2015, p. 5). Essentially, transparency not only increases the agency of the individual user with regards to data ownership, but of the collective public. One of the predominant themes in the literature pertaining to transparency was that greater levels of transparency contributed to greater levels of users' understanding of the processes involved in the appropriation of their data. Increasing users' understanding is the difference between data appropriation by corporations, and data contribution by users. Increasing

the level of users' understanding allows them to feel like they have a stake in the ownership of their data, as they are more likely to feel connected to the information when they understand it as their own.

Alim (2014) indicates that most corporations place the onus on the user, believing "it is the user who has the right to decide how they use the service, provided there are no violations of laws" (p. 67). If users can decide how they use the service and choose whether or not to provide their data, then they can be seen as agents in the process of data collection. However, many of the issues raised in the scholarly literature indicate that the question of data ownership is much more complicated.

Transparency does not only apply to making processes more apparent, but also to making users cognizant of the data that they are contributing, and the potential implications of this contribution. Reyman (2013) outlines that "although users are aware of the content they are contributing online—when sharing a photo, writing a blog post, updating a status, or entering a 140-character tweet—many are unaware of the additional, hidden contributions of data made with each act of participation" (p. 514). If users are not aware of the fact that their online behaviours create a data trail, then they are unable to claim ownership of data that they do not know exists. For a more agentic role in the data ownership process, efforts to make processes more transparent will allow users to exert more control over their data through their increased level of knowledge and understanding as to where and how their data is being used.

Another potential method discussed in the literature for allowing users a more agentic role in data ownership is that of adopting some of the principles of open source culture. Baack (2015) explains that "open source culture is associated with a transparent and collaborative form of governance that might support agency (p. 2). He outlines "the first and most fundamental

modulation of open source culture", which is "to conceive raw data as source code that should be shared openly to allow others to interpret it and to generate their own knowledge from it" (p. 4). The reasoning behind open source culture is that sharing raw data makes the process of interpreting it transparent, which then "mak[es] the biases of this data transparent" (Baack, 2015, p. 4).

Germany's Open Knowledge Foundation established a universal definition for "Open" specific to data, "according to which data is 'open' when it can be accessed, modified and shared by anyone for any purpose without restrictions" (Baack, 2015, p. 4). While it is universally acknowledged that personal data or data with security implications should not be made open (Baack, 2015; Kennedy and Moss, 2015), open source culture perpetuates the idea that data should be made accessible and available to everyone, eliminating the barriers that contribute to hierarchies between those that have access to the data and those that do not. Baack (2015) describes how open source culture addresses these hierarchies:

> Even though the idea behind the democratization of information is to potentially allow everybody to interpret raw data, activists are well aware that the average citizen does not have the time and expert knowledge to do so. They recognize that their vision of empowerment through open data can only be realized with intermediaries that make raw data accessible to the public. (p. 6)

In this way, when adopting some of the principles of open source culture, these "intermediaries" would bridge the gap between those with access and those without access, make the data accessible, and work toward disrupting these hierarchies and creating a more universal level of access.

In some instances, even researchers can be prone to disregarding users' individual privacy rights. Alim (2014) offers Facebook as an example:

> In 2012, Facebook updated their privacy policy to prohibit the collection of user information via automated means, such as harvesting bots, robots, spiders or scrapers without permission from Facebook. If information from users was collected, consent from the user had to be obtained. Also, it had to be made clear that the researcher (and not Facebook) was collecting the information. (p. 67)

This is also exemplified through "socialbots", or "dynamic fake profiles", which "can be used to extract a user's private information. Unlike a fake profile, socialbots send out friend requests. For a researcher, if users accept friend requests from a socialbot, this gives them the chance to analyse a user's private information as well as their privacy settings" (Alim, 2014, p. 8). The ambiguity lies within the question of whether or not "researchers consider online social media profiles as human participants" (Alim, 2014, p. 12); because the profile is technological in nature, it becomes easy to distance its two-dimensional representation from the human responsible for its creation. While this is also an important issue, limitations in scope do not permit its discussion in this paper.

While there are a few drawbacks to adopting the principles of open source culture, like the assumption of "voluntary participation of citizens", the relative idealism of assuming universal openness when in reality there are still differences in technological literacy among people (Baack, 2015, p. 8), and potential issues with access (Kennedy and Moss, 2015), "sharing raw data should help citizens to better understand and control their governments and to be more active and engaged in their local communities" (Baack, 2015, p. 2). Adopting some of the

principles of open source culture embodies a "type of transparency" that can allow users to become active agents in the ownership of their data (Baack, 2015, p. 4-5).

However, the scholarly literature alludes to one potential issue with user agency. Most of the literature discusses user agency with an ideal user in mind, one who wishes to have a more agentic role in the collection and analysis of their data and wants to be involved and included in the process. While there are many users who may feel this way, this outlook neglects to address the users who know their data is being appropriated, and simply do not wish to do anything about it. There are many reasons why users may be unconcerned with the appropriation of their data: they may not understand what their data is being used for and are consequently unaware of the implications of its appropriation, they may feel as if their role as users is insignificant in comparison to that of corporations and feel powerless to change rules and regulations governing the appropriation of user data, or they might be informed and simply not care, among others. While the scope of this paper does not allow for an argument that addresses each of these population groups individually, increased transparency measures can help those who are unaware or those who feel insignificant by allowing them to feel more involved in the process.

Making data mining practices more transparent would involve "requiring data-mining companies not just to show the public what they are doing, but to tell publics what they are doing, why, and with what effect" (Kennedy and Moss, 2015, p. 6). Transparency is not a "fix-all" that allows users to automatically see and understand the processes involved in collecting and analyzing their data, resulting in the compelling need to claim ownership. Rather, it is an option that can provide the greatest benefit to users when used in conjunction with other methods, like making consent measures more explicit.

RQ 2: Does a more active role in data ownership include a more explicit way to require consent, and what role do more explicit consent measures play with regards to data ownership?

In answering RQ 2, most scholars (Andrejevic, 2014; Custers, 2016; Book and Bronk, 2016) believed that, like in RQ 1, more explicit consent measures attempt to offer users a better understanding of their data, which makes them more likely to make informed decisions and feel included in rather than excluded from the process. However, there seemed to be a fine line between what constituted "more explicit consent measures", and what was an excess of information that risked confusing and isolating users. The general idea is that in asking for consent, corporations attribute users with ownership of their data by asking them permission to use it. However, many times the level of agency ascribed to users by this process is limited by the lack of choice of whether or not to consent, and again alluding to RQ 1, there is relatively little users can do about this. The question then becomes centred on what corporations can do to build accountability, and what society as a whole can do in a collective effort to regain agency.

Much of the scholarly literature acknowledges that when users are made aware of the processes surrounding the appropriation of their data, they want to play a more active role in protecting their privacy. In an Australian-based study conducted by Andrejevic (2014), results "revealed a very high level of support for stricter controls on information collection" (p. 1683). In fact, 95% of respondents supported "do-not-track" legislation (p. 1678). What's more, 56% "opposed customized advertising based on tracking" (p. 1678), 96% expressed their desire for a requirement to delete personal data upon request, and 95% wanted "real-time notification of tracking" (p. 1683). Almost three quarters of respondents felt that "they needed to know more about the ways websites collect and use their information" (p. 1683). This draws attention to the

issue of hierarchical structures discussed in RQ 2, and shows that at the user level, there is still much work to be done to increase levels of data transparency and accessibility.

A major theme in the scholarly literature was the necessity to improve consent measures. Custers (2016) notes that "when discussing consent, it is usually assumed that consent is only valid when it is informed consent. This involves that the person asked for consent should be properly informed of what exactly he or she is consenting to and to some extent (made) aware of the consequences such consent may have" (p. 2). An example of what should be specified to users is "information about which data are collected, used and shared, for which purposes the data are used, which security measures are taken, information about who is processing the data and who is accountable and information on user rights and how they can be exercised" (Custers, 2016, p. 2). Unfortunately, many authors note that this information is generally not shared with users, and if it is, it is not provided as explicitly.

The lack of disclosure by corporations as to where and how this data is being used, and even that it is being collected, is indicative of the opacity that has become a trademark of large-scale data analysis, especially in terms of marketing and advertising. One reason corporations may be reluctant to be forthcoming with this information could be attributed to a "don't ask, don't tell" mindset—there is no need to address any concerns regarding the data if users are not asking questions, which then expedites the process of "aggregating, analyzing, and selling" this data for a profit through data mining (Reyman, 2013, p. 514). This, though, becomes a paradox: increased transparency measures are necessary to allow users to become more engaged with the processes surrounding the collection of their data, which will then contribute to a greater stake in its ownership; however, the current processes surrounding data mining largely benefit

corporations and are shrouded in opacity, making it difficult for corporations to change established patterns to allow for more transparency.

It must be acknowledged that corporations are inevitably at the top of the hierarchical power structure, and changing their privacy and consent structures would greatly benefit users; users can only do so much to act on their own behalf, and sometimes it is futile to change one's actions within the system when the system itself is flawed with no prospect of repair. How, then, can corporations be motivated to change these structures when they rely on accessing user data in order to fulfill their own objectives of selling this information to ad brokers and other corporations for marketing and advertising purposes? Martin (2013) suggests emphasizing "brand and reputation" as a viable solution, in that users "could work online by creating a network that identified Web sites, allowed brands and reputations to develop based on feedback of users and experts, and gave users and Web sites a mechanism to signal preferences at a fine grain level" (p. 44). She uses the example of customer review sites like Angie's List and Trip Advisor, which evaluate a company's services for a public audience. The way that these sites are structured encourages a corporation to be held accountable for providing the best services to users; a large part of developing a successful business is how it is perceived by the public, and if corporations are potentially "written up" for impinging on user privacy rights and commodifying users, then the idea is that the public will not support them and they will consequently lose revenue. Corporations "have the power to unilaterally establish and enforce any restrictions on mobile advertising that they see fit. They also stand to benefit if their platform is perceived as more secure or respectful of user privacy than their competitors" (Book and Bronk, 2016, p. 27). This model "rewards firms who build a reputation around respecting privacy expectations.

Importantly, firms must now understand the evolving privacy expectations of users for different contexts rather than rely upon adequate notification" (Martin, 2013, p. 47).

Martin (2013) suggests that "a focus on **brand name and reputation** [*bolded in original*] shifts the exchange away from the one–shot dilemma governed by an explicit contract to a long-term relationship or repeated transactions governed by trust" (p. 34). This would also allow users to check on the reputation of a website in meeting privacy expectations before entering into a transaction (Martin, 2013, p. 44). In this way, users have more agency and control in the process, as they are placed in a position of power through their role as evaluators.

Book and Bronk (2016) refer to those who collect and analyze the data as "regulators" (p. 37) and outline a few different ways in which they can apply better transparency measures, like providing users with the details of what data they are collecting and for what purpose, or offering them the option to opt-out of data gathering processes without inhibiting online accessibility (p. 37). Alternatively, consent measures can be made more explicit by applying "opt-in" measures when collecting "personal" or "sensitive" information (p. 37), which would require users to consent to the use of their information instead of corporations assuming their consent is implied through their use of an online platform's services. This, though, sparks a debate about what constitutes a public or private space, and general or sensitive information, as "on the one hand, people may produce data in public spaces but have strong perceptions of privacy. On the other hand, people acknowledge that communication is public but the context it appears in implies restrictions" (Alim, 2014, p. 69). While there is no concrete definition of public or private in the context of Big Data as of yet, because the distinction between sensitive and general information is so subjective, it is likely that these concepts will continue to inspire debate. Nonetheless, opt-in measures can still be a simple and effective way of informing users as to how and where their

data is being used, and asking for their permission to use this data. Measures to "opt-in" or "opt-out" afford the user a more active role in data ownership, as the act of asking for permission implies that the creation, and therefore the ownership, is that of the user and not the corporation who is appropriating the data. They also offer users a clear choice between whether or not to contribute their data, which can eliminate some of the ambiguity surrounding existing consent measures.

Transparency in consent measures cannot stagnate, though, and "need[s] to be adapted because, as social media grows, technology and user expectations change" (Alim, 2014, p. 68). While measures change and adapt to new technologies and circumstances, users also have the responsibility of monitoring their own online privacy to the best of their ability, as they can with some social media sites like Facebook, and ensure that only those whom they wish to be privy to their online social media presence are so (Alim, 2014). It must be acknowledged that there is only so much users can do to regulate their own online privacy, as much of this data is appropriated by regulators. However, if users choose to remain vigilant and actively seek to ensure that their online behaviours remain, to some extent, private, then this can help in the process of remedying this lack of transparency.

Consent is integral to the analysis of privacy in the context of Big Data. As noted by Custers (2016), oftentimes corporations assume that users "have given their consent to the collection, sharing, and processing of this data through their acceptance of a privacy policy included with the application" (Book and Bronk, 2016, p. 3). It can also be assumed that "users who do not want their information collected in this way may simply choose not to install the app in question"; however, many users rely on these "essential" (p. 3) technologies and use them regularly. As such, "mobile apps (and their associated ads) have moved from being another

feature in the software marketplace to becoming a part of the fabric of society, and hence have placed themselves in need of greater regulatory scrutiny" (Book and Bronk, 2016, p. 3). Also, Martin (2013) notes that "in terms of online privacy statements, organizations are limited in effectively communicating to consumers how information flows even when individuals do read privacy notices" (p. 17). It is here that the question of adopting more explicit consent measures and how these affect user rights and roles is most apparent, as sometimes, the word "explicit" tends to imply offering more information to make consent measures clearer.

More explicit consent measures do not have to imply a higher degree of information, though; offering more information can be confusing and isolating to the user due to the "difference between [the] knowledge of those who collect and analyze data and those who generate it" (Custers, 2016, p. 3). In some cases, privacy protection measures can potentially negate the benefits of Big Data. Users may not understand these newer, more complex policies, and may simply refrain from using the site or consent without acknowledging the policy. Consequently, this can prevent the data from ever being used at all (Custers, 2016). To remedy this, Custers suggests the possibility of "assuming informed consent", instead keeping those methods that may lack transparency or prove challenging to gain consent bound in more explicit measures (Custers, 2016, p. 4).

Informed consent was also described in the literature as "notice and choice", which "allow[s] for heterogeneity in privacy expectations — each exchange develops a particular set of rules governing how, when, why, and where information is used" (Martin, 2013, p. 2). Some authors were critical of notice and choice's ability to strengthen transparency and protect users' privacy. Rubinstein (2013) states that "empirical studies show individuals neither read nor understand privacy policies, which anyway rely on ambiguous language, and are easily modified

by firms" (p. 5). While it is a valid concern that users may not read or understand privacy policies, notice and choice are beneficial in bridging the gap between being excluded from the data collection process and included in it. This affords users a more agentic role in data collection, and allows them to better understand potential privacy implications.

Notice and choice has the potential to increase transparency and protect users' privacy; it ensures that users "fully understand the terms at the point of the exchange", and is a way to "empower individuals" and "give control to consumers" (Martin, 2013, p. 2). Pertaining to the question of a more agentic role for users in data ownership, notice and choice affords users the opportunity to decide whether they wish to contribute their data. Providing proper notice as to how users' data will be used and what it will be used for allows for better consent measures to be in place without necessarily making them more complicated. Understanding where their data is going allows users to have a more active role in the data collection and aggregation process, as it gives them the power to make their own choices. Martin (2013) also believes that "notice and choice can be viewed as consistent with privacy scholarship by allowing for heterogeneity in privacy expectations — each exchange develops a particular set of rules governing how, when, why, and where information is used" (p. 2). Ultimately, more explicit consent measures allow for a more active role in data ownership.

Notice and choice may not always be effective, though, and in order for them to be so, both users and corporations need to understand the terms of the agreement, and a strategy addressing potential costs and shortcomings needs to be in place (Martin, 2013, p. 14). It seems that among users, "considerable agreement exists that transparency and choice has failed" (Martin, 2013, p. 5). Users are unaware of how their information is being used, and have no control over the process. While notice and choice does not allow them to control how the data is

being used, an informed and educated user can at least understand more about their data; creating a more informed user through improved notice and choice measures, then, will give users a greater sense of control.

While these concerns must be taken into account, it is important to note that most authors agree that notice and choice is not a universal solution to the problem. Martin (2013) cautions that "notice and choice as the sole mechanism to address privacy fail where similar explicit contracts fail: where the environment renders the transaction costs of the exchange too high. High information asymmetries, enforcement costs, and uncertainty combine to make the online environment hostile to effective explicit contracting" (p. 26). There is also an issue of the differences in access and technological literacy between users and data analysts, and "even if users had access to their own data, they would not have the pattern recognition or predictive capabilities of those who can mine aggregated databases" (Andrejevic, 2014, p. 1674). As such, notice and choice is not the definitive answer to improve consent measures, but needs to be supplemented by alternative measures to prove most effective as a way to protect users' privacy and offer them a more active role in data ownership (Martin, 2013).

As an alternative to notice and choice, Custers (2016) proposes the introduction of consent expiry so that "consent, when not renewed, expires after some time" (p. 4). He also notes that it is important to "limit the duration of consent to a maximum amount of time", and suggests "two or three years" as an appropriate target (Custers, 2016, p. 4). This allows users the option to change their minds, and the time to better understand how and why their data is being used (Custers, 2016). It also offers users the chance to re-evaluate whether they want their data to be used, allowing them to exert agency through the power of making this decision (Custers, 2016).

As with any proposed solution, there are some potential drawbacks to the "partial consent" (Custers, 2016, p. 4) in consent renewal processes. For some users, renewing their consent may just be another check-box to click on without reading it (Custers, 2016, p. 4). Moreover, "in case of software updates, providing new consent for a new purpose does not always imply that the previous consent for other purposes ends. New consent can be additional consent instead of revised consent" (Custers, 2016, p. 3). Deleting information from databases is often a difficult and complex process:

> [Deleting data] would involve the collection of a lot of metadata on which personal data can be used for which purposes before which expiry date. In fact, such metadata may also reveal privacy preferences of data subjects, yielding less privacy rather than more privacy, as privacy preferences can be used for personalization or profiling. (Custers, 2016, p. 4)

To remedy this, Custers (2016) suggests that "expiry dates may still be helpful in those situations in which users no longer actively use their accounts. Their inactivity is then automatically interpreted as a revocation of their consent, blocking further use of their data" (p. 4). Although these drawbacks must be considered, overall this concept has the potential to greatly benefit users, and the process causes users to become more engaged in the collection of their data.

The idea of the "Do Not Track" mechanism, which was briefly explained in the Literature Review, is discussed in the literature as another possibility for improving consent measures. The Federal Trade Commission's "'Do Not Track' report and commonly accepted practices are seen as attempts to recommend industry best practices" and focuses on "recommendations at the browser level, such as adding tracking protection or the voluntary

conformance to standards supplements work with the Better Business Bureau to highlight best practices within a notice and choice approach" (Martin, 2013, p. 32). There are also phone applications that can be installed to detect which applications appropriate users' data and impinge on their privacy rights (Martin, 2013). Having these applications on a cellphone allows for measures to be implemented at the user level instead of relying on a corporation to self-impose improved consent measures. This affords users the control to determine which applications they wish to use, at an accessible level.

Rubinstein (2015) proposes a number of alternative ways in which user privacy and consent measures can be strengthened. He suggests "selective disclosure, ie, the ability of customers to share their data selectively, without disclosing more personal data than they wish to" (p. 32). An example of selective disclosure is through opt-in or opt-out measures, where users can choose whether or not to consent to providing their data. He also lists "identity management, which handle tasks such as the authentication and use of multiple identifiers while preventing correlation unless permitted by the user" (p. 35), and "data-portability, ie, the ability to move all of one's data from one provider to another using standard data formats and interface protocols" (p. 37). Like Martin (2013), Rubinstein emphasizes the need for corporations to rely on "accountability and enforcement, ie, accountability for protecting and securing personal data in accordance with the rights and permissions established by agreement and/or enforced by tagging mechanisms; and enforcement under self-regulatory guidelines and legal mandates, both backed by comprehensive auditing" (p. 38). These measures have the potential to "ensur[e] a much higher degree of transparency than would be achievable with the opaque data stores maintained by businesses or third parties" (p. 39), either when implemented individually or together.

Another way to improve user privacy is through the use of "PDSes", or Personal Data Services, that offers a "secure data store for a wide variety of personal information" for everything from licenses to passwords (p. 29). Rubinstein (2015) explains that "PDSes treat FIPs [Freedom of Information Practices] not as externally imposed constraints that must be balanced against business objectives, but as a set of organizing principles, which guide business objectives and thereby set design goals from the outset" (p. 39). Rubinstein also asserts the need to establish PDSes with "the highest level of security", and explains that to be effective, "(i) all personal data must be encrypted both in storage and during transmission; (ii) all encryption keys must be stored outside the PDS; (iii) all metadata must be encrypted and digitally signed; (iv) all individuals who access a PDS must be authenticated by multi-factor authentication and authorized to perform various actions; and (v) all PDSes must ensure accountability by using secure audit mechanisms" (p. 41).

The literature explains that PDSes "help individuals organize and manage their daily lives and give them tools for realizing the inherent value of their own data" (Rubinstein, 2015, p. 54). They "readily allow users to opt-in or out of various services at any time, thereby enabling a right of exit" (p. 53), permitting users to exert control over their own data. Having "a right of exit" helps to ensure that users do not feel trapped or forced into providing their data, and instead allows them to feel free to make their own choices.

Implementing PDSes involves a high level of security, which includes encryption measures to protect users' data. Using encryption techniques, researchers have created systems that can allow for the placement of targeted advertisements with revealing individual information. If implemented, "such technical solutions could enable targeted advertising to continue, along with its associated benefits, such as apps and content that are available at no cost

to the user, while still preventing the collection of data sets used for purposes outside of targeting advertisements" (Book and Bronk, 2016, p. 32).

Martin (2013) offers some potential remedies to make privacy more explicit. She believes that websites should be designed in order to clearly communicate specific goals to users "through icons, cosmetic changes to the Web site, and reminders" (p. 66) so that users are made known of the privacy measures the website is trying to convey. Couldry and Turow (2014) propose that "the goal should be to mobilize stakeholders representing every dimension of the democratic process in a public debate about the implications of big data's deep embedding in a personalized public realm" (p. 1722). While these may be good solutions, there is still a certain level of uncertainty surrounding best practices that would be mutually beneficial to both users and corporations, which may be an area for future research into how this can be effectively and efficiently achieved. The aim is not to increase privacy measures so much, or require such explicit consent as to impede in the daily online behaviours of the user. Rather, it is necessary to achieve "balance" when "protecting individual rights to privacy versus the common need for open disclosure of information" (Book and Bronk, 2016).

Most authors note that users must be made to feel empowered in order for them to become engaged in the processes surrounding the collection of their data. Rubinstein (2015) argues for "consumer empowerment as the heart of a new business model", as it "presupposes that individuals maintain control over the creation and sharing of their personal data" (p. 29). When users become more engaged, they are more likely to feel a sense of ownership over their data. This, then, can result in their playing a more active role, and taking the necessary measures to protect their data.

While there are still issues with transparency, and "requiring companies to show their

algorithms does not mean they will or are required to revise problematic practices, nor does it necessarily lead to greater public understanding, given the levels of expertise required to make sense of the technical operations of data-mining processes", in some cases this can be counteracted by placing an emphasis on "accountability" rather than "transparency" (Kennedy and Moss, 2015, p. 6). If corporations choose to build their reputations by providing users with improved transparency and consent measures, and be perceived by the public as accountable, then this greatly benefits users by allowing them to be more informed. Some authors noted that users felt a sense of powerlessness when it came to the convoluted privacy policies put forth by corporations, and consequently refused to read them (Andrejevic, 2014). If individuals feel like they understand more about how their data is being collected and how it is being used, then maybe this sense of powerlessness will dissipate enough for them to feel empowered enough to read these privacy policies. It is necessary for users to resist simply accepting privacy structures that don't accurately reflect their best interests, and not to accept having to sacrifice their privacy in order to benefit from an online site's services, but rather to "negotia[te] over the privacy norm itself" (Martin, 2013, p. 55).

RQ 3: To what extent can the individuality of users who generate data be preserved in the process of Big Data collection and analysis, and how can users be protected from commodification?

In answering RQ 3, there seemed to be a dichotomy between "personal" data and "too-personal" data. Many authors (Reyman, 2013; Book and Bronk, 2016, Borlin and Schwarz, 2015; Tene and Polonetsky, 2013; Alim, 2014) discussed the way that aggregate data has the ability to commodify users, causing them to lose their identity as individuals and instead placing them within the context of the group. They also believed that Big Data was a commodifying

process as corporations reduce individual people to the data they generate. However, in contrast to this, some authors (Couldry and Turow, 2014; van Dijk and Poell, 2013) cautioned against data that can become too targeted and too personal, predicting people's individual likes, dislikes, locations, and behaviours to target personalized ads to them. In this way, the question of individuality seemed to be twofold: individuality must be preserved so users are not commodified and reduced to the data they generate, yet users' individual identities must be obscured enough so as not to reveal any personally identifiable information.

Agency is integral to the concept of Big Data; while users are responsible for the generation of the data that comprise Big Data sets, corporations assume ownership of this data because they are responsible for its collection and analysis. This is made explicit in the third of Richards and King's (2013) paradoxes, the "Power Paradox", where corporations assume control and appropriate user data freely and undermine the user's sense of ownership. Baack (2015) underscores the importance of agency in relation to Big Data, and notes that Big Data has raised "urgent questions about public agency" (p. 1).

A lack of user agency is inherent in the context of Big Data from the initial data mining process, where "data appears to be a neutral result of user-technology interaction, naturally occurring and available to be appropriated and mined by the technology provider" (Reyman, 2013, p. 525), to the final stages where user data is collected by "data brokers", who "re-sell that data for a variety of purposes including ad targeting" (Book and Bronk, 2016, p. 26). Reyman (2013) describes corporations' "understanding of data as objective facts that preexist rather than products of authorial agency precludes an understanding of data as authored texts to be owned in part by users themselves" (p. 525). The language of objectivity alludes to the fact that users are

stripped of their right to own this data, which is essentially subjective as it is a result of a choice that constitutes the user's behaviour.

Inherent in Big Data is the practice of predictive analytics, where individuals' data is gathered to draw conclusions and make generalizations about populations. Predictive analytics allows corporations to "predict" what goods and services should be offered to which people based on their online behaviours. Tene and Polonetsky (2013) note that the process of aggregating data, recognizing patterns, and making predictions "facilitates the masking of illegitimate or illegal discrimination behind layers upon layers of mirrors and proxies… The machine can find strong correlations, which result in discriminatory outcomes that are based on neutral factors" (p. 359). Many authors indicate that the generalizing nature of predictive analytics can result in the loss of the user's individual identity as the patterns that they generate through online behaviours are aggregated to draw general conclusions, which can then lead to discrimination against populations. Predictive analytics can be discriminatory because they focus on patterns and tend to negate or overlook those outliers that represent members of the population that do not fit the "norm". While this may be helpful for markers and advertisers, it is inherently problematic for populations as it embodies the bias that we as a society should look to avoid. Predictive analytics rely on "correlations" rather than "causation", and as such, "the newly discovered information is not only unintuitive and unpredictable, but also results from a fairly opaque process" (Rubenstein, 2013, p. 8). This opacity is exemplified in the lack of a representative population embodied by Big Data sets, as "aggregated, individual actions cannot, in and of themselves, illustrate the complicated dynamics that produce social interaction— the whole of society is greater than the sum of its parts" (Crawford, Miltner & Gray, 2014, p. 1667). Individuality must be accounted for in order to ensure that all members of a population are

represented.

Predictive analytics also embody certain practices that have implications on user privacy. Tene and Polonetsky (2013) offer Netflix and Amazon as examples, where both sites predict and recommend products based on users' past searches or orders. Another example is "Google's autocomplete and translate functions", which "are based on comprehensive data collection and real time keystroke-by-keystroke analysis" (Tene & Polonetsky, 2013, p. 11). These functions predict future searches based on the user's own search history, and those searches most frequently conducted by users in general. This represents an infringement on the privacy of both the individual user, and of general populations. In predicting audience behaviours, corporations can then sell this data to advertisers who target individual users or population groups based on their own search histories (Tene & Polonetsky, 2013). For example, if one was to search for a particular clothing item on a website, or a particular book, that website could then collect this data and sell it to advertisers, causing advertisements for these or related items to appear to these individual users while they are online. Book and Bronk (2016) compare these measures to "attaching a tracking device to their car, or by simply following them around throughout the day" (p. 16). While it seems unrealistic in this context, the essential practice of appropriating the data that users generate in order to facilitate targeted advertising embodies the same principles.

However intrusive these measures are, predictive analytics are a major source of information for advertisers today:

> The huge availability of content—and the movement of advertising dollars to new vehicles such as search engines and social media sites—means that media buyers can exploit unprecedented competition to reach people (gain 'audience impressions') at far

> lower costs per thousand impressions than with analog media. (Couldry and Turow, 2014, p. 1715)

The movement toward the "deep personalization" of advertising is "not the result of conspiracy to remove people from collective experiences", but rather "an unintended side effect—a negative externality—of how advertising, big data, and content production have come to coexist over the past two decades" (Couldry & Turow, 2014, p. 1712). Data mining also facilitates the use of predictive analytics, and is of particular concern as it is the main practice through which user data is collected. Data mining allows advertisers to better understand users through the collection and analysis of their profile data and web search histories. This allows them to then create more personalized advertisements targeted to individual users, increasing these advertisements' efficacy (van Dijk and Poell, 2013). This, then represents a dichotomy between the practice of aggregating data so the individual identity is lost, and targeting advertising based on a user's personal likes, dislikes, and behaviours as discovered through their data.

How, then, can users mitigate the effects of this challenge and preserve their individuality while becoming active and engaged in the processes of protecting their privacy and resisting commodification? Tene and Polontesky (2013) propose the adoption of "obscurity" (p. 364) as a potential remedy for privacy infringement, as "individuals are far less troubled by data analysis processes that do not *single them out* from a group" (p. 364). They explain that "de-identification", or the adoption of a non-focused approach that considers the data of an unspecified group of users rather than each individual, "can be a mitigating precaution" (p. 364). With de-identification, the idea is that users' information is "hashed" or obscured enough so that any "information that is sufficient to uniquely identify an individual", also called "PII", or

Personally Identifiable Information, is unknowable (Book and Bronk, 2016, p. 8). Similar to Tene and Polonetsky (2013), Martin (2013) suggests that "the desire to be *obscure* or hidden from view remains a driver of privacy expectations and protects information from possible leaks. Rather than focus only on identifiability, individuals online and off–line search for a state of obscurity" (p. 60). In addition to considering ways in which users' individual identities can remain anonymous online, "the key to obscurity is keeping relevant information away from those it was not intended or avoiding information being leaked" (p. 60).

However, some authors point out that "re-identification of data" (Alim, 2014, p. 28) can occur by cross-referencing data sets, "which weakens anonymization as an effective strategy, thereby casting doubt on the fundamental distinction between personal data and non-personal data" (Rubenstein, 2013, p. 13). The message put forth by these authors is that "anonymisation does not guarantee the privacy of data subjects" (Alim, 2014, p. 68). Despite this, and like the concept of notice and choice discussed in RQs 1 and 2, the anonymization of user data has the potential to greatly benefit users and help protect their privacy and their individuality as people by reducing the risk of the commodification of their data. The key is to implement these measures in conjunction with other measures in order to achieve the greatest possible benefits from each, and thus provide users with the greatest level of protection from commodification.

Richards and King (2013) discuss what they call the "Transparency Paradox", where "Big data promises to use this data to make the world more transparent, but its collection is invisible, and its tools and techniques are opaque, shrouded by layers of physical, legal, and technical privacy by design" (p. 42-3). If employed with the concerns of users in mind, Big Data can exhibit a high degree of transparency; it can prove beneficial while still maintaining the

integrity of the user. A lack of transparency contributes to the perpetuation of stereotyping and bias.

Kennedy and Moss (2015) underscore the importance of transparency, and state that "perhaps the most widely discussed way in which the public can have more control over data mining is by making data-mining practices more transparent" (p. 5). The central goal with transparency is not to make processes more difficult for users by providing them with convoluted explanations and more information than necessary, but to make the processes involved in data collection and analysis more apparent, and disclose how the data is being gathered and how it will be used. More explicit consent measures do not necessarily imply longer or more convoluted consent practices, but rather simply making processes more apparent to users.

The concepts of privacy and transparency overlap in many ways; one of the opaquest processes related to data collection is the "Privacy Policies" or "Terms-of-Use" agreements that many websites put forth. In theory, these policies typically explain that user data will be collected and won't be shared, and include a consent request asking the user to allow the website to use their data. However, these processes are convoluted and structured to benefit corporations, and in general, "informed consent is a grey area" (Alim, 2014, p. 63). Informed consent can be applied and understood in many ways, and is essentially up to the interpretation of users and corporations. For example, "empirical studies show individuals neither read nor understand privacy policies, which anyway rely on ambiguous language, and are easily modified by firms. Thus, consent is too often an empty exercise" (Rubenstein, 2013, p. 5). These policies are inherently commodified, and "such terms-of-use statements position user data and data records as information without a human author or productive force, without meaning until it is mined and put to use by the site owner" (Reyman, 2013, p. 525). The current structure of informed consent

makes it difficult for users to play an agentic role, and encourages data appropriation. As such, privacy and consent measures need to be modified and improved in order to allow users to play a more active role in the ownership of their data, and preserve their own individuality.

Data appropriation relates back to privacy, as "a claim to appropriate user data, then, is not based on intellectual property law but, rather, on the assumption of a meaningful distinction between the content and data contributed by users (and on manufactured user consent)" (Reyman, 2013, p. 524). The language that describes the manufacture of consent alludes to its production by those in dominant power roles, not as freely assented to by users. Individuality, then, can be preserved if consent measures are based on those processes that govern "intellectual property law", and acknowledge users as the generators of this data. As such, the title of "generator" implies that users are the data's owners and creators, allowing them to maintain some extent of their individuality through their role as owners. This can result in the ability to potentially resist commodification, as appropriation becomes more difficult when dealing with an active and engaged public.

Some authors refer to the concept of user commodification as "datafication", which "endowed social media platforms with the potential to develop techniques for predictive and real-time analytics" (van Dijk and Poell , 2013, p. 9). This allows data to become real, and it is no longer two-dimensional concept but a three-dimensional entity (van Dijk and Poell, 2013). The instantaneous nature of datafication creates "datafied publics" (Baack, 2015, p. 6), where social media platforms and websites use their ability to collect and aggregate this real-time data to their advantage (van Dijk and Poell, 2013).

The process of datafication is both explicitly commodifying, reducing users to the data they generate, and implicitly commodifying, embodying these processes in its structure. van Dijk

and Poell (2013) note that "an important aspect of datafication is the *invisibility* or naturalness of its mechanics: methods for aggregation or personalization are often proprietary and thus often inaccessible to public or private scrutiny" (p. 10). This excludes the user from the process, and contributes to a further lack of agency and understanding of these mechanisms.

Borlin and Schwarz (2015) state that "the nature of the mining in question has gradually shifted to highlight abstract bundles of behavioural patterns (correlations and sociograms), downplaying the referential subject" (p. 10). They express their desire for "future research to heed this shift to de-subjectification, and to even employ models of what could be called 'post-referential' agency" (p. 10). They believe the abilities of media users to be self-reflexive, specifically, their "reflective abilities to understand themselves" (p. 10). This is echoed by Baack's (2015) idea of "recursive publics" (p. 8).

Instead of simply accepting this structural bias, users and policymakers alike "should seek more fair and ethical practices that make data collection transparent and that openly recognize the value of users' data contributions to the cocreation of digital culture" (Reyman, 2013, p. 528). There are also certain degrees of opacity, especially when regarding personal or sensitive data, that should not only be applied but are recommended (Richards & King, 2013); the key is to keep the preservation of the user's integrity in mind when regulating transparency and opacity. While increasing transparency is not the definitive answer for removing bias from these structures, it is an effective way to allow the user some agency and control over their own data through keeping them informed as to how it will be used, and giving them the option of whether or not they want to contribute their data without having to sacrifice their online presence.

Discussion

In my review of the existing scholarly literature, I discovered that while issues of privacy were largely correlated with the study of Big Data in the social sciences, user commodification and user agency were not as highly represented. This could be because Big Data is such a new field of study, and has only been discussed in a social sciences context for a few years. When I searched databases for Big Data and the term "agency", it returned minimal (if any) search results; this may be because this term has not yet been widely accepted or acknowledged as a potential theme in the study of Big Data. This may change in a few years as the study of this field begins to grow, but I generally had more results when I searched for "user roles" or "user rights".

In answering RQ 1, which asks how users can play a more active role in data ownership when it comes to data appropriation, increasing transparency measures was a common theme in the scholarly literature. This would facilitate users' understanding of the processes surrounding data collection and analysis, which would consequently allow them to be more aware of these processes. Knowing that they are generating data can make users feel more involved; this might create a sense of ownership in users, causing them to seek more agentic roles in the collection and analysis of their data. Baack (2015) suggests the adoption of open-source culture as a way to make data publicly accessible, thus helping to eliminate, at least in part, the hierarchical boundaries that separate those who generate the data from those who collect and analyze it— or, essentially, users and corporations.

In answering RQ 2, it became evident while reading the scholarly literature that it is not just a question about how users can increase their agency when it comes to understanding and claiming ownership of their data, but how corporations can change their policies to better reflect

users' interests. Because much of the data appropriated by corporations is sold to other corporations for marketing purposes, corporations seem to have little incentive to get users to protect the data that is so vital to their profitability. If the development of brand name and reputation (Martin, 2013) is emphasized, though, then this will give corporations more incentive to improve their policies to better reflect users' interests. If, for example, a publicly available review system is applied to rate those corporations that best protect users' privacy, and those that commit the most egregious infractions and neglect to do so, then users will be placed in an agentic position as they attribute the rating to the corporation. Consequently, corporations may be more likely to construct their policies to empower users and reflect their best interests, as they wish to remain profitable by upholding their accountability to the public and thus establishing a good corporate reputation.

Rubinstein (2015) also suggests that methods like selective disclosure, or the choice to opt-in or opt-out of providing a website with personal information, as well as the employment of PDSes, or high-security, heavily encrypted services that store personal data, can help to establish more explicit consent measures without necessarily implying more barriers to providing data that risk isolating and confusing users. Alim (2014) states that the main issue here is not to require such explicit consent as to make using websites difficult, but to achieve a level of balance between making consent explicit, enhancing privacy measures and facilitating usability while protecting users' rights. These options allow users to claim a more active role in data ownership by allowing them to choose whether or not they wish to provide their data, and help to protect user data in an easily understood and universally accessible and applicable way.

In answering RQ 3, there was an obvious tension between protecting a user's individuality online and ensuring that individual experience was not subsumed by the whole of

aggregate patterns, while still obscuring the individual enough so that their personal identity and information is protected. The majority of the scholarly literature did not seem to focus on how users could protect their individuality, but rather emphasized the various ways in which the collection and aggregation of Big Data sets commodified users by impinging on their rights and disregarding their individuality.

I have discussed the ways that the processes of aggregation and analysis that are inherent in Big Data are essentially commodified, and identified what is being said about transparency, privacy, and agency in the existing scholarly literature. But how can these issues be remedied to facilitate data collection while still preserving the integrity and privacy of the individual user? Martin (2013) suggests that transaction economics, while they do have the potential to commodify users, may also have the ability to preserve users' integrity, as "transactions, when aligned with a governance structure and working correctly, provide order, relieve conflict, and support mutually beneficial solutions" (p. 11). In order for users to play a more active role in data ownership and for their rights to be protected, governmental structures need to decipher what is happening right now, and what corporations and data analysts, as well as users, want out of the situation. They must then figure out a way to mitigate users' concerns so that the information can be analyzed and applied in a mutually beneficial way.

While the processes embodied by Big Data tend to be commodifying toward users, Big Data is beneficial to society in many ways and it is unrealistic to assume or demand that all measures of advertising be stopped. Once this realization is made, it then becomes a priority to figure out how to experience the benefits of Big Data while preventing users from commodification.

Conclusion

Overall, while there have been many insightful articles that discuss the implications of Big Data on transparency, privacy, and user rights and roles, there is still room for further discussion regarding the effects of large-scale data analysis on users.

With regards to methodology, 16 (80%) of the total 20 articles studied in the Findings (Borlin and Schwarz, 2015; Crawford, Miltner and Gray, 2014; Crawford and Schultz, 2014; Couldry and Turow, 2014; Custers, 2016; Dalton and Thatcher, 2015; Kennedy and Moss, 2015; Martin, 2013; McFarland and McFarland, 2015; Reyman, 2013; Richards and King, 2013; Rubinstein, 2015; Tene and Polonetsky, 2013; Tene and Polonetsky, 2013; Tufecki, 2014; van Dijk and Poell, 2013) conducted purely theoretical discussions. While Book and Bronk (2015) did not perform a study, they conducted their own original research and included the results in their discussion. Only 3 articles detailed information from an original study conducted by the authors (Alim, 2014; Andrejevic, 2014; Baack, 2015). Theoretical discussions of user commodification, privacy, and agency are very important to gaining a better critical understanding of Big Data in the social sciences. However, it is necessary for future studies to determine the direct impacts these issues have on users through reception methods that deliver first-hand accounts from users, like surveys and interviews. This is essential to understanding the rights and roles of users in the context of Big Data, how users are affected by the appropriation of their data, and what they feel would help increase their agency.

Dalton and Thatcher (2015) encourage users and scholars alike "to develop critical (and self-critical) perspectives and approaches to spatial Big Data and subsequent technologies" (p. 11). It is necessary to "reflexively analyz[e]" (p. 11) the processes of Big Data, as "establishing the historical, social context of a technology is a key step in demystifying and denaturalizing it"

(p. 11). A better understanding of these practices will "allow us to learn from those earlier processes to better critically evaluate current technologies and knowledge" (p. 11).

Kennedy and Moss (2015) suggest that one possible way to bridge the gap "between social media data mining and public life is the shift from known to knowing publics (p. 9). Known publics "are subject to the data-mining practices of others", whereas knowing publics "are positioned in relation to data as more active and reflexive agents" (p. 9). In this way, "it may be possible for data mining not only to be used by elites to produce known publics, but rather for the public to be more knowing of itself and to participate in the active production of itself, the public" (p. 9).

Another possibility is "to consider whether data-mining practices can be used by publics to constitute themselves as more active and reflexive agents" (Kennedy and Moss, 2015, p. 9). In this way, users are able to use the very processes that originally commodified them to gain insight into how to become more "active" in their roles as data generators. They may then be able to reflect on these roles, and find new ways to gain agency and be included in the processes of data collection and analysis.

Borlin and Schwarz (2015) suggest that "the nature of the mining in question has gradually shifted to highlight abstract bundles of behavioural patterns (correlations and sociograms), downplaying the referential subject" (p. 10), and "ask future research to heed this shift to de-subjectification, and to even employ models of what could be called 'post-referential' agency" (p. 10). Borlin and Schwarz (2015) also ponder "whether their respective heuristics will adjust to the technological possibilities and the de-subjectified logics of the algorithm, or align with principles from earlier modes of surveillance (e.g. re-subjectification)" (p. 10). Borlin and Schwarz expand on Kennedy and Moss' (2015) idea of self-reflexivity, and encourage users to

reflect and seek to improve on the representation of their rights and roles, always be looking for ways to increase their agency, and ensure that they play an active role in data ownership, protecting their privacy, and resisting commodification.

Big Data is a growing phenomenon that will continue to be studied in both technological and social science contexts. There are many insightful discussions about how Big Data affects privacy and user rights and roles; from analyzing the transparency of privacy policies and terms of use statements, to debating the topic of data ownership, to determining in what areas users must be afforded more agency, scholars in the social sciences are increasingly turning to Big Data as a subject for analysis. There is still much to discover, though, like the outline of a tangible plan to allow users to claim an active role in the ownership of their data and reduce the risk of impinging on their privacy, while still benefitting from Big Data's contributions to social life. The central goal of this research paper is not to present an apocalyptic view of the future of Big Data, where people become numbers and privacy is all but a myth; data analysis benefits society in many important ways, like generating medical records and determining what the trending topics for future study are in emerging fields. This research paper attempts to discover what is being said about privacy, commodification, and user rights and roles in the context of Big Data, and identify what potential ethical issues are discussed in the scholarly literature and how these concerns can be addressed. Big Data has redefined the social world and expanded the limits of knowledge and information, and it is necessary to draw attention to these issues and identify places where they can be remedied so as to achieve the greatest benefit from this valuable phenomenon.

Appendix

Alim, S. (2014). An initial exploration of ethical research practices regarding automated data extraction from online social media user profiles. *First Monday, 19(7).* Available at: http://firstmonday.org/ojs/index.php/fm/article/view/5382/4099

Andrejevic, M. (2014). The Big Data Divide. *International Journal of Communication*. 1673-1689. DOI: 1932–8036/20140005

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and empowerment. *Big Data and Society*, 1-11. DOI: 10.1177/2053951715594634

Book, T. & Bronk, C. (2016). I see you, you see me: Mobile advertisements and privacy. *First Monday*, *21(3).* Available at: http://firstmonday.org/ojs/index.php/fm/article/view/6154/5215

Borlin, Goran and Schwarz, Jonas Andersson (2015). "Heuristics of the algorithm: Big Data, user interpretation and institutional translation". *Big Data and Society, 1(12).* DOI: 10.1177/2053951715608406

Crawford, K., Miltner, K. & Gray, M.L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication*. 1663-72. Available at: http://ijoc.org/index.php/ijoc/article/viewFile/2167/1164

Crawford, K. & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review, 55(1),* 93-128.

Available at:
http://search.proquest.com.ezproxy.lib.ryerson.ca/docview/1664533162/fulltextPDF/61C29BDF590148B8PQ/1?accountid=13631

Couldry, N. & Turow, J. (2014). Advertising, Big Data, and the Clearance of the Public Realm: Marketers' New Approaches to the Content Subsidy. *International Journal of Communication*.
DOI: 1710–1726 1932–8036/20140005

Custers, B. (2016). Click here to consent forever: Expiry dates for informed consent. *Big Data and Society,* 1-6.
DOI: 10.1177/2053951715624935

Dalton, C. M. & Thatcher, J. (2015). "Inflated granularity: Spatial ''Big Data'' and geodemographics". *Big Data and Society*, 1-15.
DOI: 10.1177/2053951715601144

Kennedy, H., & Moss, G. (2015). Known or knowing publics? Social media data mining and the question of public agency. *Big Data and Society*, 1-11.
DOI: 10.1177/2053951715611145

Martin, K. (2013). Transaction costs, privacy, and trust: The laudable goals and ultimate failure of notice and choice to respect privacy online. *First Monday, 18(12).*
Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4838/3802

McFarland, Daniel A and McFarland, H. Richard (2015). "Big Data and the danger of being precisely inaccurate" *Big Data and Society, 1(4).*
DOI: 10.1177/2053951715602495

Reyman, J. (2013). User data on the social web. *National Council of Teachers of English,*

513-33.

Available at: http://www.ncte.org/library/NCTEFiles/Resources/Journals/CE/0755-may2013/CE0755User.pdf

Richards, N. M. & King, J.H. (2013) "Three Paradoxes of Big Data". *Stanford Literature Review, 66(41).*

Available at: http://poseidon01.ssrn.com/delivery.php?ID=549006098115006005113083127123105064018071056080004037007125080119094113077084116004037020023014049096033103028093120006026101020055059047019083088110088090013015030036038009026002098085100031017026088117122077109090112025106000111123066088093095112&EXT=pdf

Rubenstein, I. S. (2013). Big Data: The End of Privacy or a New Beginning?. *International Data Privacy Law, 3(2),* 74-87.

DOI: 10.1093/idpl/ips036

Tene, O. & Polonetsky, J. (2013). Judged by the Tin Man: Individual Rights in the Age of Big Data. *The 5th Annual Privacy Symposium: The Technology of Privacy*. 351-368.

Available at: http://heinonline.org.ezproxy.lib.ryerson.ca/HOL/Page?public=false&handle=hein.journals/jtelhtel11&page=351&collection=journals

Tene, O. & Polonetsky, J. (2013) Privacy and Big Data: Making Ends Meet. *Stanford Law Review, 66(25).*

Available at: http://www.stanfordlawreview.org/online/privacy-and-big-data/privacy-and-big-data

Tufekci, Z. (2014). Engineering the public: big data, surveillance and computational politics. *First Monday, 19(7).*

Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4901/4097

van Dijk, J. & Poell, T. (2013). Understanding Social Media Logic. *Media and Communication, 1(1)*, 2-14.

DOI: 10.12924/mac2013.01010002

References

Aradau, C., & Blanke, T. (2015). "The (Big) Data-security assemblage: Knowledge and critique". *Big Data and Society,* 1-12.
DOI: 10.1177/2053951715609066

Borlin, G., & Schwarz, J. A. (2015). "Heuristics of the algorithm: Big Data, user interpretation and institutional translation". *Big Data and Society,* 1-12.
DOI: 10.1177/2053951715608406

Couldry, N., & Powell, A. (2014). Big Data from the bottom up. *Big Data and Society,* 1-5.
DOI: 10.1177/2053951714539277

Couldry, N., & van Dijk, J. (2015). Researching Social Media as if the Social Mattered. *Social Media and Society,* 1-7.
DOI: 10.1177/2056305115604174

Custers, B. (2016). Click here to consent forever: Expiry dates for informed consent. *Big Data and Society,* 1-6.
DOI: 10.1177/2053951715624935

Custers, B., & Ursic, H. (2016). Big Data and data reuse: a taxonomy of data reuse for balancing Big Data benefits and personal data protection. *International Data Privacy Law, 6(1),* 4-15.
Available at:
http://idpl.oxfordjournals.org.ezproxy.lib.ryerson.ca/content/6/1/4.full.pdf+html

Dalton, C. M. & Thatcher, J. (2015). "Inflated granularity: Spatial ''Big Data'' and geodemographics". *Big Data and Society,* 1-15.
DOI: 10.1177/2053951715601144

danah boyd & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15(5)*, 662-79.
DOI: 10.1080/1369118X.2012.678878

Kennedy, H., & Moss, G. (2015). Known or knowing publics? Social media data mining and the question of public agency. *Big Data and Society*, 1-11.
DOI: 10.1177/2053951715611145

Kennedy, H., Poell, T., & van Dijk, J. (2015). Data and agency. *Big Data and Society*, 1-7.
DOI: 10.1177/2053951715621569

Lewis, K. (2015). "Three fallacies of digital footprints". *Big Data and Society*, 1-4.
DOI: 10.1177/2053951715602496

Peacock, S. E. (2014). How web tracking changes user agency in the age of Big Data: The used user. *Big Data and Society*, 1-11.
DOI: 10.1177/2053951714564228

Pybus, J., Cote, M., & Blanke, T. (2015). Hacking the social life of Big Data. *Big Data and Society*, 1-10.
DOI: 10.1177/2053951715616649

Reyman, J. (2013). User data on the social web. *National Council of Teachers of English*, 513-33.
http://www.ncte.org/library/NCTEFiles/Resources/Journals/CE/0755-may2013/CE0755User.pdf

Schermer, B. W., Custers, B., & van der Hof, S. (2014). The crisis of consent: how stronger legal protection may lead to weaker consent in data protection. *Ethics Information Technology*, 171-182.

http://journals2.scholarsportal.info.ezproxy.lib.ryerson.ca/pdf/13881957/v16i0002/171_tcochstwtcoc.xml

Tene, O. & Polonetsky, J. (2013). Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property, 11(5)*. 239-273. http://heinonline.org.ezproxy.lib.ryerson.ca/HOL/Page?public=false&handle=hein.journals/nwteintp11&page=239&collection=journals

van Dijk, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society, 12(2)*, 197-208. http://search.proquest.com.ezproxy.lib.ryerson.ca/docview/1547988865/fulltextPDF/D60EF4080F4C4148PQ/1?accountid=13631

Zeller, F. (2014). Big Data in Audience Research: A Critical Perspective. *Revitalizing Audience Research: Innovations in European Audience Research*. F. Zeller, C. Ponte, and B. O'Neill, eds. New York, NY: Routledge, 261-78.