The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts

Shawn Walker

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Emma Spiro, Chair

W. Lance Bennett

Nicholas Weber

Program Authorized to Offer Degree:

Information School

University of Washington

Abstract

The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts

Shawn Walker

Chair of the Supervisory Committee:
Assistant Professor, Emma Spiro
Information School

Just as social media has permeated communication in our public and private lives, it has also become a widely used source of data and object of study in academic and commercial research. Despite widespread use, relatively little is known about how social media datasets change when observed at different points over time or how collection methods may impact the data at the core of our research projects. For example: Will results differ if social media data are collected in real-time, a few minutes after production, hours, days, or weeks later? What happens to the metadata, links to web pages, photos, and videos embedded in this content over time? If data collection methods do not preserve and archive social media posts, metadata, and linked content; are researchers venturing into a new dataset each time they engage with it? In this dissertation, a combination of quantitative and qualitative approaches are used to examine how social media datasets change over time and how change impacts the

reliability and authenticity of this data. Three Twitter-based case studies, each exhibiting prototypical elements social scientists encounter in their research are used to demonstrate the impact of research design and data collection choices. This work advances the field of information science by empirically investigating how the ephemeral nature of social media data, metadata, and linked content have significant and lasting effects on the reliability and authenticity of datasets used in research. By situating research design decisions of how and when to observe data within the frameworks of process theory, infrastructure studies, and archival theory, this work brings the importance of methodological considerations to the forefront of studies of digital and social media. Empirical observations inform a set of implications for social media research, offering researchers practical considerations to inform their research designs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Like all dissertations, this process has been a long journey for me with a lot of support from friends, family, and colleagues along the way. For everyone who supported me: THANK YOU!, I made it! I would like to express my gratitude to:

- To my advisor, Dr. Emma Spiro, I am grateful for your deep understanding, unwavering support without judgement, and tirelessly helping me defend and cross this finishing line.
- To Dr. Nicholas Weber for stepping in to support and help me with the final edits this summer.
- To Dr. W. Lance Bennett for always helping me with the big picture, providing support, and pushing my work out of my comfort zone.
- To Dr. Karine Nahon for being my first advisor, shaping so much of my work and who I am as a scholar, and providing support from afar.
- To Dr. Robert Mason for showing me how to put students first, how to write my first grant, helping a group of grad students start a lab around a crazy idea, and always championing our work.
- To my SoMe Lab partners in crime – Joe Eckert and Jeff Hemsley – where a large portion of this dissertation began to develop.

- To my friends in colleagues in the sunshine cohort – you provided such an inspiring and supportive environment, I feel so lucky to have gone through this process with all of you. Amazingly, 8 of us started the program and 8 of us competed the program!
- To my Sunday library writing group and partners in crime.
- My PhD buddies – Liz Mills, Norah Abokhodair, and Jordan Eschler.
- The support from the Social Media Collective at MSRNE and the Oxford Internet Institute's Summer Doctoral Program. The support system and lifelong connections I formed have already taken me far.
- To Dr. Sheetal Agarwal for being a source of support, inspiration, patience, understanding, and an all-around amazing person. Watching you finish your PhD such an amazing gift. Thanks for being my source of support.
- To Dr. Kristen Shinohara and Dr. John Mario for all of the blood, sweat, and tears. Thanks for going through the job process with me and watching my job talk 150 times to the point where you could present it better than I!
- To my adopted family who went on this journey with me – Elyse and Kevin Lewis, Jean Donohue, and Fred Johnson – I don't know what I'd do without you all.
- To Linda Dolive for making space for me and "adopting" me into your family, being a mom to me, and being such a champion.
- To Darwin for all of the support, wags, and unconditional love. As promised, you stuck with me through the PhD. I'll miss you on the next stage of this journey buddy.

DEDICATION

To my mom, even though she didn't get to see this, I know she's proud.

CHAPTER 1: INTRODUCTION

> For the first time, we can follow [the] imaginations, opinions, ideas, and feelings of hundreds of millions of people. We can see the images and the videos they create and comment on, monitor the conversations they are engaged in, read their blog posts and tweets, navigate their maps, listen to their track lists, and follow their trajectories in physical space (Manovich, 2013, p. 461).

"Big" and "social" data bring a substantial increase in the scale and types of data that academic researchers and practitioners can access. This digital trace data, in the form of social media posts from Facebook or Twitter for example, allows automated observation and collection of the online activities of millions of users by simply writing a short program to collect data (Freelon, 2014). Researchers use social media data is to make claims of study human activity (Zimmer & Proferes, 2014). Businesses and governments (Parmelee & Bichard, 2013) use social media data to make decisions about which potholes to repair or customers to serve. In this new research environment, how does one conduct empirical research with methodological rigor? In this dissertation I focus on one aspect underlying this question: how social media data sets change over time.

1.1 Methodological Issues Surrounding Social Media Data Collection

Motivated by direct prior experience with the challenges of collecting and analyzing social media data, videos, and web links in, this dissertation directly addresses this question. The first project examined YouTube videos and blogs during the 2008 US Presidential Election to understand the role these platforms play in propagating viral political videos (Nahon, Hemsley, Walker, & Hussain, 2011). The second project focused

on the use of social media by the Occupy Wall Street movement, where 31,000 seed URLs embedded in tweets related to the Occupy Wall Street movement were coded based on the type of resource (e.g. mainstream media site, celebrity site, government site, etc.) each referred to to (Agarwal, Bennett, Johnson, & Walker, 2014; Bennett, Segerberg, & Walker, 2014). The third study used tweets and links embedded in tweets to examine rumor propagation after the Boston Marathon Bombings (Starbird, Maddock, Orand, Achterman, & Mason, 2014). During each of these projects, the research team encountered an ephemeral and unstable social media dataset leading to a host of non-trivial methodological issues that needed to be addressed in order to meet projects aims and answer core scientific questions.

While social media data allows researchers to "study social and cultural processes and dynamics in new ways" on an unprecedented scale (Manovich, 2013, p. 461), our tendency as researchers making use of such data is often to focus on the phenomena under examination; less attention is paid to understanding the dynamic nature of these data themselves. Social media datasets present a number of challenges for researchers including, but not limited to: (1) the need for new and/or adaptation of existing methods for data collection and analysis, (2) issues of representation and sampling (Boyd & Crawford, 2012; Liang & Fu, 2015), and (3) ethical implications and risks to those whose social media data is being collected and analyzed (Light & McGrath, 2010; Zimmer, 2010; Zimmer & Proferes, 2014). Underlying these important issues are fundamental questions about the nature of social media data itself. Relatively little is known about how social media datasets change when observed at different points over

time or how choices of collection method may impact the data at the core of our research projects, and subsequent research findings. For example: Will results measuring the prevalence of rumors over time differ if social media data are collected as it is produced in real-time, a few minutes after production, hours, days, or weeks later? What happens to the metadata — links to web pages, photos, and videos — embedded in and documenting this content over time? If data collection methods do not preserve and archive social media posts, metadata, and linked content; are researchers venturing into a different dataset each time they engage with it? The findings in this dissertation show that latency, a delay in the collection of data, changes the resulting social media dataset, its metadata, and linked data. Social media posts are deleted or become inaccessible, users change their profiles, and embedded link change.

To illustrate, consider the case of a researcher using the #YesAllWomen campaign to study misogyny online. The #YesAllWomen hashtag and social media campaign was used to share stories of misogyny and violence against women following the 2014 Isla Vista killings (Valenti, 2014). Elliot Rodger, a twenty-two-year-old man, went on a shooting spree on Isla Vista, near the University of California Santa Barbara, killing six people before committing suicide. In the weeks leading up to the killings, Rodger posted a series of YouTube videos and a 137 page autobiographical "manifesto," declaring his hatred of all women for the rejection and disdain he claims they dealt him throughout his life. Responses to the campaign, ranged from support and personal stories to hateful and sexist comments. Since hate speech is against the abusive behavior policies of social

media sites like Twitter,[1] Facebook,[2] and Instagram, many of these posts were deleted and accounts suspended by the platforms.

This example demonstrates a number of challenges for researchers:

- What search terms should the researcher use to collect content related to #YesAllWomen? While the hashtag #YesAllWomen may seem like an obvious choice, would a single hashtag will not contain all of the content relevant to the research project. Some users may have tweeted content related to #YesAllWomen without a hashtag, used a related hashtag, or posted content on other social media platforms outside of Twitter. The researcher must decide how to bound the project using a set of query terms and social media platforms based upon their researcher questions and understanding of the context of the campaign. This would include the observation what keywords users include in their posts and what platforms they are utilizing.
- Will all content collected with the query terms be relevant to the research at hand? Or will some posts contain the query terms, but be irrelevant to the project?
- What processes should be used to filter out irrelevant posts and how should these processes be documented?

Since the campaign was not planned in advance, data collection cannot be setup prospectively to capture the campaign from its inception. As a result, at least a portion, if

[1] https://support.twitter.com/articles/18311
[2] https://www.facebook.com/help/216782648341460

not all, of the social media data related to the campaign would need to be collected retrospectively. The delay in data collection introduces additional questions:

- During the delay from the time a researcher started collecting and when tweets and images were posted, content, accounts, images, and links may change or be deleted.
- How will this affect the dataset and the findings from the dataset?
- Will retrospectively collected data represent the actual discussions and posts within the #YesAllWomen campaign or will missing posts provide a false account of the campaign?

Despite these challenges, there has been an explosion of research using social media data to study human behavior in almost every domain of social science. Examples range from death and memorialization (Acker & Brubaker, 2014), social movements (Agarwal et al., 2014; Bastos, Mercea, & Charpentier, 2015; Bennett et al., 2014), disasters (Starbird & Palen, 2010; 2012), epidemiology (Malik, Gumel, Thompson, Strome, & Mahmud, 2011), to many others. This body of literature is growing at a staggering rate (Williams, Terras, & Warwick, 2013; Zimmer & Proferes, 2014), but accompanying methodological contributions describing and examining the process of conducting research with social media data (SMD) is very thin. Existing methodological literature is typically tool or technologically driven, not a result of empirical examination of the data collection process (Felt, 2016; Miller, Ginnis, Stobart, Krasodomski-Jones, & Clemence, 2015), leaving researchers without an understanding of how to approach or evaluate the social media data collection process. As a result, researchers, practitioners, and students

are left to continually re-invent the wheel, learning through a process of trial and error (Brooks, 2015).

Complementary to the social media methodology literature, a body of literature coined "critical data studies" by Dalton and Thatcher (2014) has emerged. Critical data studies focuses on concerns related to the use, analysis, and ethics of big and social media data (Boyd & Crawford, 2012). boyd and Crawford (2012) provide the most cited critique of big data with six provocations ranging from questions about the (lack of) contextualization of big data to how big data changes our[3] definitions of knowledge. Often, however, such literature provides cautions and criticism without tangible solutions to issues raised.

1.2 The Process of Collecting Social Media Data

The process of collecting social media data, while seemingly simple on the surface, requires numerous competencies (Brooks, 2015; Driscoll & Walker, 2014; Felt, 2016), both technical and research design related. The process is made more complex as it involves a mixture of theory, data, and computational processes (see Goble 2008 for a bioinformatics perspective) filled with many "black-boxes" (Driscoll & Walker, 2014; Goble et al., 2008, p. 510). An algorithmic system underlies the multitude of interfaces users and programs use to consume and interact with information from social media platforms. These algorithmic systems (Ananny & Crawford, 2017) are an assemblage of "institutionally situated code, practices, and norms with the power to create, sustain, and

[3] In this document when referencing 'our' or 'we' I am referring to the community of researchers using social media data.

signify relationships among people and data through minimally observable, semiautonomous action" (Ananny, 2015, p. 93). To users and researchers outside of the platform, these algorithmic systems and databases seem like black boxes taking input from a user's action and outputting posts without giving any details of how data is processed or changed. The lack of transparency only adds complexity to the research process since the impact of forces assembling and acting on data are unknown to us.

The process often begins with the research design of a project, linking a set of research questions with the appropriate social media data. After matching a research design and data source, a data collection plan must be developed and executed. It is important to use data collection methods that preserve elements of each social media post and their accompanying metadata that are essential to answering the researchers' research question. Developing and executing this plan requires a deep understanding of the phenomena and social media site(s) under examination: that is to say the recipes used to "cook" (the collecting, cleaning, processing, and analysis of the data) (Bowker, 2013), and the technical skills to carry out the "cooking". The amount of data collected ranges from a few hundred posts to larger datasets consisting of thousands or millions of information artifacts. Methods of collecting social media data range from manually copying and pasting content from social media web sites to large-scale automated data collection via complex scripts. Large or small, manual or programmatic, the processes of research design and social media data collection require a set of empirically informed principles to guide researchers through many choices that must be made throughout the

data collection process. Literature focusing on this process is lacking and this dissertation contributes to this area.

Social media datasets continue to present challenges for researchers after data collection. These datasets also push at the boundaries of traditional research methods (Hargittai & Sandvig, 2015; Karpf, 2012). Oftentimes researchers attempt to apply existing, more traditional methods in this space, but this approach may be problematic if researchers do so blindly without first adapting methods to the unique properties of social media datasets and platforms. For example, consider the application of stratified random sampling techniques traditionally used in survey techniques (de Leeuw, Hox, & Dillman, 2012; Lynch, 2008); how can this method be applied to social media data without a proper sampling frame or observed characteristics on which to determine which strata each account or individual falls? How do we account for representation (Miller et al., 2015), political power (Nahon, 2015), algorithmic and platform bias (Gillespie, 2010), presentation of self (Goffman, 1990), and the context within which these posts are generated (Seaver, 2015)? These are important questions to address in order to assess the validity of research using social media data. A precursor to this

In the rush to collect data, the implications of observing dynamic content at a particular (arbitrary) point in time and issues of preservation of social media posts and their accompanying linked metadata aren't generally considered and rarely discussed in research publications. At its core, social media data is ephemeral – a term often used but rarely defined in research (see Bernstein, Monroy-Hernández, Harry, & André, 2011 for an example of use without a definition). When researchers use the term ephemeral in

the context of social media data it is often shorthand for instability; data is constantly changing, being updated, or deleted. As a result, it is difficult for two researchers to collect the same exact dataset in real-time and practically impossible for them to collect the same dataset retrospectively via a purchase of data from a reseller or by scraping (Burgess & Bruns, 2014). Further, researchers are often forbidden from sharing full datasets by the Terms of Service (ToS) of many platforms. While some platforms, such as Twitter, allow for sharing of each post's unique identification number; this still requires researchers to "rehydrate" or go back to the platform to recollect the most current post content and metadata, if available. The difficulty of collecting and/or sharing datasets makes it impossible to validate or replicate studies using social media data (Felt, 2016).

1.3 Gaps in the Social Media Literature

Few studies have focused on the dynamic nature of social media data itself; those that have primarily looked at specific tools or software interfaces for data collection (Driscoll & Walker, 2014; Felt, 2016; e.g. Gaffney & Puschmann, 2014; González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014), and have not considered the impact of the ephemeral nature of social media data on the data collection process and resulting dataset. In this dissertation, I examine the impact of the ephemeral nature of social media data on research datasets — how posts, their accompanying metadata documenting the post, and linked content such as videos, images, and web pages change over time. Our research community needs ways of acknowledging, understanding, stabilizing, and combating/addressing the ephemerality of social media datasets.

Without ways to measure the impact ephemerality has on datasets, researchers are unable to determine the subsequent impact on research designs and findings. If ephemerality does have an impact, how can researchers quantify and counteract or at least address it or understand the limitations of results? The first step in addressing these questions is to gather empirical data to measure the level of ephemerality over time within multiple social media based case studies at the post, metadata, and linked content levels.

1.4 Research Questions

In this dissertation I use the lenses of process theory (Crowston, 2000), infrastructure studies, and archival theory surrounding electronic records (Duranti, 1995; 1997) to examine how the ephemeral nature of social media impacts collected data, situating social media data collection within the social science research process. Quantitative approaches are applied to examine how social media datasets change over time through an examination of social media posts surrounding the three Twitter-based case studies. The first case, Occupy Wall Street movement, was a world-wide social movement observed over a 3-year timeframe. The second case, Departments of Transportation on the West Coast (Washington, Oregon, and California), represents everyday political interactions with official government accounts. The third case, the reality TV show RuPaul's Drag Race, represents a entertainment context with a high level of image and video content due to the show's visual nature. Each case study was chosen because it represents prototypical features of the types of data collection scenarios researchers

experience when collecting social media data. Examples of these dimensions include time-scale (short to long), population bounding (tight to lose), level of political contention (highly contentious to the everyday political context), and inclusion of links to media such as images and videos (high to low level of media and linking). The aim of this research is to contribute an empirically informed framework for the study of social media data. As such the research questions to be addressed in this dissertation are as follows:

RQ: How does the ephemeral nature of social media data affect social media data (SMD) sets?

- RQA: How does the ephemerality of SMD interact with the process of data collection to impact the reliability of social media data sets?
- RQB: How does the ephemerality of SMD interact with the process of data collection to impact the authenticity of social media data sets?

1.5 Chapter Summaries

The rest of this dissertation is organized as follows:

Chapter 2 provides a description of social media as a source of research data — the social media platforms and social media data collection, and the social science research process through the theoretical lenses of process theory, platform studies, and infrastructure studies. The lens of process theory situates the collection of social media data as subprocess within the process of conducting social science research. Through the lens of process theory, research and data collection are seen as a series of linked steps,

allowing for the testing and comparison of alternative choices each step — for example time between an event and collection of social media data collection or the choice collection method. This process is impacted by the affordances, or features, of each social media platform and the data collection infrastructure these platforms provide.

Chapter 3 introduces a framework to discuss and approach the design of systems to collect social media data. The framework, Social Media as Record, brings relevant concepts from archival and electronic records theory to social media and its preservation. When viewing social media posts as records within a collection, posts are connected via an archival bond and not just seen as individual posts. Posts and embedded content are also bound by time, taking into account not just the text of the social media post itself, but the metadata documenting each post and the linked content within each post as well as the related social media posts and the context within which the posts were created.

Chapter 4 introduces the overarching concept of *ephemerality* drawn from media studies, archival theory and practice, web archiving, and data curation. Three Twitter-based case studies and the methods of data collection, each exhibiting prototypical elements social scientists encounter in their research, that will be used to quantify the impact of ephemerality of social media data are described in detail. The three cases studies are: 1) the Occupy Wall Street movement, 2) Departments of Transportation on the West Coast of the US, and 3) the reality TV show RuPaul's Drag Race. Tweets, metadata, and archives of web links embedded in tweets were collected for each cast study (see Appendix B) in real-time for a period of two weeks. Descriptive statistics for each case study are also provided in this chapter.

Chapter 5 develops the concept of a *reliable* social media dataset. Drawn from the concept of statistical repeated measures, a reliable social media dataset is one in which the corpus of social media data collected by a researcher is impervious to change — collecting a dataset with the same parameters at different points in time should yield the same dataset. I operationalize the concept of reliability as the number of tweets still accessible at any point in time compared with tweets collected in real-time. Tweets that were inaccessible at the end of the observation period were examined to determine the cause of inaccessibility. Approximately 7 - 12% of tweets were no longer accessible at the end of the observation period with over 90% of tweets inaccessible due to either the deletion of the tweet itself or deletion/protection of the user's account. Less than 10% of tweets were inaccessible due to the deletion of a related retweet or account. Over 40% of tweet inaccessibility occurred within the first 48 hours.

Chapter 6 develops the concept of *authenticity*, pertaining to the stability metadata and linked data surrounding a social media post. I measure authenticity by comparing nightly changes in tweet metadata such as the user description, account username, number of followers, and number of retweets as well as the change in the content of hyperlinks embedded in the tweet text. In the Departments of Transportation and RuPaul's Drag Race case studies, over 50% of users changed profile information and there was pervasive linking to other tweets and social media platforms.

Chapter 7 summaries the main findings of this dissertation as well as the differences and similarities of each case study. Conclusions, limitations, and future work are also discussed.

Appendix A presents a general set of implications for social media researchers based upon the framework and findings of this dissertation for researchers and practitioners looking to collect and analyze social media data as part of a research project. Appendix B lists the keywords and accounts used as query terms for data collection in each case study.

Chapter 2. SOCIAL MEDIA AS A DATA SOURCE

This chapter focuses on social media as a data source for academic, industry, and practitioner research. Social media platforms, social media data collection, and the social science research process are examined through the theoretical lenses of process theory, platform studies, and infrastructure studies. The lens of process theory is used to situate the collection of social media data as subprocess within the process of conducting social science research. Through the lens of process theory, research and data collection are a series of linked steps, allowing for the testing and comparison of different choices the same step — for example the time between an event and collection of social media data collection or choice collection method. This process of data collection and is impacted by the affordances of each social media platform and the data collection infrastructure these platforms provide.

2.1 Social Media Sites as Infrastructure and Platforms

Within the contexts of social media data collection research questions and the research process intersect with the infrastructure and affordances offered by social media platforms. Affordances are the features a platform offers to users. Facebook offers users the ability to “like” posts. Other affordances are not offered — Facebook does not offer a “dislike” button — placing constraints on the activities of a user. As a result, the

affordances of a platform create a set of activities and interactions a user can and cannot perform within the platform.[4]

I argue, as illustrated in figure 2.1 below, research involving social media datasets is impacted by each of these layers. Some layers, such as the databases and algorithms within each platform that process and translate the activities of users into data structures and interfaces we see when accessing the platform, are hidden from public view. These algorithms and data structures shape the possibilities within the system (Ananny & Crawford, 2017; Gillespie, 2010; Vis, 2013) and, through their function, constrain what information researchers can easily consume and process. For example, each platform quantifies certain actions and offers these as part of the interface users see (Grosser, 2014). Figure 2.2 show examples of this quantification for two social media platforms, Facebook and Instagram, for the US National Park Service. The National Park Service's Instagram profile shows that the account has (1) 204 posts, (2) 554k followers, and (3) follows 430 other Instragram as well as the most recent posts from this account. Their Facebook profile categories the National Park Service as a (1) 'Government Organization' with 4.6 star rating as well as quantifying the number of Facebook users who have (2) liked the National Park Service, (3) follow their posts, (4) visited the page, as well as the (5) number of the my friends who have liked the National Park Service page.

[4] For a more lengthy discussion of the affordances of social media platforms, see Taina Bucher & Anne Helmond's article (2017).

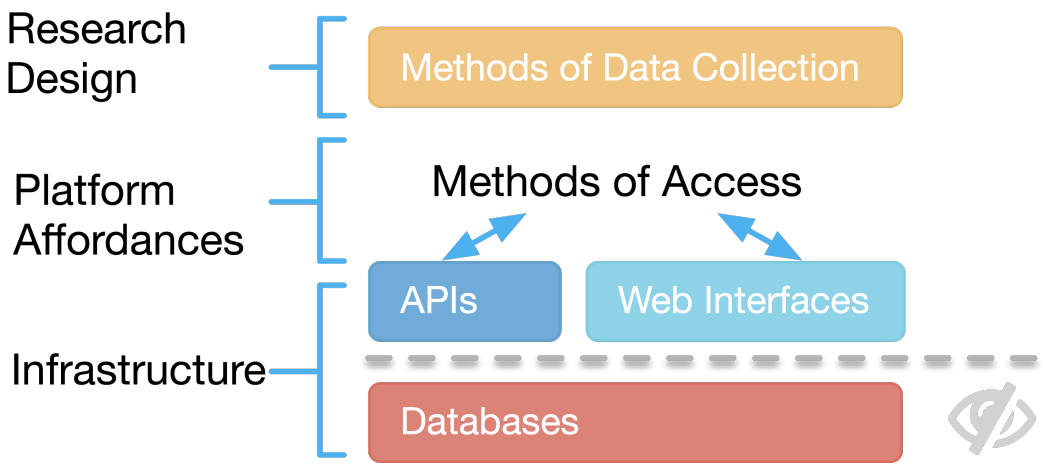

Figure 2.1: Layers impacting the social media data collection process

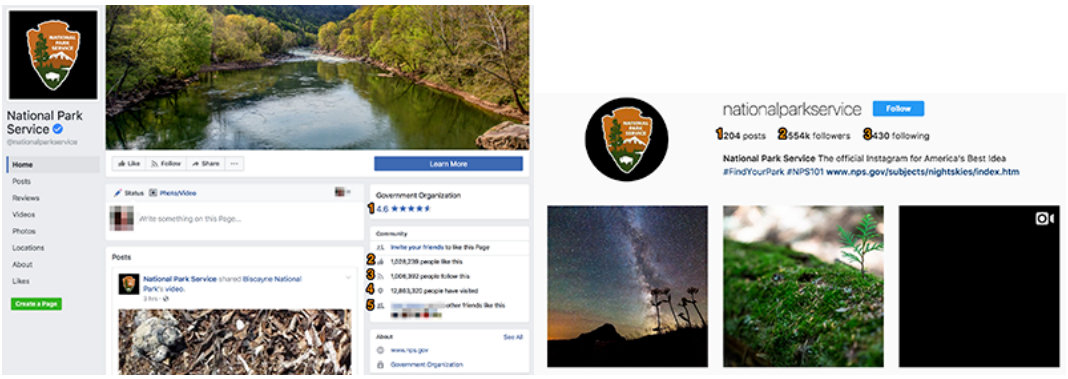

Figure 2.2: Examples of quantification offered in the US National Park Service public profile on Facebook (left) and Instagram (right) from May 2017.

The design choices made by social media sites to provide metrics for certain activities within their platform privilege some activities while limiting or preventing the visibility of other types of activities. For example, the number of times a tweet was retweeted is often used as a measure of the popularity or reach of a tweet (Starbird & Palen, 2012; Zimmer & Proferes, 2014). The number retweets are displayed prominently when viewing a tweet on the Twitter website. It is important to note that no other measures related to the number of times a tweet has been seen by users. Using the number of

retweets as a measure of popularity or influence privileges production of posts over for other types of listening (Crawford, 2009). The affordances and metrics offered in the Twitter interface are labeled in the tweet from the University of Washington shown Figure 2.3. Twitter offers three metrics for that can be used as proxies for tweet popularity[5]: (1) number of retweets, (2) number of likes, and (3) a visual proxy made up of the profile images of users who have retweeted this tweet. Below the tweet timestamp, three affordances to interact with this tweet are offered (highlighted in Figure 2.3): (1) the curved arrow allows users to rely to this tweet, (2) the square arrows allows users to retweet this tweet, and (3) the heart allows users to like this tweet.

[5] See https://dev.twitter.com/overview/api/tweets for a list of fields in a tweet, archived at https://perma.cc/439D-PKXJ.

Figure 2.3: Example of the display of activity metrics and affordances in the Twitter interface.

Social media sites take on the roles of both a platform and an infrastructure (Plantin, Lagoze, Edwards, & Sandvig, 2016) in research. Infrastructure and platform studies both refer to underlying features and structures, combined they 'take account of how rapidly "infrastructuralized platforms" have arisen in the digital age' (Plantin et al., 2016). Through these lenses are social media sites are seen as research infrastructures offering a

rigid set of affordances, or entry points, constraining our ability to access, query, format, and collect data. Entry points take two forms: (1) interfaces for human-consumption (e.g. Facebook.com, Twitter.com, and mobile applications) and (2) software interfaces designed for consumption by computer programs called Application Programming Interfaces (APIs) (e.g. Facebook Graph API, Twitter Streaming API, Instagram API, and Amazon's e-commerce APIs) (Helmond, 2015). Social media sites also offer these interfaces websites on the open web to extend their reach, decentralize data production, and centralize data collection and processing (Gerlitz & Helmond, 2013). Algorithms underlie the interfaces, mediating between users and databases. The impact of these underlying features of social media sites on research design and data collection need to be taken into account as each of these layers process and shape the resulting datasets. In this section, I briefly examine social media sites through these two lenses — as illustrated in Figure 2.1 by the bottom two layers.

The bottom layer of Figure 2.1 illustrates visible and invisible layers of infrastructures of social media sites. An infrastructure lens "makes the fundamental qualities of endurance, reliability, and the taken-for-grantedness of a technical and institutional base supporting everyday work and action" (Edwards, 2010) visible. In infrastructure studies, Ribes developed the kernel as a unit of analysis offering a lens through which to investigate the enabling capacities of an infrastructure — specifically a research infrastructure (Ribes, 2014). The kernel, a concept borrowed from computer science operating system design, is composed of (1) the "core resources and services that an infrastructure makes available" and (2) "the work, techniques, and technologies that

seek to sustain the availability of those resources over time" (Ribes, 2014). In the kernel, resources and services are entangled with the techniques and technologies used to make the resources and services available thus acknowledging the blurred nature between layers. Social media sites offer services in the form of APIs and interfaces allowing researchers to access and query data within the site. These services offer resources in the form of rendered data about users, posts, and interactions with content.

While some reverse engineering of algorithms within social media sites is possible (Ananny, 2015; Ananny & Crawford, 2017), due to the lack of transparency and speed of evolution of social media sites the majority of the processes that shape data underlying the sites' publicly accessible interfaces remain invisible and unknowable. What we can do is be cognizant of how these invisible layers constrain our ability to conduct research through the metrics, formats, and query parameters of accessible data from these platforms (Grosser, 2014; Vis, 2013). Examining social media sites through the lens of the infrastructure kernel, only two of the kernel components are visible — the resources and services that the infrastructure makes available. These resources and services are in the form of APIs and web-based interfaces sites make available to both users and researchers as well as the data presented within these interfaces. The activities of users within the site act as input into the invisible bottom layer of algorithms and data structures. These invisible components of each site's infrastructure process, shape, and render these activities into web pages and API data when researchers and users access the sites via public interfaces and APIs.

Moving up to the second layer in Figure 2.1, the platform lens, social media sites act as platforms offering a set affordances, or features, that allow users and researchers to generate and interact with data held in the data structures and algorithms of the infrastructure layer. Key features of platforms include programmability, affordances or features that allow and constrain the activities of users (Bucher & Helmond, 2017; Gibson, 1977), and accessibility of data and logic through application programming interfaces (APIs). Public APIs and web interfaces offer a set of affordances, or features, which constrain or enable users to act and interact in certain ways. For example, an API that provides items posted within the last 7 days or the Facebook web interface provides a limited number like "reactions" (like, love, haha, wow, sad, and angry) for users to respond to posts. In some cases researchers use the same web-based interfaces that users use — for example, viewing or scraping data from a user's profile or social media post — or they may use publicly available APIs to collect data from the platform. The affordances of these interfaces allow researchers entry into the infrastructure of the platform.

2.2 Collecting Social Media Data

Twitter, a social media platform founded in 2006, offers a set of computer-focused Application Programming Interfaces (APIs) for automated data collection and user-focused public web interfaces for manual data collection that researchers may use (Zimmer & Proferes, 2014). These APIs offer interfaces for scripts and applications to request information and interact with the platform. Responses are returned as a JSON

document, a computer readable format of key-value pairs. The key uniquely identifies a field and the value contains the field's data. Figure 2.4 show a tweet rendered on the Twitter website as well as the JSON document retrieved from the API. The "Name" field of the tweet displayed in figure 2.4 would contains the value "Twitter API" in the API output. It is important to note that APIs render social media posts as textual documents while the user-facing web interfaces render social media content as web documents. The JSON API output contains links to embedded content such as images and videos. The web interfaces render this content inside of the web document, as can be seen with the Twitter logo in the upper left-hand corner of the rendered tweet in figure 2.4. This is an important consideration because:

1. API data does not render content in the same format as the web interfaces platform users interact with. As a result, the interface and data researchers collect differs from the experience of platform users.
2. While the JSON API output provides pointers to linked content such as images, URLS, and videos, the content of the links is not contained within the data returned by the API. If a researcher plans to include linked content as part of their analyses, the content may change or become inaccessible between the time of data collection from the API and when linked content is accessed at a later date during analysis. For example, a researcher may collect tweets from the Streaming API in real-time and then access links when coders analyze the content weeks or months later. As a result, the content may not accurately

reflect the content users posted at the time of data collection – breaking the time bound between the social media posts and the embedded content.

API Request

```
GET https://api.twitter.com/1.1/statuses/
        user_timeline.json?screen_name=twitterapi&count=1
```

API Response

Some fields have been removed for

```
{
  "created_at": "Wed Aug 29 17:12:58 +0000 2012",
  "contributors": null,
   "text": "Introducing the Twitter Certified Products Program:
https://t.co/MjJ8xAnT",
   "retweet_count": 123,
  "id": 240859602684612608,
   "retweeted": false,
   "in_reply_to_user_id": null,
   "user": {
      "name": "Twitter API",
      "created_at": "Wed May 23 06:01:13 +0000 2007",
      "location": "San Francisco, CA",
      "favourites_count": 90,
      "utc_offset": -28800,
      "followers_count": 1212864,
      "time_zone": "Pacific Time (US & Canada)",
      "description": "The Real Twitter API. I tweet about API
changes, service issues and happily answer questions about Twitter
and our API. Don't get an answer? It's on my website.",
      "statuses_count": 3333,
      "screen_name": "twitterapi"
      ...
    }
   ...
}
```

Figure 2.4: Tweet (top), Twitter API request, and API output of tweet by @TwitterAPI.

The Twitter offers API endpoints, or connections, for the posting of tweets, modification of user accounts, and to request information about a specific user or tweet. Researchers select the endpoint relevant to the data and time period they need to collect. Twitter's API interfaces are similar to the data collection interfaces offered by other social media platforms such as Facebook (GraphAPI[6]), Instagram[7], and Baidu[8].

As described in table 2.1, each Twitter API provides access to a type of data within a specific timeframe. For example, the Streaming API provides access to up to tweets that match a set of keywords (hashtags, usernames, text, or URLs) as tweets are being posted to the platform and it rate-limited up to 1% of the entire Twitter stream. If a set of keywords match more than 1% of the Twitter stream, those tweets are not delivered and a rate limit notice is returned. Each API offers access to a specific time period of data, so the time period of the API must be matched with the time period of data access. For example, the Streaming API only allows real-time access to tweets as they are posted to Twitter, so if a researcher does not know about an event of interest in advance and is setup the data collection infrastructure prior to the event another API must be used for data collection.

[6] https://developers.facebook.com/docs/graph-api
[7] https://www.instagram.com/developer/
[8] http://developer.baidu.com/wiki/index.php?title=docs

Table 2.1: Description of the Twitter API Ecosystem

API[9]	**Time Period**	**Access**	**Description**
REST API	N/A	Public	Provides access to the current state of user profiles, timelines, and tweets. A user's screen name or a tweet's unique identifier must be known in order to be retrieve.
REST API – Search API	7 days	Public	Provides access to tweets from the last 7 days via keyword search matching the tweet's username, text, URLs, or hashtags. The documentation states that the Search API focuses on "relevance and not completeness", noting that some tweets and users may be missing from search results from the Search API. Twitter points developers and researchers are pointed to the Streaming API or GNIP for more complete datasets. The API is currently rate-limited[10] to 180 requests every 15 minutes. Each request may contain up to 100 tweets.
Streaming API – Filter	Real-time	Public	Provides real-time access to tweets via keyword matching (400 keywords), username/id (5,000 users), or a geographic bounding box (25 boxes). Researchers must maintain a constant connection to the API in order to receive data. Any disconnection will result in missing data. The API is currently rate-limited to 1% of the entire Twitter stream.
Streaming API - Sample	Real-time	Public	Provides a small random sample in real-time of all public tweets. Information is not provided on how the sample is generated or if the sample is statically random and representative.
GNIP PowerTrack	Real-time	Commercial	A commercial service from Twitter providing full-access real-time to the entire Twitter "firehose". Query options are more granular

[9] See https://dev.twitter.com/products for full technical documentation of Twitter's APIs.
[10] https://dev.twitter.com/rest/public/rate-limiting

GNIP Historical PowerTrack	Historical	Commercial	A commercial service from Twitter providing access to all non-deleted tweets from the start of twitter to the present.

Researchers may also collect data directly from Twitter's public-facing website. This has the advantage of accessing data in the same rendered format that platform's users experience as well as the inclusion of some linked content (images and videos). The search interface on the Twitter website provide access to all public tweets and is not limited to a 7-day window like the search API. Collecting data directly from the Twitter website does not lend itself to large-scale data collection like the APIs offer.

Each collection method (API vs. website) offers its own advantages and disadvantages, so researchers should choose a data collection strategy that most closely meets the requirements of their research questions and design. The best strategy may include a combination of data collection from public APIs, the public website, and archiving linked content such as media and URLs embedded in each post.

2.3 The Dimensions of Latency and Level of Automation

In the social media research space, researchers are applying existing methods to the collection and analysis of social media data. In a content analysis of the abstracts of over 500 papers focusing on Twitter from 2007 to 2011, Williams et al. (2013) found that the analysis of tweets rather than Twitter users or the Twitter site itself was the most common focus of these papers. Building on this work, Zimmer and Proferes coded 382 studies focusing on Twitter for their primary data collection and analysis published between 2006 to 2012. They created a typology of Twitter research related to the

"disciplines and methods of analysis, amount of tweets and users under analysis, the methods used to collect Twitter data, and accounts of ethical considerations related to these projects" (2014). Their findings show the amount of research utilizing Twitter data has grown from two studies in 2007 to 145 studies in 2011, with a slight dip in 2012 of only 109 studies. The fields of computer science, information science, and communications dominated. Content analysis of the text of the tweet itself was the dominant analysis with nearly two-thirds of all studies examined using this method with a majority of studies using Twitter APIs for data collection. Of the papers not using the Twitter API, manual capture or the use of a tool such as TwapperKeeper was popular. Similar work with papers focusing on Facebook as their data source found that content analysis of posts also dominated as the primary method for analysis (2016).

Based on the meta studies of social media research approaches and my experiences working with social media data, it is useful to think about the social media data collection process across two dimensions: 1) Latency (real-time vs. historical) and 2) Automation (manual vs. automated). Figure 2.5 shows data collection methods along a latency (or delay) continuum from the least (data collection in real-time) to the highest latency (data collection from a historical archive). In the middle are low latency (semi-real-time) data collection methods enabling the collection of data in near real-time — seconds to minutes after a post has been created.

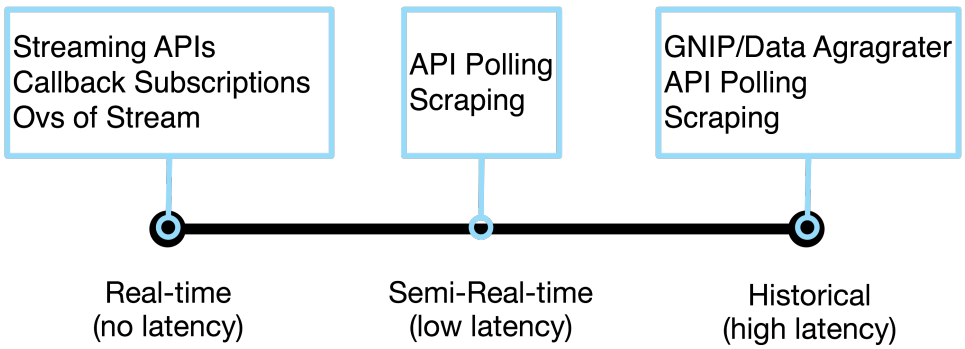

Figure 2.5: Spectrum of Social Media Data Collection Methods by Latency.

2.3.1 *Latency of Data Collection (Temporal)*

The time of data collection refers to whether social media data is collected at the time of production (real-time) or with some delay after a post has been produced (historical). Here I borrow the concept of latency, a computer networking term related to the delay in transferring information from one part of a network to anther (Gummadi, Saroiu, & Gribble, 2002; B. Zhang et al., 2006), to refer to the delay between the production of a social media post, metadata, or reference to linked content and its collection. This is separate from, but related to the time period under investigation. For real-time data collection, posts are collected immediately after they are produced or "posted"; in historical data collection, there is a latency or delay between the when a post was produced and its collection.

2.3.2 *Level of Automation (Method)*

The level of automation refers to level of manual intervention required by the method of data collection. Automated collection of social media data is normally accomplished through a small program or script freeing up researchers or their assistants from

completing the process by hand. The level of automation is separate, but rated to the method of data collection as most methods can be accomplished in an automated or manual fashion. For example, if a researcher chooses to take screenshots of social media posts, this can be done by manually loading each page and taking a screenshot or through a script that automates the process. In many cases, automation allows for the templated collection of higher volumes of data over longer periods of time since scripts can execute process faster than humans and for longer periods of time. For many researchers, who prefer high-volume, real-time data collection, automated data collection has become the "gold standard" for social media research.[11] Methods, such as grounded theory based coding (Patton, 2001), require a level of human decision making and nuisance that are less amenable to automation.

Merging the temporal and automation dimensions results in 4 possible approaches as displayed in Table 2.2.

Table 2.2: Data Collection Approaches by Time and Method

Real-Time / Manual	Real-Time / Automated
Historical / Manual	Historical / Automated

Real-Time/Manual

In this scenario, a researcher or proxy is collecting social media data at the time of production using a manual process. The data collection may be done using copy and paste, screen-shots, or by viewing the posts as they appear on the screen. For example,

[11] While some researchers treat real-time data collection as the "gold standard", I do not take a normative stance in this dissertation. Researchers should use the findings of this dissertation as they see fit in their own research.

during a political debate, a researcher could follow specific keywords and users as the debate progresses.

Real-Time/Automated

In this scenario, a researcher is collecting social media data in real-time using a social media site's Streaming API via automated script. With a streaming API, a script maintains an open connection to an API in order to receive posts related to the query in real-time — moments after they have been posted.

Historical/Manual

In this scenario, a researcher or proxy is collecting social media data after it has production using a manual process. The data collection may be done using copy and paste, screen-shots, or by viewing the posts/profiles minutes to months after they were posted. For example, months after a political debate, a researcher could search for specific hashtags and users.

Historical/Automated

In this scenario, a researcher is collects social media data after its production via a using a social media site's API via automated script. With a REST API, a script maintains polls an API with a query in order to receive posts related to the query. Each query returns a certain number of posts and the script must query the API multiple times in order to receive all of the posts related to query. This could take seconds or days depending on the rate limits imposed by the API and the number of posts matching the query.

2.4 Process Theory

The research questions guiding this dissertation are concerned with the impact of choices made during the research process, specifically how and when the collection of social media data is performed, on the resulting datasets. Process theory "argues for a patterned sequence of events [focusing on] … questions of the order and sequence of events and about the effects of that order [to determine if more] preferable outcomes can be associated with particular sequences of activities" (Abbott, 1990). A sequence is an 'ordered sample of things' that can be temporal or spatial in nature with properties of a continuous or discrete variable. These become events when tied together into temporal sequences (Abbott, 1990).

Using process theory, a given set of sequence patterns can be examined to understand why they are the way there are or the effect of a certain set of sequence patterns. Examples of the former include: "Does education determine the characteristic sequence of career? Does the size of an organization determine the shape of the status rankings we find in it?" Examples of the latter include: [Are] "those promoted before acquiring certain kinds of expertise are helped or hindered in their ultimate career success"? I focus on the second of these questions — the effect of a certain set of sequences on a particular outcome. The research questions to be addressed focus on the effect of the ephemerality of social media data on the research process and characteristics of the resulting datasets. The sequence of events I focus on is the research process, and the data collection design choices made during this process. Within this data collection subsequence, I am interested in quantifying the effect of ephemerality on the reliability

and authenticity of the social media data collected using these processes. Thus, through the lens of process theory, social media data collection by researchers becomes a sequence of events. A process theory approach allows for the examination of the impact of changes to the sequence. Thus allowing for the examination of the impact of ephemerality on social media data sets using different data collection procedures.

2.4.1 *The Social Science Research Process*

The research process is the process researchers go through in order to achieve their desired research outcome. The lens of process theory (Crowston, 2000) views processes as a way of accomplishing goals and transforming inputs into outputs, allowing the subprocess of social media data collection to be situated within the research process. This approach stands as an alternative to existing work focusing on the use of tools to collect social media data; often obscuring the methodological choices and epistemologies embodied and hidden within the tool.

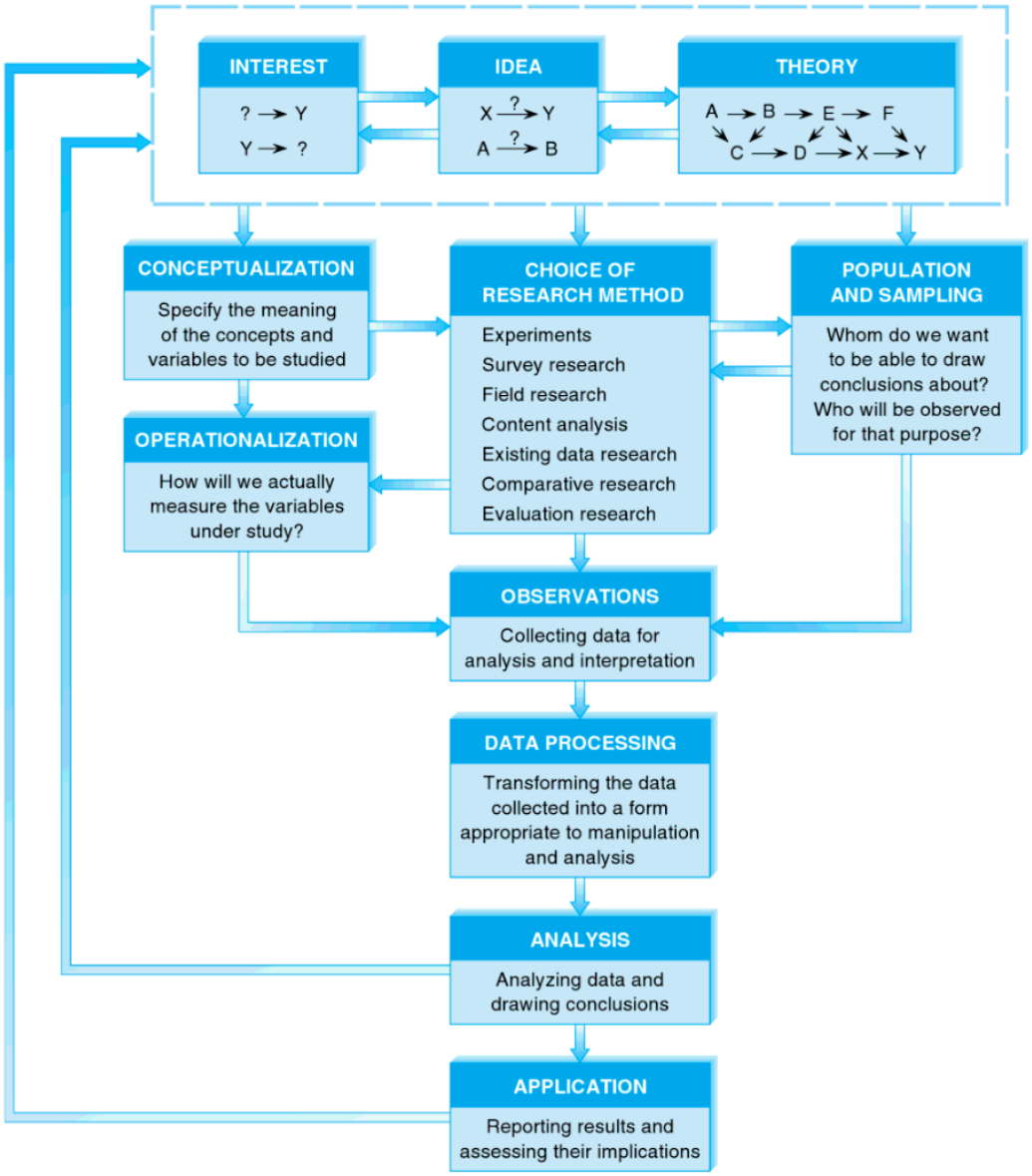

Figure 2.6: The Research Process from The Practice of Social Research (Babbie, 2007, p. 108).

Consider the above diagram (Figure 2.6) presenting a high-level overview of the research process as described by Babbie in The Practice of Social Research (2007). When viewed through a process theory lens, ideas, interests, and theories act as inputs in the research process leading to outputs — research findings and applications. Generally, the genesis of research is the ideation phase, ignited by an idea, some interest in a phenomenon, and/or a theoretical frame. From that point, a researcher may do

exploratory work and/or reading of prior studies to better understand the phenomenon and sites of observation for data collection. Thus, the diagram begins with “interests, ideas, and theory” with double arrows between them representing the bidirectional movement between the three. For example, an interest may lead to an idea which is further developed through theory, generating new ideas.

Once a researcher’s ideas, interests, and theories are honed into a more well-defined purpose and list of outcomes, the conceptualization, choice of research method, population and sampling methods, and operationalization of variables must be determined. Again, this process is iterative with each step influencing the others, occurring in any order. Conceptualization involves specifying the meaning of each concept in the research. In research designs using highly structured methods, such as surveys and experiments, concepts may need to be well-defined in advance. In other cases, such as with open-ended interviews, the goal of the research may be to uncover the meaning of certain concepts so these concepts may not be well-defined at the start of the project. Single or multiple research methods are then chosen based on their appropriateness to address the research question(s) and the constraints of the available data and skills of researchers involved in the project. Operationalization is the process of determining the measurement techniques for each variable. The population and sampling methodology details the group under investigation. Since it’s normally not possible to observe every member of a population (complete data), researchers specify a sample to be collected and analyzed.

At this point, the researcher has decided what to study among what population and to do that through a specific method or set of methods. Observations and data collection can now commence. Once data has been collected, it is often not in a form lending to analysis or interpretation, so it must be processed and cleaned. “Unprocessed” data is cleaned and reformatted for analysis. In the cleaning step any erroneous and invalid data is filtered out — it should be noted that erroneous data is different than outlying data. The “processed” dataset, if necessary, can now be reformatted for analysis and analyses performed. The final step and output of the process, application, involves packaging and communicating the results of the study. Methods of communicating the results of a research study include, but are not limited to, publishing peer-reviewed articles, presenting at conferences or public forums, or writing a blog post.

2.4.2 *Social Media Data Collection as Process*

Data collection is a subprocess occurring within the larger research process described in the previous section. Determining what data to collect, what platform(s) to collect data from, and how to collect are precursors to starting data collection — these steps occur during the development of the research idea and goals, conceptualization, operationalization of variables, selection of the population and sample, and the selection of the research method(s). As shown in Figure 2.5, each of these steps inform the process of data collection or observation of a phenomena. The data collection process occurs within this subprocess of observation, but as mentioned, the process does not occur in a vacuum but is informed by all of the stops occurring before. While social media data are

just one type of researcher data source, the context has unique features that make the data collection choices extremely impactful.

Consider the example introduced in the introduction, of the researcher interested in using the #YesAllWomen campaign to study misogyny online. Before collecting data, a researcher must conceptualize the research, choose one or more research methods, bound a population and sample, and operationalize the variables. Let's run through an example project using this case in order to show how the social media research process is situated within and connected to the other elements of the research process.

Refining these ideas into a more tangible research project, imagine that the researcher is interested in discovering common factors between Twitter accounts which are targets of hate and misogyny within the #YesAllWomen hashtag. Important terms such as misogyny and prominent users would need to be conceptualized and operationalized. After determining the indicators of a misogynistic tweet, content analysis may arise as the most appropriate research method. Tweets could then be collected and classified as to whether they contain misogynistic content in order to produce a corpus of misogynistic tweets. This corpus could be examined to find the most mentioned user accounts. Prominent user's public profiles, including the profile image, profile description, and timeline of tweets, could be examined to determine common factors in how the accounts present themselves or the types of tweets in their timeline.

This scenario illustrates multiple issues of ephemerality during social media data collection process:

- When coding tweets for misogynistic content, what content will be examined? As discussed in Will linked content or media embedded in the tweet be included in the coding process or will the coding only focus on the text of the tweet? If included in the coding process, will embedded media and URLs be archived at the time of data collection or will coders open the URLs and media at the while coding the tweet? Will the content of the URLs and images change between the time of tweet collection and coding occurs?
- Are the assumptions inherent in the methods used met? For example, content analysis assumes a certain level of stability in the dataset (Karlsson, 2012; Krippendorff, 2012; Saltzis, 2012) and parametric statistics assume certain normalized distributions of data. Also, after determining what accounts are the most prominent in the misogynistic corpus, some accounts may be deleted, made private, or public details such as profile images and descriptions change due to the level of harassment they received? How might these changes impact the research findings? These examples point to the central issues of the reliability and authenticity of social media data that lie at the heart of this dissertation.

2.5 Chapter Summary

In this chapter I have discussed social media as a data source for research and provided a framework for understanding social media platforms as data collection infrastructure. The framework helps researchers understand how the affordances of social media platforms constrain and shape their ability to collect data from these

platforms. This chapter also discusses methods of data collection from social media platforms with a specific focus on Twitter. The process of data collection was then situated within the larger social science research process.

Chapter 3. SOCIAL MEDIA AS A RECORD

Often the analysis of social media posts focuses on either volume-based metrics or the text of a posts within bounded sets of keywords/accounts (Zimmer & Proferes, 2014) on a singular platform (see for a discussion of "hashtag studies" Burgess & Bruns, 2014). These approaches do not take into account that social media posts contain an assemblage of text, images, metadata, and hyperlinks. When accessing a post via a platform's website or API, the post and its accompanying metadata are assembled at the time of the request. Embedded content and metadata surrounding a post may change independently of the text of a post, breaking the time-based bond between a post and the surrounding metadata and content. Since researchers often use social media data as a historical record or documentation of an event or phenomena occurring at a specific time , changes in the accessibility of posts or content of embedded metadata may have an impact on research findings since the content may no longer be reflective of what was posted by the user. Some changes to web pages, such as hourly updates to the BBC homepage or 404 Not Found error pages, may be more easily recognizable and quantifiable. Other changes, such as a link redirecting to a new location or the deletion of a user's account, may be less obvious. Current practices solely focus on collecting social media posts and often assume a high level of stability in social media data sets which does not reflect the experience of many researchers as evidenced by the treatment of real-time data collect as a "gold-standard" in social media research (Burgess & Bruns, 2014; Driscoll & Walker, 2014; González-Bailón et al., 2014).

Current informal research practices, derived from the choices described in the methods sections of early publications exploring the use of social media data (Bruns, 2012; ex: Bruns & Burgess, 2011), were written when the collection of social media data was experimental, novel, and the platforms were just starting to emerge. These practices coalesce around the "large-scale" collection of social media data via automated scripts and public APIs offered by social media platforms. Often the choices made in bounding a case, the related keywords and account, the method of collection, and data cleaning are briefly described in the methods section of these papers. While shedding some light on reasoning behind and the implementation of these choice, they normally do not "include enough detail about how the studies were actually conducted on the ground to allow for their replication" (Hargittai & Sandvig, 2015, p. 2), leaving researchers without a comprehensive framework through which to employ similar methods or to determine the best approach for their own research questions.

In this chapter I develop a framework for social media data collection based on relevant concepts borrowed from archival and electronic records theory. Archival theory and practice focuses on the acquisition, arrangement, description, and preservation of objects and records in library collections. Electronic records theory builds on approaches in archival science for the management and preservation of integrity of legal and business records in an electronic environment. In this chapter, I draw on elements from both theories to develop a framework to expand our strategies in collecting data, incorporating multiple components (post text, metadata, linked content) and greater environment in which posts are generated by users and platforms.

3.1 Archival Theory and Preservation

There are two dominant models in the curation process — the older lifecycle model conceives records as living organisms. It is heavily used in the records management literature and practice based on a sort of cradle to grave understanding of records where archives are part of the “end-of-life” management. In the lifecycle model, records pass through stages until they die. While life cycle concept has been taken up in studies of data ("data life cycle") and many have pointed out that the model is troubling, namely that things are born and they eventually die, or they may not mirror life stages of development. In contrast, Australian archival scholars (McKemmish, 2001) have developed the "records continuum model" which suggests that records live on in many iterations, perhaps even after they end/die. In this model, “records are 'fixed' in time and space from the moment of their creation, but record-keeping regimes carry them forward and enable their use for multiple purposes by delivering them to people living in different times and spaces” (Pearce-Moses, 2005).

Both models include at least four main stages: 1) appraising the historical value of a record, 2) accessing an item into an archive, 3) arranging and describing items in the archive, and 4) preservation of items in an archive (Acker, 2014; Daniels, Walch, & Service, 1984). Appraisal is ”the process of establishing the value of documents made or received in the course of the conduct of affairs, qualifying that value, and determining its duration” (Duranti, 1994). An archivist uses this assessment to determine if an item should become part of the collection and, if so, would move on the next stage of the

process. Inherent in archival practice is the recognition of impossibility of collecting and preserving every record, only items deemed to have a high value and relevance are accessed or brought into the archive. Also, an archive is limited to the records available. As a result, all archives are incomplete with gaps in their record (Thumim,2002).

After accessing an item, it is arranged and described. Arrangement involves the physical placement of records, often mirroring the arrangement and ordering that was given to the archive. This physical placement of records also represents the association and relation with all of the other documents received as part of that collection (Holmes, 1964). Collections are then integrated into the larger arrangement at the depository, record group, and filing unit.

A description of the record is then recorded which includes information related to the creator, dates, and content to facilitate the management and finding of the record. Finally, the physical or digital item is preserved to prevent further degradation. This is part of an ongoing process.

As an example, consider the example of a prominent politician donating her collection of letters to the local university library. An archivist would first visit the collection of items to collect or access the appropriate items into the collection. After accessing the items, they will be taken back to the library to be arranged and each record will be described.

The archival process mirrors the social media data collection process making this a good model to draw relevant concepts from. Researchers appraise the value of data collection, develop and execute a research design to collect and analyze those records,

after collect data is arranged and described for analysis, and, as part of collection, data is preserved in a format necessary for analysis. Social media data collection, like web archiving to a certain extent, collapses the traditional archival lifecycle into one step.

3.2 Applicable Concepts From Archival Theory

In the following sections, I describe the relevant archival concepts from archival and electronic records theory applying each concept to social media data.

- **Action**. A core component of every record is that they participate in some action. This falls into types: dispositive (action comes into existence with the creation of a record - contract of sale / enter of relevant information in patient record substantiates admittance to hospital), probative (record acts of proof action took place such as a marriage document), narrative (records that are the substance of non-legal actions - eg. most email), and supporting (help carry out an oral action such as lecture notes or a meeting agenda) (Duranti, Eastwood, & MacNeil, 2013). Social media posts serve as a record of the act of producing and interacting with a posts and social media platforms.
- **Archival bond.** The archival bond web of relationships that each record has at the moment it was made or received with the records that belong in the same aggregation. In a traditional collection, the archival bond carries from the implicit physical arrangement records. The archival bond in an electronic record is made up of the classification codes assigned to records, connecting it to other records belonging to the same class (Duranti, 1997; Duranti et al., 2013). In social media

data sets the archival bond consists of the (inter)relationships between social media posts in the same reply/retweet/hashtag stream and the same content aggregation (about the same topic). In social science research a content aggregation is analogous to the bounds of a case study.

- **Context.** The context is anything outside a record that has significance for its meeting. Relevant contexts could include the legal and organizations system in which the record creation took place, procedures used in the course of creating a record, and 4) *documentary context*: fonds, the whole of the records that a person naturally accumulates by its activities and the byproducts of them, and internal structure. Electronic records theory expands the documentary context to include the *technological context* which includes the technological characteristics of record keeping system (Duranti et al., 2013). Within the context of social media datasets, the context includes the documenting the technological context of the platform that the user used to produce the post — this is especially important as platforms evolve over time. Consider that Twitter of 2011 has different features and configurations than Twitter of 2017.
- **Physical Form.** In an electronic record, the physical form includes the configuration and architecture of the electronic operating system, architecture of electronic records, the software, all those parts of the technological framework that determine what the document will look like and how it will be accessed, and digital signatures and time-stamps (Duranti et al., 2013). The majority of these are invisible to the user, but any migration or small change would generate a new

and different record. In the case of social media data, this would include information about the form or format the researcher collected data.

- **Content.** The context consists of the textual, symbolic, and/or visual message that is meant to be conveyed. Content must be fixed and stable in order for record to exist and cannot be separated from its form or its medium (Duranti et al., 2013). In the context of social media data, content includes the post text, links, video, or images — making an argument for the content of a social media post to be more than just the 'text' of the post.

3.3 Social Media as a Record

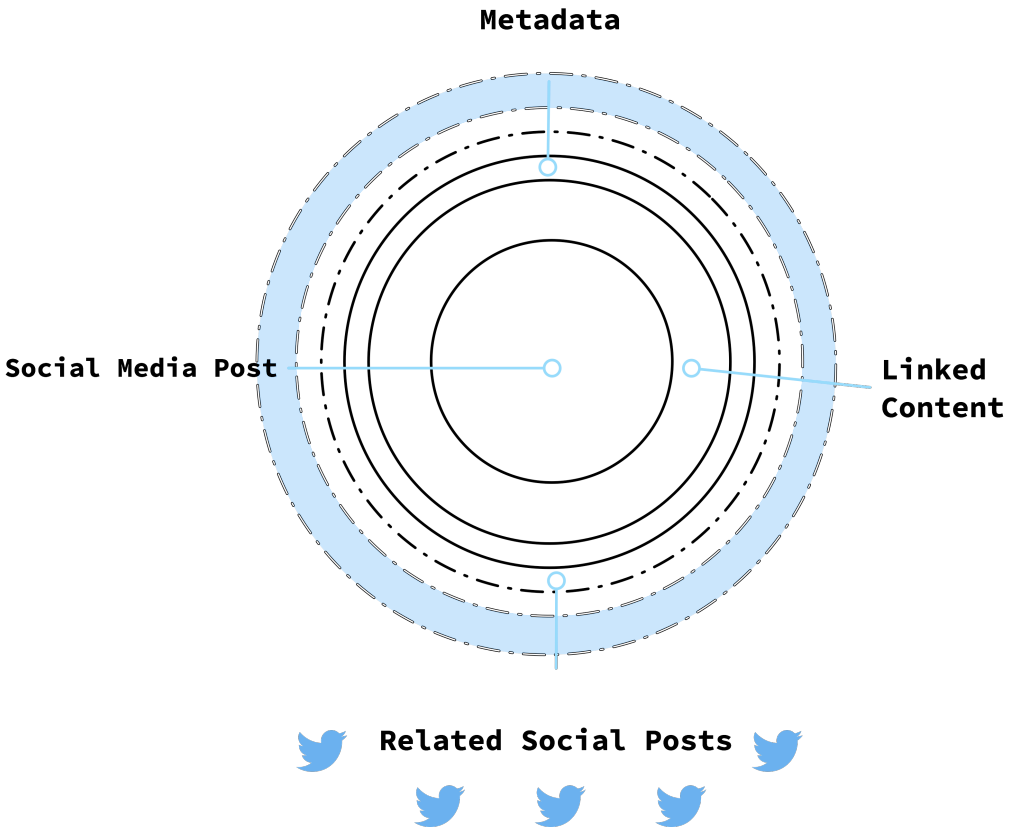

Figure 3.1: Diagram of the social media as a record framework.

A framework derived from relevant concepts discussed in the previous section is imagined in Figure 3.1. The framework has 5 elements: 1) Layers of context and infrastructure which a post is produced and embedded, 2) social media post, 3) linked content contained in the post, 4) metadata surrounding the post, and 5) related social media posts.

At the center of the framework is the content of the social media post, often rendered as “the text of the post” through social media platform APIs. The second layer consists of linked content — URLs and media content (images, videos, etc.) are often embedded within this text. Surrounding, documenting, and describing the post is the metadata associated with the post. This can take the form metadata about the time the post was created, what client was used to create the post, metrics about the post such as the number of likes, and user information. The social media post, its linked elements, and metadata are all linked by a specific point in time. The post reverses to a specific state of the user profile, embedded links and images, and other metadata connected to the post at single moment in time. Viewing these items disconnected from that bound moment in time may result in viewing a different post and content than the user intended. In the final layer, the post is connected to other posts within the same platform and in other platforms bring proceeded at that same moment in time within a specific human context.

3.4 Chapter Summary

Seeing social media posts through the lens of a record offers researchers a guide to approach both the collection and analysis of social media posts. This framework based on

relevant concepts in archival theory, expands the conception of a social media post beyond “just its text” and illustrates interconnectedness of the elements of a post as well as the time-bound nature of the elements.

Chapter 4. EPHEMERALITY

The concept of ephemerality is often used in Internet and Social Media Studies, but rarely defined. In this chapter I develop and operationalize the concept of ephemerality as well as describe the core data collection methods and cases used in this dissertation. The chapter concludes with descriptive statistics for each case under examination.

4.1 Conceptualizing Ephemerality

Researchers use the term ephemerality when discussing the ever changing or impermanent nature of social media. Issues of ephemerality in research are not new or unique to digital or social media data, scholars in many fields, including film and internet studies, have wrestled with the issue. In the case of film studies, consider representations of early histories of 1950s British broadcasting. When the BBC started archiving footage in the 1950s, it prioritized the recording of documentaries for inclusion in its early archives(Thumim, 2002). As a result, the absence of audiovisual archives created false assumptions about the reality of early television because only footage deemed “important” enough to record was archived. The “decision of what to archive and not archive created ‘particular bundles of silences’” (Thumim, 2002). These silences exist in the gap between what was broadcast by the BBC and what was archived, presenting two different images of reality — this is the ephemerality of early television.

Similar issues of ephemerality are faced by researchers using web archives. At the root of the internet is a hypertext system in which data is stored in a network of nodes

connected by links (Smith & Weiss, 1998). These links, commonly called hyperlinks (URLs), serve as the primary mechanism that connects nodes in the web to one another, and are technological affordances that allow seamless connection between one website and another (Park & Thelwall, 2003). Estimates of link decay (also known as: half-life, death, accessibility, persistence, and link rot) mainly come from studies of links in journal articles and range 31% - 39% (Dimitrova & Bugeja, 2007; Goh & Ng, 2007; Moghaddam, Saberi, & Esmaeel, 2012; Sanderson, Phillips, & Van de Sompel, 2011). These studies use a combination of automated analysis of error codes returned by web servers or rely on researchers visiting each of the URLs. These methods only detect obvious cases where the destination of the URL returns an error. In addition, focus on "404 not found" errors of links in academic journals tell us little about the other ways links can change, decay, disappear, or erode in other contexts.

Publicly accessible web archives, such as the Internet Archive's Wayback Machine (https://archive.org/web/), do not provide a "magic" solution to issue of link decay and change since only a handful of the web is archives. Even these 'saved' pages create issues such as "broken links, missing images, and code written for outdated pages" (Ankerson, 2012). Dynamic content including flash animations provide another challenge because some types of embedded content are not archived by the automated systems used by web crawlers such as the Wayback machine. This results in a "broken flash image" and unaccessible content (Ankerson, 2012).

Ankerson encourages scholars working with web histories and archives to look at strategies used by broadcast historians who, as with the aforementioned BBC example,

understand “well the difficulties in piecing together the past when so much of what was broadcast was sent out live and unrecorded” (2012). Although it is important to note that most broadcast historians relay on centralized corporate and institutional archives[12] which often archive ephemera related to broadcast shows, such as internal memoranda, letters, press cuttings, reports, and much more. This differs from web archives which are (partially) “preserved digital files with proper contextualization” (Ankerson, 2012). As Anderson acknowledges, these two examples experience different problems due to the ephemeral nature of the content (web and broadcast) — with the BBC archives containing large amounts of contextual resources but lacking extensive archives of full broadcasts while web archives offer a plethora of preserved sites, with its own complexities, but often lack contextual information.

Going back to the case of social media, researchers using social media data will stumble upon both a lack of context and a lack of archived material. Social media inherits many of the characteristics and complexities of the web since many platforms have web-based interfaces and allow the inclusion of linked elements. A post or account may be deleted; linked content such as images, videos, or web pages may change, decay, or disappear but this is rarely discussed in papers using social media data. Consider two cases – the Occupy Wall Street Movement (OWS) and the Boston Marathon bombings. In both cases, social media were used to share event specific content. During the Occupy protests, protesters posted images of police actions in order to assist other protesters in

[12] Interview with Brewster Kahle. RLG DigiNews 6(3). Available at:http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file3096.ml

avoiding these actions. Once police actions had ceased, posts and their accompanying images were deleted by users; using deletion as a protest tactic (Neumayer & Stald, 2014). Similar behavior was seen after the Boston Marathon bombings, when rumors about the identity and location of the bombers were abundant (Starbird et al., 2014). As more information was released by responding organizations, posts containing misinformation (i.e. false rumors) were deleted from timelines.

In both of these cases, the real-time record differed from the retrospective record, leading to similar 'bundles of silences' experienced by researchers using early BBC archives. Researchers collecting social media posts in real-time likely ended up with a different dataset than those who collected or purchased data weeks or months later. As a result, protest tactics during OWS would look significantly different between these two datasets since some practices were meant to be temporary in nature. Similar issues would emerge in the Boston Marathon bombings dataset— after correction, some rumors and misinformation might disappear entirely.

These two examples do not take linked content such as web links into account – news articles are updated as a story progresses (Saltzis, 2012) and web forums, such as reddit, may delete or modify content over time. An important concept developed in the previous Chapters 2 and 3 is that social media posts are made up of more than just the text of the post itself – they often contain links to web content, videos and images that extend the post and are an integral part of the post. Without the content of the link, whether that is a picture or web page, it may be impossible to understand the post. The post and link may change together, or one may change while the other remains untouched, but they

are both bounded by a specific time – the time when the post is created. If the two are out of sync then the relationship, meaning, and context of the post and link may be disrupted.

Gray et al's (Gray, Szalay, Thakar, & Stoughton, 2002) describe ephemeral data as data and metadata describing that data which cannot be replaced, reproduced, or reconstructed; therefore necessitating the archiving of that data. Stable data, in contrast, only requires the preservation of the metadata documentating its creation so it can more easily be reconstructed. Social media data easily fits into the category of data that cannot be reconstructed after the fact.

The ability to reconstruct a dataset is an important and distinct from the ability to acquire a dataset. In the case of social media data, while it is possible to collect it using a myriad of methods (manual copying and pasting, automated collection from APIs, reading of profiles/walls/timelines, or purchasing data from aggregators); these methods do not necessarily preserve the data for future data collection endeavors. For example, the SoMe Lab has an archive of over 350 million tweets related to the Occupy Wall Street movement (Agarwal et al., 2014) but it would not be possible to purchase the same exact dataset that the lab collected — even using the same collection parameters. That is because accounts, posts, and links have been deleted or modified. It is also important to note that deletes cascade in that when someone deletes an account, any retweets of their tweets are also deleted from the timelines of users who retweeted them. So while it is possible to go to a social media data aggregator such as GNIP (now a part

of Twitter and the only authorized provider of Twitter data[13]) to purchase tweets from their historical collection, the tweets are filtered through a deletion list before being delivered to the customer. This is in contrast to real-time data collection from Twitter's Streaming API. With the Streaming API, tweets matching the search criteria are delivered shortly after a user posts content (often within seconds).

Due to limitations of the Terms of Service (ToS), Twitter[14], like most social media platforms, restricts the ability of researchers to share the data they have collected from the service. Resections are similar to confidential datasets such as the US Census (Abowd, Vilhuber, & Block, 2012). These restrictions mean researchers and librarians cannot publish and archive data as suggested by Gray et al. While it is possible to share aggregations of social media data such as the number of posts, likes, retweets, and the results of analysis; the ToS for most platforms only allow for the sharing of the unique identifier associated (e.g. Tweet ID or Facebook post ID) with each post. It is not possible to share the actual content of or metadata associated with a post. Using the unique ID of the post, researchers with the requisite technical skills can use a social media site's public APIs to programmatically "rehydrate" each of the social media posts. The API returns the current content and metadata associated with the post, if it is still accessible. While this solves the issue of telling other researchers "what posts are in my dataset" and provides a method of comparing the posts in different datasets, it creates a series of problems of its

[13] While Twitter gifted a copy of the twitter archive to the Library of Congress, it has not been make available to researchers or members of the public. The limitations and filtering of deleted/inaccessible would content apply to both GNIP and the LoC Twitter archives.
[14] See section 6b of the Twitter Developer Policy at https://dev.twitter.com/overview/terms/policy, archived at https://perma.cc/Y64D-JDLC.

own. Three major issues emerge across many social media platforms: 1) deleted posts and posts from deleted accounts cannot be retrieved from the API so we can be left with orphaned data, 2) modified posts are not flagged by the API so we do not have a way to determine if a post changed since its creation, and 3) large datasets are difficult and time-consuming to rehydrate due to API request limits. For example, the Twitter REST API is rate limited to 150 requests per hour, returning a maximum of 100 tweets per request. As a result a researcher with a single Twitter account, can ideally "rehydrate" up to 15,000 per hour. While it is possible to get around these limits by using multiple account simultaneously, doing so increases the technical complexity of the rehydration process. The result is that this is not a viable solution. The post IDs themselves falls under Gray et al's definition of ephemeral data since, in most instances, it cannot be reconstructed. When posts changes or disappears, it may end up being a research "opportunity lost forever" (Lynch, 2008) or present a false account of the phenomena under study.

4.2 Ephemerality and Social Media Data

Some scholars such as Herring (2010) and Karlsson (2012), express the concern that structural features of new media (such as hyperlinks) and embedded media content created through them are simply too 'new' to be addressed by 'old' or existing methods of content analysis alone. Content analysis, as described by Krippendorf, assumes a high level of stability of the data being coding (2012). Scholars also note that sampling procedures in the context of social media analysis are far from being understood (Gerlitz

& Rieder, 2013). Without data concerning the stability of social media datasets, scholars are often forced to use the methods they are already familiar with and are unable to adapt existing methods to this new space.

The majority of the literature surrounding social media datasets and ephemerality focuses on deleted posts — especially tweets. Almuhimedi et al. (2013) preformed a large-scale analysis of tweet deletion. In their dataset of over 67 million tweets, only 2.4% of tweets were deleted, however 50% of roughly 300,000 users have deleted a tweet. Petrovich et al used tweet and account features (number of words, presence of curse words, number of followers, number of tweets, etc.) to predict deletions (2013). This is similar to the methods used to predict deleted emails (Dabbish, Venolia, & Cadiz, 2003) via certain features of the email message. Of the 200,000 tweets examined, 85.2% were manually deleted by the user, 12.2% were inaccessible due to a changing the account from public to private, and remaining 2.6% were due to deletion of the account. The authors posited that account deletions were due to Twitter acting on violations of their SPAM policy. Other studies have examined deletions due to government censorship in Chinese social media platforms (Bamman, O'Connor, & Smith, 2012), deletions as a protest tactic (Bamman et al., 2012; Neumayer & Stald, 2014), and deletions or changes by site administrators related to bullying or the posting of inappropriate behavior (W. Phillips, 2011). Changing policies of sites can also lead to deletions and changes of content and profiles — for example some public libraries created a Facebook “user” account to interact with patrons, but this violated Facebook’s policy so the user was deleted (Roblyer, McDaniel, Webb, Herman, & Witty, 2010).

A specific class of deletions that has been studied includes ‘regret tweets’. Sleeper et al. (2013) used an Amazon Mechanical Turk task to understand the types of regret users experience regarding content they have tweeted. Of the 474 responses (using the regret categories from Knapp et al. (1986)) the most common cause for regret was revealing too much in the tweet (e.g. personal information or a secret), followed by direct criticism regarding a specific person. Participants rarely reported experiencing regret due to lying or ‘behavioral edict’. Of the participants who experienced regret due to a specific tweet, only 52% of the tweets were actually deleted. This further highlights the difficulty in deletion prediction, that even if a tweet has cause for deletion, it may very well remain on twitter. Similar work by Zhou et al. also focuses on responses to regrettable tweets (2016).

The majority of social media deletion studies gather platform-wide data by connecting to a public API for a number of days to ingest publicly available deletion notices. This shows the gap between how content deletions are studied and how researchers often conduct their research. In that research has focused on “all deletes” from a stream vs. the topic bounded case studies many researchers use. Also these studies only take the deletion of a post into account — missing edits to a post (if a platform offers such affordances), changes to a user’s profile or presentation, embedded media such as images and videos, or linked content.

4.3 Research Design & Methods

As discussed in the previous sections, social media data is often labeled ephemeral, or unstable over time, based on researchers' experiences; however we lack empirical data to support or refute this claim — especially through a social science lens. Existing studies have focused on deletions, only one aspect of ephemerality, by monitoring platform-wide deletion notifications from social media APIs (Almuhimedi et al., 2013; Petrovic et al., 2013; Zhou et al., 2016). As a result, researchers have little understanding of ephemerality and impact of delays in data collection on social media datasets and therefore research findings. My research design addresses this gap through the investigation of the following research questions:

- RQA: How does the ephemerality of SMD interact with the process of data collection to impact the reliability of social media data sets?
- RQB: How does the ephemerality of SMD interact with the process of data collection to impact the authenticity of social media data sets?

To address these research questions, I conducted an empirical analysis of three Twitter-based case studies. A case study approach was chosen because it emulates the conditions, contexts, methods, and topic/population bounding commonly used in a social science approach. It also provides a foundation for generalized guidelines or best practices for social media based research designs for future researchers. The studies and their collection parameters are described in the next chapter. The case studies focus on the social media platform Twitter because: 1) the use of Twitter as an object of study and source of observational data is pervasive in academic research (Williams et al., 2013;

Zimmer & Proferes, 2014), 2) Twitter is less susceptible to algorithmic filtering, also called 'filter bubbles' (Bozdag, 2013; Bruns & Stieglitz, 2012; Bucher, 2012; Flaxman, Goel, & Rao, 2013; van Dijck, 2013, p. 75), than other platforms since the public APIs return all public, non-deleted statuses matching query terms; theoretically producing a more "accurate" record[15] (Driscoll & Walker, 2014), and 3) concepts, structures, metadata, and links (URLs) easily generalize beyond Twitter to other social media services and platforms.

Each case study was chosen because it represents specific prototypical features (e.g. time scale, account stability, context, level of image and media usage) of the types cases and therefore the types of social media datasets social science researchers might encounter. This design allows for a combination of within and between case analysis to understand ephemerality within the prototypical cases. Across case analysis allows for generalization to other social media platforms beyond Twitter.

4.4 Data Collection

Since the concepts of reliability and authenticity in social media datasets refer to the stability of specific components of a social media post over time, I collected tweets related to each case study at three different points in time. Using the framework described in the Chapter 3, I collected data in real-time, semi-real-time, and nightly. To allow for a longitudinal data collection over a period of two weeks, I used the appropriate Twitter APIs matching each of the concepts in the framework. For example,

[15] https://dev.twitter.com/streaming/overview, archived at https://perma.cc/AT9U-KEWJ.

it would be difficult to manually collect data 24/7 for a period of two weeks so automated data collection from the Twitter Streaming API is used to collect data in real-time. This should not be seen as a privileging of automated data collection, but a recognition that automated data collection most closely aligns with the continuous, longitudinal data collection strategy needed — it would be very difficult to manually copy/paste tweets for a period of two weeks. Tweets were collected in real-time from the Twitter Streaming API over a period of two weeks. Semi-real-time data was collected from the Twitter REST API two weeks after real-time collection concluded. The Twitter Search API was queried daily to check the availability of each tweet collected for a period of 90 days after each tweet was collected in real-time. This process is summarized in Figure 4.1 below and described in more detail in the sections below.

For each case study, the same query terms and parameters were used for each data collection method. As mentioned in the second chapter, it is not possible to use the same API or collection method to collect tweets at different points in time. Each API is specifically designed to support a certain level of latency. For example, the Twitter Streaming API only supports real-time data collection and, as a result, cannot be used to collect tweets after they have been posted. The publicly accessible Streaming API provides access to up to 1% of the current Twitter stream; tweet volumes over 1% are rate limited and not accessible.[16] The REST API provides access to the past 14 days of

[16] While the two case studies requiring new data collection have been chosen to avoid rate-limiting by the Streaming API, it is theoretically possible for it to be an issue during data collection.

non-deleted, public tweets (Driscoll & Walker, 2014; González-Bailón et al., 2014) using query terms or access to all non-deleted tweets and users via their unique identifiers.

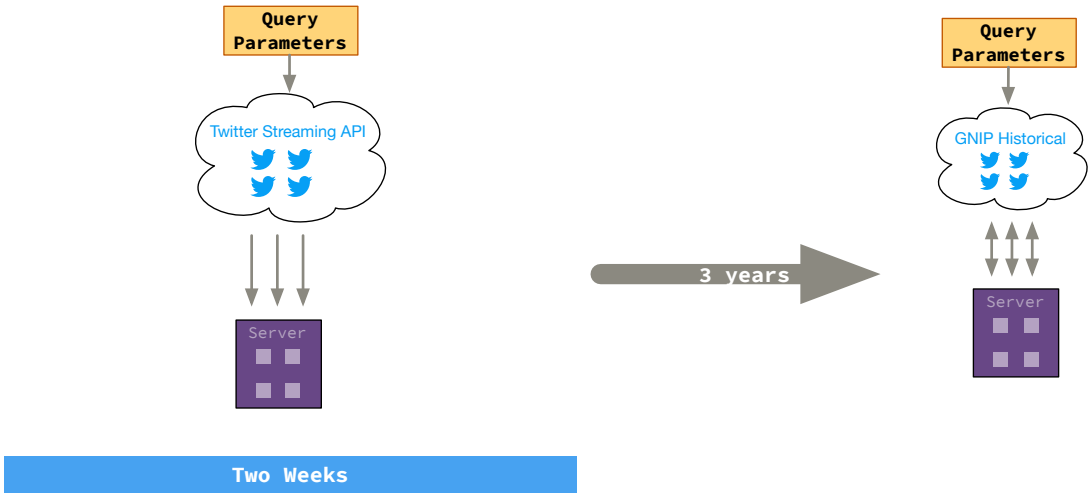

Figure 4.1: Summary of the data collection process and timeline for Occupy Wall Street case study.

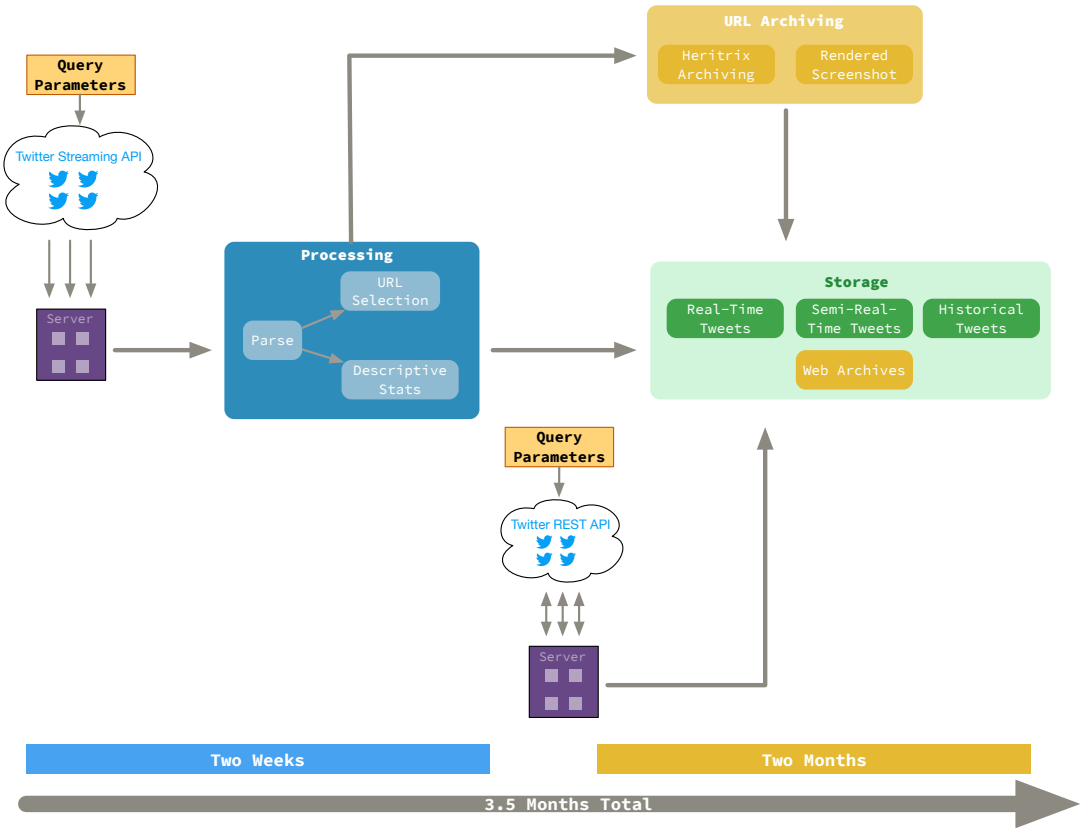

Figure 4.2: Summary of the data collection process and timeline for Departments of Transportation and RuPaul's Drag Race case studies.

4.4.1 *Real-Time Data Collection*

For real-time collection from the Streaming API, a script maintained a constant connection to the API for a period of two weeks. Since the Streaming API delivers tweets in real-time, a connection to the Twitter API must be maintained at all times during the data collection period. Any drop in connection between the script and API will result in missed tweets during the period of disconnect. As each tweet was received by the script from the API, the metadata associated with each tweet was examined for URLs and media. Tweets were randomly selected for immediate and ongoing weekly archiving of

embedded URLs and media through three processes: 1) by the heritrix (Mohr, Stack, Ranitovic, Avery, & Kimpton, 2004) web crawler, 2) automated screenshots of URLs as rendered by a web browser via the PhamtomJs CLoud service (https://phantomjscloud.com/), and 3) extraction of content contained in web pages using the the Phantom Js Cloud API. Heritrix is the web archiving engine developed by the Internet Archive to create archival-quality crawls of web sites based on the standards from the International Internet Preservation Consortium (IIPC). Since some dynamic and flash content is difficult for the heritrix engine to archive, automated screenshots of each selected URL were also taken. In addition, the main content of each URL, or the readable text on the page, was extracted using the Phantom Js Cloud API. Phantom Js Cloud is a commercial cloud-based service which renders screenshots of web pages — producing an image of what a web page looks like in desktop web browser. Once all randomly selected URLs were archived using the processes outlined above, the tweet and its associated metadata was stored for later analysis. Archiving URLs at the time a tweet was posted allowed for gathering of a baseline of what the URL looked liked at the time the tweet was produced. Weekly archiving of selected URLs continued for a period of two months after real-time data collection ended in order to track changes in linked content over time. This is illustrated by section above the blue “two week” bar in Figures 4.1 and 4.2.

In summary, the data collection procedures doing real-time data collection were as follows:

1. A script submitted the query keywords, described in Appendix B, to the Twitter Streaming API for each case study and maintained an open connection.

2. As tweets matching the query terms are received via the Twitter Streaming API, each tweet was stored in a text file for later analysis.

3. The metadata of each tweet was examined for URLs (entities.urls) and embedded media (entities.media).

4. Upon encountering a URL, the script randomly selected the tweet and its accompanying URLs for archiving. URLs and media within selected for archiving were be submitted to the heritrix crawler for immediate archiving and a rendering of the URL in a web browser will also preserved via the Phantom Js Cloud service.

5. URLs selected for archiving were re-archived on a weekly basis for a period of two months.

4.4.2 *Nightly Availability*

From the start of real-time data collection to 90 days after each tweet was collected, the Twitter REST API was queried for the status and current version of each tweet collected. Each tweet and its associated metadata were stored for later analysis.

4.4.3 *Semi-Real-Time Collection*

At the conclusion of two-weeks of real-time data collection, tweets were again collected for each case study using the Twitter REST API. A script connected to the Search API and poll (repeatedly ask) for all tweets with the keywords/accounts from the two-week time period of real-time data collection. Each tweet and its associated metadata were be stored for later analysis. This is illustrated by the data collection steps

between the blue “two week” and yellow “two month” bars (semi-real-time) in Figure 4.2.

4.4.4 *Summary of Data Collection*

At the conclusion of data collection, multiple datasets for the Departments of Transportation and RuPaul’s Drag Race Case studies were collected including:

- Tweets and metadata collected in real-time using the Twitter Streaming API for a period of two weeks (14 days).
- Tweets and metadata collected in semi-real-time using the Twitter Search API two weeks after real-time collection.
- Tweets collected nightly by requesting each tweet via its unique id via the Twitter REST API for 90 days after real-time data collection.
- Randomly selected URLs archived in real-time and on a weekly basis for two months.

For the Occupy Wall Street case study, preexisting data from an earlier study was used:

- Tweets and metadata collected in real-time from the Twitter streaming API for a period of 12 days.
- Historical purchase of tweets from GNIP three years (June, 2014) after real-time data collection.

Together these datasets allow for the examination of the reliability and authenticity of the social media datasets contained within the three case studies.

4.5 Case Study and Data Description

The data used in this dissertation is derived from three Twitter-based case studies. A case study approach was chosen to closely replicate the bounded, event/phenomena focus of the social science approach to research allowing the findings to apply to a wider range of research designs from a social science point of view. Case studies also work well when a "how" question is being asked about a set of contemporary events (Yin, 2014) — which in my case included a recent world-wide social movement, interactions between the public and West Coast Departments of Transportation, and a reality TV show. This approach also preserves the connection between the phenomenon and its context (Yin, 2014) while retaining the capacity to address a case study's complexity (Simons, 2006). Each case study was chosen because it represents prototypical features of the types of data collection scenarios researchers experience. Examples of these dimensions include time-scale (short to long), population bounding (tight to lose), level of political contention (highly contentious to the everyday political context), and inclusion of links to media such as images and videos (high to low level of media and linking).

The range of case studies provide for a triangulation of different contexts — social movements, daily interactions with government, and a reality TV show — to examine the ephemerality of social media data within, between, and across cases. Different levels of case study analysis provide different types of insights — within case analysis informs researches in situations where their case studies/data match one or more of the prototypical dimensions of one of my case studies. Between-case analysis provides information about the impact of different prototypical dimensions on the ephemerality of

social media data sets. Across case analysis allowed me to generalize across twitter and to other social media platforms. For example, due to its nature as a social movement and the use of deletion as a protest tactic (Neumayer & Stald, 2014), the Occupy Wall Street case may exhibit a different level of ephemerality than the other cases. As a result, the Occupy Wall Street case acts as a model for researchers working with politically contentious social media datasets, but may not be a good model for researchers working with non-political data. The other case studies counter this, for example the inclusion of popular culture and entertainment though the Drag Race case study, acts as a model for a variety of events/phenomena social sciences researchers encounter when working with social media data. Combined the three case studies allow for a more general understanding of ephemerality across the Twitter and to other social media platforms.

A description of each case study as well as the associated data collection procedures and proposed analyses are listed below. A list of query terms for each case study are listed in Appendix B.

4.5.1 *Occupy Wall Street - Topic Based Dataset*

On June 2nd, 2011, Adbusters proposed a peaceful demonstration, “Occupy Wall Street”, to take place on September 17th to demand a separation of money from politics. Over the course of the next three months, face-to-face working groups met in NYC to create a General Assembly to coordinate and organize action. Around the same time, in August 2011, a Tumblr site called “We are the 99%” was created in which individuals were able to upload their personal narratives as they related to the actions and message of Occupy Wall Street. The Tumblr site provided an opportunity for geographically

distant supporters to participate and connect with the localized efforts in NYC. On September 17th, roughly 1,000 protesters marched on Wall Street and set up a camp in Zuccotti Park, “occupying” the space to demonstrate their dissent.

Over the course of a few weeks, the Occupy Wall Street demonstrations grew into the global Occupy movement, as camps where set up around the world including in the United States, United Kingdom, Japan, Italy, Canada, and Mexico in solidarity with the philosophy of the movement. This digitally enabled action network closely mirrored the indignados movement of Spain, in that established political organizations so common to traditional social movements, such as unions and political parties were replaced by technology platforms and applications (Bennett & Segerberg, 2012). Facebook pages for Occupy city camps sprung up in early October, accumulating tens of thousands of likes in major cities such as Philadelphia and Chicago (Caren & Gaby, 2012). Twitter handles and hashtags such as #occupywallst, #ows, and #occupy emerged to facilitate the coordination and exchange of relevant information. A battery of local city focused websites also grew in conjunction with umbrella websites that organized cross-city coordination and information. As the movement grew and matured new tools and platforms were incorporated into the local and national level of the Occupy information ecosystem. Livestream, a tool for broadcasting live events to the web, provided an opportunity for those not on the ground to bear witness and connect to the movement by watching real-time events such as protest marches, General Assembly meetings, speeches, arrests, and evictions. Protestors used photo sharing tools such as twitpics and

yfrog (both now defunct)[17] to share pictures and document events unfolding on the ground. While on-the-ground efforts and participation were critical to sustaining and growing the movement; digital tools and technologies in these networked organizations acted as communication infrastructure, providing channels to share information, organize events, coordinate activity, and connecting participants and camps to one another (Agarwal et al., 2014).

The real-time #OWS Twitter case consists of 64,298,104 tweets collected from October 19, 2011 - June 9, 2011. Tweets were collected using Twitter's Streaming API, which returns tweets matching any of the search keywords occurring in the text, hashtags, @mentions, or URLs within a tweet. A panel of faculty and graduate students curated a list of popular hashtags, keywords, and Occupy city accounts related to the Occupy movement. The resulting data stream was examined at regular intervals for emerging terms. New terms were added to the keyword list after being reviewed by the entire research team, resulting in a dynamic archive based on a list of 355 keywords as data collection continued through the summer of 2012.

A companion historical #OWS Twitter dataset collected from GNIP consists of tweets from October 17, 2011 - October 31, 2011. Clemson University collected this dataset using the GNIP PowerTrack Historical Search API, which returns non-deleted tweets matching any of the search keywords occurring in the text, hashtags, @mentions, or

[17] See https://blog.twitpic.com/2014/10/twitpics-future/, archived at https://perma.cc/WR7V-QB3V; https://en.wikipedia.org/wiki/Yfrog.

URLs within a tweet. The initial list of 205 keyword terms[18] from the contemporaneous data collection was used to collect this data in June of 2014.

4.5.2 *West Coast Departments of Transportation - Account Based*

This case focuses on Department of Transportation (DoT) accounts on the West Coast of the United States including WA, OR, and CA. It represents an everyday form of every day political talk (J. Kim & Kim, 2008) about traffic and transportation issues with government (S. Zhang, 2015); less contentious than a social movement but still involving communication with government. Unlike the other two cases, this case does not consist of a set of keywords, but a list of accounts. Tweets collected consist of tweets produced by each account and tweets from users interacting with these accounts (replies, mentions, and retweets).

This case study is prototypical of research projects with a well-defined population of users, high metadata stability, and high link stability. I posit this level of account metadata and link stability since government accounts properties such as profile text and usernames are unlikely to change and the majority of links tweeted by these accounts point to public facing .gov sites.

This case contains only official staffed and automated accounts used by the Departments of Transportation in Washington, Oregon, and California. Twitter accounts were gathered from public facing websites of the Departments of Transportation and will

[18] A list of keywords used to collect data can be found at https://github.com/somelab/SoMeToolkit/blob/master/collection.terms, archived at https://perma.cc/7E58-PWGB.

include the primary account for the state DoT, regional DoT accounts, and automated traffic bots. Account profiles were reviewed to ensure that the account was active at the time of data collection and related to the state DoT. Unofficial, county, and city Department of Transportation (non-state) accounts were excluded from the list. Accounts without any activity in the last year were also excluded from the list.

Data was collected in real-time for two weeks from September 19, 2016 - October 2, 2016 via the Twitter Streaming API using the list of list of accounts and hashtags as generated above. During real-time collection, tweets with URLs and media (embedded images and video) were randomly selected for archiving as the tweet was received from the API. Embedded content for the selected tweets was archived within minutes of receipt of the tweet via the Twitter Streaming API and regular at weekly intervals for two months — for a possible total of 9 archives, 1 real-time with 8 weekly additional archives. Upon conclusion of the two-week real-time data collection period, the Twitter REST API was queried using the account and hashtag list. The Twitter REST API was queried nightly for 90 days after each tweet was collected to determine the accessibility of each tweet and any changes to its associated metadata.

4.5.3 *RuPaul's Drag Race - Mixed Account/Topic Based Dataset*

RuPaul's Drag Race is a reality competition television show in which a group of 12 drag queen constants seek the title of "America's next drag superstar"[19] with a grand price of $100,000. The show, currently in its 8th season, is the highest-rated television on

[19] http://www.logotv.com/shows/rupauls-drag-race/cast

its parent network, Logo[20], also airing in Australia, Canada, and the United Kingdom. A panel of regular and guest judges, led by RuPaul, critique constants as they progress through a series of weekly challenges. Each show concludes with the top two constants competing in a "lip sync for your legacy" event to win the weekly challenge and to select a competitor for elimination. The winner of the lip sync competition selects a competitor for elimination from the bottom two contestants as selected by RuPaul. The contestant selected for elimination "sashays away", while the competitor not selected for elimination stays in the contest for another week.

This case study is prototypical of research projects with a semi-defined population of users bounded by a set of twitter accounts and hashtags, high level of media (image and video) usage, or a focus on an entertainment context.

The query terms for this case contain a combination of Twitter accounts and hashtags related to the show including the Twitter accounts of judges, constants, and guests appearing in the episodes during data collection. The list of judges, constants, and guests was gathered from the show's website and Twitter account handles were obtained via Google and Twitter searches. Hashtags related to the show were after observing a sample of tweets from the official Drag Race Twitter account, @RuPaulsDragRace. A full list of account, hashtags, and keywords are listed in Appendix B.

Data was collected in real-time for two weeks from September 19, 2016 - October 2, 2016 via the Twitter Streaming API using the list of list of accounts and hashtags as generated above. During real-time collection, tweets with URLs and media (embedded

[20] http://www.etonline.com/tv/160480_for_rupauls_drag_race_mainstream_is_jumping_the_shark/

images and video) were randomly selected for archiving as the tweet was received from the API. Embedded content for the selected tweets was archived within minutes of receipt of the tweet via the Twitter Streaming API and regular at weekly intervals for two months — for a possible total of 9 archives, 1 real-time with 8 weekly additional archives. Upon conclusion of the two-week real-time data collection period, the Twitter REST API was queried using the account and hashtag list. The Twitter REST API was queried nightly for 90 days after each tweet was collected to determine the accessibility of each tweet and any changes to its associated metadata.

4.6 Summary of Case Study Data Collection and Analysis

Because data was collected for the Occupy Wall Street case in prior work, only a subset of analyses can be applied to that case study. Specifically, archiving URLs or the collection nightly availability of tweets was not part of the preexisting dataset. All data collection procedures and analyses were applied to the other two case studies. A short summary and table are below.

Table 4.1: Case Collection and Analysis Summary

Dataset	**Streaming API**	**GNIP API**	**REST API**	**URL Archiving**	**Prototypical Dimensions**	**Duration of Analysis**
1. OWS	X	X			multi-year time scale, contentious political context, keyword based query, high account and metadata instability, unbounded population	3 years
2. DoT	X	X	X	X	bounded population, account based query, high metadata and account stability, high URL stability, every-day political context	2 Months
3. Drag Race	X	X	X	X	mixed keyword and account query, entertainment context, media intensive, semi-bounded population	2 Months

4.6.1 *Analysis of Occupy Wall Street Case Study*

A subset of both datasets with the same timeframe and hashtags was compared. Tweets ids and metadata from October 19, 2011 to October 31, 2011 with the hashtags #ows and #occupy were compared from the real-time and historical datasets. Since this is a preexisting dataset, URLs were not archived during real-time data collection so the analysis of the URLs cannot be performed in this case study.

The comparison of these Occupy Wall Street datasets provides insight into the ephemerality of social media data from a social movement over a three-year timeframe — from its (near) inception to three years later. This case study is prototypical due to the timespan between real-time and historical collection (three years) and its semi-

contentious political nature; thus providing insight into the ephemerality of situations when researchers collect social media data about a historical political event multiple years after it occurred.

Tweets missing from either set were further examined using a process similar to the one developed by Petrovic et. al (2013). Missing tweets were requested via the Twitter REST API — to determine if the tweet was deleted, modified, still available, or the user account was deleted. Since URLs were not captured in real-time during data collection, it was not possible to track the change of URLs over time. It would be possible to do an automated analysis of HTTP codes or code a random sample to get the current state of URLs, but this would provide little insight into the change over time. Also, the protest nature of this dataset may lead to different patterns of ephemerality vs other less political datasets.

4.6.2 *Analysis of DoT and Drag Race Case Studies*

Since the DoT and Drag Race cases were specifically collected for this dissection, all of the data collection and analysis procedures outline in the previous chapter have been conducted on the case studies. This includes a comparison of real-time (Streaming API), semi-real-time (Search API), and nightly availability check (REST) datasets within each case study.

4.6.3 *Case Descriptive Statistics*

Data for each case study was collected in real-time for a period of 12 - 14 days via the Twitter Streaming API using a set of query terms including keywords and/or account

accounts as described in the previous section. The set of query terms for each case study are listed in Appendix A. Figure 4.3 shows a graph of the number of tweets collected each day during real-time data collection.

Due to the high-volume of tweets collected in the OWS case study, the number of tweets exceeded the rate limits of the Twitter Streaming API. When rate-limiting occurs, the API reports an estimated number of rate-limited tweets since the connection to the API was opened. This number is reset with reconnecting to the API. The estimated number of rate-limited tweets reported by the Twitter Streaming API for this case was 25,052. As seen in Figure 4.3, no tweets were collected on October 29, 2011 due to API connection issues.

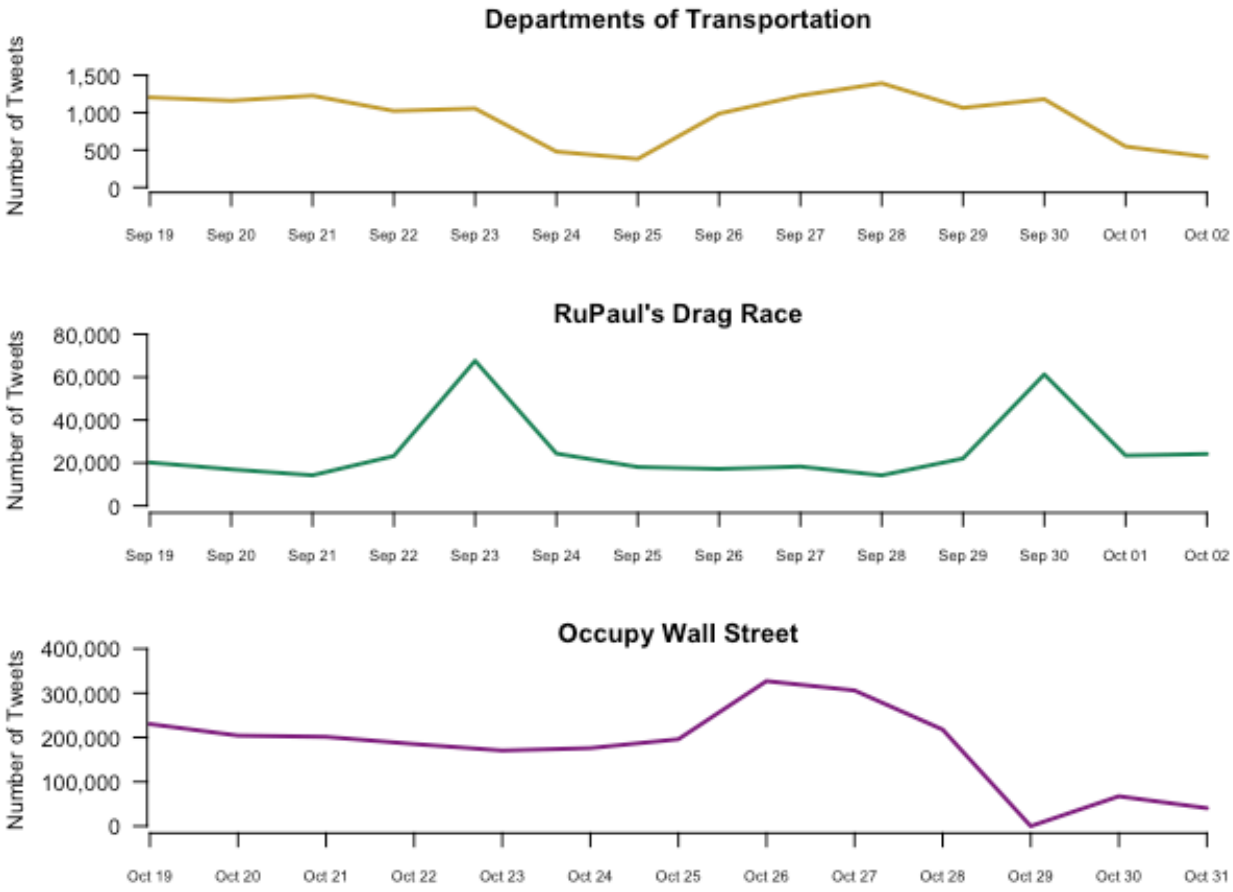

Figure 4.3: Daily tweet volume collected for each case study during real-time collection. Timestamps are in UTC.

Table 4.2: Summary of Case Study Descriptive Statistics

Case	Start Date	Duration	Unique Users	Tweets	Retweets	Replies
Occupy Wall Street	10/19/11	12	536,912	2,310,038	972,208 (42.1%)	142,452 (6.2%)
Departments of Transportation	9/19/16	14	3,464	13,330	5,567 (41.8%)	2,032 (15.2%)
Drag Race	9/19/16	14	106,602	356,147	179,346 (50.3%)	38,477 (10.8%)

Table 4.2 displays basic descriptive statistics for each case study. The graph in Figure 4.3 shows the daily tweet volume collected for each of the case studies. The Occupy Wall

Street case contains the most tweets due the collecting period taking place as the movement was ramping up during a high period of new media coverage. The Departments of Transportation case contains the least number of tweets of the three cases since the majority of the accounts in the dataset were information accounts automatically tweeting information related to traffic conditions. The lower volume of tweets is also due to the query parameters used — tweets in the case are either from the set of accounts mentioned in Appendix B or from an account retweeting or replying to one of those accounts.

Table 4.3: Proportion of tweets with entities: hashtags, URLs, and mentions in each case study.

Case	**With hashtags**	**With URLs**	**With mentions**
Occupy Wall Street	65.1%	52.5%	65.4%
Departments of Transportation	24.7%	27.3%	65.1%
Drag Race	35.4%	21.6%	78.5%

Table 4.3 lists the proportion of tweets containing hashtags, URLs, and mentions of other users for each case study. The RuPaul's Drag Race case study contains the highest number of mentions (78.5%) as Twitter users watching the show mentioned the Twitter handles of judges and contestant in their tweets. The OWS (65.4%) and Departments of Transportation (65.1%) case studies contain a similar percentage of mentions. The percentages of hashtags and mentions are highly impacted by the construction of case keywords — for example, the lower number of tweets with hashtags in the Departments of Transportation case study (24.7%) may be due to the fact that the query terms for the

case include only accounts and not keywords or hashtags. The high number of tweets with URLs in the Occupy Wall Street case study (52.5%) may be due to the nature of a social movement as protesters tweet out different types of informational resources in responses to external events (Bennett et al., 2014; Segerberg & Bennett, 2011).

Table 4.4: Summary of Case Study Descriptive Statistics - User Statistics

Case	**Mean tweets/user (SD)**	**Max tweets/user**	**Min tweets/user**
Occupy Wall Street	4.3 (28.3)	7,364	1
Departments of Transportation	3.8 (32.5)	1,214	1
Drag Race	3.4 (27)	8,250	1

Table 4.4 lists descriptive statistics related to user accounts in each of the case studies. Mean tweets per user range for each of the case studies between 4.3 to 3.8 with the majority of users having only one tweet. As shown by the high standard deviation and high max tweets per user value, users central to the case often tweeted out more tweets with the majority of users entering the dataset with only 1 tweet.

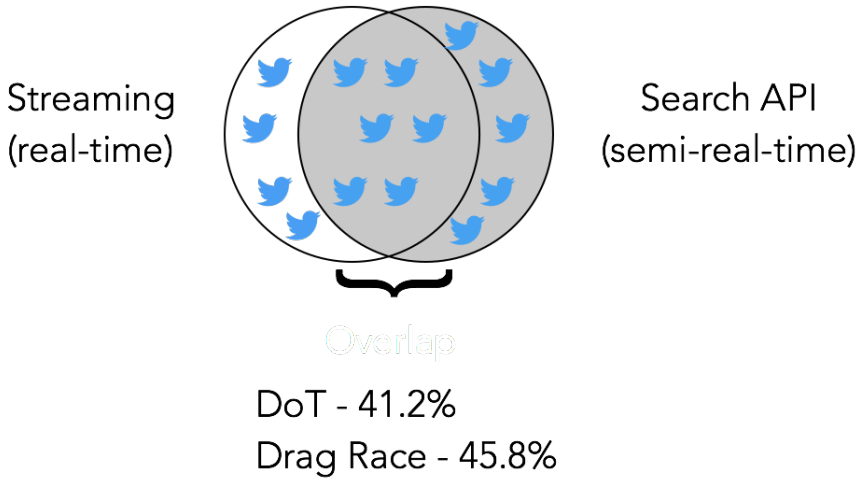

Table 4.5: Visualization of overlap between data collected real-time (Twitter Streaming API) and semi-real-time (Twitter REST API).

Case	Streaming	Search	Overlap	Difference
Departments of Transportation	13,330	6,457	5,490 (41.2%)	7,840 (58.9%)
Drag Race	356,147	179,131	163,220 (45.8%)	192,927 (54.2%)

Figure 4.4: Visualization of overlap between data collected real-time (Twitter Streaming API) and semi-real-time (Twitter REST API).

The Twitter Streaming API only supports real-time data collection so any delay in data collection forces researchers to use only data collection points such as scraping the Twitter website, purchasing data, or using the Twitter REST API. The Twitter REST API supports the collection of tweets via a search interface[21] against a sampling of tweets posted in the last 7 days. The documentation states that the Search API focuses on "relevance and not completeness", noting that some tweets and users may be missing from search results from the Search API. The documentation points developers and researchers to the Streaming API or GNIP for more complete datasets. The lack of "completeness" and short result window (7-days) make the Search API problematic for research data collection (Driscoll & Walker, 2014; González-Bailón et al., 2014).

Figure 4.4 and Table 4.5 compares tweets collected in real-time via the Twitter Streaming API and in semi-real-time (Twitter REST API) for the Departments of Transportation and RuPaul's Drag Race case studies. The same query terms were used

[21] See https://dev.twitter.com/rest/public/search for more information about the Twitter Search API.

for the Streaming API and Search API as listed in Appendix A. Search API data was collected for each study after real-time data collection had concluded, or 15 days after data collection began. Tweet ids were matched between the Streaming API and Search API datasets as listed in the overlap column in Table 4.5. The difference column lists the number of tweets missing from the Search API dataset. For both case studies, approximately 40% - 45% of tweets were accessible via both the Streaming and Search APIs.

4.7 Chapter Summary

In this chapter I introduced the concept of ephemerality, or unstable nature, of social media data over time. This concept is important for researchers working with social media data since any latency, or delay, in data collection may lead to changes in the resulting dataset. These changes may be caused by posts becoming inaccessible, changes to a user's profiles, or changes to linked content embedded in posts and user profiles. Depending on the scope of the changes and their interaction with the research design, the content of the social media dataset my no longer accurately reflect the phenomena under investigation.

This chapter also described my research design, data collection procedures, and the case studies under investigation. Each case study was chosen because it represents prototypical features of the types of data collection scenarios researchers experience. Examples of these dimensions include time-scale (short to long), population bounding (tight to lose), level of political contention (highly contentious to the everyday political

context), and inclusion of links to media such as images and videos (high to low level of media and linking). The range of case studies provide for a triangulation of different contexts — social movements, daily interactions with government, and a reality TV show — to examine the ephemerality of social media data within, between, and across cases. Descriptive statistics were provided for each case study.

Finally, the overlap between real-time (Streaming API) and semi-real-time was calculated for the Departments of Transportation and RuPaul's Drag Race case studies. For both case studies, approximately 40% - 45% of tweets were accessible via both the Streaming and Search APIs showing that the Search API is a poor choice for data collection when a research design requires the collection of all tweets related to a set of keywords.

Chapter 5. RELIABILITY

The first concept closely connected with ephemerality is reliability. From a statistical standpoint, reliability is concerned with consistency in obtaining the same measurement or finding upon repeated measurements under similar conditions with a research instrument. Within the context of social media, high reliability would imply capturing a social media dataset in such a way that the number of posts in the dataset is stable and each post, after collection, is the same post each time it is accessed. In other words, a social media dataset that is observed with a particular measurement device is impervious to change — deletion, modification, or change in privacy settings (public/private). I measure the reliability of a social media dataset through the level of change in tweet identification numbers (appearance/disappearance) in the dataset over time when the same parameters are used to collect the data at different points in time (real-time, semi-real-time, and historical). To measure the level of change, I compare the unique tweet ids, identifiers which uniquely identify the individual tweets over time, between these datasets.

5.1 Operationalizing Reliability

To examine the reliability of the social media datasets in each case study, I tracked the availability or unavailability of each individual tweet throughout the three periods of data collection (see Chapter 4). Since Twitter does not offer users the ability to modify a tweet posting, this analysis only focuses on the accessibility or inaccessibility of a tweet. This approach is similar to the one used by Driscoll and Walker (2014) employed to

compare tweets collected via different real-time APIs. The availability of each tweet was determined by matching unique tweet ids from the real-time and nightly datasets. Tweets found in the real-time dataset but not in the nightly datasets were further analyzed to determine the cause of its inaccessibility. The Twitter API was queried to determine if the individual tweet was deleted, the account was deleted, or the account was made private. This is similar to the method used by Petrović , Osborne, and Lavrenko (2013). If the inaccessible tweet was a retweet, the original tweet was also queried via the Twitter REST API to determine if the inaccessibility is a result of a deletion cascade. In a deletion cascade, the deletion of a retweet in another user's timeline is not related to the timeline owner's actions, but a propagation of the deletion of a tweet, account deletion/suspension, or privacy change by the original producer of the retweet; thus, without classification, magnifies the impact of a single deletion by the number of times the tweet was retweeted. This analysis, along with basic descriptive statistics, provides a picture of the reliability within each case study as the availability of each tweet collected in real-time is monitored for a period of 90 days after collection. This approach shows data a researcher would be able to collect as the latency of data collection increases from 1 to 90 days.

5.2 Reliability Analysis of Each Case Study

The accessibility of each tweet collected in real-time was checked each day for a period of 90 days after collection. Each night a script requested every tweet collected during the real-time data collection period for the Departments of Transportation and

RuPaul's Drag Race case studies by requesting each tweet from the Twitter REST API via its unique id. After the 90-day data collection period, the data was processed using the following process:

- Since tweets were collected in real-time over a two-week period, time periods were calculated for each tweet. The date of collection was represented as time point t0, the 1st day as time point t1, and the 90th day as time point t90. This transformation allows the accessibility of each tweet to be compared while taking the two-week collection window into account.
- For each time point (t1 - t90), the accessibility of each tweet was determined by noting if each individual tweet id appeared in the nightly data corresponding the time point for that tweet.
- Any gaps in the accessibility of a tweet were backfilled and the missing time point was assumed to be due to an error in the data returned from the Twitter API. For example, if a tweet was accessible at time points t9 and t12 but not at time points t10 and t11, time points t10 and t11 were re-coded as accessible. Backfilling missing time points masks changes users made to their private settings, moving their account from protected to not protected. More granular data on the reason for the inaccessibility of each tweet would be required in order to detect daily changes in account protection settings.

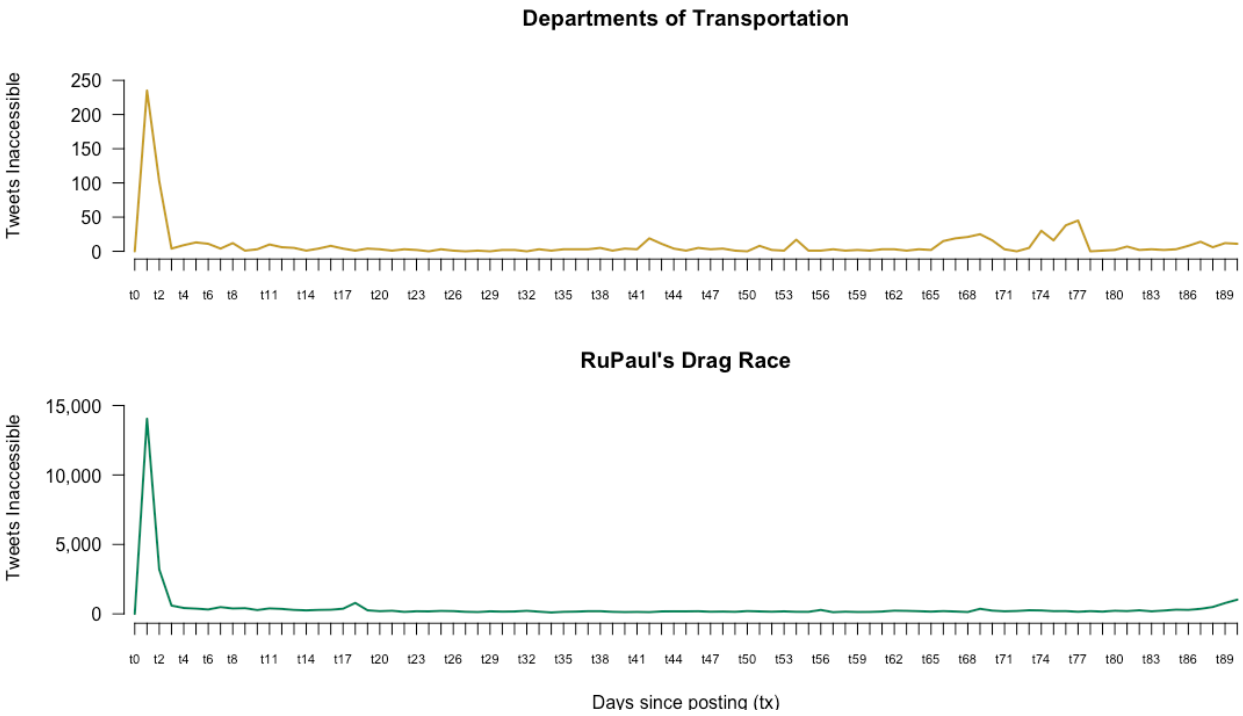

Figure 5.1: Tweets inaccessible per time period during the 90-day observation period for the Departments of Transportation and RuPaul's Drag Race case studies.

Table 5.1: Proportion Tweets Accessible After Time Periods Under Investigation.

Case	**Time Period**	**Total Tweets Collected in Real-Time**	**Tweets Accessible**	**Tweets Inaccessible**
Occupy Wall Street	3 years	2,310,038	2,029,074	280,964 (12.7%)
Departments of Transportation	90 days	13,330	12,448 (93.4%)	882 (6.7%)
Drag Race	90 days	356,147	276,204 (89.3%)	38,943 (10.7%)

The proportion of tweets available 90 days after real-time collection for the Departments of Transportation and RuPaul's Drag Race case studies are shown in Table 5.1. After 90 days, 6.7% of tweets in collected in the Departments of Transportation case study were unavailable. For the RuPaul's Drag Race case study, 10.7% of tweets were inaccessible after 90 days.

For the Occupy Wall Street case study a different process was used to assess the reliability of the dataset. Tweets were collected at only two points in time: (1) real-time and (2) purchased from GNIP three years later in June 2014, requiring a different approach for comparison. The same process was used as noted above, but only two time points were compared (see Chapter 4 for a data collection methods). After three years, 280,964 (12.7%) tweets were in no longer accessible.

Table 5.1 confirms the expected ranking of cases with respect to the proportion of inaccessible tweets when considering the prototypical features of each case. When constructing the prototypical features of the case studies, it was expected that the Occupy Wall Street case study would exhibit the highest level of inaccessible tweets (12.7%) due to the highly contentious nature of the social movement, the query terms used to collect data were solely keyword based (no accounts were followed), and the three-year timeframe. RuPaul's Drag race was expected to have the second-highest level of inaccessible tweets (10.7%) due to its reality TV and entertainment context. The Departments of Transportation case study was expected to have the lowest proportion of inaccessible tweets (6.7%) due to the "everyday political context" and its focus on tweets from or replying to official government accounts.

5.3 Mechanisms of Inaccessibility

Tweets become inaccessible due to a variety of mechanisms[22] related to the tweet itself as well as the tweets and accounts a tweet is related to. Each tweet inherits the

[22] See https://support.twitter.com/articles/18906?lang=en for a description of what happens when a tweet is deleted. Link archived at https://perma.cc/38XY-RQFR.

accessibility properties of the Twitter account that created the tweet. As illustrated in Figure 5.2, when a user deletes their account, this action cascades deleting all tweets contained in the account. This same process takes place with retweets, but a retweet is also impact by the availability of tweet the original tweet that was retweeted.

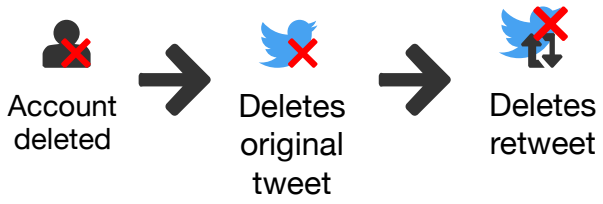

Figure 5.2: Illustration of how a tweet inherits the accessibility properties of the tweets it is related to. In example shown, a retweet is deleted because the account that produced the original retweet was deleted.

The mechanisms tweets become inaccessible by actions of the user/account that produced the tweet. The following actions impact the accessibility of a tweet:

1. Tweet deletion. A user deletes a tweet from their timeline.
2. User account deletion. A user deletes their account thereby removing all their tweets.
3. Account set to protected. Tweets in an account are set to 'protected' thereby effectively making all tweets by that user inaccessible.

If the tweet is a retweet, the retweet also inherits the accessibility properties of tweet that was retweeted. The following actions impact the accessibility of a retweet:

1. Deletion of an original tweet. If this tweet happens to be a retweet, if the retweeted tweet (original tweet) is deleted, the deletion of original tweet cascades

to all retweets of the original tweet. This is illustrated in figure 5.2, starting at the middle, “deletes original tweet” stage.

2. Deletion of the account of the user who created the original tweet. A user whose tweet was retweet, deletes their account thereby removing all of the tweets in their account trigging the previous step (#1).
3. Protection of the account of the user who created the original tweet. A user whose tweet was retweet, protects their account thereby removing all of the tweets in their account trigging the previous step (#1). The difference between protection and deletion of an account is that all tweets become publicly accessible again if the protection setting is turned off.

In order to determine the reason a tweet became unavailable, a list of tweets inaccessible 90 days after collection were generated for the Departments of Transportation and RuPaul’s Drag Race case studies. It was not possible to complete this analysis on the Occupy Wall Street case study since nightly data was not collected for the case study. Further data was collected on each inaccessible tweet by querying the Twitter REST API to determine the accessibility of the account that produced the tweet and its public or protected status. If the tweet was a retweet, the status of the original tweet and the account that produced the original tweet was also queried.

Using the 6 actions (deleted tweet, deleted user, protected user, deleted retweet, deleted retweet user, and protected retweet user) that can cause a tweet to become in accessible, each inaccessible tweet was analyzed using the data retrieved from the Twitter REST API. In cases where there could be multiple causes for a tweet to be

inaccessible, the reason for inaccessibility was attributed to the highest cause. For example, if a tweet was deleted and the account was also deleted, then the cause would be attributed to the deletion of the account. If the tweet was a retweet, and the account of the user who retweeted the tweet was deleted and the original tweet was also deleted, the cause of the inaccessibility of the tweet is attributed to the deletion of the original tweet. This process will not detect instances when a retweet was deleted before original tweet was deleted during the 90-day window of data collection.

Table 5.2: Reason for tweet inaccessibility - Departments of Transportation and RuPaul's Drag Race

Case	**Total Inaccessible Tweets**	**Deleted Tweet**	**Deleted User**	**Protected User**	**Deleted Retweet**	**Deleted Retweet User**	**Protected Retweet User**
Departments of Transportation	882	475 (53.8%)	326 (37%)	21 (2.4%)	26 (2.9%)	33 (3.7%)	1 (0.1%)
Drag Race	38,943	20,811 (53.4%)	10,536 (27%)	3,816 (9.8%)	2,280 (5.9%)	980 (2.5%)	520 (1.3%)

In both cases, the main cause (53%) of tweet inaccessibility is the deletion of the tweet itself. This indicates that the majority of inaccessible tweets were deleted by users themselves. The second highest cause of tweet inaccessibility is due to the deletion of user's account. It is important to note that in both cases, the accounts deleted represented transient users who either mentioned the accounts included in each case study or the, in the case of RuPaul's Drag Race, the keywords contained in the query parameters. None of the accounts contained in the query parameters were deleted during data collection.

Table 5.3: Tweet inaccessibility categorized by changes to user account vs. a retweeted account.

Case	**Total Inaccessible Tweets**	**Inaccessibility due to:**	
		Changes in user account	**Changes to a retweeted account**
Departments of Transportation	882	823 (93.3%)	59 (6.7%)
Drag Race	38,943	35,163 (90.3%)	3,780 (9.7%)

It is helpful to think of the root of the case a tweet's inaccessibility: is it due to changes made by a user to their own account or changes to other accounts. This collapses the reasons for inaccessibility into two categories: (1) changes to the user account (deleted tweet, deleted user, and protected user) into one category and (2) changes to the account related to an original tweet (deleted tweet, deleted retweet user, and protected retweet user). The 6 categories from Table 5.2 have been collapsed into these two categories in Table 5.3. In both cases, the majority (over 90%) of tweets were inaccessible due to changes to the account of the user who tweeted the tweet.

5.4 Chapter Summary

In this chapter I introduced the first concept closely connected with ephemerality: reliability. From a statistical standpoint, reliability is concerned with consistency in obtaining the same measurement or finding upon repeated measurements under similar conditions with a research instrument. Within the context of social media, high reliability would imply capturing a social media dataset in such a way that the number of posts in the dataset is stable and each post, after collection, is the same post each time it is

accessed. In other words, a social media dataset that is observed with a particular measurement device is impervious to change — deletion, modification, or change in privacy settings (public/private). I measured the reliability of a social media dataset through the level of change in tweet identification numbers (appearance/disappearance) in the dataset over time when the same parameters are used to collect the data at different points in time (real-time, semi-real-time, and historical comparing the unique tweet ids between these datasets.

At the end of the time period under observation, 6.7% - 10.7% of tweets were no longer accessible. The cases followed the expected ranking with respect to the percentage of tweets that were inaccessible – 1) Occupy Wall Street, 2) Rupaul's Drag Race, and 3) Departments of Transportation.

Two of the case studies, Departments of Transportation and RuPaul's Drag Race, were examined to determine the mechanism that caused each tweet to become inaccessible. In both cases, the main cause (53%) of tweet inaccessibility is the deletion of the tweet itself. This indicates that the majority of inaccessible tweets were deleted by users themselves.

Chapter 6. AUTHENTICITY

The second concept closely connected with ephemerality is authenticity. Authenticity of a dataset is "the extent to which it accurately (precision) and faithfully (fidelity) represents what it is meant to. Establishing and documenting data quality and veracity is a key aspect of data lineage" (Kitchin, 2014, p. 153). Similarly, from an archival point-of-view, it is the authority and trustworthiness of a record as proof and memory of the activity of which they constitute a natural byproduct (as social media trace data is a "natural" byproduct of a post). In other words, a "record that can stand for the facts it is about" (Duranti, 1995). This is "linked to a record's state, mode and form of transmission, and to the manner of its preservation and custody" (Duranti et al., 2013 Ch. 2). Within the context of social media, the authenticity of a post involves capturing and stabilizing the surrounding metadata so a post or digital object can stand on its own. I measure the authenticity of a social media dataset through the level of change in metadata and linked URLs embedded in a post over time when the same parameters are used to collect the data at different points in time. To measure the level of change, I compared specific metadata elements (e.g. username, profile description, post statistics) and URLs of the same tweets collected at different points in time. Textual metadata elements such as username and profile description were compared at different collection times to quantify how the extent of change. For example, changing cats to cat requires one change — the removal of the 's'. The difference in numeric metadata elements was be calculated over time. Weekly URLs archives of URLs will be manually compared using

a qualitative coding process over time. Text content of URLs will also be extracted and compared using a method similar to the textual metadata elements.

6.1 Operationalizing Authenticity

The authenticity of a social media dataset is concerned with the stability of the metadata and linked URLs surrounding each post. This analysis focuses on the stability of the metadata embedded in each tweet and the URLs randomly selected during real-time data collection. For example, when requesting a tweet via one of the Twitter APIs or viewing a tweet via the Twitter website or mobile client, the current and not historical metadata, such as a user's profile description, is displayed. As a result, there could be a mismatch in the metadata between a tweet collected in real-time and one displayed or requested at a later point in time.

Figure 6.1: Tweet from the US National Park Service as display on the Twitter website in May 2017 with metadata fields labeled.

To examine this change, I compared tweet and the user profile metadata of the real-time and nightly data from the Departments of Transportation and RuPaul's Drag Race case studies. For reference, metadata fields analyzed are labeled in a tweet in Figure 6.1.

Since the screen name (Twitter handle), profile description, and location are editable by the user, these fields were compared using edit or Levenshtein distance (Levenshtein, 1966). The Levenshtein distance measures the number of insertions, deletions, and substitutions required to change one string into another. For example, a change from 'cat' to 'bat' has a Levenshtein distance of 1 since 1 edit is required to change the 'c' in 'cat' to a 'b'. This provides a measure of the extent of change, if any, between each metadata field in the real-time and nightly datasets within the two cases.

In addition, I also checked for changes in the profile picture URL and homepage URL listed on each user's profile by comparing values returned in each dataset. Tweet and user statistics were also compared over time, including the number of: followers, following, retweets, and favorited count.

It is important to note that the backfilling of missing days of data collection as described in section 5.2 may have an impact on the above measure of authenticity since backfilled days are treated as if there was no change in a user's profile or metadata. Any changes occurring on backfilled days will be detected during the next non-backfilled day thus shifting the recorded day of change by the number of backfilled days. In the case of indicators measuring the extent of change each day to the profile of a user (user description, location, and name), the extent of change may be overrepresented on days

occurring after backfilled days since that day is also recording changes made to a user's profile during those gaps.

To measure the stability of linked URLs, the web archives and extracted page content were compared across the highly weekly sampling points. The process involved comparing the baseline URLs archived at the time of tweeting to the weekly archived pages using web archives produced by the heritirx crawler. The extracted page content was compared using a similarity hash. This analysis provides a measure of the change of the content of URLs embedded in each tweet thus lending empirical data to the level of change of linked content embedded in social media datasets.

6.2 Authenticity Analysis of Tweet and User Metadata

Authenticity of the metadata associated with each tweet was analyzed at two levels: (1) the metadata associated with the tweet including the number of times a tweet was retweeted and favorited by other users and (2) the metadata associated with the user tweeting the tweet. As mentioned in the previous section, changes in tweet metadata were tracked for a period of 90-days following its collection. Since some users tweeted multiple tweets in each case study, user profile information was tracked for 90 days from the first time a user tweeted in each case. Nightly user profile information was extracted from the last tweet the user tweeted each day. As shown in Table 4.2, 3,464 users were tracked as part of the Departments of Transportation case study and 106,602 users were tracked as part of the RuPaul's Drag Race case study.

Table 6.1: Number and proportion of users changing profile metadata

Case	Unique Users	Screen Name	Name	Profile Image	Location	Description
Departments of Transportation	3,464	41 (1.2%)	325 (9.4%)	731 (21.1%)	173 (5%)	809 (23.3%)
Drag Race	106,602	5,978 (5.5%)	34,313 (31.5%)	57,846 (53.3%)	13,689 (12.6%)	45,763 (42.2%)

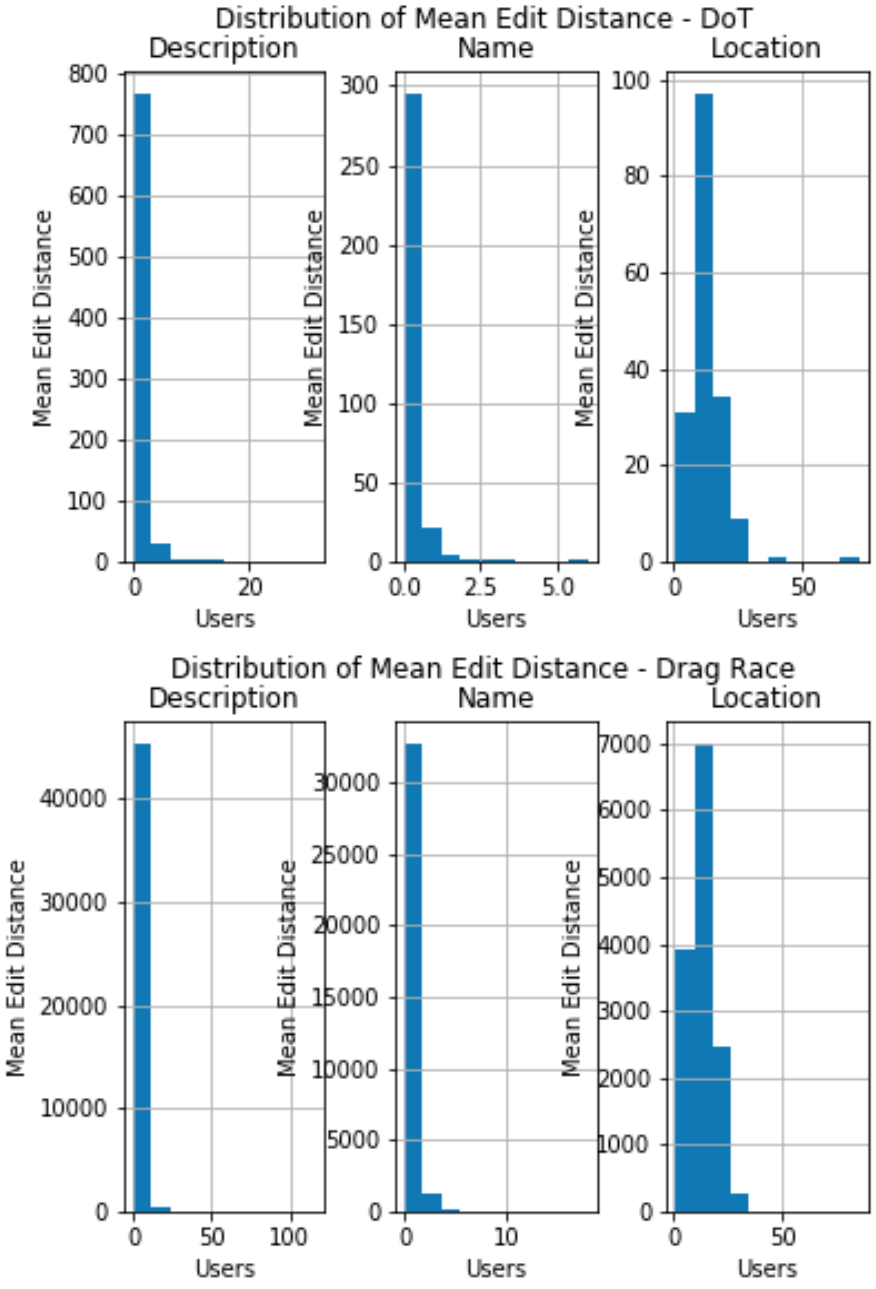

Figure 6.2: Distribution of mean edit distance of change to user profile metadata - user description, user name, and user location. Only users with an edit distance > 0 are displayed.

Table 6.2: Extent to which users changed profile metadata as measured by mean edit distance between changes.

Case	**Description**	**Name**	**Location**
Departments of Transportation	0.99 (1.96)	0.28 (7.28)	13.5 (0.48)
Drag Race	1.47 (2.57)	0.5 (0.79)	12.92 (6.1)

As shown in Table 6.1, the highest percentage of users in both case studies made changes to their profile image. In the RuPaul's Drag Race case study, the second most edited piece of user metadata was the description field (42.2%) with only 2.4% of users editing their profile description in the Departments of Transportation case study. This points to the impact of the construction of each case study and its prototypical features have on the level of authenticity. The Departments of Transportation has a lower level of user profile change due the more stable nature of official government accounts.

User level metrics such as the number of followers are often used as proxies of user centrality and popularity. The three user-level metrics were tracked across the 90-period included:

1. Followers Count: indicates the number of users following this user's posts, (2) friends count: indicates the number of users this user follows. This is often used as an indicator of a user's potential reach or the number of other Twitter users who have the potential to see a user's tweets in their timelines.

2. Friends Count: indicates the number of users that this user follows. While used less often than the followers count, this is normally used as an indicator of which users a user is connected to.
3. Statuses Count: indicates the number of non-deleted tweets a user has posted. This is often used as an indicator of the level of activity of a user. Note that deleting and posting tweets will decrease or increase this number.

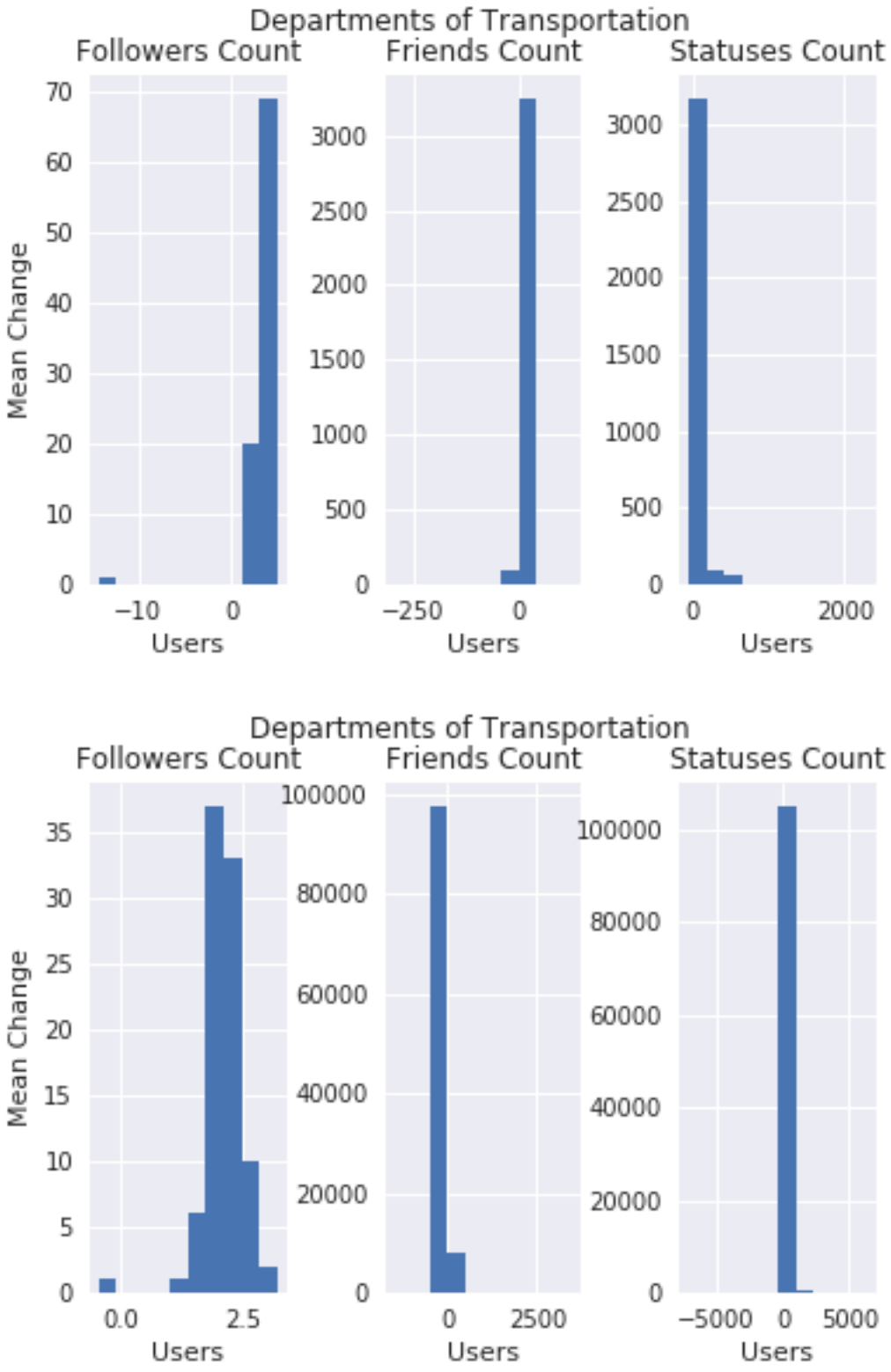

Figure 6.3: Distribution of mean change in user metrics per user.

Table 6.3: Mean change in user-level metrics

Case	Followers Count	Friends Count	Statuses Count
Departments of Transportation	2.8 (24.06)	0.47 (6.53)	36.7 (120.0)
Drag Race	1.97 (82.75)	0.47 (19.9)	25.57 (117.67)

Table 6.3 lists the mean change in these user level metrics. It is important for researchers to understand how user-level metrics change over time since each time a tweet is viewed or retrieved from the Twitter API, the user-level metrics reflect the statistics at the time of the request. For example, a user may not be very popular as measured by the number of followers at the time of posting a tweet, but if that tweet went viral (Nahon et al., 2011) after being posted, the number of followers and retweets may increase significantly due to the viral event. As a result, the updated statistics may show the user as being more popular than they actually were at the time of posting. Status count has the highest average change which would be expected as users posted more tweets over time during the 90-window of data collection and observation. For the Departments of Transportation case study, the change in status count ranged from -52 to 2,297. It is important to note that each of these metrics may increase or decrease as users lose followers, unfollow users or users delete tweets.

6.3 Tweet Linked Data

Table 6.4: Top 10 URLs by volume.

Departments of Transportation	**RuPaul's Drag Race**
twitter.com: 1,635 bit.ly: 686 tpck.us: 369 ow.ly 197 remmont.com: 148 wa.gov: 59 usa.gov: 46 ca.gov: 38 goo.gl: 38 youtube.com: 33	twitter.com: 36,289 dlvr.it: 8,679 youtu.be: 4,698 instagram.com: 2,847 vine.co: 2,588 youtube.com: 2,333 bit.ly: 2,172 fb.me: 1,236 logo.to: 1,195 nyti.ms: 1,195

Starting at the URL level, Table 6.4 lists the top 10 domains linked to from all tweets with URLs in the Departments of Transportation and RuPaul's Drag Race case studies. For both case studies, other tweets (twitter.com) were the top destination. Other social media sites such as YouTube, Instagram, Vine, and Facebook were also popular sites linked to. As expected, government domains were in the top 10. It is important to note that the top linking destinations for both case studies point to interconnectedness of tweets to other tweets and social media sites.

Table 6.5: Descriptive statistics for archived URLs.

Case	**Contain URLs**	**Selected for Archiving**	**Successfully Archived**	**Average Weeks Archivable**
Departments of Transportation	3,639	2,189 (60.2%)	1,984 (90.6%)	7.23
Drag Race	76,928	48,767 (63.4%)	43,421 (89%)	7.16

Moving to the tweet level, tweets with URLs where randomly selected for archiving during real-time data collection. Table 6.5 provides descriptive statistics for the Tweets with URLs that were selected. Approximately 60% of all tweets with links were selected to have their links archived. All URLs within a tweet were grouped as a single unit. Of those, approximately 90% were able to be archived.

For those that were successfully archived, the context was extracted from each URL embedded in a tweet. For tweets with multiple links, the content of each link was combined into a single document. A simhash (SalahEldeen & Nelson, 2013; Sood & Loguinov, 2011), or similarly hash was calculated. A simhash was used instead of an edit distance because many web pages will have slight changes in content, for example an automatically updated date, to which an edit distance is too sensitive. The distance between two simhashes indicates how similar two documents are to each other.[23] The lower the simhash distance, the more similar the two documents. Two identical documents would have a simhash distance of 0. Two simhash distances of 40 would indicate that the two documents are not very similar. The simhash distance between weekly archives will give us an idea of the extent of change each week. The mean simhash distance between the content in weekly archives of all URLs in tweets selected for archiving is displayed in Figure 6.4.

[23] See http://matpalm.com/resemblance/simhash/ for a non-technical description of the simhash algorithm.

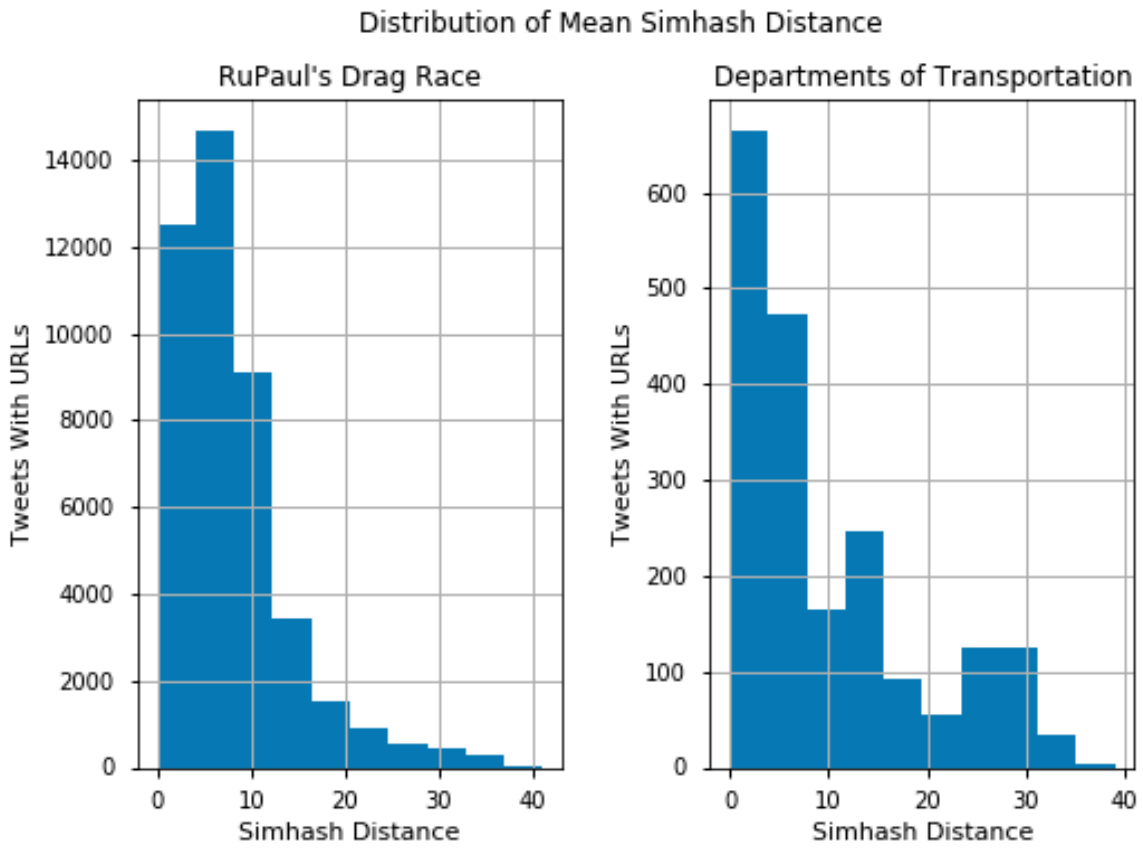

Figure 6.4: Distribution of mean simhash distance between the content of weekly archives of URLs within tweets selected for archiving. Content in all URLs for each tweet was grouped into one unit. Tweet URLs archived for less than two weeks excluded.

In the Departments of Transportation case study, 90% of all the URLs within a tweet (where all URLs within a tweet were treated as a single unit) were archived for the full 8 weeks. Over 98% of URLs that were achievable in real-time were also archivable over the entire 8-week period. As shown in Figure 6.4, the majority of URLs were identical over the archivable time period. The mean simhash distance between archives was 9.8. For the RuPaul's Drag Race case study, 89% of URLs were archivable for the entire 8-week period. As expected, the stability of URLs is higher in the Departments of Transportation case.

6.4 Chapter Summary

This chapter introduced the second concept closely connected with ephemerality: authenticity. Within the context of social media, the authenticity of a post involves capturing and stabilizing the surrounding metadata so a post or digital object can stand on its own. The authenticity of a social media dataset was measured via the level of change in metadata and linked URLs embedded in a post over time when the same parameters are used to collect the data at different points in time.

The descriptive statistics for metadata change in the RuPaul's Drag Race and Departments of Transportation case studies support the hypothesis that the level of metadata stability would differ for each case studies due its context. Official government accounts did not change their profile information while over 50% of users in the Drag Race case study changed their profile image and over 42% of users changed their profile description. Follower, friend, and status count had a high rate of change for tweets in each of the case studies. For research analyzing user profile information or tweet statistics, these results point to the important of preserving the state of this metadata at the moment of data collection or taking the high rate of change into account during data analysis.

The link analysis points to the interconnected nature of social media platforms and posts through the links users embedded within the post content. This is demonstrated by the top 10 domains for the Departments of Transportation and RuPaul's Drag Race case studies. The top 10 domains included other posts within Twitter as well as links out to other platforms including Instagram, YouTube, Vine, and Facebook. Links within the

Departments of Transportation case also included many .gov or URL shorteners offered by the US Federal Government when linking to government content. This points to the limitations of current methods focusing on a single social media platform for analysis, which does not accurately reflect how users actually use social media platforms.

Chapter 7. IMPACTS OF EPHEMERALITY

In this chapter, I summarize findings related to the impacts of ephemerality on the reliability and authenticity of the social media datasets within and across each of the three case studies. The observations collected in this dissertation are descriptive in nature. This was a necessary first step to understand the effects of ephemerality on social media dataset since we lack empirical data relating to the impact, if any, of latency on of the social media posts, metadata, and linked data collected for research purposes. The analysis is based on a combination of the descriptive statistics described in Chapters 4-6, my experience collecting and analyzing social media data, and when possible, inferential statistics.

7.1 Reliability - The Relationship Between Time and Ephemerality

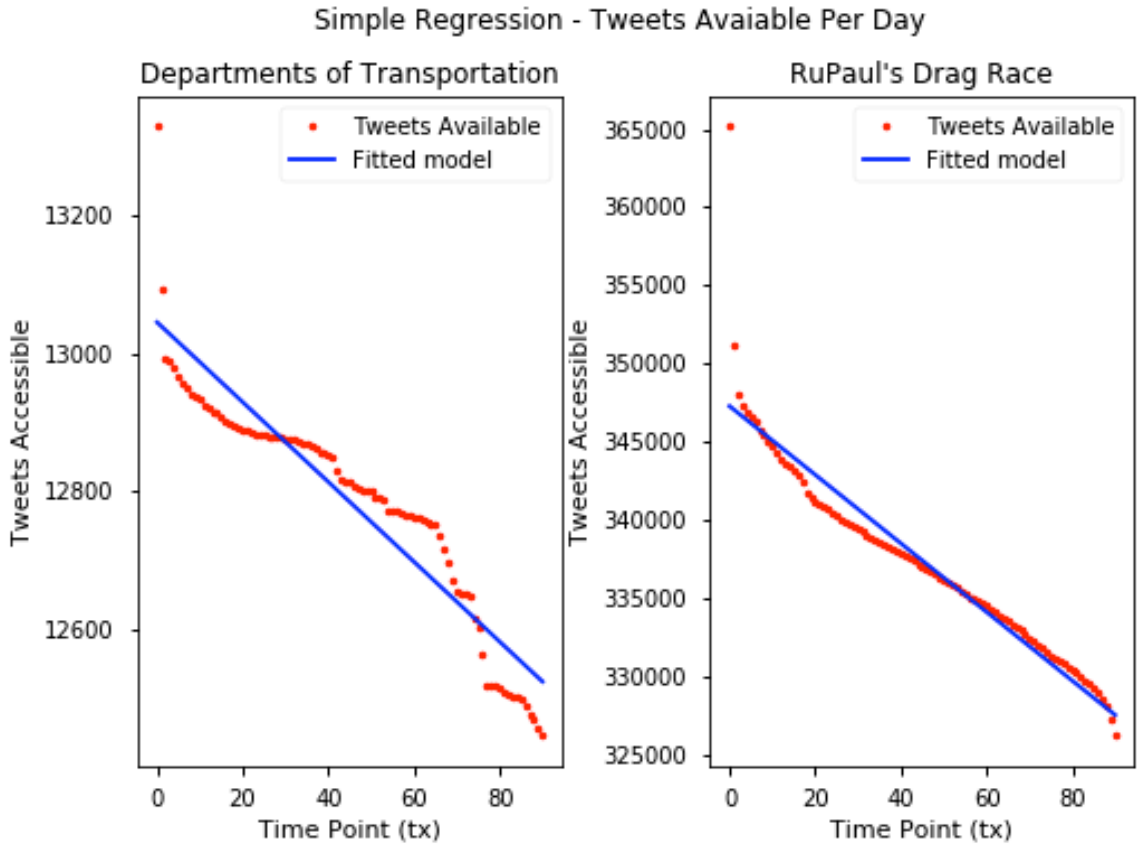

Figure 7.1: Simple regression of tweet accessibility at time points t0 - t90 for the Departments of Transportation and RuPaul's Drag Race case studies.

As illustrated in Figure 7.1, latency in data collection is a significant predictor of tweet accessibility. A simple regression model was calculated to predict the number of tweets accessible in each case based on the days since a tweet was posted. A significant regression equation was found ($F(1, 89)=676$, $p<.000$), with an R^2 of .884 for the RuPaul's Drag Race case study. The predicted number of tweets available is equal to 347,300 + -5.7857 (days) when days is measured in days since a tweet was posted. The number of tweets available decreased -5.7857 for each day since posting. A significant regression equation was also found ($F(1, 89)=651.6$, $P<000$), with an R^2 of .880 for the

Departments of Transportation case study. The predicted number of tweets available is equal to 13,050 + -219.4831 (days) when days is measured in days since a tweet was posted. The number of tweets available decreased -219.4831 for each day since posting. A regression equation was not calculated for the Occupy Wall Street case study because data was only collected for two time points (real-time and three years later). As latency in data collection increases, the number of inaccessible posts increases. The highest number of inaccessible posts occur within the first 48 hours after tweets were created in both the Departments of Transportation (38%) and RuPaul's Drag Race (44%) case studies. This points to the importance of collecting social media data in real-time and indicates that datasets may contain large gaps with as little as a 24-hour latency in data collection. Large gaps in social media datasets may result in a dataset that no longer represents the social media posts at the time under investigation. For example, when examining the spread of rumors after disasters, users may delete posts containing misinformation resulting in the disappearance of rumors in datasets collected hours or days later. This is a concern for researchers as social media data collected with a high latency may no longer accurately reflect the social media posts at that time.

7.2 Authenticity: The impact of the prototypical features

As noted in Chapter 4, a case study approach was chosen to closely replicate the prototypical features of data collection scenarios social science researchers commonly use. While it is not possible to determine which prototypical features have the most impact, it is possible to determine if the differences between the three cases are

statistically significant. A chi-square test of independence was calculated between each of the three case studies on the core measure of reliability — the number of tweets inaccessible at the end of the period of observation (described in Table 5.1) — finding that the cases are independent $\chi^2(2, N=2{,}667{,}515)$, $p<.001$.[24] The significant results of the test can be interpreted to mean that the differences between cases is not due to randomness, but due to the prototypical features of each case study (described in Table 4.1). If the differences were not significant then the differences in the cases would not be meaningful. I posit that the following prototypical features have an impact on the level of ephemerality in each case study:

- **Context.** The context of each case study is related to how users of within the case study utilize the platform, including the types of content posted as well as their relationship to other users. For example, the Departments of Transportation case study has the lowest level of inaccessible tweets and this may be to the low level of political contention of the everyday political context — the discussion and sharing information related to state transportation infrastructure is potentially less contentious than the case contexts. The Occupy Wall Street case study has the highest level of tweet inaccessible tweets. This is unsurprising due to the highly contentious nature of social movements and, as the literature has shown, deletion is used as a protest tactic (Neumayer & Stald, 2014). The reality TV context of RuPaul's Drag race sits between the other two case studies resulting in the second-highest level of tweet inaccessibility. The impact of context confirms work

[24] With a N of over 2 million, it would be expected to find significance.

examining rumor spread (Starbird et al., 2014) and behavior around “regret” posts (Knapp et al., 1986; Petrovic et al., 2013; Sleeper et al., 2013) on social media sites.

- **Query Terms.** While each case study was collected in real-time via Twitter’s Streaming API, the construction of query terms differed significantly. The Occupy Wall Street case study was based solely on keywords, the Departments of Transportation case study was based solely on following a fixed set of official government Twitter accounts, and the RuPaul’s Drag Race was based on a combination of keywords and a bound set of accounts. A focus on following users vs. keywords may also relate to the impact of cascading account and tweet deletion — where a deleted account causes retweets in other accounts to be deleted — is slightly higher in the Departments of Transportation case. Notably, query construction may have an impact on the distribution and change of tweets with certain types of entities (URLs, mentions, and hashtags) as evidenced by the results in Tables 6.1 and 6.2.
- **Metadata Stability.** Part of the criteria for each case study was an expectation that there would be a different level of stability of the metadata surrounding each case study. The descriptive statistics in the RuPaul’s Drag Race and Departments of Transportation case studies support those assumptions. I’ll discus the implication of changes to user and tweet-level metadata separately:
    - **User-Level Metadata — Screen Name.** Of special importance to note is the small percentages of users who changed their screen name during the

time of observation. The screen name servers a dual purpose on most social media platforms: (1): a descriptive element similar to other items in a user's profile such as the profile image or user description and (2): a unique identifier for each user that can be used to collect data. In Twitter, and other social media platforms, a user can be identified by a screen name and a unique numerical identifier. This unique numerical identifier is not editable by a user, but the screen name is. If researchers use a user's screen name as a query term in their data collection, the user may drop out of their data collection if they change their screen name. For the two case studies, 1-5.5% of users changed their screen names. This supports a best practice of using a user's unique numerical identifier instead of their screen name as a query term as any change to the user's screen name will impact account-focused data collection. Users who change their username would no longer be part of data collection after the change unless the unique identifier is used.

- **User-Level Metadata — Profile Metadata.** User-level metadata such as the screen name, name, profile image, location, and description are often used to categorize users. Over half of the users in the Drag Race case study change their profile image or profile description and a quarter of users changed their profile information in the Departments of Transportation case. The implication is that a user may change their presentation of themselves thereby impacting how a user is categorized, pictured in their

profile photo, or described in their profile text. If the profile data analyzed is not part of the initial data collection — for example, viewing user profile photos at a later data, the bond between user profile image and the post may be broken. If a research design utilizes the user profile metadata, a deal in collection of this information may result in incorrect categorization.

- **Tweet-Level Metadata.** Tweet-level metadata such as the number of retweets or favorites provide a similar set of challenges as user-level metadata. Since this data can change very quickly, it is important to note when it was collected for analysis and contextualized within that timeframe.

- **Linked Content.** Social media platforms are interconnected through the links users embedded within the post content. This is demonstrated by the top 10 domains for the Departments of Transportation and RuPaul's Drag Race case studies. The top 10 domains included other posts within Twitter as well as links out to other platforms including Instagram, YouTube, Vine, and Facebook. Links within the Departments of Transportation case also included many .gov or URL shorteners offered by the US Federal Government when linking to government content. This points to the limitations of current methods focusing on a single social media platform for analysis, which does not accurately reflect how users actually use social media platforms.

7.3 LIMITATIONS

The focus of this dissertation was to further describe and explore the problem space around the ephemerality of social media datasets specifically focusing on the reliability of posts and the authenticity of metadata surround those posts. Heavy reliance on descriptive statistics, the data collection environment, and case construction while providing new insights, also create a set of limitations:

- In the Occupy Wall Street case, data collection was limited to two time points (real-time and three years later). As a result, the most politically contentious case study was excluded from the majority of the within case study analyses
- Due to the daily granularity of nightly data collection, it is not possible to detect multiple changes occurring within each day or disambiguate missing posts due to API errors vs. changes in user privacy settings. It was not possible to collect data more often than once-a-day due to API request rate limits since checking the status of all Tweets in the larger case studies took 6-7 hours.
- While the difference between case studies were significant, it is not possible to determine which prototypical features has the highest impact on the ephemerality of each dataset.
- The descriptive and exploratory nature of this work limited the use of inferential statistics to determine causality.
- While Twitter shares many concepts, structures, metadata, and links (URLs); patterns of user activity within Twitter may differ from social media platforms, limiting my ability to generalize outside of Twitter.

- Between and during period of data collection, Twitter made changes to the affordances of the platform. For example, Twitter introduced the simplified replies and media attachments where the @mentions at the beginning of a tweet and URLs linking to media (photos, videos, and GIFs) at the end of a tweet do not count toward the 140-character limit.[25] While none of the metadata fields analyzed in this dissertation were changed, changes in affordances may have impacted user behavior.

7.4 Contributions

Returning to the guiding research questions of this work, the findings address the interaction between ephemerality and the process of data collection. This dissertation advances the field of information science by empirically investigating how the ephemeral nature of social media data, metadata, and linked content have significant and lasting effects on the reliability and authenticity of datasets used in research. Situating research design decisions, specifically choices made on how and when to observe data, within the frameworks of process theory and archival theory, this work brings the importance of methodological considerations to the forefront of studies of digital and social media.

Key contributions of this work include:

- The introduction of a new framework detailing typical methodologies for sampling data from digital and social media platforms.

[25] See https://dev.twitter.com/overview/api/upcoming-changes-to-tweets for a description of changes made to tweets, archived at http://perma.cc/K8D6-BHM3.

- Demonstrates through an empirical analysis of descriptive data related to the reliability and authenticity across three illustrative case studies, how the challenges of ephemerality of social media translate to consequences for research.
- A design and technical system for archiving links embedded in social media datasets during data collection.
- Guidelines and design considerations for social media-based research studies to aid in limiting and understanding the impact of ephemerality.
- Limitations of social media datasets and the importance of these limitations within the field of information science.

7.5 Future Work

This research addressed gaps in the current literature related to social media methods and data collection. Invitation of this problem space revealed directions for future work including:

- Integrating ongoing re-conceptualizing of the archival record to take into account the concept of non-fixed, event-oriented records. Seeing social media as a performance that cannot be separated from its creator (Anderson, 2013, p. 362).
- Examining the extent to which the changes in the reliability and authenticity due to latency in data collection impact the results of an analysis within the same research project — this could take the form of analyzing social media data addressing the same research question at different data collection latencies.

- While this work found that the prototypical features of a case study result in different patterns of inaccessibility, it was not possible to determine which have the most impact. Future research designs could further examine which prototypical features have the most impact.
- Conduct sensitivity testing and random modeling of inaccessibility to better determine the extent of impact within each case study.
- Addressing ethical questions surrounding ephemeral social media data sets — both from a research methods as well as a human-subjects angles.

7.6 Conclusion

The process of collecting social media data presents a number of challenges for researchers as we attempt to add rigor to the field. In this dissertation I developed the concept of ephemerality as it relates to social media data sets – quantifying the levels of reliability and authenticity within three cases studies observed over a 90 day to 3 year timeframe. To me, the most surprising results were the levels of change of user profile metadata with over 50% of users in the RuPual's Drag Race case study changing their profile images and a large number of users completely rewriting their profile descriptions. The empirical results lay the foundation for future work examining what impact latency in data collection and the resulting change in a social media data set have on findings.

The results point to the need for researchers to more closely align the latency and methods of data collection with their research design. Some researchers may see these

results as a call to strengthen their data collection methods to prevent change and stabilize their data sets. Other researcher may see ephemerality as an inherent property of social media data itself. I do not have a normative stance on either view, but the results point to the importance of taking the impact of data set change into account when describing findings and limitations of research using social media data. These findings also point to the importance of more clearly describing our data collection procedures when publishing research so readers may evaluate findings in light of the research design choices that were made. Both of steps will go a long way to increasing the rigor of social media research.

WORKS CITED

Abbott, A. (1990). A Primer on Sequence Methods. *Organization Science*, *1*(4), 375–392. http://doi.org/10.1287/orsc.1.4.375

Abowd, J. M., Vilhuber, L., & Block, W. (2012). A proposed solution to the archiving and curation of confidential scientific inputs. *Privacy in Statistical Databases*.

Acker, A. (2014). *Born networked records: A history of the short message service format* (Order No. 3623371). Available from ProQuest Dissertations & Theses Global. (1549977698). Retrieved from https://search.proquest.com/docview/1549977698?accountid=14784

Acker, A., & Brubaker, J. R. (2014). Death, Memorialization, and Social Media: A Platform Perspective for Personal Archives. *Archivaria*, (77), 1–23.

Agarwal, S. D., Bennett, W. L., Johnson, C. N., & Walker, S. (2014). A Model of Crowd Enabled Organization: Theory and Methods for Understanding the Role of Twitter in the Occupy Protests. *International Journal of Communication*, *8*, 27.

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., & Acquisti, A. (2013). Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 897-908). ACM.

Ananny, M. (2015). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology & Human Values*, *41*(1), 93–117. http://doi.org/10.1177/0162243915606523

Ananny, M., & Crawford, K. (2017). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *33*(4), 146144481667664–17. http://doi.org/10.1177/1461444816676645

Anderson, K. (2013). The footprint and the stepping foot: archival records, evidence, and time. *Archival Science*, *13*(4), 349–371. http://doi.org/10.1007/s10502-012-9193-2

Ankerson, M. S. (2012). Writing web histories with an eye on the analog past. *New Media & Society*, *14*(3), 384–400. http://doi.org/10.1177/1461444811414834

Babbie, E. R. (2007). The Practice of Social Research (11 ed.). Belmont: Thomson Wadsworth.

Bamman, D., O'Connor, B., & Smith, N. (2012). Censorship and deletion practices in Chinese social media. *First Monday*, *17*(3), 259. http://doi.org/10.5210/fm.v17i3.3943

Bastos, M. T., Mercea, D., & Charpentier, A. (2015). Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, *65*(2), 320-350.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication &*

Society, *15*(5), 739-768.
Bennett, W. L., Segerberg, A., & Walker, S. (2014). Organization in the crowd: peer production in large-scale networked protests. *Information, Communication & Society*, *17*(2), 232–260. http://doi.org/10.1080/1369118X.2013.870379
Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., & Vargas, G. G. (2011, July). 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *ICWSM* (pp. 50-57).
Bowker, G. C. (2013). Data flakes: An afterword to "Raw Data"is an oxymoron. In *Raw Data Is an Oxymoron*. MIT Press.
Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.
Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227. http://doi.org/10.1007/s10676-013-9321-6
Brooks, M. (2015). *Human centered tools for analyzing online social data* (Order No. 10000022). Available from ProQuest Dissertations & Theses Global. (1760603707). Retrieved from https://search.proquest.com/docview/1760603707?accountid=14784
Bruns, A. (2012). How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi. *Information, Communication & Society*, *15*(9), 1323–1351. http://doi.org/10.1080/1369118X.2011.635214
Bruns, A., & Burgess, J. (2011). #Ausvotes: How twitter covered the 2010 Australian federal election. *Communication, Politics & Culture*, *44*(2), 37.
Bruns, A., & Stieglitz, S. (2012). Quantitative Approaches to Comparing Communication Patterns on Twitter. *Journal of Technology in Human Services*, *30*(3-4), 160–185. http://doi.org/10.1080/15228835.2012.744249
Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, *14*(7), 1164–1180. http://doi.org/10.1177/1461444812440159
Bucher, T., & Helmond, A. (2017). The affordances of social media platforms. In J. Burgess, T. Poell, & A. E. Marwick (Eds.), *The SAGE Handbook of Social Media*. London and New York: Sage Publications Ltd.
Burgess, J., & Bruns, A. (2014). Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn. In G. Elmer, J. Langlois, & J. Redden (Eds.), *Compromised Data From Social Media to Big Data* (pp. 1–27).
Caren, N., & Gaby, S. (2012). Sociologist Tracks Social Media's Role in Occupy Wall Street Movement. University of North Carolina.

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continuum*, *23*(4), 525–535. http://doi.org/10.1080/10304310903003270

Crowston, K. (2000). Process as Theory in Information Systems Research. In *Organizational and Social Perspectives on Information Technology* (pp. 149–164). Boston, MA: Springer US. http://doi.org/10.1007/978-0-387-35505-4_10

Dabbish, L., Venolia, G., & Cadiz, J. J. (2003). Marked for deletion: an analysis of email data. *CHI '03 extended abstracts* (pp. 924–925). New York, New York, USA: ACM. http://doi.org/10.1145/765891.766073

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Society and Space open site*.

Daniels, M. F., & Walch, T. (1984). *Modern archives reader*. Washington, DC: National Archives and Records Service, US General Services Administration, 1984.

de Leeuw, E. D., Hox, J., & Dillman, D. (2012). International Handbook of Survey Methodology. Routledge.

Dimitrova, D. V., & Bugeja, M. (2007). Raising the dead: Recovery of decayed online citations. *American Communication Journal*, *9*(2), 2.

Driscoll, K., & Walker, S. (2014). Big Data, Big Questions Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, *8*, 20.

Duranti, L. (1994). The concept of appraisal and archival theory. *The American Archivist*, *57*(2), 328-344.

Duranti, L. (1995). Reliability and authenticity: the concepts and their implications. *Archivaria*, 39.

Duranti, L. (1997). The Archival Bond. *Archives and Museum Informatics*, *11*(3-4), 213–218. http://doi.org/10.1023/A:1009025127463

Duranti, L., Eastwood, T., & MacNeil, H. (2013). Preservation of the Integrity of Electronic Records. Dordrecht: Springer Science & Business Media. http://doi.org/10.1007/978-94-015-9892-7

Edwards, P. N. (2010). A Vast Machine. MIT Press.

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, *3*(1). http://doi.org/10.1177/2053951716645828

Flaxman, S., Goel, S., & Rao, J. M. (2013). Ideological Segregation and the Effects of Social Media on News Consumption. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2363701

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media , 58*(1), 59-75.

Gaffney, D., & Puschmann, C. (2014). Data collection on Twitter. Twitter and Society.

New York.
Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society*, *15*(8), 1348–1365. http://doi.org/10.1177/1461444812472322
Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: collections, baselines, sampling. *M/C Journal*, 16(2).
Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*.
Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, *12*(3), 347–364. http://doi.org/10.1177/1461444809342738
Goble, C., Stevens, R., Hull, D., Wolstencroft, K., & Lopez, R. (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, *9*(6), 506–517. http://doi.org/10.1093/bib/bbn034
Goffman, E. (1990). The Presentation of Self in Everyday Life. Penguin Books, Limited (UK).
Goh, D. H. L., & Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, *58*(1), 15–24. http://doi.org/10.1002/asi.20513
González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16–27. http://doi.org/10.1016/j.socnet.2014.01.004
Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C., & Vandenberg, J. (2002). Online scientific data curation, publication, and archiving. *arXiv preprint cs/0208012*.
Grosser, B. (2014). What do metrics want? How quantification prescribes social interaction on Facebook. *Computational Culture: a journal of software studies,* 4.
Gummadi, K. P., Saroiu, S., & Gribble, S. D. (2002, November). King: Estimating latency between arbitrary internet end hosts. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment* (pp. 5-18). ACM.
Hargittai, E., & Sandvig, C. (2015). Digital Research Confidential. MIT Press.
Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media & Society*, *1*(2). http://doi.org/10.1177/2056305115603080
Herring, S. C. (2010). Web Content Analysis: Expanding the Paradigm. In *Web content analysis: Expanding the paradigm* (pp. 233–249). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-1-4020-9789-8_14
Holmes, O. (1964). Archival Arrangement—Five Different Operations at Five Different Levels. *The American Archivist , 27(1), 21-42*.
Karlsson, M. (2012). Charting the liquidity of online news: Moving towards a method for content analysis of online news. *International Communication Gazette*, *74*(4), 385–402. http://doi.org/10.1177/1748048512439823

Karpf, D. (2012). SOCIAL SCIENCE RESEARCH METHODS IN INTERNET TIME. *Information, Communication & Society*, *15*(5), 639–661. http://doi.org/10.1080/1369118X.2012.665468

Kim, J., & Kim, E. J. (2008). Theorizing Dialogic Deliberation: Everyday Political Talk as Communicative Action and Dialogue. *Communication Theory*, *18*(1), 51–70. http://doi.org/10.1111/j.1468-2885.2007.00313.x

Kitchin, R. (2014). The Data Revolution. SAGE.

Knapp, M. L., Stafford, L., & Daly, J. A. (1986). Regrettable Messages: Things People Wish They Hadn't Said. *Journal of Communication*, *36*(4), 40–58. http://doi.org/10.1111/j.1460-2466.1986.tb01449.x

Krippendorff, K. (2012). Content analysis: An introduction to its methodology, Sage.

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady,* 10(8), 707-710.

Liang, H., & Fu, K.-W. (2015). Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science. *PLoS ONE*, *10*(8), e0134270. http://doi.org/10.1371/journal.pone.0134270

Light, B., & McGrath, K. (2010). Ethics and social networking sites: a disclosive analysis of Facebook. *Information Technology & People*, *23*(4), 290–311. http://doi.org/10.1108/09593841011087770

Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*(7209), 28–29. http://doi.org/10.1038/455028a

Malik, M. T., Gumel, A., Thompson, L. H., Strome, T., & Mahmud, S. M. (2011). "Google Flu Trends" and Emergency Department Triage Data Predicted the 2009 Pandemic H1N1 Waves in Manitoba. *Canadian Journal of Public Health / Revue Canadienne De Sante'e Publique*, *102*(4), 294–297. http://doi.org/10.2307/41995614?ref=no-x-route:af550292fd45cb5ef4f567324da26478

Manovich, L. (2013). *Software takes command* (Vol. 5). A&C Black.

McKemmish, S. (2001). Placing records continuum theory and practice. *Archival Science*, *1*(4), 333–359. http://doi.org/10.1007/BF02438901

Miller, C., Ginnis, S., Stobart, R., Krasodomski-Jones, A., & Clemence, M. (2015). The road to representivity, a Demos and Ipsos MORI report on sociological research using Twitter. *London: Demos. Available at: http://www. demos. co. uk/files/Road_to_representivity_final. pdf*, *1441811336*.

Moghaddam, A. I., Saberi, M. K., & Esmaeel, S. M. (2012). Availability and half-life of web references cited in Information Research Journal: a citation study. *International Journal of Information Science and Management (IJISM)*, *8*(2), 57–75.

Mohr, G., Stack, M., Ranitovic, I., Avery, D., & Kimpton, M. (2004). An Introduction to Heritrix An open source archival quality web crawler. In *In IWAW'04, 4th*

International Web Archiving Workshop.

Nahon, K. (2015). Where there is Social Media there is Politics. In A. Bruns, E. Skogerbo, C. Christensen, O. A. Larsson, & G. S. Enli (Eds.), *Routledge Companion to Social Media and Politics* (pp. 39–55). NYC, NY.

Nahon, K., Hemsley, J., Walker, S., & Hussain, M. (2011). Fifteen Minutes of Fame: The Power of Blogs in the Lifecycle of Viral Political Information. *Policy & Internet*, *3*(1), 1–28. http://doi.org/10.2202/1944-2866.1108

Neumayer, C., & Stald, G. (2014). The mobile phone in street protest: Texting, tweeting, tracking, and tracing. *Mobile Media & Communication*, *2*(2), 117–133. http://doi.org/10.1177/2050157913513255

Park, H. W., & Thelwall, M. (2003). Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer-Mediated Communication*, *8*(4). http://doi.org/10.1111/j.1083-6101.2003.tb00223.x

Parmelee, J. H., & Bichard, S. L. (2013). Politics and the Twitter Revolution.

Patton, M. Q. (2001). Qualitative Research & Evaluation Methods (3rd ed.). Thousand Oaks, Calif: SAGE Publications, Inc.

Pearce-Moses, R. (2005). A glossary of archival and records terminology. Society of American Archivists.

P Petrovic, S., Osborne, M., & Lavrenko, V. (2013). I wish i didn't say that! analyzing and predicting deleted messages in twitter. *arXiv preprint arXiv:1305.3107*.

Phillips, W. (2011). LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*, *16*(12). http://doi.org/10.5210/fm.v16i12.3168

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *18*(1), 1–18. http://doi.org/10.1177/1461444816661553

Ribes, David. "The kernel of a research infrastructure." In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 574-587. ACM, 2014.
http://doi.org/10.1145/2531602.2531700

Roblyer, M. D., McDaniel, M., Webb, M., Herman, J., & Witty, J. V. (2010). Findings on Facebook in higher education: A comparison of college faculty and student uses and perceptions of social networking sites. *The Internet and Higher Education*, *13*(3), 134–140. http://doi.org/10.1016/j.iheduc.2010.03.002

SalahEldeen, H. M., & Nelson, M. L. (2013). Reading the correct history?: modeling temporal intention in resource sharing. *the 13th ACM/IEEE-CS joint conference* (pp. 257–266). New York, New York, USA: ACM. http://doi.org/10.1145/2467696.2467721

Saltzis, K. (2012). Breaking News Online: How news stories are updated and maintained

around-the-clock. *Journalism Practice*, *6*(5-6), 702–710. http://doi.org/10.1080/17512786.2012.667274

Sanderson, R., Phillips, M., & Van de Sompel, H. (2011). Analyzing the persistence of referenced web resources with memento. *arXiv preprint arXiv:1105.3459*.

Seaver, N. (2015). The nice thing about context is that everyone has it. *Media, Culture & Society*, *37*(7), 1101–1109. http://doi.org/10.1177/0163443715594102

Segerberg, A., & Bennett, W. L. (2011). Social Media and the Organization of Collective Action: Using Twitter to Explore the Ecologies of Two Climate Change Protests. *Communication Review*, *14*(3), 197–215.

Simons, H. (2006). The Paradox of Case Study. *Cambridge Journal of Education*, *26*(2), 225–240. http://doi.org/10.1080/0305764960260206

Sleeper, M., Cranshaw, J., Kelley, P. G., Ur, B., Acquisti, A., Cranor, L. F., & Sadeh, N. (2013). i read my Twitter the next morning and was astonished: a conversational perspective on Twitter regrets. *the SIGCHI Conference* (pp. 3277–3286). New York, New York, USA: ACM. http://doi.org/10.1145/2470654.2466448

Sood, S., & Loguinov, D. (2011, October). Probabilistic near-duplicate detection using simhash. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1117-1126). ACM.

Starbird, K., & Palen, L. (2010). *Pass it on?: Retweeting in mass emergency*(pp. 1-10). International Community on Information Systems for Crisis Response and Management.

Starbird, K., & Palen, L. (2012). (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. *the ACM 2012 conference* (pp. 7–16). New York, New York, USA: ACM. http://doi.org/10.1145/2145204.2145212

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *iConference 2014 Proceedings: Breaking Down Walls. Culture - Context - Computing*. http://doi.org/10.9776/14308

Thumim, J. (2002). 'Mrs. Knight must be balanced': Methodological problems in researching early British television. In S. Allan, B. G, & C. C (Eds.), *News, Gender, and Power* (pp. 91–104). London: News.

Valenti, J. (2014, May 28). # YesAllWomen Reveals the Constant Barrage of Sexism That Women Face. *The Guardian*. The Guardian. Retrieved from http://www.theguardian.com/commentisfree/2014/may/28/yesallwomen-barage-sexism-elliot-rodger

van Dijck, J. (2013). The Culture of Connectivity. Oxford University Press.

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10). http://doi.org/10.5210/fm.v18i10.4878

Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*, *69*(3), 384–410. http://doi.org/10.1108/JD-03-2012-0027

Yin, R. K. (2014). Case Study Research. SAGE Publications.

Zhang, B., Ng, T. S. E., Nandi, A., Riedi, R., Druschel, P., & Wang, G. (2006). Measurement based analysis, modeling, and synthesis of the internet delay space. *the 6th ACM SIGCOMM* (pp. 85–98). New York, New York, USA: ACM. http://doi.org/10.1145/1177080.1177091

Zhang, S. (2015). Using Twitter to Enhance Traffic Incident Awareness. *2015 IEEE 18th International Conference on Intelligent Transportation Systems - (ITSC 2015)*, 2941–2946. http://doi.org/10.1109/ITSC.2015.471

Zhou, L., Wang, W., & Chen, K. (2016). Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones (pp. 603–612). International World Wide Web Conferences Steering Committee. http://doi.org/10.1145/2872427.2883052

Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, *12*, 313–325. http://doi.org/10.1007/s10676-010-9227-5

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. http://doi.org/10.1108/AJIM-09-2013-0083

APPENDIX A: IMPLICATIONS OF THIS RESEARCH FOR SOCIAL MEDIA RESEARCH

In this appendix, I move beyond the context of Twitter to discuss general implications for social media research arising out of this work. For researchers who have not yet started data collection, these implications act as a set of considerations for approaching data collection. For researchers who already collected data, these implications provide a set limitations of their data collection procedures. These implications emerge from the findings in this dissertation as well as my direct experience with the challenges of collecting and analyzing social media data.

As previously noted in Chapter 5, the data in this dissertation comes from three Twitter-based case studies. Twitter was chosen as the objects of study because: 1) the use of Twitter as an object of study and source of observational data is pervasive in academic research (Williams et al., 2013; Zimmer & Proferes, 2014), 2) Twitter is less susceptible to algorithmic filtering, also called 'filter bubbles' (Bozdag, 2013; Bruns & Stieglitz, 2012; Bucher, 2012; Flaxman et al., 2013; van Dijck, 2013, p. 75), than other platforms since the public APIs return all public, non-deleted statuses matching query terms; theoretically producing a more "accurate" record[26] (Driscoll & Walker, 2014), and 3) concepts, structures, metadata, and links (URLs) easily generalize beyond Twitter to other social media services and platforms. By focusing on the concepts, structures,

[26] https://dev.twitter.com/streaming/overview

metadata, and links within each case study, a general set of implications for social media research emerged.

1. **The Impact of Latency in Data Collection.** The time between data collection and the event/phenomenon under investigation and data collection is important consideration since latency is a significant factor in predicting the availability of posts within social media data sets. Data from the case studies in this dissertation point to the first 48 hours as a critical time period, but posts continue to become unavailable over time.

   In addition to the posts, metadata surrounding posts also changes over time. Users change their profiles, images, and locations. Statistics related to users and posts also change over time. This content is bound to a specific point in time, so breaking is may result in the analysis of data not related to the original post content. For example, if an Instagram post was collected in September but the content of user profile was viewed and analyzed in January of the following year, the user profile may no longer represent the user at the time of posting. To address this issue, collection of the Instagram post as well as the user profile would need to be integrated into the research design.

2. **Posts are Assemblages of Content.** Social media platforms are not just filled with text. Posts and user profiles are assemblages of content of content made up of the post content, post metadata, user metadata, and linked content. Depending on the platform, the content of the post could be text — Tweets are 140

characters — or images and text — Instagram posts contain an image as well as descriptive text. Metadata about the user and other users' interactions with the posts are also displayed — the number of retweets on Twitter or the number of links for an Instagram post. Content is also linked-to via URLs or embedded into the rendered post. When a post is rendered for a platform's API or user-facing web interface, subsets of the content is rendered and organized by the logic of the algorithms within the platform. When collecting data, the interface used by a researcher may render all or part of this content and metadata leaving the researchers with a partial view of the post.

3. **The Affordances of Platforms Create Constraints for Users and Researchers.**

The affordances, or features, of social media platforms create constraints for users as well as researchers collecting data from the platforms. Platforms offer users a specific set of interactions, as a result, users are unable to perform activities and actions not offered by a site. For example, Facebook users are provided with a set of emotional reactions (like, love, haha, cry, angry, and wow) to respond to each post, tweets are limited to 140 characters, and Snapchat messages can be viewed for a limited amount of time before self-destructing. Users sometimes develop practices to get around these limitations, such as including links in tweets or taking screenshots of snaps. The affordances and practices within a platform must be closely matched to the research questions and phenomena under investigation.

Researchers are also constrained by the affordances of the interfaces each platform offers for data collection. Data can be collected from user-focused web interfaces and software-focused APIs — each offering their own sets of constraints and subset of data. For example, collecting screen shots of posts would provide a rendering of the post similar to the experience of platform users, but may contain a limited set of metadata about the post and user. Collecting the same post via a platform's API may provide extended metadata, but the format not provide information about how posts are rendered and presented to users of the platform.

Similarly, the metrics provided by each platform privilege some activities while limiting or preventing the visibility of other types of activities. For example, the number of times a tweet was retweeted is often used as a measure of the popularity or reach of a tweet. The number retweets is displayed prominently when viewing a tweet on the Twitter website. It is important to note that no other measures related to the number of times a tweet has been seen by users. Using the number of retweet as a measure privileges production of posts over for other types of listening.

4. What data may I want to collect and analyze?

Below is a list of components of a social media post. Considerations for data collection are given for each component:

- Post content. The content of posts on most social media platforms consist of more than just the text -- posts often include images and links. If this content will be included in your analysis, it should also be included as part of your data collection. If you're using an API to collect data, mapping the fields received from

the API to the rendered post on the public web-interface can help locate missed content and differences between the text-based API response of the API and the rendered version of the post.

- User profiles. Users change the name, description, image, location, and URLs within their profiles. If user profile information is included in your analysis, consider if your data collection strategy includes all of this information. As users change their profile information, their roles and presentations may change significantly.
- URLs. Embedding URLs in posts and user profiles is a common affordance of social media platforms. Links extend the reach and content of a post, so examining the content of URLs should be considered when analyzing the content of a social media post. URLs change or become inaccessible over time, so an archiving strategy should be considered as part of your data collection strategy.
- Query terms. A platform may also allow users to change their usernames, impacting your data collection if you use usernames as part of your query terms. Most platforms assign users unique numeric ids that do not change when a user updates their screen names. The unique user id is a more stable reference to use as a query term than usernames.

APPENDIX B: CASE STUDY QUERY TERMS

Occupy Wall Street - Keywords

A list of keywords used to collect tweets surrounding the Occupy Wall Street movement can be found at https://github.com/somelab/SoMeToolkit/blob/master/collection.terms.

RuPaul's Drag Race Case - List of Accounts and Keywords Followed

Keywords	#DragRace, #DragRaceAllStars, #AllStars, #AllStars2, #RPDR, #RuPaul - popular show hashtags DragRace OR DragRaceAllStars OR AllStars OR AllStars2 OR RPDR OR RuPaul OR RuPaulsDragRace OR DragRaceAllStar OR RuPaul OR michellevisage OR AdoreDelano OR Alaska5000 OR AlyssaEdwards_1 OR cocomontrese OR TheOnlyDetox OR TheGingerMinj OR katya_zamo OR PhiPhiOhara OR roxxxyandrews OR TATIANNANOW - hashtags used to express support for the final 3 constants
Accounts	@RuPaulsDragRace - OFFICIAL @LogoTV #DragRace Twitter account @RuPaul - RuPaul's official account and head judge @Michellevisage - Drag Race Judge Show contestants: @AdoreDelano @Alaska5000 @AlyssaEdwards_1 @Cocomontrese @TheOnlyDetox @TheGingerMinj @katya_zamo @PhiPhiOhara @Roxxxyandrews @TATIANNANOW

Departments of Transportation Case - List of Accounts Followed

State	Accounts
Washington	@wsdot - Statewide updates @wsdot_traffic - Traffic and construction reports for King, Snohomish, Skagit and Whatcom counties @wsdot_sw - Traffic reports for Vancouver and southwest Washington @wsdot_passes - Mountain pass reports @wsdot_tacoma - Traffic and construction reports for Pierce, Thurston, Mason and Kitsap counties @goodtogowsdot - Good To Go! tolling information @snoqualmiepass - Snoqualmie Pass conditions and project info @wsferries - Ferry alerts and updates @wsdot_east - Traffic and highway news and information east of the Cascade Mountains @GoodToGoWSDOT - Washington state's toll system @BerthaDigsSR99 - Official account of the tunneling machine digging the SR 99 tunnel to replace Seattle's Alaskan Way Viaduct @wsdot_520 - Official WSDOT feed for 520 construction updates
Oregon	@OregonDOT - Official Oregon Dept. of Transportation Twitter account @MyOReGO - Road usage charge program of the Oregon Department of Transportation @TripCheckPDX - TripCheck Portland @TripCheckSalem - Tripcheck Salem @TripCheckEugne - TripCheck Eugene @TripCheckNCascd - TripCheck Cascades @TripCheckS_OR - Tripcheck S Oregon @TripCheckI_84 - TripCheck I-84 @TripCheckSE_OR - TripCheck SE Oregon
California	@CaltransHQ - The official Twitter of Caltrans @CaltransDist1 - Del Norte, Humboldt, Lake, and Mendocino @CaltransD2 - Counties of Shasta, Siskiyou, Trinity, Tehama, Modoc, Lassen, Plumas, and parts of Butte and Sierra @CaltransDist3 - Butte, Colusa, El Dorado, Glenn, Nevada, Placer, Sacramento, Sierra, Sutter, Yolo and Yuba @CaltransD4 - Bay Area @CaltransD5 - Monterey, San Benito, San Luis Obispo, Santa Barbara, and Santa Cruz Counties @Caltransdist6 - Fresno, Madera, Kings, Tulare and Kern counties

	@CaltransDist7 - Los Angeles & Ventura County @Caltrans8 - Riverside and San Bernardino Counties @Caltrans9 - Eastern Sierra Nevada and California @CaltransDist10 - Alpine, Amador, Calaveras, Mariposa, Merced, San Joaquin, Stanislaus and Tuolumne Counties @SDCaltrans - San Diego and Imperial Counties @Caltrans12 - Orange County