IMPRESSION FORMATION AND IDENTITY MANAGEMENT IN SOCIAL MEDIA

by

CHARLES FILE

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Communication, Information, and Library Studies

Written under the direction of

Dr. Marie Radford

And approved by

____

____

____

____

New Brunswick, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Impression Formation and Identity Management in Social Media

by CHARLES FILE

Dissertation Director:

Dr. Marie Radford

This project explores the connection between information-seeking strategies used in impression formation and on self-presentation in social media. The goal is to amplify and quantify earlier findings of a recursive relationship between seeking and providing interpersonal information in social media environments (Ellison, Heino, & Gibbs 2006). This study builds on two trends identified by researchers. Using Ramirez, et al.'s (2002) model as a guide, it notes that due to changes in the ways in which social media is presented and consumed, the active and interactive strategies – while still important – are becoming increasingly less dominant over the passive and extractive strategies (Antheunis, Valkenburg, & Peter, 2010). Second, it notes that the "vocabulary" of social interaction in social media has expanded to include a whole raft of indirect interactions (liking, sharing, emoji, etc.) that carry meaning and information in ways that differ from the primarily text-based overt communications that characterized social media previously (McEwan, 2013; Oh, Ozkaya, & LaRose, 2014).

This study supposes that there exists some connection between the increasing prevalence of passive and extractive types of social information seeking in social media, and the increasing prevalence of the use of indirect communication for the purposes of self-presentation and impression management. This study seeks to understand the nature of this connection by using a mixed-method approach in order to establish a holistic understanding of the interaction between information seeking, uncertainty reduction, and identity management in a single, cohesive theoretical structure. As such, qualitative and quantitative techniques were used to compose a model of information seeking and impression formation. Termed the “Identity Formation / Information Seeking (IF/IS)” Model, it is designed specifically for the investigation of behavior related to these phenomena in social media. It identifies five use types or “personas” that are both drawn from and applicable to quantitative and qualitative data. Further, this model identifies a set of variables that can be used to efficiently model behavior in each of these interaction modes. Finally, the model illustrates and measures changes in use patterns surrounding both information seeking (increasingly passive/extractive), and communication (increasingly indirect), providing some explanation for the convergent trends identified above.

Acknowledgement

Dedicated to Karen Maru, Ph. D.

The only reason I finished my degree at all is because I had the best dissertation committee chair on the planet: Marie Radford. I was very close to giving up many times; she believed in me when nobody – not even myself – did. The other committee members, Jeffrey Boase, Mary Chayko, and Matthew Weber were incredibly helpful and supportive. No other committee would have held a defense at 7:00 AM during a snow storm. Previous advisors Susan Keith, Mor Naaman, and Nick Belkin all helped me along as my interests wandered across disciplines. Thanks also are due Alex Galloway at NYU and Tim Murray at Cornell, my Master's and undergraduate thesis advisors, respectively.

I would like to recognize the financial support of the Rutgers Presidential Fellowship, the US Department of Homeland Security (DHS) Fellowship, and the Rutgers School of Communication and Information (SCI)'s Library and Information Science (LIS) dissertation support fund. Fred Roberts and Tamra Carpenter were responsible for the DHS Fellowship, and Ross Todd made the decision for the LIS dissertation support award.

I had support from a number of other short-term teaching and research contracts. Thanks to Chirag Shah, Paul Kantor, Mor Naaman, Mary Chayko, Anselm Spoerri, and Sharon Stoerger for these opportunities at SCI. In addition to SCI, a large part of the time, energy, and effort I spent at Rutgers was shared with CCICADA, where my DHS Fellowship was overseen. I was supported financially and intellectually by such

wonderful collaborators as Fred Roberts, Paul Kantor, Dennis Egan, and many others there.

I had support from the Communication Department at St. Peter's University in Jersey City, New Jersey. I am grateful to Barna Donovan and Cynthia Walker for inviting me to teach their graduate students.

Rutgers SCI was an amazing place to earn a Ph. D: a large part of the reason why I chose that department over others was because the students I met while visiting there seemed remarkably sane and normal compared to others I'd met. Thanks to everyone who helped make it that way, especially my friends at the Wine and Cheese Appreciation Club. Someday we'll have to actually try some cheese.

Without my family's constant support this project never would have happened. Joe File is the most patient man I've ever met, and that is no hyperbole. He saw me, my mother, and my brother through some of the darkest times anyone could have in life. Without him stalwart guardianship this family would have fallen apart years ago. Karen Maru was an intellectual and emotional inspiration, combining singular determination and grit with boundless compassion. A trailblazing feminist, a leader of industry, a full-time university professor, a Zen Buddhist, a fiber artist, a passionately attentive and caring mother – and this list doesn't begin to describe her. I miss her every day. Mike File has the heart of a lion. Fiercely loyal, protective, and sincere, his total lack of guile never ceases to amaze me, while his generosity and spirit never cease to inspire me. I hit the family lottery jackpot when I was born.

Jessica Crowell is the favorite person of just about everyone that’s met her, especially me. A paragon of positivity and love, she has an implacable capacity to see the goodness in everyone and every situation. She has a determination and drive that has led her to overcome innumerable trials, yet a grace that makes these victories appear effortless. Her trust and faith in me has many times been the only thing that’s kept me going through this process.

Table of Contents

List of Tables

List of Illustrations

Chapter One

Introduction

Problem statement

The purpose of this research is to study the ways in which users in a Social Networking Service (SNS) environment learn about one another; that is, how they form impressions of other individuals. Furthermore, to study how this learning process influences the ways in which individuals view themselves, and how this information-seeking process influences – in a conscious and deliberate way or otherwise – the types of decisions they make with respect to their own SNS use. Impression formation in Computer Mediated Communication (CMC) is a long-studied topic (see below) but the differences between SNS and older Information and Communication Technology (ICTs), as outlined above, add heuristically rich new possibilities for research. Specifically, the semi-permanent, semi-public nature of communication that takes place within an SNS means that there are far greater potential sources of information available to users when they want to form an impression of another based on their profile or past communications. (For the purposed of this study, SNS "profile" is defined by combining two early definitions as such: a public or semi-public self-authored collection of personal information that can include photos, video, audio files, and blogs [Boyd & Ellison, 2008; Kaplan & Haenlein, 2010]). So too do these factors influence how people use SNS for communication in general, and self-presentation in particular. Thus the two processes influence one another, and inasmuch as individuals are making anything beyond completely naïve judgments about their own and other's communications, the process of searching for information about others cannot help but influence the ways in which they present information about themselves. Researchers have long known about this phenomenon (Berger & Calebrese, 1975; Berger, et al., 1976) but the characteristics of SNS make these processes both much more explicit and salient to users, and much better recorded for researchers, thus opening up an array of research possibilities.

Statement of significance.

While both self-presentation and identity management have been extensively studied both within the CMC context and without, the interaction between the two represents a gap in the literature, particularly with regard to CMC. Further, SNS offer an environment well-suited to the study of the interaction between self-presentation and identity management, particularly due to the prevalence of SNS: 69% of American adults currently maintain a presence on at least one form of social networking, and 76% of Facebook users say they check the site at least daily (Greenwood, Perrin, & Duggan, 2016). This growing presence of SNS has rendered understandings of self-presentation and identity management inadequate, given that most of the extant literature in CMC is not concerned explicitly with SNS contexts.

Defining social networking sites.

Danah boyd and Nicole Ellison (2007) provided a first definition of SNS and did much to coalesce a nascent and inchoate smattering of literature into something properly resembling an academic discipline. As such, their foundational text “Social network sites: Definition, history, and scholarship” provides as good a place to begin this discussion. Their original definition was:

> We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system (p. 231).

This definition provided a generally accurate one for SNS at the time, and was the de facto standard for several years. However, as both the technologies and patterns of use around SNS began to change, this definition became less useful for several reasons, many of which Boyd and Ellison (2013) note in their revision. Primarily, the notion of a profile as a generally static representation or projection of one's self is a less common feature of SNS. Examples abound (most notably Twitter) of SNS that incorporate little to no such profile information. More commonly SNS that do incorporate profile pages contain an aggregated stream of the residue of interaction, communication, and sociability. Dubbed by Naaman and Boase (2010) "social awareness streams," these types of pages, that aggregate and display a history of connected social interaction that an individual user has engaged in on the site, have largely replaced the more traditionally static "profile page." Various SNS have various levels of privacy controls that help (to some extent) determine who is able to access these data, but in (almost) all cases there is some potential for an "outside" audience (that is, those not directly engaged in the interaction) to traverse a record of a user's communication history. It is this feature that separates modern SNS from other CMC technologies that existed for decades prior to 2007, like email or a private message boards.

In this sense, one can consider the “social network” in SNS as both a noun and verb. An SNS allows users to both create a social network and to engage in the process of social networking. Boyd and Ellison’s (2007) early definition focused almost entirely on SNS as noun. It identifies what an SNS is, but did not address what an SNS did; that is, what types of communication patterns were afforded by the new technology. When it was introduced, email, being both instantaneous and asynchronous was qualitatively different from existing communication technologies, opening up whole new use and interaction patterns. Social media, being instantaneous, asynchronous, and (semi-)public is again entirely different from email, and again opens up whole new ways for communication to emerge. It is the addition of this third adjective – the (semi-)public record of the communication that can be browsed by other users of the SNS, whether they were directly involved in or addressed by the communication or not, that is the differentiating factor, and also the most interesting one from a research perspective. In addressing these concerns, boyd and Ellison (2013) provided an updated definition:

> A social network site is *a networked communication platform* in which participants 1) have *uniquely identifiable profiles* that consist of user-supplied content, content provided by other users, and/or system-level data; 2) can *publicly articulate connections* that can be viewed and traversed by others; and 3) can consume, produce, and/or interact with *streams of user-generated content* provided by their connections on the site. [Emphasis original] (p. 158).

This definition better articulates the differences in communication patterns afforded by SNS, in contrast to existing communication technologies. Considering point

1: though previous communication technologies provided for asynchronous group communication – message boards, for instance – SNS allow this content to be aggregated and viewed by each user and for each user, such that each has their own distinct feed of information both displayed to them and displayed about them. Point 2 is something of a sine qua non of SNS, and perhaps their original defining feature. It is what makes SNS different from other forms of social communication technologies like blogs. Though they create, in a sense, the type of individualized content stream discussed with respect to point 1, they do not permit (at least nearly so easily) the type of immediate social transmission of that information to the content stream of other users. Finally, point 3 addresses how SNS differ from email, which does afford the immediate social transmission of information to the content stream (read: inboxes) of others. It is the ability to "consume, produce, and/or interact with streams of user-generated content," (boyd & Ellison, 2013, p. 158) regardless of the awareness or intentionality of the original producers.

This definition and the call for research that appeared alongside it sparked a wave of research in to SNS. Across a variety of fields, researchers began examining the ways in which new technologies specifically designed for the projection of digital identities across electronic media were changing how identity is created, maintained and interpreted in online settings. The influence of this work has informed the present study; the most germane is summarized below.

Literature Review

Among the earliest identifications of the recursive and cyclical interaction between self-presentation and information seeking in social media systems in the CMC literature is a series of studies of online dating led by Ellison, Gibbs, and Heino and their colleagues (Ellison, Heino, & Gibbs 2006; Gibbs, Ellison, & Heino 2006; Heino, Ellison, & Gibbs 2010; Gibbs, Ellison, & Lai 2011; Ellison, Hancock, & Toma 2012). Specifically in the 2006 study, the authors state that their research subjects had a number of set strategies to gauge the veracity of information presented in the profiles of others on an online dating site, and that these strategies would influence their decisions when setting up their own profiles. They note, "The twin concerns that resulted from these factors — the challenge of establishing the credibility of one's own self-descriptions while assessing the credibility of others' identity claims — affected one another in a recursive fashion" (Ellison, Heino, & Gibbs 2006, p. 420). For example, "some participants constructed rules of thumb for assessing others (e.g., an inactive account indicates a lack of availability or interest) while simultaneously incorporating these rules in their own messages (e.g., frequently making slight adjustments to the profile)" (Ellison, Heino, & Gibbs 2006, p. 432). Online dating represents an attractive venue for an examination of the process of recursive adjustment of self-presentation: it is difficult to imagine an environment in which users searching for information about others as extensively, while simultaneously managing the information the present about themselves in so consciously strategic a manner. It is therefore perhaps fitting that the phenomenon would be first there identified in a CMC context. It is arguable that this recursive self-presentation adjustment process is not unique to a dating environment; it is perhaps amplified in importance there, but this can be viewed as a difference of degree rather than of kind with respect to other SNS. This process is manifest in many – perhaps all – SNS environments, and is a necessary result of their characteristics as communication platforms. Specifically, again, these characteristics are the openness of communication by one user to many others, and the enduring nature of those communications. These characteristics create an environment in which users can learn about others while managing what others can learn about them in a uniquely conscious and strategic way.

Foundational theories.

This literature review will next discuss existing, foundational theories of self-presentation, identity management, and information seeking in CMC, and then discusses each of these topics in greater detail.

Social psychologists have documented a strong link between self-presentation and identity. Much of this work stems from that of Erving Goffman and his foundational book *The Presentation of Self in Everyday Life* (1959). In this work, Goffman relies on dramaturgy as a metaphor for social interaction, with each actor in the social sphere taking the role of performer, audience, or outsider. More recently, this work has been extended to develop links between self-presentation and different types of self-knowledge. For instance, researchers have found that by asking individuals to portray various personality types, such as introversion or extroversion, they could manipulate how individuals viewed themselves in that domain (Fazio et al., 1982). Researchers have demonstrated similar links between self-presentation and other types of self-knowledge, including self-concept (Tice, 1992), self-appraisals, (Schlenker & Trudeau, 1990), and a sense of personal autonomy (Schlenker & Weigold, 1990).

Researchers have also noted the close relationship between the publicness of a social performance and its effect on identity self-knowledge. For example, Tice (1992) varied the publicness of a self-presentation by manipulating factors such as the presence of an audience, or personal information about the individual. Participants were then asked to portray an introverted or extroverted individual as in the Fazio et al., (1982) study. Those in the public conditions reported actually possessing the assigned trait to a greater degree than those in the private condition. In a more recent study, Kelly and Rodriquez (2006) asked participants to create two conflicting self-presentations on video tape, one introverted and the other extroverted. Participants were told that only one tape would be viewed by others, and it was only the personality traits presented on that tape that participants integrated into their self-concepts. This early finding underlines the importance of the physical act of self-presentation as a key moment in identity building and maintenance.

Another major theme of SNS research that informs this study is the nature of information seeking in online environments, particularly as has to do with uncertainty reduction with respect to impression formation. That is, questions of how individuals get to know one another in CMC, and what effects the channels have on this process. Early theoretical approaches failed to consider the active role of individuals in the communication process, much less their information-seeking behaviors, opting instead to emphasize channel effects. The theories collectively described as “cues filtered-out” by

Culnan and Markus (1987) posit that one does not form impressions of others online, but due to the flattening effects of the medium on message richness, users' attention is turned away from others towards the self and the task. As a result, these approaches assume that due to the starkly reduced number of social cues available in online environments, the ability to acquire social information and to form impressions of others is likewise limited.

Later theories would largely reject these hypotheses about the influence of a "reduced cues" environment on social information seeking and uncertainty reduction. A number of more recent approaches argue that individuals engage in cognitive deliberations and communicative behavior to compensate for media limitations. Lea and Spears (1992; 1995) introduced a model called the Social Identity Model of Deindividuation effects (SIDE) based on self-categorization theory (Tajfel & Turner, 1986) which held that the individual identity is composed of a range of self-categories including both social and personal identities. SIDE posits that the extent to which group or social identity is more salient will have an impact on the types of attributions and perceptions made about a target. That is, the deindividuating effects of, for instance, anonymity in CMC, will have varying effects depending upon the extent to which the features of the channel and the nature of the interaction make salient the embodied, offline self.

Walther proposed two additional models that also disputed some of the core assumptions of the cues filtered-out approaches. These include the social information processing theory (SIPT) (Walther, 1992; Walther & Burgoon, 1992) and the hyperpersonal perspective (Walther, 1996). SIPT proposes that social actors in CMC adapt to the limited cues environment and are able to infer a good degree of social information, despite the limited set of cues, like eyes adapting to seeing in the dark, perhaps. For instance, Walther demonstrated that CMC users were able to infer social information from the metadata created by others' use of the system itself. For instance, Walther (1992) also demonstrated that chronemic factors, which have to do with the speed of a reply, the order messages were received, and so on, can be used as a source of social information. Walther's (1996) hyperpersonal perspective extends SIPT by focusing more explicitly on the processing of information sought and given online. In this research, Walther noted that users tended to overestimate the attractiveness of their online relational partners relative to people they knew offline, making dialogue partners in CMC more socially desirable than those in face-to-face interaction. He offers several explanations for this phenomenon: that the lack of visual cues allows users to freely idealize their partners; that others have the ability to freely and strategically choose which aspects of themselves to disclose, leading to a slanted information being received from them; that others have the ability to apply increased attention to message formation, allowing for more effective communication; and that these factors may combine such that computer-mediated messages show more self-awareness and introspection.

Moving beyond core theoretical models, the rest of this review will focus on research themes germane to the topic at hand: the recursive and cyclical interaction between self-presentation and information seeking in social media systems. These topics are: self-presentation, with particular focus on how this is accomplished in social media profiles; how users manage their identities while interacting with others in social media; uncertainty reduction in social media, with particular focus on the information seeking strategies that individuals use to learn about others, including a discussion of warranting theory.

Self-presentation in SNS: profiles.

Researchers have identified a number of ways that self-presentation can occur through the creation of online profiles in SNS: by conveying messages of social maintenance and awareness (Donath, 2007), by conveying personal tastes (Liu, 2007), or by using the social network as a contextualizing structure to convey information about the individual based upon their place in it (boyd, 2006). But perhaps the most common use of profiles in SNS is to convey personality information. Counts and Stetcher (2009) reported an experiment wherein subjects created a hypothetical online profile in a step-wise manner. Participants were then ask to rate their profile according to the personality traits it exhibited, as measured by the Ten-Item Personality Inventory (TIPI) (Gosling, Rentfrow, & Swan 2003), a Likert scale measurement of the “Big Five” personality traits: openness, conscientiousness, extroversion, agreeableness, and neuroticism. Subjects then also completed personality inventories of their imagined ideal self. Counts and Stetcher reported that after completing four items on the profile, correlation between these two measures was over .6, and after thirteen items was over .8, indicating a relatively strong relationship between personality type a measured by a psychological assessment and personality type portrayed in the hypothetical profiles. They reported that free-form test areas (as opposed to limited-response fields like “location” or “political party”) carried by far the most personality information. Finally, the researchers surveyed their subjects using an instrument based on work by William James (1890) designed to discover the type of information they wanted to portray in their profile. Over 90% reported the desire to portray their spiritual self (perceived abilities), 25% desired to portray their social selves (friendships, occupations) and only one respondent desired to portray their material selves (physical attributes). This finding comes in stark contrast to others, which have focused on the more social aspects of SNS (e.g., Donath, 2007; Liu, 2007; boyd, 2006; Donath & boyd, 2004) and highlighted the importance of social context in SNS. Counts and Stetcher (2009) therefore underline the points made in the discussion earlier with respect to the use of profiles as a defining characteristic of SNS: the profile is a place of personal expression primarily related to the individual and their personality. The truly social aspect of SNS occurs, at least for 75% of subjects in Counts and Stetcher (2009), in some other affordances of the medium.

Other research has shown a strong link between profile creation and identity. Stern (2004) goes so far as to parallel the process of building a personal web page to the very act of identity creation, in that each portion of the site appears to be carefully constructed so as to portray a particular impression of the self. Gosling, Gaddis, and Vazire (2007) rated Facebook profiles according to their perceived personality on the TIPI scale, and compared the results with assessments given by the individuals those profiles represented. They found that Facebook-based personality impressions showed some consensus for all Big Five dimensions, with particularly strong consensus for extraversion. In an earlier iteration of the study (Vazire & Gosling, 2004) that examined personal websites rather than profile pages, the same researchers were able to demonstrate a similar result.

The connection between the publicness of a self-presentation and its impact on identity and self-knowledge was noted above. Researchers have been able to replicate Goffman’s (1959) assertion that public performance of a self-presentation is tightly linked to the creation of a concomitant identity in both face-to-face (Tice, 1992) and video (Kelly & Rodriquez 2002) contexts. Gonzales and Hancock (2008) conducted a series of experiments to extend these findings into an explicitly CMC environment. In their experiment, they manipulated both the personality being presented (introvert or extrovert) and the publicness of the presentation (writing in a word processor or writing on a putative blog). Echoing earlier findings, Gonzales and Hancock found that individuals who portrayed both personality types in the public setting rated themselves more extroverted or introverted (depending upon their experimental condition) than those who wrote in the private condition.

This research demonstrates that the links that exist between self-presentation and identity are present in CMC, and underlines the finding that this link is amplified by the publicness of that self-presentation. While SNS are certainly not unique with respect to other ICT in that communication that takes place within them can be public, they are unique that much or all of this communication is – to some degree – public. This is an important defining characteristic of SNS, and as such they provide a useful advantages when conducting research on self-presentation.

Impression management: interaction.

Based on boyd and Ellison’s (2013) definition, a defining characteristic of social media systems is the ability for individuals to interact with and learn about others in ways unavailable, or at the very least not nearly so easily, in previous CMC channels. Specifically, the ability to look up and review past communications of individuals, or to monitor their current conversations unobtrusively via a social awareness stream. Building upon this insight, based upon the research of Ellison, Heino, and Gibbs (2006) that the awareness of this affordance must surely engender in SNS users a concordant awareness of the potentially wide audience of their own communications, which must have an influence on their content, tone, message, etc. While this specific "recursive" relationship between information seeking and self-presentation is not yet fully explored, CMC researchers certainly have investigated impression formation and management in general in both CMC and SNS environments.

Early research in CMC focused generally on the question of reduced cues and the influence that this had on the ability for communicators to engage one another, and the ability to be understood. O'Sullivan (1996) proposed an "ambiguity-clarity" dialectic between "the contradictory and interdependent needs to create ambiguity and clarity about the self and a partner in relational interactions" (p. 12). He asserted that openness and closed-ness were not patterns themselves, but rather represented a continuum along which users could exist as the channel, the communicative task at hand, and their temperaments suited them. Based on these insights, he surmised that varying levels of media richness (Daft & Lengel, 1986; Trevino, Lengel, & Daft, 1987) – defined by the presence or absence of social cues – could provide various benefits depending on the communicative task. This sort of approach to the study of CMC focused on issues like media richness, deception, and impression management and characterized much of the early research in the field (e.g., Camden, Motley, & Wilson, 1984; Lippard, 1988; Chovil

1994; Coupland, Giles & Wiemann, 1991; Parks 1982). In a study investigating these claims, O'Sullivan (2000) found that, contrary to the predictions of media richness theory, individuals would sometimes intentionally suggest leaner channels specifically in order to increase the potential ambiguities, in order to match interactional needs. Further, he found that "Specifically, situations in which individuals seek to regulate very closely what information is exchanged – such as when self-presentations are on the line – are when mediated channels can provide advantages over face-to-face" (O'Sullivan, 2000, p. 21).

Later researchers, drawing upon these results, began to investigate in greater detail the ways in which lean media might be used by individuals to modify, manage, or influence their self-presentations. Turkle (1984) perhaps first popularized the notion that online interactions might offer individuals the opportunity to explore normally hidden aspects of their personalities, or to try on whole new ones. This insight, which is related in some ways to Walther's (1996) hyperpersonal model discussed above, is based on research that has demonstrated potentially negative effects of truly honest self-presentation in social situations (Goffman, 1959; Rogers, 1951) and potential costs to revealing taboo or negative aspects of one's personality (Pennebaker, 1990; Derlega, Metts, Petronio, & Margulis, 1993). This has lead Turkle, Walther, and others to propose (and others to test in work including O'Sullivan, 2000) that the very absence of social cues, the presence of anonymity – the very ambiguity of it all in CMC – may have certain advantages in the realm of personality expression. Much of the research based on this work employs a framework developed by Higgins (1987) that postulates three domains of the self: the actual self – one's basic self-concept – the ideal self – that persona to which one aspires – and the ought self – that persona that one feels social pressures to portray. Much of this research focuses on how the lack of cues and the relative anonymity of CMC allow individuals to express themselves in a fashion closer to their true selves.

For example, Bargh (2002) had dyadic pairs interact in either CMC or face-to-face contexts over the course of several experiments. Prior to each experiment, he used rapid recall of character traits in order to determine the accessibility of personality characteristics based on the Me/Not-Me procedure of Markus (1977). He also asked subject to list aspects of themselves they present to others in public (their “actual selves”) and aspects they possess but generally do not portray (their “true selves”). After each interaction, he asked subjects to describe the personalities of their partners. In one experiment, he found that pairs’ descriptions of one another in the CMC condition aligned more closely with self-described “true selves” than in the face-to-face condition, demonstrating that individuals are better able to act in accordance with their less socially-ascribed personality in CMC. In order to test the reverse directionality of this result, he repeated the process but had subjects describe their ideal (non-romantic) interaction partner, or friend. He found that after interactions subjects ascribed more of these characteristics to their partners in CMC, and surprisingly ascribed almost none in the face-to-face communication, indicating that the reduced social pressures of the CMC allow for more “true” attributions to both self and other.

An important point to consider is also the extent to which such online interactions are genuine expressions of identity. Research has demonstrated that online groups are often considered to be genuine communities by their members in part because it gives those who use it a strong “sense of place” (Meyrowitz, 1985). By providing a venue for the sharing and telling of stories and experiences, CMC provides a mechanism for “transporting” (Biocca & Levy, 2995) individuals who share similar viewpoints and thoughts into shared environments. These shared stories provide a powerful mechanism by which relationships are established, to the extent that such digital environments are often characterized as neighborhoods or communities (Kim & Biocca, 1997). Such language become even more prevalent in the realm of SNS, as in: “Facebook helps you share and connect with the people in your life.” (Parks, 2011, p. 106). Thus the notion of the neighborhood or community provides a communal metaphor via which individuals are empowered to make their collective experience visible and real.

The implications of findings like these with regard to social media specifically are difficult to establish precisely. On the one hand, if the CMC findings like those of Bargh hold, one might assume that users of SNS would be more in touch with their “true selves” and make connections more easily in social media. On the other, if the “publicness” implications of research discussed in the previous section like that of Gonzalez holds, then one might expect, based on the public nature of communication in SNS that users would see their existing, socially mediated “ought” or “actual” selves amplified rather than reduced. These types of conflicting results make predictions of in-the-wild SNS behavior based on laboratory results difficult.

Indeed, researchers studying SNS in that very wild have, perhaps to be expected, discovered that behaviors are quite a bit more complex outside the laboratory. In negotiating the apparent conflict between the self-revealing benefits of anonymity and the public demands of the medium, SNS users adopt a number of strategies when engaging in impression management via interaction. Lange (2007) describes the many strategies in which YouTube users engage in order to maintain what he describes as a “fractal” sense of public/private spaces. Adapted from Gal (2002), Lange employs an analogy of the home, in which one might feel private in a bedroom relative to the living room, but when in the living room private relative to the street outside. Lange describes interaction strategies that are “publicly private” (Lange, 2007, p. 11) in which only a subset of the YouTube audience that is “in the know” is able to find a user’s video, or “privately public” (p. 12) in which video sharers make their content widely available, but are less forthcoming with personal details, candid thoughts, and their “true selves.” In sum, it appears that, much as O’Sullivan detailed, users adopt a number of strategies when interacting with others, and often intentionally use ambiguity and obscurity in order to manage the information conveyed in those interactions.

In another ethnography of social media, Marwick and boyd (2011) studied Twitter users in order to discover how they manage their identity via their interactions with others. Marwick and boyd were particularly interested in how Twitter users approach their “audience,” and how they stay “authentic” to that audience. They note that Twitter, like many other SNS, collapse all the social networks an individual has (work, family, church, etc.) into a single large one. They discovered that in the face of this “context collapse” (Marwick & boyd, p. 10), managing identity via communication and interaction requires a good deal of strategy. Employing Goffman’s (1967) notion of symbolic interactionism, they discovered that “Twitter users negotiate multiple, overlapping audiences by strategically concealing information, targeting tweets to different audiences and attempting to portray both an authentic self and an interesting personality" (Marwick & boyd, 2011, p. 10). These results echo those of Lange, and highlight the extent to which identity management is linked closely to notions of audience and publicness. In SNS, the open, public, and recorded nature of communication – its availability for later retrieval by any potential information seeker – seem to strongly influence the strategies users employ when presenting themselves. Again, this research underlines the recursive nature of information seeking and identity management in SNS.

Uncertainty reduction and information seeking.

The discussion now shifts towards a more explicit information-seeking task, the process by which SNS users actively search for information about other users of the service. Impression formation as an information-seeking process is typically discussed in the CMC literature as "uncertainty reduction" and is conceptualized as a cognitive process by which new acquaintances attempt to increase mutual understanding by observing and communicating with one another. A lack of knowledge about other social actors creates an information need, which in turn cues these individuals to engage in information-seeking behaviors. These behaviors are comprised of a number of strategies by which individuals gather information about other social actors in order to decrease their levels of uncertainty (Berger & Calabrese, 1975; Berger, et al., 1976).

Ramirez, Walther, Burgoon, and Sunnafrank (2002) categorized these social uncertainty-reduction information-seeking strategies as that occur specifically in the realm of CMC into four basic types: interactive, active, passive, and extractive. Interactive strategies are those that involve direct interpersonal communication between the information seeker and the individual he/she is attempting to learn more about. Active strategies involve collecting information on the target without direct interrogative interaction with that target. In a face-to-face environment, these types of strategies were, to a large extent, the only ones available outside a direct interrogatory encounter with the target. However CMC environments in general, and SNS in particular, offer other avenues. Extractive strategies involve looking back at the record of stored communications that a target has made in order to gain information about how the target behaved in social situations. Finally, passive strategies involve unobtrusive monitoring and evaluating of a target's social interactions as they occur ("lurking," perhaps).

One specific interactive strategy bears a bit more discussion: self-disclosure. As an uncertainty-reducing strategy, self-disclosure has been shown to have a strong reciprocity expectation (Goffman, 1959). In face-to-face communication, research has shown that this expectation of self-disclosure reciprocity is used as an uncertainty-reduction strategy: social actors would disclose some personal detail about themselves, and this would, in turn entice the interlocutor to reciprocate (Berger, et al., 1976). These findings have been replicated in CMC contexts, and are used by researchers to elicit more personal information from research subjects (Joinson, 2001). From this perspective, it is reasonable to conclude that self-disclosure – an intimate variation of self-presentation – is, in fact, an interactive information seeking strategy. This conclusion again highlights the reciprocal nature of self-presentation and information seeking that Ellison, Heino, and Gibbs (2006) first identified.

Writing in 2002, Ramirez, et al. allowed that it may be “difficult to imagine” (p. 216) how passive information seeking strategies might be employed in a CMC environment. At the time of their writing, the dominant online medium was e-mail, a channel in which unobserved information gathering or “lurking” was difficult, at best (though certainly possible in certain cases – such as email listservs and online discussion boards). Contemporary authors working within the same framework tended to find results in line with these expectations. Tidwell and Walther (2002) compared uncertainty reduction strategies in a CMC environment to a control group in a face-to-face environment. They discovered “CMC interactants appeared to employ a greater proportion of more direct, interactive uncertainty reduction strategies—intermediate questioning and disclosing with their partners—than did their FtF counterparts” (Tidwell & Walther, 2002), highlighting the importance of interactive strategies in CMC environments compared with face-to-face. They found that these interactions tended to be more intimate, again demonstrating the degree to which self-disclosure can act as an information-seeking strategy. Though Tidwell and Walther found the interactive strategy to be the most important in CMC, they were, again, to some extent bound by the limitations of the medium.

More recent studies that deal more specifically with SNS have altered the picture. Antheunis, Valkenburg, and Peter (2010) found in that SNS users were most likely to learn about others using passive techniques like lurking and monitoring the social awareness stream. The researchers posit this is due in large measure to the nature of SNS, which often include detailed profiles that encourage self-disclosure. This study did not specifically test extractive strategies as different from passive, so their prevalence remains an open question. Interestingly, this study also found that although SNS users employed a number of information-seeking behaviors, only the interactive uncertainty reduction strategy was effective. The authors offer no speculation, and thus the efficacy of each strategy also remains an open question.

Other studies have found differing results in different contexts, however. Extending the findings of their 2006 study, Gibbs, Ellison, and Lai (2011) conducted surveys of online dating site users in order to establish potential links between the use of uncertainty-reduction strategies and self-disclosure. The authors also examine privacy concerns, self-efficacy, and internet experience as potentially variables that may potentially be mediated by uncertainty reductions strategies, as they relate to levels of self-disclosure. The authors found that privacy concerns and self-efficacy did predict the use of uncertainty reduction strategies, which in turn led to greater self-disclosure. Based on these findings, they assert that information reduction may be "an important middle step in reducing privacy concerns and providing verification of identity claims that then provokes self-disclosure in online contexts" (Gibbs, Ellison, & Lai 2011, pp. 91). The authors further found that most of these information-seeking strategies were of the interactive variety, though they note that this may be due to a contextual specifics of online dating sites, which limit the availability of information that can be used to investigate another user (e.g., their name or email address).

Warranting.

In tracing the conceptual links between information seeking, through self-disclosure, and back to information seeking again, this review has examined the various ways researchers have grappled with questions regarding what influences self-disclosure in SNS environments. To complete this discussion, however, it is worthwhile to consider factors that influence information seeking. One theory that has been proposed by Walther and Parks (2002) is the warranting principle. Based on the work of Stone (1995), they proposed that in online environments that allow for some level of anonymity, the connections between self and self-presentation exist along a continuum, rather than a binary. Reasoning that anonymous environments can allow for discrepancies between the self and self-presentation, Walther and Parks hypothesized that the greater the potential for these discrepancies, the more likely users would be skeptical of any information they received. Warrants are those social cues that are perceived to provide reliable evidence that identity claims made online are valid. Warrants have a varying degree of warranting value inasmuch as they are perceived to be difficult to manipulate. Those warrants that are difficult for the claimant to alter (particularly in Walther and Park’s formulation the opinion of third parties) are perceived to be more reliable than those which are more easily altered (such as first-person claims).

Much of the extant literature on warranting has focused on examining the influence of other-generated information. For example, Walther, Van Der Heide, Kim, Westerman, and Tong (2008) studied the effect of warranting cues on levels of social attraction. Facebook profiles were generated that were neutral in content but contained two comments, either positive or negative in nature. The comments were accompanied by profile pictures that were either attractive or unattractive. The researchers found a strong correlation between levels of social attraction to the profile page and the physical attractiveness of the commenters, indicating that the mere presence of these cues influenced the judgments of participants. Interestingly, the actual content of the comments yielded inconclusive results.

In a follow-up study, Walther, Van Der Heide, Hamel, and Shulman, (2009) more directly compared the relative warranting value of self-generated versus other-generated cues. In a pair of experiments, they created Facebook profiles that contained generated personas designed to be introverted or extroverted in the first study, and attractive or unattractive in the second. These profiles also contained warranting cues in the form of comments made by putative friends. In the first study, while other-generated statements did have some effect on judgments, the self-generated statements had a more powerful effect. In the second study, the results were reversed: other-generated statements had the greatest effect. Walther, Van Der Heide, Hamel, and Shulman propose that when making judgments about internal claims, like introversion or extroversion, participants weighed self-generated claims more heavily, however when gauging external claims, they weighed other-generated claims more. This claim remains to be thoroughly tested, however.

Community.

Until now this survey of the literature has focused on characterizing the features of SNS and how they are described and defined. The discussion now turns from the “what” towards the “why.” A number of researchers have examined the ways in which social media are used by individuals, and their motivations. Among these are senses of community, identity, and presence.

Community membership is frequently cited as a strong motivation for online membership. Communities provide for their members regular, patterned, and personalized social engagement, as well as shared engagement and culture. These help individuals feel connected in a meaningful way, and provide shared feelings of togetherness and belonging. Particularly in online settings, ideas of community have been central themes that researchers have discovered. (Parks 2011, Hampton & Wellman, 2003). Chayko (2008) notes that when asked to characterize the nature of the social groups that they encounter or form online, most deployed the term "community." Furthermore, they described these communities as close and meaningful: "My group is an extremely tightly bonded community that simply cannot be found in normal daily life," (Chayko, 2008).

It is important to note that this finding is not universal. Tufekci (2010) points out that here are some individuals that are more likely, "to accept online friendship formation as possible, or even desirable," (Tufekci, 2010, p. 176). But for others, "face-to-face interaction has inimitable features that simply cannot be replicated or replaced by any other form of communication," (Tufekci, 2010, p. 176). This provides an important counterpoint to the notion that SNS provide a substitute for, or replacement of, other forms of interaction. Tufecki's research demonstrates that it is important to be careful when generalizing research regarding the online community to broader application.

Yet for many, research has clearly demonstrated that online groups are commonly referred to and experienced as communities (Baym, 2010; Parks 2011). Furthermore, these online communities provide many of the same mechanisms for social cohesion that other social networks provide. Hampton and Wellman (1999) assert that social networks are pathways along which support, empathy, and resources can flow. They demonstrate that online communities provide mechanisms via which these networks can be maintained and reinforced. In their study of these online groups, Hampton and Wellman (1999) assert that these “communities are clearly social networks,” (p. 648). This insight provides the reasoning by which research into SNS specifically can have application to broader themes of online community.

Connectedness.

Research has indicated that an important aspect of social media use is the feeling of, or awareness of, connectedness to one’s social group. For certain users of social media, the feeling of being always-on or constantly plugged-in can provide individuals with a strong sense of connection to a group, or membership in a community much larger than oneself. (Baron, 2010). Several researchers have noted that the anxiety over being disconnected can, itself, be a motivating factor in social media use. Termed “the fear of missing out,” those accustomed to constant connection and availability can report feelings of anxiety and being overwhelmed by the apprehension of disconnection (Bawden & Robinson, 2009; Przyblyski, et al., 2014). Digital technology can be a community-building, uniting mechanism in some circumstances, a space of alienation for others, or a compulsion for others. Therefore, there are both positives and negatives to be found that accompany continuous digital connectedness.

On the upside, there are powerful gains to be felt from integration into online communities, including access to information, emotional support, feelings of connectedness, and sense of belonging. Research has shown the power of these online communities to support individuals during emergencies or crisis (Shklovski et al., 2008, 2010). Even without the imminence of danger, online communities can provide the means of comfort knowing that others in the community are available and care for others' well-being (Castells, 2011; Ling 2014). Furthermore, online communities provide a means by which individuals that might feel excluded from their physical social circles can find inclusion in virtual ones. Research has shown that those who feel marginalized or discriminated against can often find inclusive spaces in online communities. Online communities that individuals create with like-minded or similarly-identifying individuals can be more open and less judgmental than those that exist in a physical space (Hillier et al., 2012; Daniels 2009).

Digital technology can feel burdensome, too. Keeping up with the sheer volume of information, tasks, appointments, and obligations to others can feel taxing and overwhelming. Feelings of information overload are commonly reported in the literature, combined with frequent decision dilemmas regarding which pieces of information are the most pertinent, and most immediately important (Katz & Aakhus, 2002). Further, digitally connected individuals must consider how and when to approach differing audiences in differing contexts (Marwick & boyd, 2011).

Research Objective and Significance

While both self-presentation and identity management (Goffman, 1959) have been extensively studied within the CMC context and without, the interaction between the two represents a gap in the literature, particularly with regard to CMC. Further, SNS offer an environment well-suited to the study of the interaction between self-presentation and identity management, particularly due to the prevalence of SNS: 69% of American adults currently maintain a presence on at least one form of social networking, and 76% of Facebook users say they check the site at least daily (Greenwood, Perrin, & Duggan, 2016). This growing presence of SNS has rendered understandings of self-presentation and identity management inadequate, given that most of the extant literature in CMC is not concerned explicitly with SNS contexts.

This project is designed to test the relative magnitude of the effect that various information-seeking strategies have on self-presentation, and in particular the passive and extractive strategies. The goal is to amplify and quantify earlier findings of a recursive relationship between seeking and providing interpersonal information in SNS environments. Earlier research has looked at the effects of information seeking on self-presentation (e. g., Tidwell & Walther, 2002; Westerman, et al., 2008; Gibbs, Ellison, & Lai, 2011), though certain gaps in the literature remain. Specifically, while there has been ample study of active and interactive information seeking strategies (e.g., O'Sullivan, 2000; Bargh, McKenna, & Fitzsimons, 2002; McKenna, Green, & Gleason, 2002; Yee & Bailenson, 2007; Davies 2007), the literature is relatively light with examples of the effects of passive and extractive information-seeking strategies. This study attempts to fill this gap in the literature by focusing on the study of passive and extractive strategies in its design. In so doing, the effects these strategies on self-presentation can be directly studied. Further, examines the type and valence of the information participants are exposed to, in order to gain further insight based on the predictions of warranting theory.

It is suggested here that due to changes in the ways in which social media is presented and consumed, the active and interactive strategies – while still important – are becoming increasingly less dominant. While specific data on the relative frequency of different types of browsing behavior is scarce, exceptions include work by Gibbs, Ellison, and Lai (2011) which found some use of passive strategies in a minority of their sample. Antheunis, Valkenburg, and Peter (2010), had similar findings in their study of uncertainty reduction. This study seeks to explore the concept that increasingly the nature of social media interaction involves the passive, occasional browsing of an ongoing stream of social trace data, rather than the purposive, directed, active dyadic interaction that is the more common subject of experimental research in CMC on uncertainty reduction, attribution, and warranting.

Further, this gap in understanding is due not solely to changes in the technological and social context in which social media interactions take place, but is also partly due to the experimental method that much of the research concerning this topic (as summarized above) takes. Investigating passive, undirected, unintentional information seeking and interaction behavior is a difficult thing to study using these methods, which could also partly provide an alternative explanation for lack of coverage in the literature.

Finally, not only has the nature of information seeking in social media changed due to new features available in the platforms that provide it, self-presentation and identity management have also been affected by these and other factors. What specifically constitutes interaction in the context of SNS has grown from the sharing of text and images to liking behaviors, sharing or reposting, tagging and linking, using emoji and icons, and so forth. That is, the “vocabulary” of social interaction in SNS has expanded to include a whole raft of interactions that carry data, meaning, and intentionality in ways that are inherently tacit and indirect when compared to the primarily text-based overt communications that characterized SNS during their formative years (McEwan, 2013; Oh, Ozkaya, & LaRose, 2014).

It is the undergirding assumption of this project that there is some link between these two trends. This study supposes that there exists some connection between the increasing prevalence of passive and extractive types of social information seeking in SNS environments, and the increasing prevalence of the use of tacit, indirect, communication for the purposes of self-presentation and impression management. This study seeks to understand the nature of this connection by taking on three tasks: First, to measure and understand the importance of passive and extractive information seeking strategies as an approach to uncertainty reduction, attribution, and impression formation in social media. Second, to measure and understand the importance of passive and extractive information seeking strategies as an approach to self-presentation and identity management in social media. Lastly, to determine what link, if any, exists between the first and second trends.

Based upon these goals, and following from the above literature review, the following research questions (RQ) have been developed:

RQ 1. What strategies or procedures are used in the maintenance of online identities in SNS?

RQ 2. How are online identities managed in SNS?

RQ 3. What is the relationship between information seeking behavior and self-presentation and identity management in SNS?

RQ 4. What is the role of indirect forms of communication, such as liking and following behaviors, in the creation of online identity in SNS?

RQ 5. What is the role of indirect forms of communication, such as liking and following behaviors, in the maintenance of online identity in SNS?

Conclusion

This chapter introduced the research problem being studied, and outlined the factors impacting it. It discussed the foundational theories upon which this dissertation stands, as well as reviewing key literature in the area upon which this study builds. Five research questions were presented, which served to drive the development of the research model outlined in subsequent chapters. The details surrounding this development are described in the next portion of this dissertation, Chapter Two.

Chapter Two

Methods

Procedure

To address the research questions listed in Chapter One, a mixed method approach will be adopted in this study. Mixed method approaches allow for a more nuanced understanding of the processes involved in the participants, particularly germane to the current topic with respect to the levels of intimacy of self-disclosure that individuals are making (Connaway & Radford, 2017). Such approaches also more readily lend themselves to inductive reasoning. Finally, given the topic of this dissertation, the use and analysis of extant social media data seems a natural step. The most compelling rationale for this approach is that it offers external validity, combined with statistical reliability, given the generally quite large data sets it involves.

Specifically, this study used a three-phase approach that relied upon both qualitative and quantitative data. The ultimate goal of this project was to develop a technique for social trace content analysis along the lines of that implemented by Boase and Naaman (2011). This technique used human coders to develop a taxonomy of content types in what they termed “social awareness streams” – data akin to the type that is most suited to the type of passive and extractive strategies that is discussed here. Once the coding scheme was created they used a machine-learning technique called supervised classification to code the remaining, much larger, data set. This dissertation’s methods were based on this approach: first, using qualitative research techniques to uncover in the data the specific techniques and behaviors that individuals use in the process of social information seeking and identity management, and then using quantitative techniques to reach statistical conclusions about the relative prevalence of each of these techniques, and the relationship between them and other factors, including demographics, psychographics, and network factors. Below, the dissertation’s first two phases, which involve qualitative approaches, are described.

Phase One Study Method

Phase 1 of this project included the collection of qualitative research data related to the information seeking behaviors involved in the browsing of SNS information, and the behaviors related to self-presentation and impression management. Again, the goal is to identify the specific types of behaviors, actions, techniques, strategies, etc. that participants engage in when interacting with others on social media. Specifically, the contribution that this phase of the project makes is to uncover those types of actions and interactions that other methods often missed – passive and extractive information seeking on the one hand and tacit, indirect communication on the other. The goal ultimately was to create a newer, updated taxonomy of interaction via SNS, building upon that of Boase and Naaman (2011) that allows for and accommodates a larger set of communication and interaction patterns than simply messages posted as text. For a variety of technological and methodological reasons (see Chapter One) to date the understanding of how these types of interactions fit in models like that of Boase and Naaman, for instance, is incomplete. Additional goals for phase one included the development of phase two interview protocols, with a specific focus on testing of the talk-aloud protocol (described below).

For this purpose this study made use of the talk-aloud protocol for the analysis of verbal data, a mainstay of user interface design analysis, system interaction analysis, process analysis, and so forth that was first introduced by Ericsson and Simon (1984). Phase one involved individual interviews lasting approximately 30 minutes with five participants that involved questions about their use of passive and interactive information-seeking strategies while using SNS. Participants were be recruited from the academic communities of Rutgers University and St. Peter’s University, both in New Jersey. They were recruited using posted flyers and email solicitations on respective university listservs using the same flyer. Participants were chosen based on their age (over 18 as per the approved Institutional Review Board [IRB] protocol), and for their use of Twitter (at least semi-weekly use), and so that there was rough gender parity in the participants. During the procedure, participants were asked to talk-aloud, while they scroll through their SNS feeds, while the investigator was present (see Appendix A for the interview protocol). Their verbalizations were audio-recorded and later transcribed for analysis. Participants were asked to sign an informed consent form (see Appendix D) and each was compensated with a $20 gift card funded by a dissertation support grant.

Ericsson and Simon’s (1984) technique is an approach to qualitative research that is designed to complement quantitative or experimental approaches to psychological inquiry, particularly in the context of systems design, interaction, etc. They attempted to reinstate the use of verbal data as a valid resource for understanding cognitive processes. Specifically, they set about this by introducing a methodology that would allow for using verbal data to discover phenomena of interest, but also interpret this data within a theoretical framework. A major assumption of this approach is that all cognitive processes pass through short-term memory, and therefore the study of the vocalization of this thought is indistinguishable from other measures of its effect. They argue that a sentence is a verbal realization of an idea, and that verbs in a sentence can be used to identify different kinds of information and cognitive processes. This assumption is contingent, however, on the corollary assumption that everything a participant might know has, at some point, passed through short-term memory, and they were conscious of it at one point. If the participant is asked to vocalize this thought as they are in the process of perceiving it – or shortly afterwards while it is still in short-term memory – then this vocalization is as good a measure of that thought or cognitive process as any other.

As a consequence of this understanding, and because the aim of the protocol is the study of task-directed processes and interactions, Ericsson and Simon suggest that this protocol is only appropriate when used in conjunction with concurrent verbal data – collected in the process of the task – and certain types of retrospective data. Specially, they identify three types:

- Vocalizations of thoughts that are already encoded in verbal form. This they term "talk aloud" data.

- Verbalizations of a sequence of thought is held in some other form, such as visual information, procedures for carrying out a task, special or geographic representations, etc. This they term "think aloud" data.

- Other verbalizations. (Any other data, particularly retrospective reports on thoughts not held in short term memory.)

For Ericsson and Simon, it is primarily talk-aloud and think aloud that are of interest, because these verbalizations express the content of short-term memory. Phase one used a “think aloud” protocol designed to collect data concerning the interactions that participants have with social media platforms. Participants were asked to interact with these platforms as they would normally, and speak through their interactions in order to discover not only the actions they are taking, but the ideas, motivations, concepts, and models that support and contextualize them.

Specifically, the focus was on in studying participant’s interactions with respect to two general sets of behavior: a) information seeking behaviors related to learning about other actors in the participants’ network, and also b) identity management and self-presentation behaviors, that is, the creation of the data that the participant intends others to use to learn about themselves. While Phase one was intended to be a primarily explorative study that uses an inductive approach to generate the codes that would ultimately be used for the categorization of data in phases two and three, Ericsson and Simon are clear that their Protocol Analysis approach is only valid as a means for analyzing data within a specific theoretical framework. As such, this study began analysis with the application of existing models for the two general sets of behavior that are being studied. For the description of information-seeking behaviors, the study uses the model created by Ramirez, et al. (2002) discussed in Chapter One, which categorizes information seeking behavior in the context of impression formation and uncertainty reduction into interactive, active, passive, and extractive modes. As a framework for coding self-presentation, this study used that developed by Boase and Naaman (2011), which, as discussed in Chapter One, creates categories of SNS use types based on categorizations of message content, and build on the foundational work on self-presentation by Goffman (1956).

Phase Two Study Method

Phase Two builds upon findings from Phase 1 and differs primarily in the focus of analysis. As originally developed by Ericsson and Simon (1984), protocol analysis consists of two parts: first, the transcription and editing of the data records of a research session, and then second (and mainly) the encoding of verbal reports. This encoding is done by using a priori determined coding categories. Each segment must be treated independently of the surrounding text. Hence the method requires a high awareness of the relation between raw data, interpreted data, and theory. Phase two included a similar set of interviews as Phase one, though it involves a longer interaction session on the order of 30 minutes, as well as 15 additional participants. This process used an extended protocol (see Appendix B). Their verbalizations were audio-recorded and later transcribed for analysis. Participants were asked to sign an informed consent form (see Appendix D) and each was compensated with a $30 gift card funded by a dissertation support grant.

In the talk-aloud approach, moving from raw data to theory involves three steps. The first is the rote transcription of verbalizations, though they point out that this needs to take place in the context of a theoretical structure. The next step is segmentation, which involves dividing the transcript into individual assertions or propositions, such that each segment constitutes one “instance” of a general process. They note that cues for these divisions can include meta-linguistic signs like pauses or intonation. Finally, the last step is the actual encoding, which needs to be done using an a priori vocabulary that was developed from an initial task analysis and an examination of the protocols. As noted above, for both Phases one and two participants’ sessions were audio recorded, and transcribed for analysis. Coding was be done using the NVivo software package, with 20% of the coding also done by another coder in order to ensure inter-coder reliability.

Thus the division of the qualitative portion of the research into two phases is informed by Ericsson and Simon's (1984) guidelines for protocol analysis for talk-aloud research protocols. Phase one represents this initial task analysis, where the main goal was to develop the vocabulary necessary to encode the behavior data uncovered in the analysis of a larger pool of participants. Then, in phase two, based upon this encoded data it was determined which specific types of social trace data were to be collected in order to measure and track this type of behavior. From there, building upon the initial two phases, the study moved to phase three, a primarily quantitative approach that is described below.

In order to test the validity of the codes and categories of use types discovered in phases one and two, this study then tested them on a set of publicly available communications that were made on one of the most popular online social networking sites, Twitter (Greenwood, Perrin, & Duggan, 2016). Twitter was chosen as the platform for this research because it constitutes the largest publicly-available set of SNS data. While other services, such as Facebook, are more popular, their access policies disallow the collection and aggregation of its users' data. Again, the ultimate goal of this project was to more completely integrate the types of passive and extractive strategies of information seeking, and the tacit, indirect forms of self-presentation, into existing models of SNS use typologies. That is, to build a more complete picture of the types of behaviors and strategies that are used for identity management, and furthermore to demonstrate how this more generalized theoretical framework can be applied to existing social media data in order to uncover evidence of this behavior that might have previously gone unnoticed.

Phase Three Study Method

Proceeding from phases one and two to phase three required translating the actions and behaviors of study participants into a set of specific data traces that were collected from raw SNS data. This is not always straightforward. For instance, operationalizing "social information seeking" as a variable that can be systematically measured using the type of data typically available in most data sets is murky. In order to be done properly, log data that recorded how often an individual visited the pages of others, and how much they read, would be required; this type of data is typically unavailable. Instead, other data will have to be discovered. Analysis from previous phases will be used to indicate passive or extractive information seeking activities. For example, these traces could take the form of liking, sharing, and commenting behavior. Types of information seeking activities might also be detectable through network patterns detectable in the data. Individuals with deeply-connected network topologies are likely to be exposed to, and therefore seek out, different types of social information than those with less dense networks. Results from phases one and two informed the development of analysis protocol for phase three and was used to demonstrate how a more complete picture of uncertainty reduction and impression formation proceeds. This study used existing social media data sets in order to underline the point that the current approach may uncover additional insights that others may have missed.

Detecting the presence and use of tacit or indirect communication was more straightforward. This type of data has long been used to detect and measure, for instance, the strength of ties between individuals in social networks. For example, Gilbert and Karahalios (2009) were able to demonstrate a method for estimating tie strength that relied upon signals present in a dyadic interaction on Facebook, including commenting, sharing photos, tagging, etc. Working from this example, a more considered and nuanced understanding of the differing behaviors involved in the use of tacit and indirect communication for the purposes of identity management and impression formation improved upon this model in several ways. First, their model is fairly naïve with respect to messaging: it factors in merely the number and frequency of these interactions, not their content. In this project a model was be created for identity management in SNS that is based on a study of use behavior discovered via qualitative and quantitative means. This study applied this model to the type of tacit, indirect interactions in order to measure their relative frequency and understand the implications of this frequency. By reasoning about data related to social ties strength (for instance) using this model, a framework was be derived for describing patterns of interaction, and the strategies that drive them.

More specifically, this approach helped explain not just that certain types of interactions tend to be associated with one another, or that certain patterns of information seeking or browsing are associated with certain types of use personas, but rather why these associations tend to occur. This is the rationale behind proceeding first with a qualitative method in order to develop a set of codes that can be used to categorize the strategies and behaviors present in the larger set of SNS data. Because previous studies have tended to focus on overt indications of self-presentation, it may be that much may be going unexplored. This study used data regarding indirect interactions to build a more complete framework for understanding self-presentation and identity management, and did so using existing data sets in order to underline this point.

Because Twitter does not allow the distribution of data sets that include message content – which includes liking and retweeting behavior – it will be necessary to crawl for these tweets. Given the size of Twitter, and the large number of bots, companies, and generally other-than-useful accounts that exist, establishing a set of legitimate users to base the crawl from would be a very difficult task. In addition, even if it were possible to construct such a list, in order to discover the signal data needed in order to establish tie strength, connectivity, etc. as discussed above, the additional task of crawling the entire Twitter graph would need to be completed in order to establish these links. Given the amount of time that these tasks would take, it seemed sensible to use an existing Twitter user corpus for this purpose (especially considering that determining an expedient way to separate actual Twitter users from corporate actors, Public Relations firms, newswires, bots, etc. is beyond the scope of this research). For this purpose this study will use the social network dataset for Twitter that is available via Stanford University's SNAP (http://snap.stanford.edu/). These data contain a list of 81,306 nodes and the 1.7 million edges between them, which provided the basis for the network connection and social circle inferences that were drawn as part of the analysis. In order to obtain actual tweets and interaction behavior, this data was then crawled independently. Then, specific tweets were crawled and stored using Twitter crawling code that was adapted from the python-twitter Application Program Interface (API) wrapper.

Once crawled, tweets were analyzed for content using an application of Latent Dirichlet Allocation (LDA) that groups text from a corpus into clusters based upon a probability distributions. Topic modelling has a rich history in application to Twitter (e.g., Rehurek & Sojka, 2010, Zhao, et al. 2011, Hong & Davison, 2010) and has proven to be a useful tool for categorizing and grouping large corpora of text. LDA is an approach to topic modeling that forms a series of probability distributions over progressively larger subsets of text, and then samples from those distributions in order to determine a set of topics that explain similarities between the texts. LDA supposes that these topics account for the systematic differences between the texts, and by applying statistical inference, is able to categorize the texts into those topics.

The quantitative portion of the study was designed to integrate into the proposed cohesive, mixed-method approach in three ways. First, by identifying categories and groups of interaction types that can be specifically addressed in the qualitative portion. For example, in a mixed-method study, Gilbert (2009) attempted to predict tie strength based on signals present in social media. He found several instances in which his mathematical model didn’t line up with his participants’ reported tie strength. Investigating this disconnect via qualitative interviews, Gilbert discovered that some relationships had special sets of characteristics for which is model didn’t account. For example, some participants reported a very high degree of social intimacy with their priest or other religious leader, but rarely if ever interacted with them on social media. Identifying the presence of these edge cases is an important goal of the quantitative portion, to be sure that our qualitative interviews capture as much of the potential interactions as possible.

Second, the use of quantitative data was designed to provide a set of correlations between various types of passive interaction, and various network characteristics, user characteristics, etc. While qualitative interviews are well-suited for reasoning about personal intention and desire, they are not as well suited for measuring other types of variables that may impact the type and nature of passive social interaction in SNS. For instance, preliminary anecdotal evidence suggested that some SNS users are uninterested in politics, and use the service only to maintain relationships with friends and family. In such cases, it might be expected these individuals to have social circles that are relatively dense, and relatively highly embedded. By converse, those who use the services to help spread a particular political message might be more liberal with the number and closeness of their SNS connections, resulting in a different network structure. The quantitative portion of this approach was therefore designed to give context and background to the qualitative research. However, it bears noting that in no way is it possible to draw specific conclusions about intention (and even the above hypotheticals tread a bit close to this line) merely about correlation. As such the direct connections, as relates to individual behavior, may be limited, however both portions are needed to provide a full accounting.

Finally, the quantitative portion was designed to provide some measure of magnitude, prevalence, and effectiveness of each passive interaction style or strategy. Returning to the earlier discussion about the interactive, active, passive, and extractive styles proposed by Ramirez (2002), the authors not that it is “difficult to imagine” a context in which passive or extractive strategies could be useful or even effective. Little research (Antheunis & Valkenburg, 2010 is a notable exception) since then has done much to question this claim. One of the core suppositions of this project was that changes in both the nature of the systems used to host CMC (moving increasingly to mobile devices that may be use almost anywhere) and the patterns of use in SNS have created an environment where passive and extractive uncertainty reduction, and indirect interaction styles are far more common than in 2002, and perhaps even the dominant modes now. Testing this assumption was a major goal of this project, and can best be done by using data resulting from actual interactions in SNS.

Phase Three Study Method

This is a documentation of Twitter data aimed discovering use types and patterns among users of the service. The corpus is derived from a public dataset maintained by SNAP. This data set lists 81,306 Twitter users and their follower network as of 2012. This data was chosen primarily due to its quality. It has been vetted by SNAP scientists to ensure that it contains only data related to actual Twitter users, and not fake accounts, bots, or other sources of potential noise that would impact the results of this study. However the SNAP data is dated. Many of the users that appear in the data are no longer active on Twitter, and In order to update the data the original users mentioned in the data were crawled again in order to obtain the users that the original SNAP users follow. From this list, only those users that followed the original SNAP users in return were considered, on the assumption that these are likely to be actual social connections and not the accounts for brands, celebrities, news outlets, and the like. This expanded list of some 1.2 million accounts was the initial starting point. From this list were removed inactive or redundant accounts that appeared, which reduced the total number of accounts in this study to 447,318. From these accounts, the friend network was crawled again in order to calculate network measures, and finally the most recent tweets for each were collected.

Variable selection.

The first step in the study was to create a classifier to categorize the tweets by content. The purpose of this was two-fold. First, to create a set of variables that could be used to measure and classify users based upon their behavior with others on twitter. Second to compare these measured variables to the findings of a partnered qualitative study in order to extend and amplify the findings of each.

However, arriving at a set of quantifiable variables that can be used to categorize the behavior of an unknown Twitter user is a non-obvious task. Furthermore, it was of theoretical importance that these measure be selected purposefully in order to provide the resultant categorization scheme some level of face validity. That is, because the method used in this study would be relying upon supervised clustering techniques to categorize individuals, it was important that the variables used to accomplish this clustering emerge from observations in the data. These clusters were used to compare with conclusions made during qualitative study, so in order for that comparison to make sense the clusters observed in the quantitative analysis needed to be generated using the same logic. Furthermore, while several studies have categorized tweets using various inferential techniques that often rely upon unsupervised learning, such an approach would not be appropriate in the present study.

Because the purpose of this study was to use a mixed-method approach in order to better refine the method of the quantitative portion and to validate the insights of the qualitative portion, again it was important that these approaches follow the same logic and trajectory so that these comparisons are valid. As such, when determining a set of variables that would be used to measure and quantify each user, decisions were made based upon data that select the variables that arose during qualitative interviews. These interviews were conducted using a structured format known as the talk-aloud protocol, in which participants were asked to conduct tasks related to information practice and communication in social media. This data was then coded specifically with the intention of discovering variables that could be applied to quantitative analysis. That is, as individuals discussed themselves and their use patterns, the protocol cued them provide information about their thoughts, idea, and intentions. In this way, the variables discovered in the qualitative portion, and the attendant information with respect to intention that is associated with it, can be tied to the qualitative data.

Variables were identified from the qualitative data; a table depicting each is provided below. Once potential variables were identified, the next step was operationalization, which was handled as described in the chart. Once variables were operationalized, each user was analyzed according to each variable, and given a value for each. Table 1 lists some of these variables.

Table 1

Variables

Variable	Description	Operationalization
Directionality	Is the message directed at a particular audience	Relative ratio of at-mentions
Reflex	Frequency of reflexive replies	Ratio of replies to at-mentions
Intentionality	Ratio of indirect to direct communication	Ratio of likes to posts
Audience	How personal is audience	In-degree to out-degree ratio
Familiarity	How close is the social group	Adjusted network density
Content	What the content concerns	Topic modelling

Frequency	How often does the activity take place	Logarithmic rate measure
Linking	Relative frequency of informational posts	Relative ratio of retweets containing links to those not
Co-activity	Synchronicity between communications	Logarithmic decay time ratio of replies to at-mentions
Message hold	How varied is the content that is posted	Mean number of topics

Content analysis.

Further notes on content operationalization is included here. Each user's tweets were categorized using Latent Dirichlet Allocation (LDA), a topic modeling algorithm (Rehurek & Sojka, 2010, Zhao, et al. 2011, Hong & Davison, 2010). In LDA, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. LDA then defines a generative process for each document in the collection. First, for each document pick a topic from its distribution over topics. Sample a word from the distribution over the words associated with the chosen topic. Finally, repeat this process for all the words in the document. The words in the document are the observable variables, while all other variables are latent. The problem then becomes inferring the values of these latent variables which is a problem of Bayesian inference. As each word in each document is sampled, the probability priors are updated as they converge towards the distributions theoretically present in the corpus.

LDA, as originally designed, is not ideally suited for application to Twitter data. A number of researchers have proposed methods of improving topic modeling for application to the specific requirements of Twitter. In the interest of exigency, however, this study used an approach that while not ideal has shown to be relatively effective:

collapsing all the tweets for each individual into a single document. Since this study is not primarily concerned with the content of each individual tweet but rather with the overall nature of the expression that each user makes as a part of their communication practice, this seemed like a reasonable approach.

Another major roadblock in applying topic modeling to Twitter data is the presence of a large number of misspellings, abbreviations, emoticons, and so forth present in the data. This, too, has been the object of considerable study for researchers. Several approaches have been proposed for normalizing Twitter data to reliable language tokens that can be used in natural language processing. As an exigent solution, this study employs the Carnegie Mellon University Tweet Natural Language Processing tokenizer to resolve language that appears in tweets to their proper tokens (“smh” becomes “shaking my head”, etc.)

Once parsed, cleaned, and tokenized the actual LDA analysis began. Each user’s tweets were analyzed in order to create a set of topics common in their tweets. These topics are those inferred by LDA to exist given the relative frequency of co-occurrence in the document, assuming that the document is sampled over some distribution of topics. Some example topics are presented in the table below. Each user’s document had a distribution of topics inferred from the word count, and this distribution was recorded as a vector for the purposes of comparison and further analysis. For illustration, Table 2 includes some examples of some of the topics returned.

Table 2

Example Topics

Family-1	News-1	Weather-1	Social-1

Dog	(Donald) Trump	Hurricane	Dinner
Wife	Congress	Harvey	Phone
Kids	Obamacare	Houston	Picture
Today	(Paul) Ryan	Irma	Movie
Dinner	DACA	Flooding	Friend
Play	(Steve) Bannon	Storm	Party
Husband	Speech	Texas	Tonight

Finally, note that for the purposes of cluster analysis, the content variable was recoded into a number of smaller variables. This is due to the fact that a relatively simple form of cluster analysis that works best with scalar data was employed in this study so as to limit complications and speed up analysis. However, it does not work well with vectored data, unlike many more sophisticated approaches. As a result, the content variable was re-coded to a number of sub-variables, each representing a scalar measure of the frequency of appearance of each group of topics. For instance, there were several topics related to news, world events, the weather, etc. These were grouped into a larger category, and the weighted mean of the appearance of each topic – adjusted for its overall frequency of appearance across the entire corpus of tweets – was calculated and used as a scalar measure of general tweets about current events. A similar process was conducted for topics dealing with home and family, topics related to celebrities, arts, culture, and so forth.

Primary data collection of tweets occurred during July, 2017. At this time, data was monitored and initially analyzed for topic distribution. This early analysis revealed a great deal of fluctuation among topic distributions, which appeared to vary wildly. After analysis revealed these effects, an additional data collection period began in designed to collect long-term data that might serve to smooth these effects. Therefore, this study continued data collection for eight weeks after the main collection period, between

August and September of 2017 in order to at least attempt something like a rational cumulative distribution function. During this additional data collection period, a randomly selected 25% of the accounts tracked in the main data set were additionally followed, and their tweets again sent through LDA analysis, including their initial data included in the main data set. This allowed for the normalization of the topic distributions by comparing them to historical trends.

Cluster analysis.

There are many approaches to cluster analysis, including the Random Forest algorithm, Support Vector Machine, k-Means, and Naïve Bayes (Scott, 2017). Each has their strengths and weaknesses, and for the purposes of analyzing Twitter data density-based clustering using SVN is commonly employed in large-scale environments. However, given the focus on moving the analysis forward quickly, this study employs k-means. K-means is a well-documented (Ding & He, 2004) approach to clustering that divides a data stream into *k* clusters. In the first step, *k* points are randomly selected as centers for the clusters and used to associate all remaining points. Once every point is associated with a center, the mean for every cluster is calculated and named the new center. This process continues until the centers no longer move. Although k-means is able to effectively cluster similar points together through the Euclidean distance, the researcher needs to know how many clusters should be found. If *k* is selected incorrectly, improper clustering is likely to result. Typically, therefore, many values of *k* are used and goodness-of-fit measures calculated in order to find the best result. There are several other shortcomings of k-means, particularly as it pertains to large streams of data. To minimize these errors, several improvements on the k-means approach have been proposed. This study relies upon one such approach termed k-means++, as implemented in the StreamKM++ software package used in this study (http://huawei-noah.github.io/streamDM/docs/StreamKM.html).

K-means++ and its implementation in StreamKM++ are quite fast compared to other clustering approaches, but computation still proved a substantial roadblock. Any adjustments to the model would require days of computation to validate, and the approach was therefore intractable given the time frame. As such this study reduces the number of variables as an intermediate step before cluster analysis. This was accomplished using Principal Component Analysis, or PCA (give URL or citation for this analysis). PCA is a variety of factor analysis that is useful in situations involving multiple variables that may have different scales and distributions. PCA works by creating a translation matrix of a set of observed data, and then casting these observations onto the new variables (components) to create an estimation of the observations. The data is then fit to these components beginning with the variable containing the most variance and ending with that with the least. Once all the variables are fit it is then possible to determine which variables are providing useful amounts of measures of variance and which have redundant information. In many instances, including the present one, it is possible to reduce the number of variables present in a model significantly, with acceptably small loss in fidelity to the original. Based on this adjusted model, the cluster analysis could be run on just five components of the model, rather than the ten initial variables, dramatically speeding up calculation time. Applying PCA reduction prior to clustering has an additional major advantage: as Ding and He (2004) have demonstrated, PCA component loadings tend to track quite closely to k-means clustering variance, on the order of less than 1% in many cases. This means that the PCA component vectors can be used as a starting point for the cluster centers in the k-means clustering algorithm, saving a massive amount of time. This estimation saved thousands of iterations of the procedure per run, and since it provided a useful starting value for k (that is the number of components discovered, in this case five) saves thousands more by eliminating the number of values of *k* that must be tested. Without the time-saving estimations and data reductions that came by applying PCA to the clustering problem, it is unlikely that the calculations performed in this study would have been feasible.

The analysis then proceeded to final clustering using k-means. One of the principal rationales for its selection was its speed and reliability over a large number of iterations. Using and approach developed by Ding and He (2004), it is possible to estimate the initial number of clusters and the initial central vector for each cluster using PCA. By then using the PCA-reduced data in the cluster, a large reduction in the dimensionality of each component vector can be achieved. This reduced computation for each run of the analysis by several orders of magnitude, from days to mere hours. This consideration was the decisive one, as the mixed-method approach demanded near-constant iterations on and improvements of the model. As new data came in from interviews, variables would be added or removed or changed in the quantitative data, which required a re-run of the cluster analysis. This, in turn, altered the number, size, shape, and characteristics of the clusters, which fundamentally altered the overall model. This then altered the approach to coding and analysis taking place in the qualitative portion of the study, which in turn necessitated changes in the quantitative analysis, and around it went. The ability to for the model to be flexible, reactive, and adaptable was a decisive consideration, and the utility of this PCA/*k*-means approach is difficult to overstate.

An initial value for *k* was calculated using Ding and He's (2004) technique of using the eigenvector dimensionality of the variance of the data determined via PCA. The study then undertook several iterations and tests using a variety of initial cluster means and k-values. Ultimately the model settled to five clusters. The values for these clusters was according to the components created during PCA. At this point the analysis was run again, using the full variable set rather than the components.

Ethical Considerations

This research was subject to review and oversight by the Rutgers University Institutional Review Board (IRB). It has been approved for conduct because minimal potential harm was possible to qualitative participants, and the quantitative data came from an existing public data set. The primary investigator successfully completed the required ethics training and conducted the research in a safe and conscientious manner under the oversight of the IRB.

Nonetheless, ethical considerations remain. The IRB can not be expected to be fully up-to-date with current trends in online ethics and privacy regarding social media data. Recent events, including the leak of the private data and communications of some 87 million Facebook users to Cambridge Analytica, and the effect that this leak had on the 2016 U.S. Presidential election, have called individuals at all levels of involvement with social media – professionals, researchers, etc. – to re-examine their practices for ethical considerations (Rosenberg, Confessore, & Cadwalladr, 2018).

Though the IRB calls only for the anonymization of social media data, the Cambridge leaks have demonstrated that even anonymous data can be used in unethical ways. This research, however, differs in both its techniques and goals. First, it is a primarily behavioral model. All demographic data collected for qualitative participants has been kept confidential. Participant numbers were assigned for reporting quotations and any connections to demographic information are being kept in a secure, password-protected file. No demographic data was collected for the quantitative data from Twitter. Second, the behavioral data that was collected was abstracted through several layers of analysis. For instance, the actual language contained in each Tweet was never directly used in the model, merely the topic of that conversation, and under what general category of content (information, social, etc.) that topic fell. Finally, this model is a description of types of communication and information seeking present in those that participate with social media. As such, its utility outside its direct sphere of application is limited and practical application of this model for political or economic gain is difficult to imagine.

Conclusion

The method this study uses is a mixed-methods approach designed to leverage a variety of techniques in order to examine the link between self-presentation, identity management, and information seeking in social media. By combining qualitative and quantitative data, the goal was to create a procedure that could generate a model for the understanding of patterns of interaction behavior that take place in SNS settings using either quantitative or qualitative means. That is, the model should be useful in application in a variety of settings and approaches.

For a variety of reasons related to cost, time, resource availability, and scale, the project was somewhat limited in scope, and several adjustments have been made to the procedure in order to ensure that is a feasible project. Chief among these is the reliance upon an existing dataset, a limited level of statistical and computation analysis, and a comparatively attenuated qualitative interview schedule. Given that this is something of a speculative study – designed to explore a technique and generate a model – rather than one based upon hypothesis generation and testing, if not constrained data gathering and analysis can quickly become haphazard and onerous. As such, strict guidelines for the procedure were set down and followed so as to limit the potential for mistakes. Moving from these guidelines to their application, the subsequent chapters outline the results of the data gathering, as well as its synthesis into a cohesive model.

Chapter Three

Qualitative Findings and Discussion

Introduction

This chapter discusses the outcomes of the qualitative portion of the research project, as detailed in the methodology section above. It was designed as a complement to the quantitative portion, and was the starting point for the development of a model of identity management and impression formation in social media. This study was conducted in two phases: a preliminary set of 15 to 20 minute interviews with five participants, and a longer set of 30 minute interviews with an additional fifteen participants. Across all 20 participants, 12 were men, 8 women. The median age was 33, and 17 had a Bachelor's degree or higher. As part of the selection screening, all participants reported using Twitter at least once a week or more often.

As both the quantitative and qualitative study progressed, the overall model was adjusted as data came in. In this chapter in particular there will be noted several places where adjustments to the model were made, though this occurred in both the qualitative and quantitative portions of the study, as tends to be the case in mixed-method design.

Phase one study outcomes.

As noted in Chapter Two, the main purposes of the phase one pilot study were:

1) Verify the interview guide as a data collection tool

2) Preliminarily investigate the research questions

3) Establish a baseline code

The following discussion addresses the outcome of each of these processes.

Verify the interview guide.

As described in Chapter Two, The interview guide was based primarily upon the method of the talk-aloud protocol. In this method, participants are encouraged to provide an ongoing narrative of their thoughts, motivations, reasoning, etc. behind actions as they take them. While talk-aloud has seen wide application in a variety of research fields, it has been particularly popular in the field of User Experience (UX) design and Interaction design. In this arena, the goal has typically been to understand the assumptions that users make with respect to the affordances provided by the interface. From there, designers can engineer interfaces designed to support and reinforce these assumptions in order to create and maintain a fluid and intuitive interaction experience.

In this dissertation, the underlying goals are much the same. That is, to uncover some of the assumptions that participants make with respect to the individuals with whom they interact in social media, and to use those assumptions, ideas, and motivations in order to build a plan for studying these interactions in a mixed-method setting. More specifically, the main purposes of this dissertation are to create something approaching a framework for the analysis of indirect or signaled communication that often takes place in the idiom of social media: likes, dislikes, hearts, re-tweets, and the like. Given the relative dearth of extant literature on this topic specifically as a mode of communication, this seems appropriate. Further, another goal is to examine how these interactions influence and inform the information seeking process of individuals that use social media as a mechanism for learning about the social world they inhabit.

For the purposes of gathering data regarding how individuals both use cues in social media in order to establish their identities, and also use the information they discover via social media to inform those choices, the interview instrument (see Appendix A for the interview instrument) had both strengths and weaknesses. It worked well as an instrument for the gathering of reasoning and practice in social media communication. That is, having participants discuss aloud their thought processes served to provide valuable insight into the rationale behind their actions. Furthermore, this method encouraged discussion and provided impetus for expanded discussion. As participants vocalized their thoughts and actions, they often discovered or rediscovered the motivations behind those actions. After varying degrees of coaxing, these sorts of discussions seemed to come readily to participants. Given the mixed-method approach that this study took, this type of data could prove to be a valuable addition to the signals evidence gathered in the quantitative social media data.

A weakness of the interview guide was readily apparent, however: it was insufficiently robust to address research questions related to impression formation and information seeking habits that participants engage in while browsing social media. A large part of this project was to focus on the information seeking habits that users of social media engage in when learning about their social environments, specifically with respect to its expression in their networks present in social networking systems. Based upon existing literature, this project sought to identify several specific categories of social information seeking behaviors, including the passive, active, extractive, and interactive modes. While there was some data present regarding each of these modes, the instrument as originally devised did not generate sufficient data to arrive at substantive conclusions regarding the frequency, nature, and character of each of these modes.

To address this and other issues, the survey instrument has been updated to more fully account for these shortcomings in data (see Appendix B). Specifically, several information seeking tasks were added to the protocol, encouraging individuals to act, think, and speak about their tendencies and behaviors with respect to social information seeking. Furthermore, the instrument was redesigned in such a way so that data with regard to these modes or strategies would arise in context, occurring organically from the participants’ behavior. In other words, the instrument did a satisfactory job encouraging discussion and generating data with respect to communication and meta-communication. Several probes were added to each of the questions regarding communication to extend the discussion towards information seeking. In this way the instrument was re-designed to more closely integrate the communication tasks and the information seeking tasks, in order to more closely couple these two concepts in the thoughts and actions of participants. The goal here was to generate data not merely regarding communication, and not merely regarding information seeking, but also the synthesis between the two. As this synthesis is a point of emphasis in the research project, the instrument would be better able to generate data relevant to answering these research questions.

Phase one research question discussion.

Briefly, this portion of the dissertation discusses some of the findings from the initial preliminary Phase one investigation. Primarily because the purpose of this initial round of interviews was to validate and improve the interview protocol, and to gain some initial insights that might drive the focus of the more in-depth subsequent investigations, these initial insights are included here, prior to a more substantive discussion of the dissertation’s findings which appears in Chapter Five. By research questions a brief summary of the insights gleaned from the results of the preliminary study follow:

RQ 1. What strategies or procedures are used in the maintenance of online identities in SNS?

Data generated in the preliminary Phase 1 study suggest that multiple strategies are used, and that these strategies stem from a variety of purpose-driven activities that participants engage in. Use case seems to vary quite a bit among preliminary Phase 1 subjects, though this could easily be an artifact of the sampling method, which intentionally designed for the integration of a variety of use personas. Nonetheless, each participant was able to articulate a specific set of goals and motivations for the use of social media, and these motivations factored heavily into their expressed desire for self-presentation.

RQ 2. How are online identities managed in SNS?

This research question needed further investigation in the main quantitative and qualitative studies based upon preliminary research results. The management of online identity did not appear, in the five interviews conducted to be substantive different from the establishment of online identity. Once a patterns of behavior was solidified, and once a set of practices for communication and information seeking became the norm, these behaviors appeared to be relatively stable. That said, there was evidence of the varying of their strategies based upon the specific information most salient at the time. When cued for information seeking practices, for instance, individuals tended to take a more passive or extractive viewpoint.

RQ 3. What is the relationship between information seeking behavior and self-presentation and identity management in SNS?

The research instrument as designed in the Phase 1 study did not sufficiently provide for this research question. Amendments to the instrument, as outlined above, sought to improve the dearth of information gathered in the preliminary study. That said, there was some data gathered related to this research question. Several participants, for instance, noted the existence of a set of normative behaviors expected from them, and reciprocated by others, that served as guardrails for their behavior. Members of close-knit social groups reported being expected to respond to or react to the content posted by others. Further, one participant reported that these expectations were so entrenched that failing to respond to the content that a close communication partner posted would be interpreted as a slight. Analogous behavior could be seen in some of the more politically-motivated participants. Norms of reciprocity again manifested here in the form of expectations of liking, sharing, and re-tweeting of congruent political views. Such participants noted that one of their primary motivations for using social media was to spread their political views, and these tacit norms of automatic sharing and re-tweeting were useful, in that they were able to leverage the network of their political comrades in order to expand each other’s audiences.

RQ 4. What is the role of indirect forms of communication, such as liking and following behaviors, in the creation of online identity in SNS?

This research question is another for which the interview instrument has been improved and expanded. Initial talk-aloud tasks asked participants to discuss their active and direct communication, to the detriment of indirect communication. Several interview prompts were been added to improve the quantity and quality of data gathered regarding this research question. That said, preliminary data gathered in this area seems to amplify and reinforce the data gathered for the research questions one and two. Particularly in relationships whereby norms dictate regular patterns of liking, sharing, and reinforcing the content posted by others, indirect meta-communication served as an important mechanism for this to occur, particularly for those participants that received a high frequency of messages and interactions. Participants noted that reacting to the content posted by others becomes far easier and less mentally onerous if meta-communication is used, in that it only requires a single click. At the same time, however, this lowered barrier for interaction has only served to reinforce the normative expectations of communication. Given that the effort required for interaction itself (ignoring the effort required for constant monitoring of new content, which has a far higher cognitive cost) is lower, the slight felt by participants' interlocutors should they ever fail to meet their expectations is all the higher. Indeed, several participants noted using this implied slight as a form of message-sending: using silence as a response.

RQ 5. What is the role of indirect forms of communication, such as liking and following behaviors, in the maintenance of online identity in SNS?

Participants discussed this topic with good detail, and the instrument gathered satisfactory data in this regard. Generally, they noted that such forms of communication served as useful ways to maintain social ties, reinforce reciprocal relationships, and traverse temporal or geographic distance. Several participants noted that such forms of indirect communication had supplanted more active forms of interaction. Further, they noted that the comparatively low level of effort required in such interactions encouraged them to maintain a broader level of infrequent communication.

Code identification.

A final purpose of the preliminary round of five interviews was the creation of an initial codebook, which would be used as a guide for establishing research themes identified in the larger sample (see Appendix C for the final codebook, which was developed from this initial one). Given the strictures of the talk-aloud method, it was important to have a set of lines of inquiry established before conducting long-form interviews using this method.

This codebook was initially established based upon the five interviews conducted in Phase 1. Given that this study was conducted with an existing set of goals, data was coded in such a way that it could inform the established research questions (RQ 1-4, see above). It was important that the interviews stay relevant to the questions under examination, so that the data could inform later portions of the study. Ultimately, data from the short-form interviews were be used to devise a longer-form interview protocol and data code. These data, in turn, informed the creation of the quantitative model that would be used to further investigate these research questions in the context of a large set of social media data. This process, however, was not serial. Insights from the quantitative findings also helped shape and focus the interview guide. As the model began to take shape, certain questions began to arise. What types of data should be included in the quantitative model, for instance, or what should be left out. As the model began to focus on a specific set of variables, it created the need for a greater level of precision with regard to the data being examined in the qualitative portion of the project.

The codebook was therefore generally organized around the five research questions themselves. In order to answer the research questions, data was gathered about three general concepts:

1. Identity formation and management
2. Information seeking and impression formation
3. Indirect communication, especially as it pertains to 1) and 2) above.

These three general categories formed the basis for the initial codes by which the data was recorded. As the initial study progressed, a number of additional codes were created, which both added breadth and depth to the overall data design. Further details about these additional findings are provided below.

Phase two outcomes.

Phase 2 involved 15 longer-form interviews with more directed protocol that aimed at uncovering the types of strategies that individuals used to create and maintain their identities in social media (see Appendix B). The demographics for this group are combined with those for phase one, and listed above. Specifically, the protocol was amended to more directly interrogate the strategies that participants used in the process of learning about others and forming impressions of them, as well as how these practices impacted their own information management practices.

Data was coded using an existing framework based on previous work as outlined above. However, this coding scheme was amended and altered as data was analyzed. While this prompted no major substantive changes in the overall theoretical framework, a number of additional themes were identified that opened the analysis to new possibilities (see Appendix C for final codebook).

While completing the tasks assigned in the information seeking portion of the protocol, participants displayed aspects of each of the active, passive, interactive, and extractive modes identified in the theoretical framework (see Chapter Two). The interactive mode, characterized by direct communication with others, was used to respond to posts or to ask questions. The active mode, characterized by intentional gathering of information about a subject based information present in their profile, network, etc. was used to establish the nature of the relationship participants had with a target individual. The extractive mode, characterized by examining the history of another's behavior, was often used to form impressions regarding interaction behavior. The passive mode, characterized by undirected monitoring of communication as it occurred, was used to maintain social relationships and gauge the behavior of others.

While each of these strategies were used in varying contexts, there was also variance among participants in their application. Some participants favored one strategy over another, while others were more catholic. Participants that reported high levels of communication frequency, for example, tended to rely upon extractive strategies less. They reported that since their involvement was frequent, their need to rely upon historical information to form judgments or make impressions was less of a concern. "I have so many new tweets coming in, I almost never go back through people's history. I never really seem to need to," Said Participant 10. Of greater interest was the actions taken recently. As a result, the tended to rely on active strategies, being more interested in how the individual fit into their various networks of friends and family, rather than diving into archival history of that individual's posts. "This guy seems to know my sister, and he's on constantly, so I'll give him a like and a follow," she continued.

Discussion of findings.

Through the use of a talk-aloud protocol, respondents were cued to discuss how they think about and establish their online identities via the behaviors they engage in when communicating with others in their social circles. That is, rather than addressing in an overt manner the idea, concepts, and designs of the individual and how they thought about their identities and those of others, this protocol was designed to elicit these ideas on a tacit level by cueing participants via the performance of their interaction with the SNS. More specifically, individuals were given a number of interaction and information seeking tasks but no specific instructions on how to carry them out. In this manner, they were given the opportunity to demonstrate their own ideas about identity formation through the mechanism of their actions. This is the goal of a talk-aloud protocol: it seeks to uncover the mental constructs that individuals have with respect to the subject of study in a tacit, even automatic manner.

Political discussion.

Nearly all respondents mentioned the prevalence of political discourse on Twitter to one degree or another, with varying levels of satisfaction. While some respondents clearly saw Twitter as an environment well-suited to political dialogue and discussion, others had a different point of view. Specifically, several participants intentionally avoided political discussion or attempted to ignore it when this was impossible. Thus the extent to which political dialogue comprised the message traffic and interaction patterns of an individual represented an important place of divergence among the respondents. On Twitter, as in many other places, politics was a divisive issue.

Collecting data regarding political discourse was not one of the main purposes of this study. The extant literature contains a large number of studies that specifically deal with this topic, and quite well (Small, 2011; Parmelee & Blanchard, 2011). Furthermore, though political ideology has been suggested to be intimately tied to individual notions or feelings of identity, this topic was outside the scope of the current project. The aim of this project was to create a model of identity expression in a social networking system, and to ground this model in a qualitative understanding of the types of behaviors that drive this model. That is, rather than using naïve approaches to model generation of this type (which is not uncommon in many applications of a large data sets) the idea was to develop a model grounded in data gathered from directed interaction. While it is easy to overstate questions of causality, the idea in this dissertation was to specifically identify interaction behavior tied to patterns of identity formation and maintenance. By creating a model based on such directed behaviors, the aim was to identify variables and measures of interactions specifically related to identity formation. In other words, to hone in on a specific set of variables directly related to the research questions at hand. Thus this study could both more directly address research questions in both the qualitative and quantitative portions, while at the same time neatly skirting issues of message content and specifically political content. This would, incidentally, save a great deal of analytical effort, as ideally this study could simply ignore message content and focus on network-centric variables in the quantitative phase of this dissertation.

However, political discourse and its place on Twitter became a matter of discussion for nearly every interview subject at some point during the sessions. Thus the findings demanded that at least some measure of message content be included in the quantitative model, as the data was too prevalent to simply be ignored. As such, the analysis introduced variables based on this reality. However, since the purpose of this dissertation is to indicate the general type of behavior demonstrated, rather than its inherent substance, data falling under political codes was examined in this manner.

In many cases, this was a retweet or a like given to a particular post. Others would react to posts that others had made or comment upon them. In general, the data indicates that their activity in the political sphere was no more specifically directed than any other. Even once, by necessity, specific codes regarding political activity were added to the data scheme, most individuals did not demonstrate a marked change in their overall data based upon the amended scheme. This seems to indicate that for many, responding to specifically political messages carries no more or less psychic weight than responding to any other type of message. Rather, they tended to show the same overall patterns of behavior and interaction with any type of message. What is more important for them are some of the other variables identified in this study: message frequency, audience closeness, etc. In other words it is more important how and why and when they react, rather than to what they react.

Two respondents in particular viewed Twitter specifically as a venue for political discussion and message passing. “I’m not really interested in people’s dogs. I’m interested in talking about things that matter,” as Participant 18 put it. Their data is especially striking when compared to others in that nearly all their activity was devoted to political activity of some kind or another. Their activity was markedly different from that of others in that they specifically sought out political discussion to engage in throughout the course of the session, while others tended to take a more contextual approach to their activities, following whatever discussions caught their attention at the moment. For the two more politically active participants, however, the interaction activity never strayed far from their main purpose, and they often specifically avoided what they saw as spurious or mundane activities. Indeed, while the data for most respondents tends to be fairly widely coded in terms of its purpose and pattern, nearly all the data for the political participants fell under the political category.

Given that these politically active individuals seem to be outliers, they warrant greater attention. In both interaction sessions, they mentioned that they use Twitter primarily as a means for communicating political messages that they deem to be important. Twitter provided a "soap box" for them to express themselves and their views. For both, their ultimate goal was to build as large a following as possible, in order to broadcast their messages as widely as possible. "I have no idea who half of these people are," continued Participant 18, "I just know they see what I post. That's what matters." Neither was intent on establishing and maintaining a network of social immediacy, but were approaching haphazard in their following activity. Both followed at least two new accounts during the interaction session, while no other participants were observed doing so. Furthermore, while other participants had a specific imagined audience (Marvick & Boyd, 2011) in mind when crafting their posts, the political participants did not. One was quite blunt:

> "I usually write my posts ahead of time. Then I have a bunch ready to go. I always have something ready so that no matter what happens. If [the news of the day] is about the budget or immigration or whatever, I have something ready to go and can post it immediately." – Participant 3

Contrasted with other respondents, Participant 3 provided an extreme counter-point. While nearly all used Twitter in a fairly casual, haphazard manner, he was calculating in the extreme. While most used it as a means to keep up with others, to read and respond to what they said, he rarely, if ever, actually took the time to read very many posts. Instead, his every effort was building up as large and broad an audience as possible, and he saw every other activity – liking, retweeting, following, etc. – merely as service to that goal. His existence caused several changes to the model, which needed now account for this type of behavior. However, these changes proved worthwhile, as the data demonstrated another, similar type of strategic interaction: professional development.

It is critical to also acknowledge the role of visibility – of just being seen – as a key issue here. That is, there is something critical to identity formation in its display before an audience. Self-presentation as a process includes more than simply the connotative meaning of the information conveyed, but a whole raft of additional information present in the context and environment in which that information is conveyed. The mere process of communication carries with it goals, risks, and unspoken norms the use or violation of which carries with them large amounts of meaning. According to Erving Goffman (1967), "When a person volunteers a statement or a message, however trivial or commonplace, he commits himself and those he addresses, and in a sense places everyone present in jeopardy" (p. 1) . Communicators are always engaged in impression management, and acting with the goal that others might have or form a positive impression of them. There is a constant risk of violating the norms and expectations that structure these interactions, and, in so doing, offending the other person in some unintended way, or presenting oneself in ways that are not desired. As such it is important to consider the structure and nature of an audience when evaluating the meaning and importance of acts of communication.

Professional development.

Several respondents discussed the place of Twitter in their lives as one primarily aimed at professional development, professional networking, and other job-related activities. While these respondents were not nearly so mercenary with their activities – even, on occasion, actually taking the time to read what they retweeted – they still demonstrated similar levels of forethought and planning as the political participants. One participant in particular was notable in this regard, as she planned all her Tweets and links ahead of time, unlike most other respondents. Participant 12 pointed out, “Tweeting for me I consider part of my job. It’s how I get new leads and meet new clients. So I take it pretty seriously.” She noted that it was important as part of her occupation to maintain an active social media presence, as it was import to her for generating new employment opportunities. She also noted that it was important to her that she limit personal communication and contact on the site for this reason, so as to limit the potential exposure of personal details to potential business colleagues. “I try to keep things pretty professional. That’s important to me,” Participant 12 explained.

As such, her use of Twitter also differed from that of most other participants. She tended to be deliberate and even strategic of her use of the site. For example, it was important to her that she maintain a balance between the number of people she followed and people who followed her. Were either of these numbers to get too high, she would take steps to bring them back in line. She also spent a good deal of time deciding on whom to follow. During the interview, she explained that she would often seek out individuals on Twitter that posted quality content or made interesting comments. She would then re-tweet these posts, and use this as a way to increase the amount of information she was posting. “I like to find things that maybe people haven’t seen before. I think it helps me stand out, rather than just being the 9 millionth person to tweet out whatever was at the top of Reddit that day,” Participant 12 explained. This contrasts starkly with the political posters, both of whom were much more interested in the raw size of the audience they interacted with. Rather the professional developer was much more concerned with her imagined audience, and deliberately sought out new information or links to Tweet out that they had perhaps not seen before. She saw this both as a way to increase the size of her network, but also as a way to solidify her reputation within that network (see also Madden, 2010).

Other participants would occasionally discuss employment matters, but for most it was not as central to their activities. Particularly in the information seeking task, several individuals decided to browse their colleagues’ networks in order to learn about other individuals in their own company, their business contacts, and so on. The use of social media for this task was a common one, and all the participants who chose a business contact to learn about mentioned that it was common practice. While most echoed the developer’s point of view that personal and private lives are best kept separate in public forums like Twitter, several pointed out that this is difficult at times to achieve. This was indicated by the research. While during more structured interviews most participants repeated this concern, during the interaction task it was apparent that in practice this boundary is a difficult one to maintain. For most individuals, browsing their personal and professional networks was an almost identical task, and with the exception of those few (especially the networker) that maintained the salience of this difference at all times the behavior observed was identical. Most individuals exhibited the same general interaction patterns regardless of the general category of relationship; however within each category interaction behavior varied greatly, though again to similar degrees.

Social connection.

Nearly all respondents said that Twitter was an important venue for them to both establish and maintain relationships with others (see also Burke, Marlow, & Lento, 2011). However the nature of this connection varied widely. This study found that there were in the data some exceptions to general patterns of behavior, however for the majority of participants difference was more a matter of degree than of type. All participants tended to follow similar patterns of interaction with Twitter. The differences were primarily in the nature, type, and frequency of these patterns.

For instance, respondents varied in the type of social groups they communicated with most using Twitter. For some, the platform served primarily as a means to maintain relationships with existing, offline social connections. These individuals tended to use the platform to build and maintain these relationships using a variety of interaction types, but most commonly indirect methods, including liking. Individuals noted that these were important means of maintaining and active “presence” or “activity” with their social groups, in order to maintain existing bonds. Participant 4 noted somewhat facetiously that without platforms like Twitter, “my friends might forget I even exist!” For others, however, Twitter provided a platform to interact with a wider variety of people that they know only causally. “That’s the great thing about it, you can follow literally anyone, you don’t actually have to even know them,” pointed out Participant 18. For individuals like these, their communication networks tended to be far broader, and their following patterns far less often reciprocal. For these participants, their audience was much broader and covered a range of interaction types, from celebrities they follow with whom the communication is entirely one-way, to work colleagues, to family members. Thus the size and nature of their audience is much different.

It also provided a mechanism of keeping up to date with the activity of others. Despite the type of audience that each individual has, their mechanisms for communication also change. Data indicate a variety of interaction types, falling along a spectrum of intentionality. Some participants, like the politicians and the networker, were incredibly strategic and forward thinking in their communications. However others tended to be much more carefree. Most respondents stated that they didn’t think too hard about what they communicated, or when, especially when it came to indirect communication. They stated that their communications tended to follow fairly predictable patterns. For example, some only ever posted when there was major news or a life event. Others, however, would post several times a day no matter what the circumstances. Thus the intentionality of posting was another variable that indicated interaction behavior.

Usage patterns also diverged along temporal lines, and these differences could have stark meaning in the minds of participants. Use appeared to be fairly contiguous, but those at the higher end of the frequency spectrum tended to be much more aware of these types of patterns. Those respondents who reported the most use of the platform on the initial screening survey (daily use or higher) reported deep meaning conveyed in the chronemic effects of message-passing. They pointed out that because they used the platform so often, they were aware of who also used it the most, as they saw these messages appear in near-real time. Furthermore, they noted that there was a strong expectation among this group that message response and interaction would be prompt. They noted that compared to other CMC platforms (and specifically email) Twitter moves with great rapidity, with messages leaving the front page of their screen in minutes or less. As a result, in order to both see and be seen by those with large amounts of activity on their streams, responses, comments, retweets, and likes all must come in very short order. Because of this, the most active Twitter users point out that the custom is that interaction must come within minutes or not at all. This expectation is so severe, that a few even noted that late or non-existent replies to messages might be seen as a deliberate slight. Participant 7 pointed out, “You know immediately. If people aren’t liking or commenting on your post then you probably screwed up.” Thus, to those at the highest end of Twitter usage, not only frequency of reply but its reciprocation is a deeply-ingrained norm.

Identity maintenance.

Individuals also discussed how they establish and maintain a digital representation of the identities they wish to project online. As has been discussed, the level of salience of the strategic thinking that goes in to how carefully individuals craft their identities varied widely in the data. On one extreme was the networker, who has spent years carefully cultivating her image on Twitter, ensuring that it meets the high standards she has set for her performance on the platform. On the other hand there were at least three respondents that, though they reported using Twitter at least several times a week on the selection survey, seemed to post very little at all, meaning that what exists of their identity on Twitter is very scant. These individuals all reported using the site primarily to follow the news and events of others, and one made explicit mention of attempting to remain as anonymous as possible. To a certain extent, perhaps, this pattern of behavior represents its own identity.

Despite these variances in strategy, the data make clear that all respondents were aware of their online persona to at least some extent. When asked to make judgments about the personalities of others in the directed task, none had any trouble doing so, and nearly all immediately linked this behavior to identity formation in some sense. That is, it was common for individuals to shift the discussion from judgments about others to judgments about themselves. Indeed, oftentimes these sorts of judgments were related to online behaviors. Participants typically characterized individuals based upon how they interacted with others on the service. For instance, Participant 20 pointed out that one member of his feed posted pictures of his lunch almost every day. “He’s one of those types of people,” he said. When pressed further, the participant became a bit reticent, but finally indicated that he judged this person to be a bit conceited and even obnoxious. “There are certain types of people that think that everyone will care about what they eat. Like, who cares? I would never post a picture of my lunch,” he pointed out.

Information seeking.

Information seeking practice is a central research question, and explored in some depth in all of the interviews. Each participant was provided with an information seeking task to perform, and also interviewed in a structured format about how they learn about others in SNS in general and Twitter in particular. This protocol was designed to investigate links between information seeking and uncertainty reduction as described above. Specifically, the goal was to discover the patterns of information seeking behavior that exist in SNS interaction, and to investigate how they can be modeled in a quantitative setting.

The interaction task was performed first in order to reduce the introduction of bias due to priming. Participants were each asked to browse Twitter as they normally would for a period, then also to carry out specific tasks. For the majority of the participants, activity on Twitter could be characterized as a semi-directed search. Individuals typically would begin the task with some ideas about how they would carry it out, while modifying their trajectory based upon the context and the results that had discovered along the way. For example, when asked to select an individual in their network and describe their impressions of that individual, almost all would immediately go to that individual's particular feed to browse back through previous messages. However, very seldom was any of this information read or examined. For the majority of participants, there was a clear indication of the target's conduct and nature on Twitter. Especially for those most active on the site, locating an individual about whom they had concrete mental images was not a difficult task.

On the other hand, when asked to investigate a new person or someone not well known to them, a similar set of steps was carried out, but the intentionality of the reading was much higher. Individuals paused, sometimes for a full minute or more, to read back through the individual's posts in order to make fully-formed opinion. Oftentimes, participants would go back through the target’s previous interactions with others, noting the type of relationships they seemed to form on the site. Principally, this took the form of browsing through at-reply threads that are common on the site. Another important resource was the browsing of that individual’s followed list. This turned out to be far more important to respondents in forming an opinion of the individual than browsing through whom followed the target. Several respondents would use what they knew about prominent personalities on Twitter in order to locate their impression of the target. By determining which type of content the target was exposed to most, participants felt they could gauge what was important to the person to know about, and thus help form an impression of them.

When asked specifically about how they browse Twitter and gather information on the individuals they encounter there, participants generally indicated that their most common form of information seeking was of the passive and extractive varieties. Several patterns of responses indicate each. Participants often noted that they liked to keep Twitter “on in the background” so as to provide something like a pseudo-social interaction while they were engaged with other tasks. By having Twitter open on their phone or in a browser window, participants indicated that they felt a sense of something like “community” no matter where they were. This was often manifest in interactions sessions, when even during directed information seeking tasks individuals would often note other messages that were appearing in their feed, even as they undertook their main task. As such, the platform provides a mechanism for passive interaction with one’s social circle, even as more directed, active tasks were being taken. Indeed, several participants noted that the directed tasks seemed “strange” or “unusual.” Actively seeking out information about an individual was foreign to many participants. Rather, they would generally find new people to follow, or new stories to investigate, via their passive consumption of the awareness stream.

Indirect communication.

Another central theme of the research was the use of indirect means of communication in the establishment and maintenance of online identity, as well as its use in both information seeking and impression formation. As noted in the literature review, existing literature is relatively scarce concerning the use of indirect forms of communication such as liking, retweeting, etc. (e.g., see McEwan, 2013; Oh, Ozkaya, & LaRose, 2014). This is particularly true with respect the use of these forms in modern SNS platforms, which are increasingly reliant upon their use. In many contemporary platforms such as Instagram, Snapchat, etc., video, graphics and emoji have come to complement a.nd even supplant text as the dominant form of communication.

The data support the importance of indirect forms of communication in SNS, particularly in one as high-volume as Twitter. Nearly all participants reported relying upon liking, retweeting, and other forms of indirect communication as one of their primary means of message passing. Interestingly, the data further demonstrate that this importance is highlighted at the heaviest and lightest ends of the integration spectrum. Those who use Twitter heavily were both reported relying upon likes using words like “heavily” and “frequently,” while the same was also true for those that used it rarely. However there were obvious differences. Those who were sporadic users reported wanting to remain relatively anonymous on the service, preferring primarily to read what others post rather than contribute much of their own, and as such limited the exposure of their contribution in this manner. On the other end, those who contribute a great deal often report that they want to feel as integrated as possible, but fully commenting or direct messaging a response would feel like too much effort or commitment. Indeed, much like the unspoken rules of chronemics that proved a great deal of unspoken meaning to heavy users of Twitter, the amount and content of a reply was also reported to carry similar weight. One particularly heavy user pointed out that the more time and effort a reply appeared to take, the more weight or importance it might be given, and thus one must carefully gauge the response that others might have against this expectation. Participant 13 explained, “You can kinda tell what kind of tell what kind of a person is by how they tweet. Like if you get a huge, multi-tweet reply you know its one thing, but if it’s just a heart you know it’s another. I try to keep things pretty light, so I’m usually just doing hearts.”

Conclusion

This chapter presented major findings from the qualitative data collection sessions, conducted using a talk-aloud protocol. It first provided details concerning phase one of the study, which primarily served to validate the survey instrument and to establish a baseline codebook. Next it documented what occurred during phase two of the qualitative interviews, the more substantive portion. Several major research themes were presented related to the research questions.

One major purpose of the qualitative phases of this dissertation was to provide insight into behavior patterns in social media in order to begin the process of selecting a set of variables that would drive the development of a model during the quantitative portion of the analysis. Furthermore, the qualitative results would amplify, extend, and color the quantitative results. As each process continued, these phases informed one another. For instance, as the importance of certain types of content in tweet messages became apparent, the method used for the quantitative phase was adapted to better account for this. Further details on these and other features of the quantitative phase results and discussion can be found in the next chapter.

Chapter 4

Quantitative Findings and Discussion

Introduction

This chapter describes phase three of the dissertation, a quantitative phase which was built upon the insights gained during the qualitative phases and analysis. Phase three aims to reinforce and amplify the findings of the qualitative talk-aloud protocol phase, while at the same time developing a framework for analysis that can be applied to future studies in this area. It is the presumption of this dissertation that there are a set of detectable and replicable strategies that individuals use when constructing their online identities, and in turn when inferring impressions of others based upon the signals that are presented via the application of these strategies. Based upon this presumption, the goal of the quantitative portion of this study is to develop a model that can be used to establish and measure these strategies. Though the dissertation is primarily driven by research questions rather than hypotheses, it is hoped that this model can lead to the development of testable hypotheses with regard to identity management and impression formation in SNS.

Overview of phase three

To move from the qualitative data phase to the quantitative phases, the first important step was to set the bounds of the subject of inquiry. It is often said that the great gift that social media data presents researchers is that it enables them to measure nearly anything on a monumental scale, while the great curse of this data is that it enables researchers to measure nearly anything on a monumental scale. The sheer vastness not only of the data itself, but of the potential variables, measures, and statistics that can be drawn from it, makes establishing strong theoretical boundaries (as detailed in Chapter One) a critical first step in any such research endeavor.

This chapter outlines the main findings of the quantitative phase of the analysis. Primarily, this analysis was aimed at building a model of interaction behavior in social media with an emphasis on identity formation and impression management. Working in concert with the qualitative phases, this analysis was an iterative process. As data was collected and analyzed from both the qualitative and quantitative phases, the model was updated and refined, then tested again. Once the new model was in place, more data was sought to further refine and sharpen its focus. In this way the model was in some senses both the product and the process of the quantitative analysis.

The Development of a Model

The development of a model that could be used for the analysis of social media data was the primary goal of the quantitative analysis. As this dissertation is primarily one based on the addressing of research questions rather than null hypothesis significance testing, this development was the main purpose. That is, the goal was to develop a framework which could be used to answer such hypotheses in the future, based on the understanding of the data that the model provides. This development process was an iterative one, and each step was repeated several times as the model was improved. These steps include the identification of variables, the operationalization of variables, the coding of data on these operationalizations, data reduction, data clustering, and interpretation.

Variables

Several variables related to identity management were included in the model. These variables were ultimately chosen due to their suitability as bases for the model for both theoretical and statistical reasons.

The first major group of variables that were measured are related to Latent Dirichlet Allocation (LDA) topic analysis (see Rehurek & Sojka, 2010, Zhao, et al. 2011, Hong & Davison, 2010). As discussed in Chapter Two, LDA is a method for selecting groups of words that are related to one another due to their inferred relationship to a hypothetical topic. This association is inferred through a Bayesian calculation of their priors. As also noted in Chapter Two, there are several challenges associated with using LDA on Twitter data, which is further discussed below.

A final challenge was settling on the number of topics to be measured. As the number of topics increases, the relative probability distribution of each begins to decrease, as does their expected associations. As such, it is important to determine the number of topics to choose in order to establish an accurate model. One common method of establishing the number of topics to model is using perplexity. In the specific realm of natural language processing, perplexity is a measure of the amount of information carried in language, typically measured per word, using “information” here in the Shannon (Hogan, 1995) sense: the decrease in entropy that each word conveys. In this case, the greater the entropy of each word, the more difficult it would be to select randomly from a bag of words, and thus the more information it conveys. Perplexity is a way of evaluating models of language in natural language processing by calculating the extent to which the model decreases the entropy of each word. The better the model, the better a job it will do at reducing the amount of uncertainty there is in randomly guessing the next word, thus reducing the overall entropy of the entire document and providing a measure to evaluate the model.

The first important step in LDA is to decide upon some total number of topics to detect, a value typically called *k*. However in order to avoid confusion with a separate value *k* that will be used in the k-means cluster analysis below, in the present discussion the number of topics chose in our LDA analysis *t*. In order to determine a reasonable value for *t*, the number of topics in our corpus to use in the LDA analysis, perplexity was calculated for several values of *t*. Note that LDA across large data sets is a very computationally-intensive process. Ideally, this calculation would be done for every possible value of *t*, however computational resources limited the potential measurements that could be calculated in a reasonable amount of time. Figure 1 is a plot of perplexity values for values of *t*.

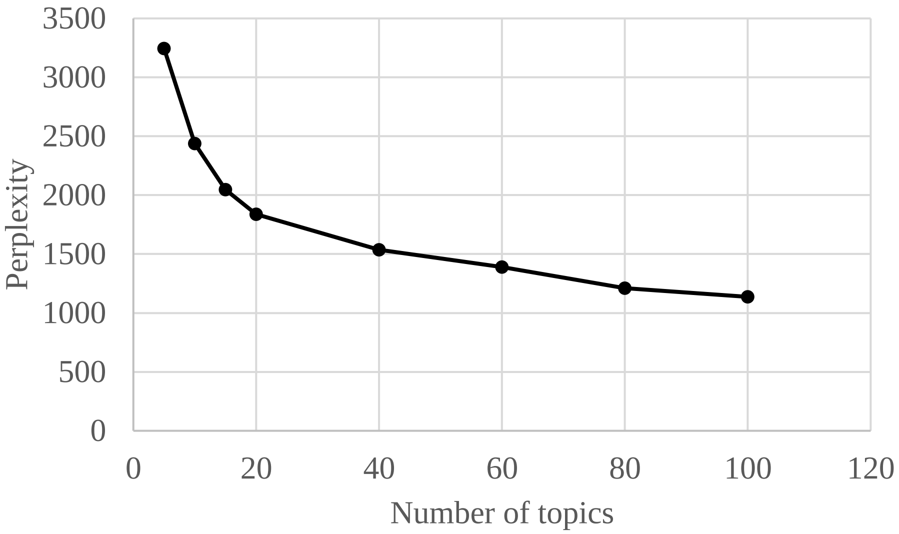

Figure 1. Perplexity. The Perplexity of the Twitter data as topics in the LDA analysis increased.

While perplexity is a useful guideline, it does not provide a definitive value that is "best" for LDA analysis. As Figure 1demonstrates, with increasing values of *t*, the perplexity is strictly decreasing. While the analysis was halted at *t*=100, it is feasible that for even larger values of *t*, the perplexity may decrease even further. However, other factors must be included in the ultimate choice. First, there is the cost of the analysis. The run of LDA at 100 topics took nearly eight hours to complete, and ultimately would be infeasible to run multiple times in order to, for instance, perform some validation or post-hoc analysis. Second and more important is the amount of data that is added to the final analysis. For the purpose of the current research, the specific nature of, for instance, political discussion, is largely irrelevant. Rather it is more important that some type of political discourse is taking place, not its specific content. While increasing the number of topics that are chosen does decrease perplexity, it does little to increase the accuracy of a measure of political discussion. Indeed, in many cases it decreases the accuracy. Consider Table 3 which shows some of the topics detected during the *t*=100 run of LDA.

Table 3

LDA Topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Trump	Obamacare	Kelly	Sun	Bacon
Charlotte	Health	Jones	Forest	Sink
Blame	Repeal	Sandy	Candle	Garbage
President	Insurance	NBC	Watching	Wrap
Rally	Fair	Conspiracy	Summer	Chicken

While some of these topics have clear bearing on the inclusion of political discourse, they are more specific than is called for, or even desired, in this dissertation. And at least two seem to be mostly noise, which can happen in large-topic corpora runs of LDA, particularly in the case of Twitter data where vocabulary size tends to be quite a bit smaller than typical English writing. In these cases, the topics identified are often spurious or coincidental correlations, and not indicative of an actual, existing underlying topic. After reviewing the topics present in each of the runs, the 20-topic run was selected for having a good balance between accurately tracking with discourse in the corpus, capturing the discourse that is germane to the variables being measured accurately, and computation limitations.

For the purpose of the model, these topics were divided into several categories, each of which could be tracked with a separate variable. Again, the purpose of these variables is not to track the specific nature of the content, but to identify its mere presence. As such, these variables were made a broad as possible, based upon the data gathered in phases one and two of this dissertation. Specifically, individuals in phases one and two spoke to different “types” or “personas” of individual behavior they observe. Based upon these findings, general categories of use were identified, and topics assigned to each Table 3 displays the relative frequency of each of these topic categories, as well as the cumulative distribution of each of these categories across the topic model distribution. These frequencies are functions both of the distribution of these topics in the data, as well as the overall probability distribution of each topic across the entire data.

It is important here to note that the frequency of topic distribution has very little correlation with overall communication content. LDA is designed to identify topics, not general content. As such, political communication – which by nature tends to be quite topical and narrow in its vocabulary – is over-represented. Users that engage in more social conversations, which use a far broader vocabulary, will be under-represented. As such, Table 4, below, also includes coverage: the extent to which each topic category is represented in the corpus of users.

Table 4

Topic Coverage

Topic category	Overall frequency	Coverage
News / information	.28	.21
Social discourse	.20	.28
Personal updates or opinions	.35	.40
Other	.17	.11

The final model included variables measuring topic frequency in each of these categories for each user, based on the topics identified at k=20, excluding the "other" category, which for the purpose of this study are considered noise. Again, the purpose here is to measure relative frequency of general types of content, based on the findings of the qualitative study which suggest that these types of content are of particular importance with respect to identity management. Further, it is worth noting that this analysis can be characterized as vague at best. Many other studies have done more work into modeling the content of Twitter messages (e. see Zhao, et al., 2011; Hong & Davidson, 2010). This dissertation's goals are to generally categorize content, rather than specifically model it. Rather, the goal is to model the types of interaction behaviors that lead to the production of that content.

A number of other variables were considered for inclusion in the model based on the findings in the qualitative study. As detailed above, this study identified several strategies and behaviors that individuals both used to form impressions of others, and also used themselves to manage their identities. Those included in the model are presented in Table 5.

Table 5

Variables

Variable	Description	Operationalization
Directionality	Is the message directed at a particular audience	Relative ratio of at-mentions
Reflex	Frequency of reflexive replies	Ratio of replies to at-mentions
Intentionality	Ratio of indirect to direct communication	Ratio of likes to posts
Symmetry	What is the shape of the individual's audience	In-degree to out-degree ratio
Density	How dense is the individual's network	Adjusted network density
Content	What the content concerns	Topic modelling
Frequency	How often does the activity take place	Logarithmic rate measure
Linking	Relative frequency of informational posts	Relative ratio of retweets containing links to those not
Co-activity	Synchronicity between communications	Logarithmic decay time ratio of replies to at-mentions
Message hold	How varied is the content that is posted	Mean number of topics

Several variables were also examined for inclusion, but excluded for either theoretical or statistical reasons. One variable that was examined, but not included in the model's direct associations, is account creation recency. For several of the other variables measured, there was a strong correlation between recency of account creation and other measures. For instance, Figure 2 shows activity frequency as a function of time after account creation for those accounts for which such data was available (given the age of the data examined, this was 23% of accounts studied).

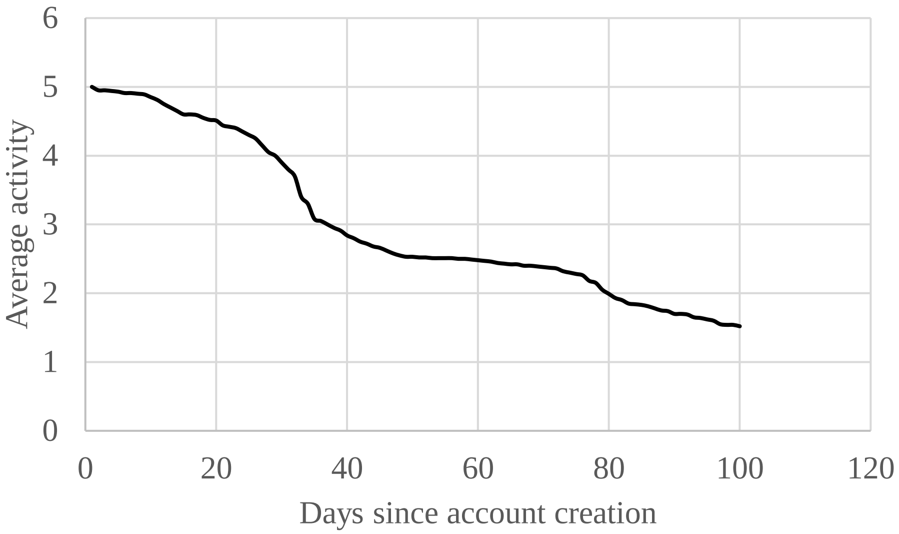

Figure 2. Average Daily Activity. Activity of Twitter users after account creation.

Account creation recency correlates with account activity at a Pearson r = -.82, which could clearly alter the model. However, based upon the qualitative, phase one findings, this dissertation is not designed to address questions of recency. Rather, it is interested in examining questions of behavior and identity management strategy regardless of temporal effects. Given the strong correlation between recency and several other variables, including activity, it is important to control for this variable. In order to do so, an ordinary least squares (OLS) regression (Freedman, Pisani, & Purves, 2007) is calculated between the account recency distribution and each of the variables considered in the model. This results in a partial relationship between the model and the two variables under examination. When considering the beta correlations between each variable and the main effect, can then control for the effect account recency by partialling out its variance. This same basic technique was applied to all other variables that had correlation effects but were not included for theoretical reasons.

Additionally, there were several instances of increased activity in political and news discussion in the corpus during the period of data collection. These include a June 29, 2017 Twitter post by President Donald Trump criticizing television host Mika Brzezinski and her co-host and fiancé Joe Scarborough, in which they were characterized as “low I.Q. Crazy Mika, along with Psycho Joe” and claimed that during a meeting with Ms. Brzezinski that, “she was bleeding badly from a face-lift.” Figure 3 shows the frequency of tweets in each of this study’s topic models around this date.

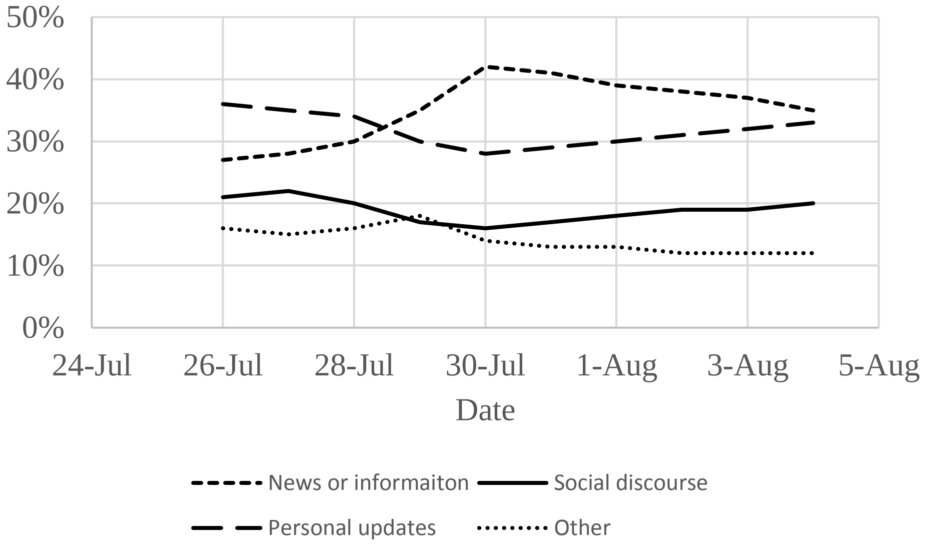

Figure 3. Topic Categories. Topic category proportion over data collection period.

Because data collection occurred during a relatively narrow time window of four weeks, such exaggerated spikes could have a clear bias on the data. It is therefore important to control for such effects. Further, these spikes are likely greater due to the enormous influence the President has on Twitter. As the above data demonstrates, when he Tweets, it has an influence. Yet at the same time, this is not so simple. Statistically discounting political tweets simply because they occur in spurts is plainly wrong. Politics occurs in spurts, as issues are raised, discussed, debated, and then the discussion moves on. Again, the data demonstrates this. At the same time, it is important for the sake of validity to attempt to normalize these effects. During the data collection period, as described in Chapter Two, there were several major new items that vastly eclipsed the Brzezinski affair in terms of Twitter activity, including racially charged violence in Charlotte, NC, several major scandals that resulted in firings at the White House, and a total solar eclipse. Therefore, additional longitudinal data was used to smooth data in the smaller sample window in order to attempt to normalize it to historical trends. While it is impossible to adjust topic distribution, coverage factors can be adjusted, and these were so according (say how) to a moving average across the entire extended sample period.

Ultimately, the model was constructed using the above variables while statistically discounting or controlling for the others. This creates a measure of identity maintenance based upon the qualitative findings of this study while managing, as best as possible, for other interactions. Once the model was in place, the next step is to test it against the corpus, in order to see if it provides insight. There are several ways to proceed, including regression, factor analysis, and others. Ultimately, cluster analysis is employed in order to attempt to identify and replicate identity management strategies that were discovered in the qualitative research.

Clustering

This study uses k-means clustering in order to identify clusters (Scott, 2017). The goal being to identify patterns of identity formation and maintenance behaviors suggested during the qualitative analysis. By relying upon the qualitative data in phase two to create a system of measures that indicate these behaviors, phase three attempts to provide a framework for measuring the prevalence and nature of these strategies. Thus, following the research questions (see Chapter One), the aim was to create a framework that may be useful for identifying the strategies discovered during qualitative analysis in phase two.

K-means clustering is a process of determining data grouping based upon a central point within each sub-group of data. As discussed in Chapter Two, based upon some value *k*, a random set of central points are created with the data matrix. Based upon the Euclidean distance between each point in the data and these central points, new estimates are created to new central points. This process is iterated until no data points are ambiguous to their central points above some threshold, implying that the central points accurately describe the clusters.

The central question, then, is *k*. There are several methods to evaluate an appropriate k-value. The first, and least practicable in the current project, is using every possible value for *k*, running the analysis, and calculating fit. A better alternative is using the "elbow method," (add cite) which relies upon the F-test to arrive at a satisfactory level of variance explained by the model. More specifically, when the relative ratio of variance explained by the k-cut each increasing value of *k* is less than the overall variance of the variable in question, this is denoted as the "elbow point" and is considered a satisfactory indication of k-cut estimation. From there, more in-depth k-cut analysis may continue in order to arrive at the optimal. In this study's model, this point was reached at k=5, after which the model at k=6 increased variance explained by .04, while variance within the distribution was .15. Figure 4 illustrates the variance explained in the model as a factor of *k*.

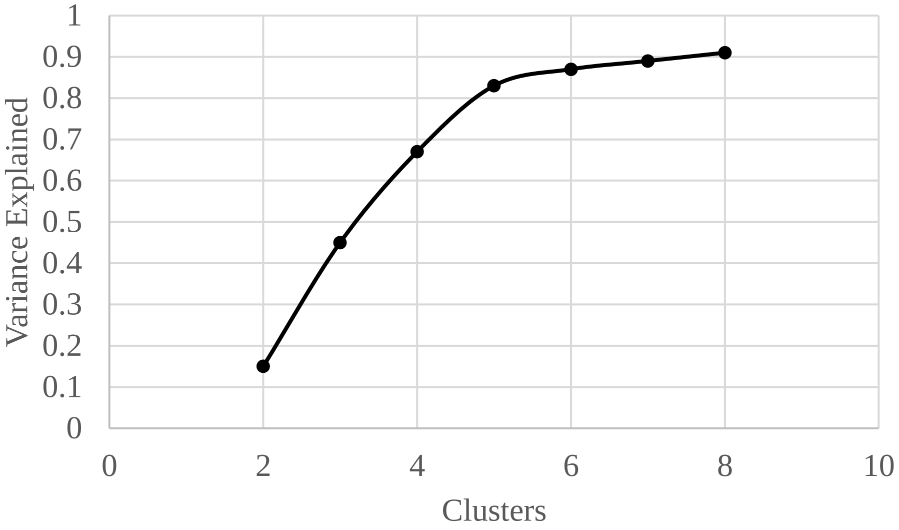

Figure 4. Variance Explained. Variance explained by the model for increasing *k*-cut clusters in the model

A third alternative for the selection of *k* was also employed for verification purposes. Ding and He (2004) have suggested that using the component vectors calculated during PCA data reduction as the basis for a cluster analysis provides consistent results (often with a difference of less than .5% in variance explained) at a fraction of the computing cost. Given that computing time and efficiency was a priority in this project, this approach was adopted. Usefully, this also gave validation to the perplexity-based selection of cluster size. After reaching a satisfactory value of *k*, the clustering is applied. The central points for each of the five identified clusters, and their values, is presented in Table 6.

Table 6

Cluster Values.

-	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Direction	0.87	0.56	0.34	0.23	0.64

Ratio of mentions	0.85	0.63	0.54	0.49	0.68
Intention	0.23	0.68	0.89	0.13	0.66
Reply ratio	0.86	1.10	1.28	0.80	1.09
Symmetry	0.56	0.13	0.45	0.45	0.78
In-out degree ratio	1.03	0.80	0.98	0.98	1.17
Mean news topics	1.58	2.62	4.22	2.82	2.21
Mean social topics	5.02	3.05	1.88	3.43	3.64
Mean personal topics	3.39	1.75	2.55	3.58	4.22
Frequency	0.91	0.23	0.85	0.24	0.43
Mean post per day	0.72	0.25	0.62	0.26	0.34
Linking	0.23	0.34	0.67	0.23	0.55
Mean links per post	0.25	0.30	0.46	0.25	0.39
Co-activity	0.75	0.56	0.55	0.33	0.39
Mean time between replies	3.33	2.84	2.82	2.38	2.50
Hold	0.34	0.46	0.77	0.42	0.53
Mean Topics	3.65	4.26	6.48	4.05	4.65

This representation of the model at *k*=5 cuts was tested for goodness-of-fit using a one-way ANOVA for each of the cluster centers against the means for each variable for each cluster. That is, for each variable, the cluster center estimate based on the k-cut in each of the 5 clusters was compared against and the true calculated mean for each variable. The results of this test are presented in Table 7.

Table 7

Goodness-of-Fit.

	F	p
Direction	2.45	0.05
Ratio of mentions	2.67	0.04
Intention	2.64	0.04
Reply ratio	3.34	0.02
Symmetry	2.73	0.04
In-out degree ratio	3.1	0.03
Mean news topics	3.6	0.01
Mean social topics	3.31	0.02
Mean personal topics	2.56	0.05
Frequency	3.72	0.01
Mean post per day	3.64	0.01
Linking	2.95	0.03
Mean links per post	3.56	0.01

Co-activity	2.87	0.04
Mean time between replies	2.86	0.04
Hold	2.56	0.05
Mean Topics	2.45	0.05

This model appears to have a solid basis for goodness-of-fit along statistical lines, while at the same time satisfying this project's aims of developing a model of identity management in social media driven by ground truth data generated during qualitative interviews. While several other variables might include the, for instance, predictive power of the model, they would fail to reflect the goal of the project in by sacrificing theoretical validity. Based on this statistical analysis, as well as the interview data, it seems likely that this research has identified a useful model for understanding identity management in social media.

Measures

Presented here are measures of the clusters resulting from the application of the model. These measures represent the outcome of the model as applied to the data. The first measure is the overall proportion of the population of the sample attributable to each cluster. The second is the between-ness centrality of the cluster, as calculated by the mean minimum distance between nodes in the subgraph representing each cluster. The third is the mean variance load carried by the vector representing the center point of each cluster. The fourth and final is the goodness-of-fit parameter as determined by an ANOVA test.

Table 8

Cluster measures

Cluster	Proportion	Centrality	Load	Fit (F)
One	.13	.0077	.93	3.76
Two	.24	.0134	.87	3.51
Three	.11	.0099	.81	2.87
Four	.26	.0105	.76	2.31
Five	.26	.0089	.91	3.03

From these measures we can see that personas four and five where the largest, comprising more than half the total population. However, persona two was nearly as large, indicating that at least half the individuals included in that portion of the model are not active posting participants on Twitter, but nonetheless are heavy consumers. Given that these individuals were most closely identified with the passive mode of information seeking in the qualitative portion of the study, this finding lends support to the assertion that passive strategies are of increasing importance. That said, centrality was also the lowest for this group, indicating that their social network tended to be looser in organization than those of others. Those in the other personas all had lower centrality, especially persona one. The data therefore suggests some correlation between passive information seeking and a lower level of connection to social circles. No definite conclusions can be drawn at this point, but this correlation provides an interesting avenue for future research.

Conclusion

This chapter discussed the quantitative aspects of the development of the behavior model. It included discussion of the data, a description of the variables applied, and a summary of the steps taken in the analysis. In order to move from the adjusted data to a cluster model, the Ding and He (2004) approach was used to apply PCA as an indicator of cluster variance, which helped establish a guide for the cluster cut as well as provide efficient data reduction. This chapter also summarized the content analysis procedures using LDA on Twitter data, and some of the challenges this presents. This discussion for the quantitative findings, along with the Chapter Three’s discussion of the qualitative findings, concludes the major reporting of research findings. In the remaining chapters, data findings will be synthesized and interpreted.

Chapter 5

Synthesized and Discussion

Introduction

This chapter discusses the implications that the qualitative and quantitative phases of data collection and analysis had with relation to the research questions. This chapter serves as a discussion of how these two phases informed a set of conclusion drawn from a synthesis of data generated by each of the methods. It presents and outline of the main research findings presented in the context of the model that was developed as a result of analysis of data from all phases. It then provides further insights that were gained through the application of this model to the directed information seeking task that qualitative participants from phase two were asked to complete.

This chapter begins with a summary of the findings from the qualitative and quantitative phases of the study. It then reviews the important variables in the model, and discusses how these variables combine to form a framework for the categorization of interaction behavior in SNS. The chapter discusses how both qualitative and quantitative findings were used in the creation of five "personas," which substantively comprise the model that was the generative purpose of this dissertation, a model that here is termed the Identity Formation / Information Seeking (IF/IS) model. These five categories were created through a synthesis of the findings. Once a set of variables for analysis was identified in the qualitative portions of the study, these variables were measured and grouped in the quantitative portion. As described in Chapter Four, by using cluster analysis based on k-means clustering via PCA reduction, the model became stable at five clusters. By supplementing this numeric data with the words and ideas of the qualitative participants, the model seeks to move beyond simply numeric description to a more cohesive understanding of the behavior patterns that characterize each of these clusters. To this end, for the purposes of the continued analysis and to highlight the mixed-method approach that informs these categories, this dissertation employs the term “personas” to describe them, rather than “clusters.” For orientation, a brief description of each is provided in Table 9.

Table 9

Personas

Persona	Description
One	Always-on, heavy Twitter users. Members of dense, very active social circles
Two	Heavy browsers, infrequent posters, rely heavily on indirect communication
Three	Use Twitter to engage with information and personalities, engage with interest groups, collect and share content
Four	Light Twitter users. Sporadically use the service to check on and discuss major news or life events
Five	Moderate Twitter users. Socially engage with their circles, occasionally follow a couple celebrities. Interact in “batch mode.”

Mixed-Method Synthesis and Discussion

This project relied upon a mixed-method approach in order to attempt to build a model of identity management and impression formation in social media using the lens of information seeking as a way to approach the problem. In order to do so, both quantitative and qualitative phases were undertaken in order to build a model based on a deductive process, whereby insights grounded in qualitative data could be used to make inferences about quantitative data. Furthermore, the quantitative data analysis was also used to inform the qualitative data analysis by opening new questions and lines of inquiry.

One key aspect of the qualitative data collection and analysis was the identification of a set of variables grounded in observational data that could be operationalized to build a quantitative framework for the modeling of large data sets (e.g., Wakefield, Warren, & Alsobrook, 2011; McPferson, et al., 2012). This process was informed by other research in this area, which has successfully applied this approach. However it is important to bear in mind that the IF/IS model is not meant to be a panacea for the study of interaction behavior writ large in social media. Rather, it is a targeted framework meant to describe a specific type of social interaction related to identity management and information seeking. This is key to the purpose of the dissertation: to establish a connection between these phenomena and present a method for their examination. As such, variables present in the quantitative model were informed by the qualitative data, as they represent patterns of behavior observed during the qualitative sessions that could be operationalized in the quantitative setting. These variables were introduced in Chapter Two and discussed in greater detail in Chapter Four. It is important to note that these variables do not represent an exhaustive list of what could be studied in either a qualitative or quantitative examination of identity management and information seeking. By nature, this list represents a compromise between the exigencies of each.

A number of potential ideas rose from the qualitative phases that could not feasibly be tested in the quantitative portion given the limitations of Twitter as a platform. One interesting example that arose from the qualitative studies was the idea of negation, or what is not said. In several instances respondents mentioned that not reacting, commenting, etc. was sometimes as important a part of their interaction patterns as overt communication. Participant 7 noted, "People notice when you don't like one of their posts and everyone else in the family does." In particular, the heaviest users of Twitter noted that not receiving replies from individuals that were specifically mentioned in a post could be seen as an affront. Others mentioned that they will often deliberately tailor the type and frequency of their posts so as to maintain a particular identity on the site. One participant in particular was extremely strategic in this regard. However, many noted that they would not make certain posts based on the norms of the platform. One respondent noted that they would never post pictures of their children, for instance, out of privacy concerns. Others pointed out that attempt to avoid using "Twitter-speak" because they associate it with younger users of the site. For example, Participant 15 noted, "I try to use decent spelling so it doesn't come off too casual.” In sum, nearly all participants noted how they tailor their actions in various ways in order to project a particular identity, and that this tailoring process necessarily means the elimination of certain actions and practices as much as it means the inclusion of others. Testing for what is not said or done using trace data from social media streams is, quite difficult. The only data available regards the actions that were taken, not those that were not. For this reason, the qualitative data regarding negation is particularly relevant to the mixed-method approach.

However, there were patterns of behavior identified in the qualitative sessions that were able to be operationalized into quantitative variables that could be measured. It is important to note, however, that there exists a large gap between the conception of these variables and their operationalization. Twitter data is, by nature, sparse. The data available includes little more than the Tweet itself, some metadata, and lists of followers. As such, the data itself is not a direct measure of the attitudes and actions this project asserts they actually represent. It is asserted here that the operationalization of these variables is, at least, some indicator of the interaction behavior in question, and, as such, provides a means to identifying its presence in the interaction patterns of the individual in question. The following sections discuss these variables, the data that was gathered regarding them, and the implications surrounding this data, in greater detail.

Directionality

Directionality was the first variable identified. During the phase two qualitative sessions, individuals displayed a wide variety of imagined audiences (see Marvick & Boyd, 2011). Some individuals used the platform almost as a public messaging service, and included at-mentions in nearly all of their messages. Participant 6, in particular, noted that she almost always used at-mentions because it increased the likelihood of garnering interactions. She noted, “Sometimes if I don’t include mentions it feels like I’m just talking to myself.” For her, Twitter was one of her primary messaging platforms, supplanting text messaging and even email in many cases. For other individuals, this practice was less common. They tended to include at-mentions occasionally, but far from habitually. Several individuals pointed out that if they were Tweeting a piece of information, or a link that they thought someone in particular might find interesting, they would include an at-mention. This would indicate that there might be a correlation between mentions and linking, but this was tested during the data reduction process described above and not found to be significant, as reported in Chapter Four. However, the use of Twitter as a type of public messaging service was interesting and anomalous enough to warrant inclusion in the IF/IS model. Even though the majority of participants did not use Twitter in the way that the messenger did, including it would enhance the robustness of the model. As such, this variable of "directness" was included and operationalized as the log-normalized relative ratio of at-mentions per Tweet.

When included in the quantitative model, directionality became an important indicator of persona one, to the point that they might be characterized by this variable alone. Persona one members showed dramatically higher likelihood means for directionality and for ratio of at-mentions than the other groups. This group was characterized by a high amount of frequent interaction with Twitter, primarily for the purposes of social interaction, as they tended to post far less news-related content. This group also demonstrated a very high level of co-activity, indicating that they frequently interact with others in something approaching real-time. These individuals seem to reflect the behaviors that others like Participant 6 displayed during the qualitative interviews. For Participant 6, Twitter served as a primary mode of conversation, and she valued it for its immediacy and flexibility.

Intentionality

The second variable was intentionality, a measure of commenting and posting behavior as opposed to liking and retweeting behavior. During the qualitative interviews, particularly around the topic of indirect communication, individuals displayed wide variance in their behavior with respect to commenting and liking. For many, the use of indirect communication provided a convenient means for maintaining social relationships, but the way it was used varied greatly. For example, the heaviest use participants all noted that they will like almost anything that comes across their awareness stream from one of their close friends. “It’s sort of automatic,” Participant 7 noted, “If it comes from one of your friends or family, you’ve got to give it a like.” Others were more deliberate. One of the lower-use participants noted that they would like or retweet a big piece of news or major life event that one of their connections posted, but that was about it. They reasoned, “If you like everything, it sort of loses its specialness after a while.” Still others saw the feature as a way to get noticed by those they had aspirations to impress, be they celebrities, business contacts, etc. Several individuals noted that since the inclusion of the like feature, they have nearly ceased posting entirely. They use liking as the primary means of interacting with others on the site, and as an indicator that they are still connected and participating. These users all said that they use Twitter at least daily on the screening survey (see Appendices A & B), yet during the interaction tasks their posts histories were remarkably bare. For them, Twitter was a way to keep up with those in their social circle, and not primarily a means of communication. The use of indirect communication therefore carried a good deal of meaning for participants, and was used in a number of ways.

Intentionality differed as widely in the quantitative data as it did in the qualitative. Persona one, again mirroring the impressions of the heaviest users in the interviews, had a very high intentionality, indicating that they commented or posted rarely compared to the number of likes and retweets tweets sorted into this persona made. However, though higher than other groups, persona one was still lower in intentionality than persona two. If directionality defines persona one, than intentionality defines persona two. For these individuals grouped into persona two, the level of intention is lowest of all, yet their level of co-activity is relatively high. In other words, when they are messaged or mentioned by others, their reply comes quite fast, but it is almost always in the form of a like or retweet. Interestingly, this group was the second largest of all the persona groups, indicating that for a large percentage of Twitter users, this behavior patterns is common. Not surprisingly, this group also had among the lost post frequency, and also the lowest audience reciprocity, meaning that they had the lowest number of followers relative to the number of people they followed.

Symmetry and Density

Grouped together as measures of participants' audience network, symmetry and density measure network features of the Twitter graph. Specifically, symmetry is a ratio of in-degree to out-degree for each user, while density is the density of the subgraph made up of each user's 2-degree following network (that is, followers of followers). Initially, during the data reduction phase of the qualitative analysis, a correlation was discovered between these two metrics, and they were considered for removal or combination into a single variable in order to reduce the IF/IS model's complexity. However, as the qualitative interviews continued, it became increasingly apparent that the liking/lurking pattern of interaction behavior was an important one for several of the participants. Again, they noted that they mainly browsed Twitter to get updates from people they were interested in, both privately and publicly, and never made many posts themselves. They saw Twitter as a news resource – again meaning both news in their social circle and news related to public events – rather than a communication resource. However, they also pointed out that they did like items fairly consistently, and saw that form of indirect communication as their main form of input to the site. It became quickly apparent that capturing this mode of interaction in the model was important, and therefore both measures remained as they capture some interesting features of this mode.

While most personas have relatively stable normalized values for their symmetry and density, persona two does not. While persona two is characterized by a relatively low symmetry value, indicating that the ratio of the number of people they follow to the number of people that follow them is relatively high, they have a higher-than average density value, indicating that their social network is relatively dense. What this indicates is that though their connections to a particular subgraph are relatively loose and uni-directional, the connections within that subgraph are relatively tight. This is an unusual feature among the personas.

Among the other personas, difference between symmetry and density was negligible. However, between them there were contrasts. Personas one and five, the most frequent users of Twitter in terms of posting and commenting behavior, also had the highest symmetry score, underlining a well-documented link between network integration and communication activity. However, between the two, despite being the more frequent, more co-active, and therefore generally more “always-on” type of user, persona one had both lower symmetry and frequency values. This might indicate that persona one tends to have a wider, looser group of friends, or tend to be members of more social circles, but none of these can be entirely supported by the quantitative data. When reviewing this difference in the qualitative data, other possibilities emerge. Though they appear somewhat similar in the quantitative data, between those that demonstrated use patterns in line with personas one and five in the interaction sessions, behavior varied markedly. Persona one types seemed to leave likes and follows on just about everything they came across. For the most constantly engaged Twitter user, being maximally connected to the site seems to be the priority, and even during directed interaction tasks they would often take time away from the task to follow new individuals that they came across, especially if they appeared to be other frequent Twitter users. “This guy seems to know my sister, and he’s on constantly, so I’ll give him a like and a follow” was one such line of reasoning given by Participant 10. The persona five types – the frequent, but less-than-always-on pattern of behavior – tended to be far more discerning with whom they followed and what they liked. Even in the directed task when asked to learn more about someone in their social circle, the majority of respondents decided not to follow the target they chose, even after learning more about them as directed in the protocol. Thus, the qualitative data analysis seems to indicate a differing patterns of interaction behavior, which helps clarify the difference between personas one and five even though this difference is murky and not well explained by the quantitative analysis.

Content and Message Hold

As mentioned earlier, initial plans for this dissertation did not include content analysis. However, even during the phase one preliminary interview stage it became clear that post content played a particularly strong part in the interaction patterns of participants. Particularly during directed tasks, participants would note that they formed an impression of an individual primarily based on what they posted about and what they said. Secondarily individuals would look at who their closest contacts were, where they fell in the participants' social network, what their profile said about them, etc. Clearly content was an important part of impression formation, and therefore needed to be accounted for in the IF/IS model. Additionally, even after quantitative analysis began without content analysis, data returned from the qualitative analysis emphasized its importance. As individuals who used Twitter exclusively for building a professional network or espousing political beliefs began to share their experiences, it became clear that accounting for this type of behavior patterning was important. Qualitative data analysis demonstrated that Twitter was being used as far more than simply a "social" platform for connecting with others, but also as a political platform, as business service, a reporting outlet, etc. Thus a model of interaction patterns in Twitter related to identity maintenance would be incomplete if it did not account for these additional use cases, and this in turn would be impossible without content analysis. As such, the Latent Dirichlet Allocation procedure was added to the IF/IS model, which obviously slowed down the analysis substantially.

The two variables examined related to this were content and message hold. Both were based on the LDA procedure described earlier in Chapter Two. Content was calculated by categorizing the most frequent topics for each Twitter user into one of four groups: news/information, social discourse, personal updates or opinions, or other. The ratio of the number of topics falling into each category then defines the extent to which each individual posts about that particular topic. Related to this, message hold was the extent to which users deviated from their most frequent topic. In other words, content and hold can be thought of as the amplitude and the frequency of posting in each topic group, respectively. Those who post frequently about social discourse and rarely about anything else would have a high score in social content and high score in message hold, for example. Conversely, one that posts equally in a number of topic groups would have a low level of message hold.

Message hold was highest among the persona that these measures were more-or-less purpose-built to detect: those who carefully manage their identities and Twitter activities. Examples of such behavior from the qualitative analysis include the political activists and the professional networker. Each of these individuals explained that they are quite careful and strategic in their use of Twitter in order to maximize the capabilities it has to serve their goals. For the politicians, this meant increasing the size of their network so that their message reaches the broadest audience possible. For the networker, this meant carefully curating her activity so that she presented the best possible example of herself and her work. For these and others in persona 3, Twitter is primarily a means to communicate about a single topic or set of related topics, and also to gather opinions about that topic from others. This persona had by far the highest values in the news/information content variable, and therefore also comparatively low scores in the social and personal content variables. Moreover, their message hold was dramatically higher than other personas. These individuals are much more likely to eschew anything unrelated to their primary interest from the account. Compare this with persona five, which had the second highest value in social content (after persona one), but the lowest value in message hold. In other words, though compared to other groups, persona five was more likely to post about their social activities, because members of this group also posted a large amount about a wide variety of topics (particularly in the case of persona five those that fell in the “other” category) their message hold is very low.

Frequency and Co-activity

An early identified variable for measuring engagement with Twitter was post frequency. Once again, however, this proved insufficient on its own to developing a satisfactory model. During the qualitative analysis, participants noted that Twitter was different from other asynchronous forms of Computer-Mediated Communication in that it moved extremely rapidly, and often times posts would only remain on an individuals’ message feed for minutes at most. As such, there was a strong emphasis on the immediacy of the medium, and participants in the high-use categories noted that this played a large role in determining their use patterns: they needed to be quick with their responses, or else they might never see a message again. This in no small part played into their heavy use of liking and re-tweeting, which carry considerably less effort that composing a reply.

However it was among the lower-use groups that this difference became even more apparent. For example, the lowest scoring groups in the frequency variable were personas two and four, both of which maintain relatively low post counts compared to the others. However, analysis of qualitative interviews demonstrated a clear difference between lurkers and likers. For one group of interview participants, the fact that they rarely posted belied frequent use of the site in other ways, particularly in liking and retweeting. Though they rarely posted, they still attempted to remain engaged with the social circles that they were a part of using these forms of indirect communication. Participant 20 said: “I don’t really like to post my own stuff, but I am on Twitter a lot and people know that.” Participants cited a variety of concerns – from privacy and employment issues to the wish to reduce their digital footprint – that made them reticent to make their own posts. Nonetheless, these individuals were clearly active in the identity management process and the IF/IS model needed to account for them, which lead to measures that would such as the differing symmetry and density measures. Frequency and co-activity are another set of such measures meant to separate this use pattern from genuine lurkers. Because they both have relatively similar scores in terms of post frequency, co-activity was designed to differentiate them. Persona four captures this difference. Though members of this group have nearly identical post frequency as persona two, persona four has the lowest co-activity score of any group, while persona two has the second-highest. This indicates that the posts that come from those in persona two tend to be timely in response to the tweets of others. Posts made by persona four, however, have very little correlation to the timing of posts made by others. This distinction was an important one to make in order to separate out those characterized by persona two – who are very much engaged with their social circles – from the sporadic, inconsistent, and infrequent Twitter users that appeared in the data sample. Before the analysis of the qualitative interviews revealed this distinction, these two groups would have been captured in the same cut during the qualitative analysis, which would have hurt the model.

It is worth pointing out for emphasis that no users of persona four were part of the qualitative data collection. As part of the selection screening questionnaire, potential participants were asked their frequency of Twitter use, and those who responded less than “once a week” or “less than once a week” were excluded. In hindsight, this may have been a mistake, as gathering at least some qualitative data on those in persona four may have been useful. At the same time, it is doubtful to what extend such hypothetical data may have been helpful.

Frequency and co-activity also serve as a useful point of distinction between personas one and five, the high use and moderate use personas, respectively. Though they both have post frequencies at the top of the frequency measure, persona five has the second-lowest co-activity measure. This is partly due to how this measure is calculated. Co-activity is defined as the log length of time between any post made by a user in reply to any post made by any other. Naturally, those who are on Twitter nearly constantly did the best in this measure, but those who are on “merely” daily scored worse than the information posters in persona three, and nearly as bad as the lightest users in persona four. Initially, this finding made little sense. Once again the qualitative protocol was adjusted slightly to address this question. Two individuals likely to fall into persona five were asked additional questions about their posting and reply habits. They described a pattern of use whereby they scroll back through their stream history for several days at a time and post replies to each message that they care to all at once. They describe this process happening sometimes once or twice a day at the high end, and sometimes only once a week or at the low. As such, though they eventually get around to responding to just about everyone that they care to, often times this process is delayed by one or more days. Compared to the interaction patterns of other groups, this is unusual. Those in persona three, for instance, would rarely scroll back more than several hours during their normal interaction sessions. Thus even though their social circles were much less close than those in persona five, their co-activity scores were much higher. Thus the qualitative findings related to low-frequency users that were still heavily engaged with the site led to the creation of additional quantitative measures that could differentiate these individuals from those not so engage, which in turn lead to the development of qualitative data that further helped describe the distinction between the always-on and the daily-on interaction patterns.

Linking

Linking behavior was included in the IF/IS model based on existing literature that demonstrates a strong correlation between information sharing practices and use personas in Twitter (e.g., Boase & Naaman, 2010). During the qualitative interviews, no explicit instructions were given with regard to including links in posts, and no participants were observed doing so. In general, the inclusion of links in tweets observed in the quantitative data was low. This may be explained by the fact that the data set used was deliberately created to capture high engagement Twitter users, whose volume of tweets over the relatively short collection period would be sufficiently numerous to draw meaningful statistical conclusions about. This set was also selected to match at least vaguely the selection criteria for the qualitative portion of the study.

Linking was most readily apparent in the informational tweeters of persona three, whose members had double the frequency of links in their tweets as the next lowest group, persona five. Again, this result is a bit surprising given that individuals in persona five make far fewer tweets overall than persona one. Again returning to the qualitative data, participants described their posting process as happening at most once or twice a day. At that time, they report that they may make one or two posts, related to whatever they feel like is the most interesting thing happening at that time. As such, there appears to be something of a time-delay filter to their posting habits, such that they mainly post about whatever is the most important or most popular news item, celebrity gossip, etc. The thing that catches their eye, that they decide to post about, therefore, is much more likely to include a link. Those in persona one, on the other hand, tend to post far more reflexively, and in reaction to a smaller temporal collection of events.

Information Seeking

Findings from this dissertation indicate that individuals utilized a variety of information seeking strategies in undertaking the directed tasks. During the directed information seeking tasks of phase two, participants were asked to identify a target individual to learn more about, then pursue that task by whatever means they chose (see Appendix B for phase two interview protocol). Actions taken were then recorded, coded, and analyzed for interaction patterns in their behavior in line with the Ramirez model of social information seeking (Ramirez, et al., 2002). Using the mixed-method approach described in Chapter Two, both qualitative and quantitative data were used to categorize participants based upon their general interaction behavior with Twitter, and then learn how these behavior patterns influenced their information seeking approaches in the task. The results of this are presented in Table 10 below, and each persona is further described following the table.

Table 10

Persona details

Persona	Description	Key variables	Information seeking	Qualitative participants (phase 2; n=15)
One	Always-on, heavy Twitter users. Members of dense, very active social circles	High direction, High social content, High frequency, High co-activity	Interactive	4
Two	Heavy browsers, infrequent posters, rely heavily on indirect communication	Low intention, Low symmetry, High density, High co-activity	Passive	2
Three	Use Twitter to engage with information and personalities, engage with interest groups, collect and share content	Low direction, High intention, High information content, High linking, High message hold	Active	3
Four	Light Twitter users. Sporadically use the service to check on and discuss major news or life events	Low direction, High intention, Low co-activity, Low message hold	Unknown	0
Five	Moderate Twitter users. Socially engage with their circles, occasionally follow a couple celebrities. Interact in "batch mode."	High symmetry, High social content, High frequency, Low co-activity, Low message hold	Extractive	6

Persona One

Individuals grouped in persona one were the most frequent users of Twitter. They are characterized primarily by a high degree of directed communication as well as a high degree of co-activity. For them, Twitter is an "always-on" application and their use patterns reflect this. When cued to perform the directed search tasks, individuals in this persona tended to take a circuitous route. They were often set off-track by items or individuals that caught their eye, and would interrupt the task to interact with others. For them, this provided the most successful route to learning more about others on the site. It was through interaction and co-activity that they maintained a keen awareness of what was occurring in their social circles and the world. Rather than directly searching for information about an individual, they tended to browse their stream in general in order to discover when and how they appeared in it. Contextualizing the target individual in this manner provided them with important clues as to their relationship to the participant's own social world. When asked to select a target, one even had some difficulty finding someone that they were not at least partially aware of. Through their constant connection to the service, these individuals have deep perspectives about how people in their circles fit together. When learning about a new individual, they tended to use this perspective in order to form an impression of the individual. Rather than read through a long archive of posts, or even visit the target's profile page, they often visited the pages of others in order to see how the target and others in the social circle interacted.

Thus the primary means of social information seeking for those in persona one could be described as an interactive technique. Rather than judging the target individual based on their posts or history, they tended to use contextual cues provided by others. Often, this took the form of at-mentions and hashtags. When an individual was mentioned frequently with others, this was a relationship cue that was relied upon heavily by persona one. Additionally they may look to liking behavior; frequently like-links between individuals was another factor that persona one used to form impressions. Unlike all other groups, those in persona one were likely to open a tweet in order to see whom else liked it, and use this as a rationale for locating them in the social circle, and in turn forming impressions of them. This may be reflected in the relatively high directionality, density, and co-activity scores that characterize persona one. Their network is both dense and close, and they interact with it frequently. This provides them with a near-constant set of cues by which they learn to make impression about others in their social circles.

Persona Two

Persona two is oftentimes as frequent a visitor to Twitter as persona one, but is far less likely to make comments or posts. Though they oftentimes consume nearly as much information, they do so primarily without providing much original content, instead relying upon indirect forms of communication such as likes and retweets in order to demonstrate their connections with others. These individuals rely upon Twitter as a news source in order to keep up-to-date about what is happening in the worlds around them, both socially and broader. They see Twitter as a resource for keeping up and staying current, and less as a messaging platform. They prefer to stay relatively anonymous, and prefer not to be the center of discussions. When they make posts, they tend to be limited to major life events or major news items, rather than the daily catalogue of their lives that persona one tends to provide, or the daily catalogue of news events, that persona three like to provide. When performing the directed information seeking tasks, members of this persona tended browse their social streams in a similar way as persona one. However, rather than bounce from person to person while gathering context about a target individual's conversations and interactions, persona two was more likely to visit the individual's profile page directly and simply read more about them. Because persona two tends to have a much more asymmetric network topology than persona one – having far more follows than followers – these participants received fewer messages and had far fewer at-mentions directed at them. As a result, the likelihood of their co-inclusion with the target individual was far lower. They therefore had little, if any, messaging context in which they themselves were included in order to coordinate the location of the target individual in one of their various social circles. Furthermore, because their overall networks tended to be smaller, they were less likely to simply run across the target during directed searches of conversations taking place in their social circle.

As a result of having fewer such contextual cues, individuals in persona two were less likely to be able to make interactive searches of the feeds of themselves and their follows and find useful information. They therefore tended to take a more passive approach, learning about new individuals as it was presented to them. While persona one was likely to have history of communication with others surrounding the target individual from which they could draw, persona two did not have as ample an opportunity. They reported that when following people for the first time, or engaging with new acquaintances on Twitter, they would simply wait for that individual to start posting, rather than engaging directly with them and those around them. While persona one actively sought to set up a strong network of interactive partners, persona two was more interested in developing a meaningful, useful stream of data that would keep them informed while not demanding an excessive amount of communicative effort. Characterized by a low audience value but a relative high density value, persona two is connected to relatively tight social circles, but only loosely so. They tend to have less direct interaction with individuals in these circles, instead relying primarily upon indirect communication as denoted by their low intentionality value. Rather than being able to interactively engage with their target individuals – cueing them with questions or at-mentioning them in posts, for instance – they therefore tend to rely upon passive observation of their activities.

Persona Three

Persona three individuals use Twitter less as a social platform than as an information platform. They use the service to gather, collect, generate, and share information about a select topic or group of topics that they care about. This can take a variety of forms. In the qualitative sample there were individuals that used Twitter as an information source for politics and business. However this persona equally applies to the many other category interests that permeate Twitter, including sports, entertainment, food, celebrities, and so forth. Increasingly Twitter is becoming an important resource for interest groups to gather and spread information, and particularly so as an increasing number of athletes, entertainers, politicians, authors, scientists, etc. have begun using Twitter. Whole communities have sprung up around these interest groups, and for many within them Twitter is their primary source of information regarding their interest of choice. Persona three are members of one or more of these communities, engaged in discussion with other fans and aficionados like them.

Persona three individuals are the most deliberate with both their identity management and their information seeking strategies. Note, however, that this is not necessarily to mean thorough. In anything, compared to personas one and two those individuals within this group tended to rely more upon appeals from expertise, rather than gathering their own information. They reported that they had sources for information that they trusted a great deal, and would generally rely upon this information implicitly. During the directed information seeking tasks, persona three participants were the most likely to seek out individuals directly connected to the target individual, rather than intermediaries within their own social circles. When deciding upon an individual to learn more about, for instance, all three persona three participants in the qualitative, talk-aloud phases first went to the page of one of their trusted “advisors” and scanned that individual’s feed for mentions or links to others that they might also learn from. In this case of the politicians, this meant going to the feed of a trusted pundit or blogger, in the case of the professional networker this meant going to the feed of a thought leader in her industry. Regardless, persona three individuals were most likely to rely upon this type of recommendation when learning about and judging others.

This strategy falls best within the active form of social information seeking. In both selecting their targets and forming opinions about them, persona three relied heavily upon the opinion of others. They seek out information from others and use this information to gather sources. Of all the groups, persona three was most likely to follow someone during the directed tasks simply on the recommendation from someone else that the target “was a good follow.” To some extent this may create something of an echo-chamber, as other researchers have noted. During the qualitative interviews, the social streams of persona three individuals were frequently noted to have duplicate links and long chains of re-tweets. As the persona type most engaged with some of the most active and most popular areas of Twitter, this makes some sense. Characterized by relatively low density and directionality values, persona three was unlikely to be engaged in direct communication with individuals. Rather, their posts and replies took place in more of an open, community atmosphere. They were more likely to include hashtags than any other persona, indicating their tweets were intended for an audience of many rather than an audience of a few. Their messaging content was also much less likely to be personal or social than other personas, and their message hold was much higher. They use Twitter primarily to engage with the information and entertainment they care about, and others who care about similar things.

Persona Four

As mentioned previously, persona four represents the lowest-activity members of the Twitter data set, and as such individuals in this category were selectively screened from the qualitative portion of the study. However, ample quantitative data was collected about them. Compared to other groups, persona four has the highest intentionality, indicating a large number of posts and replies compared to likes and retweets. This may be an artifact of the relatively sparse amount of data collected about them during the sampling period or may indicate a trend; without further qualitative investigation it is difficult to make assertions. They were also characterized by a relatively low directionality, indicating infrequent direct communication with others. Their co-activity was also quite low, though again this likely due to the same reasons that persona five had relatively low co-activity as previously discussed: the fact that check their streams infrequently means that they are often replying to tweets made days earlier.

Persona Five

Persona five individuals use Twitter in some ways similar to all those mentioned above. They regard it as a social platform for messaging those in their social circle, an information platform for learning about their own world and the world around them, and as a way to interact with others. Of the interview participants, they represented the largest proportion of participants, and were also the largest group in the quantitative analysis. To some degree, this can be seen as an artifact of the cut method used. Because the method chosen seeks to maximize between-group disparity rather than maximizing in-group similarity, oftentimes this can result in several groups that are quite different than one or two larger, more homogenous ones. Nonetheless, they represent a use pattern different from others in the qualitative study. Though they were in some ways similar to the individuals who are “always-on” in persona one, because they did not interact with the platform as frequently, their values in many of the variables were different, and the qualitative observations supported this.

Persona five individuals were the most likely of the group to read through previous threads and Twitter feeds, sometimes going back days or even weeks when completing the directed information seeking task. As discussed earlier, because they want to be caught up with everything that happens in their social circles on the site, but because they only interact with it sporadically, they are more likely to scan through past messages and comments, and respond to them in a time-delayed manner. They were the most likely group to use previous posting history as a cue to make inferences about a target individual, and several respondents noted that this is typically how they use Twitter: by reading back through a long history of posts and replying to one or more that catches their eye, rather than attempting to keep up with every post or message in real time. This approach most close lines up with the extractive social information seeking strategy, in that it is based on the use of past history of interactions made by the target individual.

Conclusion

This chapter presents in detail the composition of the IF/IS model that is this dissertation’s primary research outcome. It discussed the development of this model via mixed-methods consisting of both qualitative and quantitative research, and the iterative process by which this development took place. It discussed the variables and measures included in this model, and how each was expressed in the data. Finally, it presents five “persona” that represent categories of behavior present in both the qualitative and quantitative data. In the next chapter, the final issues related to the project are discussed, including the research’s contribution to the field and future directions for subsequent projects.

Chapter Six

Conclusion

This chapter includes a final discussion of the outcomes of this dissertation. A final review of the research questions and the major findings and resolutions that were determined for each of them is presented. Next the discussion moves to limitations to this research project, and concludes with a discussion of future directions in research projects moving forward from this one. Finally, methodological, disciplinary, and theoretical implications are discussed in order to situate this project in a broader dialogue with other research.

Summary

This study proposed a number of research questions related to the phenomena of information seeking, identity maintenance, and indirect communication in social media. A literature review revealed potential links between information seeking and identity maintenance. This project proposed that as individuals learn the mores and norms of a particular online culture or subculture, they tend to adjust their self-presentation in line with those norms, and that this adjustment process in turn modifies their information seeking behavior and other interaction patterns in a recursive manner. As such, it was proposed that there would be some link between interaction patterns in general, and information seeking patterns in particular, and the modes or strategies by which individuals present themselves online. Further, this project suggested that modeling this behavior could provide a means by which these information seeking strategies could be better studied and understood, as this understanding is lacking in the extant literature. This chapter provides the results of these efforts.

Research question one.

All participants in the study recognized in some degree that they present themselves on Twitter using a variety of strategies, and that these strategies can be used to provide insight into their owner's personality. In the course of this research, using both qualitative and quantitative methods this study attempted to create the beginning of a framework for understanding the interaction patterns present in Twitter users, particularly as this interaction relates to identity maintenance. This model – termed Identity Formation / Information Seeking or IF/IS – includes a number of variables that can be tested on Twitter data which were specifically chosen due to their application and use in identity maintenance, supported by data gathered during interviews with heavy Twitter users that manage their online identities on a daily basis.

First, findings indicate that individuals manage their audience in various ways. Some attempt to increase it maximally, in order to reach as broad an audience for their posts as possible. Others deliberately keep it small and restricted to a select group of individuals in order to maintain a sense of community. The IF/IS model therefore includes measures of the individual's audience, including its size, density, and direction. The results also demonstrate the importance of message content and variety. Respondents were quick to point out that post content played a large part in how they managed their identities and formed impressions of others. This not only includes the topics involved, but also the language or diction used (e.g. "Twitter-speak", inclusion of links and other information, and the consistency of topicality). For this reason the IF/IS model also includes several measures of post content, including topic nature and topic frequency. The research findings also indicate that interaction type and frequency is an important means by which participants manage their identities. Some of the heaviest users made it a point to reply immediately to the posts made by others, and expected the same promptness in return. Others made it a point to check the site only a few times a week, specifically to avoid that sort of expectation. Messaging frequency and co-activity were therefore also included in the IF/IS model. Finally, a special point of this research was to include indirect communication into the model, as a literature review identified this as a potential gap in the literature (see Afifi & Weiner, 2004; Chi, 2008; Ramirez, 2009). Research underscored this need, as nearly all respondents noted that liking and retweeting were important parts of their interaction behavior patterns. Therefore measures of this behavior are also included in the IF/IS model, such as the ratio of direct to indirect communication.

Research question two.

Respondents varied widely in their approach to this question, ranging from having no clear picture or strategy for how they wanted to present themselves, to having a very clear set of rules managing their own interaction behaviors. Indeed, to a certain extent even having no strategy at all might be considered a strategy in itself. This study applied the IF/IS model derived above to a large set of Twitter data in order to discover patterns in interaction behavior. This was done by operationalizing each of the variables for use on the raw data, then converting the data to scores for each measure included in each variable. As noted above, several of these variables had several measures associated with them. Once these values were determined, each user in the data set then had a vector of numerical scores for each category. These vectors were then grouped using a k-means analysis, until a stable value for *k* was found (Scott, 2017). This was then confirmed via the PCA approach discussed in Chapter Two.

This resulted in a group of five clusters, each of which was linked to a general pattern of interaction behavior. This numerical system was validated by, and built up alongside, a qualitative coding system based upon the interview data. The qualitative and quantitative findings were synthesized into a final group of “personas” that represented the identity management and information seeking behavior patterns of frequent Twitter users. As shown in Table 2 in Chapter Two, persona one was characterized by very low intentionality and high co-activity, as these users are constantly replying to and re-tweeting one another. They demonstrate a high degree of activity that seems in line with maintaining social relationships. Persona two is characterized by a low in-out degree ratio, as these users tend to follow lots of others but post infrequently, which limits the amount of social engagement they exhibit. They tend to rely upon indirect communication in favor of direct, and limit the amount of text and image content that they post. Persona three is characterized by a high frequency of posting on a related group of topics. These users seem to use Twitter primarily as a means to communicate about news, politics, entertainment, etc. Persona four is characterized by low intention, co-activity, and frequency. These users are on Twitter sporadically at best. It’s something they might check every once in a while, but not a main part of their media mix. Persona five is characterized primarily by their social post content and dense social network. These individuals primarily use Twitter as a social tool for connecting with friends and colleagues, and less as an information source. Similar in some respects to persona one, though their less frequent use creates differences in their activity patterns.

Research question three.

Information seeking was one of the primary points of study for this dissertation, particularly in the qualitative portion of the research. Individuals were provided with tasks to seek out social information about others in their social circle, and the strategies they employed were observed, and then categorized using the Ramirez (2002) model of the active, interactive, passive, and extractive modes. Data analysis discovered that individuals used these modes in varying degrees and contexts based on a number of factors, including the nature of the relationship they had with the target individual, the degree of existing familiarity they had with the target individual, the available interaction history of that individual, and other factors.

All participants pointed out that they used many of these strategies at various times and for various purposes, again depending upon some of these same factors. Indeed, during the directed information seeking task portion of the research, individuals were observed doing precisely this. It was common for participants to begin a search in a passive mode, merely scanning their awareness streams for an individual or post that interested them. Next, they may transition to a more active mode, directly reading the individual’s profile and posts. Next, they might move to an interactive mode, looking for interactions that took place between their friends and the target individual, in order to learn more about the context of how the individual fit within their social circles. Finally, they might use an extractive mode by reading through the post history of the target individual in order to gauge what they posted about most. This order itself was quite fluid, and individuals were observed moving between these modes quite quickly and easily. Thus there is clearly no specifically set strategy that works best in any particular situation (Westerman, et al., 2009).

Those caveats aside, findings did record some links between patterns of interaction behavior on Twitter and patterns of information seeking strategy. Individuals that were the heaviest, most active communicators on Twitter tended to used interactive strategies the most. They would commonly look for cues that indicated how the target individual fit in with their existing social circles to make judgments about them. These participants were part of persona one. Persona two participants tended to use Twitter almost as much, but mainly relied upon indirect forms of communication to maintain their presence, rather than making text comments or replies. These individuals tended to have weaker ties to their networks of social circles, and thus had fewer interactive cues available to them than those in persona one. As such, they tended to rely upon the passive strategy more often, particularly because it fit with in their natural mode of interaction with the site. Other participants used Twitter mores as an information resource and interest group community than did other groups. For these members of persona three, sharing ideas and information about a particular hobby, culture, or topic was a primary reason for using Twitter. These individuals relied upon the active mode of information seeking more than others did, and often would rely upon the opinions of thought leaders in their community when determining who was important, trustworthy, or “a good follow.” Finally, those in persona five used Twitter regularly, but much more sporadically compared to personas one and two. They might typically sign on a few times a week, and answer make all their posts and likes in those sessions. During that time, they typically will read back through several days of posts to see what they might like to reply to. Using this same strategy in the information seeking task, these individuals were most likely to look back through the post history of the target individual in order to learn more about them, applying the extractive strategy.

Research question four.

Respondents varied widely in their approach to the use of indirect communication and how it fit into the creation of their online identities. All respondents utilized indirect communication to a certain extent, but the way it fit into their interaction patterns depended quite a bit upon the ideas of identity that were important to them. For instance, Those in persona two reported feeling a strong desire to monitor and observe Twitter behavior, but felt reticent for a variety of reasons to post content themselves and become actively involved. These reasons varied within the group, ranging from concerns over privacy and advertising policies to issues regarding potential monitoring by their employers to simple lack of the time, energy, or desire to maintain an active social media presence. In fact most respondents in persona two noted this final factor in particular as one that made the use of indirect communication a useful and important one to them.

Other respondents used indirect communication as a form of aspirational signaling behavior. For instance, those in persona three often took part in discussions with or about their favorite celebrities, politicians, athletes, etc. A common practice among those in this group is to monitor, gather, collect, share, and discuss the posts and content of high-profile individuals within their interest group, from sports stars, to opinion columnists, to artists. Liking and especially retweeting behavior was particularly prominent in the interaction behavior patterns of this group, as well, as it provided a mechanism to directly connect with and respond to those high-profile individuals that created the content that was shared in the interest group directly. It also provided a means to signal to others both in their interest groups and outside it their membership in it, as a means of providing bona fides, perhaps. By retweeting their favorite musician’s latest post, they signal to others the importance of this musician in their lives. By liking the latest post by a Congressman, they signal to others that have also liked that post that they are also a supporter of her policies. Thus indirect communication provides a mechanism for flagging in-group membership that carries a surprising amount of weight given its relatively light effort load. Because it leverages the cachet or reputation of an existing high-profile individual on Twitter, the individual is able to apply some of that reputation for themselves.

Research question five.

This dissertation examined indirect communication primarily in two ways. First, as it pertains to the link between information seeking and identity formation. Second, as it pertains to identity maintenance among users of SNS as a means of conveying information. In the directed information seeking task, participants were asked to use whatever strategies they preferred in order to learn about an individual either in or related to their social circle. While most of the participants noted the use of information seeking to some degree, it was particularly important to those individuals that use indirect communication the most themselves. To these individuals, indirect communication provided a means of determining how individuals fit within the context of their social relationships. Respondents noted that determining how individuals fit within these relationships on the basis of following behavior alone is a difficult task, due to the sheer volume and complexity of these following connections. However active behavioral cues similar to liking and retweeting behavior can make manifest those connections in far stronger and more immediately salient ways.

All the individuals studied used indirect communication to a certain extent. However there were marked differences between them in how, why, and when it was used. Particularly for the heaviest Twitter users, indirect communication was an important part of their communication patterns. Particularly for persona two, indirect communication was the primary means by which they interacted with their social circles. While others among the heaviest Twitter users – primarily persona one – did use indirect communication as well, it made up a far smaller proportion of their overall communication pattern than it did for persona two. For them, indirect communication provided a mechanism by which they could maintain a sense of connection and activity with a social circle while still maintaining a sense of relative anonymity or privacy. Furthermore, it allowed them a quick and easy way to maintain a temporally immediate social presence with minimal effort. All participants among the heaviest Twitter users noted that the speed with which messages are passed on Twitter make response delays especially noticeable. Those users that wished to maintain a high level of engagement with their peers at this speed all pointed out that leveraging indirect communication was a useful way to maintain this sense of connection without feeling overly burdened by the norms and expectations of immediacy that are hallmarks of the service’s cultural expectations.

Limitations

As in all research, this project has several limitations with regard to its procedure and analysis. The first and perhaps most serious of these has to do with sampling, for both the qualitative and quantitative phases of the study. The quantitative portion used a sample based upon a 2013 study conducted by the Stanford Network Analysis Project (SNAP, http://snap.stanford.edu/). This data set lists 81,306 Twitter users and their follower network as of 2012. This data was chosen primarily due to its quality. It has been vetted by SNAP scientists to ensure that it contains only data related to actual Twitter users, and not fake accounts, bots, or other sources of potential noise that would impact the results of this study. However the SNAP data is dated, although it does provide a snapshot from 2012. Many of the users that appear in the data are no longer active on Twitter, and in order to update the data the original users mentioned in the data were crawled again to obtain the users that the original SNAP users follow. From this list, the data was expanded to include only those users that followed the original SNAP users in return, on the assumption that these are likely to be actual social connections and not the accounts for brands, celebrities, news outlets, and the like. This expanded list of some 3.5 million accounts was the initial starting point. From this list inactive or redundant accounts that appeared were removed, which reduced the total number of accounts used in this study to 247,318. From these accounts, the friend network was crawled again in order to calculate network measures, and finally the most recent tweets for each were collected. This data chosen for a number of reasons. First, initial analysis indicated that those in this data set were more frequent posters than average, which would line them up with the qualitative sample. Second the longevity of the accounts in question would limit the effects of account creation recency bias as discussed in Chapter Four regarding results from the quantitative analysis. Finally, it would also limit the potential for anomalous or infrequent posting patterns, which could skew the results particularly since the sample period was a relatively short period. Nonetheless, this sample is biased due to its collection procedure. For the reasons above, this bias may in fact be a net positive given the goals of the research project. Nonetheless, these issues remain.

The quantitative sample was also intentionally biased toward heavy users of Twitter, as discussed in Chapter Two that described the qualitative method. The sampling method used to recruit participants involved primarily two methods: soliciting at three Northeastern US universities for participants, and a generative snowball sample from these participants and from personal contacts of the researcher. The demographic and socioeconomic skew of this data is difficult to overstate as a result of this collection method. Though the initial screening survey asked only very basic details, the participants were all dramatically younger and had a higher socio-economic status than the US population as a whole. Though the Twitter population also skews younger and more affluent, the interview sample passed even this threshold. Due to the inclusion of an urban university in the solicitation sample, ethnic diversity was fairly in line with US averages. However, continuing the above reasoning, it may not necessarily be a fault that demographics of the respondents skew in the direction of heavy Twitter users; indeed, this was largely by design. Nonetheless, the sampling method was not random because it was based on the existing SNAP data. This necessarily hurts the generalizability of the findings presented here. Further, there was no form of 3rd-party verification of any of the analysis performed. Ideally, several clustering algorithms would be applied and their results compared for differences. Due to the fairly involved nature of the calculations, however, this was not time-feasible.

The second and most glaring limitations come from the application of the quantitative data analysis. These issues are handled more fully in Chapter Four where the specific technical rationale for each decision made, as well as its relative pros and cons, is handled more fully. However to summarize, LDA as a topic modeling technique was never initially designed to handle short-form text corpora. Though it has been modified extensively and proven robust in both this and many other studies that apply it to Twitter data (see Rehurek & Sojka, 2010, Zhao, et al. 2011, Hong & Davison, 2010), several core problems remain with its validity as a topic modelling approach to Twitter data, or even if topic model as an approach at all makes sense with respect to Twitter. Nonetheless, LDA has been applied successfully to corpora of Twitter data for many years. The other major issue is the k-cut clustering algorithm and its selection. There are a wide number of clustering algorithms, each of which attempt to group data in various ways. Some attempt to make the groups as even as possible, some as different as possible from one another externally, some as similar as possible to themselves internally, etc. This project selected an algorithm that maximizes external differences in order to dovetail into the qualitative portion of the procedure that also sought to detect and describe differences between information seeking patterns. Other algorithms could – and almost certainly would – have yielded different results. Again the nature of these considerations is described in greater detail above.

The final limitation was, as it always seems to be, sheer computation time. Especially with the introduction of the content analysis, quantitative calculations began to reach levels of complexity that quickly outstripped the computation power available to this project. As such, a data reduction technique was applied in order to identify and remove redundant variables and limit the scale of the data the clustering algorithm needed to work with. In the end, this helped simplify and clarify the IF/IS model, so it may have been useful, but it is irrefutable that some information was lost when some of the data due to computational necessity was ignored.

Future Directions

One of the main purposed of this research was to create a model of identity management and information seeking in social media that could be used to further investigate the link between these phenomena. Using qualitative and quantitative techniques, this model, termed IF/IS, has been validated in the course of this research. An accompanying goal was also to be able to create a model that would be usable for other research projects.

On the qualitative end, future efforts might investigate more fully the link between information seeking and identity management. This study and others (Ellison, Heino, & Gibbs, 2006; Westerman, et al., 2008) have begun to explore this link, but without a solid framework for discussion, it has largely been relegated to asides and footnotes. Future studies might examine precisely what types of strategies individuals employ, especially with regard to their imagined audiences.

On the quantitative side, future research could further examine indirect communication and its use in social media. How specifically it is used in what contexts, by whom, and for what purpose. Again the goal of this project was to provide a framework for this analysis. With it in place, a study of this practice in greater depth could take place. Particularly as SNS platforms are shifting towards formats devoid of any text at all (Instagram, SnapChat, etc.) these practices need to be better understood.

Implications

This study was an exploratory and descriptive one, aimed at developing a framework for analysis via which future predictions might be made and hypotheses tested. Based on research questions rather than hypothesis testing, this study was intended to use existing data in order to demonstrate a method and procedure via which future studies might approach similar problems. It provides the IF/IS model for prediction, analysis, and interpretation that may be applied to hypothesis testing, and perhaps more importantly demonstrates a method by which such interdisciplinary, mixed-method models may be developed. When considering the impact of this study, it has three primary areas of effect: methodological, disciplinary, and theoretical.

The methodological implications begin with the choice of use of a mixed-method design. Though far from novel, mixed-method approaches are capable of opening up research fields in exciting new ways (Creswell & Plano Clark, 2011). Considering especially the area of SNS research, an important example is Gilbert and Karahalios’s (2009) work on Facebook, which demonstrates that though it is possible to mathematically model tie strength with social trace data, it is impossible to understand these ties fully without conducting interviews. The current study continues this line of inquiry, moving it from an analysis of relationships to the analysis of individuals.

Since the early days of research into SNS, many previous studies have examined self-presentation in online profiles using qualitative data (Cross & Madsen, 1997; Donath, 1998, Jacobson, 1999) and many have examined it using quantitative data

(Gosling & Gaddis, 2007; Lewis, Kaufman, & Christakis, 2008) there are relatively few examples that combine the two approaches. This represents a gap in the discipline's understanding. Snelson's (2016) survey of social media research identified that only 55 of the 299 projects analyzed utilized mixed methods. She further observed that of the methods mixed in even this small sub-category of 55, the use of quantitative analysis of social networks and SNS interaction was rare. She observes:

> The majority of the qualitative and mixed methods social media studies were conducted with established methods such as interviews, surveys, focus groups, or content analysis. […] Emergent social media research designs such as those that couple network analysis with qualitative analysis were present but uncommon in the literature sampled for this review. – Snelson (2016, p. 12)

While both qualitative and quantitative approaches have major advantages, too often they are relegated to separate panels, separate journals. While the reasons behind this lack of integration between qualitative and "big data" approaches – both of which are touchstones of the SNS research community's method – are unclear, the current study attests to at least one: it is hard. As this dissertation demonstrates, framing the IF/IS model in such a way that it makes sense in both quantitative and qualitative contexts is not an easy task. The amount of behind-the-scenes cajoling of the data that took place in this study was substantial, and time and again the model needed to be revised, variables added and removed, operationalizations adjusted, qualitative findings revisited, additional interviews conducted, and on and on.

However, the juice is worth the squeeze: on several fronts, this dissertation moves the methodological needle forward. First it demonstrates the effectiveness of procedures that can be used to generate insight on users: the use of the talk-aloud protocol to learn about interaction behaviors, and the use of personas to codify them. Importantly, it is worth noting that this procedure is not novel: the use of the talk-aloud protocol and the use of personas have for many years been a staple of the user interface / user experience (UI/UX) design community (Garrett, 2010). However, these approaches are infrequently seen outside the professional research domains in which they are most often applied. Second, it demonstrates the interplay between quantitative and qualitative approaches, and how data and analysis gathered in one domain can inform and improve approaches in the other. Finally, it attempts to systematically codify a specific type of interaction behavior into a general taxonomy or framework about which predictions can be made. This important first step to the null hypothesis significance testing approach is too often missing from many contemporary studies of SNS, particularly in the quantitative realm. Computational Social Scientists, awash with more data than they could ever hope to analyze in their lifetimes, too often begin with results and move backwards toward a framework for analysis.

The domain-based contributions of this study stem from its contributions to current understanding of identity management and impression formation. Many early attempts at codifying identity formation online focused on self-presentation in its most literal form, via the signaling of status (Donath, 2007), depiction of personal tastes (Liu, 2007) or the content of profiles (Counts & Stetcher, 2009). Increasingly, the field has come to appreciate some of the foundational theories of identity upon which this dissertation rests: theories like those of Goffman (1959) that hold that identity is a symbolically constructed metanarrative that is asserted, reinforced, and ossified in the dynamic of the communicative act. That is, interaction behavior is what describes and determines notions of identity, more so than any internal homunculus.

This dissertation rests on this theoretical insight, and holds that any attempt to understand identity without the close examination of how it is established, maintained, and interpreted in the contextualized act of interaction is incomplete. As such, the IF/IS model presented here deliberately eschews typical lay markers of identity in its conception – gender, nationality, religion, occupation, socio-economic status, etc. – anything that might be included on a profile page. Rather, it is built entirely upon data that stems from interaction. This is a deliberate and important point. Researchers in CMC have long noted its power to allow for, cope with, and empower alternative expressions of individuality, or even multiple selves in ways that the "real world" never can (Yee & Bailenson, 2007). While many existing quantitative models of identity presentation and management online focus on these intransigents, this dissertation is informed by scholars that perform primarily qualitative research, many of whom (Donath & Boyd, 2006; Turkle, 2011) underline the importance of transience of identity online. Echoing Goffman, they note that identity is something that can be adopted, performed, and then discarded quickly and easily. They therefore note that to focus on the intransigent markers of identity – particularly in online settings – is to largely miss the point. Following this logic, the IF/IS model presented here presents a notion of identity based upon the ripples it leaves in the quotidian pond of everyday interaction, rather than scaffolding it around some stony artifice of self-memorialization like a profile page or geo-location tag.

The theoretical insights developed over decades of qualitative research have a great deal to teach the “quants” of the world. It is the fervent purpose of this document to demonstrate the value of these insights. The revolution in data that lead to the creation of the field of Computational Social Science has empowered researchers to answer questions of “who” and “what” and “when” with accuracy and precision long thought impossible. Twenty years ago, the *very best* theories of sociology and psychology could explain around one-third of the variance in behavioral data. Most did far worse. Today, the very best theories can intuit with near certainty when a woman becomes pregnant before her parents or partner are even told. “Who” and “what” and “when” are increasingly solved problems in behavioral research.

The questions of “why” and “how,” on the other hand, remain largely open. It was the pursuit of these question that inspired this dissertation. Why do individuals tailor their identities in various ways online? How does the information seeking process alter their impressions of others, and how does this alteration effect their own ideas of identity? The sophistication of understanding in Computational Social Science has lagged dramatically behind in answering questions of “how” and “why” when compared to “who” and “when.” The purpose of this dissertation was to attempt to close this gap.

The theoretical contributions of this study are related to both the process of information seeking and identity management. Of the many models of information seeking, Joseph, Debowski, and Goldshmidt (2013) note that the models of Ellis (1989), Meho and Tibbo (2003) and Marchionini (1995) most closely reflect the search behaviors of users of electronic systems. These models are all primarily based on action and interaction, and when combined provide a useful framework for the analysis of information seeking behavior in online environments. Each built on the last, providing additional moments of contact between the model and its application. This dissertation seeks to extend the thrust of this work by providing additional places of contact between these existing models of online information seeking and the new contexts and domains created via SNS.

Social information seeking is a new and burgeoning field in Information Science (IS), providing new insights into how individuals leverage social information to arrive at decisions (Chi, 2008). Complicit in this discussion – but until recently largely ignored in the IS literature (Antheunis, Valkenburg, & Peter, 2010) – is the manner in which information seeking is used to establish relationships in the first place, such that they then may be leveraged to accomplish other information seeking tasks. This dissertation has relied upon the theories of uncertainty reduction (Hogg, 2000) to describe this process, and highlight the parallels between it and information seeking writ large. It is an undergirding supposition of this study that these processes are linked in this manner, and the demonstration of this linkage was a central goal of the study.

In other words, this dissertation argues that information seeking is a vital part of the process of impression formation (conceived here via the lens of uncertainty reduction), itself a vital part of identity formation. That is, this study argues that the process of establishing identity in the self is inexorably linked to the process of evaluating identity in others, and this process of evaluation is inherently an information seeking task. As Goffman (1967, p. 227) observes, “If a person is to employ his repertoire […], obviously he must first become aware of the interpretations that others have placed upon his acts and the interpretations that he ought perhaps to place upon theirs.” This dissertation seeks to extend these crucial insights into the quantitative, digital realm. By evaluating the information seeking process of impression formation, this study attempts to model identity management practices in social media.

Taken together, these links in their various forms comprise what this study intends to be its main contribution. This dissertation established a framework for analysis drawing upon theories from numerous disciplines, including Communication, Information Science, Sociology, and others. It built the IS/IF model for evaluation useful to and generated by both quantitative and qualitative methods and demonstrated the intrinsic value in linking them. It utilized cutting-edge techniques in computation to explore and extend some of the oldest and most important lines of inquiry in social science – lines often ignored by computational techniques. Finally, this study demonstrated that by employing mixed methods and multi-disciplinarily, exciting new avenues of research are open.

Social media provide innumerable new avenues for individuals to share and connect with one another, but their habits vary widely. Rapid changes in these technologies have created an environment where new forms of both the consumption and production of communication are possible, and new modes of interaction with a community are possible. This study has argued that such changes include increased use of indirect communication for the maintenance of social relationships, and an increased reliance upon passive forms of social monitoring that involve keeping a social stream on “in the background” in order to provide an inchoate sense of co-presence at all times, across wide physical, spatial, and even socioeconomic distances. To date, research into these new forms of communication and new modes of information seeking has been scant. This study breaks new ground and informs the development of innovative models of identity management and self-presentation in social science that more readily account for these changes, and others that may appear in the future.

Appendix A

Phase One Interview Instrument

Introduction

I'm collecting information for a project on self-expression in social media. I will be sharing the information that I gather here with the other member of my research team and with the Rutgers University Institutional Review Board, but your name will not be associated with any other information.

First I'm going to ask you some questions about your use of social media.

Demographic questions

1. How frequently do you visit Twitter? (circle one)

   a) Never used it
   b) Once or twice a year
   c) A few times a month or less
   d) Weekly
   e) Daily
   f) More than once a day (but less than 5)
   g) Five to 10 times a day
   h) Too many times to count

2. How frequently do you update your status (or profile) on Twitter? (circle one)

   a) Never updated my status or profile

b) Once or twice a year.
c) A few times a month or less
d) Weekly
e) Daily
f) More than once a day
g) More than five times a day
h) Too many times to count

3. About how many friends do you have on Twitter? [Open-ended]
4. How old are you? [Open-ended]
5. What's the highest level of education you received? [Open-ended]

Talk Aloud procedure

I'm going to ask you to perform a few different tasks on the computer, and I want you to talk aloud as you do these tasks. I would especially like to hear where you're planning to go, how you plan to get there, why you decide to take a certain path, anything that confuses or frustrates you, and anything that you think is really easy. This session will take between 10 and 15 minutes. I'll be timing you so that we stay within the 15 minute time limit, but I want you to focus on what you tell me about what you're thinking as you do the tasks and not on doing the tasks quickly. It's not as important that you finish the task as it is for you to explain to me why you did things a certain way. Do you have any questions?

We're going to start with a couple of practice runs so that you understand what it means to do a talk aloud. First, let's say I'm doing a talk aloud. My interaction centers on my Twitter use [new task chosen each time for interviewer so it does not become routine]. [Interview does a task while describing thoughts, confusions, emotions, plans, etc.]

Walkthrough Outline

3 tasks involving your use of Twitter

Prompts:

- What's going on?
- What are you thinking?
- What are you looking for?
- You look confused, are you?
- It looks like you're thinking, what are you thinking?

1. Start at the home page for your favorite social media presence. Describe what you see and your typical interaction with the site.

2. Start at your own profile page on Twitter.

Describe the posts you have made most recently, how and why you made them.

3. Start at the home page for Twitter.

Describe the most recent interactions/messages/replies/likes you've made on content posted by others. Describe why and how you decided to do each.

Conclusion

We’ve talked a bit today about how you use Twitter to learn about what others are thinking and doing, and also about how you use Twitter to express yourself. Is there anything you think we missed? Or anything else you might want to tell me?

Thank you very much for your participation today. If you have any further questions or concerns, please feel free to reach me using the contact information listed on your informed consent form.

Appendix B

Phase Two Interview Instrument

Introduction

I'm collecting information for a project on self-expression in social media. I will be sharing the information that I gather here with the other member of my research team and with the Rutgers University Institutional Review Board, but your name will not be associated with any other information.

First I'm going to ask you some questions about your use of social media.

Demographic questions

1. How frequently do you visit Twitter? (circle one)

   a) Never used it
   b) Once or twice a year
   c) A few times a month or less
   d) Weekly
   e) Daily
   f) More than once a day (but less than 5)
   g) Five to 10 times a day
   h) Too many times to count

2. How frequently do you update your status (or profile) on Twitter? (circle one)

   a) Never updated my status or profile

b) Once or twice a year.
c) A few times a month or less
d) Weekly
e) Daily
f) More than once a day
g) More than five times a day
h) Too many times to count

3. About how many friends do you have on Twitter? [Open-ended]
4. How old are you? [Open-ended]
5. What's the highest level of education you received? [Open-ended]

Talk Aloud procedure

I'm going to ask you to perform a few different tasks on the computer, and I want you to talk aloud as you do these tasks. I would especially like to hear where you're planning to go, how you plan to get there, why you decide to take a certain path, anything that confuses or frustrates you, and anything that you think is really easy. This session will take between 20 and 30 minutes. I'll be timing you so that we stay within the 30 minute time limit, but I want you to focus on what you tell me about what you're thinking as you do the tasks and not on doing the tasks quickly. It's not as important that you finish the task as it is for you to explain to me why you did things a certain way. Do you have any questions?

We’re going to start with a couple of practice runs so that you understand what it means to do a talk aloud. First, let’s say I’m doing a talk aloud. My interaction centers on my Twitter use [new task chosen each time for interviewer so it does not become routine]. [Interview does a task while describing thoughts, confusions, emotions, plans, etc.]

Walkthrough Outline

8 tasks involving your use of Twitter

Prompts:

- What’s going on?
- What are you thinking?
- What are you looking for?
- You look confused, are you?
- It looks like you’re thinking, what are you thinking?

1. Start at the home page for your favorite social media presence. Describe what you see and your typical interaction with the site.

2. Start at your own profile page on Twitter.

Describe how and why you might update your profile

3. Start at your own profile page on Twitter.

Describe the posts you have made most recently, how and why you made them.

4. Start at your own profile page on Twitter.

Describe the posts or messages that others have sent to you, and how and why they were made.

5. Start at the home page for Twitter.

Describe what appears in your news feed, and what you are most interested in seeing

6. Start at the home page for Twitter.

Describe the most recent interactions/messages/replies/likes you’ve made on content posted by others. Describe why and how you decided to do each.

7. Start at the home page for Twitter.

Suppose you’ve met someone new at a recent dinner party. How would you learn more about them using Twitter?

8. Start at the home page for Twitter.

Suppose you’ve met someone new at a recent party or gathering of friends. How would you interact with them using Twitter?

Conclusion

We’ve talked a bit today about how you use Twitter to learn about what others are thinking and doing, and also about how you use Twitter to express yourself. Is there anything you think we missed? Or anything else you might want to tell me?

Thank you very much for your participation today. If you have any further questions or concerns, please feel free to reach me using the contact information listed on your informed consent form.

Appendix C

Qualitative Codebook

Table 9

Qualitative codebook

Code	Title	Notes
1	Components of network	
1.1	Friends	
1.1.1	Disparate	
1.1.2	Local	
1.2	Family	
1.2.1	Disparate	
1.2.2	Local	
1.3	Co-workers	
1.3.1	Disparate	
1.3.2	Local	
1.4	Anonymous	
1.5	Interest groups	
1.5.1	Religious	
1.5.2	Sports	
1.5.3	Entertainment	
1.5.4	Internet	
1.5.5	Polititcs	
1.7	Suggested	Network member discovery: active
1.7.1	Implied	Due to co-activity with existing network members
1.7.2	Automatic	Due to combined weight of many existing connections
1.7.3	Expected	Due to weight of a few strong existing connections
1.8	Inferred	Network member discovery: passive
1.8.1	Implicit	Due to platform/algorithmic suggestions
1.8.2	Discovered	Due to discovery of existing offline connections
1.8.3	Recovered	Due to recovery of existing offline connections
2	Content of communication	
2.1	Political	
2.1.1	Prepared	
2.1.2	Ad hoc	

2.3	Social	As opposed to personal: directly involving others
2.3.1	Cohesion	Expressions of co-group membership of others
2.3.1.1	Bonding	Reinforcement of in-group value
2.3.1.2	Planning	
2.3.1.2.1	Extemporaneous	
2.3.1.2.1	Deliberate	
2.3.1.3	Reinforcing	Expressions that validate self group membership
2.3.1.4	Memories	
2.3.1.5	Exclusive	Expressions of denial of out-group members
2.4	Personal	As opposed to social: not directly involving others
2.4.1	Opinion	
2.4.1.1	Planned	
2.4.1.2	Ad hoc	
2.4.2	Location	
2.4.3	Emotion	
2.4.3.1	Positive	
2.4.3.1	Negative	
2.5	Here/now	Self-updates including selfies, meal photos, etc.
2.5.1	Location	
2.5.2	Activity	
2.5.3	Desire	"Can't wait to x"
2.5.4	Fulfillment	"Can't believe I just y"
2.6	Other	
3	Device	
3.1	Phone	
3.1.1	Convenience	
3.1.2	Difficulty	
3.2	Tablet	
3.2.1	Convenience	
3.2.2	Difficulty	
3.3	Computer	
3.3.1	Convenience	
3.3.2	Difficulty	
4	Information seeking mode	
4.1	Extractive	Review of archived data
4.1.1	Search	Selective review based on filtered search

4.1.2	History	Unselective review of all history related to target
4.1.3	Back-pathing	Linked traverse of interaction history
4.2	Interactive	Review of contextual records of interactions between target and others
4.2.1	Social circles	Examination of target's network
4.2.2	Integration	Examination of ties between target and target's network
4.2.3	Ephemera	Examination of interaction records
4.2.4	Bridging	Examination of ties between target and investigator's network
4.3	Passive	Undirected monitoring of message traffic stream
4.3.1	Lurking	Unfiltered
4.3.2	Matching	Filtered to include target
4.3.3	Action	Responses designed to stimulate target
4.3.3.1	Liking	
4.3.3.2	Anonymity	
4.3.3.3	Participation	
4.4	Active	Directed monitoring of message traffic of target, examination of target profile, etc.
4.4.1	Integration	Inference based on interactions between unknown third parties and the target
4.4.2	Implication	Inference based on interactions between known proxies and the target
4.4.3	Interrogation	Inference based on directly messaging the target
5	Indirect communication	
5.1	Commenting	
5.2	Reacting	
5.2.1	Expectation	Due to reciprocity, time scale, etc.
5.2.2	Non-reaction	Implicit communication meaning in non-reaction
5.2.3	Automatic	Normative enforcement, e.g.
5.3	Reinforcing	Communication aimed at encouraging repeat behaviors
5.4	Retweeting	
5.4.1	Promotion	To raise awareness
5.4.2	Comment	To continue a debate
5.4.3	Visibility	To assist a task
5.4.4	Reinforcement	To demonstrate support or encouragement

5.4.5	Proxy	To leverage the cachet/reputation of original poster
5.5	Interest	
5.5.1	Hobby	
5.5.2	Entertainment	
5.5.3	Politics	
5.5.4	Activity	
5.5.5	Sports	
5.5.6	News	
5.5.7	Religious	
6	Purpose of communication	
6.1	Employment	
6.2	Expression	
6.3	Family connection	
6.4	Separation	
6.5	Political	
6.6	Proxy	To leverage the cachet/reputation of original poster
6.7	Share	
6.8	Connection	
6.9	Monitoring	
6.10	Religious	
6.11	Activism	
7	Purpose of information seeking	
7.1	Employment	
7.1.1	Professional networking	
7.2	Family connections	
7.2.1	Monitoring	
7.2.2	Updating	
7.2.3	Sharing	
7.2.4	Memories	
7.2.5	Memorials	
7.2.6	Planning	
7.2.7	Opinion	
7.2.8	Caring	
7.2.9	Religious	
7.3	Social connections	
7.3.1	Monitoring	
7.3.2	Trolling	
7.3.4	Interest	
7.3.5	Political	

7.3.6	Friends	
7.3.7	Family	
7.3.8	Entertainment	
7.3.9	Question	
7.3.10	Buying/Selling	
7.3.11	Planning	
7.3.12	Memories	
7.3.13	Memorials	
7.3.14	Caring	
7.3.15	Religious	
7.3.16	Activism	

Appendix D

Interview Consent Form with Audio/Visual Recording

I am a PhD candidate in the School of Communication and Information at Rutgers University, and I am conducting interviews in order to gather data for my dissertation research. I am studying how individuals use social media to express their identities and learn about the identities of others.

During this study, you will be asked to answer some questions concerning your use of social media. This interview was designed to be approximately a half hour in length. However, please feel free to expand on the topic or talk about related ideas. Also, if there are any questions you would rather not answer or that you do not feel comfortable answering, please say so and we will stop the interview or move on to the next question, whichever you prefer.

This research is confidential. Confidential means that the research records will include some information about you and this information will be stored in such a manner that some linkage between your identity and the response in the research exists. Some of the information collected about you includes your use of social media in your daily life, and the types of people you interact with via social media. Please note that we will keep this information confidential by limiting individual's access to the research data and keeping it in a secure location. We will not collect any data directly from any social networking service related to you or your identity. The data gathered in this study are confidential with respect to your personal identity unless you specify otherwise.

The research team and the Institutional Review Board at Rutgers University are the only parties that will be allowed to see the data, except as may be required by law. If a report of this study is published, or the results are presented at a professional conference, only group results will be stated. All study data will be kept for three years and then destroyed.

You are aware that your participation in this interview is voluntary. You understand the intent and purpose of this research. If, for any reason, at any time, you wish to stop the interview, you may do so without having to give an explanation.

There are no foreseeable risks to participation in this study.

You will receive no direct benefit from taking part in this study. However, you will receive compensation in the form of a $20 VISA gift card for completing the entire study.

The recording(s) will be used for analysis by the research team.

The recording(s) will include not include your name or other identifying information, and will not record images of your face or any other distinguishing features. If you say anything that you believe at a later point may be hurtful and/or damage your reputation, then you can ask the interviewer to rewind the recording and record over such information OR you can ask that certain text be removed from the dataset/transcripts.

The recording(s) will be stored on a password-protected computer in a locked office, and will not include any personally identifying information**.** The recordings will be kept for one year and then destroyed.

If you have any questions about the study or study procedures, you may contact the researchers at**:**

Charles File
Rutgers University School of Communication and Information
4 Huntington St.
New Brunswick, NJ
08901

Telephone: (908) 208-9757

Email: chasfile@rutgers.edu

Marie L. Radford, Ph.D.
Rutgers University School of Communication and Information
4 Huntington St.
New Brunswick, NJ
08901

Telephone: (848) 932-8797
Fax: (732) 932-6919
Email: mradford@comminfo.rutgers.edu

If you have any questions about your rights as a research participant, you can contact the Institutional Review Board at Rutgers (which is a committee that reviews research studies in order to protect research participants) at:

Institutional Review Board
Rutgers University, the State University of New Jersey
Liberty Plaza / Suite 3200
335 George Street, 3rd Floor

New Brunswick, NJ 08901

Phone: 732-235-2866

Email: humansubjects@orsp.rutgers.edu

You will be offered a copy of this consent form that you may keep for your own reference.

Once you have read the above form and, with the understanding that you can withdraw at any time and for whatever reason, you need to let me know your decision to participate in today's interview.

Your signature on this form grants the investigator named above permission to record you as described above during participation in the above-referenced study. The investigator will not use the recording(s) for any other reason than that/those stated in the consent form without your written permission.

Subject (Print) ____

For IRB Use Only. This Section Must be Included on the Consent Form and Cannot Be Altered Except For Updates to the Version Date.

IRB Stamp Box	**IRB Stamp Box**

Subject Signature _________________________________ Date _______________________

Principal Investigator Signature ______________________ Date ___________________

For IRB Use Only. This Section Must be Included on the Consent Form and Cannot Be Altered Except For Updates to the Version Date.

IRB Stamp Box	**IRB Stamp Box**

References

Antheunis, M. L., Valkenburg, P. M., & Peter, J. (2010). Getting acquainted through social network sites: Testing a model of online uncertainty reduction and social attraction. *Computers in Human Behavior, 26*(1), 100-109.

Bargh, J. A., McKenna, K. Y., & Fitzsimons, G. M. (2002). Can you see the real me? Activation and expression of the "true self" on the Internet. *Journal of Social Issues*, *58*(1), 33-48.

Baron, N. S. (2010). *Always on: Language in an online and mobile world*. New York: Oxford University Press.

Bawden, D. & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science 35*(2), 180-191.

Baym, N. (2010). *Personal connections in the digital age*. Cambridge, UK: Polity.

Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, *1*(2), 99-112.

Berger, C. R., Gardner, R. R., Parks, M. R., Schulman, L., & Miller, G. R. (1976). Interpersonal epistemology and interpersonal communication. *Explorations in Interpersonal Communication, 5*, 149-172.

Boyd, D. (2006). Friends, Friendsters, and MySpace top 8: Writing community into being on social network sites. *First Monday 11*, 12. Retrieved from: http://firstmonday.org/article/view/1418/1336

Biocca, F. & Levy, M. R. (1995). *Communication in the Age of Virtual Reality*. Hillsdale, NJ: Erlbaum.

Boyd, D., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer‐Mediated Communication, 13*(1), 210-230.

Boyd, D. & Ellison, N.B. (2013). Sociality through social network sites. In Dutton, W. H. (Ed.)*, The Oxford Handbook of Internet Studies* (pp. 151-169). Oxford University Press, Oxford, UK.

Camden, C., Motley, M. T., & Wilson, A. (1984). White lies in interpersonal communication: A taxonomy and preliminary investigation of social motivations. *Western Journal of Speech Communication, 48*(4), 309-325.

Castells, M. (2011). *The rise of the network society, 6th ed.* Malden, Mass.: Blackwell Publishers.

Chayko, M. (2008). *Portable communities*. Albany, NY: State University of New York Press.

Chovil, N. (1994). Equivocation as an interactional event. In B. H. Spitzberg & W. R. Cupach (Eds.), *The dark side of interpersonal communication* (pp. 105-123). New York, NY: Routledge.

Counts, S. & Stetcher, K. (2009). Self-presentation of personality during online profile creation. *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM).*

Coupland, N., Giles, H., & Wiemann, J. M. (Eds.). (1991). *"Miscommunication" and problematic talk.* Newbury Park, CA: Sage.

Culnan, M. J., & Markus, M. L., (1987). Information technologies. In F. M. Jablin, L. L. Putnam, K. H. Roberts, & L. W. Porter (Eds.), *Handbook of organizational communication: An interdisciplinary perspective* (pp. 420-443). Thousand Oaks, CA: Sage.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Daniels, J. (2009). *Cyber racism: White supremacy online and the new attack on civil rights*. New York: Rowman & Littlefield Publishers.

Derlega, V., Metts, S., Petronio, S., & Margulis, S. T. (1993). *Self-disclosure*. Thousand Oaks, CA: Sage.

Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning* (p. 29). ACM.

Donath, J. (2007). Signals in social supernets. *Journal of Computer*‐*Mediated Communication, 13*(1), 231-251.

Donath, J., & Boyd, D. (2004). Public displays of connection. *BT Technology Journal, 22*, 71–82.

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self presentation processes in the online dating environment. *Journal of Computer-Mediated Communication, 11*, 415–441.

Ellison, N. B., Hancock, J. T. & Toma, C. L. (2012). Profile as promise: A framework for conceptualizing veracity in online dating self-presentations. *New Media & Society, 14*(1), 45-62.

Fazio, R. H., Chen, J. M., McDonel, E. C., & Sherman, S. J. (1982). Attitude accessibility, attitude-behavior consistency, and the strength of the object-evaluation association. *Journal of Experimental Social Psychology, 18*(4), 339-357.

Gal, S. (2002). A semiotics of the public/private distinction. Differences: *A Journal of Feminist Cultural Studies, 13*(1), 77-95.

Garrett, J. J. (2010). *The elements of user experience: User-centered design for the web and beyond*. Pearson Education.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528.

Gibbs, J. L., Ellison, N.B., & Heino, R.B. (2006). Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in Internet dating. *Communication Research, 33*, 152–177.

Gibbs, J.L., Ellison, N.B., & Lai, C-H. (2011). First comes love, then comes Google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. *Communication Research, 38*, 70-100.

Gilbert, E., & Karahalios, K. (2009, April). Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 211-220.

Goffman, E. (1959). *The presentation of self in everyday life*. New York, NY: Doubleday

Goffman, E. (1967). *Interaction ritual: Essays in face to face behavior*. Chicago, IL: AldineTransaction.

Gosling, S. D., Gaddis, S., & Vazire, S. (2007). Personality impressions based on Facebook profiles. *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM), 7*.

Gonzales, A. L., & Hancock, J. T. (2008). Identity shift in computer-mediated environments. *Media Psychology, 11*(2), 167-185.

Greenwood, S., Perrin, A., & Duggan, M. (2016, November). Social Media Update 2016. Pew Research Center. Retrieved from: http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/11/10132827/PI_2016.11.11_Social-Media-Update_FINAL.pdf

Hampton, K. & Wellman, B. (1999). Netville online and offline: Observing and surveying a wired suburb. *American Behavioral Scientist 43*(3), 475–492.

Hampton, K. & Wellman, B. (2003). Neighboring in netville: How the Internet supports community and social capital in a wired suburb. *City & Community 2*(4), 277-311.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011, May). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237-246). ACM.

Heino, R.D., Ellison, N.B., & Gibbs, J.L. (2010). Relationshopping: Investigating the market metaphor in online dating. *Journal of Social and Personal Relationships, 27*, 428-447.

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*(3), 319.

Hillier, L., Mitchell, K. J., & Ybarra, M. L. (2012). The Internet as a safety net: Findings from a series of online focus groups with LGB and non-LGB young people in the United States. *Journal of LGBT Youth 9*(3), 225-246.

Hogg, M. A. (2000). Subjective uncertainty reduction through self-categorization: A motivational theory of social identity processes. *European review of social psychology*, *11*(1), 223-255.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80-88). ACM.

Horgan, J. (1995). From complexity to perplexity. *Scientific American*, *272*(6), 104-109.

James, W. (1890). *The principles of psychology (Vol. 1).* New York: Holt.

Joinson, A.N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology, 31*, 177–192.

Katz, J. E. & Aakhus, M. A. (2002). *Perpetual contact*. Cambridge: Cambridge University Press.

Kelly, A. E., & Rodriguez, R. R. (2006). Publicly committing oneself to an identity. *Basic and Applied Social Psychology, 28*(2), 185-191.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web* (pp. 591-600). ACM.

Kim, T. & Biocca, F. (1997). Telepresence via television: Two dimensions of telepresence may have different connections to memory and persuasion. *Journal of Computer-Mediated Communication, 3*(2).

Lange, P. G. (2007). Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication, 13,* 361–380.

Lea, M., & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing and Electronic Commerce, 2*(3-4), 321-341.

Ling, R. (2004). *The mobile connection*. San Francisco, CA: Morgan Kaufmann.

Lippard, P. V. (1988). “Ask me no questions, I'll tell you no lies”: Situational exigencies for interpersonal deception. *Western Journal of Communication, 52*(1), 91-103.

Liu, H. (2007). Social network profiles as taste performances. *Journal of Computer-Mediated Communication, 13*(1), 252-275.

Madden, M., & Smith, A. (2010). *Reputation management and social media*. Pew Internet & American Life Project.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*(2), 63.

Marwick, A.E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*, 114-133.

McEwan, B. (2013). Sharing, caring, and surveilling: An actor–partner interdependence model examination of Facebook relational maintenance strategies. *Cyberpsychology, Behavior, and Social Networking*, *16*(12), 863-869.

McPherson, K., Huotari, K., Cheng, F., Humphrey, D., Cheshire, C., & Brooks, A. L. (2012, February). Glitter: a mixed-methods study of twitter use during glee broadcasts. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion* (pp. 167-170). ACM.

Meyrowitz, J. (1985). *No Sense of Place*. New York: Oxford University Press.

Miller, L. C., Berg, J. H., & Archer, R. L. (1983). Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology, 44*, 1234-1244.

Naaman, M., Boase, J., Lai, C. (2010). Is it really about me? Message content in social awareness streams. *Proceedings of the ACM Conference on Computer-Supported Collaborative Work.* ACM.

Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge, UK: Cambridge University Press.

Oh, H. J., Ozkaya, E., & LaRose, R. (2014). How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Computers in Human Behavior*, *30*, 69-78.

O'Sullivan, P.B. (2000). What you don't know won't hurt me: Impression management functions of communication channels in relationships. *Human Communication Research, 26*, 403-431.

Parks, M. R. (1981). Ideology in interpersonal communication: Off the couch and into the world. *Annals of the International Communication Association*, *5*(1), 79-107.

Parks, M. (2011). Social network sites as virtual communities. In A. Paparachissi (Ed.), *A networked self: Identity, community, and culture on social network sites*. New York: Routledge. 105-123

Parmelee, J. H., & Bichard, S. L. (2011*). Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public*. Lexington Books.

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, *31*(6), 539-548.

Przybylski, A. K., Murayama, K., DeHaan, C. R., & Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior 29*(4), 1841-1848.

Qian, H., & Scott, C. R. (2007). Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication*, *12*, 1428-1451.

Ramirez, A., Walther, J. B., Burgoon, J. K., & Sunnafrank, M. (2002). Information-seeking strategies, uncertainty, and computer-mediated communication. *Human Communication Research*, *28*(2), 213-228.

Rogers, C. R. (1951). *Client-centered therapy: Its current practice, implications and theory*. London: Constable.

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018, March 17) How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times,* p. A1.

Schlenker, B. R., & Trudeau, J. V. (1990). Impact of self-presentations on private self-beliefs: Effects of prior self-beliefs and misattribution. *Journal of Personality and Social Psychology, 58*(1), 22.

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology, 43*(1), 133-168.

Scott, J. (2017). *Social network analysis*. New York: Sage.

Shklovski, I., Palen, L., & Sutton, J. (2008). Finding community through information and communication technology during disaster events. *ACM 2008 Conference on Computer Supported Cooperative Work,* San Diego, CA.

Shklovski, I., Burke, M., Kiesler S., & Kraut, R. (2010). Technology adoption and use in the aftermath of Hurricane Katrina in New Orleans. *American Behavioral Scientist 53*(8), 1228-1246.

Small, T. A. (2011). What the hashtag? A content analysis of Canadian politics on Twitter. *Information, Communication & Society, 14*(6), 872-895.

Snelson, C. L. (2016). Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods, 15*(1).

Stern, S. R. (2004). Expressions of identity online: Prominent features and gender differences in adolescents' World Wide Web home pages. *Journal of Broadcasting & Electronic Media, 48*(2), 218-243.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In W. G. Austin & S. Worchel (Eds.), *Psychology of intergroup relations* (pp. 7-24). St. Louis, MO: Burnham.

Tice, D. M. (1992). Self-concept change and self-presentation: the looking glass self is also a magnifying glass. *Journal of Personality and Social Psychology, 63*(3), 435.

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human Communication Research*, *28*(3), 317-348.

Trevino, L. K., Lengel, R. H., & Daft, R. L. (1987). Media symbolism, media richness, and media choice in organizations a symbolic interactionist perspective. *Communication Research*, *14*(5), 553-574.

Tufekci, Z. (2010.0 Who acquires friends through social media and why? "Rich get richer" versus "seek and ye shall find." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.* Menlo Park, CA: AAAI Press, 170–177.

Turkle, S. (1984). *The second self*. New York: Simon and Schuster.

Vazire, S., & Gosling, S. D. (2004). e-Perceptions: personality impressions based on personal websites. *Journal of Personality and Social Psychology, 87*(1), 123.

Wakefield, J. S., Warren, S. J., & Alsobrook, M. (2011). Learning and teaching as communicative actions: A mixed-methods Twitter study. *Knowledge Management & E-Learning, 3*(4), 563.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction a relational perspective. *Communication research, 19*(1), 52-90.

Walther, J. B. (1996). Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*(1), 3-43.

Walther, J. B., & Burgoon, J. K. (1992). Relational communication in computer-mediated interaction. *Human Communication Research, 19*(1), 50-88.

Walther, J., & Parks, M. (2002). Cues filtered out, cues filtered in: Computer-mediated communication and relationships. *Handbook of interpersonal communication (3rd ed)*, 529–563.

Walther, J., Van Der Heide, K., Westerman, & Tong (2008). The role of friends' appearance and behavior on evaluations of individuals on Facebook: Are we known by the company we keep? *Human Communication Research 34(1),* 28.

Walther, J., Van Der Heide, B., Hamel, L., & Shulman, H. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook". *Communication Research 36(2),* 229–253.

Wu, T. (2003). Network neutrality, broadband discrimination. *Journal on Telecommunications & High Tech, 2*.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval* (pp. 338-349). Springer, Berlin, Heidelberg.