EXAMINING ADAPTATION IN COMPLEX ONLINE SOCIAL SYSTEMS

by

Ross Schuchard
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Social Science

Committee:

_____	Dr. Andrew Crooks, Dissertation Director
_____	Dr. Robert Axtell, Committee Member
_____	Dr. Arie Croitoru, Committee Member
_____	Dr. Anthony Stefanidis, Committee Member
_____	Dr. A. Trevor Thrall, Committee Member
_____	Dr. Jason Kinser, Chair, Department of Computational and Data Sciences
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Summer Semester 2019 George Mason University Fairfax, VA

EXAMINING ADAPTATION IN COMPLEX ONLINE SOCIAL SYSTEMS

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Ross Schuchard
Master of Arts Interdisciplinary Studies
George Mason University, 2015
Bachelor of Science
United States Military Academy, 2004

Director: Andrew Crooks, Associate Professor
Department of Computational and Data Sciences

Summer Semester 2019
George Mason University
Fairfax, VA

Copyright 2019 Ross Schuchard
All Rights Reserved

DEDICATION

To Sarah, John, Lila and Anna.

ACKNOWLEDGEMENTS

First and foremost, I must thank my advisor and dissertation chair, Dr. Andrew Crooks, for guiding me through this journey. His tireless work ethic and incredible mentorship inspired me beyond words and cement his legacy as one of the best leaders I have encountered in my life. I must also recognize the amazing talent comprising the remainder of my dissertation committee: Dr. Rob Axell, Dr. Arie Croitoru, Dr. Tony Stefanidis and Dr. Trevor Thrall. This diverse group of experts provided me not only essential feedback, but constantly challenged me to seek greater depths of understanding beyond my comfort level. I cannot imagine a better sounding board to validate my PhD efforts. I will never forget the "Lads in the Lab" sessions.

I would be remiss if I did not directly thank Karen Underwood for all of her efforts over the years. It is not easy dealing with the everyday workload of the CSS-CDS-GMU machine, but having to account for the emergent bureaucratic requirements of a DoD-sponsored student added an additional layer of true 'complexity' to her plate. Thanks for always taking care of me.

Finally, I must thank the U.S. Army for funding my graduate studies. Specifically, I would like to acknowledge the Functional Area 49 (Operations Research) leadership for allowing me to traverse a graduate studies path outside of the traditional norms for an Army officer.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND/OR SYMBOLS

Agent-Based Model ..... ABM
Application Programming Interface ..... API
Bot-hunter ..... BH
Botometer ..... BT
Carnegie Mellon University ..... CMU
Computational Social Science ..... CSS
Defense Advanced Research Projects Agency ..... DARPA
DeBot ..... DB
European Union ..... EU
Information and Communications Technology ..... ICT
Online Social Network ..... OSN
Organisation for Economic Co-operation and Development ..... OECD
Russian Federation ..... RU
Social Identity Model of Collective Action ..... SIMCA
Social Network Analysis ..... SNA
Turkey ..... TR
Uniform Resource Locator ..... URL
United Kingdom ..... UK
United States ..... US
Virtual Private Network ..... VPN
World Wide Web ..... WWW

ABSTRACT

EXAMINING ADAPTATION IN COMPLEX ONLINE SOCIAL SYSTEMS

Ross Schuchard, Ph.D.

George Mason University, 2019

Dissertation Director: Dr. Andrew Crooks

Online social systems, comprised of social media services and platforms including social networking (e.g. Facebook, LinkedIn) and microblogging (e.g. Twitter, Sina Weibo) applications, continue to gain traction among an ever-increasing global user base. The growing reliance upon online social systems to augment an individual's daily workflow and the resulting interdependence between human and technical systems provide sufficient evidence to classify them as socio-technical systems. These interdependencies are complex in nature and are best defined from a complex adaptive system (CAS) perspective.

It is through a CAS lens that this dissertation examines two types of adaptation in online social systems using an array of Computational Social Science (CSS) tools. In the first type of adaptation, human actors are no longer the sole participants in online social systems, since social bots, or automated software mimicking humans, have emerged as potential threats to stifle or amplify certain online conversation narratives. The first part

of the dissertation addresses adaptation to these new types of actors by presenting a novel social bot analysis framework designed to determine the pervasiveness and relative importance of social bots within various online conversations. In the second form of adaptation, individual citizens and government entities modify their behaviors in relation to each other through censorship circumvention or detection. This second form of adaptation in the dissertation investigates the rise of digital censorship in online social systems, creating a new agent-based model inspired by the findings from an evaluation of a Turkish digital censorship campaign.

The social bot analysis framework results consistently showed that while users identified as social bots only comprised a small portion of total accounts within the overall research corpus, they account for a significantly large portion of prominent centrality rankings across all observed online conversations. Furthermore, bot classification results, when using multiple bot detection platforms, exhibited minimal overlap, thus affirming that different bot detection algorithms focus on the various types of bots that exist. Finally, the results of the Turkish digital censorship campaign showed marginal effectiveness as some Turkish citizens circumvented the censorship policies, thus highlighting an individual decision cycle to risk punishment and engage in online activities. The recognition of this citizen decision cycle served as the basis for the adaptation to digital censorship model, which used empirical evidence to stylize and template a simulation censorship environment. In all, this dissertation presents a unique CSS methodology to observe, measure and simulate social adaptation that exists in complex online social systems.

CHAPTER 1. INTRODUCTION

The following introductory chapter provides the overarching motivation behind this dissertation. Section 1.1 presents a framework dedicated to analyzing adaptation in online social systems. Section 1.2 follows with two areas of focus and associated research questions for those topics. Section 1.3 introduces relevant background literature to place the key components of the dissertation into context, while Section 1.4 concludes with an overview of the dissertation's structure.

1.1. Motivation of the Dissertation

The evolution of the World Wide Web (WWW) throughout the past decade and a half has ushered in a wave of new online social systems that are pervasive around the world. These systems, comprised of social media services and platforms, to include social networking (e.g. Facebook, LinkedIn), microblogging (e.g. Twitter, Sina Weibo) and crowdsourcing (e.g. Wikipedia, OpenStreetMap) applications, have come to characterize the participatory nature and the user-generated content driving the Web 2.0 (O'Reilly, 2005) period and beyond. The explosive and sustained worldwide growth in online social system participation has resulted in these systems becoming an increasingly primary source for news (Pew Research Center, 2017) as well as enabling modern protests and political discussions (Pew Research Center, 2018). The 'digital exhaust' created by these online social systems produce tangible metadata that are readily accessible for

researchers to investigate social interactions and contributed content at unprecedented scales. Although the role of academia in analyzing online social systems effects on real-world physical activities is relatively young and requires much additional scrutiny to develop associated research standards (Ruths & Pfeffer, 2014; Tufekci, 2014), many recent studies have produced noteworthy results examining the hybrid characteristics that emerge at the intersection of these 'cyber' and 'physical' interactions, to include political participation (e.g. Bode & Dalrymple, 2016; Gibson & Cantijoch, 2013; Vaccari et al., 2015), emergency event response (Crooks et al., 2013; Imran et al., 2015; Sakaki et al., 2013) and extremist activities (Berger & Morgan, 2015; Ferrara et al., 2016).

The ability to discover potential causal linkages between online cyber actions and offline physical activities is a logical research endeavor that could produce substantial benefits. Theocharis and Van Deth (2018) propose an entire multi-dimensional taxonomy to both capture and differentiate between offline and online citizen participation in a standardized fashion. However, Croitoru et al. (2015) and Althoff et al. (2016) point out the significant scientific challenges and associated elusiveness in delineating tractable certainty between online and offline activities. Such an endeavor may not be justified given the increasing ubiquity and fusion of online social systems in the everyday activities of individuals. The growing reliance upon online social systems to augment an individual's daily workflow and the resulting interdependence between human and technical systems will only further complicate how to define where a cyber and physical divide exists, if it exists at all. Kleinberg (2008) declares a convergence of technological and social networks due to the fact that social forces drove changes to the underlying

technical operating specifications of online social systems. Scholtes et al. (2014) goes so far as to argue that underlying technical systems (such as the WWW) and their associated users are inherently coupled and thus inseparable. As will be discussed below, such points of view suggest looking at the activities emanating from online social systems not from a differentiated cyber or physical standpoint, but from the standpoint that the activities are artifacts of a socio-technical system.

The concept of a socio-technical system arose in the 1950s (Trist & Bamforth, 1951), when industrial systems research sought to improve both manufacturing efficiency and employee workplace satisfaction. The goal of the socio-technical approach was to view technological machinery and human workers as combined systems, rather than isolated components (Mumford, 2006). Given the inherent coupling of technology and humans in today's WWW applications, one could easily interpret online social systems as emblematic of socio-technical systems. In examining the evolutionary history of media, Stöber (2004) described the historical emergence of new media technology as not merely a consequence of technical inventions, but as a coupled process between innovation and social institutionalization. Fuchs (2005, 2007) has argued extensively that the WWW, especially the Web 2.0 period and beyond, should be viewed as a self-organizing socio-technical system, as opposed to a purely technical system, in which human communicative actions continually refine the purpose and structure of the underlying technical specifications of the Internet. Figure 1 captures the intricacies of a socio-technical system and the inherent interdependencies that exist between the social and technical subsystems. This identification of the WWW as self-organizing and adaptive in

nature presupposes the greater identification of the WWW, and online social systems, as complex adaptive systems. Fuchs (2007) has not been alone in making this assertion, as numerous works have followed in describing the WWW and online social systems as socio-technical systems that are real-world complex adaptive systems (Borge-Holthoefer et al., 2013; Niederer & van Dijck, 2010; Sayama et al., 2013).

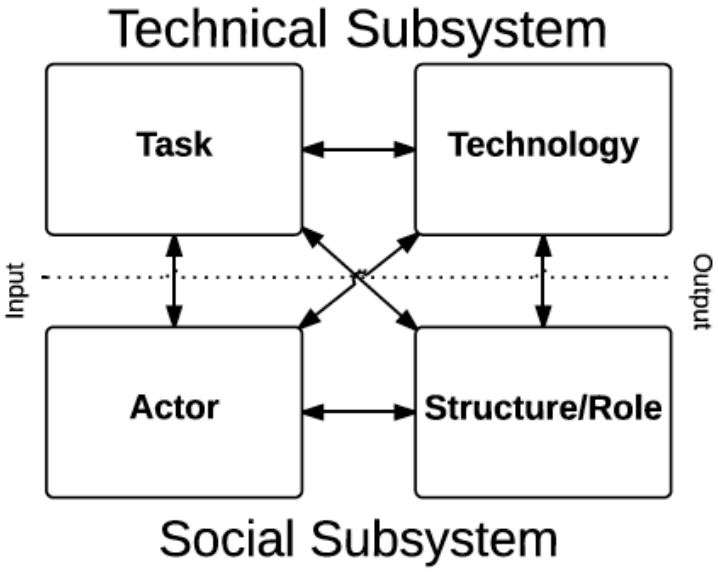


Figure 1: Socio-technical system comprised of interdependent technical and social subsystems (Source: Calero Valdez, Brauner, & Ziefle, 2016)

The concept of near-decomposability (Simon, 1996), in which interdependence of system components generates complexity, is highly applicable in classifying the deeply interdependent nature of the social and technical components of online social systems. It adds further credence to the earlier assertions that view them as complex adaptive

systems. Just as we have observed the WWW evolve[1] through a Web 1.0 to Web 2.0 transition, we anticipate further evolution through a Web 3.0 and beyond (Berners-Lee et al., 2001; Hendler, 2009), since persistent adaptation continues to drive online social systems into artifacts of increasing importance in today's global society. Given that adaptation is a key observable component of complex adaptive systems (Miller & Page, 2007; Mitchell, 2011; Sayama et al., 2013), the examination of adaptation may serve as a more realistic framework through which to view the impact of evolving online social systems.

1.2. Research Questions

The primary theme of this dissertation research is the identification and analysis of adaptation in complex online social systems through the application of computational social science (CSS) methodologies and tools such as automated information extraction, agent-based modeling (ABM) and social network analysis (SNA). Specifically, this research examines social adaptation in complex online social systems from two distinct perspectives: (1) the adaptation of traditional social actors (i.e. humans) to new types of actors (i.e. social bots) in online social networks (Section 1.2.1) and (2) the adaptation of social actors (i.e. citizens, governments) to digital censorship practices in online social networks (Section 1.2.2). A general description of each adaptation perspective follows and serves as a primer to more in-depth discussions of the topics put forth in the

[1] Academic and commercial publications attempting to formally describe the evolution of the WWW are not in universal agreement as to the precision of terms and associated timelines, but a general consensus agrees to the categorization of Web 1.0 as a system in which users accessed content online and Web 2.0 as a system in which users became content creators in addition to consumers (Cormode & Krishnamurthy, 2008).

subsequent Background Literature of Interest (Section 1.3) and the Dissertation Structure (Section 1.4) portions of this introductory chapter.

1.2.1. Adaptation to Social Bot Actors in Online Social Networks

Social bots, or automated software or computer algorithms designed to mimic human behavior and/or engage with human actors, have become ubiquitous actors in OSNs (Howard et al., 2018). The implications of human actors engaging, intentionally or not, with bot actors are numerous given the relative ease of deploying bots at scale. As I highlight in Section 1.3.1, there is a growing body of research dedicated to detecting bots of ever-increasing sophistication, but there is a complementary requirement to assess the impact bots are having within online conversations. In Section 1.4, I identify three distinct research efforts that contribute to the greater understanding of how human actors are adapting to bot actors within OSNs, while also attempting to characterize the adaptation of bots themselves. Specifically, this dissertation addresses the following research questions by analyzing harvested Twitter data from various online conversations enriched with classification data from multiple state-of-the-art bot detection services:

How does social bot pervasiveness and relative importance compare across different online social network conversations?

To what extent do different social bot detection platform classification results overlap in the identification of bots within the same online conversation?

1.2.2. Adaptation to Digital Censorship in Online Social Networks

We have witnessed the power of OSNs to enable individuals to communicate and exchange ideas globally in a near-instantaneous fashion, especially in times of social

unrest (e.g. Arab Spring, Ukraine Crisis). The amplification of unfettered individual opinion stands as a challenge to authoritarian governments that seeking to quell the spread of specific information through the implementation of censorship practices. As I highlight in greater detail in Section 1.3.2, certain governments and their citizens continue to evolve their participative behaviors in OSNs in an attempt to enforce or circumvent censorship practices. It is this adaptation to censorship in OSNs from both a government and citizen perspective that I seek to analyze using harvested OSN data during specified periods of government censorship. Specifically, I follow the initial methodology presented in Section 1.4.2 to answer the following research question in relation to an observable Turkish government censorship campaign initiated to block Turkish citizens from using the Twitter platform:

To what extent can an authoritarian government be effective in blocking its citizens from using an online social network during periods of social unrest?

I then extract learned outcomes from this initial digital censorship analysis to inform the creation of an agent-based model, as described in Section 1.4.2, to analyze adaptation to censorship via simulation experimentation in an effort to answer the final research question of the dissertation:

How do government entity digital censorship practices affect the decisions of individual citizens to continue participating in online social activities?

1.3. Background Literature of Interest

The following section provides pertinent background information on the two adaptation areas of focus this dissertation addresses. Although not intended to be a

comprehensive introduction to each topic, this section serves as a foundational piece for the more detailed literature reviews that accompanying each chapter within this dissertation. Section 1.3.1 provides an extensive introduction to the topic of social bots and ongoing works analyzing bot presence in online social systems. Section 1.3.2 provides background information on digital censorship practices observed in online social systems, along with a brief introduction to the few ABMs that seek to address some aspect of digital censorship or the concept of risky collective action emanating from online activities.

1.3.1. Online Social Bot Research

According to recent Internet security reports, industry experts estimate that 'bots' have consistently accounted for approximately half of all web traffic in the past five years (Zeifman, 2017). The term bot, however, has broad meaning in the context of technological applications, since all forms of automated services or applications could potentially be construed as bots. For the remainder of this dissertation, I restrict the definition of bots, or social bots, to automated software or computer algorithms designed to mimic human behavior and/or engage with human actors within online social systems.

Current social bot research primarily continues to focus on advancing initial bot detection techniques and classification efforts (e.g. Chavoshi & Mueen, 2018; Cresci et al., 2018; Stukal et al., 2017; Varol et al., 2017). While there is great need for detection techniques to keep pace with the rapidly evolving sophistication of social bots (Cresci et al., 2017), Abokhodair et al. (2015) observed that relatively simple bots are able to avoid detection by advanced bot detection efforts and operate freely, thus suggesting that social

bot detection methodologies cannot focus solely on sophisticated bots. Overall, there is ample opportunity for methodological improvement in bot detection research. The 2015 Defense Advanced Research Projects Agency (DARPA) Twitter Bot Challenge recognized the difficulty of detecting bots in OSNs early and highlighted the need for bot detection systems to be semi-supervised in order to account for the far-ranging types of bots (Subrahmanian et al., 2016).

In addition to simply publishing research describing bot detection methodologies and findings, some researchers have also transitioned their detection algorithms to open-source bot detection platforms for other researchers to use. Davis et al. (2016) provides researchers access to its bot detection framework Botometer (formerly known as BotOrNot) by allowing researchers to submit questionable Twitter author names for classification. Botometer, in turn, accesses the current Twitter profile and activity of a queried user account and assesses the likelihood of the account being a bot by using a supervised Random Forest approach. Chavoshi et al. (2016) developed and launched DeBot, which provides researchers open access to a platform that uses an unsupervised warped correlation model to detect bots as opposed to relying on feature extraction.

While detection algorithm research garners the most attention in the nascent field of social bot research, analysis dedicated to evaluating the prevalence of bots has risen recently. Social bot analyses have included examining bot evidence in the following OSN conversation use-cases: the 2016 U.S. presidential election (Bessi & Ferrara, 2016; Howard et al., 2018), the Brexit Referendum (Duh et al., 2018; Howard & Kollanyi, 2016), financial trading markets (Cresci et al., 2019), the ongoing Ukrainian conflict

(Hegelich & Janetzko, 2016) and vaccinations (Broniatowski et al., 2018). Most methodologies, however, are limited to simplistic descriptive statistical and temporal analyses of observed bot tweet volumes. Furthermore, these analyses typically focus on single OSN conversation use-cases resulting from the employment of a single bot detection platform. As Kušen and Strembeck (2018) points out, bot studies focused on sole events make it difficult to generalize findings across this growing research area of interest. A problematic example of social bot research not heeding this critique is the recent proclamation by Pew Research (Wojcik et al., 2018) claiming that social bots are responsible for posting two-thirds of all tweeted website links to popular websites. Such a broad, generalized headline statement presents a misleading finding as the Pew Research team relied upon a bot detection sampling method using estimated results from a single bot detection platform service (i.e. Botometer) (Wojcik et al., 2018).

1.3.2. Online Censorship Research

Historically, governments have suppressed political dialogue in media through the implementation of various forms of censorship (Briggs & Burke, 2009). Such practices are typically customized to the environments in which information flows within a given society (Esarey & Xiao, 2011). The advent of OSNs (e.g. Twitter, Facebook, Sina Weibo and VKontakte), fueled by the Web 2.0 revolution, has significantly impacted global events throughout the past decade. For example, OSNs have been a contributing catalyst for the major social unrest in northern Africa and the Middle East (i.e. Arab Spring) and Ukraine (i.e. Euromaidan) and have therefore become the target of certain censorship practices. Governments have had to adapt their censorship practices beyond the

traditional sources of print, radio and television (Fourie, Bothma, & Bitso, 2013; Nunziato, 2010) to account for this new information media that has been shown to enable rapid collective action and potential political unrest. Certain governments have developed a wide range of options to attempt to control Internet dialogue, ranging from simple messaging of appropriate online behavior discourse to sophisticated content monitoring and filtering as well as fully restricting access to the WWW (Clark et al., 2017; Dainotti et al., 2014). Shirky (2011) provides one of the earliest extensive reviews of authoritarian regimes and despotic governments viewing social media platforms as highly problematic and highlights initial attempts by governments to implement digital censorship controls.

The readily accessible nature of online social system data has afforded researchers the opportunity to evaluate Internet censorship to a fine level of granularity. As Meserve and Pemstein (2017) highlighted, even democracies have not been immune from government-level digital censorship when internal dissent became evident. Censorship is of course an incredibly broad field that covers a vast array of topics, but in the case of this dissertation I limit the focus to political censorship of social media by authoritarian governments.

Any discussion of current digital censorship practices must begin with the most expansive digital censorship campaign, the Great Firewall of China. The Chinese government has instituted extreme measures to restrict access to social media platforms via the Great Firewall and to maintain control of political narratives via a vast array of surveillance programs (King et al., 2017). In highlighting the flexible constraint of Chinese censorship practices, King et al. (2013) discovered that the Chinese government

surprisingly tolerated some level of disparaging social media remarks directed at the Chinese government but immediately filtered or blocked messages tied to collective action or mobilization of protest efforts. Adding to the evidence of dynamic censorship practices, Bamman et al. (2012), while evaluating 56 million Sina Weibo messages and 11 million Chinese language tweets, discovered non-uniform patterns of deletion based on message analyses from the provincial perspective. Further examples of Chinese explicit digital censorship include the specific restriction of citizen access to a litany of web services (e.g. Google, Facebook, Twitter and YouTube) (Bamman et al., 2012; Xu & Albert, 2014).

Digital censorship exists in various forms globally and past examples include Azerbaijan (Pearce & Kendzior, 2012), Ukraine (Metzger & Tucker, 2017) and Turkey (Tanash et al., 2015). Metzger and Tucker (2017), while analyzing social media during the Ukrainian Euromaidan protests, concluded that the 2013 Ukrainian government's censorship attempts to suppress social mobilization were ultimately unable to control social media messaging content. Tanash et al. (2015) conducted the first substantive evaluation of social media censorship in Turkey, finding evidence that the censorship rate of tweets in Turkey was at least two orders of magnitude higher than Twitter's own transparency report. In a follow-up study, Tanash et al. (2017) presented a heuristic for observing self-censorship rates of Turkish twitter users immediately following the failed Turkish coup attempt in July 2016.

Evidence has shown that attempting to circumvent or ignore digital censorship policies can lead to a wide variety of punishments. Surveying the evolving nature of

global censorship, Clark et al. (2017) detailed government-imposed penalties by the Turkish government on citizens breaching government policy via social media activity. The study highlighted that more than 1,600 social media-related arrests took place following the attempted Turkish coup in the summer of 2016 and that a further 10,000 other citizens were actively investigated for their online activities. Yesil and Sözeri (2017) drew further attention to the legal ramifications associated with bypassing social media policies by providing an expansive analysis of the evolution of Turkish legal policies that enable OSN censorship.

Sometimes digital censorship practices do not produce the effects anticipated by government entities. Nabi (2014) discovered that censorship practices were not only ineffective in restricting access to specific OSN content in Syria and Turkey, but that the attempted censorship ultimately produced the unintended effect of popularizing the very topics those governments were trying to censor. Through the analysis of Alexa, Google Trends, and YouTube statistics data, Nabi (2014) dubbed the ineffectiveness of these state-level censorship activities as the 'Streisand Effect.' Katz (2014) further warned that attempting to censor or restrict access to social media, while not the sole reason, can serve as a primary enabling factor behind social movement causes that lead to the potential mobilization of citizens into action. Consequently, social adaptation to digital censorship practices can be viewed as a collective action problem (Katz, 2014).

The adaptive decision-making of citizens of heterogeneous populations attempting to circumvent the digital censorship practices and governments correspondingly trying to detect and prevent such censorship avoidance represents a

complex social situation. As such, it is an ideal subject for agent-based modeling. Few ABMs in the current literature focus specifically on digital censorship. Casilli and Tubaro (2012) developed an ABM that is an extension of Epstein's (2002) model of civil violence to simulate the effects of online social media censorship during the 2011 London Riots and to determine the extent to which the censorship would propel agent populations to physical acts of violence. Waldherr and Wijermans (2017) reviewed the Casilli and Tubaro (2012) ABM as the only current effort dedicated to online censorship, while presenting their own ABM design to model street protests that are informed or influenced by social media engagements. In an extension of the Granovetter (1978) threshold model, Funcke and Franke (2016) viewed online social network participation as an initial participation cue that could lead to eventual acts of greater physical consequence. While there are few models dedicated explicitly to social media interactions, we can look to more traditional ABMs to draw inspiration in modeling interactions of social conflict that are similar to censorship or its effects on a system. Lemos et al. (2013) provides an extensive review of ABMs focused on social conflict topics, to include collective action activities of civil disobedience, riots and revolution. The overarching theme among these mentioned models is that they focus exclusively on collective action events emerging in physical environments and not cyber environments.

1.4. Structure of Dissertation

This dissertation presents each chapter beyond the introduction (Chapter 1) and conclusion (Chapter 8) as stand-alone research papers. Each chapter can be classified as a published paper, awaiting publication decision or pending publication submission. The

chapters are grouped according to adaptation topic as shown in Figure 2, with Chapters 2-5 presenting social bot analysis research and Chapters 6 and 7 focusing on digital censorship. The following two sections provide brief introductions to the chapters comprising each adaptation topic.

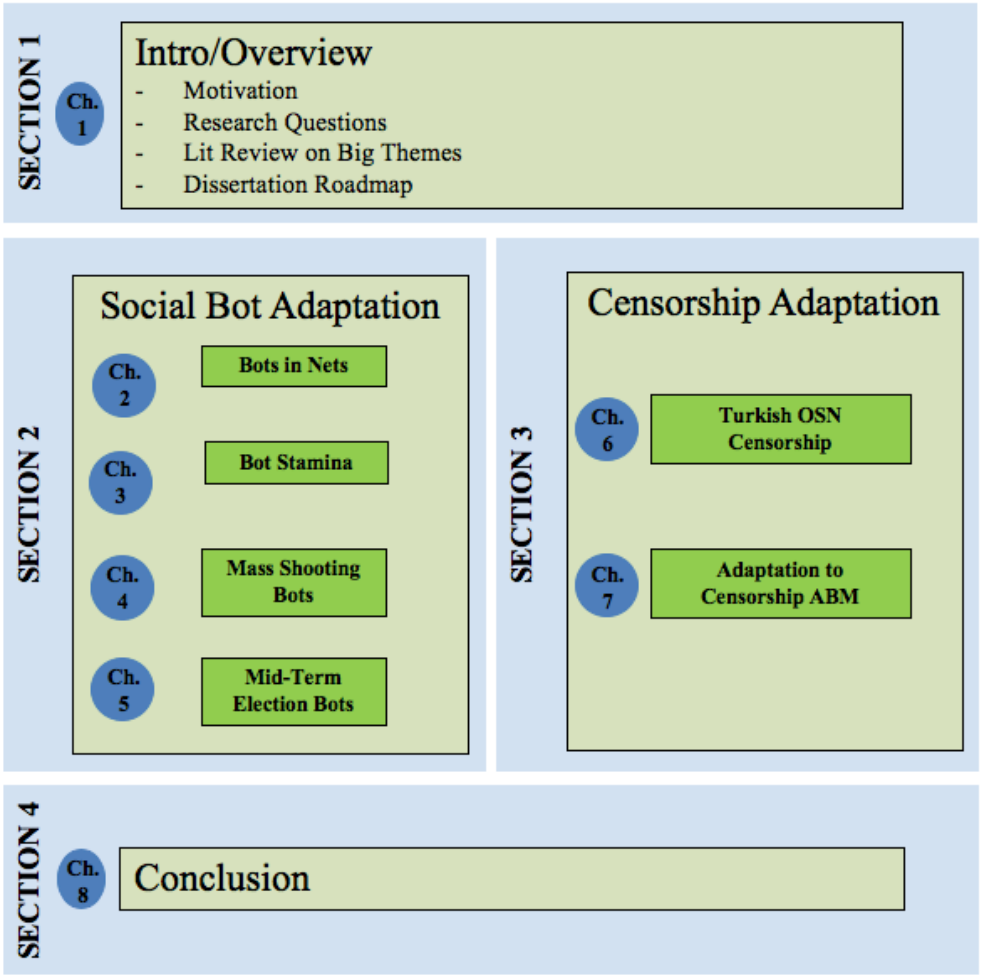


Figure 2: Research components comprising chapters of this dissertation.

1.4.1. Adaptation to Social Bot Actors in Online Social Networks

As discussed in Section 1.2.1, online social bots have emerged as observable actors within online social networks (OSNs), and their potential to diffuse information at scale and to influence opinion has given rise to many efforts to detect them. I view the growing emergence of social bots from an adaptation perspective in the sense that bots are adaptive actor themselves (i.e. they are continually evolving to detection methods), while regular human actors have to adapt to non-human bot actors within these systems. While methodologies employed to detect the evolving sophistication of bots continue to improve, much work can be done to characterize the impact of bots within OSNs and to potentially increase overall detection efforts. In the remainder of this section, I introduce three distinct use-cases presented in this dissertation that determine the pervasiveness, relative importance and effect of social bots in various OSN conversations. These use-cases address the research questions posed in Section 1.2.1.

Chapter 2 and Chapter 3 comparatively analyze bot evidence from the same Twitter corpus comprised of more than 30 million tweets stemming from three major global events in 2016 (the U.S. Presidential Election, the Ukrainian Conflict and Turkish Political Censorship). Chapter 2 serves as the initial social bot analysis research effort. Here I compare the conversational patterns of bots and humans within each event and examine the social network structure of each conversation to identify social bots exhibiting particular network influence, while also determining bot participation in key emergent network communities. Chapter 3 is an extension of Chapter 2 that introduces additional analysis techniques. The subsequent bot-related chapters build upon the initial

social bot analysis framework, but through the lens of different OSN conversation use-cases.

Chapter 4 presents the second bot use-case, which focuses on OSN conversations surrounding mass shootings in the United States. Specifically, this effort analyzes bot evidence in the Twitter conversations for the following five mass shooting events: the Santa Fe High School shooting (Santa Fe, Texas - May 2018), the Parkland High School shooting (Parkland, Florida - February 2018), the First Baptist Church in Sutherland Springs shooting (Sutherland Springs, Texas - November 2017), and the Route 91 Music Festival shooting (Las Vegas, Nevada - October 2017).

While the previous bot use-cases rely upon a sole bot detection platform to identify social bots, the final bot use-case presented in Chapter 5 presents an ensemble bot detection framework using three different platforms to identify bots within the 2018 U.S. midterm election OSN conversation. This is the first known effort to simultaneously use three separate bot detection services within the same study on near real-time data. Chapters 2-4 rely upon a sole bot detection platform for bot classification due to the historical nature of the observed OSN conversation data in those chapters.

1.4.2. Adaptation to Digital Censorship

As discussed in Section 1.2.2, the enabling features and increased usage of online social systems in times of social unrest (e.g. Arab Spring, Ukraine Crisis) have led to actors at both the individual and government levels to adapt their participative behaviors in social media networks. This includes the initiation of censorship practices by governments to restrict communication and the self-censoring of individuals out of fear

of reprisal from these governments. The study presented in Chapter 6 analyzes an online censorship use-case that eventually helps inform development of the adaptation to censorship ABM introduced in Chapter 7.

Chapter 6 comparatively analyzes political social media dialogue from Twitter prior to and during a period of extreme dynamic censorship in an effort to determine the effect of dynamically-changing digital censorship policies on OSN participation. Historically, digital censorship studies focus on censorship practices such as the surveillance of Internet traffic patterns (Dainotti et al. 2014; Florio et al. 2014), or specific content filtering and/or content removal (Bamman, O'Connor, and Smith 2012; Meserve and Pemstein 2017; Parks et al., 2017; Tanash et al. 2017, 2015; Zhu et al. 2013). This chapter, however, examines a censorship campaign to completely block access to an entire OSN platform. Through the application of social network analysis-based framework, the chapter analyzes Turkish political online social media conversations harvested from Twitter in December 2016 when the Turkish government abruptly blocked access to Twitter twice in a one-week period. The analysis results evaluate the effectiveness of the Turkish government's censorship implementation by identifying observable social network artifacts at the regional and global level of the OSN conversation. The results provide direct insights into the research question posed in Section 1.2.2.

Chapter 7 serves as the final research contribution of this dissertation with the introduction of the adaptation to censorship ABM. Drawing upon the models reviewed in Section 1.3.2 and social psychology theory, Chapter 7 presents a novel model that looks

not only at how the social actors adapt to one another in a given locale but also at the finite technical adaptation components each actor can use. This includes technical cyber options such as an individual's use of obfuscation technologies and a government's ability to detect or deter such technologies.

CHAPTER 2. BOTS IN NETS: EMPIRICAL COMPARATIVE ANALYSIS OF BOT EVIDENCE IN SOCIAL NETWORKS[2]

2.1. Introduction

The increased dependency on online social networks (OSNs) for information and the unprecedented ability to instantaneously message global populations provides an opportunity to control or exploit the narrative of online conversations. Attempting to control or exploit the narrative of a certain topic becomes much easier in OSNs as 'digital gatekeepers' can employ social bots—computer algorithms designed to mimic human behavior and interact with humans in an automated fashion—to amplify a specific position or drown out its opposition at scale. This includes increasing the spread of fake news by orders of magnitude through a directed bot campaign (Lazer et al., 2018). The evolvement of social bot sophistication is a primary concern, as it has become very hard for humans to discern whether they are engaging in dialogue with a human or a bot (Ferrara et al., 2016). Given that recent studies estimate that social bots account for 9-15% of all Twitter accounts (Subrahmanian et al., 2016; Varol et al., 2017), it is essential to understand the implications associated with human and machine dialogue, either intentional or not.

[2] This chapter was published in: Schuchard, R., Crooks, A., Stefanidis, A., & Croitoru, A. (2019). Bots in Nets: Empirical Comparative Analysis of Bot Evidence in Social Networks. In L. M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, & L. M. Rocha (Eds.), *Complex Networks and Their Applications VII* (pp. 424–436). Springer International Publishing.

Recent social bot research continues to build initial essential knowledge on the classification and detection of social bots (Chavoshi et al., 2016; Chu et al., 2012; Davis et al., 2016; Stukal et al., 2017; Varol et al., 2017). However, the establishment of social bot norms is difficult and predictively elusive given the evolving nature of bot sophistication. For this reason, studies continue to discover bot activity that does not align with previously published conceptions (Abokhodair et al., 2015). Beyond the necessary continued work associated with improved bot detection methods to move closer to ground truth discovery, there is also a growing need to present novel evaluation methodologies to better understand the effects of currently detected bots within social media conversations. Promising recent studies applying multidisciplinary approaches to social bot analysis include classifying bot emotion (Kušen & Strembeck, 2018), determining the political agenda of bots (Hegelich & Janetzko, 2016) and distorting political discourse with bots (Bessi & Ferrara, 2016; Forelle et al., 2015; Howard & Kollanyi, 2016).

This chapter presents a unique methodological framework to comparatively analyze evidence of social bots found within OSN Twitter conversations about three major global events in 2016: (1) the United States Presidential Election, (2) the Ukraine Conflict and (3) Turkish Online Political Censorship. First, a comparative descriptive statistical analysis (Section 2.4.1) of these Twitter conversations determines the characteristics of human and social bot tweeting patterns. Next, applied social network analysis techniques sought to determine the relative influence of social bots within each of the associated conversation's constructed retweet networks (Sections 2.4.2 - 2.4.4). In

total, this study evaluated more than 30.4 million tweets generated by 5.2 million distinct Twitter users, of which, bot enrichment processing recognized 14,661 users as bots responsible for 2.1 million tweets.

The results of this study showed that social bot communication patterns were fairly consistent across the various observed online conversations. Furthermore, bots were found to have a higher engagement rate than humans for both in-group and cross-group communication. Most interestingly, although online conversation participants recognized as social bots comprised only 0.28% of all OSN users observed in this chapter, they accounted for a significantly large portion of prominent centrality rankings across the three online conversations. In total, this work provides a new contribution to the growing study of social bots by applying social network analysis techniques across multiple online conversations to help determine the relative pervasiveness and importance of detected bots.

2.2. Related Work

The term bot has broad meaning in the context of technology and Internet applications, since all automated services or applications could be construed as bots. For the purpose of this chapter, we restrict the definition of bots, or social bots, to automated software or computer algorithms designed to mimic human behavior and/or engage with human actors within online social networks. Many recent works have contributed to the growing corpus of knowledge capturing social bot features that differentiate social bot-generated activity from human-generated activity in OSNs (Boshmaf et al., 2013; Chu et al., 2012; Davis et al., 2016).

Some researchers have not only published their research on bot detection methodologies and findings but have also transitioned their work to open-source bot detection platforms for other researchers to use via a web application or an application programming interface (API). Davis et al. (2016) provide access to Botometer (formerly known as BotOrNot), which assesses the likelihood of a Twitter account being a bot by using a supervised Random Forest applied to extracted account features. Chavoshi et al. (2016) published DeBot, which employs an unsupervised warped correlation model to detect Twitter bots rather than feature extraction.

Published research analyzing detected bots in specific OSNs has increased as the prevalence of bots has risen. Such studies include examining bot evidence in the following use-cases: the 2016 U.S. presidential election (Howard et al., 2018; Varol et al., 2017), Venezuelan political public opinion (Forelle et al., 2015), the Syrian civil war (Abokhodair et al., 2015), the Brexit Referendum (Howard & Kollanyi, 2016), the Ukrainian conflict (Hegelich & Janetzko, 2016; Zhdanova & Orlova, 2017) and Russian politics (Stukal et al., 2017). Most methodologies are limited to initial descriptive statistical and temporal analyses of the human versus bot tweet volumes. Although highly relevant contributions, these efforts focus on single events. As Kušen and Strembeck (2018) point out in their recent analysis of bot emotion across multiple events, bot studies focused on sole events make it difficult to generalize findings across this growing topic of interest.

2.3. Methodology

In order to understand the patterns of bots across multiple global events and determine the relative bot impact within associated online conversations, this study employed a combination of comparative descriptive statistical analysis and social network analysis applications. This multi-faceted approach expands the literature of social bot analysis by comparatively analyzing multiple OSN use-cases and contributes new techniques to the field of bot research by adapting social network analysis methods to measure and define the impact or influence of social bots. The remainder of this section presents in detail the methodology steps used in this study as depicted in Figure 3.

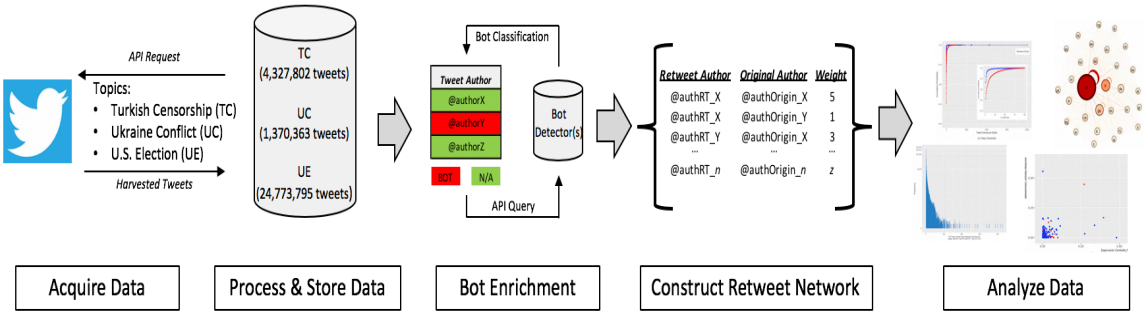


Figure 3: Overall methodology to analyze bot evidence across multiple Twitter OSN conversations.

2.3.1. Data

This study presented in this chapter focused on three major global online conversations harvested solely from Twitter in 2016. Summarized descriptions of each event conversation are as follows: (1) U.S. Presidential Election (Feb. 1-29, 2016): a one-month period which captured the narrative surrounding the Republican and Democratic

party primary races prior to the U.S. general election when it became apparent that then-candidate Donald Trump could win his party's nomination, (2) Ukraine Conflict (Aug. 1-31, 2016): a one-month period which captured the narrative surrounding the ongoing conflict in Ukraine as military activity and political rhetoric intensified between Russia and Ukraine around the 25th anniversary of Ukrainian independence from Russia, (3) Turkish Political Censorship (Dec. 1-31, 2016): a one-month period which captured Turkish political conversations before, during and after two distinct periods of censorship when the Turkish government banned Turkish citizens from using Twitter.

Relevant keywords representing each of these events were crafted and submitted to extract associated tweets from the Twitter Standard Search API. The volumes of tweets returned were as follows: 24.8 million (U.S. Presidential Election), 1.4 million (Ukraine Conflict), 4.3 million (Turkish Censorship). Given the resulting large tweet volumes, all initial data storage and pre-processing took place in an Amazon Web Services EC2 t2.2xlarge instance (8 vCPUs/32GiB). This allowed for rapid processing and the creation of individual graph objects for more rapid data analysis use at the local compute level.

2.3.2. Bot Enrichment

To determine the presence of bots within the acquired Twitter conversations, this study leveraged the DeBot open-source bot detection platform (Chavoshi et al., 2016). The decision to use DeBot was two-fold. First, the corpus of tweets came from 2016, so access to historical bot evidence was a requirement, which only DeBot currently provides. Second, the performance of DeBot's unsupervised warped correlation process has outperformed other bot detection platforms to date (Chavoshi et al., 2017). To

determine bot presence, automated data processing procedures extracted tweet author names from the harvested tweet corpus and submitted them for classification via the DeBot API. The classification results were then merged with the existing database and labeled each tweet user as a bot (or not) and annotated the source of bot classification. This study produced and archived automated scripts to execute this enrichment phase with the hope of accounting for other bot detection services in the future.

In total, this enrichment process classified 14,661 Twitter users as bots, which accounted for just 0.28% of total tweet corpus users. This relatively small population of users classified as bots was responsible for publishing 2.1 million tweets, or 6.8% of all tweets in this study. Table 1 provides detailed values for each event conversation.

2.3.3. Construct Retweet Network

Retweets accounted for 57.8% of all tweets in this study, with the Turkey Censorship conversation exhibiting the highest retweet density at 65.6%, followed by 57.8% for the U.S. Election conversation and 49.8% for the Ukraine Conflict conversation. The parsed retweets from the originally harvested tweets served as the basis for the construction of retweet networks for each conversation. These resulting retweet networks serve as the primary artifacts required to examine the conversation via social network applications that include centrality analysis and community detection.

To reveal the network structure from the harvested Twitter conversations, the study relied upon the constructed retweet networks for each of the events. The act of a Twitter user ‘retweeting’ a message of an originally authored tweet establishes the basis for an edge between two nodes, or users, in the retweet network. Specifically, when a

Twitter user (*X*) retweets an original tweet message from a given user (*Y*), one can then assign a directed edge weight value of 1 for initial retweets or add to the cumulative weight for existing edges. The resulting directed networks for each of the conversations were as follows: 2,557,805 nodes / 8,985,736 edges (U.S. Election), 250,541 nodes / 537,459 edges (Ukraine Conflict), 1,075,833 nodes / 2,224,939 edges (Turkish Censorship).

Table 1: Harvested Twitter corpus overview

Corpus	Tweets	Retweets	Users
United States Election	24,773,795	14,321,387	3,472,114
Bot Source *(% of total)*	1,882,809 *(7.60%)*	1,452,155 *(10.14%)*	6,875 *(0.20%)*
Ukraine Conflict	1,370,363	681,806	383,237
Bot Source *(% of total)*	55,718 *(4.07%)*	34,938 *(5.12%)*	2,486 *(0.65%)*
Turkey Censorship	4,327,802	2,837,059	1,390,362
Bot Source *(% of total)*	126,352 *(2.92%)*	83,582 *(2.95%)*	5,300 *(0.38%)*

2.3.4. Analyzed Data

The final phase of this study's methodology was the application of a multi-faceted data analysis approach to the processed data from the three online conversations. Recall that the main purpose of this work was to identify potential common characteristics of social bots across multiple online conversations and ascertain any in-group (bot-to-bot) or cross-group (bot-to-human/human-to-bot) tendencies. Additionally, this study sought to classify the overall relative importance of bots within the conversations by examining bot positions within the social structure of the retweet networks and associated bot

membership within any emergent communities of said networks. Section 2.4 follows with detailed subsections discussing the specific methods used to achieve the purpose described above.

2.4. Results and Discussion

2.4.1. Bot and Human Participation Rates

Cumulative distribution frequency (CDF) plots depicting tweet volume per author for each of the online conversations served as visual evidence to directly compare the conversation participation rates between bot and human authors. The resulting CDFs serve as comparative artifacts between the author types and the various conversations. In addition, a two-sample Kolmogorov-Smirnov (KS) test returned a D statistic metric to capture the absolute max distance between the bot and human distributions for each of the conversations.

The CDFs, depicted in Figure 4, show similar general participation rate trends for both bots and humans across all conversations. The resulting distributions all exhibit a 'many-some-few' fat-tail distribution, with most of the authors having extremely low tweet volume (i.e. fewer than 10 tweets), some authors with higher tweet volumes (i.e. $10<x<1000$) and very few authors with high tweet volumes (i.e. $x = 1000+$). Additionally, the results showed that human authors account for the largest tweet volumes per author across all conversations and have a higher concentration of low volume authors accounting for all tweet volumes.

The KS test results between bot and human authors highlight the major difference in low tweet volume authors accounting for much larger portions of the entire tweet

conversation by humans. The conversations returned D statistic values of 0.529, 0.408, and 0.419 for the U.S. Election, the Ukraine Conflict and the Turkish Censorship conversations, respectively. These maximum values were all observed where the tweet volume per author was a single tweet as shown in each plot's associated inset zoom.

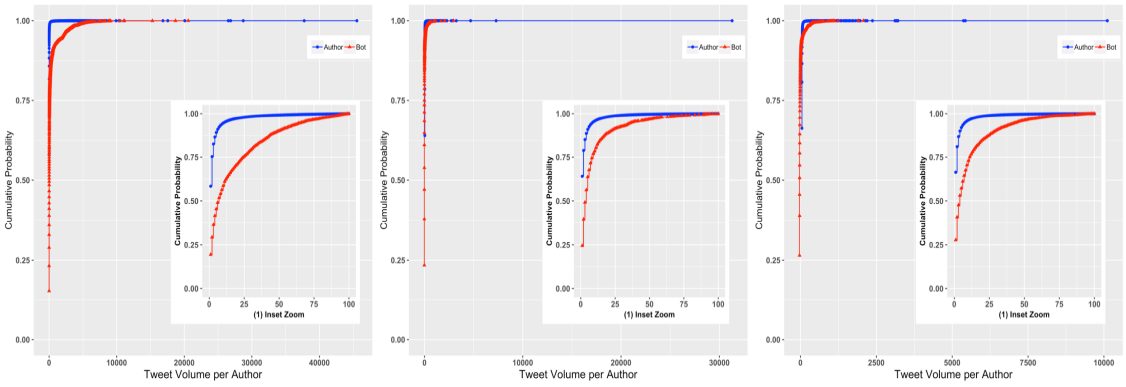
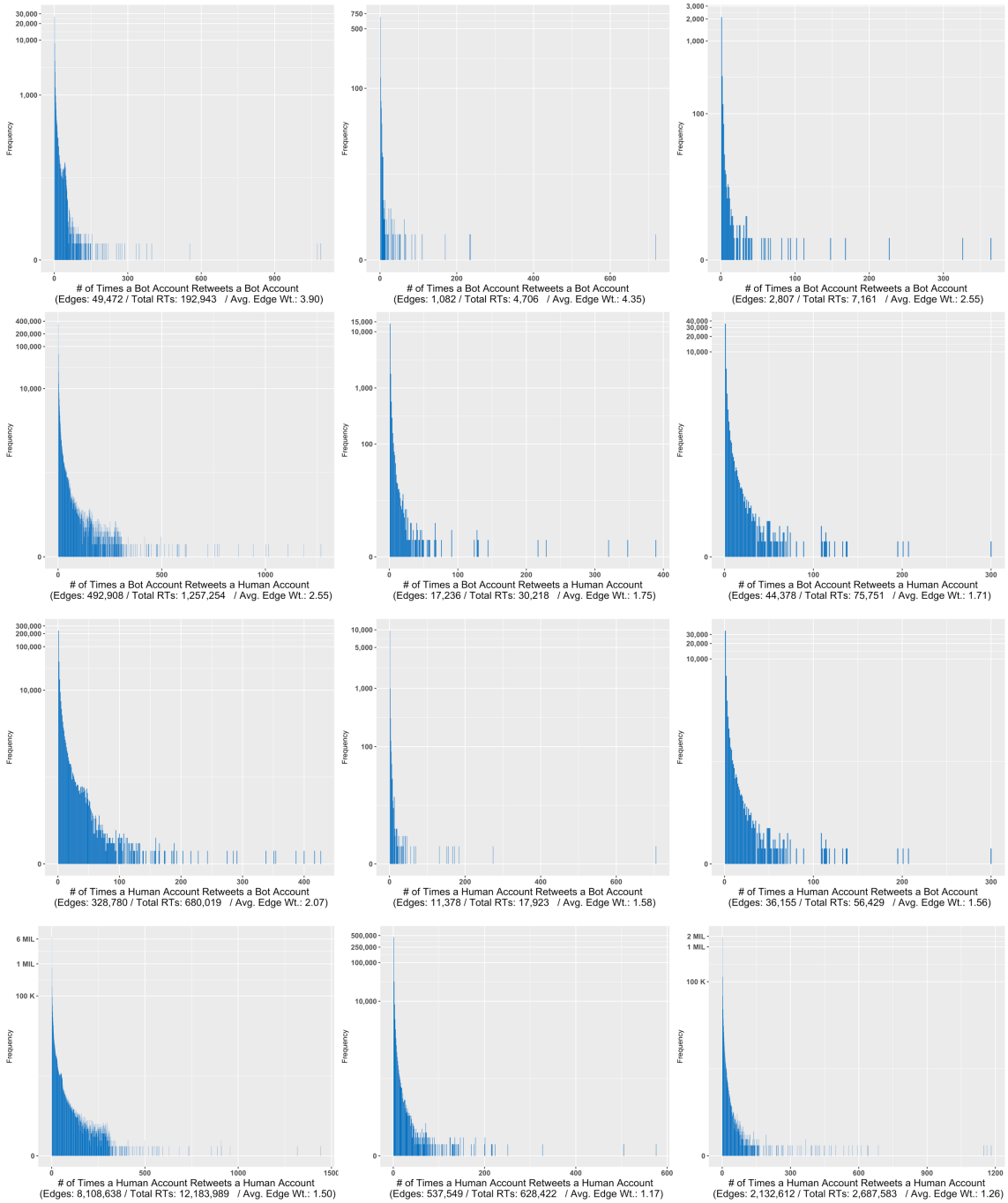


Figure 4: Cumulative distribution (CDF) plots of tweet volume per human (blue) and bot (red) for each online conversation: (a) U.S. Election, (b) Ukraine Conflict and (c) Turkish Censorship. Inset zooms provide granularity to capture the high density of authors with low tweet volumes.

2.4.2. In-Group and Cross-Group Communications

Figure 5 presents a consolidation of all in-group and cross-group communication frequencies observed in this study. This work defines in-group communication as retweet edges between like types of authors (i.e. bots retweeting bots or humans retweeting humans), while cross-group communication refers to retweets between different types of authors (i.e. bots retweeting humans or humans retweeting bots). While low retweet volumes appear to dominate for in-group and cross-group conversations across all of the online conversations, there exists a noticeable increase in retweet rates for all

conversations initiated by a bot author, as opposed to a human author. For all three online conversations, each bot-to-bot in-group and bot-to-human cross-group conversation has a relatively higher average edge weight. The bot-to-bot author average edge weight is 160%, 272% and 102% higher than the human-to-human author average edge weight for the U.S. Election, the Ukrainian Conflict and Turkish Censorship, respectively. This suggests that either bots seek persistent contact more so than humans, or the high rate of single retweet volumes between so many different human edges dilutes any persistent human-to-human connections that exist.



(a) U.S. Election (b) Ukraine Conflict (c) Turkish Censorship

Figure 5: Frequency distribution plots for (a) U.S. Election, (b) Ukraine Conflict and (c) Turkish Censorship retweets of in-group bot conversations (row 1), cross-group bot and human conversations (rows 2 and 3) and in-group human conversations (row 4)

2.4.3. Centrality Analysis

In social network analysis, centrality measurements allow for us to distinguish nodes in a network as more prominent, or important, than other nodes based on their relative position in the structure of the network (Wasserman & Faust, 1994). This study sought to classify the overall relative importance of bots within the online conversations of interest by using centrality measures. To do so, three relatively common centrality measures (degree, eigenvector, and betweenness) were calculated for each online conversation. Degree centrality is the most straightforward centrality, as it is calculated from the total number of direct connections a node shares with other nodes throughout the network. One could view degree centrality as a level of popularity in a network. Eigenvector centrality is a weighted sum of both direct and indirect connections for a given node that is based on the individual degree centrality score of each node with which it shares an edge (Bonacich, 2007). Thus, we can infer eigenvector centrality as a level of entire network influence. Betweenness centrality is the degree to which a node falls on the shortest path between other nodes in the network (Freeman, 1977). Therefore, we can characterize betweenness as a potential measure of information flow in a network.

The consolidated results for the three centrality measure calculations across all three conversations are presented in Figure 6. The binned the results capture the density of bots falling within the top-N centrality valuations (*where*, *N= 1000, 100, 50 or 10*). Of note, Figure 6 provides both the raw number of bots and the total percentage of bots comprising the given population of top-N centrality values. The results clearly show that authors identified as bots, though they comprise just 0.28% of total conversation authors

in this study, account for a significantly large portion of prominent centrality rankings for each of the centrality measures across all conversations. Showing penetration into conversations as an influencer, the eigenvector valuations show that bots account for 43% of the top-100 nodes in the U.S. Election conversation, to include four of the top-10 centrality value positions. In the Ukraine Conflict dialogue, bots show a gaining dominance of top eigenvector values, as the bot population accounts for 21%, 30% and 50% at the top-100, top-50 and top-10 bins respectively.

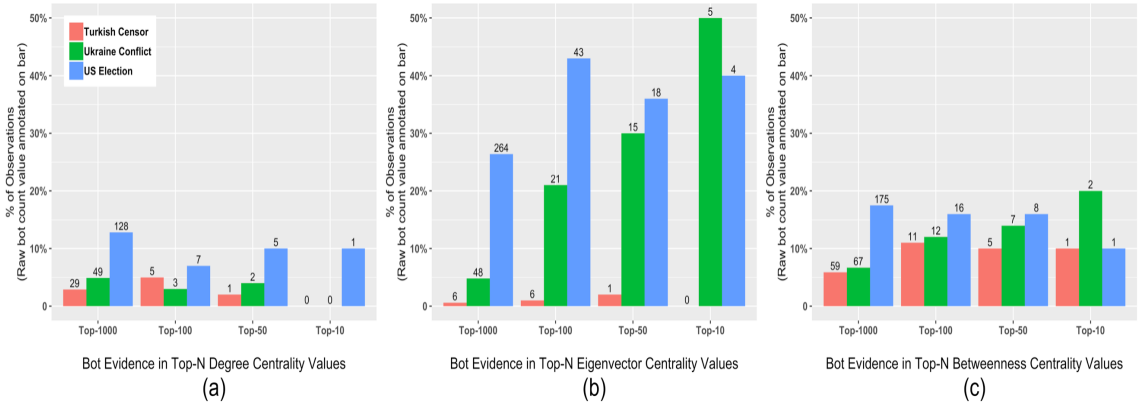
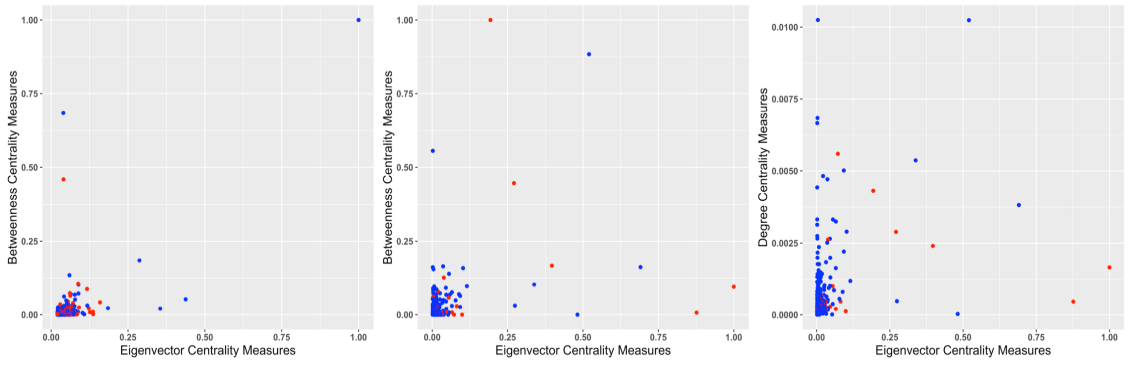


Figure 6: Bot evidence in top-N (*N= 1000 / 100 / 50 / 10*) [(a) degree (b) eigenvector (c) betweenness] centrality values for: U.S. Election (blue), Ukraine Conflict (green) and Turkish Censorship (red)

Many studies point to the positive correlation of computed centrality values given the conceptual overlap that exists between the inputs required of the calculations (Valente et al., 2008). Given an expected correlation of centrality values, lack of correlation evidence provides an opportunity to further investigate a node for interesting behavior. This study conducted such an analysis by plotting correlation plots against each other as depicted in Figure 7.



(a) U.S. Election (Top-1000 accounts) (b) Ukraine Conflict (Top-1000 accounts) (c) Ukraine Conflict (Top-1000 accounts)

Figure 7: Correlation of centrality measures for select centrality comparison: (a) U.S. Election eigenvector versus betweenness analysis, (b) Ukraine Conflict eigenvector versus betweenness analysis and (c) Ukraine Conflict eigenvector versus degree analysis.

The depicted centrality correlation plots in Figure 7 provide compelling insights into some of the observed conversations. First, in the U.S. Election conversation plot (Figure 7a), we see very few correlation outliers on the plot. Interestingly, the top eigenvector and betweenness centrality node is the same human author, in this case, then-candidate Donald Trump (@realDonaldTrump). Conversely, we see far more correlation outliers in the Ukraine conflict conversations. Specifically, the most divergent nodes are bots, which could be cause for greater investigation as to their specific tweeting behavior. In the eigenvector versus degree Ukraine plot (Figure 7c), the two most 'influential' nodes according to eigenvector centrality, which are bots, are actually not that popular given low degree centralities. This suggests these bots were able to infiltrate the conversation network by acquiring connections with popular nodes, while avoiding popularity, or detection, themselves.

2.4.4. Community Detection

Community detection is another common application in social network analysis that allows researchers to uncover localized sub-graphs, or communities, of highly connected nodes that are otherwise less connected to the remainder of the network (Girvan & Newman, 2002). The Louvain (Blondel et al., 2008) method is one such community detection algorithm that is highly applicable for the identification of emergent community structure in large-scale network analyses. It seeks an undefined number of emergent communities by executing a two-stage greedy heuristic that iteratively optimizes modularity locally and culminates when global network modularity reaches a maximum value. For the purposes of this study, the resulting analysis sought to observe the density of bots within any defined community structure of the online conversations. Specifically, it was desirous to determine if bots clustered among themselves or if they dispersed among the larger human author communities, which would provide further explanation for the in-group and cross-group communication findings in Section 2.4.2.

Table 2 outlines the evidence of bot density within the most populated emergent communities detected for each online conversation. In total, the community bot density analysis discovered 71.2% of all bots within the top-5 most populated communities for the U.S. Election conversation, with 75.9% and 53.1% for the Ukraine Conflict and Turkish censorship conversations, respectively. Although we see a dispersal of bot populations throughout all of the top communities, there are multiple instances in which the bot density is much greater than the community population percentage in relation to the total network population. This is representative of the higher in-group communication

rates found between bots in Section 2.4.2, while the general dispersal of bots supports the findings of cross-group communication evidence.

Table 2: Bot density of largest emergent communities.

Comm.	U.S. Election		Ukraine Conflict		Turkish Censorship	
	Bot Count *(% of comm.)*	Comm. Size *(network %)*	Bot Count *(% of comm.)*	Comm. Size *(network %)*	Bot Count *(% of comm.)*	Comm. Size *(network %)*
1	901 (15.69%	1,009,872 (39.48%)	454 (21.25%)	58,397 (23.30%)	787 (16.39%)	268,311 (24.94%)
2	1,305 (22.73%)	900,076 (35.19%)	166 (7.78%)	45,330 (18.09%)	277 (5.77%)	146,350 (13.60%)
3	1,345 (23.43%)	308,040 (12.04%)	267 (12.50%)	29,310 (11.69%)	1,172 (24.40%)	107,224 (9.97%)
4	337 (5.87%)	84,733 (3.31%)	12 (0.06%)	15,536 (6.20%)	287 (5.98%)	85,550 (8.04%)
5	242 (4.20%)	59,441 (2.32%)	616 (28.84%)	15,439 (6.16%)	27 (0.56%)	48,813 (4.54%)

2.5. Conclusion and Future Work

In summary, this chapter presented a framework to characterize the pervasiveness and relative importance of bots in various OSN conversations of three significant global events in 2016. In total, over 30 million harvested tweets capturing the U.S. Presidential Election, the Ukrainian Conflict and Turkish Political Censorship OSN conversations served as the foundational data to compare the conversational patterns of bots and humans within each event. The study further examined the social network structure of each online conversation to determine if bots exhibited particular influence in a network, while also determining bot participation in key emergent network community subgraphs. The results showed that although Twitter participants identified as social bots comprised only 0.28% of all OSN users in this study, they accounted for a significantly large portion of prominent centrality rankings across the three conversations. This includes the

identification of individual bots as top-10 influencer nodes out of a total corpus consisting of more than 2.8 million nodes. Additionally, observed results showed that the most influential social bots had relatively low popularity, or degree centrality, suggesting influence can be obtained without popularity. In the case of social bots, popularity could be seen as a negative characteristic if trying to avoid detection. This finding is supported by previous findings in social media studies showing influence in a network is not necessarily driven by popularity (Cha et al., 2010).

While this chapter contributes to the nascent literature of social bot analysis by introducing a comparative analysis framework based on social network analysis techniques, there are limitations to take into consideration. As Tufekci (2014) asserts, social media analyses must state their limitations in terms of validity and representativeness when attempting to account for issues such as the overemphasis of single platforms and sampling biases. These issues are not unique to the study presented in this chapter, however, relied on just one OSN platform (i.e. Twitter) that includes a sampling bias. Though the methodology presented is not bound to a particular social media platform type, this analysis was limited to currently available bot detection sources, which focus solely on Twitter. As the literature expands in the near future, this effort can hopefully expand to account not only for additional bot detection services using Twitter, but additional social media platform sources as well. Specifically, future efforts could seek to determine if the findings produced here hold with other bot detection algorithms. Further extensions of this initial work will closely examine any observable characteristics differentiating the emergent communities of interests. This will include

narrative analysis through natural language processing to determine any attempts by bots to polarize particular populations within the conversations. The results from such an analysis could increase the relevancy of this study by potentially extending the observable influence of social bots beyond online social networks and into other social activities.

CHAPTER 3. BOT PERSISTENCE

3.1. Introduction

The previous chapter (Chapter 2) presented a study that focused on developing an initial framework to characterize the pervasiveness and relative importance of social bots in OSNs. This current chapter extends this earlier work by providing a more robust contribution along three lines of effort. First, this chapter provides a more extensive centrality analysis in Section 3.4.2 by including additional centrality measures that are specific to complex communicative networks. Second, temporal centrality rank persistence results are presented for each online conversation to determine the relative staying power of certain social bots over time (Section 3.4.2). Finally, the evolution of ego networks for the most structurally relevant bots over time is conducted in an effort to better characterize the user types communicating with social bots (Section 3.4.3).

As online social network (OSN) platforms (e.g. Twitter, Instagram, Sina Weibo) continue to attract dramatic global participation in terms of active user rates, they are becoming indispensable components of the online ecosystem (Blackwell et al., 2017). In the same sense that Fuchs (2005) describes the Internet as a socio-technological system, user devotion to OSNs has led to usage patterns that transcend simple messaging activities among networks of friends. In the United States (U.S.), OSN platforms recently surpassed print newspapers as a primary source for news, and they continue to gain

traction in relation to other traditional news sources such as television and radio (Mitchell, 2018). While the convenience of receiving 'news' within a multipurpose communication system is understandable, the sharing of real-world news in a social interaction environment may lead to unintended consequences. Sunstein (2018) suggests that the homophily-driven nature of OSNs results in the formation of echo chambers, which serve as fertile ground for the amplification of perpetuated false information, or fake news, among their members.

Recent studies have pointed to evidence of fake news within OSN conversations (e.g. Grinberg et al., 2019; Lazer et al., 2018). Furthermore, while examining news stories within Twitter from 2006 to 2017, Vosoughi et al. (2018) discovered that false stories spread more rapidly and to a greater audience than true stories. In addition to struggling to decipher the veracity of news, OSNs also have trouble accounting for the veracity of user accounts. This is largely due to the proliferation of accounts belonging to social bots, which are computer algorithms designed to mimic human behavior and interact with humans in an automated fashion. While automated in nature, social bots are not universally designed for intentional malice, as many bots serve in benign or even helpful roles (e.g. news aggregator) (Ferrara et al., 2016). The increasing sophistication of bots has made it difficult for human users to discern fellow human users from social bots in OSNs (Ferrara et al., 2016; Ruths & Pfeffer, 2014). While Vosoughi et al. (2018) argued that social bots were responsible for spreading both false and true news at the same rates as humans, Shao et al. (2018) discovered that social bots amplified news stories from low-credible sources in a disproportionate fashion. Although such studies

have demonstrated the strong presence of social bots in OSNs, the full extent to which these bots introduce, spread or amplify information remains elusive. For this reason, it is essential to gain greater understanding of the implications associated with human and machine dialogue, either intentional or not.

Initial social bot research continues to build upon a foundation of the classification and detection of social bots in OSNs (e.g. Chavoshi et al., 2016; Chu et al., 2012; Davis et al., 2016). The increasing sophistication of bots and the ability of some bots to mimic human behavior are proving to be too complex for current passive detection methods (Cresci et al., 2017). Even some simple rules-based social bots continue to gain an influential role in networks and go undetected for extended periods of time (Abokhodair et al., 2015). Recent promising advances in active bot detection algorithm development follow an adversarial learning approach by employing genetic algorithms to detect evolving bots (Cresci et al., 2018, 2019b, 2019a). While bot detection methodologies are improving with respect to keeping pace with evolving bot sophistication, there exists ample opportunities to develop and test necessary social bot analysis techniques to better characterize currently detectable social bots. Recent initial social bot analysis studies, which rely upon an array of multidisciplinary approaches, have provided positive insights into social bot influence within OSN conversations involving healthcare issues (Broniatowski et al., 2018), elections (Howard et al., 2018; Stella et al., 2018), financial trading markets (Cresci et al., 2018, 2019) and protests (Suárez-Serrato et al., 2016). Given that social bots aim to mimic and replicate human behavior, some researchers suggest that a computational social science (CSS) paradigm

could provide a compelling framework for characterizing the influence that bots may have on OSN conversations (Ciampaglia, 2018; Strohmaier & Wagner, 2014).

It is from a CSS perspective that this chapter extends the unique methodology and analysis framework presented in Chapter 2 to observe human and social bot behavior and interactions within OSN conversations in greater detail. This chapter observes the same OSN conversations but includes additional analytic techniques, so some of the following data description commentary serves as a refresher to the reader. This extended study relies upon acquired Twitter data associated with three major global events in 2016: the 2016 U.S. presidential election primary races, the ongoing Ukrainian conflict involving Russia and Ukraine, and the Turkish government's implementation of censorship practices against its own citizens. Bot enrichment procedures applied to the Twitter data then classify the bot status of all user accounts within the corpus. This enables a multi-faceted data analysis approach that includes comparative descriptive statistical analysis methods and social network analysis techniques to determine the relative importance and persistence of social bots within each global conversation. Overall, the constructed corpus consists of over 28.6 million tweets produced by approximately 5 million distinct users, of which, the bot labeling process identifies 14,386 of those users as likely social bots producing more than 1.9 million tweets. This reproducible framework, which can be extended to other OSN conversations and additional bot detection algorithms, creates an opportunity to better describe currently detected bots, while also providing essential feedback loops to bot detection research.

The results of this study show that suspected social bot users, on average, attempt to initiate contact with other users via retweets at a rate far higher than human users. Through the application of social network analysis centrality measurements, the results discover that social bots, while comprising less than 0.3% of the total user population, display a profound level of structural network influence by ranking particularly high among the top eigenvector centrality users within the U.S. Election and the Ukraine Conflict OSN conversations. Further, in observing the temporal persistence of suspected social bots, the presented findings show that bot users maintain their density of top centrality rankings over the cumulative OSN conversations of interest. Finally, the most relatively influential social bots from the Twitter corpus display a distinct ability to attract higher in-degree edge connections from human users that retweet their original bot messages. These results are quite promising given this study relied upon one open-source bot detection platform which provides limited total conversational coverage, but precise positive bot classification.

The remainder of this chapter is structured as follows. In the Background section (Section 3.2), a brief synopsis introduces current social bot detection methods and social bot analysis efforts. Enabling a Social Bot Analysis Framework (Section 3.3) provides a detailed overview of this study's processes, which acquire and fuse the data sources to enable the subsequent analysis section. Analysis Results and Discussion (Section 3.4) focuses on the results of the comparative descriptive statistical analysis methods and social network analysis techniques from this study and discusses their implications across

the global event use-cases of interest. Finally, the chapter concludes with the Conclusion section (Section 3.5) and highlights potential future research opportunities.

3.2. Background

In the past 15 years, the digital data exhaust created by increasing OSN usage rates and the relative ease at which researchers can gain access to such data has led to the rapid emergence of social media research. As social media research norms continue to develop, studies have produced insights from OSN-extracted data on topics including disaster response (Avvenuti et al., 2016; Crooks et al., 2013; Sakaki et al., 2013), mental illness forecasting (Reece et al., 2017) and political polarization (Conover et al., 2011). The limitations and risks associated with using OSN data for research are well documented (Ruths & Pfeffer, 2014; Tufekci, 2014), but the adaptive nature of social bots participating in OSNs amplify these concerns and may lead to many additional research implications (Morstatter et al., 2016).

The evidentiary rise of social bots in OSNs has led to a corresponding increase in research dedicated to bot detection (Murthy et al., 2016).The motivation and design methods associated with bots can vary dramatically, so a myriad of detection methods is necessary to account for the potential characteristics or activities attributable to certain social bots. In the following, we focus on two bot detection platforms, Botometer (Davis et al., 2016) and DeBot (Chavoshi et al., 2016), which exhibit very dissimilar design criteria but are both widely used in research due to the fact that they provide open access through web applications and application programming interfaces (APIs).

The Botometer (formerly named BotOrNot) bot detection platform employs a supervised ensemble Random Forest classification technique, which classifies potential Twitter accounts as bots according to six different classifiers based on more than 1,000 extracted features from an associated Twitter account (Davis et al., 2016). Botometer assigns a probabilistic [0,1] score representing the likelihood that a Twitter account is a bot, with simple and sophisticated bots falling within score ranges of 0.8-1.0 and 0.5-0.7, respectively (Varol et al., 2017). The DeBot bot detection platform, on the other hand, relies upon an unsupervised warped correlation method to find correlated Twitter accounts that have more than 40 synchronous events within a given time window (Chavoshi et al., 2016). DeBot provides a binary positive or negative bot classification for a Twitter account at incredibly high levels of precision (Chavoshi et al., 2017), but at a cost of recall due to evaluating smaller populations of Twitter accounts (Morstatter et al., 2016). In contrast to Botometer, DeBot archives its detection results, which allows researchers to ascertain potential bot status for previously detected accounts from a historical perspective (Chavoshi et al., 2017). As Cresci et al. (2017) aptly asserts, individual bot detection methodologies are not designed to detect the wide range of operational social bot types, and they require continual refinement to keep pace with evolving bot sophistication.

Social bot analysis is gaining traction as a means to better understand the impact of social bots and potentially provide essential feedback to bot detection research efforts. While social bot analysis currently lacks a formal definition, we submit an informal definition to be a multidisciplinary research effort employing quantitative and/or

qualitative methods with a stated purpose of better understanding detectable social bots and their behaviors in OSNs. Recent initial social bot analysis contributions examine the presence of detected social bots in Twitter conversations involving the 2016 U.S. presidential election (Bessi & Ferrara, 2016; Howard et al., 2018), the United Kingdom Brexit referendum (Duh et al., 2018; Howard & Kollanyi, 2016), the ongoing Ukrainian-Russian conflict (Hegelich & Janetzko, 2016), financial trading markets (Cresci et al., 2018, 2019) and the debates on vaccination (Broniatowski et al., 2018). These works have built the initial corpus of social bot analysis research, but much work is left to be done to introduce more advanced evaluation methods across greater use-cases of interest. One path to advancing these evaluation methods are social network analysis (SNA) techniques.

Observable human and bot interactions in OSN platforms such as Twitter provide a prime opportunity to employ SNA techniques to evaluate the relative importance of detected bots in comparison to human users. A key finding in Boshmaf et al. (2013), Aiello et al. (2014) and Mønsted et al. (2017) is that social bot infiltration and subsequent interactions with human users in OSNs occur at surprisingly high rates. Learning from Cha et al. (2010) that relative influence in Twitter by users is not necessarily gained through popularity (i.e. associated follower volume), we can look to SNA techniques to derive influence in OSNs (Bakshy et al., 2011; Kwak et al., 2010; Riquelme & González-Cantergiani, 2016; Weng et al., 2010).

Initial social bot research employing advanced social network analysis techniques to evaluate bot influence in OSNs is limited but growing. Aiello et al. (2014) applies the

PageRank and Hypertext Induced Topic Search (HITS) link analysis algorithms to judge the relative importance of an experimental bot. In observing the Catalan referendum Twitter conversation, Stella et al. (2018) uses an average PageRank valuation to compare suspected bot and human accounts, while also showing that bot interactions targeting human accounts positively correlates with the in-degree of human-to-human interactions. Perna and Tagarelli (2018) present the most promising effort to quantify social bot relevance with their ensemble machine learning framework, Learning-To-Rank-Social-Bots (LTRSB). The LTRSB framework aims to provide a unifying method to rank bots based on the extracted features present in the available bot detection platforms (e.g. Botometer, DeBot, BotWalk) (Perna & Tagarelli, 2018).

OSN research has turned into a burgeoning field in academia that has risen in stride with the overall rapid advancement in global social media usage. The increasing reliance upon OSN platforms as primary news sources by today's digitally-focused citizens, however, highlights the need to better identify and analyze the implications of social bot actors participating in online dialogue. Therefore, there is an immediate need to develop research methods to account for social bot implications within the larger field of OSN research.

3.3. Enabling a Social Bot Analysis Framework

This study employs and extends a social bot analysis framework that focuses on the aggregation of multiple harvested Twitter conversations and bot detection results to better characterize the relative influence and persistence of social bots in OSNs. This section describes the processes to transform these data, which enable the ensuing

ensemble application of comparative descriptive analysis methods and SNA techniques in this study. Figure 8 summarizes the processes comprising the social bot analysis framework and detailed subsections constitute the remainder of this section. Data Acquisition and Processing (Section 3.3.1) presents the details behind each OSN conversation of interest and the associated keywords serving as the input parameters to harvest tweets. Bot Enrichment (Section 3.3.2) describes the bot labeling process for each Twitter user in this study's corpus. Retweet Network Construction (Section 3.3.3) explains the process to build network objects for each OSN conversation, while Data Analysis Overview (Section 3.3.4) concludes the section by introducing the analysis methods comprising the subsequent sections of this chapter.

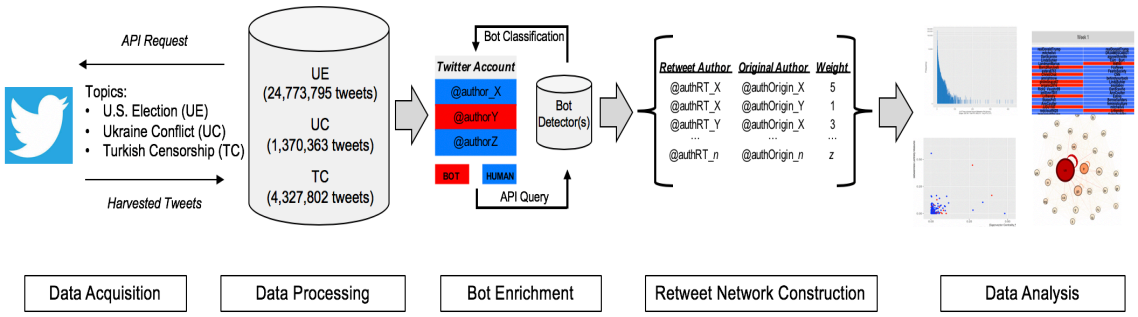


Figure 8: Social bot analysis methodological framework overview depicting the processes required to transform data to enable comparative analysis. OSN conversation data focused on three 2016 global events served as the example use cases, including: the 2016 U.S. Election (UE), the Ukraine Conflict (UC) and Turkish Censorship (TC).

3.3.1. Data Acquisition and Processing

In the same fashion as Chapter 2, three major global events from 2016 serve as the OSN conversation use cases in this chapter's study. Focusing solely on Twitter, the

analysis examined harvested tweets from four weeks of different topical conversations, to include a political election (2016 U.S. Presidential Election), a war/conflict (2016 Ukraine Conflict) and censorship (2016 Turkish Censorship). By analyzing varied topics, this study seeks to determine social bot behavioral differences across a diverse set of conversations. The following introduces and briefly summarizes the three OSN conversations of interest:

U.S. Presidential Election (February 1-28, 2016): This OSN conversation observes four weeks of tweets in February 2016 based on keywords associated with the 2016 U.S. Presidential Election. During this period, the election's primary races to determine the Republican and Democratic party candidates for the general election are well underway. The Republican primary race attracts considerable social media attention as then-candidate Donald Trump gains substantial momentum towards securing the Republican nomination over Texas Senator Ted Cruz. The Democratic race develops into a two-candidate race between former Secretary of State Hillary Clinton and Vermont Senator Bernie Sanders.

Ukraine Conflict (August 1-28, 2016): This OSN conversation observes four weeks of tweets in August 2016 based upon keywords associated with the ongoing conflict between Ukraine and Russia. At this point in time, it has been fewer than three years since the anti-Russian Euromaidan protests and the

subsequent annexation of Crimea by Russia. Military bravado and political rhetoric between these nations increases dramatically as the 25th anniversary of Ukrainian independence from Russia approaches (August 24, 1991).

Turkey Censorship (December 1-28, 2016): This OSN conversation observes four weeks of tweets based on keywords associated with Turkish government censorship of OSNs, specifically Twitter, in December 2016. Following a failed coup attempt against the sitting Turkish government in July 2016, government officials are keen to monitor and suppress messaging campaigns on OSNs. In December 2016, the Turkish government explicitly blocks Turkish citizens from using Twitter immediately following two events. The first block period takes place in the aftermath of the public assassination of Andrei Karlov, the Russian Ambassador to Turkey, on December 19, 2016. Turkey initiates a second block on December 23rd immediately following the release of a video that shows two Turkish soldiers being burned alive.

Based on the OSN conversation overviews as described above, Table 3 displays the representative keywords used associated with each topic. These keywords serve as the filter parameter to harvest associated tweets via the Twitter Standard Search API. Overall, the keyword harvest yields more than 28.6 million total tweets produced by approximately 5 million unique accounts with a breakdown for each OSN conversation as follows: U.S. Presidential Election ~23.3 million tweets (~3.3 million unique accounts),

Ukraine Conflict ~1.3 million tweets (~0.4 million unique accounts) and Turkish Censorship ~4.0 million tweets (~1.3 million unique accounts). In order to account for the storage and computation demands for such a large data corpus, all initial data processing took place within an Amazon Web Services (AWS) EC2 t2.2xlarge instance consisting of 8 vCPUs and 32GiB of RAM. In doing so, the AWS instance enabled the rapid creation specified data objects for processing at the local level, while also maintaining a scalable compute/storage platform to account for future data expansion.

Table 3: Key word list of terms for submission to Twitter Standard Search API for each OSN conversation of interest in this study

U.S. Election* *(Language)*		Ukraine Conflict *(Language)*		Turkish Censorship *(Language)*	
trump (English)	clinton (English)	ukraine (English)	киев (Russian)	turkey (English)	erdoganblockedtwitter (English)
@realdonaldtrump (English)	hillary (English)	ukrainian (English)	київ (Ukrainian)	türkei (Turkish)	erdoganblockstwitter (English
cruz (English)	sanders (English)	украина (Russian)	crimea (English)	turkish (English)	twitterisblockedinturkey (English)
@tedcruz (English)	bernie (English)	russia (English)	spetsnaz (English)	erdogan (English)	direntwitter (English)
makeamericagreatagain (English)	gop (English)	russian (English)	specnaz (English)	erdoğan (Turkish)	resisttwitter (English)
trump2016 (English)	gopdebate (English)	kiev (English)	putin (English)	turkeycoup (English)	occupytwitter (English)

*Keywords submitted for the 2016 U.S. Election skew heavily toward capturing tweets associated with then-candidate Donald Trump. We provide representative inclusion of other candidates but attempt to capitalize on the almost celebrity status of Mr. Trump by including additional Trump-related keywords.

3.3.2. Bot Enrichment

To individually label each unique Twitter account user in the tweet corpus as a human or a suspected bot, the bot enrichment step relied upon the open-source DeBot bot detection platform (Chavoshi et al., 2016). DeBot was the logical bot detection platform to use as the detection service proof of concept for this study since, as the Background section (Section 3.2) details, the archival nature of DeBot allows for the classification of

the historical Twitter user accounts. Further, DeBot, via its unsupervised warped correlation method, detects bots at much higher precision rate than other bot detection platforms (Chavoshi et al., 2017). While such precision comes at the cost of lower recall and increases the risk of false-negative bots (i.e. automatically assessing non-assessed accounts as human accounts) as Morstatter et al. (2016) notes, DeBot is the logical platform to initially test this social bot analysis framework given the historical nature of our Twitter corpus. As further stressed in the Conclusion section (Section 3.5), future improvements in social bot analysis research will rely upon the increased availability of additional bot detection algorithms to researchers which will allow for a more comprehensive coverage of all types of bots.

The bot enrichment process entails extracting all unique tweet author names from this study's tweet corpus and passing them for classification via the DeBot API[3]. The returns simply classify the tweet author name as a suspected social bot or not (i.e. a human author). Parsing scripts then automatically label each user account and merge the bot classification results with the tweet corpus. This process is easily extendible to account for other bot detection platform results. While beyond the scope of this study due to the historical nature of the tweet corpus, future work should also consider tracking the suspension/deletion statuses of accounts as the typical activities of social bot accounts make them primary targets of such actions by Twitter (Ferrara, 2017).

DeBot ultimately labels 14,386 Twitter user accounts as likely social bots based on the classification results. This includes restricting positive bot labels to accounts only

[3] Accessible at https://www.cs.unm.edu/~chavoshi/debot/.

evaluated by DeBot prior through the dates of the Twitter corpus. While this population represents just 0.29% of the total unique user accounts in the corpus, social bots are very active, and account for an over twentyfold share of the corpus of published tweet (1,966,623 tweets, or 6.80% of the total) and over thirtyfold share of the corpus of published retweets (1,495,388, or 8.84% of the total). Table 4 below provides weekly and cumulative corpus metrics for each of the OSN conversations of interest. At the specific OSN conversation level, the U.S. Election corpus shows much greater weekly and cumulative tweet and retweet percentage contributions from social bot user accounts in comparison to the Ukraine Conflict and Turkey Censorship corpuses, even though the relative percentage of total bot accounts is much smaller in the election corpus. Further, social bots account for a higher percentage of total retweets in comparison to tweets across all conversations.

Table 4: Twitter corpus overview at the weekly and cumulative perspective for each OSN conversation in this study.

Corpus	Week(s)	Tweets	Retweets	Users
United States Election Bot Source *(% of total)*	Week 1 *(Feb. 1-7, 2016)*	4,054,560 315,540 *(7.78%)*	2,280,176 235,815 *(10.34%)*	1,029,090 4,229 *(0.41%)*
	Week 2 *(Feb. 8-14, 2016)*	4,991,968 423,976 *(8.49%)*	2,802,381 326,808 *(11.66%)*	997,107 4,260 *(0.43%)*
	Week 3 *(Feb. 15-21, 2016)*	5,704,997 474,652 *(8.32%)*	3,284,436 373,496 *(11.37%)*	1,215,948 4,314 *(0.35%)*
	Week 4 *(Feb. 22-28, 2016)*	8,580,214 573,950 *(6.69%)*	5,071,862 442,897 *(8.73%)*	1,661,688 4,720 *(0.28%)*
	Cumulative	*23,331,739* *1,788,118 (7.66%)*	*13,438,855* *1,379,016 (10.26%)*	*3,313,230* *6,776 (0.20%)*
Ukraine Conflict Bot Source *(% of total)*	Week 1 *(Aug. 1-7, 2016)*	306,544 12,605 *(4.11%)*	155,151 8,059 *(5.19%)*	75,653 1,445 *(1.91%)*
	Week 2 *(Aug. 8-14, 2016)*	305,796 13,764 *(4.50%)*	141,193 8,272 *(5.86%)*	107,281 1,200 *(1.12%)*
	Week 3 *(Aug. 15-21, 2016)*	381,146 17,012 *(4.46%)*	210,047 11,212 *(5.34%)*	143,684 1,647 *(1.15%)*
	Week 4 *(Aug. 22-28, 2016)*	280,761 8,772 *(3.12%)*	133,985 5,247 *(3.92%)*	122,845 1,126 *(0.92%)*
	Cumulative	*1,274,247* *52,153 (4.09%)*	*640,376* *32,790 (5.12%)*	*364,422* *2,436 (0.67%)*

Bot Source *(% of total)*				
	Week 1 *(Dec. 1-7, 2016)*	709,530 18,243 *(2.57%)*	442,429 11,976 *(2.71%)*	305,239 1,892 *(0.62%)*
	Week 2 *(Dec. 8-14, 2016)*	894,900 27,349 *(3.06%)*	591,209 18,189 *(3.08%)*	410,035 2,705 *(0.66%)*
	Week 3 *(Dec. 15-21, 2016)*	1,486,289 44,833 *(3.02%)*	1,036,436 29,696 *(2.87%)*	635,535 3,807 *(0.60%)*
	Week 4 *(Dec. 22-28, 2016)*	917,937 25,575 *(2.79%)*	571,393 17,379 *(3.04%)*	425,435 2,702 *(0.64%)*
	Cumulative	*4,008,656* *116,000 (2.89%)*	*2,641,467* *77,240 (2.92%)*	*1,322,010* *5,174 (0.39%)*

3.3.3. Retweet Network Construction

The practice of retweeting can produce a diverse range of conversational implications, but Twitter users that deliberately retweet are more likely trying to engage in conversation or directly share information (Boyd et al., 2010). In this study, retweets account for 58.4% (~16.7 million) of the total tweet corpus, with the specific conversation retweet densities of 57.6%, 50.3% and 65.9% for the U.S. Election, the Ukraine Conflict and the Turkish Censorship conversations, respectively. The act of a retweet between two Twitter users (i.e. nodes) results in an observable directed network connection (i.e. an edge). Each directed edge receives a weight value of '1' for each initial directed retweet connection between two users and the edge weight increases for each additional number of retweets between the appropriate directional pair of users.

A retweet serves as the primary artifact to extract a 'node-edge' network construct from a Twitter conversation and ultimately enables the application of the SNA methods introduced in the subsequent Data Analysis Overview section (Section 3.3.4). In total, each four-week OSN conversation of interest produces a fairly large cumulative directed retweet network to analyze: 2,431,030 nodes / 8,437,925 edges (U.S. Election), 238,714

nodes / 509,614 edges (Ukraine Conflict) and 1,030,381 nodes / 2,088,524 edges (Turkish Censorship).

3.3.4. Data Analysis

The following concludes the introduction of the social bot analysis methodological framework by discussing the last step, data analysis. While data analysis is an entirely broad characterization of a step, it is the noted culmination point of acquisition, normalization, fusion and transformation of the harvested Twitter conversation data that enables us to address the overall research questions by applying the methods put forth in the subsequent sections comprising the Analysis Results and Discussion (Section 3.4) of this chapter's study. Furthermore, as this chapter seeks to contribute to the expansion of social bot analysis techniques, it does not portend that the proposed analysis methods are comprehensive, but merely foundational building blocks paving the way for future application methods.

3.4. Analysis Results and Discussion

This section presents the findings of the comparative descriptive statistical analysis methods and social network analysis techniques of this study and discusses the resulting implications of social bot evidence across the global event conversations of interest. By analyzing multiple significant global OSN conversations this analysis expands current social bot analysis literature. Further, the deployed methodological framework shows how the adoption of SNA techniques can provide quantifiable and comparative results to determine the relative impact or influence of suspected social bots in OSN conversations. Bot and Human User Communication Participation (Section 3.4.1)

compares the communication trends of human and bot Twitter users by observing participation volume and identifying the proclivity to engage with certain types of users. Temporal Persistence of Bot Centrality Rankings (Section 3.4.2) conducts centrality measurements within the retweet networks and evaluates the persistence of social centrality rankings over time. This section concludes with Prominent Bot Ego Networks (Section 3.4.3) dissecting the associated ego networks of the highest ranking eigenvector centrality bot from each OSN conversation.

3.4.1. Bot and Human User Communication Participation

The analysis first compares the communication participation patterns of bot and human users by examining the associated tweet and retweet volume rates. Table 5 summarizes the corresponding average and median volume rates across all three OSNs. We see social bots exhibit much higher average and median participation rates, which is not surprising given the large volume of contributions made by such a small bot population. Of interest though, we see that the top human user account tweet volumes dominate the top bot account tweet volumes across all OSNs, while top bot account retweet volumes are dominant except in the case of the Turkish Censorship OSN.

Table 5: Overall conversation tweet contribution volumes by human and likely social bot users within each OSN conversation corpus of interest.

Corpus	User Type	Average User Tweet Volume	Median User Tweet Volume	Min/Max User Tweet Volume *(# of users)*	Average User Retweet Volume	Median User Retweet Volume	Min/Max User Retweet Volume (# of users)
U.S. Election	Humans	6.5	1.0	Min: 1 *(1,920,647 users)* Max: 45,542 *(1 user)*	5.4	1.0	Min: 1 *(1,428,380 users)* Max: 8,016 *(1 user)*
	Bots	263.9	13.0	Min: 1 *(1,032 users)* Max: 19,905 *(1 user)*	267.0	10.0	Min: 1 *(950 users)* Max: 19,905 *(1 user)*
Ukraine Conflict	Humans	3.4	1.0	Min: 1 *(231,623 users)* Max: 28,072 *(1 user)*	2.8	1.0	Min: 1 *(146,879 users)* Max: 1,265 *(1 user)*
	Bots	21.4	4.0	Min: 1 *(569 users)* Max: 2,623 *(1 user)*	16.8	3.0	Min: 1 *(523 users)* Max: 2,508 *(1 user)*
Turkish Censorship	Humans	2.9	1.0	Min: 1 *(882,172 users)* Max: 10,065 *(1 user)*	2.7	1.0	Min: 1 *(647,261 users)* Max: 28,072 *(1 user)*
	Bots	22.4	4.0	Min: 1 *(1,416 users)* Max: 1,936 *(1 user)*	17.3	4.0	Min: 1 *(1,331 users)* Max: 4,966 *(1 user)*

Figure 9 presents the cumulative total tweet contribution percentages by human and bot users over the four weeks of harvested tweets for each OSN conversation. The U.S. Election (Figure 9a) and the Ukraine Conflict (Figure 9b) conversations both exhibit a gap between bot and human contribution percentages that begins to widen at approximately two weeks into the conversation and closes over the final days. A similar gap between users does not exist in the Turkish Censorship conversation (Figure 9c), while its initial conversation trajectory is much shallower until a spike in contributions takes place corresponding to the onset of the first censorship event in Turkey on December 19, 2016. This latter contribution spike, coupled with lower overall social bot tweet/retweet volumes and participation rates observed in Table 4 and Table 5, might be symptomatic of the Turkish Censorship conversation being an emergent topic during the

period of observations, as opposed to the already established U.S. Election and Ukraine Conflict conversations.

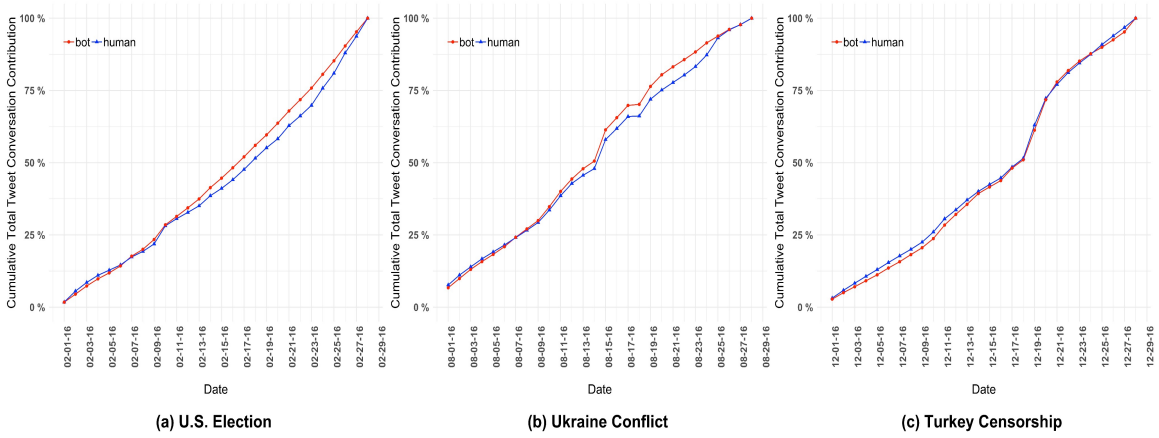
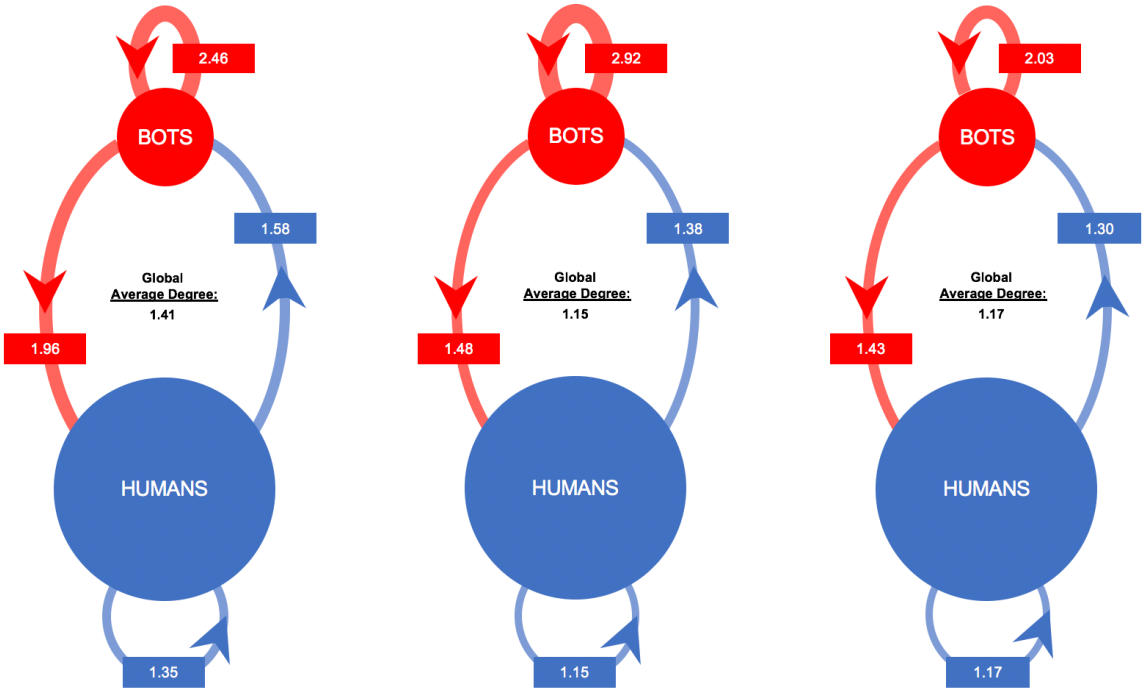


Figure 9: Cumulative total tweet contributions over the four-week Twitter conversation span for: (a) U.S. Election (February 1-28, 2016), (b) Ukraine Conflict (August 1-28, 2016), (c) Turkish Censorship (December 1-28, 2016).

The volume of in-group and cross-group communication within OSN retweet conversations provides an additional opportunity to classify communication patterns. This study defines in-group communication as retweets between like types of users (i.e. humans retweeting humans and bots retweeting bots), while cross-group communication denotes retweets between different types of users (i.e. humans retweeting bots or bots retweeting humans). In terms of total retweet volume percentage for each conversation, humans dominantly retweet other human accounts at total volume rates of 84.92% (U.S. Election), 92.12% (Ukraine Conflict) and 94.74% in (Turkish Censorship), while bot in-group retweet rates occur at relatively low rates of 1.38% and lower. To overcome the human dominance volumes, retweet interactions are normalized by average edge weight

of specified group pairings. Figure 10 summarizes the resulting average weighted edges of all inter-group and cross-group communication pairs for each of the OSN conversations of interest. We see that bots, from an average edge weight perspective, engage in higher intra-group and cross-group communication rates across all three conversations, with the U.S. Election conversation showing the highest cross-group and intra-group engagement edge weights of 1.96 and 2.46, respectively.



(a) U.S. Election (b) Ukraine Conflict (c) Turkish Censorship

Figure 10: In-group and cross-group retweet communication average edge weights of human (blue) and social bot (red) users within each OSN conversation: (a) U.S. Election, (b) Ukraine Conflict and (c) Turkish Censorship. Arrows express communication directionality (e.g. bot directed engagements with human accounts with an average retweet edge weight of 1.96 for the U.S. Election OSN).

To further place these overall in-group and cross-group interactions into context, Table 6 presents the average retweet edge weight for all communication pairings over time in. The results show that from the weekly and cumulative perspective social bots engage with their in-group bot and cross-group human edge pairs at higher rates, except for the third week of the Turkish Censorship conversation. These across the board higher rates suggest social bots, on average, are hyper-social in comparison to humans: they are much more persistent in attempting to initiate contact with other users in Twitter as opposed to average human users.

Table 6: Average retweet edge weight for all inter-group and cross-group communications by human and bot users for each OSN conversation.

		Average Retweet Edge Weight			
Corpus	**Week(s)**	**Bot-to-Bot**	**Bot-to-Human**	**Human-to-Bot**	**Human-to-Human**
U.S. Election	Week 1 *(Feb. 1-7, 2018)*	2.04	1.83	1.44	1.29
	Week 2 *(Feb. 8-14, 2018)*	2.61	2.08	1.69	1.47
	Week 3 *(Feb. 15-21, 2018)*	2.68	1.94	1.61	1.32
	Week 4 *(Feb. 22-28, 2018)*	2.42	1.96	1.58	1.35
	Cumulative	*2.46*	*1.96*	*1.58*	*1.35*
Ukraine Conflict	Week 1 *(Aug. 1-7, 2018)*	3.02	1.47	1.39	1.15
	Week 2 *(Aug. 8-14, 2018)*	3.84	1.71	1.55	1.21
	Week 3 *(Aug. 15-21, 2018)*	2.42	1.32	1.31	1.12
	Week 4 *(Aug. 22-28, 2018)*	2.64	1.58	1.36	1.15
	Cumulative	*2.92*	*1.48*	*1.38*	*1.15*
Turkish Censorship	Week 1 *(Dec. 1-7, 2018)*	2.78	1.46	1.28	1.21
	Week 2 *(Dec. 8-14, 2018)*	1.85	1.47	1.26	1.16
	Week 3 *(Dec. 15-21, 2018)*	1.88	1.38	1.25	1.14
	Week 4 *(Dec. 22-28, 2018)*	2.23	1.47	1.48	1.19
	Cumulative	*2.03*	*1.43*	*1.30*	*1.17*

3.4.2. Temporal Persistence of Bot Centrality Rankings

Degree centrality is the total number of direct edges a node shares with other nodes in a network and does not recognize edge directionality. In a retweet network, degree centrality is synonymous with a Twitter user's popularity in the network. In-degree and out-degree centrality are simply degree centrality that take into account edge directionality. Nodes with higher in-degree centrality receive more directional edge contact from other nodes, while higher out-degree centrality signifies nodes that initiate more directional edge contact. In a retweet network, higher out-degree centrality equates to a Twitter user initiating more retweets, while higher in-degree means a Twitter user has more users retweeting its original messages. Eigenvector centrality is the weighted sum of all direct and indirect edges for a node that takes into account the individual degree centrality of each node in the network (Bonacich, 2007). From a retweet network perspective, eigenvector centrality is a global measure of influence within a conversation. Betweenness centrality measures the propensity of a given node falling on the shortest path between all other node pairs in a network (Freeman, 1977). We can view the betweenness centrality of a retweet network node as a measure of communication that flows through that specific node. Finally, PageRank is a derivation of eigenvector centrality, but places more importance on the degree value of the nodes that initiate edges with a node of interest (Brin & Page, 1998). Therefore, in a retweet network, a node with higher PageRank value receives more retweets from Twitter users that have greater popularity in the network.

To determine the relative importance of social bot users compared to human users, the study calculated the chosen centrality measures for the entire duration of each OSN conversation using the applicable centrality functions provided in the networkx Python package (Hagberg et al., 2008). Scale tests by the authors on larger Twitter datasets of at least twice the volumes of the events in this study (i.e. ~ 50 million tweets) comprising networks with cumulative edge volumes that are three times larger (i.e. ~25 million edges) returned efficient centrality processing times (i.e. PageRank calculation was most time intensive calculation at ~5 min 20 sec) within in a cloud environment with the same specifications detailed in the "Data acquisition and processing" section (Section 3.3.1). The rank order of the centrality results present the density of social bots within the top-N centrality ranking positions (*where, N = 1000 / 500 / 100 / 50*). The results (Figure 11) clearly show that suspected bot users, while representing only 0.28% of all corpus users, account for a significant number of high centrality rankings, especially out-degree and eigenvector centrality rankings. The prevalence of social bots among the top ranks of out-degree nodes shows the above-mentioned hyper-social attitude of bots: they attempt to induce interaction by retweeting other users at a significantly higher rate that their human counterparts. In terms of influence, the results show that bots infiltrate some of the highest eigenvector centrality rankings within the U.S. Election and the Ukraine Conflict conversations, where bots account for 36.0% and 30.0% of the top-50 influential accounts, respectively. These results are quite substantial given the employment of just one bot detection source.

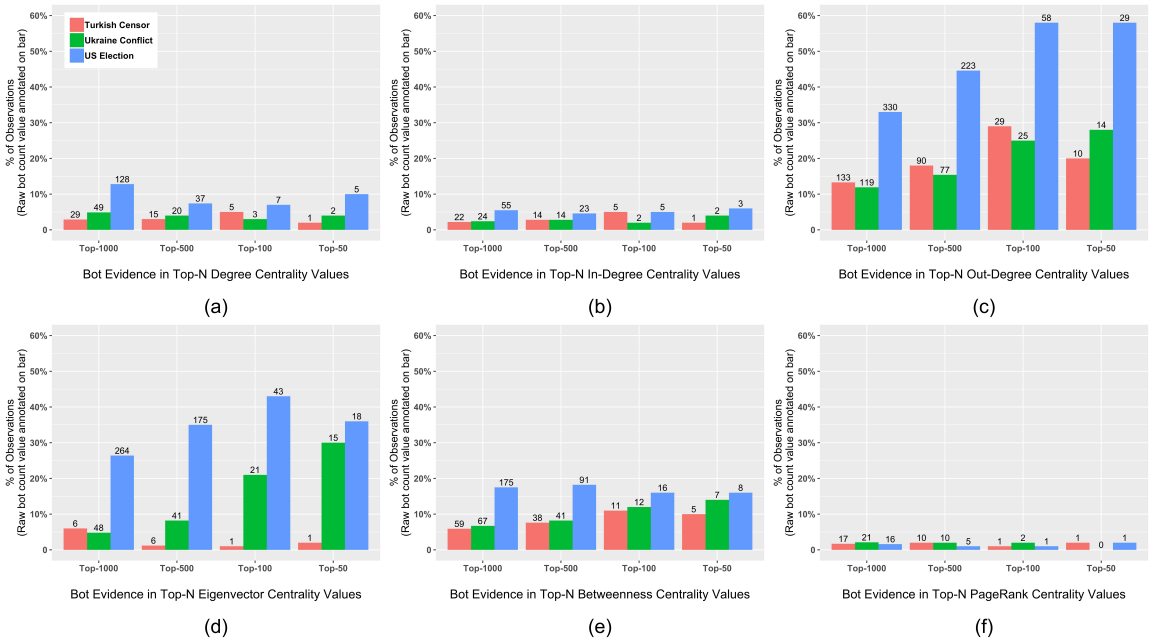


Figure 11: Social bot user evidence within the Top-N (where, N = 1000 / 500 / 100 / 50) [(a) degree (b) in-degree (c) out-degree (d) eigenvector (e) betweenness (f) PageRank] centrality rankings for the U.S. Election (blue), the Ukraine Conflict (green) and the Turkish Censorship (red) OSN conversations. Each bar chart represents the total social bot percentage of the range of accounts with a raw social bot account atop each bar.

To evaluate the temporal persistence of social bot centrality rankings, we recalculate and directly compare centrality rankings in a cumulative fashion over the four weeks for each OSN conversation. In doing so, we are able to analyze the centrality ranking staying power of identified social bot accounts over time, as opposed to an overall snapshot of the entire corpus timeframe. Figure 12 (U.S. Election), Figure 13 (Ukraine Conflict) and Figure 14 (Turkish Censorship) present a consolidated visualization depicting the density of bot (red block) and human (blue block) users as each conversation progresses on a weekly cumulative basis, while also annotating the individual accounts within each block.

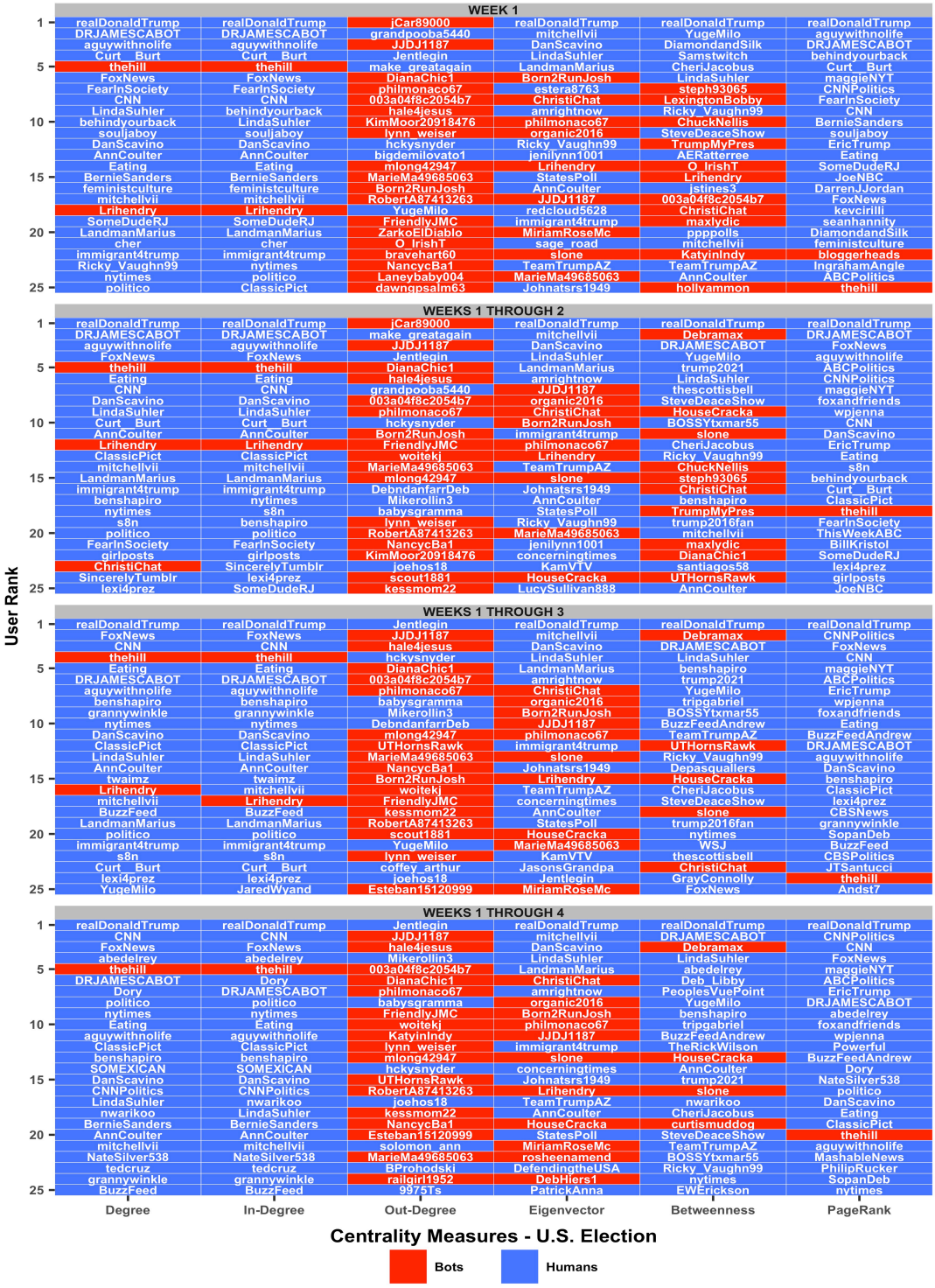


Figure 12: Centrality ranking of top-25 bot (red) and human (blue) users over a cumulative four-week period for the U.S. Election OSN conversation for six centrality measures: (1) degree, (2) in-degree, (3) out-degree, (4) eigenvector, (5) betweenness and (6) PageRank.

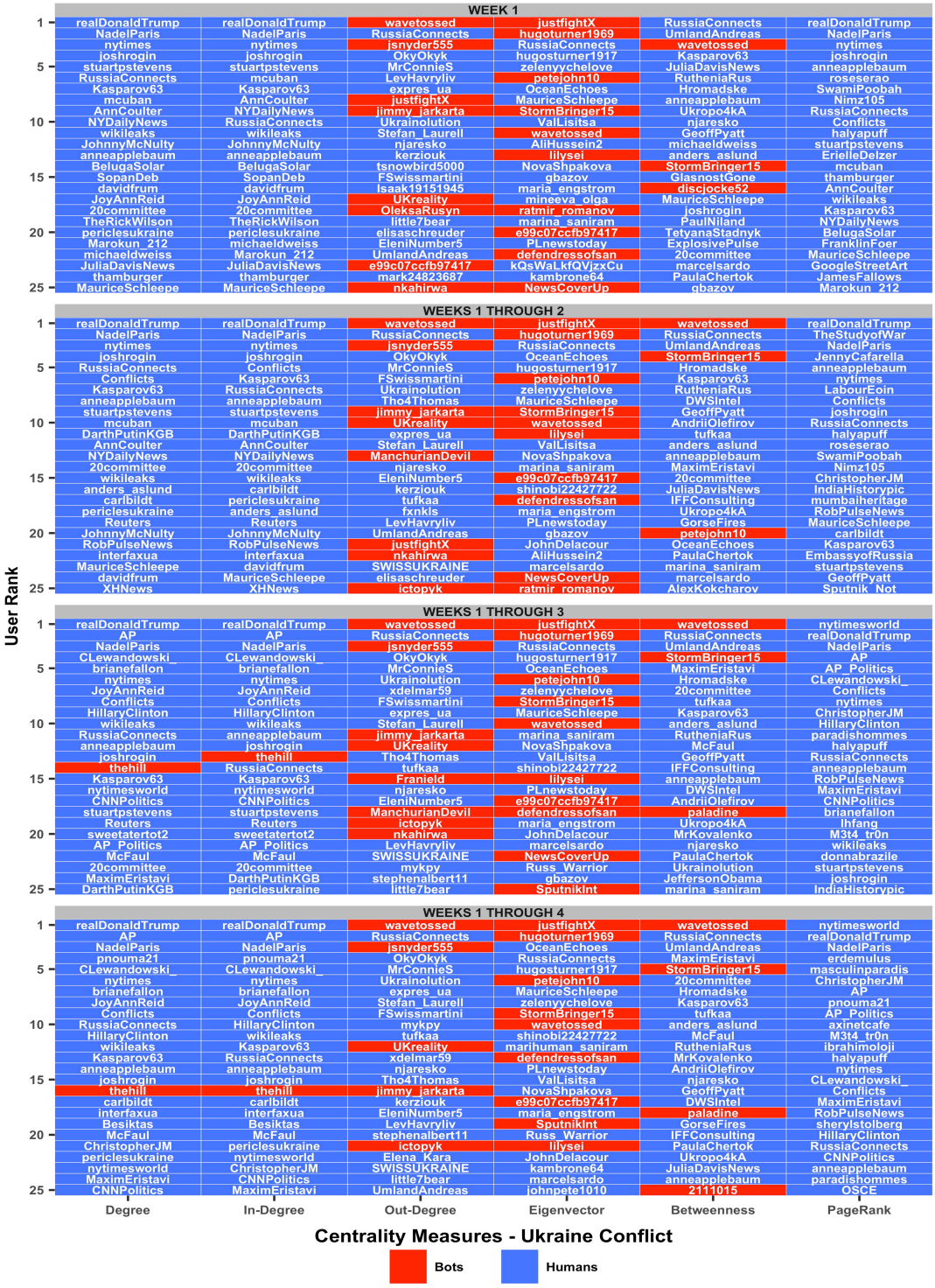


Figure 13: Centrality ranking of top-25 bot (red) and human (blue) users over a cumulative four-week period for the Ukraine Conflict OSN conversation for six centrality measures: (1) degree, (2) in-degree, (3) out-degree, (4) eigenvector, (5) betweenness and (6) PageRank.

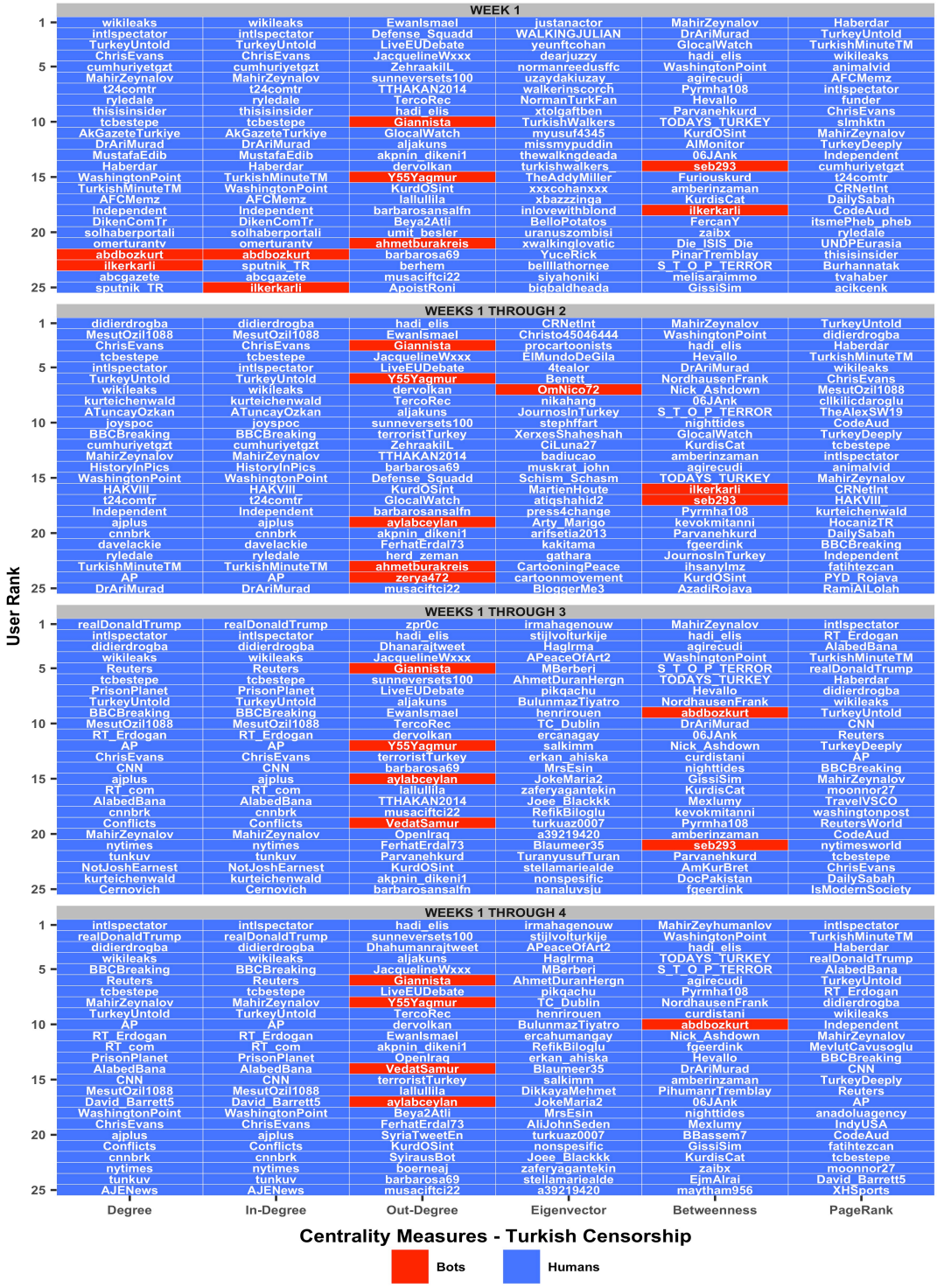


Figure 14: Centrality ranking of top-25 bot (red) and human (blue) users over a cumulative four-week period for Turkish Censorship OSN conversation for six centrality measures: (1) degree, (2) in-degree, (3) out-degree, (4) eigenvector, (5) betweenness and (6) PageRank.

The centrality ranking persistence of suspected bot users is visually evident over time across the cumulative conversations. The results show persistent bot density within each centrality ranking with especially high density associated with the out-degree and eigenvector centralities for the U.S. Election and the Ukraine Conflict conversations. This includes social bots achieving extremely high-rankings to include two of the top-5 out-degree, eigenvector and centrality rankings within the Ukraine Conflict conversation (Figure 13) and seven and four of the top-10 out-degree and eigenvector centrality rankings, respectively, within the U.S. Election conversation (Figure 12).

Observing the classification results of popular news source accounts (e.g. @CNN, @thehill, @AP) highlights the shortcomings of using only one bot detection service. For example, DeBot classifies @thehill as an automated bot account, but does not for @AP or @CNN. One can only assume, therefore, that coverage by DeBot has not evaluated those accounts by the time of this study. The account @FoxNews was later evaluated after this study by DeBot and determined to be an automated account on May 5, 2018, but this account maintained its original label given the evaluation dates of this study. Further extensions of this proof-of-concept work should include additional bot detection services, while consideration should be taken into potentially removing verified accounts from evaluation.

3.4.3. Prominent Bot Ego Networks

The following final Analysis Results and Discussion section (Section 3.4) investigates the ego networks of the highest ranking eigenvector centrality social bots from the U.S. Election (Twitter ID: *732980827*, Username: *ChristiChat*) and Ukraine

Conflict (Twitter ID: *3346642625*, Username: *justfightX*) OSN retweet conversations. No Turkish social bots achieved a high sustained eigenvector centrality ranking, so Turkish bots are excluded from consideration in this section. The *ego_graph* function provided in the Python networkx package derives the ego-networks based on immediately adjacent neighbors for each bot node of interest. The Bot and Human User Communication Participation section (Section 3.4.1) directly compares the extracted observable retweet network characteristics of these most relatively influential social bot users. Figure 15 provides a proportionally-scaled ego network that depicts the inter-group and cross-group neighbor interactions of these top eigenvector social bots. While both of these influential bots engage in differing levels of inter-group communication with other bots and cross-group communication with humans, both the U.S. Election and the Ukraine Conflict top eigenvector bots are able to establish in-degree and out-degree retweet connections with other top eigenvector ranking users. Further, each of these bot accounts are able to successfully solicit attention from human users that results in humans accounting for retweet rates 69.84% and 45.12% within the U.S. Election and Ukraine Conflict ego networks, respectively.

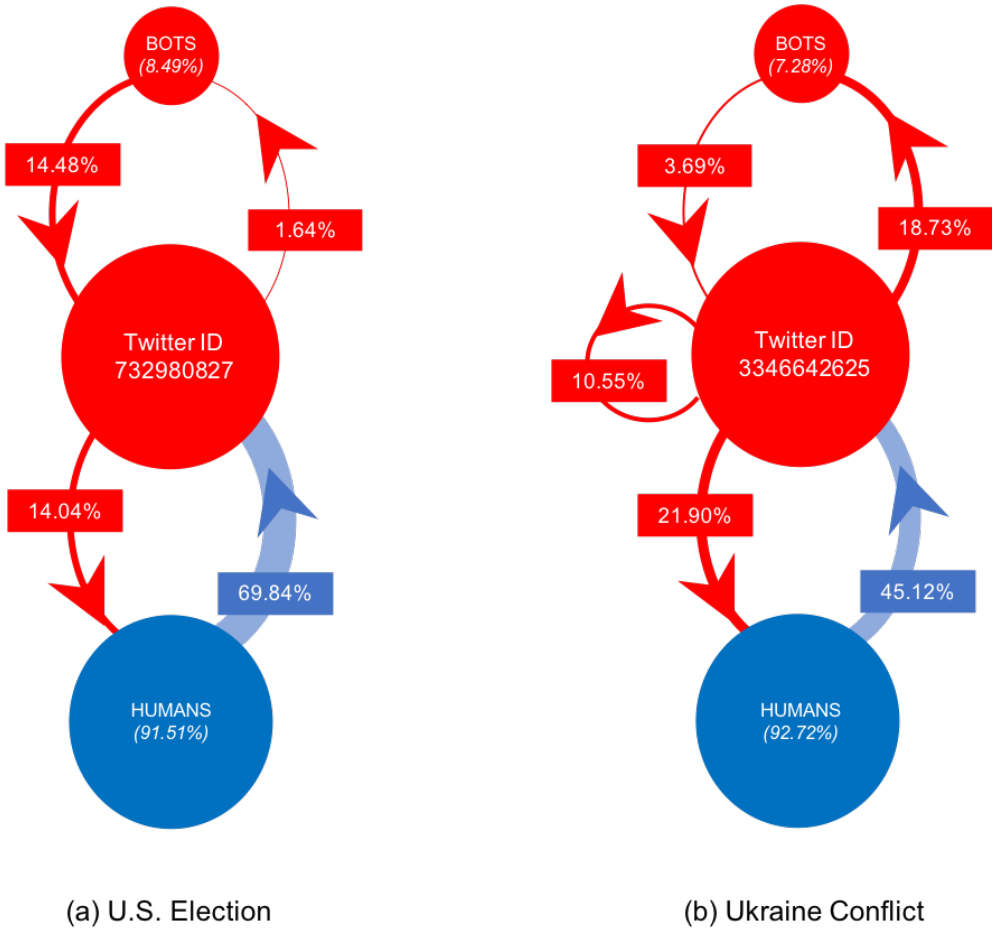


Figure 15: Ego network retweet patterns for the top-ranking eigenvector centrality bot accounts from the (a) U.S. Election and (b) Ukraine Conflict OSN conversation.

3.5. Conclusion

This chapter presented novel extensions of the social bot analysis methods presented in Chapter 2 and contributes to the expansion of the emergent area of social bot research. The unique social bot analysis methodological framework put forth enables the inclusion of additional bot detection platform services, while also opening the analysis window to account for new OSN conversations of interest. Through the lens of three major global event OSN conversations in 2016, the results confirmed the hyper-social

nature of bots: suspected social bots users make far more attempts on average than human users to initiate contact with other users via retweets. Social network analysis centrality measurements discover that social bots, while comprising less than 0.3% of the total user population, display a profound level of structural network influence by ranking particularly high among the top eigenvector centrality users within the U.S. Election and the Ukraine Conflict OSN conversations. Further, the results show that social bots exhibit temporal persistence in centrality ranking density across all of the OSN conversations.

While this chapter's study reports promising findings, it must account for its many limitations. Relying upon a single bot detection platform helped validate this study's applied network analysis methods, but a sole source detection algorithm is not sufficient for overcoming known specific limitations that currently challenge all open-source bot detection results (Cresci et al., 2017; Subrahmanian et al., 2016). Also, solely using data from a single OSN platform induces a litany of associated biases to include representativeness and sampling bias shortcomings (Tufekci, 2014). Ruths and Pfeffer (2014) further expands on social media data issues, while also singling out the inability to properly determine the presence of bots. While it is also in the spirit of this study to help improve overall bot detection methods, it is a reasonable perspective to state the current difficulties to determine ground truth effectiveness in detecting bots (Chavoshi & Mueen, 2018; Cresci et al., 2017; Subrahmanian et al., 2016). Further, a binary classification between bots and humans is not entirely sufficient as cyborg accounts also exist, which Chu et al. (2012) coins as bot-assisted human or human-assisted bot account.

Immediate primary extensions of this work should expand beyond the proof of concept framework demonstrated here and aggressively seek the inclusion of additional bot detection algorithms for a more holistic bot labeling perspective. While there currently exists a limited number of open-source bot detection algorithms, a comprehensive collection of detection sources would ideally include access to the continually improving pre-existing detection platforms (Beskow & Carley, 2018; Chavoshi et al., 2017; Varol et al., 2017), as well as recent novel detection algorithms based on detecting evolving bot signatures (Cresci et al., 2018; Mazza et al., 2019). Further extensions of this work could aim to incorporate additional social media sources beyond Twitter as Hecking et al. (2018) describe in a cross-media information diffusion example sourcing data from Twitter, Wikipedia edits and other web-based sources. In the case of this study, if the analysis does not observe centrality measures beyond just degree and PageRank centrality, then we miss the important social rankings made available via out-degree and eigenvector centrality. Therefore, it is important to maintain an expansive centrality analysis to account for social bots by potentially incorporating additional centrality measures, such as percolation centrality (Piraveenan et al., 2013), that may perform well in ranking social bot prominence within networks. On its own, this study is a unique stepping stone that adds to the growing research efforts focused on understanding social bot behavior in global event conversations.

CHAPTER 4. BOTS FIRED: EXAMINING SOCIAL BOT EVIDENCE IN ONLINE MASS SHOOTING CONVERSATIONS

4.1. Introduction

Mass shootings have become their own distinct phenomenon separate from the likes of general homicide and mass murder due to their continued prevalence and the natural draw of media attention to extreme events (Schildkraut et al., 2018). While not a formally defined government statistic, a mass shooting has generally been defined as an incident resulting in the death of four or more victims, not including the killer (Dahmen et al., 2018; Silva & Capellan, 2019; Towers et al., 2015). Moffat (2019) tallies that 620 people have been killed and more than 1,000 wounded from 70 mass shooting events beginning with the Columbine shooting in 1999 and concluding with the Parkland shooting in 2018. While the media reporting environment has changed drastically since the Columbine shooting with the advent of online social networks (OSNs) driven by the Web 2.0 paradigm, the general public's interest in mass shooting coverage remains high given the recent historical increase in mass shootings and the associated debate on the polarizing topic of gun control (Newman & Hartman, 2017). Media research has shown that particular newsworthy events, such as mass shootings, lend themselves to journalistic framing, which is the purposive highlighting of different attributes of a single event to attract or sustain interest (Chyi & McCombs, 2004). Guggenheim et al. (2015) points to a reciprocal relationship in framing mass shooting narratives between traditional and OSN

media sources, but also highlights how OSN content creators and users transcend the typical journalistic gatekeeping norms of traditional media by openly expressing emotional reactions.

In the United States, OSNs recently surpassed traditional print newspapers as a primary source for news and continue to gain traction on other traditional news sources such as television and radio (Mitchell, 2018). Mahabir et al. (2018) describes the current news ecosystem, comprised of both traditional news sources and online platforms (e.g. news websites/apps and social media), as highly participatory and fostering digital activism. Edwards et al. (2013) defines digital activism as an organized public effort orchestrated by supporters using digital media to make collective claims against a target authority. Given the highly contentious policy debates surrounding gun control that are a typical conversational byproduct in the immediate aftermath of a mass shooting event (Merry, 2016; Newman & Hartman, 2017), OSN conversations about mass shootings are a salient topic for digital activists. While a convenient means to access and publish content, OSNs have proven to be complicit in spreading and amplifying manipulated and/or blatantly falsified narratives (Bolsover & Howard, 2017; Lazer et al., 2018; Starbird, 2017; Vosoughi et al., 2018). A primary factor contributing to the skewed narratives in OSNs is the existence of vast populations of social media accounts controlled by social bots (Boshmaf et al., 2013).

Social bots are computer algorithms that automatically produce content and interact with human OSN users (Ferrara et al., 2016). Social bot pervasiveness has led to numerous research efforts focused on developing novel bot detection methods (Beskow

& Carley, 2018.; Chavoshi et al., 2016; Davis et al., 2016) and examining how bots spread information (Aiello et al., 2014; Mønsted et al., 2017; Shao et al., 2018). Further introductory works have analyzed the presence of social bots in various polarizing OSN conversations such as elections (Bessi & Ferrara, 2016; Howard & Kollanyi, 2016), conflict (Schuchard et al., 2019) and vaccinations (Broniatowski et al., 2018; Subrahmanian et al., 2016).

While some promising recent works have touched in various ways on the topic of social bots in OSN mass shooting conversations in various forms (Kitzie et al., 2018; Nied et al., 2017; Starbird, 2017), there is much depth that needs to be added through quantitative social bot analysis. In light of this, this chapter contributes to the literature by examining suspected social bots within Twitter conversations associated with four recent mass shooting events: the Las Vegas concert shooting (October 1, 2017), the Sutherland Springs church shooting (November 5, 2017), the Parkland school shooting (February 14, 2018) and the Santa Fe school shooting (May 18, 2018). Specifically, this study analyzed the presence and contribution patterns of social bots in relation to human users in an effort to determine potential cross-conversational norms of bot behavior. The applied analysis sought to quantify and classify the mentioning rate of previous mass shooting events in subsequent events to potentially classify certain events with persistent salience. Finally, social network analysis centrality measures measured the relative structural importance of social bots in relation to other users within each of the OSN conversation networks.

This study's results show that social bots participate and contribute to online mass shooting conversations in a manner that is distinguishable from human contributions. The cumulative conversation contribution rates of bots outpace humans throughout the Sutherland Springs and Santa Fe conversations. In the conversations involving the highly salient Las Vegas and Parkland shootings, human contributions initially outpace social bots, but an inversion takes place within the first week of each conversation as social bots become the dominant contributor for the remainder of the conversation. In terms of cross-group communications, human accounts engaged suspected bot accounts at higher rates than bots engaged humans in all of the conversations. Finally, bots, while accounting for fewer than 1% of all corpus users, displayed significant prominence in the conversation networks, densely occupying many of the highest eigenvector and centrality measure rankings, to include 82% of the top-100 eigenvector values of the Las Vegas retweet network.

The remainder of this chapter is as follows. First, the Background section (Section 4.2) presents relevant associated literature. Next, the Data and Methods section (Section 4.3) presents a detailed overview of the data acquisition and processing steps, as well as introducing the methods employed in the study. The Results and Discussion section (Section 4.4) presents the findings of applied methods to answer the study's research questions, followed by the Conclusion section (Section 4.5).

4.2. Background

A recent study by Dahmen et al. (2018) found that a majority of journalists agreed that the traditional news coverage of mass shooting events has become routine due to

perceived formulaic reporting. However, some mass shooting events garner more initial and sustained attention than others, and traditional media studies have focused much effort on identifying how certain media reports are able to succeed at this (Schildkraut et al., 2018; Silva & Capellan, 2019). One means of achieving both initial and sustained attention is through dynamic frame changing, or emphasizing different aspects of a news event over its life span (Chyi & McCombs, 2004). In terms of mass shootings, Muschert and Carr (2006) applied and extended the dynamic framing concept by analyzing frame changes over the course of media reporting on nine school shootings from 1997 to 2001. This resulted in the ability to directly compare the highly salient Columbine school shooting event with eight less salient shooting events (Pearl, MS; Paducah, KY; Jonesboro, AR; Edinboro, PA; Springfield, OR; Conyers, GA; Santee, CA; El Cajon, CA) using the frame-changing spatial (community, regional, societal) and temporal (past, present, future) categorizations (shown in Figure 16) of Chyi and McCombs (2004). Schildkraut and Muschert (2014) extended Muschert and Carr (2006) to include the Sandy Hook Elementary School shooting in an effort to be able to directly compare to the Columbine shooting to another mass shooting of extremely high salience.

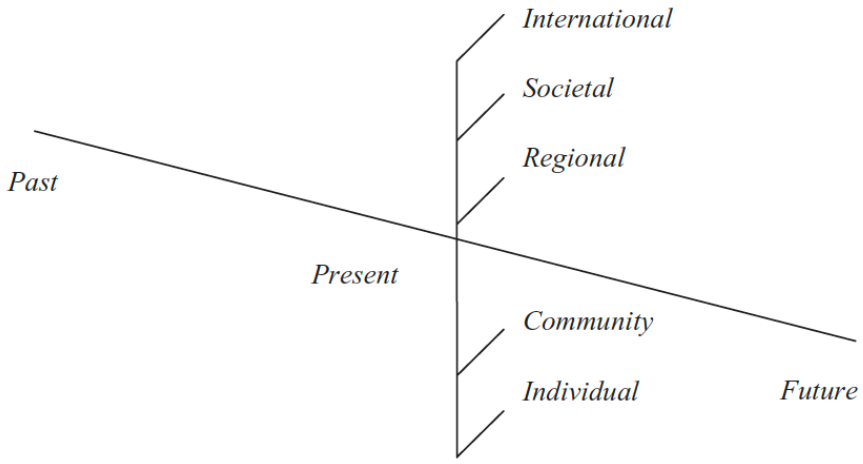


Figure 16: Two-dimensional analytical framework of Chyi and McCombs (2004) for comparing frame changes across similar media events. Source of figure: Schildkraut and Muschert (2014)

Although emerging from traditional media research, the concept of media framing serves as a primary analysis method in OSN conversations involving mass shootings as well. Guggenheim et al. (2015) examined the framing of mass shooting events in both traditional and OSN media coverage and concluded that there is a reciprocal relationship between the two. That is, tweets respond to traditional media reports just as traditional media reports respond to tweets. To account for additional narrative contributors, Merry (2016) extended the concept of mass shooting narrative framing to include framing by interest groups in OSN Twitter conversations (i.e. National Rifle Association, Brady Campaign to Prevent Gun Violence). In other work, Starbird (2017) examined the propagation and shaping of alternative narratives emanating from tweeted URLs related to mass shooting events, exposing how OSN interactions can enable a conspiratorial ecosystem of alternative media.

Overall, Varol et al. (2017) estimated that social bots account for 9-15% of all Twitter user accounts. This relatively large bot population estimate and the associated unknown implications of human users engaging with non-humans in Twitter have led to the rapid emergence of bot detection and classification research. While the ever-increasing sophistication of social bots mimicking human behavior has proven to be quite difficult phenomenon for researchers to keep pace with (Cresci et al., 2017), multiple bot detection research platforms have been launched to aid researchers in the overall detection and analysis of social bots in Twitter. Botometer[4], formerly named BotOrNot, is a widely used open-source bot detection platform that employs a supervised random forest detection algorithm against more than 1,100 extracted unique account features (Davis et al., 2016; Varol et al., 2017). Botometer then returns a classification score on a normalized scale identifying an account as more 'human' or 'bot-like'. Chavoshi et al. (2016) developed the open-source DeBot[5] bot detection platform that employs an unsupervised warped correlation bot detection algorithm. Rather than relying on feature extraction, the DeBot platform provides a binary bot classification based solely on the synchronous temporal activities of Twitter accounts.

While there is ample room for growth beyond the promising initial social bot analysis research identified in the Introduction section (Section 4.1) of this chapter, substantial introductory work is still needed to be started in analyzing bot activity in OSN conversations involving mass shootings. Kitzie et al. (2018) has produced the most recent bot-centric research covering mass shootings. It examined the retweet patterns and

[4] Botometer is accessible at https://botometer.iuni.iu.edu.

[5] DeBot is accessible at https://www.cs.unm.edu/~chavoshi/debot/.

associated narratives of more than 400 social bot accounts—identified by submitting a random sample of total user accounts for classification via Botometer—which were active in the Parkland school shooting Twitter conversation. Nied et al. (2017) conducted an exploratory analysis by hand-labeling suspected bots and examining alternative narratives—derived from the alternative narrative work of Starbird (2017)—in detected community clusters of OSN conversations of late 2015 discussing the Paris Attacks and the Umpqua Community College mass shooting.

4.3. Data and Methods

To address the research goals of discovering potential behavioral norms of social bots and the relative importance of bot accounts in relation to regular human contributors within and across mass shooting OSN conversations, this study relied upon the mixed methodology social bot analysis framework put forth in Schuchard et al. (2019). Figure 17 presents that adopted framework with annotated modifications to account for mass shooting events, while the following subsections provide a detailed overview describing each stage of the framework. First, Data Acquisition and Processing (Section 4.3.1) introduces the data sources and the processing steps used to transform the mass shooting event conversation data and enable the subsequent applied analysis methods. Bot Enrichment (Section 4.3.2) details the bot identification and labeling process of the harvested Twitter user accounts. Retweet Network Construction (Section 4.3.3) outlines the steps taken to create a network graph object of the retweet network for each OSN mass shooting conversation. Finally, Data Analysis (Section 4.3.4) introduces the

methods employed to comparatively analyze the evidence of suspected bots across this study's OSN mass shooting conversations of interest.

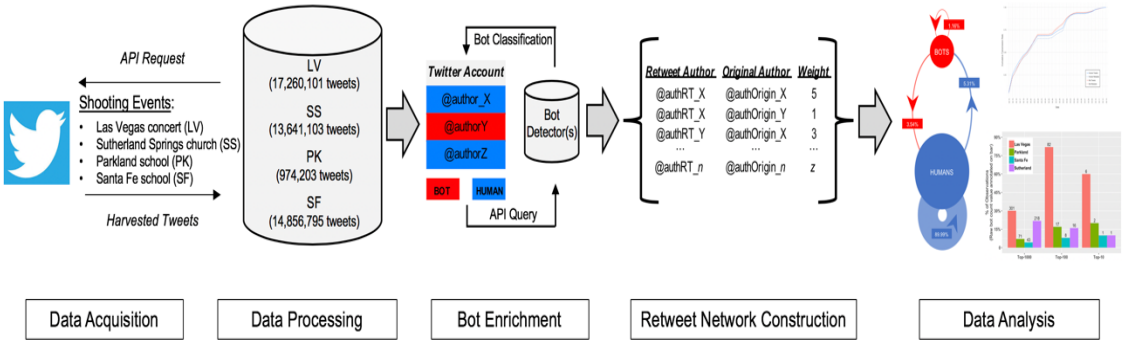


Figure 17: Overview of social bot analysis framework illustrating methodological steps taken to analyze social bots within online social network conversations involving mass shooting events from October 2017 through May 2018.

4.3.1. Data Acquisition and Processing

Four mass shooting events that took place within an eight-month period from October 2017 through May 2018 serve as the mass shooting use-cases analyzed in this study. While additional shooting events meeting the generally accepted mass shooting threshold of at least four or more deaths in a single event (Dahmen et al., 2018; Silva & Capellan, 2019; Towers et al., 2015) occurred during this period, the analysis deliberately focus on events that resulted in 10 or more deaths, since total victim counts serve as the most salient predictor of increased media coverage (Schildkraut et al., 2018). Table 7 lists these events in chronological order along with additional pertinent details.

Table 7: Summary of mass shooting events resulting in more than 10 deaths from October 1, 2017 through May 18, 2018.

EVENT	DATE	KILLED	INJURED	DESCRIPTION
Las Vegas Concert Shooting (Las Vegas, Nevada)	Oct. 1, 2017	59	418	[1] On October 1, 2017, gunman Stephen Paddock opened fire from his Las Vegas Mandalay Bay hotel room onto the Route 91 Country Musical Festival crowds killing 58 and himself, while injuring 418.
Sutherland Springs Church Shooting (Sutherland Springs, Texas)	Nov. 5, 2017	27	20	[2] On November 5, 2017, gunman Devin Kelley fired into the congregation of the Sutherland Springs First Baptist Church, killing 26 and himself, while injuring 20.
Parkland School Shooting (Parkland, Florida)	Feb. 14, 2018	17	17	[3] On February 14, 2018, gunman Nikolas Cruz opened fire within the Marjory Stoneman Douglas High School in Parkland killing 17, while injuring 17.
Santa Fe School Shooting (Santa Fe, Texas)	May 18, 2018	10	13	[4] On May 18, 2018, gunman Dimitrios Pagourtzis opened fire within the Santa Fe High School in Santa Fe killing 10, while injuring 13.

https://www.nytimes.com/interactive/2018/10/01/us/las-vegas-shooting-victims.html
https://www.cnn.com/2017/11/05/us/texas-church-shooting/index.html
https://www.cbsnews.com/feature/parkland-florida-school-shooting/
https://www.cnn.com/us/live-news/santa-fe-texas-shooting

To derive the associated OSN conversations, the study examined harvested streaming tweets from the available Twitter public application programming interface (API) for a one-month period (28 days) following the date of each mass shooting event. In an effort to maintain a consistent collection paradigm for each event, the collection effort used Twitter API request parameters which relied on the same filter keywords: *shooting*, *shot*, *shots*, *gunman*, *gunfire*, *shooter* and *activeshooter*. The resulting corpus harvest for all four mass shooting events returned approximately 46.7 million tweets produced by approximately 13.6 million unique Twitter users. Table 8 provides volume metrics for each mass shooting event. We should note that the overall Parkland collection effort, although it employed the same search parameters and method, returned a substantially smaller total tweet volume due to the fact that the collection took place in a different storage and compute environment with different resource constraints.

4.3.2. Bot Enrichment

The open-source DeBot bot detection platform (Chavoshi et al., 2016), which enables retrospective analysis of historically detected bots through an archival repository, served as the primary source for labeling likely social bot accounts within the harvested tweet corpus. This is because the historical dates of the tweets were beyond the temporal classification constraint of the Botometer open-source bot detection platform (Davis et al., 2016). DeBot has proven to classify bots at extremely high precision rates in relation to other social bot detection efforts, including Twitter (Chavoshi et al., 2016; Chavoshi et al., 2017). While DeBot classifies bots with great precision, one should acknowledge, in agreement with Morstatter et al. (2016), that such high precision comes at a cost to recall performance.

The identification and labeling of social bot accounts followed a three-step process. First, the unique Twitter account name and numeric identification for each account present in the mass shooting event corpus are submitted for classification to the DeBot API. Next, DeBot returns a binary (True or False) bot classification for each user account. Finally, the DeBot classification results are merged with the existing corpus data by creating a true or false bot attribute for each account. In total, DeBot classified fewer than 1% of all corpus tweet account users, or contributors, as likely social bots responsible for producing ~1.63 million tweets and ~1.40 million retweets, or 3.49% and 3.91% of the tweets and retweets in the corpus, respectively. Table 8 provides applicable social bot volume details for each of the OSN conversations.

Table 8: Overall tweet corpus volumes and suspected social bot contributions for each associated OSN mass shooting event conversation.

CORPUS	COLLECTION DATES	TWEETS	RETWEETS	CONTRIBUTORS
Las Vegas	Oct 1 – Oct 28, 2017	17,260,101	13,258,233	2,925,808
Bot source (% of total)		*719,509 (4.17%)*	*622,620 (4.70%)*	*39,956 (1.37%)*
Sutherland Springs	Nov 5 – Dec 3, 2017	13,641,103	10,095,006	4,996,779
Bot source(% of total)		*491,214 (3.60%)*	*418,775 (4.15%)*	*34,497 (0.69%)*
Parkland	Feb 14 – Mar 13, 2018	974,203	802,227	425,941
Bot source (% of total)		*52,207 (5.36%)*	*45,508 (5.67%)*	*8,441 (1.98%)*
Santa Fe	May 18 – Jun 14, 2018	14,856,795	11,688,269	5,262,635
Bot source (% of total)		*367,200 (2.47%)*	*316,374 (2.71%)*	*28,072 (0.53%)*

4.3.3. Retweet Network Construction

The deliberate act of retweeting has been viewed as an artifact demonstrating a particular Twitter user's propensity to share information or attempt to engage in direct conversation with other users (Boyd et al., 2010). Retweets accounted for 76.7% of the total mass shooting corpus, with approximately 35.8 million tweets identified as retweets. This overall high density of retweets permeates across each individual OSN event conversation, with retweet densities of 76.8%, 74.0%, 82.3% and 78.7% for the Las Vegas, Sutherland Springs, Parkland and Santa Fe shooting conversations, respectively. A retweet between two Twitter users (i.e. nodes) is an observable conversational activity that can be viewed as a directed connection, or edge, within a network construct. For example, one can assign a directed edge weight value of '1' for an initial retweet between two users and increment previously established edges by '1' for each subsequent directional retweet between the same two user nodes.

By iterating through each retweet in the corpus, the retweet network construction process ultimately created a social network graph of each OSN mass shooting conversation. The transformation of conversations into network graph objects enables the application of an array of social network analysis (SNA) methods, which is detailed in the subsequent Data Analysis Methods section. Overall, the OSN retweet conversations produced directed networks with the following node-edge characteristics: 4,926,906 nodes / 11,864,672 edges (Las Vegas), 4,105,206 nodes / 8,987,800 edges (Sutherland Springs), 382,797 nodes / 751,255 edges (Parkland) and 5,264,937 nodes / 13,133,371 edges (Santa Fe).

4.3.4. Data Analysis Methods

The following subsections provide a comprehensive introduction to the specific methods employed to comparatively analyze the evidence of suspected social bots across the OSN mass shooting conversations of interest. Each subsection describes the fundamental data requirement for each analysis method, a detailed characterization of the analysis method and any pertinent theoretical underpinnings. The combined effort of these Data Analysis Methods (Section 4.3.4) subsections provides necessary interpretative context to the presented findings in the subsequent Results and Discussion section (Section 4.4).

4.3.4.1. Conversation Participation Rate Analysis

To determine any potential contribution patterns of social bot accounts in comparison to regular human accounts, this section examined the cumulative tweet and retweet rates over the course of each observed online mass shooting conversation. The analysis accomplished this by bifurcating each mass shooting event corpus into separate

human and bot activity, then temporally indexing the contribution activity for each of these subsets. This allowed for a quantifiable and visual comparative analysis between human and bot temporal conversation contributions. The Conversation Contribution Inversion subsection present the results of this comparative analysis and discusses the effects of bots as potential narrative drivers within the Results and Discussion section.

4.3.4.2. Analysis of Subsequent Mention of Previous Mass Shooting Events

Many traditional media studies have focused on comparatively analyzing media attention between highly salient mass shooting events (e.g. Muschert & Carr, 2006; Schildkraut et al., 2018; Schildkraut & Muschert, 2014). To extend a comparative framework perspective to multiple OSN mass shooting conversations, this section sought to determine the sustained attention paid to previous mass shooting events in subsequent mass shooting events. The analysis accomplished this by observing the explicit mention rates of keywords associated with past events in subsequent events from both the human and suspected bot perspective. As previously explained, the tweet collection effort maintained a consistent collection paradigm by using the same filter keywords to harvest the overall conversations for each mass shooting event but required event-specific keywords to determine specific mentions within other event conversations. The most common descriptive words within the corpus for each mass shooting event served as the emergent keywords to tally subsequent mentions in other events. The resulting mention keywords followed an "*event name* / *shooter name*" paradigm, as the most common distinguishable words for each previous event included a derivation of the specific event (Las Vegas {'*vegas*'}; Sutherland Springs {'*sutherland*'}; Parkland {'*parkland*'}) and reference to the identified shooter of each previous event (Las Vegas {'*paddock*'};

Sutherland Springs {'*kelley*'}; Parkland {'*cruz*','*nikolas*'}). In the Parkland case, duplication issues arose that required the differentiation between the shooter, Nikolas Cruz, and the Texan politician, Ted Cruz. Therefore, a bigram consisting of the first name '*nikolas*' combined with '*cruz*' served for the mining mention counts within the corpus. Figure 18 provides a visual framework describing the process used for determining mention counts across events. The Previous Event Mention Rates subsection presents the normalized mention rate results and discusses the perceived implications of previous mass shooting event mentions by humans and bot accounts in the Results and Discussion section.

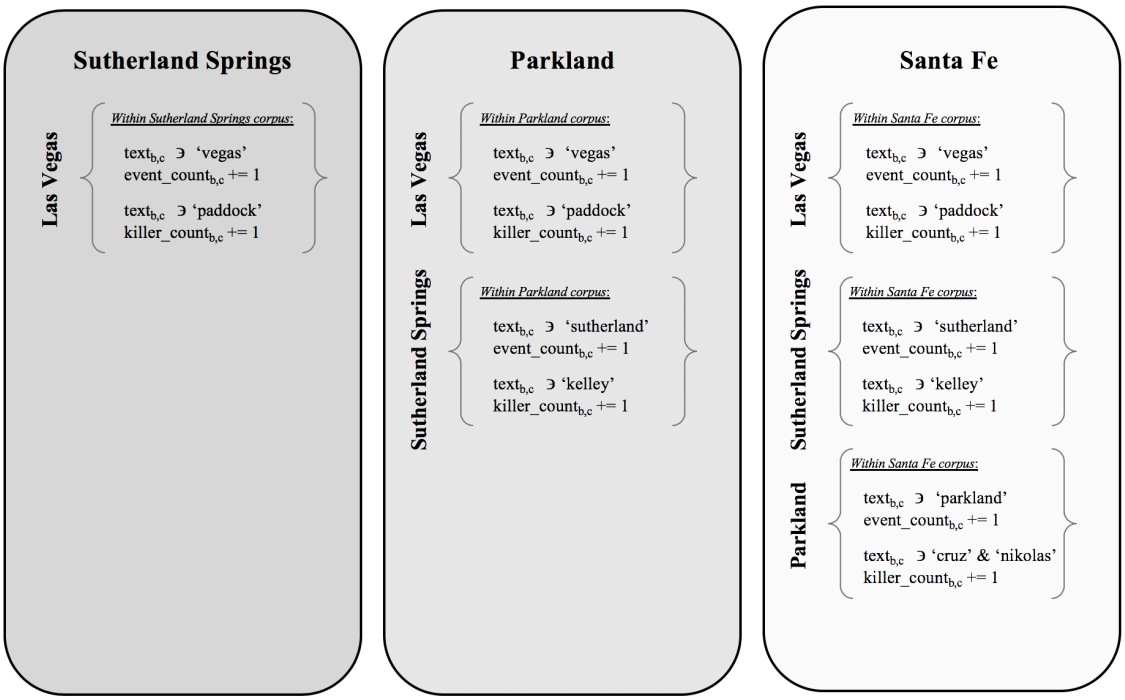


Figure 18: Mention count discovery of previous mass shooting events within subsequent online mass shooting event conversations. For example, within the Santa Fe conversation corpus, a previous event mention count by both humans and bots is determined by references to event location (i.e. 'vegas', 'sutherland', or 'parkland') and event shooter (i.e. 'paddock', 'kelley', or 'cruz & nikolas').

4.3.4.3. Intra-group and Cross-group Interaction Analysis

Having created retweet networks from the retweets of each OSN conversation, this section sought to determine if the interactions between (i.e. social bots retweeting humans or humans retweeting social bots) and among (i.e. social bots retweeting social bots or humans retweeting humans) the different accounts types (i.e. humans or bots) produced observable patterns. To do so, the analysis had to subset each mass shooting event network into separate retweet network edgelists representing each potential cross-group and intra-group network edge relationship (i.e. bot-retweets-bot, bot-retweets-human, human-retweets-bot, human-retweets-human). Table 9 presents the consolidated volumes associated with all derived edgelist relationships for each online mass shooting event. These edgelist volumes serve as the basis for the presented intra-group and cross-group conversation rate results detailed in Intra-group and Cross-group Conversation Patterns subsection of the Results and Discussion section.

Table 9: Retweet volumes of intra-group and cross-group conversation activity across all online mass shooting events.

	Intra-Group Retweet Volume		Cross-Group Retweet Volume	
Corpus	Bot-to-Bot	Human-to-Human	Human-to-Bot	Bot-to-Human
Las Vegas	153,468	11,931,552	704,064	469,152
Sutherland Springs	67,219	9,307,052	369,181	351,556
Parkland	4,552	711,027	45,692	40,956
Santa Fe	23,953	11,027,195	344,700	292,421

4.3.4.4. Relative Importance of Conversation Contributors through Centrality Analysis

This section takes further advantage of the graph construct of each derived retweet network through the application of SNA centrality measures. Centrality measures serve as a proxy to determine the relative importance of a node based on a given node's structural network position vis-a-vis other nodes (Wasserman & Faust, 1994). Riquelme and Gonzàlez-Cantergiani (2016) presented a comprehensive survey examining the wide variety of available centrality measurements that can be used to measure the relative influence of contributing users in online Twitter conversations. In an effort to determine the relative importance of social bot actors in relation to human actors, the author chose to calculate the following four network centrality measurements based on their recognizability and efficient scalability to large-scale networks for all nodes within each of the mass shooting retweet networks: in-degree centrality, out-degree centrality, eigenvector centrality and PageRank centrality.

In-degree and out-degree centrality are directional variants of degree centrality, which simply measures the total number of direct edges that a node shares with other nodes in a given network. In the context of a retweet network, in-degree centrality provides a cumulative inward activity tally of all inbound edges to a particular node, or the number of times a message from a particular Twitter account is retweeted by nodes in the network. The reverse is true of out-degree, as it measures all outward activity of a particular node, or the number of times a particular Twitter account initiates a retweet of other messages produced by nodes in the network. Measures of degree centrality can be viewed as a proxy for network popularity given the quantifiable number of direct connections, or conversation engagements. Eigenvector centrality, a more complex

derivation of degree centrality, queries the individual degree centrality of all nodes in a network and returns a weighted sum based on a particular node's set of direct and indirect edges (Bonacich, 2007). Given the completeness of the eigenvector calculation across an entire network, we can view eigenvector centrality as a measure of global network influence in a retweet network. Lastly, the PageRank centrality measurement, derived from eigenvector centrality, places a weighted premium on the degree value of nodes that initiate edges with other nodes of the most relative importance (Brin & Page, 1998). From a retweet network perspective, user accounts with higher PageRank valuation receive more retweets from the more popular user accounts in the retweet network. The subsection Relative Importance of Social Bots in Online Mass Shooting Conversations within the Results and Discussion section presents the formal centrality analysis results and discusses the overall density of social bots within the highest ranking centrality measures.

4.4. Results and Discussion

The following section presents and discusses the results of the applied methods described in the previous Data and Methods section (Section 4.3). Through the acquisition of a Twitter data corpus from multiple online mass shooting conversations and the identification of social bots within this corpus, this study was able to apply the previously described analysis methods to comparatively analyze social bot conversation participation in relation to human user behavior and presents those findings in the subsequent Conversation Contribution Inversion (Section 4.4.1) and the Previous Event Mention Rates (Section 4.4.2) subsections. Furthermore, through the application of SNA

techniques to the mass shooting event retweet networks, the Intra-group and Cross-group Conversational Patterns (Section 4.4.3) subsection presents the findings of directional conversation interactions and the Relative Importance of Social Bots in Online Mass Shooting Conversations (Section 4.4.4) subsection presents the centrality analysis rankings of bots in relation to humans.

4.4.1. Conversation Contribution Inversion

Figure 19 provides a consolidated visualization of the cumulative contribution profiles for human and suspected bot accounts over the course of each mass shooting conversation. The results show that the pace of cumulative tweet contributions from bots exceeds that of humans through the entire conversation timeframe for the Sutherland Springs (Figure 19b) and Santa Fe (Figure 19d) shootings. However, with the Las Vegas (Figure 19a) and Parkland (Figure 19c) shootings, an inversion occurs as bots begin to outpace humans after five and seven days (annotated as gray shaded areas), respectively, and continue on as the dominant contributor in terms of a normalized contribution rate. To provide a comparative benchmark between the observed human and bot contribution rates, a two-sample KS test (introduced in Chapter 2) was conducted between the bot and human distributions for each of the conversations. The KS test results return a D statistic value representing the maximum difference between the bot and human distributions, along with a corresponding p-value assessing whether the bot and human distributions came for the same overall distribution. The D statistic results showed a fairly similar closeness between bot and human contributions across all conversations with D statistic values of 0.107 ($p = 0.995$), 0.1786 ($p = 0.720$), 0.1786 ($p = 07205$) and 0.0714 ($p =$

1.000) for the Vegas, Sutherland Springs, Parkland and Santa Fe shootings, respectively. Further

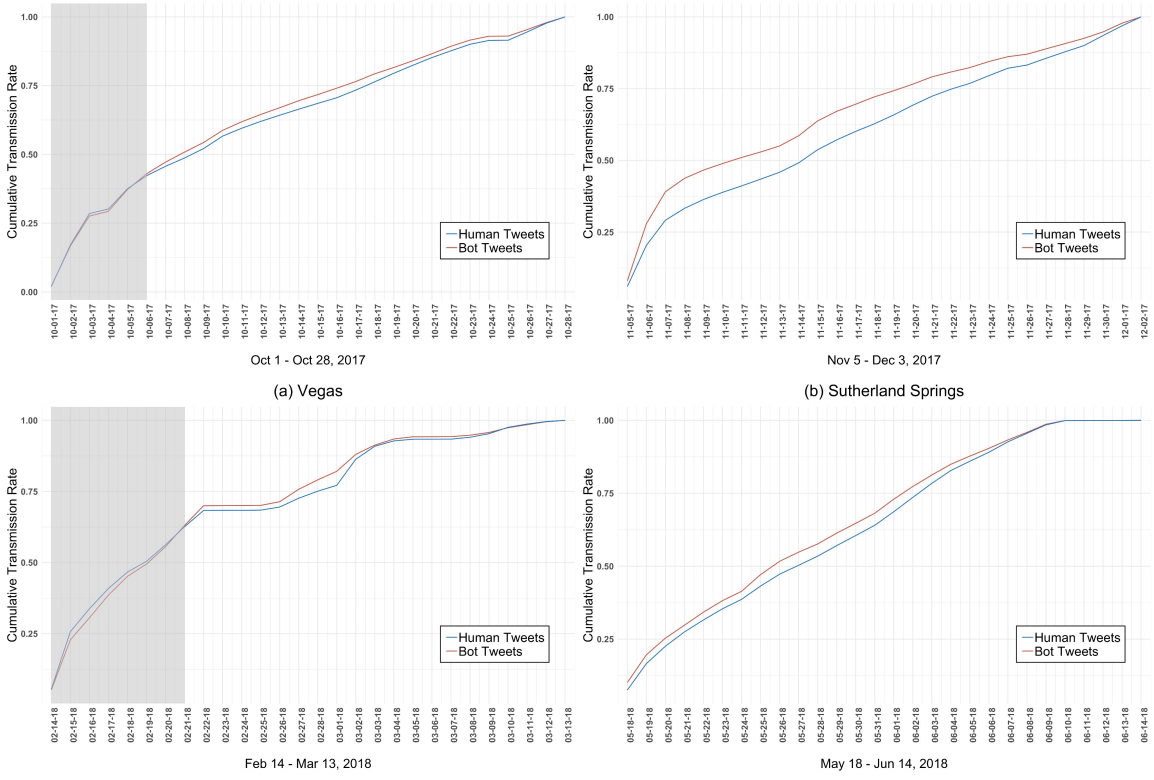


Figure 19: Cumulative tweet conversation contributions of both human (blue) and bot (red) accounts for the one-month online conversations of the following mass shooting events: (a) Las Vegas concert shooting (October 1 - 28, 2017), (b) Sutherland Springs church shooting (November 5 - December 3, 2017), (c) Parkland school shooting (February 14 - March 13, 2018), (d) Santa Fe school shooting (May 18 - June 14, 2018). Gray shaded areas depict human-led contribution rate periods.

Persistent human contribution latency in relation to bots for the entirety of the Sutherland Springs and Santa Fe conversations suggests that human accounts lacked general interest compared to bots for these particular events. In contrast, the Las Vegas and Parkland events draw immediate human interest, but this initial interest subsides as

sustained bot contribution rates bypass humans after less than a week. While further investigations are required to validate such claims, they are beyond the scope of this study. However, in discovering the clear contribution rate differences between bots and humans in the Sutherland Springs and Santa Fe events, and more interestingly, the contribution inversion in the Las Vegas and Parkland events, we can conclude that social bots are an explicit sub-population of actors in online mass shooting event conversations. Further, in the same light that Merry (2016) identified special interest groups as narrative framing agents in social media, social bots should be considered as a potential actor capable of framing online narratives.

4.4.2. Previous Event Mention Rates

By executing the mention count discovery process introduced and illustrated (Figure 19) in the Data Analysis Methods section (Section 4.4.1), we are able to ascertain the rates at which bots and humans mentioned previous events in the subsequent mass shooting events of this study's corpus. Within this paradigm, the observed historical event mention relationships are as follows: Las Vegas mentions within the Sutherland Springs, Parkland and Santa Fe conversations; Sutherland Springs mentions within the Parkland and Santa Fe conversations; Parkland mentions within the Santa Fe conversation. Table 10 presents the consolidated total mention volumes and mention rates of the Las Vegas, Sutherland Springs and Parkland mass shootings within the applicable subsequent mass shooting conversations. The author normalized the mention rates according to the unique bot and human populations within each conversation. The results show a clear partiality by both bots and humans towards mentioning associated event

names (i.e. '*vegas*', '*sutherland*', '*parkland*') in lieu of the identified shooter (i.e. '*paddock*', '*kelley*', '*cruz*') when discussing previous mass shooting events. Furthermore, there are no mentions of Devin Kelley, the Sutherland Springs gunman, in any other mass shooting event conversation. Finally, while Sutherland Springs struggles to garner any attention by the time of the Santa Fe conversation, the results display a drastic increase in both human and bot mention rates of Las Vegas. While there is little research available to properly classify these observable patterns from an OSN-specific view, one can look to traditional media studies to potentially contextualize these findings. For example, Levin and Wiest (2018) discovered that media consumers paid significantly more attention to shooting events when the narrative focused on courageous bystanders as opposed to a victim or killer, while Silva and Capellan (2019) presented an extensive overview of observable media attention patterns in mass shooting media research.

Table 10: Bot and human mention rates of previous mass shooting events in subsequent mass shooting conversations.

SUTHERLAND CONVERSATION				
Bots	34,497			
Humans	4,962,282			
			Mentions	Rate
	'vegas'	Bot	12,431	0.36035
		Human	165,927	0.03344
	'paddock'	Bot	1,031	0.02989
		Human	9,858	0.00199
			Mentions	**Rate**
Total Las Vegas Mentions		**Bot**	**13,462**	**0.39024**
		Human	**175,785**	**0.03542**

PARKLAND CONVERSATION				
Bots	8,441			
Humans	417,500			
			Mentions	Rate
	'vegas'	Bot	191	0.02263
		Human	2,362	0.00566
	'paddock'	Bot	0	0
		Human	0	0
			Mentions	**Rate**
Total Las Vegas Mentions		**Bot**	**191**	**0.02263**
		Human	**2,362**	**0.00566**
	'sutherland'	Bot	47	0.00557
		Human	674	0.00161
	'kelley'	Bot	0	0
		Human	0	0
			Mentions	**Rate**
Total Sutherland Mentions		**Bot**	**47**	**0.00557**
		Human	**674**	**0.00161**

SANTA FE CONVERSATION				
Bots	28,072			
Humans	5,234,563			
			Mentions	Rate
	'vegas'	Bot	3,860	0.13750
		Human	71,016	0.01357
	'paddock'	Bot	367	0.01307
		Human	2,723	0.00052
			Mentions	**Rate**
Total Las Vegas Mentions		**Bot**	**4,227**	**0.15058**
		Human	**73,739**	**0.01409**
	'sutherland'	Bot	226	0.00805
		Human	3,399	0.00065
	'kelley'	Bot	0	0
		Human	0	0
			Mentions	**Rate**
Total Sutherland Mentions		**Bot**	**226**	**0.00805**
		Human	**3,399**	**0.00065**
	'parkland'	Bot	8,334	0.29688
		Human	141,176	0.02697
	'cruz' & 'nikolas'	Bot	426	0.01518
		Human	5,664	0.00108
			Mentions	**Rate**
Total Parkland Mentions		**Bot**	**8,760**	**0.31205**
		Human	**146,840**	**0.02805**

4.4.3. Intra-group and Cross-group Conversation Patterns

Figure 20 presents an overall visualization of normalized intra-group (i.e. bots retweeting bots or humans retweeting humans) and cross-group (i.e. bots retweeting humans or humans retweeting bots) conversation patterns for each of the study's OSN conversations. As previously mentioned, the author normalized the intra-group and cross-group retweet volumes by the total retweet volume in each conversation. For example, the self-loops for humans and bots depicted in Figure 20 translate to 89.99% of all the retweets in the Las Vegas corpus resulting from human to human interaction, while bot-to-bot retweets only account for 1.16% of retweets. Additionally, the directed cross-group activity between bots and humans in the Las Vegas conversation shows that humans retweeting bots comprises 5.31% of retweets, while bots retweeting humans comprises 3.54% of retweets. In general, we see human-to-human interaction as the dominant

relationship across all of the mass shooting conversations. Moreover, human users retweet bots at a higher rate than bots retweet humans in each of the conversations, which demonstrates that humans are more responsible for spreading bot-generated content than bots themselves in each of the mass shooting conversations. This is an interesting finding, as previous social bot analysis has found bots to be more, on average, hyper-social than humans: they attempt to engage humans at persistently higher rates in retweet networks associated with election, conflict and political Twitter conversations as opposed to average human accounts (Schuchard et al., 2019; Stella et al., 2018).

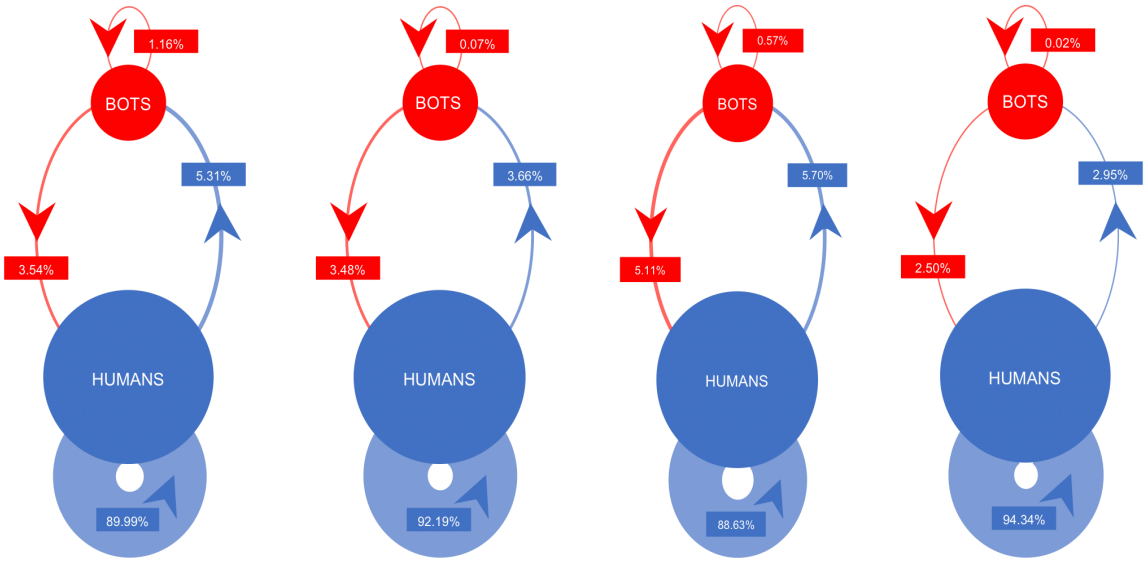


Figure 20: Intra-group and cross-group retweet interaction rates among and between human (blue) and suspected social bot (red) user accounts for a one-month period following the (a) Las Vegas, (b) Sutherland Springs, (c) Parkland and (d) Santa Fe shooting events.

4.4.4. Relative Importance of Social Bots in Online Mass Shooting Conversations

The results of the centrality measurement ranking analysis, introduced in the Data and Methods section (Section 4.3), showed that many social bots, while accounting for just 0.82% of all contributors in this study's corpus, displayed structural network importance by achieving high centrality ranking positions, especially in the eigenvector and out-degree centrality rankings. Figure 21 presents the consolidated centrality results, depicting the density of social bot accounts falling within the top-N, where N = 1000 / 100 / 10, eigenvector, in-degree, out-degree and PageRank centrality rankings for each online mass shooting conversation. The out-degree ranking persistence shows the hyper-social nature of these particular social bots across all conversations. More interestingly, social bots display exceedingly high eigenvector centrality valuations in the Las Vegas conversation, accounting for 82% and 60% of the top-100 and top-10 rankings, respectively. While also earning numerous high rankings across the other conversations, but not nearly as dominant. Given that eigenvector centrality serves as a potential proxy for total network influence, social bots could be construed as the most structurally influential nodes in the Las Vegas mass shooting conversation.

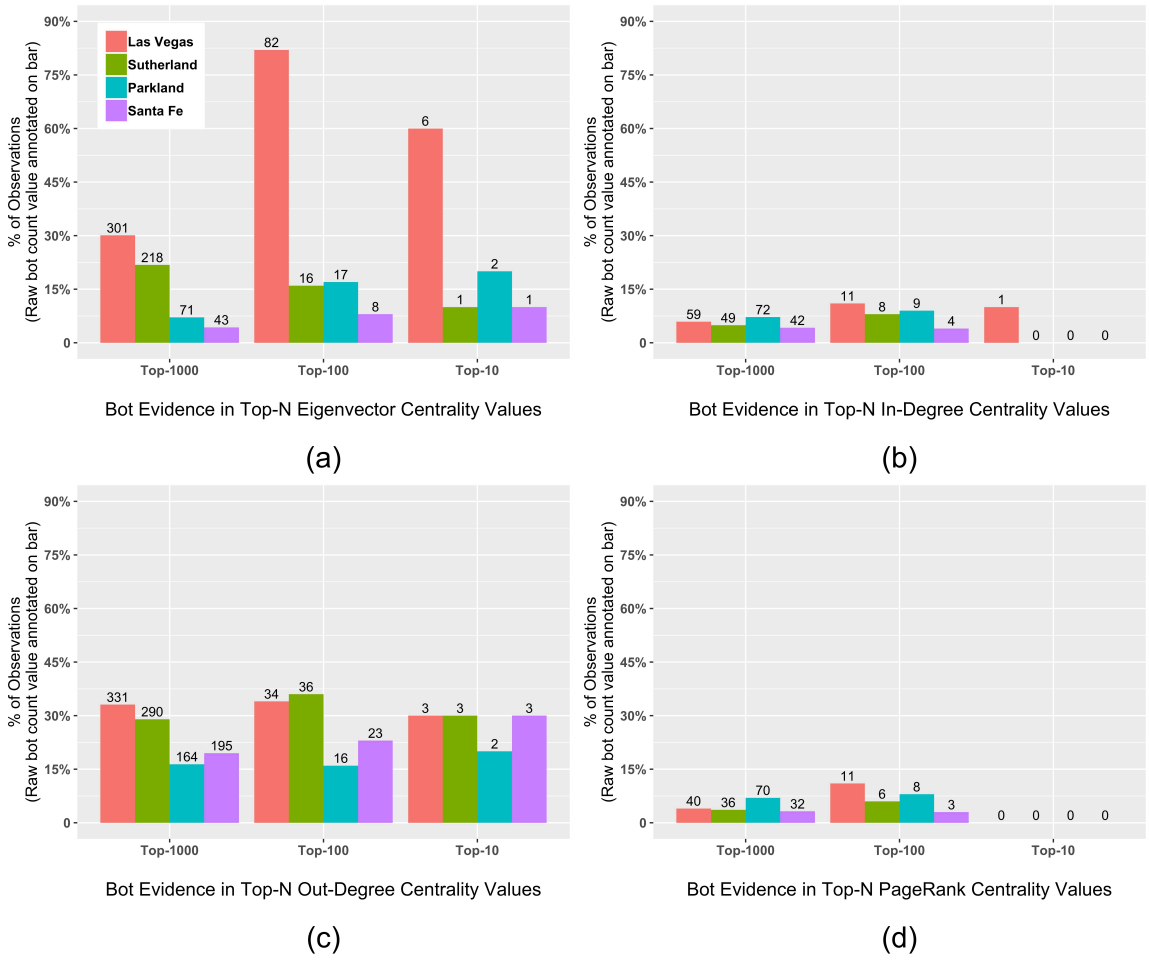


Figure 21: Social bot accounts in the top-*N*, *where N = 1000/100/10*, (a) eigenvector, (b) in-degree, (c) out-degree and (d) PageRank centrality measurement rankings within OSN mass shooting retweet networks discussing the Las Vegas (red), Sutherland Springs (green), Parkland (blue) and Santa Fe (purple) shooting events.

4.5. Conclusion

This chapter examined the presence, contribution patterns and relative importance of suspected social bots within four different online mass shooting conversations. By following a mixed methodology process focused on the normalization and fusion of associated Twitter conversation data with social bot detection results, this study presented a repeatable and agnostic process that can be extended to evaluate additional online mass shooting use cases of interest. While analyzing the cumulative contribution patterns of

both humans and bots, this study found that social bot accounts outpaced human contributions throughout the entirety of the Sutherland Springs and Santa Fe shooting conversations, while a reversal took place with the Las Vegas and Parkland cases as human accounts ceded initial rate dominance to bots after less than one week. Both bots and humans displayed a strong tendency to mention previous shooting events by referencing event locations as opposed to shooters themselves. The construction of retweet networks allowed us to observe the intra-group and cross-group engagements showing humans engaging bot accounts at a higher rate than bots engaging humans in all four conversation retweet networks. Moreover, the retweet network graph construct enabled the application of SNA centrality measurements to investigate the relative importance of social bots, which showed large populations of bots ranking prominently in overall eigenvector and out-degree centrality across all conversations. In the Las Vegas mass shooting conversation, social bots dominated the eigenvector centrality ranking results, accounting for 82% and 60% of the top-100 and top-10 accounts, respectively.

This study is not immune from limitations. First, previous traditional media research efforts have shown significant bias in mass shooting media coverage based on event factors such as the race/ethnicity of the both the shooter and the victims, as well as the number of associated casualties (Duxbury et al., 2018; Schildkraut et al., 2018). Guggenheim et al. (2015) described the reciprocal relationship between traditional and social media discussing mass shooting events and one can assume the transference of this coverage bias. In other works, Tufecki (2014) points to well-known inherent biases associated with data emanating from OSNs, to include sampling and representativeness.

In addition, Ruths and Pfeffer (2014) echoed a similar critique of OSN data, while also stressing the inability of OSN providers to prevent or limit the distortion of social bot actors.

This study serves as a compelling first step forward in providing the social bot analysis research necessary to identify and distinguish automated social bot contributions from intentional human dialogue in mass shooting OSNs. To date, it is the most comprehensive social bot analysis involving OSN mass shooting conversations. Given the evidence of contagion in the aftermath of mass shootings (Towers et al., 2015), it is essential to detect and prevent the potential amplification or glorification of such events by social bots. While social bot analysis is still in a nascent state and bot detection methodologies continue to evolve to account for the growing sophistication of bot developers (Cresci et al., 2017; Subrahmanian et al., 2016), this work provides a repeatable framework that is extendable to other OSN conversations and additional bot detection platforms. Finally, it provides requisite feedback to bot detection algorithm developers on associated detection performance against an array of different OSN conversations.

CHAPTER 5. BOTS IN ELECTIONS: AN ENSEMBLE BOT DETECTION COVERAGE FRAMEWORK

5.1. Introduction

The 2016 U.S. presidential election broke traditional campaign communication norms, as legacy institutions such as mainstream media sources (e.g. print, television and radio) and political-party organizations ceded much power and influence to unmediated, Internet-based technological platforms (e.g. online social networks (OSNs), online political blogs) (Persily, 2017). Previously, Gibson and Cantijoch (2013) identified the increasing active participatory nature of political engagement in OSNs and described such behavior as a new type of expressive political engagement. Since the 2016 U.S. election, OSNs have surpassed print newspapers as a primary news source and continue to gain traction in relation to television and radio sources (Mitchell, 2018). While the rapid rise of OSN platforms has reduced the barrier for individuals to actively participate in political dialogue, the relatively unsupervised nature of OSNs increases susceptibility to misinformation campaigns, especially with respect to political and election dialogue (Bovet & Makse, 2019; Grinberg et al., 2019; Howard et al., 2018).

Social bots—automated software agents designed to mimic or impersonate humans— are prevalent actors in OSN platforms and have proven to amplify misinformation by orders of magnitude (Lazer et al., 2018). While the original design or purpose of social bots is not always nefarious, their impact can directly lead to the

intentional or unintentional spreading of false narratives (Ferrara et al., 2016). The inability for humans to readily discern whether they are engaging in dialogue with a human is a newly intractable problem with unknown implications. The rapidly evolving social bot problem has led to the recent emergence of numerous research efforts dedicated to the development of novel bot detection algorithms (e.g. Beskow & Carley, 2018; Chavoshi et al., 2016; S. Cresci et al., 2018; Davis et al., 2016). Beyond detection algorithm development, introductory social bot analysis efforts have examined the prevalence and activities of detected social bots within general Twitter and Facebook conversations (e.g. Boshmaf et al., 2011; Mønsted et al., 2017; Shao et al., 2018). Further social bot analysis works have focused on detected bots within Twitter conversations involving specific topic areas such as the Brexit referendum (Duh et al., 2018; Howard & Kollanyi, 2016), stock market trading (Cresci et al., 2019), conflict (Schuchard et al., 2019) and political elections (Bessi & Ferrara, 2016; Boichak et al., 2018; Stella et al., 2018).

The constantly evolving sophistication of social bots has proven challenging for even the most promising detection algorithms developed to date (Cresci et al., 2017). The ever-expanding range of potential bot characteristics and activity patterns demands continual refinement to existing detection methods or the development of entirely new methods to account for the most sophisticated bots. In summarizing the array of different detection approaches, Jiang et al (2016) cautioned that detection applications, while looking to maximize the detection of the most 'suspicious' behaviors, employ different definitions of suspicious behaviors. In effect, the design parameters of bot detection

algorithms will return results to which the algorithms are trained, and, thus, different detection strategies should detect different types of social bots. Recent efforts have focused on the evolving nature of bots by introducing adversarial learning detection algorithms (Cresci et al., 2018, 2019a). While such detection advances are quite promising, they serve no immediate role in assisting broad, multidisciplinary social bot analysis efforts, since they are not readily accessible to the larger research community. Therefore, most current social bot analysis research efforts rely primarily upon an open-source bot detection platform service such as Botometer (Davis et al., 2016; Varol et al., 2017) or DeBot (Chavoshi et al., 2016).

As the results of the 2015 DARPA Twitter Bot Challenge summarized, no single detection algorithm is able to account for the myriad of social bots operating in OSNs (Subrahmanian et al., 2016). It is from this perspective that the following study expands current social bot analysis research by incorporating multiple social bot detection services to determine the prevalence and relative importance of social bots within an OSN conversation. Through the lens of the 2018 U.S. midterm elections, harvested tweets capturing the election conversation are analyzed for evidence of bots using three bot detection platform services: Botometer (Varol et al., 2017), DeBot (Chavoshi et al., 2016) and Bot-hunter (Beskow & Carley, 2018). The resulting suspected bot evidence serves as the basis for an ensemble of applied social network analysis (SNA) methods to determine the relative structural importance of bots in the conversation. Finally, a comprehensive bot detection coverage analysis evaluates the resulting overlap in performance among the employed bot detection services.

The results of this study show that bot and human accounts contributed temporally to the 43.5 million tweet election corpus at relatively similar cumulative rates. The multi-detection platform comparative analysis of intra-group and cross-group interactions shows that bots detected by DeBot and Bot-hunter persistently engaged humans at rates much higher than bots detected by Botometer. Furthermore, while bots accounted for less than 8% of all unique accounts in the election conversation retweet network, bots accounted for more than 20% of the top-100 and top-25 ranking out-degree centrality, thus suggesting persistent activity to engage with human accounts. Finally, the bot coverage overlap analysis shows that there existed minimal overlap among the bots detected by the three bot detection platforms, with only eight total bot accounts detected by all.

The intra-group and cross-group analysis of the constructed retweet network shows that bots detected by DeBot and Bot-hunter persistently engaged humans at rates much higher than bots detected by Botometer. In addition, the intra-group and cross-group interactions, when viewed from a consolidated bot account perspective, provide the first piece of evidence that minimal overall overlap existed between the set of bots detected by each detection platform. The centrality ranking results showed that bots, from an overall perspective, achieved large volumes of high centrality ranking positions despite their relatively small population size. The classification of relative importance by social bot accounts was most noticeable with bots detected by DeBot in the out-degree rankings and with bots detected by Botometer in the eigenvector rankings. Analysis of the overlap of bots detected by the detection platforms showed that no overlap existed

between the bots ranking in the top-50 centrality results. Moreover, the Jaccard similarity index showed little bot detection overlap from a pairwise perspective, while only eight bots out of a total of 254,492 unique bots in the overall tweet corpus were detected by all three detection platforms.

The roadmap of this chapter is as follows. The Background section (Section 5.2) provides the necessary context for this study by introducing applicable previous works involving social bot detection and analysis. Next, the Data and Methods section (Section 5.3) details the specific data acquisition and processing, as well as the applied methods, used in this study. The Results and Discussion section (Section 5.4) presents the pertinent findings of the study, and the chapter closes with the Conclusion section (Section 5.5).

5.2. Background

OSN research has emerged and evolved rapidly in concert with the global adoption of social media platforms throughout the past decade. While the limitations, biases and risks associated with using OSN data are widely discussed (Ruths & Pfeffer, 2014; Tufekci, 2014), there have been many positive insights gained from OSN research contributions. Such works include OSN-findings related to disaster event detection (Crooks et al., 2013; Sakaki et al., 2013), suicide prevention and detection (Luxton et al., 2012; Won et al., 2013) and cyberbullying (Hamm et al., 2015; Whittaker & Kowalski, 2015). OSNs have even been described as transformational media in creating new avenues of political participation and dialogue (Persily, 2017; Theocharis & Deth, 2018). In a 61-million person Facebook experiment during the 2010 U.S. congressional elections, Bond et al. (2012) showed how social human ties are instrumental in spreading

both online and offline political behavior. Vaccari et al. (2015) identified that lower-threshold political engagement activities in OSNs, such as posting political views, are strongly associated with higher-threshold activities such as campaigning for particular parties/candidates and attending offline political events. In a survey of active political Twitter users, Bode and Dalrymple (2016) discovered that a primary reason for engaging in political discourse on Twitter was due to a general lack of trust in mainstream media sources.

The increasing use of OSNs for political communication dialogue has led to the rightful criticism of the transparency and validity not only behind the how social media platforms operationally promote certain narratives, but also of how the platforms verify accounts as humans or bots (Woolley & Howard, 2016). Not surprisingly, given the propensity for polarization and the observed emergence of echo chambers within political conversations in OSNs (Conover et al., 2011), social bot campaigns view the manipulation of political dialogue as a natural attack vector. With the emergent role of OSNs in the 2016 U.S. presidential election, as previously mentioned, recent social bot analysis efforts have expanded their focus greatly into political OSN conversations. These works include the examination of detected bots within the 2016 U.S. presidential election (Bessi & Ferrara, 2016; Boichak et al., 2018; Howard et al., 2018), the UK-EU Brexit referendum (Duh et al., 2018; Howard & Kollanyi, 2016), the 2018 Italian general election (Stella et al., 2018) and the 2017 Catalan referendum (Stella et al., 2018). While these election-focused social bot analyses relied upon an assortment of bot detection algorithms, they all used a single method to classify bots. This study significantly

expands this body work by aggregating the classification results of three bot detection platforms in an effort to provide a more holistic social bot analysis framework. The following introduces and highlights the three detection platform services employed in this study to classify bots within the 2018 U.S. midterm Twitter conversation.

Botometer[6], a widely used open-source bot detection platform created by researchers at Indiana University, is based on a supervised Random Forest ensemble classification technique that evaluates more than 1,000 extracted features for each analyzed Twitter account (Davis et al., 2016; Varol et al., 2017). Given the supervised nature of the underlying algorithm, Botometer requires and has updated its detection classification algorithm multiple times by retraining against new data (Varol et al., 2017; Yang et al., 2019). Botometer ultimately provides a likelihood estimate score on a [0,1] scale that an account is a bot, with simple bots scoring (0.8 – 1.0) and more sophisticated (i.e. human-like) bots scoring (0.5 – 0.7) (Varol et al., 2017). While popular, Botometer is limited by several significant factors, which have been thoroughly documented in previous works (Ferrara, 2017; Stella et al., 2018; Stukal et al., 2017). These limiting factors include an inability to retrospectively analyze historical tweets and to classify suspended/protected Twitter accounts, while its publicly available application programming interface (API) does not support large-scale analyses given inherited Twitter API rate limits.

[6] Botometer is accessible at https://botometer.iuni.iu.edu/.

DeBot[7], an open-source bot detection platform developed by researchers at the University of New Mexico, adopts an unsupervised warped correlation method to detect and label as bots those Twitter accounts having more than 40 synchronous events in a given window of time (Chavoshi et al., 2016). The DeBot binary classification scheme detects bots with high precision, but it does so at a cost of total recall due to the limited sample size of overall Twitter accounts it evaluates (Chavoshi et al., 2017). While limited in coverage and susceptible to the precision/recall tradeoff of bot detection highlighted by Morstatter et al. (2016), historical DeBot results are easily accessible and have produced relevant results in social bot analyses (e.g. Kušen & Strembeck, 2018; Schuchard et al., 2019).

Finally, Bot-hunter, a newer bot detection platform developed by researchers at Carnegie Mellon University (CMU), applies a supervised Random Forest classification method to previously extracted Twitter data in a multi-tiered fashion with successive tiers incurring higher computational costs (Beskow & Carley, 2018). This deliberate tiered approach overcomes the limitations observed with Botometer (i.e. scalability and the classification of suspended accounts) by allowing bot classification to occur locally and against historical tweets, as opposed to classification in coordination with the Twitter API. In a similar fashion to Botometer, Bot-hunter returns a bot classification score for each Twitter account of interest on a normalized scale between 0 and 1. While Bot-hunter is not currently accessible via a public API, it was made available to this study upon request by the CMU research team.

---

[7] DeBot is accessible at https://www.cs.unm.edu/~chavoshi/debot/.

5.3. Data and Methods

This study breaks new ground in its use of multiple bot detection platforms to identify and analyze the presence of social bots within the 2018 U.S. midterm election OSN conversation. The following section details the study's overall methodological framework as depicted in Figure 22. First, Twitter Data (Section 5.3.1) provides the essential background describing the capture, storage and processing stages required to develop the election midterm tweet corpus. Bot Enrichment (Section 5.3.2) details the steps taken to label the accounts within the election corpus with the three chosen bot detection platforms. Retweet Network Construction (Section 5.3.3) explains the process to derive a network structure out of the original election conversation corpus. The section concludes with Bot Analysis (Section 5.3.4), which introduces the applied analysis methods used in the remainder of the study.

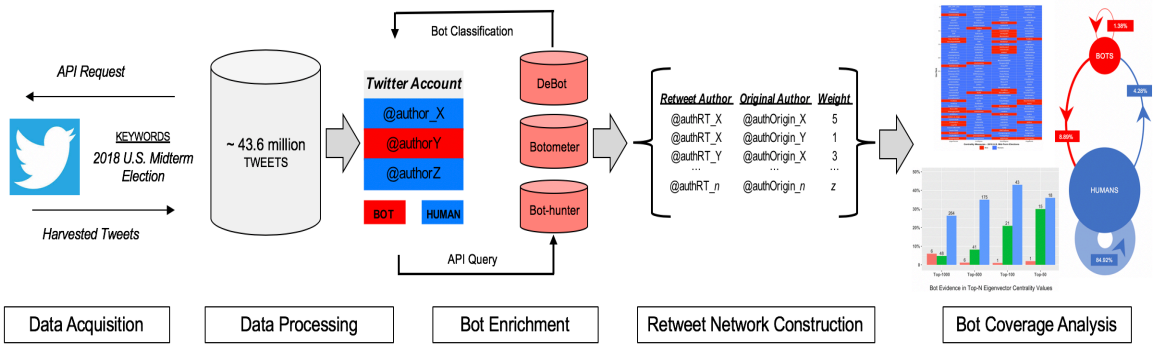


Figure 22: Social bot analysis framework employing multiple bot detection platforms. The framework enables the application of ensemble analysis methods to determine the prevalence and relative importance of social bots within OSN conversations discussing the 2018 U.S. midterm elections.

5.3.1. Twitter Data

The 2018 U.S. midterm elections provided a new opportunity to build upon previous social bot analyses dedicated to examining the role of bots within OSN election conversations. Given the specific limitations of bot detection platforms as described in the Background section (Section 5.2), it was essential to properly prepare a collection plan well in advance of the planned 30-day collection window leading up to election day (November 6, 2018). As Zhang et al. (2018) asserts, keyword selection in social media studies can induce varying levels of selection bias. To mitigate this risk, this study chose the comprehensive panel of keywords shown in Table 11 to capture the 2018 midterm election corpus. This panel included generic keywords associated with the election (e.g. Election2018, midterms2018) as well as keywords referencing campaign phrases and high-profile races in order to account for both major U.S. political parties.

The tweet collection process consisted of submitting the keyword panel to the publicly available Twitter standard streaming API for four weeks prior to the election day (October 10 thru November 6, 2018). The overall tweet collection process yielded a consolidated corpus consisting in excess of 43.5 million tweets produced by approximately 3.2 million unique accounts. Retweets accounted for approximately 83.2% of the tweet corpus with more than 36.2 million retweets produced by more than 2.3 million unique accounts. Due to the large volume of harvested tweets and the subsequent data processing requirements as detailed in the remainder of this section, all immediate data processing and storage took place in a scalable 16vCPU and 64GB RAM Amazon Web Services (AWS) m5a.4xlarge instance.

Table 11: Election-related keywords submitted to capture relevant tweets associated with the 2018 U.S. midterm elections via the Twitter API.

Generic Election		Campaign Phrases		Key Races
Election2018	midterms	BlueWave	RedWave	@ScottWalker
midterms2018	2018midterms	FlipTheSenate	maga	@WISuptTonyEvers
democrat	republican	FlipTheHouse	kag	@tedcruz
DNC	RNC	VoteThemOut	buildthewall	@BetoORourke
DNC2018	RNC2018	HandsOffOurCare	takeitback	@SenatorHeitkamp
@TheDemocrats	@GOP			@KevinCramer
@SenateDems	@SenateGOP			@FLGovScott
@HouseDemocrats	@HouseGOP			@ SenBillNelson

5.3.2. Bot Enrichment

To detect and label social bots in the collected election conversation corpus, this study relied upon three bot detection platforms: Botometer, DeBot and Bot-hunter. While the Background section (Section 5.2) provided a general overview of these platforms and their underlying detection algorithms, the remainder of this subsection presents the technical details explaining how the study used each detection platform to detect and label bots within the election conversation corpus of tweets. First, a technical explanation describes the processing and environmental considerations associated with each platform. Next, given the scoring scales of Botometer and Bot-hunter, a scoring analysis explains the chosen cutoff threshold for labeling accounts as bots. Finally, an aggregate and specific detection platform perspective presents the bot detection results.

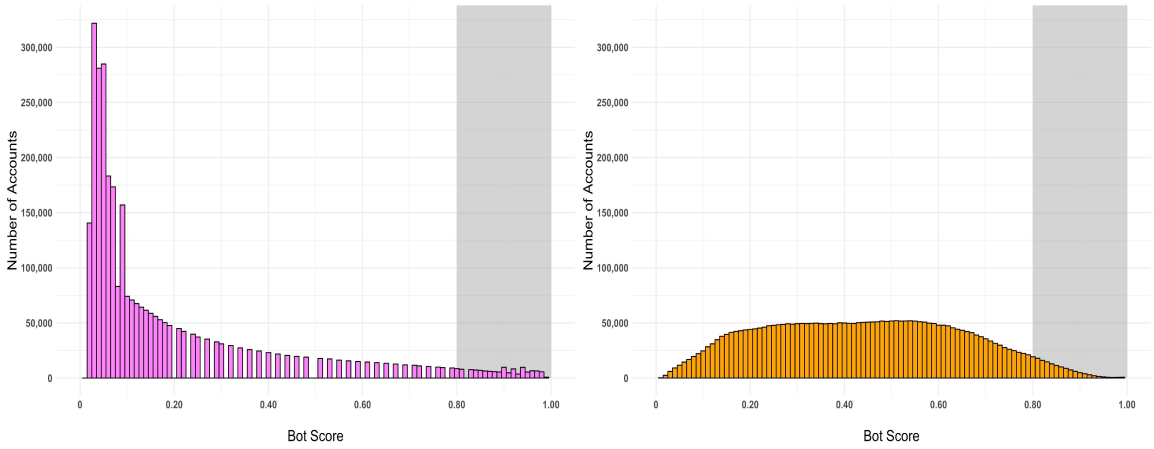
Currently, both DeBot and Botometer provide researchers open-source access to their hosted detection platforms via an API. However, due to individual API limitations, these two platforms required special access considerations to scale to the size of this

study's tweet corpus. Upon request, the DeBot development team provided access to the entire DeBot archival repository. The resulting detection processing simply consisted of matching unique tweet account information from the election conversation corpus to discovered bot profiles in the DeBot repository. The Botometer API[8] provides both an open-access free tier with a rate limit of 17,280 requests per day and a 'professional' paid tier, which aligns to the publicly available Twitter standard API rate limits, with a rate limit of 43,200 requests per day. Due to the size of the election corpus and Botometer's reliance on evaluating associated tweet data directly via the Twitter API, this study required three Botometer professional paid tier licenses in order to process the entire corpus volume in a timely manner. The faster execution tried to help mitigate Botometer's inability to process suspended or deleted accounts by evaluating accounts prior to their potential removal by Twitter. Bot-hunter does not currently provide a publicly available API, so the Bot-hunter team provided access to their platform upon request to process the raw tweets comprising the election conversation corpus.

Both Botometer and Bot-hunter return a classification score for each of the accounts they evaluate that falls within a [0,1] distribution, with a higher valuation constituting a greater likelihood that an account is a bot. DeBot, as previously mentioned, provides a simple binary classification for an account. Many studies using Botometer have historically used a 0.5 score threshold to classify bots (Badawy et al., 2018; Boichak et al., 2018; Shao et al., 2018). While a clear binary cutoff threshold is a challenging decision to make, platforms like Botometer are providing the necessary transparency for

[8] Botometer API information accessible at https://botometer.iuni.iu.edu/#!/api.

researchers to make an informed decision (Yang et al., 2019). This study used a highly conservative cutoff threshold of 0.80 to 1.00 to label accounts as detected bots, in a similar categorization paradigm of 'most likely' bots put forth by Broniatowski et al. (2018). This decision reflected a desire to determine the coverage overlap of the most certain bot accounts between different bot detection platforms. Figure 23 depicts the distribution of classification scores for both Botometer (Figure 23a) and Bot-hunter (Figure 23b), with the shaded gray areas highlighting the 0.80 to 1.00 score range.



(a) Botometer Score Distribution (b) Bot-hunter Score Distribution

Figure 23: Resulting distribution of scores for Twitter accounts present within the 2018 U.S. midterm election tweet corpus using the (a) Botometer (pink) and the (b) Bot-hunter (orange) bot detection platforms.

Table 12 provides a summary of the bot detection classification volume results across all three bot detection platforms, as well as an aggregate classification volume. The aggregate classification method labels an account as a bot if at least one of the bot detection results declares that account to be a bot. In total, the aggregate bot classification process labeled 254,492 unique accounts, or 7.95% of all accounts, as bots that were

responsible for contributing more than 5.7 million tweets (13.23% of all tweets) in the election corpus. From the specific detection platform perspective, Bot-hunter led all platforms by labeling 6.26% of all accounts as bots, followed by Botometer and DeBot with labeling rates of 3.80% and 0.64%, respectively. In terms of retweets, aggregate and specific platform bot labeling occurred at approximately that same rates; however, Botometer-labeled bot accounts retweeted at far lower rates in comparison to their regular tweet contribution rates.

Table 12: Twitter corpus volume and contributor populations from the 2018 U.S. midterm election OSN conversation with associated bot detection platform classification results.

Corpus	Detection Platform	Volume	% of Total	Contributors	% of Total
Tweets		43,565,164		3,201,996	
Humans		*37,800,157*	*86.77%*	*2,947,504*	*92.05%*
Bots		*5,765,007*	*13.23%*	*254,492*	*7.95%*
	DeBot	2,201,858	5.05%	20,605	0.64%
	Botometer	4,239,870	9.73%	121,780	3.80%
	Bot-hunter	2,729,354	6.26%	130,553	4.08%
Retweets		36,264,206		2,588,956	
Humans		*31,242,038*	*86.15%*	*2,388,447*	*92.26%*
Bots		*5,022,168*	*13.85%*	*200,509*	*7.74%*
	DeBot	1,991,654	5.49%	19,466	0.75%
	Botometer	920,675	2.54%	87,590	3.38%
	Bot-hunter	2,337,760	6.45%	107,861	4.17%

5.3.3. Retweet Network Construction

A retweet serves as an observable interaction within a Twitter conversation that has been shown to promote trust (Metaxas et al., 2015) and increase engagement between users (Boyd et al., 2010). This study focused on retweets as the primary interaction of interest between accounts within the election conversation corpus. By extracting the

directional nature of a retweet between two accounts, a logical node-edge paradigm emerges that can lead to the construction of an overall retweet network. For example, an initial retweet between two accounts receives a directional edge weight of '1' and the edge weight increases by '1' for each subsequent directional retweet between the same two accounts. Overall, the election corpus produced a retweet network, which served as the inherent graph object to enable the application of the SNA techniques described later in this study, consisting of 2,820,898 nodes and 24,511,110 edges.

5.3.4. Bot Analysis Methods

The following subsections introduce the specific analytic methods used to determine the prevalence, characteristics and relative importance of detected bots within the 2018 U.S. midterm election conversation corpus. Each method accounted for bots from an aggregate labeling perspective, as well as for each bot detection platform. The description for each associated analysis method includes the specific data requirement and any theoretical references necessary to enable the most interpretive context of results presented in the Results and Discussion section (Section 5.4).

5.3.4.1. Contribution Rate Analysis

Comparatively analyzing the temporal contribution patterns of bots and humans over time provided an opportunity to directly observe potential behavioral differences between the two sub-populations. Furthermore, this comparative context applied to differentiating the contribution patterns of bots detected by the various detection platforms used in this study. To accomplish this analysis, the entire election tweet corpus was divided into aggregate bot and human sub-populations. The resulting bot and human tweet contribution activities were then temporally indexed, resulting in a daily

contribution rate. This same process was extended to the individual detection platform bot classification results. The Results and Discussion section (Section 5.4) presents the consolidated findings of the cumulative contribution rate analysis.

5.3.4.2. Intra-group and Cross-group Participation Analysis

The constructed retweet network of the election conversation corpus enabled the observation of a multitude of communication interactions between bot and human accounts. These specific interactions can be reduced to intra-group (i.e. bots retweeting bots or humans retweeting humans) or cross-group (i.e. bots retweeting humans or humans retweeting bots) communication. To quantify the intra-group and cross-group communication volumes, applicable edgelists were created for each potential interaction. This included edgelists capturing the aggregate bot and human population interactions, as well as bot and human populations resulting from the individual bot detection platform results. These edgelists served as the foundational data source used to construct the visualization and associated results narrative presented in the Results and Discussion section (Section 5.4).

5.3.4.3. Centrality Ranking and Bot Coverage Analysis

Beyond the examination of prevalence and behavioral characteristics, it is reasonable to attempt to ascertain whether social bots can be construed as 'important' actors within an OSN conversation. SNA centrality measures provide an efficient means to make such an assessment. Centrality measures can imply relative node importance based on a given node's structural position in relation to other nodes within a network (Wasserman & Faust, 1994). Social media research includes numerous applications of centrality analysis to determine the relative influence of contributing users in tweet

networks (Riquelme & González-Cantergiani, 2016). Following the aforementioned node-edge characterization of retweets between accounts, this study applied the following four centrality measures that are efficiently scalable to the election corpus retweet network: eigenvector, in-degree, out-degree and PageRank.

Each of the applied centrality measures is a proxy for a specific form of relative importance within a retweet network. In-degree and out-degree centrality serve as a basis of popularity, given the cumulative direct inbound and/or outbound edges, or communication interactions, associated with each user account. Eigenvector centrality, which can be viewed as global measure of influence, is a more complex variant of degree centrality derived from the weighted sum of a given node's complete set of direct and indirect edge connections. Finally, PageRank, is an extension of eigenvector centrality that weights a degree valuation higher for nodes that initiate edges with nodes that have the highest relative importance values (Brin & Page, 1998). Therefore, user accounts with the highest PageRank valuations in a retweet network are the recipients of more retweets from the most popular user accounts. Ranking the centrality results then allowed for the identification of the specific bots with relative structural importance, while also providing an opportunity to observe any redundant coverage between the detection platforms. In addition, the proposed method of ranking centrality results maintains the integrity of the ordinal ranking results of measures such as PageRank, which cannot produce an average global interpretation as attempted in other studies (Stella et al., 2018). The Centrality Ranking and Bot Coverage subsection (Section 5.4.3) within the Results and Discussion section (Section 5.4) presents these results.

5.4. Results and Discussion

The following section presents the detailed results of the applied analysis methods described in the previous Data and Methods section (Section 5.3). Based on the bot detection results from three bot detection platforms, the Cumulative Bot Contribution Rates subsection (Section 5.4.1) facilitated the comparative analysis of bot and human temporal contributions to the overall 2018 U.S. midterm election OSN conversation. The Intra-group and Cross-group Comparison subsection (Section 5.4.2) details the interaction patterns between human and bot accounts. This section concludes with the Centrality Ranking and Bot Coverage subsection (Section 5.4.3) identifying social bots within the centrality analysis ranking results, while also presenting a bot coverage assessment based on the results of the detection platforms used in this study.

5.4.1. Cumulative Bot Contribution Rates

Figure 24 presents the cumulative contribution rates of bot and human accounts to the 2018 U.S. midterm election OSN conversation. The results shown in Figure 24a directly compare human and bot contributions rates, with an account being classified as a bot if any of the study's three detection platforms positively detected it as such. Visually, the contribution patterns of both human and bot accounts are quite consistent throughout the four weeks, although bot accounts slightly outpace the daily cumulative contributions of human accounts for the entire period. Figure 24b directly compares the cumulative contribution rates of bot accounts according to the bot detection classification results for each of the detection platforms. The results initially show similar cumulative contribution rates by bots from each detection platform, but bot accounts detected by DeBot and Bot-

hunter outpace Botometer-detected bots from September 25th through the November 6th election day. It is surprising to see the relatively consistent contribution rates across both analysis scenarios, which could suggest that the OSN election conversation elicited stable attention from both bot and human account contributors. While requiring further analysis, the observed cumulative contribution divergence by Botometer bots from DeBot and Bot-hunter bots midway through the conversation collection period could potentially suggest that bots detected by Botometer shift their interest over time to conversational topics beyond the election discussion.

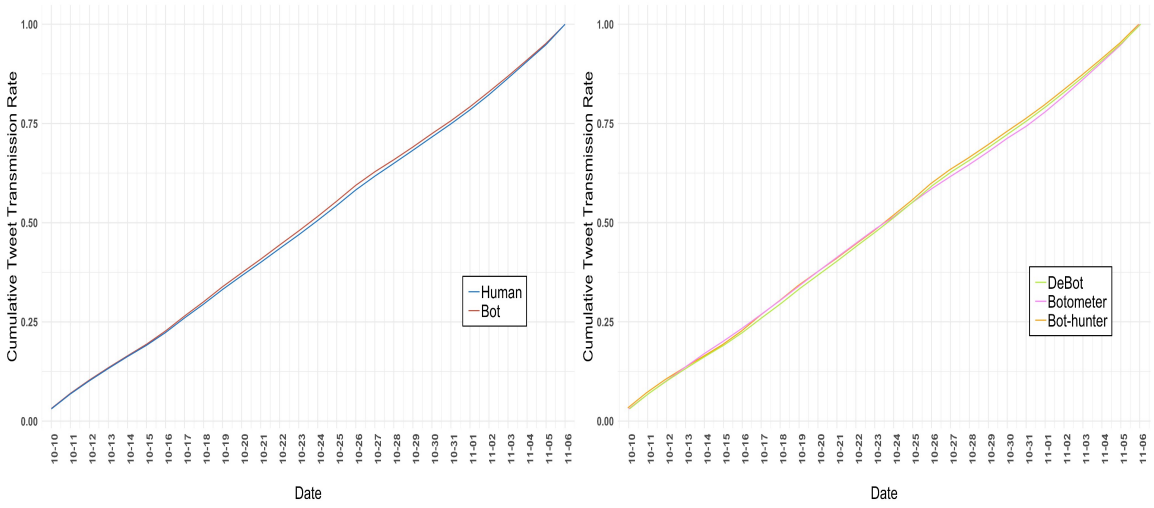
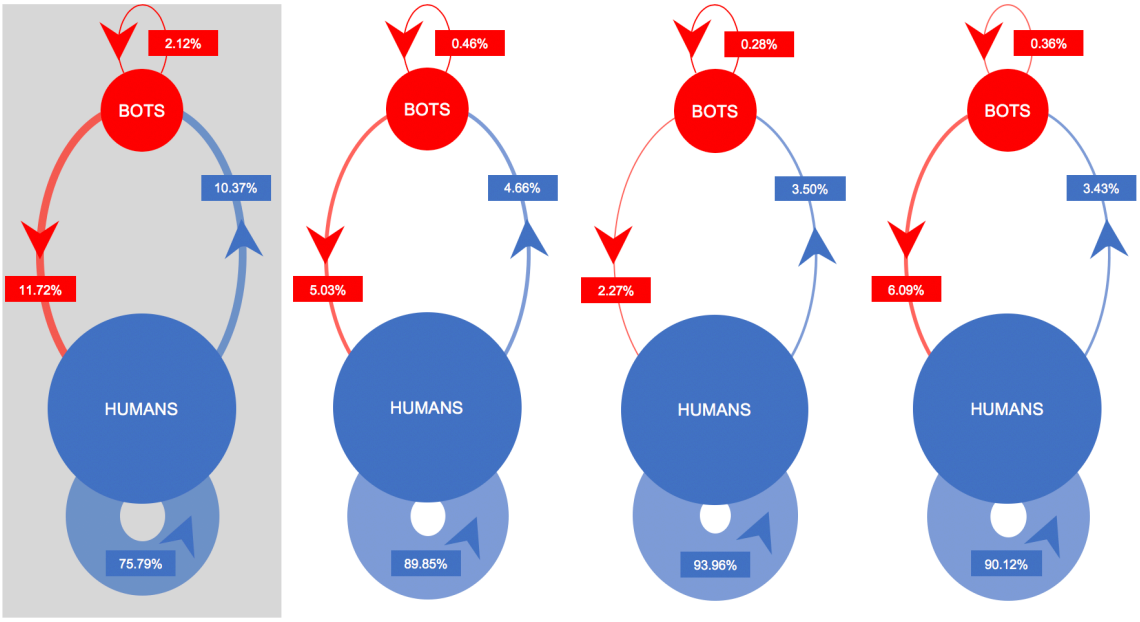


Figure 24: Cumulative tweet contribution rates for the 2018 U.S. midterm OSN conversation (October 10 – November 6, 2018) from the (a) human (blue) / bot (red) and (b) DeBot (green) / Botometer (pink) / Bot-hunter (orange) account classification perspectives.

5.4.2. Intra-group and Cross-group Comparison

The construction of the election corpus retweet network allowed for the observation of communication interaction patterns between detected bot and human

accounts. Figure 25 presents the consolidated intra-group (i.e. bots retweeting bots or humans retweeting humans) and cross-group (i.e. bots retweeting humans or humans retweeting bots) patterns between bot and human accounts from the consolidated aggregate bot perspective, shown in Figure 25a (shaded in gray), as well as individual detection platform perspectives in Figure 25b-d. Across all bot detection platforms, bot accounts initiate interaction with human accounts at a much higher rate than with other bot accounts, with intra-group bot rates all below 0.50% for from the individual detection platform perspective. Social bot accounts detected by DeBot (Figure 25b) and Bot-hunter (Figure 25d) attempt to engage with human accounts at much higher rates than observed with bot accounts detected by Botometer (Figure 25c), thus suggesting the DeBot and Bot-hunter classification algorithms more readily identify bot accounts that are more social. Most interestingly, the combined bot sources perspective (Figure 25a) shows that when combining the individual bot detection platform results, there exists minimal overlap, or redundancy, in the consolidated set of detected bots due to the substantially decreased human intra-group rate and increasing rates for all other interactions involving bots. This initial bot coverage assessment is further investigated and discussed in the following Centrality Ranking Coverage subsection (Section 5.4.3).



(a) Combined Bot Sources (b) DeBot (c) Botometer (d) Bot-hunter

Figure 25: Intra-group and cross-group retweet communication patterns of human (blue) and social bot (red) users within the 2018 U.S. midterm election OSN conversation according to each bot detection classification platform: (a) Combined Bot Sources (b) DeBot (c) Botometer (d) Bot-hunter. The combined bot sources results (shown in gray) classified an account as a bot in aggregate fashion if any of the three detection platforms classified the account as a bot.

5.4.3. Centrality Ranking and Bot Coverage

Figure 26 presents the centrality ranking analysis results by displaying the density of social bots within the top-*N,* (*where N = 1000 / 500 / 100 / 25*) centrality rankings according to each bot detection platform for the eigenvector, in-degree, out-degree and PageRank centrality measurements. Although social bots detected by DeBot and Botometer accounted for just 0.75% and 3.38% of all unique accounts in the retweet network, respectively, many displayed structural network importance by achieving top centrality out-degree and eigenvector rankings. Specifically, bots detected by DeBot accounted for more than 20% of the top-100 and top-25 out-degree ranking accounts, indicating a persistent social nature for these types of bots. Botometer-detected bots

achieved at least 50% more of the top-ranking eigenvector valuations than the other bot detection services. This could imply that Botometer detection techniques discover bots that are highly influential from a structural perspective in a network given their developed direct and indirect relationships with other accounts.

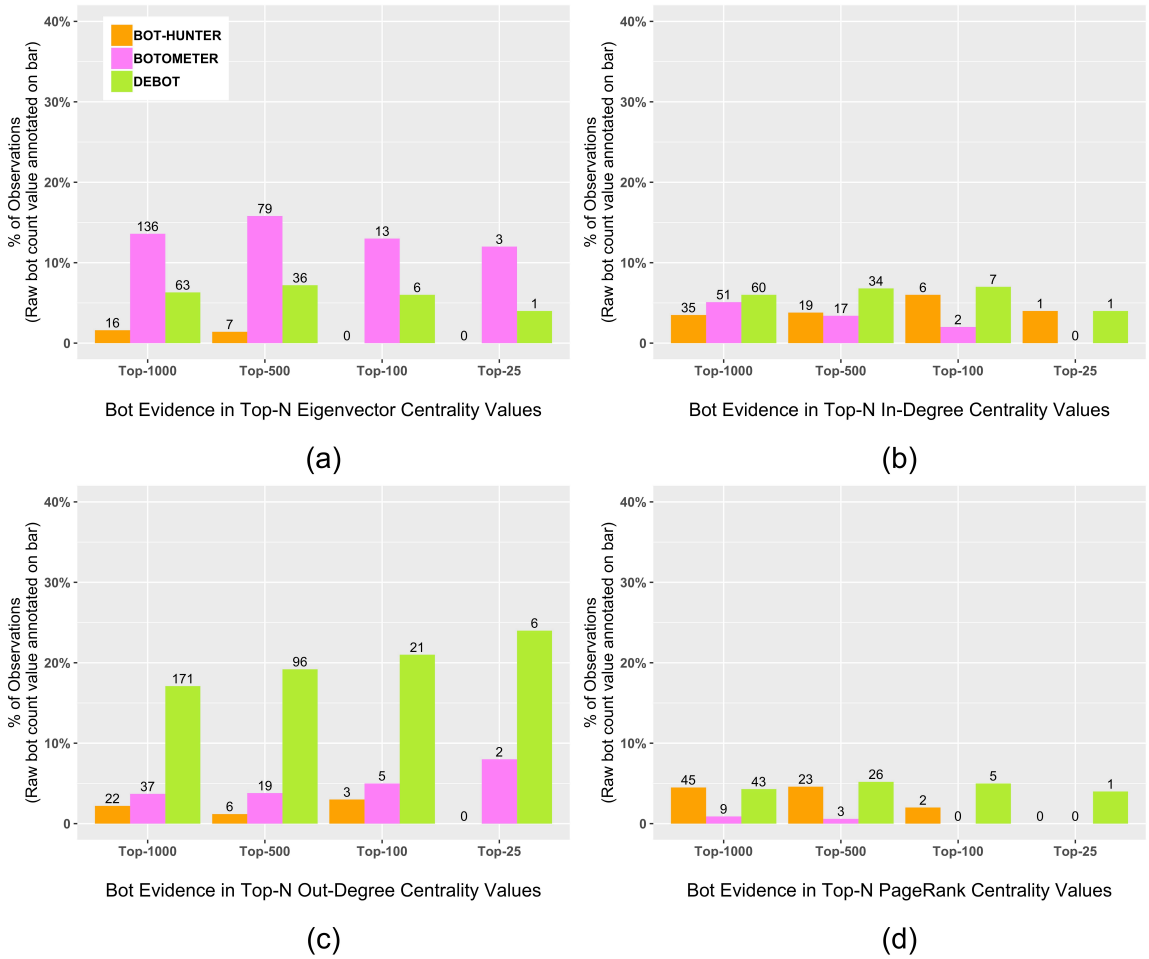


Figure 26: Social bot account evidence within the top-N (where, N = 1000 / 500 / 100 / 25) centrality rankings [(a) eigenvector (b) in-degree (c) out-degree (d) PageRank] according to bot classification results from Bot-hunter (orange), Botometer (pink) and DeBot (green).

While all of the bot detection platforms detected few bot accounts within the in-degree and PageRank centrality ranking results, the large variances shown between the out-degree and eigenvector results imply that specific detection methods detect specific types of bots. This concept is further evaluated by directly identifying each bot within the top-50 centrality rankings according to bot detection source and observing potential detection overlap. Figure 27 presents a detection classification ranking visualization with humans colored in blue and suspected bots colored according to their platform detection source. Interestingly, no bots detected within the top-50 rankings for each centrality measurement were detected by more than one detection source. This is further evidence that different detection algorithms are designed to identify different types of bots.

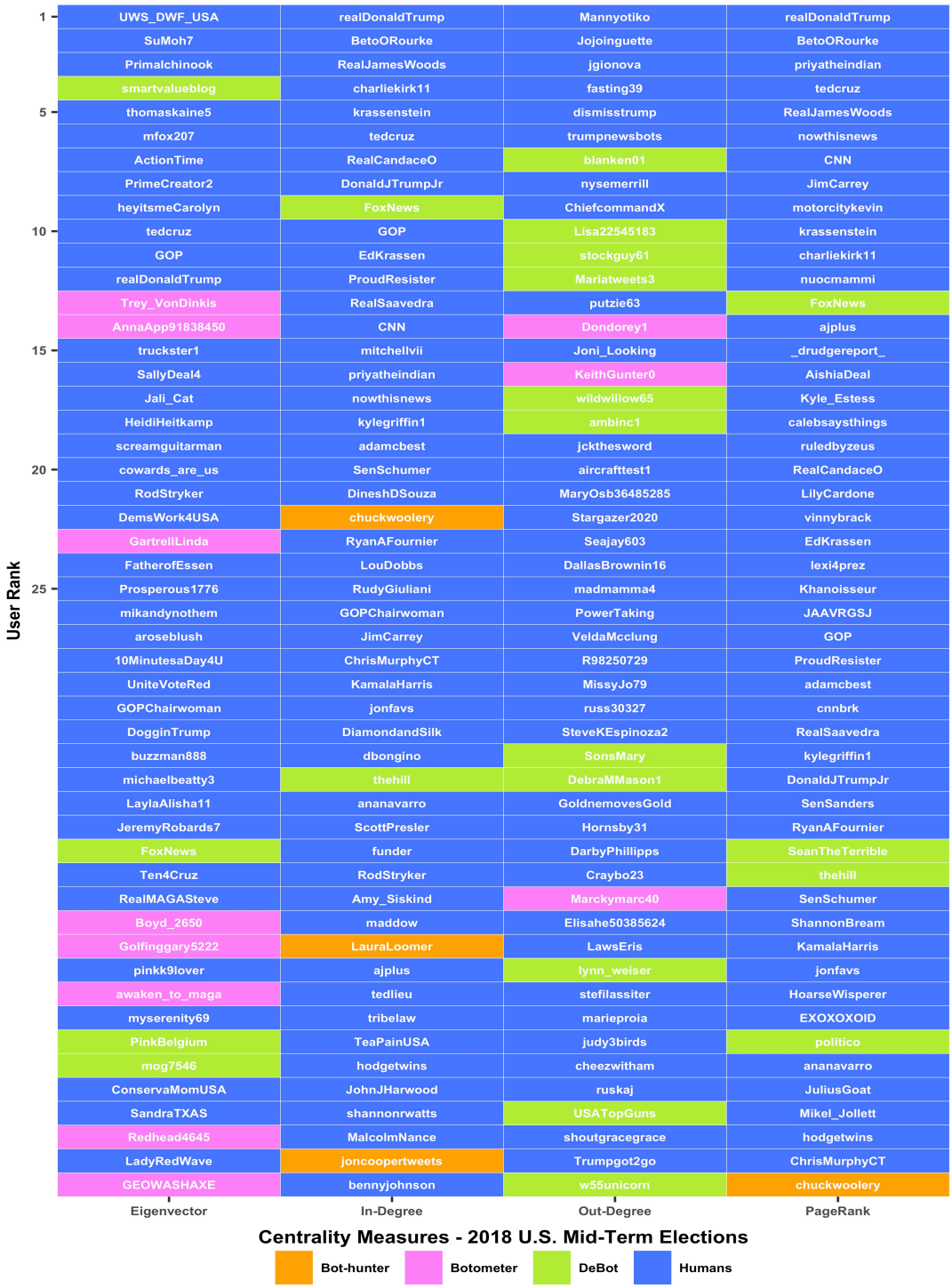


Figure 27: Top-50 bot (orange | pink | green) and human (blue) Twitter accounts within the 2018 U.S. midterm election OSN retweet network ranked by the following four centrality measures: (1) eigenvector, (2) in-degree, (3) out-degree and (4) PageRank.

The observation of minimal overlap within the consolidated set of detected bots from the retweet network discussed in the Intra-group and Cross-group sub-section (Section 5.4.2), coupled with the lack of detection overlap in the resulting centrality rankings, inspired a final bot coverage assessment of the entire election tweet corpus. The first step of this analysis consisted of a similarity assessment of the bot detection results derived from each of the bot detection platforms used in the study. The Jaccard index ($J_{A,B}$) is a similarity valuation between two sets {A, B} resulting from dividing the intersection of the two sets $|A \cap B|$ by their union $|A \cup B|$ as shown in Equation 1.

Equation 1: Jaccard similarity index

$$J_{A,B} = \frac{|A \cap B|}{|A \cup B|}$$

Table 13 presents the Jaccard similarity index results for all possible bot detection platform pairwise comparisons. Overall, there exist minimal levels of overlap between detection platforms as the highest observed similarity value is 7.62% observed between Botometer and Bot-hunter and the similarity values including DeBot are just 0.31% (DeBot and Botometer) and 1.13% (DeBot and Bot-hunter). The UpSet plot (Lex et al., 2014) shown in Figure 28 visually presents the intersection values used to calculate the Jaccard index values, while also identifying a global bot detection overlap of just eight bot accounts between all three bot detection platforms. The top bar chart of the UpSet plot represents the intersection set size between detection results, while the connected dot plots below represent the detection platforms comprising each intersection set volume.

Table 13: Jaccard similarity index values representing the pairwise comparison results of the same bots detected between each bot detection platform: Botometer (BT), Bot-hunter (BH) and DeBot (DB).

$\{A, B\}$	$\lvert A \cap B \rvert$	$\lvert A \cup B \rvert$	$J_{A,B}$
DB , BT	388	123,551	0.314%
DB , BH	1,477	131,235	1.125%
BT , BH	16,565	217,322	7.622%

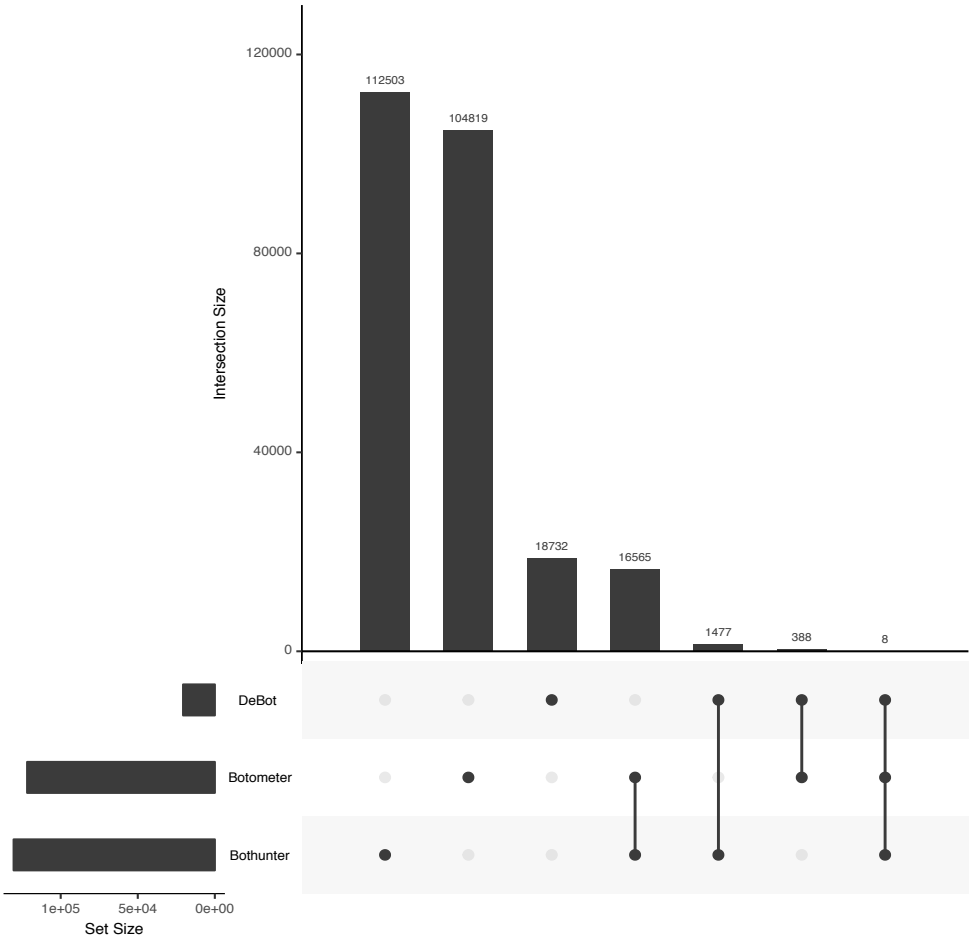


Figure 28: Bot detection coverage analysis for bots detected within the 2018 U.S. midterm election OSN conversation using the Botometer, Bot-hunter and DeBot bot detection platforms. This figure is based on the UpSet intersection of sets visualization paradigm introduced by Lex et al. (2014).

5.5. Conclusion

In summary, this chapter examined the prevalence and relative importance of detected social bots present within the 2018 U.S. midterm election OSN conversation. In expanding upon other social bot analysis works, it incorporated the use of three bot detection platforms in an unprecedented fashion, which enabled a comprehensive comparative analysis of bot coverage across the OSN conversation. Bot and human accounts contributed temporally to the 43.5 million tweet election corpus at relatively similar cumulative rates. The intra-group and cross-group analysis of the constructed retweet network showed that bots detected by DeBot and Bot-hunter persistently engaged humans at rates much higher than bots detected by Botometer. Additionally, the intra-group and cross-group interactions, when viewed from a consolidated bot account perspective, provided the first piece of evidence that minimal overall overlap existed between set of bots detected by each detection platform. The centrality ranking results showed that bots, from an overall perspective, achieved many high centrality ranking positions despite their relatively small populations size. The classification of relative importance of social bot accounts according to certain centrality results was most notable, with bots detected by DeBot in the out-degree rankings and with bots detected by Botometer in the eigenvector rankings. Analyzing the overlap of bots detected by the detection platforms showed that no overlap existed between the bots ranking in the Top-50 centrality results. Moreover, the Jaccard similarity index showed little bot detection overlap from a pairwise perspective, with only eight bots out of a total of 254,492 unique bots in the total tweet corpus having been detected by all three detection platforms.

The overall findings of the study are promising, but not immune from limitations. First of all, the analyzed OSN election corpus relied upon a single platform, Twitter. This reliance surely introduces platform representativeness and sampling bias issues as described in other works (Ruths & Pfeffer, 2014; Tufekci, 2014). Secondly, the keyword categorization of a midterm election is much harder to efficiently account for than to a more specific election like a single congressional or even presidential election. Thus, the keyword filters used to harvest tweets, while attempting to be representative and balanced, surely introduce an unknown level of potential selection bias as detailed by Zhang et al. (2018). Finally, while the focus of the study was on the cross-platform detection of bots via different sources, the ultra-conservative cutoff threshold focused on high bot precision undoubtedly contributed to an overall lower recall. While acceptable for the scope of this study, future work should seek to extend the cutoff threshold to account for more classification results.

Future extensions of this work should seek to apply this multi-detection platform framework to other OSN use-cases of interest. This study focused on the most readily available and accessibly bot detection platforms in 2019, but the rapidly evolving research area of bot detection algorithms can hopefully contribute more accessible detection platforms to the greater research community soon. New options such as these would ideally include emerging detection methods that account for the evolving nature of bots, such as the adversarial approach put forth by Cresci et al. (2019a). In addition, detection work must begin accounting for other OSN platforms beyond Twitter. Ultimately, this study expands current social bot research by putting forth a reproducible

framework to evaluate bots from a multi-detection platform perspective, and the novel analysis methods produce actionable results for analysts to better understand the prevalence and relative importance of detected social bots.

CHAPTER 6. BLOCKING TURKISH VOICES: MEASURING THE IMPACT OF CENSORSHIP

6.1. Introduction

The advent of social media has undoubtedly had a significant impact on global events throughout the past decade. Social media platforms such as Twitter, Facebook, Sina Weibo and VKontakte have made near instantaneous global communication possible to all those who have access to an Internet connection. This access to a new form of media has given individual citizens an opportunity to voice their opinions, which has helped enable and fuel significant social movements including the Arab Spring movement across the Middle East (Khondker, 2011; Wolfsfeld et al., 2013), the Euromaidan demonstrations in Ukraine (Onuch, 2015) and the Gezi Park protests in Turkey (Budak & Watts, 2015; Kuymulu, 2013). As individual citizens have learned to harness these platforms to further mobilize and sustain these social movement efforts, governments have had to learn how to account for this new avenue of discourse that can bring about rapid collective action and even political unrest. Some authoritarian regimes and despotic governments view social media platforms as highly problematic and subsequently go to great lengths to constrain or censor collective sources of political views outside of government control (Shirky, 2011).

Political censorship is by no means a new phenomenon, as many governments have a long history of controlling political discourse in the media (Briggs & Burke,

2009). The introduction of the Internet as a new medium to disseminate and access information has thoroughly complicated traditional censorship practices. Historically, governments seeking to control media could centralize authority over traditional sources such as print, radio and television (Nunziato, 2010). The decentralized nature of the Internet, amplified by the Web 2.0 revolution in which individuals became a primary contributor of content via social media platforms, thwarts these attempts at centralized control (Deibert et al., 2008; Meserve & Pemstein, 2017). Therefore, certain governments have developed a wide range of options to attempt to control Internet dialogue, specifically social media, ranging from simple messaging of appropriate online behavior discourse to sophisticated content monitoring and filtering as well as fully restricted access to the Internet (Clark et al., 2017). One instance is China's (in)famous Great Firewall, which effectively prevents its citizens from visiting a litany of web services (e.g. Google, Facebook, Twitter and YouTube) while at the same time taking great pains to monitor political discourse (Xu & Albert, 2014). Meserve and Pemstein (2017) determined that even democracies were not immune from government-level digital censorship when internal dissent became evident. The various levels of censorship efforts throughout the world have led to different categorization frameworks of censorship at the country level and have been used to inform numerous censorship studies (Nisbet et al., 2012; Warf, 2011).

While the categorization of general censorship trends can provide useful insights for specific research efforts, there are limitations to such analyses. Given the proven ability of certain nations to drastically escalate the severity of their censorship practices,

general categorizations are fixed and may therefore not apply. For example, the expansive Internet surveillance actions of the Turkish government since the late 2000s and the rapid escalation of social media censorship tactics following a 2016 coup attempt as detailed by Yesil and Sözeri (2017) would be extremely difficult to categorize in a static fashion.

To account for such dynamically changing censorship practices and their effects on regional and global social media conversations, this chapter puts forth a framework to comparatively analyze political social media dialogue prior to and during a period of extreme dynamic censorship. While other studies have looked at censorship via Internet traffic patterns (e.g. Dainotti et al. 2014; Florio et al. 2014), or specific content filtering and/or removal (e.g. Tanash et al. 2015; Zhu et al. 2013), this analysis focuses on a censorship campaign to completely block access to the Twitter social media platform. Specifically, a social network analysis-based framework is applied to Turkish political online social media conversations harvested from Twitter in December 2016 when the Turkish government abruptly blocked access to Twitter within Turkish-controlled Internet service twice in a one-week period. In doing so, the analysis of this chapter evaluated the effectiveness of such a censorship tactic on a given population by seeking observable social network artifacts at the regional and global level of the online conversation. In all, the analysis evaluated 4,257,556 tweets from November 27, 2016 through December 26, 2016.

The analysis results found the blocking campaign enacted by the Turkish government against the use of Twitter by Turkish citizens to be mildly effective. The

broad mechanism of blindly blocking access to the platform suppressed overall Turkish tweet volume activity during the censorship period in relation to the explosive volume growth observed globally. However, numerous Turkish Twitter users maintained their status as some of the most influential nodes in the Twitter network, while still discussing similar topics in the same fashion as the rest of the world.

This chapter proceeds as follows. First, Section 6.2 presents the relevant background associated with digital censorship. The Methodology section (Section 6.3) introduces the applied methods used in this chapter. The Discussion and Results section (Section 6.4) presents the findings of the applied methods, while the Conclusion and Future Work section (Section 6.5) concludes the chapter.

6.2. Background

Censorship is an incredibly broad field that covers a vast array of topics, but in the case of this study we limit the focus to political censorship of social media by authoritarian governments. The Chinese government is well known for its extreme efforts to restrict access to social media platforms via the Great Firewall and to maintain control of political narratives via a vast array of surveillance programs. King et al. (2013) made the surprising discovery that the Chinese government displayed a higher than expected tolerance for disparaging social media remarks directed at the Chinese government but immediately silenced messages tied to collective action or mobilization of protest efforts. While evaluating 56 million Sina Weibo messages and 11 million Chinese language tweets, Bamman et al. (2012) discovered a non-uniform pattern of deletion practices based on message analysis at the provincial level.

Clark et al. (2017) surveyed the evolving nature of global censorship, providing details of government-imposed penalties on individuals based on their social media activity. The study points out that more than 1,600 social media-related arrests took place following the 2016 Turkish coup attempt and 10,000 were under active investigation. The Turkish government has shown a strong penchant for instituting rapid policy changes that enable government-led social media censorship practices to evolve with citizen usage patterns (Yesil & Sözeri, 2017). Recognizing the substantial fear instilled by the Turkish government with these social media penalties, Parks et al. (2017) conducted qualitative interviews to seek ground truth perspectives from Turkish citizens on their decisions to use or avoid social media.

It is imperative to recognize Tanash et al. (2015) as the earliest substantive work to evaluate Twitter censorship in Turkey. The study examined tweet censorship requests submitted by the Turkish government to Twitter from late 2014 to early 2015. The findings showed that actual censored tweets from Turkey were two orders of magnitude higher than Twitter's own transparency report and that most of these censored tweets contained political content that was often critical of the Turkish government. In a follow-up study, Tanash et al. (2017) observed high rates of self-censorship by Turkish Twitter users immediately following the failed Turkish coup attempt in July 2016.

In a precursor to Tanash et al.'s (2017) findings of self-censorship rate increases in Turkey, Nabi (2014) claimed that censorship is not only ineffective in restricting social media access, but it also produces an unintended effect of popularizing topics governments are attempting to censor. Classifying it as the 'Streisand Effect,' Nabi

(2014) showed through analysis of data from Alexa, Google Trends, and YouTube statistics, the level of ineffectiveness of past state level censorship activities in both Pakistan and Turkey. Katz (2014) further warned of unintended consequences from social media censorship practices by concluding that although social media itself is just one of many factors behind social movement, one cannot discount its potential to mobilize social actors into action.

6.3. Methodology

While other research has focused on limited censoring activities such as content filtering and deletion, this effort evaluates an attempt at total censorship. To determine the effectiveness of a social media censorship campaign dedicated to completely blocking access of a country's population to a specific online social media platform, the following analysis applied methods borrowed from social network analysis to evaluate pre- and post-censorship network characteristics of the associated social media conversation. In this case, the rapid application of severe social media censorship against Twitter by the Turkish government in late 2016 served as the primary use case of interest. To provide further context for the discussions at both the regional and global level, natural language processing applications determined the convergent or divergent topics of discussion. The following subsections provide a detailed overview of the overall methodology and applied analytic techniques depicted in Figure 29.

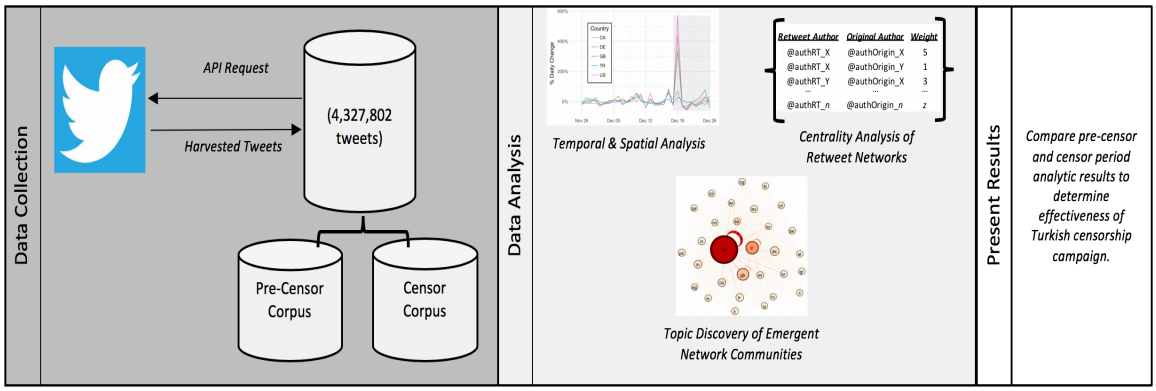


Figure 29: Overview of methodology examining OSN censorship

6.3.1. Data

The primary data source for this study were online social media conversations on Twitter related to the ongoing Turkish political activities in the aftermath of the failed 2016 coup attempt. Using the Twitter Standard Search API, English and Turkish language tweets were harvested from November through December 2016 based on relevant keywords associated with ongoing political activities (e.g. Gezi, coup, protests) and leaders (e.g. Erdogan, cumhurbaskani, Gülen) in Turkey. Truncation of the original corpus allowed for the creation of a subset corpus tweets created during the period immediately surrounding the Turkish censorship activities as described below.

In December 2016, approximately six months after a failed coup attempt against the sitting government of Turkey, Turkish citizens faced two periods during which the Turkish government intentionally blocked access to Twitter. The first blocking instance took place in the aftermath of the December 19, 2016 assassination of Andrey Karlov, the Russian Ambassador to Turkey (McGoogan, 2016). In this instance, depicted with screenshot evidence in Figure 30, technical assessments estimated the duration of blocks

to Twitter, Facebook, YouTube and others to have lasted approximately 12 hours following the assassination (Hatmaker, 2016). Three days later, the Turkish government instituted an additional series of social media blocks in response to Islamic State fighters posting a propaganda video purporting to show two Turkish soldiers being burned alive (Solomon & Srivastava, 2016). The duration of these blocks was much longer, with reports claiming outages in Turkey for up to four days.

Figure 30: Technical evidence informing initial blocking of Twitter within Turkey via tweet by @TurkeyBlocks[9].

In total, the subset censorship-related corpus covered a four-week period from November 27 through December 26, 2016 and included more than 4.2 million tweets, of which 2,802,127 (65.8%) were retweets. Two distinct epochs differentiate the corpus into distinct bins: the final week (December 19-26) serves as the censorship corpus and the preceding three weeks (November 27 through December 18) serve as the pre-censorship

[9] The Twitter account @TurkeyBlocks maps evidence of Internet censorship in Turkey. Access at https://twitter.com/turkeyblocks.

corpus. Figure 31 depicts these epoch distinctions with a gray shaded box over the censorship corpus period in the timeline presented in, along with daily tweet volumes for the entire corpus duration.

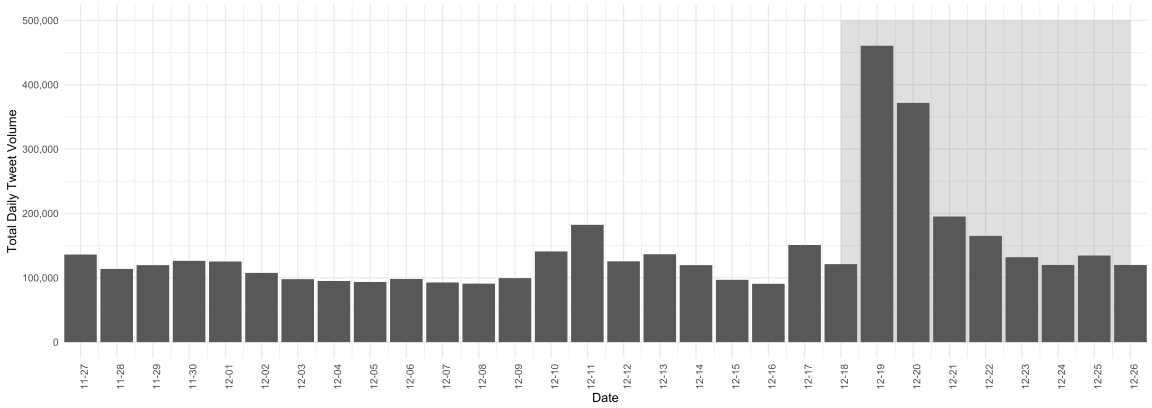


Figure 31: Total daily tweet volume of online Twitter conversations harvested from keywords associated with Turkish political events from November 27, 2016 through December 26, 2016. The daily volumes highlighted in gray (December 19-26) depict the censorship period.

6.3.2. Temporal and Spatial Patterns of Participation

As Figure 31 shows, an obvious spike of tweet volume takes place in coordination with the events of the Russian ambassador assassination on December 19, 2016. Subsequently, there appears to be sustained high volume on December 20, with volume tapering precipitously through the second period of Twitter blocks in response to the ISIS video release. To initially assess the effectiveness of Turkey blocking its citizens from accessing Twitter, country-level aggregation allowed for the classification of tweet volumes at the country level as presented in Table 14. This required reliance upon those tweets with available geolocation information. Since country-level granularity was

necessary, the country location field available in the harvested tweets served as the primary filtering key. Previous studies have used this method to successfully determine country-level activity within social media analyses (Iman et al. 2017; Zhu 2017).

Table 14: Overview of tweet corpus with geolocation features at the country-level perspective.

	Full 30-Day Corpus (27 NOV - 26 DEC 16)		Pre-Censorship Corpus (27 NOV - 18 DEC 16)		Censorship Corpus (19 DEC - 26 DEC 16)	
Country	*Tweets (% of Total)*	*Retweets (% of Total)*	*Tweets (% of Total)*	*Retweets (% of Total)*	*Tweets (% of Total)*	*Retweets (% of Total)*
United States	700,022 (37.64%)	445,684 (39.97%)	389,571 (36.40%)	233,564 (36.49%)	310,451 (39.32%)	212,120 (41.16%)
Turkey	229,612 (12.35%)	132,231 (11.86%)	168,535 (15.75%)	96,363 (15.06%)	61,077 (7.74%)	35,868 (6.96%)
Great Britain	196,114 (10.54%)	111,553 (10.00%)	101,062 (9.44%)	58,203 (9.09%)	95,052 (12.04%)	53,350 (10.35%)
Germany	82,969 (4.46%)	55,589 (4.99%)	57,367 (5.36%)	37,417 (5.85%)	25,593 (3.24%)	18,172 (3.53%)
Canada	56,223 (3.02%)	32,136 (2.88%)	29,474 (2.75%)	16,502 (2.58%)	26,749 (3.39%)	15,634 (3.03%)
Top-5	1,264,940 (68.01%)	777,193 (69.70%)	746,009 (69.71%)	442,049 (69.06%)	518,922 (65.72%)	335,144 (65.04%)
All Countries	*1,859,802*	*1,115,086*	*1,070,201*	*640,051*	*789,601*	*515,305*

Overall, the full 30-day tweet corpus contained 1,859,802 tweets (43.7%) with country-level geolocation features. To gain insight into whether the Turkish Twitter block had an effect on Turkish volumes, the time series chart depicted in Figure 32 displays the daily tweet volume percentage change for the top five volume producing countries, to include Turkey. An abrupt spike appears in the daily percentage volume change on December 19 for the United States, Canada and Great Britain, with a moderate increase for Germany and a fairly low increase for Turkey. This data suggests there was indeed a

drastic difference in interested volume for Turkey, the actual site of the event, relative to other countries.

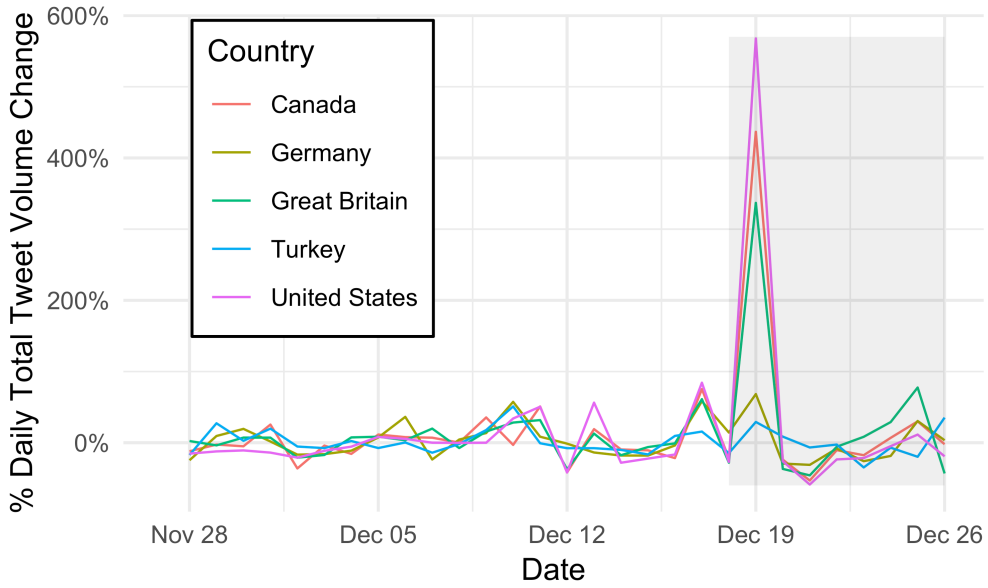


Figure 32: Daily percentage tweet volume change for the top five tweet producing countries within the tweet corpus.

To extend this analysis to the author level, the analysis extended to observe specific tweet author volume rates for the top Turkish Twitter contributors (by tweet volume) during the pre-censor period and identified any volume rate differences observed in those author tweet volumes during the ensuing censorship period. In total, there were 59,974 unique authors in the total authorship group identified as being from Turkey. Table 15 presents the top 20 Turkish authors (with anonymized names) accounting for 10% of all Turkish tweets based on pre-censor corpus volumes. Additionally, Table 15 provides a direct comparative metric between the pre-censor and censor periods in the

form of a daily average tweet rate. The results show a dramatic increase for the top author (tr_author_1) during the censorship period, while a considerable decrease (i.e. greater than a deviation 20 daily tweets) for tr_author_2, tr_author_4 and tr_author_8.

Table 15: Top-20 pre-censor Turkish authors with associated pre-censor and censor period average tweet rates.

Tweet Author	Pre-Censor Tweet Volume	% of Total Tweets from Turkey	Pre-Censor Avg. Daily Tweet Rate	Censor Tweet Volume	Censor Avg. Daily Tweet Rate
tr_author_1	3249	1.93%	135.4	1983	247.9
tr_author_2	2158	1.28%	89.9	489	61.1
tr_author_3	1237	0.73%	51.5	372	46.5
tr_author_4	1122	0.67%	46.8	62	7.8
tr_author_5	1005	0.60%	41.9	374	46.8
tr_author_6	939	0.56%	39.1	346	43.3
tr_author_7	821	0.49%	34.2	357	44.6
tr_author_8	795	0.47%	33.1	93	11.6
tr_author_9	650	0.39%	27.1	171	21.4
tr_author_10	559	0.33%	23.3	274	34.3
tr_author_11	529	0.31%	22.0	155	19.4
tr_author_12	493	0.29%	20.5	216	27.0
tr_author_13	454	0.27%	18.9	72	9.0
tr_author_14	427	0.25%	17.8	179	22.4
tr_author_15	423	0.25%	17.6	143	17.9
tr_author_16	422	0.25%	17.6	160	20.0
tr_author_17	413	0.25%	17.2	115	14.4
tr_author_18	392	0.23%	16.3	240	30.0
tr_author_19	383	0.23%	16.0	109	13.6
tr_author_20	364	0.22%	15.2	129	16.1

6.3.3. Centrality Analysis of Retweet Networks

While the basic temporal analysis at the country and author levels provided artifacts confirming associated volume changes from the pre-censor to censor periods, this analysis sought to infer the relative importance of those countries' and authors'

activities in the online conversation. To do so, SNA applied centrality measures provided such a perspective on both the country and author discussion networks derived from messages that were retweets within the harvested Twitter collection.

Centrality analysis, emanating from the larger field of social network analysis, seeks to distinguish the relative importance of actors, or nodes, based on their structural position in a given network (Wasserman & Faust, 1994). The goal of centrality analysis in this work is to determine which actors, from the country and author perspective, played the most important roles in creating, disseminating and influencing information flow throughout the associated Turkish retweet networks. To accomplish this, centrality analysis evaluated the in-degree, out-degree and eigenvector centrality measures of authors and countries during both the pre-censor and censor period of the study. Both in-degree and out-degree centrality are derivatives of the basic degree centrality measurement. Degree centrality is the summarized accounting of all connections, or edges, that an actor, or node, has within a network. In-degree and out-degree imply associated direction for a given edge, with in-degree accounting for inbound connections to a node and out-degree the opposite. Eigenvector centrality, still a derivative of degree centrality, is more elaborate as it presents a weighted sum of direct and indirect connections of a node that takes into account the individual degree centrality value of each node with which it connects in the network (Bonacich, 2007).

The construction of a retweet network is straightforward. If *Author A* reads a tweet posted by *Author B* and retweets the original message of *Author B*, then a node-to-node connection, or edge, results between *Author A* and *Author B*. An initial retweet

connection between two nodes is assigned an edge weight value of one, with subsequent retweets adding to the edge weight at increments of one. In terms of this analysis, we can view retweets as broadcasting or amplifying messages from original authors with directional implications. Therefore, the in-degree and out-degree centrality values served as the basis to quantitatively identify the actors who create or amplify the most information in the network, while using eigenvector centrality to determine top influencers.

In terms of the total tweet corpus, retweets accounted for 65.8% (2,802,127 tweets) of all tweets. As previously listed in Table 14, retweets with identifiable country-level geolocation attributes totaled 1,115,086 tweets, or 59.9% of all geolocated tweets. To evaluate pre-censorship and censorship centrality results in a comparative fashion, we created separate retweet networks for each period. The resulting pre-censorship country and author networks consisted of 231 nodes / 6,186 edges and 238,682 nodes / 357,119 edges, respectively. For the censorship period, the country and author networks consisted of 234 nodes / 6,053 edges and 214,757 nodes / 331,942 edges, respectively.

The country centrality analysis sought to determine the most prominent and influential countries participating in the Turkish Twitter conversation before and during the censorship activities. Table 16 presents the findings for comparative centrality analysis by listing the calculated centrality values for all participating countries during the pre-censorship and censorship periods. To judge the effectiveness of the Turkish censorship campaign against its own citizens, a relative drop in the prominence of Turkey's centrality values should be observable. Concurrently, specific countries should

countries serving as the top amplifier of Turkish-originated messages between the periods should also be observable. Finally, any considerable positive movement by any country in the centrality measure rankings could signify the country is filling part of the narrative void left by blocked Turkish sources.

Table 16: Top 10 pre-censorship and censorship country centrality rankings.

In-Degree		Out-Degree		Eigenvector	
Pre-Censorship	*Censorship*	*Pre-Censorship*	*Censorship*	*Pre-Censorship*	*Censorship*
United States	United States	United States	United States	United States	United States
Great Britain	Great Britain	***Turkey***	Great Britain	Great Britain	Great Britain
Turkey	***Turkey***	Great Britain	Germany	***Turkey***	***Turkey***
France	Russia	Germany	Canada	France	Russia
Germany	Germany	Canada	France	Germany	Syria
Russia	Syria	France	***Turkey***	Russia	Germany
Canada	Canada	Italy	India	Belgium	Canada
Belgium	Qatar	India	Italy	Syria	Qatar
Syria	France	Australia	Australia	Canada	France
Israel	India	Spain	Spain	Israel	India

Overall, the results from the pre-censorship country rankings are to be expected. First, United States participation rates typically dominate global Twitter conversations, so as the United States leads all centrality values. Second, Turkey ranks within the top three across all measurements, which is expected given that the focus of the harvested tweets was Turkish politics. The interesting results appear during the censorship period in which Turkey maintains its in-degree rank but falls precipitously in out-degree rankings. Furthermore, Turkey maintained a steady eigenvector or influencer status during both periods. To further classify observed retweet network characteristics, Table 17 provides a

detailed overview of the top country retweet pairs below in. In all, country pairs associated with Turkey fall in ranking dramatically during the censorship period, as Turkish retweet pairs account for slightly more than 7% of all retweet pairs during the censorship, down from more than 15% previous to censorship.

Table 17: Top pre-censorship and censorship retweet country pairs.

Pre-Censorship Corpus Retweets					Censorship Corpus Retweets				
Retweeting Country	*Original Tweet Country*	*# of Retweets*	*% of Total Retweets*	*Cumulative %*	*Retweeting Country*	*Original Tweet Country*	*# of Retweets*	*% of Total Retweets*	*Cumulative %*
United States	United States	101,562	15.87%	15.87%	United States	United States	108,669	21.09%	21.09%
Turkey	***Turkey***	47,310	7.39%	23.26%	United States	Great Britain	28,849	5.60%	26.69%
United States	***Turkey***	21,684	3.39%	26.65%	Great Britain	Great Britain	18,450	3.58%	30.27%
Great Britain	Great Britain	21,073	3.29%	29.94%	***Turkey***	***Turkey***	18,010	3.50%	33.76%
United States	Great Britain	18,152	2.84%	32.78%	United States	***Turkey***	14,203	2.76%	36.52%
Great Britain	United States	13,243	2.07%	34.84%	United States	Italy	12,697	2.46%	38.98%
Turkey	United States	11,182	1.75%	36.59%	United States	Saudi Arabia	12,167	2.36%	41.34%
Germany	***Turkey***	9,534	1.49%	38.08%	Great Britain	United States	12,096	2.35%	43.69%
Germany	Germany	8,674	1.36%	39.44%	United States	Venezuela	11,523	2.24%	45.93%
United States	Germany	6,945	1.09%	40.52%	***Turkey***	United States	5,947	1.15%	47.08%
Turkey	Germany	6,140	0.96%	41.48%	India	United States	5,537	1.07%	48.16%
Germany	United States	5,403	0.84%	42.33%	Canada	United States	5,518	1.07%	49.23%
Great Britain	Turkey	5,073	0.79%	43.12%	India	India	5,365	1.04%	50.27%
Canada	United States	4,885	0.76%	43.88%	United States	Germany	4,832	0.94%	51.21%
United States	Canada	4,064	0.63%	44.52%	United States	Russia	4,488	0.87%	52.08%

The author centrality analysis consisted of discovering the top 100 authors according to each centrality value during the pre-censorship and censorship periods. From there, the analysis determined the representation of Turkish authors within the resulting top 100 centrality rankings for both periods. The results showed that Turkish author population representation for in-degree and out-degree density dropped from 23 to 11 and 14 to 8 authors, respectively, between the two periods. Eigenvector centrality representation stayed steady as Turkish authors' representation rose minimally from 39 to 40 authors in the top 100. Therefore, Turkish authors showed a decrease of in- and out-

degree prominence during the censorship period but maintained significant influence across the network with steady eigenvector centrality values.

6.3.4. Topic Discovery within Emergent Network Communities

Community detection in networks provides researchers an opportunity to uncover underlying structural sub-graphs, or clusters, within networks. The algorithms that fall under the classification of community detection seek to discover these clusters, or communities, by focusing on the comparatively higher rate of connections between some nodes that are otherwise more isolated from the rest of the network (Girvan & Newman, 2002). There exist a wide range of community detection applications that can be used for different types of structured networks (Fortunato & Hric, 2016). The Louvain community detection method (Blondel et al., 2008) served as applied detection algorithm used to determine the existence of communities within both the pre-censorship and censorship retweet networks. The Louvain method has been quite popular for researchers given its ability to quickly scale to networks of immense size, while also not constraining the number of emerging communities to a predetermined number. The Louvain algorithm itself employs a two-stage greedy heuristic that seeks to first optimize modularity locally, then globally, while iterating until modularity reaches a maximum point.

The analysis further sought to determine community structure during both the pre-censorship and censorship periods for comparative purposes, while also examining the distribution of Turkish authors throughout each of the largest communities. The community detection results returned four primary communities for each period's network. The four identified communities accounted for 74% and 85% of all retweet

traffic for the pre-censorship and censorship networks, respectively. In terms of Turkish author distribution among the top communities, we found the maximum Turkish author membership of a top community to be 38.2% and 26.2% for the pre-censor and censorship periods, respectively. However, the total membership rates in the top communities for Turkish authors decreased from 13.2% to just 6.4% during the period of censorship. A breakdown of author and tweet volume counts is provided in Table 18.

Table 18: Top conversational topics for the most populated emergent communities.

Pre-Censorship Community Topics

Community 1		Community 2		Community 3		Community 4	
Author Population = 66,041	*Turkish Authors = 6,698*	*Author Population = 28,734*	*Turkish Authors = 945*	*Author Population = 25,808*	*Turkish Authors = 9,858*	*Author Population = 12,288*	*Turkish Authors = 35*
Total Tweets = 314,290	*Turkish Tweets = 32,797*	*Total Tweets = 47,703*	*Turkish Tweets = 2,219*	*Total Tweets = 96,176*	*Turkish Tweets = 39,891*	*Total Tweets = 27,347*	*Turkish Tweets = 169*
Topics (All Authors)	*Topics (Turkish Authors)*	*Topics (All Authors)*	*Topics (Turkish Authors)*	*Topics (All Authors)*	*Topics (Turkish Authors)*	*Topics (All Authors)*	*Topics (Turkish Authors)*
Turkey	Erdoğan	Turkey	Turkey	Erdoğan	Erdoğan	Trump	Turkey
Erdoğan	Turkey	intlspectator	intlspectator	Turkey	Cumhurbaşkanı	Turkey	Gulen
Kurdish	Mavi	Syrian	people	Cumhurbaşkanı	Turkey	blackmail	Trump
Syria	Marmara	Aleppo	killed	tcbestepe	Katıldı	business	SashaToperich
attack	davası	missing	stadium	Aleppo	Aleppo	Maddow	year
killed	Cumhurbaşkanı	stadium	Besiktas	attack	Emine	president	journalists
US	İsrail	killed	Erdoğan	DailySabah	Şehit	attention	schools
ISIS	AKP	explosion	Istanbul	PKK	Törenine	kurteichenwald	US
journalists	düştü	Trump	Blasts	Katıldı	Cenaze	extradite	Worst
die	Sonuna	US	bus	against	hope	explosive	jailed

Censorship Community Topics

Community 1		Community 2		Community 3		Community 4	
Author Population = 69,589	*Turkish Authors = 1,267*	*Author Population = 42,364*	*Turkish Authors = 218*	*Author Population = 32,224*	*Turkish Authors = 3,742*	*Author Population = 20,452*	*Turkish Authors = 5,351*
Total Tweets = 146,296	*Turkish Tweets = 3,094*	*Total Tweets = 104,328*	*Turkish Tweets = 362*	*Total Tweets = 138,516*	*Turkish Tweets = 13,032*	*Total Tweets = 50,789*	*Turkish Tweets = 14,744*
Topics (All Authors)	*Topics (Turkish Authors)*	*Topics (All Authors)*	*Topics (Turkish Authors)*	*Topics (All Authors)*	*Topics (Turkish Authors)*	*Topics (All Authors)*	*Topics (Turkish Authors)*
Turkey	Turkey	Turkey	Turkey	Turkey	Erdoğan	Turkey	Erdoğan
Russian	Russian	Russian	terror	Russian	Turkey	Erdoğan	Turkey
ambassador	ambassador	attacks	attacks	ambassador	Russian	Russian	Russian
shot	intlspectator	terror	Germany	Erdogan	ambassador	ambassador	ambassador
Ankara	shot	Germany	realDonaldTrump	ISIS	ISIS	Cumhurbaşkanı	ISIS
killed	Erdoğan	civilized	Switzerland	killed	soldiers	Aleppo	soldiers
assassination	Ankara	realDonaldTrump	worse	soldiers	video	AlabedBana	video
gunman	Reuters	Ambassador	civilized	shot	Putin	shot	Putin
attack	killed	terrorist	Russian	media	blocked	killed	killed
police	gunman	Allahu	ambassador	police	killed	Ankara	Karlov

Beyond the identification of emergent community structure, this work sought to infer the particular conversational focus areas of each identified community in order to determine if common themes existed across the top communities and during the different periods of consideration. Furthermore, by observing the specific topics of focus by

Turkish authors, the results could serve as another avenue to assess whether the Turkish government block of Twitter was successful.

Natural language processing applications allowed for the derivation of the most discussed topics for each of the top communities by tokenizing the text field for all tweets created by a community, while also creating an additional subset for Turkish authors of each community. Table 18 shows the consolidated results listing the top 10 topic words for each community and the Turkish subset. The results showed that while there was a diversity of topics in the top communities during the pre-censorship period, the assassination of the ambassador dominated the censorship period discussions across all communities, to include the Turkish community subsets.

6.4. Discussion and Results

The following section links together the most pertinent results from the analytical methods deployed in this study in order to shed light on the original research question to determine the effectiveness of a social media censorship campaign dedicated to completely blocking access to a specific online social media platform. As observed in Figure 32, Turkish tweet volumes lagged behind the sharp increases of other countries immediately following the assassination of the Russian ambassador. One can assume that an event of such magnitude would drive additional social media interest in the given locale, so the governmental block appeared to successfully contain expected growth volume. Observations of top author volume were inconclusive as the top author had a sharp increase in volume during the censorship period, while some volumes dropped, but the remainder of the top 20 stayed nearly the same. This result suggests the effectiveness

of the blocking campaign was not entirely successful, as the blocks did not account for many of the top volume authors in Turkey.

To characterize relative importance of countries and authors within the tweet conversations, centrality measures results showed that Turkey's ability to initiate retweets was severely diminished and country retweet pairs dropped substantially during the censorship period. However, Turkey maintained consistency in popularity with outside countries by maintaining a high in-degree rate and eigenvector ranking. Therefore, the blocks appeared to have succeeded only in blocking the initiation of retweets, while many Turkish nodes maintained their influence across the network during the censorship period. Finally, Turkish participation in the four most populous emergent communities fell from 13.2% to just 6.4% during the censorship period. Interestingly though, Turkish authors tweeted exactly about the topic (i.e. the Russian ambassador assassination) that the Turkish government was trying to censor throughout the censorship period.

Overall, there is evidence to support claims that the Turkish government blocking campaign of Twitter in December 2016 was indeed successful, but in a limited sense. The campaign stymied explosive usage within the Turkish population during an extreme event, but by not targeting the most influential Turkish authors in a more specific manner, the primary Turkish voices in Twitter remained largely influential.

6.5. Conclusion and Future Work

In summary, this chapter examined a social media censorship case in which a national government blocked access to an entire social media platform. The conclusions showed that the effort displayed mild success by harnessing message volumes in response

to an extreme political act of violence, but Turkish Twitter users maintained their status as influential nodes in the observed networks. Further, community detection results coupled with natural language processing showed that active Turkish Twitter users during the censorship period continued to discuss topics related to the incidents the Turkish government was trying to censor.

This analysis effort is not void of challenges and limitations. As Tufekci (2014) states, studies involving social media data must clearly state their limitations in terms of validity and representativeness. In this case, the representative sampling of harvested tweets emanated from the platform specifically targeted by the Turkish government, but collection efforts clearly did not capture the entire conversation taking place on Twitter due to API limitations. Furthermore, while resolving location at the country-level should result in a fairly reliable resolution level, one cannot fully validate the geolocation data provided by Twitter.

A primary extension to this work would be to incorporate a deliberate collection plan focused on harvesting tweets from countries beyond Turkey that have a high propensity to also censor or block access to Twitter. This would allow for a comparative analysis of censorship effectiveness between the countries. An additional extension would be to examine rates of self-censorship as posed by Tanash et al. (2017) and Nabi (2014). Such an analysis would provide insight into when populations recognize that the cost of participating in a censored environment is simply not worth the effort or risk. Still, the unique methodology put forth in this chapter adds to the field of literature

investigating the evolutionary practices of censorship taking place in today's online social networks.

CHAPTER 7. ADAPTATION TO DIGITAL CENSORSHIP: A SOCIAL SIMULATION APPROACH

7.1. Introduction

Throughout history, governments have turned to the practice of censorship as a means to suppress political dialogue (Briggs & Burke, 2009). The specific censorship tools and techniques employed have been driven by the media environments through which information can potentially be obtained (e.g. print, radio, television, Internet) (Esarey & Xiao, 2011). The accessible and decentralized nature of Internet-related information sources such as online social network (OSN) platforms (e.g. Facebook, Twitter, Sina Weibo) provides a new challenge for governments seeking to use censorship practices (Fourie et al., 2013). Recent observable challenges to governmental authority by citizens include OSN-enabled collective action ranging from non-violent digital activism (Edwards et al., 2013) to physical mass protests against oppressive regimes (Bohdanova, 2014; Tufekci & Wilson, 2012). Correspondingly, governments, in some cases, have enforced censorship efforts through physical means by punishing (e.g. prison, legal constraints) citizens who ignore or bypass censorship measures (Parks et al., 2017; Yesil & Sözeri, 2017).

Digital censorship can assume many forms, but generally consists of filtering available content, restricting access to certain sources or even implementing country-wide Internet outages (Clark et al., 2017; Dainotti et al., 2014). Examples of OSN

censorship are globally numerous and well documented. The most prominent example, the Great Firewall of China, employs a hybrid censorship strategy that fully restricts citizens from many web services the government cannot advise or control (e.g. Google, Facebook, Twitter, YouTube), while also closely monitoring citizen usage of permittable web services (Bamman et al., 2012; King et al., 2013; Xu & Albert, 2014). Other blatant digital censorship implementations include Azerbaijan (Pearce & Kendzior, 2012), Ukraine (Metzger & Tucker, 2017) and Turkey (Parks et al., 2017; Tanash et al., 2015). Clark et al. (2017) noted, unsurprisingly, that repressive regimes are more likely to implement digital censorship practices, but Meserve and Pemstein (2017) noted specific cases in which even democratic governments circumscribed the digital participation of its citizens.

The ever-increasing digital participation of citizens to contribute and access information in OSNs globally and the emerging digital censorship practices of certain governments to limit or prevent OSN usage has led to a dichotomous situation. Determining the implications of such a situation where citizens and governments (i.e. social actors) are adapting to each other's cyber actions (e.g. digital activism, digital censorship) and physical actions (e.g. mass demonstrations, punishment) is a complex problem that cannot be easily understood. Therefore, the chapter introduces an agent-based model (ABM) to examine the adaptive dynamics of a complex adaptive system of citizens choosing to engage or not engage in online discussions, while a government entity attempts to enforce censorship policies. Inspired by previous ABMs focused on emergent collective action from digital participation (Borge-Holthoefer et al., 2013;

Piedrahita, Borge-Holthoefer, Moreno, & González-Bailón, 2018) and the social identity model of collective action (SIMCA) framework (Van Zomeren et al., 2008), this model introduces a novel perspective of evaluating both cyber and physical environment adaptation to censorship practices by examining a particular population's technological skillset (i.e. technical savviness), in addition to the population's social identity and perception of its government's legitimacy. The resulting adaptation to censorship model serves as an easily extendable template to explore any digital censorship environment given the acquisition of pertinent input data as described in this chapter. Such a model provides an opportunity to simulate certain scenarios that can potentially answer how and why Turkish citizens were able to circumvent digital censorship as described in Chapter 6.

The remainder of this chapter is as follows. First, the Background section (Section 7.2) presents pertinent background literature discussing the communication theory relevant to collective action in censorship environments, obfuscation techniques to bypass online censorship tactics and a brief review of collective action ABMs using OSN data. Section 7.3 introduces and explains the model, while Section 7.4 presents an experiment comparing the results of the model when incorporating different data input parameters representative of two government entities known to participate in digital censorship practices. Finally, Section 7.5 concludes the chapter with a summary of findings and potential areas of further work.

7.2. Background

Political censorship is a pre-emptive tactic aimed at suppressing information flow to prevent the collective action of a population (King et al., 2013). OSNs have enabled major collective action events, such as the 2011 Tahrir Square protests in Egypt (Howard et al., 2011) and the 2013 Euromaidan protests in Ukraine (Bohdanova, 2014), by serving as a conduit of information which, in turn, helped mobilize citizens. To further delve into the genesis of collective action from an OSN media participation perspective, one must look to a theoretical underpinning to further analyze explicit factors that led to a tipping point threshold of action. Social identity—an individual's sense of belonging to a certain group (e.g. culture, race, economic class) (Tajfel, 1979)—serves as a primary driver of inter-group conflict that can lead to collective action (Polletta & Jasper, 2001; Tajfel et al., 1979). While social identity is a primary-enabling factor of collective action, it is not the sole direct or indirect factor (Van Zomeren et al., 2008). To account for additional observable factors of collective action, numerous research efforts have focused on developing multi-factor integrative theories of collective action (Kawakami & Dion, 1995; Stets & Burke, 2000; Van Stekelenburg & Klandermans, 2013; Van Zomeren et al., 2008). The implementation feasibility of each integrative theory into a model is data dependent and the Van Zomeren et al.'s (2008) social identity model of collective action (SIMCA) framework was the most compatible with the adaptation to censorship model data used in this chapter. Van Zomeren et al. (2008) based the SIMCA model on the three dominant socio-psychological perspectives discovered across 180 analyzed collective action studies: (1) perceived injustice (2) efficacy (3) social identity. Recent research by

Chan (2017) provided an exemplar on evaluating how alternative media, such as OSNs, potentially shaped the SIMCA model (2008) antecedents of collective action.

In terms of preventing collective action, censorship efforts can occasionally lead to unpredictable or unintentional outcomes. Using data from Google Trends, You Tube Video Statistics and Alexa Web Rankings, Nabi (2014) discovered that state-level digital censorship campaigns in Pakistan and Turkey were not only ineffective at restricting citizen access to online content, but actually popularized the content. Self-censorship, an additional potential consequence of censorship affecting an individual's efficacy, is a socio-psychological 'filter' causing an individual to withhold information intentionally or voluntarily (Bar-Tal, 2017). Cook and Heilmann (2013) categorized self-censorship resulting from political censorship practices as public self-censorship. In evaluating tweets published by Turkish citizens before and after the failed 2016 coup, Tanash et al. (2017) attributed the measurable decline in the volume of tweets censored by the Turkish government to Turkish citizens self-censoring their online Twitter participation. Finally, pluralistic ignorance is an additional factor that can potentially have a detrimental effect on collective action due to shared misperceptions of injustice or social identity. Pluralistic ignorance is a social psychology term used to describe the shared false perceptions of individuals about the internal preferences of others (Miller & McFarland, 1987; O'Gorman, 1986). Evidence of such dissonance has been used in research to explain prevailing observed sentiment associated with topics such as segregation (O'Gorman, 1975), climate change (Geiger & Swim, 2016) and college alcohol consumption (Prentice & Miller, 1993).

Citizens originally countered censorship in traditional media sources by turning to OSNs to share information, and now digital censorship practices are causing citizens to develop methods to bypass information restraints imposed on OSNs (Behrouzian et al., 2016). Parks et al (2017) described the dynamic OSN strategies citizens use to avoid digital censorship as 'transmit-trap' dynamics. Citizens can potentially bypass censorship practices by deploying an array of circumvention tools. Mou et al. (2016), for example, presented an overview of circumvention tools ranging in technical sophistication level, concluding that micro-level characteristics of individuals (e.g. technical savviness, demographics, gratifications) served as the primary factor for circumvention tool usage.

ABMs have been used to study a wide range of complex social phenomena within the research fields of economics (Epstein & Axtell, 1996), conflict (Geller & Alam, 2010; Pires & Crooks, 2017) and political science (Axelrod, 1993). In the case of this chapter, ABMs serve as a logical modeling framework to account for the complex human dynamics characterizing the emergence of collective action. Lemos et al (2013) presented a thorough review of ABMs that investigated social conflict—largely extensions inspired by Epstein's (2002) civil violence model—to include examples accounting for the role of digital communication (i.e. actions in a cyber environment) influencing physical collective action (i.e. actions in a physical environment) with the purpose of informing a future model that includes additional attributes and roles for people in protests. In attempting to explain the emergence of the Arab Spring uprisings, Makowsky and Rubin (2013) developed an ABM to observe the institutional, technological and social mechanisms responsible for revolution and discovered that access to online

communications play an important role in the sustainability of physical power for authoritarian regimes. Casilli and Tubaro (2012) directly addressed online censorship through an ABM implementation of the 2011 United Kingdom riots, determining that online censorship (i.e. cyber environment activity) of any level resulted in civil protest (i.e. physical environment activity) with higher levels of violence. Additional proposed model efforts related to censorship and OSN participation scenarios include the role of social media (i.e. cyber environment activity) in street protests (i.e. physical environment activity) (Waldherr & Wijermans, 2017) and the extension of the Granovetter (1978) threshold model to account for online interactions (i.e. cyber environment activity) as a form of participation that could lead to collective action in a physical environment (Funcke & Franke, 2016).

While previous works that directly focused on modeling online censorship are sparse, the adaptation to censorship model presented in this chapter not only adds to the literature, but also differentiates itself from the few online censorship-related models. First, previous online censorship models (Casilli & Tubaro, 2012; Funcke & Franke, 2016; Makowsky & Rubin, 2013; Waldherr & Wijermans, 2017) viewed collective action as a physical environment outcome (e.g. street protest, riot, government change), while the focused outcome in this work is a community's collective action decision to participate in an online cyber environment during a censorship event. Additionally, this model explicitly models the heterogeneous technical capabilities of a given population and government entities based on available empirical data, which is has not been attempted other censorship model efforts.

7.3. Adaptation to Censorship Model

The following sections introduce and explain the adaptation to censorship model presented in this chapter using the Overview, Design concepts and Details (ODD) protocol (Grimm et al., 2010). Section 7.3.1 introduces the model's purpose and entities. Section 7.3.2 presents the underlying concepts associated with the model's design, while Section 7.3.3 describes the model's implementation and resulting output. The source code of the model, developed using NetLogo 6.0.4 (Wilensky, 1999), is available for download at https://www.comses.net/codebase-release/df2fa006-6f2a-4ec8-b29b-7e358f43b2e1/.

7.3.1. Overview

7.3.1.1. Purpose

The purpose of the model is to explore how a population adapts to government-imposed digital censorship practices and to illustrate the extent to which the decisions individual citizens make to participate in further online activities are based upon the perception of their physical and cyber environments. Specifically, the model maps citizen attributes to the three antecedents (i.e. perceived social identity, efficacy and injustice) of the SIMCA framework (Van Zomeren et al., 2008) as shown in Figure 33. This model is designed as a foundational framework that is adaptable to specific digital censorship use-cases or situations, employing an abstraction of reality that seeks to be in qualitative agreement with obtainable empirical data sources (i.e. Level 1 ABM classification (Axtell & Epstein, 1994)). The simplification of real complex systems into an observable model form requires a litany of assumptions to be made (Batty & Torrens, 2005), which the following sections detail.

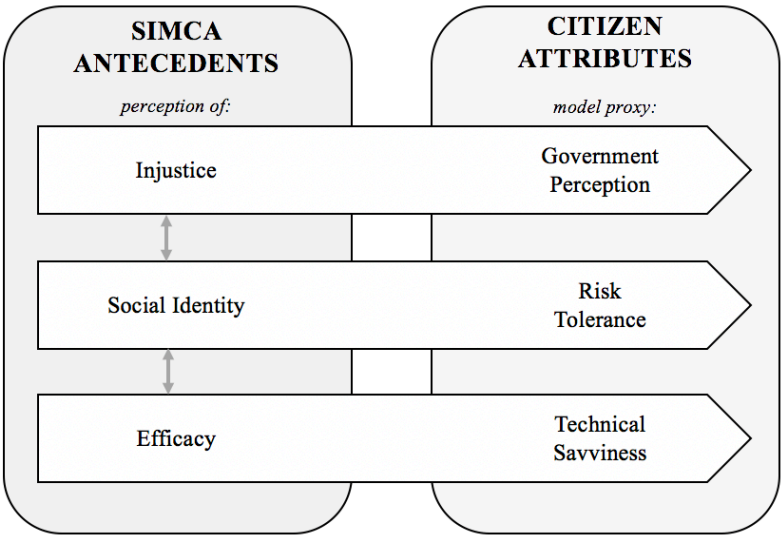


Figure 33: Mapping of SIMCA antecedents to citizen attributes in censorship model.

7.3.1.2. Entities, State Variables and Scales

The model observes citizens making decisions to engage in online activities when facing the digital censorship practices of a government entity of a specified country or locale. Citizens ultimately decide to participate in online communications (see Section 7.3.3.3) by evaluating their physical and cyber social network environments. The government, in turn, monitors and punishes citizens engaging in online conversations as described in Section 7.3.3.3. Figure 34 provides a flow diagram capturing the model's logic and processes that are further detailed in the remainder of Section 7.3.

The primary entity in the model is an agent representing an individual citizen. The citizen agents are heterogeneous actors characterized by unique state variables, or attributes, such as technical savviness, risk tolerance and government perception. The

technical savviness attribute is a measure of a citizen's general ability to use Internet technologies, where 'technically-savvy' citizens with higher valuations are more likely to use tools to obfuscate their online actions from government detection. The assignment of initial technical savviness levels is based on a given locale's available empirical distribution of technical talent (OECD, 2016) and serves as a proxy for efficacy. OECD reports the percentage of population that falls within each ICT proficiency level and the model GUI allows for the user to input a population percentage for each proficiency level. Risk tolerance is a citizen attribute representing a personal threshold to act in the face of potential retribution and serves as a determinant in a citizen's decision to participate in online communication. While initial risk tolerance levels are assigned randomly, research suggests that individual risk tolerance is a stable attribute over time, predominately determined by localized socialization (Dohmen et al., 2012; Sahm, 2012) and represents this model's social identity proxy component. The government perception attribute serves as a proxy of an individual citizen's perceived sense of injustice resulting from a government's decisions to digitally censor online communication. The initial government perception value for a given locale is based on results from World Economic Forum empirical survey data (Schwab, 2018).

Further citizen attributes include binary classifiers such as 'participant' and 'punished' that account for the status of a citizen throughout a model run. The initial designation of citizens as online participants is proportional to the total model population according to the Internet participation rate of a given locale's population as reported by the World Bank (2017). Online participants engage with one another according to a scale-

free distribution of degree connections, which is an observable distribution in online participation (Johnson et al., 2014). The assignment of punishment to a citizen is condition-based as described in Section 7.3.3.3 and symbolizes any type of digital censorship punishment (e.g. prison, probation, fines) as described by Parks et al. (2017) and Yesil and Sözeri (2017).

The government entity in the model is an exogenous actor and not explicitly observable via specific government agents. Future extensions of this model could include such government agents dispersed throughout the model environment, but that is beyond the scope of this initial model. The government 'acts' through a punishment function (Section 7.3.3.3) and derives its capability to detect online citizen participants through a government technical capability valuation that is a variable model input parameter. The author used available information and communications technology (ICT) expenditure data (United Nations, 2018) as a proxy to assign a particular government's technical capability.

Within the model, citizens interact in both a physical and cyber environment. The highly stylized physical environment is meant to simulate a citizen's physical social network, which consists of a 30 by 30 cell grid that has a uniform density of one citizen per cell. Connections in the physical environment are based on direct connections with citizens populating directly adjacent cells, and physical position does not change during a model run. The cyber environment represents an online social network consisting of edge links between citizens choosing to participate in online activities in the model. The cyber network exhibits a scale-free degree distribution in the simulated OSN and is recreated in

each step in a model run to account for citizens' daily opportunity to participate online. The temporal scale for each time step is currently notional. Future improvements to this model will introduce a temporal element based on the granularity of available data, which the author anticipates to potentially be from the day, hour and minute perspective. Therefore, the current default run time (i.e. max censorship period) of the model is 365 days, however this remains a notional threshold in the model's current development stage.

7.3.1.3. Process Overview and Scheduling

In each time step, each citizen makes three sequential observations and one decision. First, the localized technology skill spillover model (Section 7.3.3.3) enables a citizen to potentially increase its technical savviness attribute by observing the technical savviness of the local physical neighbors (i.e. adjacent eight neighbors). Next, each citizen conducts a physical and cyber environment observation to inform its decision to participate online, which comprises the steps of the overall online participation decision model. Upon completion of the online participation decision model by all citizens, the government executes the exogenous government punishment model to punish those citizens it can detect as participating in online activities. The citizen and government decision models continue to repeat themselves in the same order until the model reaches the maximum censorship time limit set by the user prior to initialization of the model.

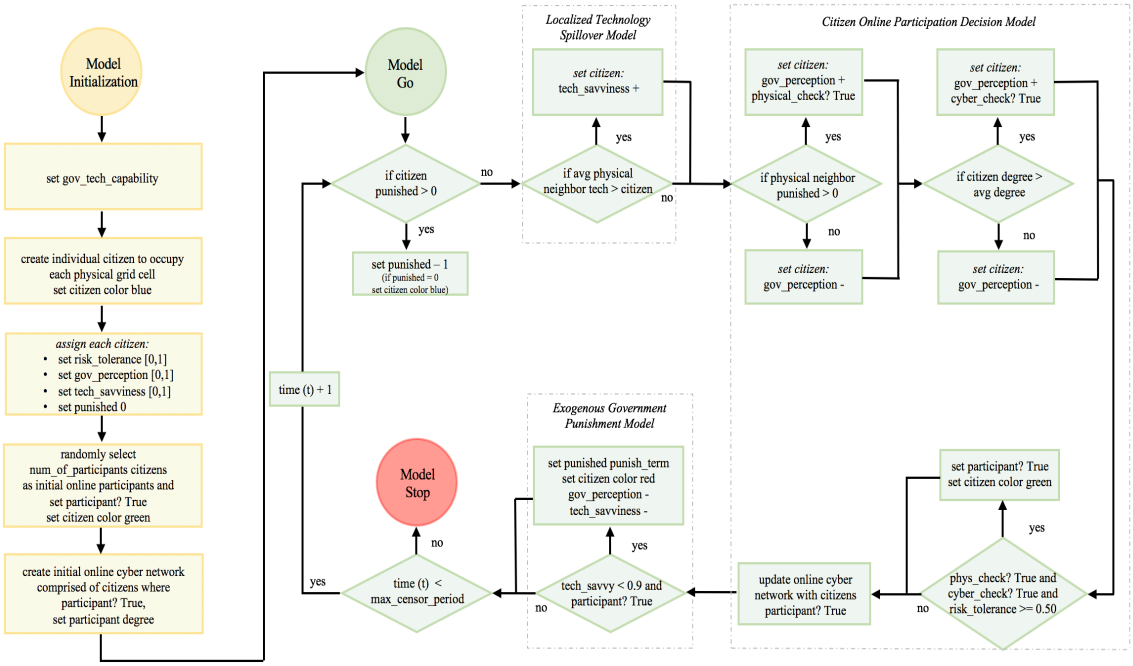


Figure 34: Adaptation to online censorship model logic and processes flow diagram.

7.3.2. Design Concepts

Aggregate monitors track and report global community statistics based on average and total citizen attribute calculations. These include the community-level measures of risk tolerance, tech savviness and government perception. Furthermore, dynamically-updated global counts report each step the different citizen classification populations of non-participant, participant and punished, while also graphically depicting each citizen classification in color (i.e. non-participant = blue, participant = green, punished = red). The changing classification of a citizen is based upon interactions among citizens within the cyber and physical environments. These state variables update state variables each step of the model.

As further detailed in Section 7.3.3.3, citizens sense the physical and cyber environments in two different ways. In the physical environment, citizens sense the state variables of their adjacent neighbors and modify their own state variables accordingly. In the cyber environment, a citizen observes participation via its degree value, which is the number of total connections it has with other citizens within the online social network. If the citizen's degree is lower than the average total degree of the online social network, then it senses there is a general overall community participation issue and decides not to continue to interact in the cyber environment.

Stochasticity exists in many forms within the model. The initial parameterization of the model includes the random assignment of individual attribute values according to the distributions as detailed in Table 19. The initial participant selection at the onset of the model is random. Additionally, the first citizen edge pair is chosen randomly to begin the construction of the online social network during each time step. While the entities execute fairly simple decision processes, the overall model enables potential emergence from a global perspective given the multitude of individual interactions based on stochastic attributes.

7.3.3. Details

7.3.3.1 Initialization

Upon initialization of the model, a citizen is created for each cell in the default size grid representing the physical environment, thus resulting in a total community population of 900 citizens. Each citizen then receives its initial attributes according to the model input parameter details listed in Table 19. The assignment of initial technical savviness values for each citizen follows an empirical distribution derived from data

detailing ICT proficiency levels across a specific country's population as reported by OECD (2016). Initial government perception values follow a similar assignment process as citizens receive assigned values according to country-specific opinion data of government performance as reported by the World Economic Forum (Schwab, 2018). The classification of citizens as online participants is initially assigned randomly and is proportional to the reported country-level Internet usage population rates by the World Bank (2017). Given these three attributes are reported with country specificity, the model user must modify the input parameters observed for the desired country of interest. The experiment presented in Section 7.4 provides country-specific examples on how to account for these parameter inputs. Finally, risk tolerance is assigned randomly to each citizen to induce stochasticity to the model and remains constant based upon observed theoretical factors (Dohmen et al., 2012; Sahm, 2012). The final initialization step is the assignment of the government technical capability attribute. This value represents the volume of citizens government entity can observe during the model and potentially initiate the punishment process. The model uses international ICT trade volume data reported by the United Nations (2018) as a proxy to determine an individual country's technical capability and categorically bins countries according to total ICT trade volume as follows: *very high*, *high*, *moderate*, *low* and *very low*.

7.3.3.2. Input Data

This model relied entirely on publicly available open-source data. The primary data sources used to stylize the model parameters include OECD (2016), the World Economic Forum (Schwab, 2018), the World Bank (2017) and the United Nations (2018). An extensive literature review shaped the estimation of certain parameters such

as censorship punishment (Parks et al., 2017; Yesil & Sözeri, 2017), risk tolerance (Dohmen et al., 2012; Sahm, 2012) and the scale-free characterization of the online social network (Johnson et al., 2014). Table 19 provides a consolidated summary of the model's input parameters. The default parameters settings represent data for Turkey, which serves as the primary censorship entity in Chapter 6 and further analyzed in the remainder of this chapter.

Table 19: Adaptation to censorship model input parameters.

Parameter	Range	Default	Reference
Citizens			
Technical savviness	0-1	Empirical Distribution	OECD (2016)
Risk tolerance	0-1	Normal (0,1)	Dohmen et al. (2012); Sahm (2012); Author estimation
Government perception	0-1	0.41	World Economic Forum (Schwab, 2018)
Online participant	T/F	T/F	World Bank (2017)
Punishment	T/F	T/F	Parks et al. (2017); Yesil & Sözeri (2017); Author estimation
Participants			
Network connections (degrees)	≥ 0	0	Johnson et al. (2014)
Total online participants	0-1	0.65	World Bank (2017)
Government			
Technical capability	0-1	0.20	United Nations (2018); Author estimation

7.3.3.3. Sub-models

The citizen agent and exogenous government entity decisions made in this model are human in nature, thus must be grounded in relevant theory. The following section ties the theoretical concepts introduced in the background section (Section 7.2) to the sub-

model processes occurring within the model and observes how each process potentially affects the SIMCA (Van Zomeren et al., 2008) core antecedents of perceived injustice, efficacy and social identity antecedents. The dashed gray boxes shown in Figure 34 annotate the logic and processes comprising each sub-model.

Localized Technology Skill Spillover. Relative physical proximity to human capital expertise can lead to a spillover of knowledge and skill (Jovanovic & Rob, 1989; Malmberg & Maskell, 2002; Moretti, 2004). In the specific instance of technology skill, knowledge from skilled humans spills over to less-skilled humans, thus resulting in the general overall increased knowledge of a localized labor population (Fang et al., 2008). In this sub-model, citizens can potentially improve their technical savviness level based upon technical knowledge spillover from adjacent neighbors. From a SIMCA perspective, a citizen can view a technical savviness increase as a potential boost to its overall perceived efficacy. At each step in the model, if a citizen observes that the average technical savviness of its neighborhood (i.e. eight adjacent neighbors) in the physical environment exceeds its current individual technical savviness level, then it can increment its current level by a fixed value 0.1. This sub-model process repeats itself at each step of the model until a citizen attains the maximum allowed technical savviness of 1.0.

Citizen Online Participation Decision. A citizen's decision to participate in online social network activities within the cyber environment is based upon three factors: the physical environment, the cyber environment and individual risk tolerance. First, a citizen observes its physical environment to determine if any of its neighbors (i.e. eight adjacent

cells) have been punished by the government. If a citizen does not observe a punishment in the neighborhood, then it increases its overall government perception value by 0.01 and sets its physical environment check status to true, while decreasing its government perception by 20% if observing neighborhood punishment. The corresponding changes to government perception affect an individual's overall feeling of perceived injustice. Next, a citizen observes its cyber environment by comparing its degree value (i.e. number of connections to other digital participant citizens) with the average degree of the entire cyber online network. If the individual citizen degree exceeds the average network degree, then it increments its government perception value by 0.01 and sets its cyber environment status check to true. A degree value of 0 indicates that the citizen either was not an online participant or did not make any online connections as a participant during the previous step. If either the physical or cyber check fails, then the applicable status check is set to false and the government perception is decreased by 0.01 per failure. Finally, the citizen has all of the information required to make a decision to participate in online activities. If a citizen passed both environmental checks and its individual risk tolerance is above 0.50, then it will decide to participate in the online social network activities. While each citizen will execute the online participation decision model for each step of the model, the government perception parameter values will not surpass the maximum or minimum threshold of 1.0 and 0.0, respectively.

Exogenous Government Punishment. Historical evidence shows the proclivity of some government entities to administer a range of punishments to citizens attempting to circumvent digital censorship policies (Clark et al., 2017; Parks et al., 2017; Yesil &

Sözeri, 2017). This model simulates the act of government punishment as the last sub-model process in a time step. The exogenous government entity observes a given population of citizens in the physical environment based on the assigned government technical capability attribute. If an observed citizen is an online participant and has a technical savviness value less than 0.90, then the government will punish the citizen. Those citizens with a technical savviness level above 0.90 are considered to have obfuscation tools that prevents the government from detecting their activities. A punishment results in an immediate 50% degradation of a citizen's technical savviness and government perception, thus effectively destroying a citizen's perceived efficacy to participate online in the cyber environment and instilling the highest level of perceived injustice. The punishment lasts according to the user-defined number of time steps.

7.4. Proof of Concept Experiment and Results

Before introducing the proof of concept experiment and results, it must be stated that this model underwent extensive verification processes. Each function passed localized logic testing to ensure all code performed as expected. Furthermore, incremental scale tests were conducted to test the scalability of code as larger populations and more interactions were introduced to the model. Finally, parameter sweeps were run to account for all potential parameter selections. These verification steps were intentional and essential to set the foundation for this model to be used as a reproducible template for various experimentation as shown in the remainder of this section. Section 7.4.1 presents a sensitivity experimental analysis to better understand the relationship between the baseline model parameters. Section 7.4.2 follows with an introduction to a comparative

country analysis, which is followed by the presentation of the experiment's results in Section 7.4.3.

7.4.1. Baseline Model Parameter Analysis

To better understand the relationship among the model's parameters in a baseline setting, this section presents a sensitivity analysis. Table 20 provides a consolidated list of the baseline parameters for the simulation runs comprising the sensitivity analysis. Following the framework (Section 7.3.3.2) to use Turkish-specific data for the model's baseline parameters, additional baseline parameters decisions for the sensitivity analysis simulation runs were made that included fixing the punishment term to be 90 time steps, running the simulation for 365 time steps and repeating a simulation run for 100 times for each particular parameter set. The sensitivity analysis focuses on how the mean community government perception and technical savviness values change over time when varying the initial government technical capability and technical savvy tiers model parameters.

Table 20: Baseline simulation settings for parameter sensitivity analyses

Simulation Variable	*Parameter Setting*
Number of simulation runs (per setting)	100
Time steps (t) (per run)	365
Punishment term	100
Number of participants	0.65
Perception of government	0.41
*Government technical capability**	0.20
*Technical savvy tiers**	
Very high	0.09%
High	6.90%
Moderate	18.6%
Low	15.9%
Very low	57.7%

*Variables serve as varied parameter test cases as explained in Section 7.4.1.

The government technical capability is the assigned attribute of the exogenous government entity to detect citizens participating in online activities during a digital censorship period. This is a fixed attribute that does not change during a simulation run. This analysis varies the government technical capability value across the following values: 0.20, 0.40, 0.60 and 0.80. Figure 35 visualizes the effect that the assigned government technical capability value has on the mean government perception and technical savviness values over the course of the simulation runs. Generally, we see a similar relationship exhibited between the mean government perception and technical savviness values for each of the government technical capability settings over time. An immediate stability between the government perception and technical savviness is observed, followed by a sharp increase in both values this is initially led by technical savviness after 90 time steps, but eventually surpassed by government perception as both values appear to stabilize after approximately 200 time steps. The primary effect observed from the variance of government technical capability is the difference between the government perception and technical savviness during the initial (0-90 time steps) and subsequent (200 to 365 time steps) periods of stability. As the government technical capability increases from Figure 35a through Figure 35d, the difference observed between government perception and technical savviness during the initial stability period diminishes, while it greatly expands in the subsequent period.

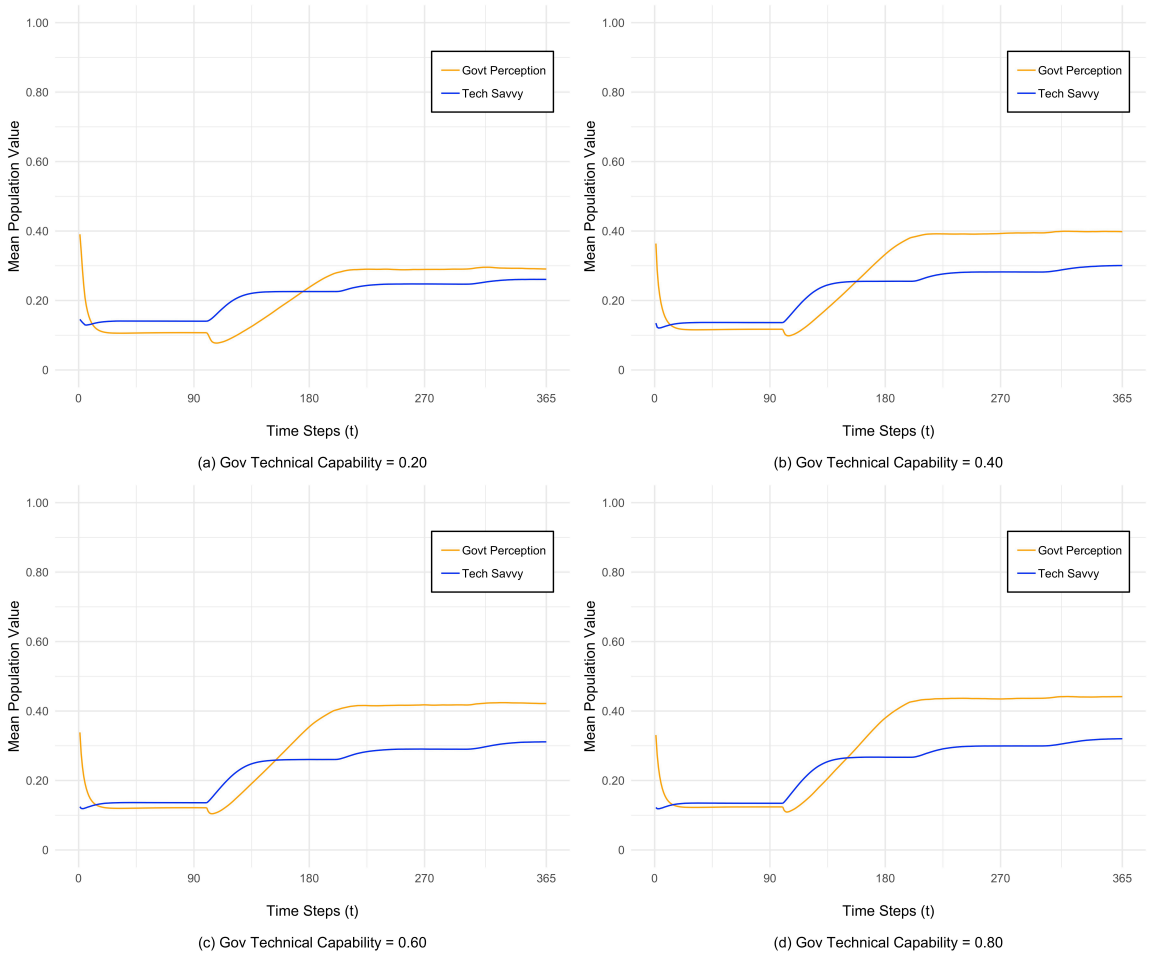


Figure 35: Resulting effect on the mean population government perception and technical savviness values over time due to varying the initial technical capability parameter of the government entity.

In contrast, the following analysis examines the effect of varying an overall population's initial technical savviness level and how it influences the mean government perception and technical savviness levels over time. This particular analysis assigns initial technical savviness levels proportionally across the five technical savviness levels in an effort to emulate a very high, moderate and very low technically savvy population. Specifically, the very high technically savvy population consists of 60% very high, 30% high, 8% moderate, 2% low and 0% very low technical savvy citizens. The moderate

technically savvy population consists of 2% very high, 23% high, 50% moderate, 23% low and 2% very low technical savvy citizens. Finally, the very low technically savvy population consists of 0% very high, 2% high, 8% moderate, 30% low and 60% very low technical savvy citizens. Figure 36 visualizes the effect that these different initial technical savviness population assignments have on the mean population government perception and technical savviness levels over time. All three populations show an initial spike in technical savviness levels corresponding with the expiration of the initial punishments at the 100^{th} time step. However, we see a primary difference with the very high technical savvy population as it gradual decays the remainder of the simulation run after the initial spike after 100 time steps, while the very low and moderate populations continue to increase the mean technical savvy level. Additionally, upon initialization, the very low population shows that it has very little technical savviness to lose as there is a minimal initial decrease in comparison to the larger initial drops observed in the moderate and very high populations.

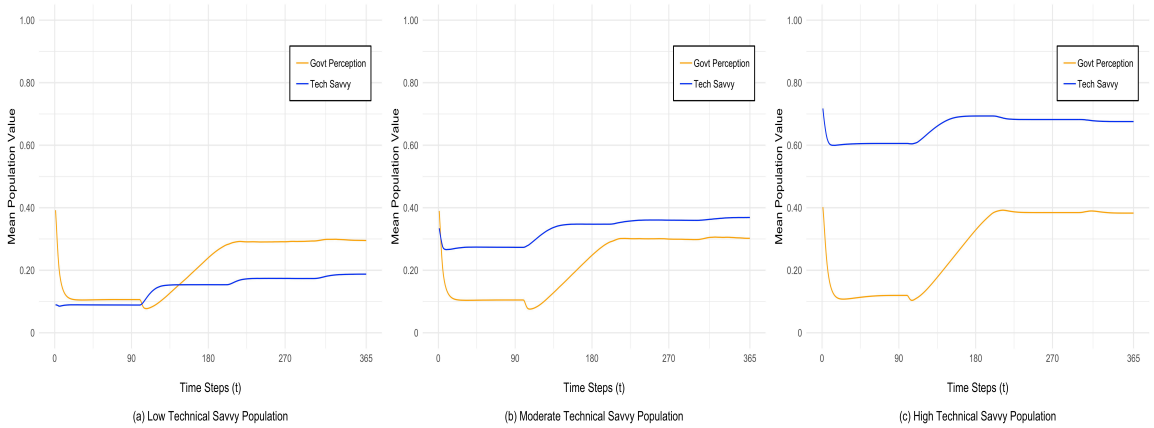


Figure 36: Resulting effect on the mean population government perception and technical savviness values over time due to varying the initial model population technical savviness levels (low (a), moderate (b) and high (c)).

7.4.2. Experiment Overview

The following overview introduces an experiment using the adaptation to censorship model with available open-source data from two countries that engage in digital censorship practices: Turkey[10] and Russia[11]. According to the input parameter descriptions and sources presented in Section 7.3, Table 21 provides the country-specific input parameters that the user must input to the model's front end graphical user interface (GUI), shown in Figure 37, prior to initialization. There is no requirement to input a *group_1* (i.e. Very Low) technical savviness population percentage as the model automatically assigns this level to all citizens that do not receive a rating of *group_2* or above.

[10] Chapter 6 presents details on past Turkish digital censorship practices.
[11] Freedom House country-level report details past Russian digital censorship practices available at https://freedomhouse.org/report/freedom-world/2019/russia.

Table 21: Country-specific input parameters for adaptation to censorship model experiment.

Input Parameter	GUI Selector	Russia	Turkey
Internet Users	*num_of_participants*	76%	65%
Government Technical Capability	*gov_tech_cap*	60%	20%
Perception of Government	*perception_of_gov*	49%	41%
Technical Savviness			
Very High	*group_5*	5.50%	0.09%
High	*group_4*	20.4%	6.90%
Moderate	*group_3*	25.6%	18.6%
Low	*group_2*	14.9%	15.9%
Very Low	--	33.6%	57.7%

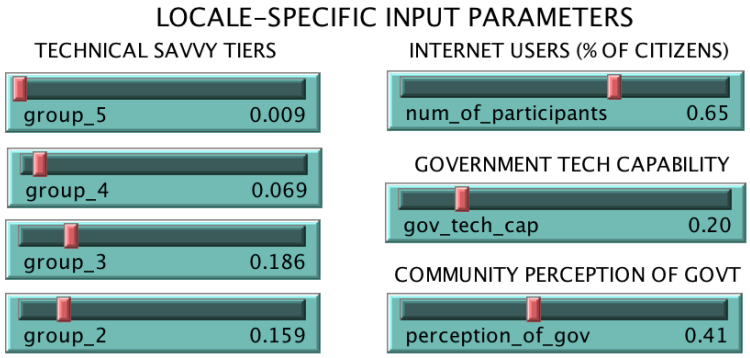


Figure 37: Snapshot of model GUI displaying input parameters for Turkey.

The experiment compares the impact of increased punishment duration on the collective action antecedent proxies for perceived injustice (i.e. community perception of government) and efficacy (i.e. community technical savviness) after a fixed-period of digital censorship. The purpose of such an experiment is to gain potential insights into how the censorship practices of two governments with significantly different technical capabilities can affect their own distinct citizen populations. In addition to the parameters listed in Table 21, the country-specific simulation experiments varied the punishment

duration parameter from 5 to 365 time steps in increments of 40 time steps. Overall, 100 simulation runs were conducted for each of the 10 different punishment durations for each country, thus, the experiment consisted of 2,000 total simulation runs.

7.4.3. Experiment Results

Figure 38 depicts how varying punishment durations affect the overall mean community perception of the Russian (black) and Turkish (brown) governments over 365 times steps. While distinctly different populations, the overall trend of both country perspectives follow a similar gradual decay in the mean perception of government by citizens, with a precipitous drop occurring when punishments extend beyond 285 time steps. The more positive view of the Russian government by its citizens in comparison to the Turkish government and its citizens is maintained consistently across the simulation runs, but it should be noted that the standard deviation ranges (depicted with error bars) are much greater for the Russian results. If preservation of positive government perception is a goal of these countries, then the simulation results show that any potential digital censorship benefits gained when punishments extend beyond 285 time steps come at a steep cost of citizens viewing their government's actions as increasingly unjust.

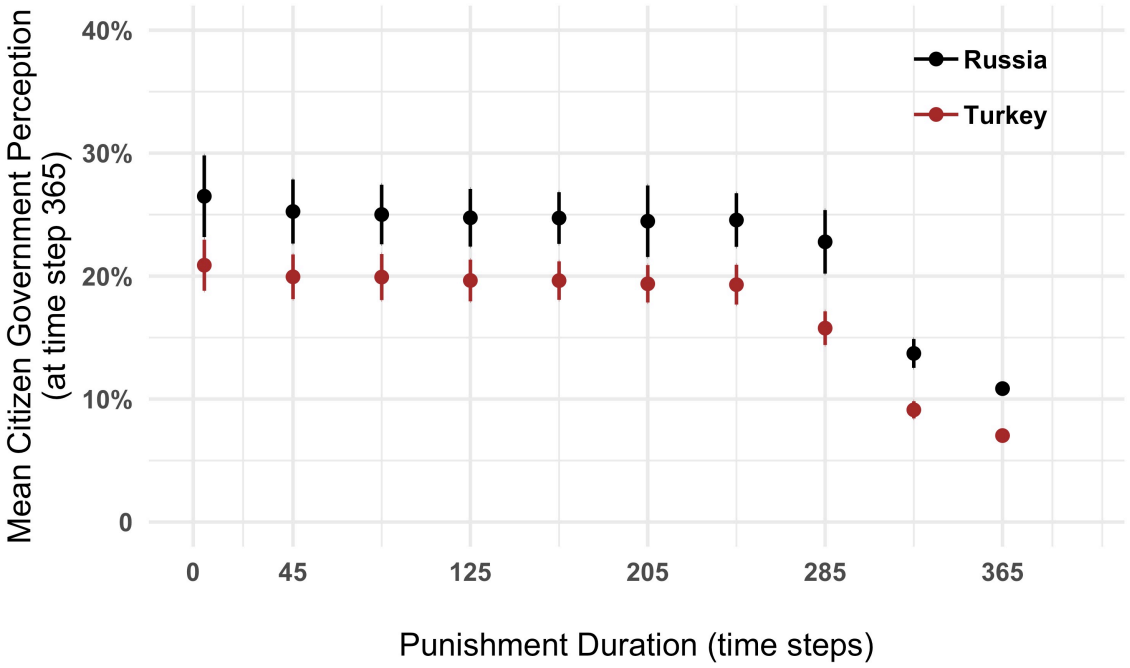


Figure 38: Impact of varying punishment duration on mean citizen government perception. Solid dots represent the mean value, while the error bars represent the standard deviation over 100 simulation runs.

The resulting effects of varied punishment duration on community technical savviness levels, as shown in Figure 39, follow somewhat similar patterns to the effects displayed on government perception. Overall, the mean citizen technical savviness levels exhibit much lower variability in terms of standard deviation. Furthermore, while government perception levels decayed somewhat steadily with increased punishment, mean technical savviness levels appeared to increase when punishment duration increased to 45 and 285 time steps, respectively. Finally, the similar precipitous drop observed in government perception is not observed with mean technical savviness until the final increase to 365 time steps. However, the same interpretation exists from a government perspective as extending punishment durations beyond 285 time steps induces a severe drop in mean citizen technical savviness. In contrast to government

support though, a particular government might want to restrict the advancement of their population's skillset to circumvent the very censorship practices instilled by it.

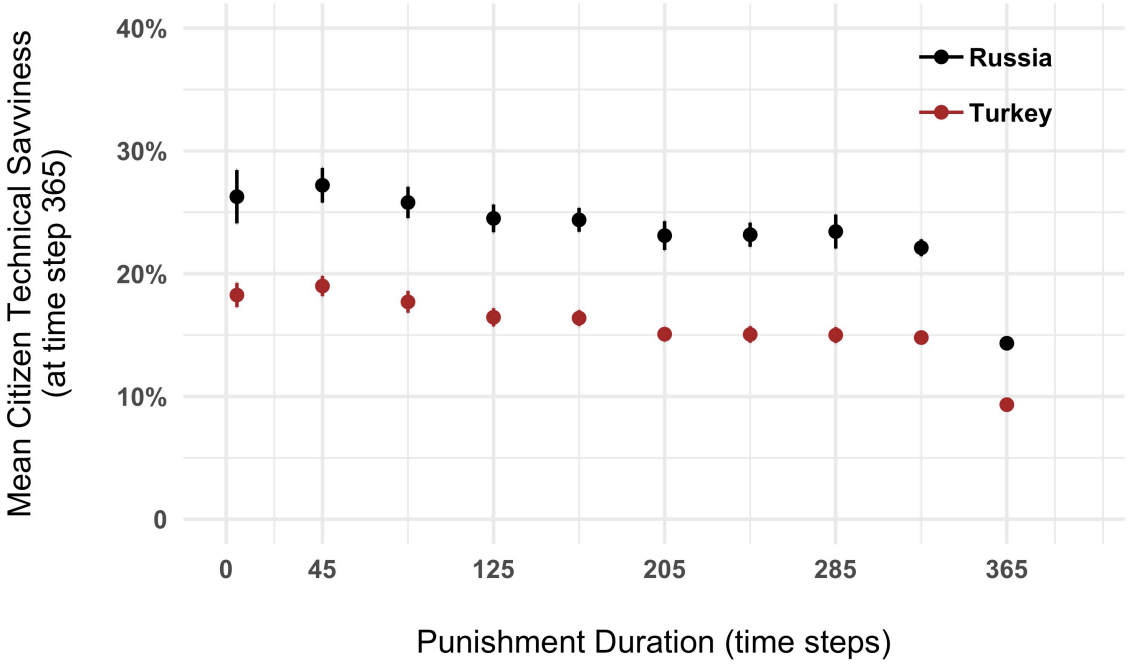


Figure 39: Impact of varying punishment duration on mean citizen technical savviness.

The precipitous declines observed in government perception and technical savviness after punishment durations of over 285 time steps require further investigation. Figure 40 shows the observation patterns of the SIMCA antecedents (i.e. risk tolerance, government perception and technical savviness) over the course of a typical simulation run where punishment durations exceed 285 time steps. Following a relative long period of stable community-level metrics, which could be characterized as rampant self-censorship, the red box annotated in Figure 40 captures a sharp divergence between the

technical savviness and government perception after 200 time steps, followed by a rapid recovery in government perception as highlighted by the red box in Figure 40. Upon further investigation, this interesting stability change is the result of many citizens reaching the 0.90 technical savviness threshold over time, thus preventing the government from detecting their online activities moving forward. The rise in mean government perception is also interesting as fewer punishments take place due to the government not being able to detect citizens that have reached the 0.90 technical savviness threshold, thus resulting in a corresponding rise in government perception.

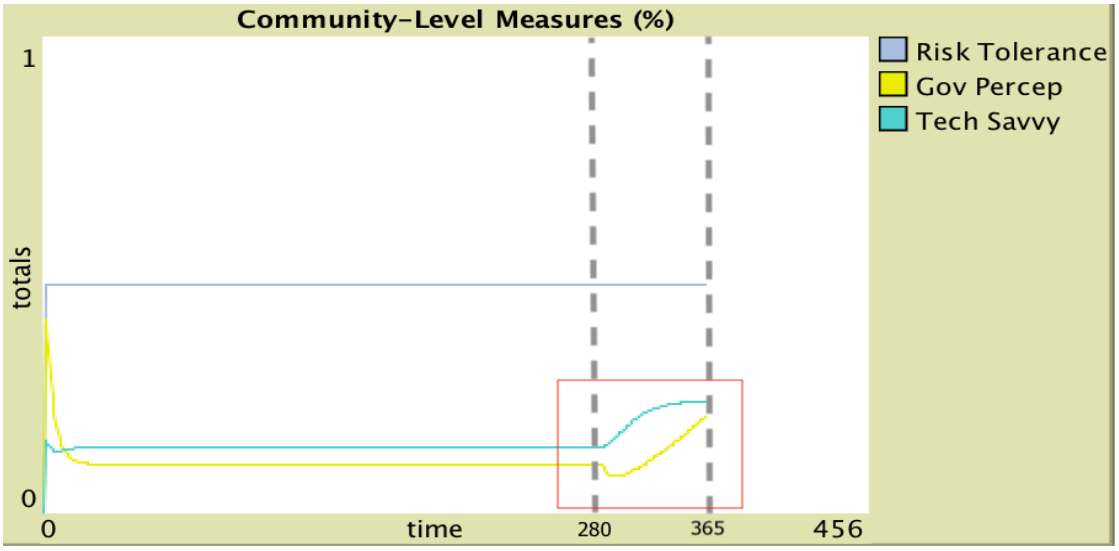


Figure 40: Typical adaptation to censorship model simulation run results from 365 time steps for punishment durations greater than 285 time steps.

7.5. Conclusion

This chapter introduced a novel ABM to examine the decisions citizens make to participate in online activities when facing digital censorship policies imposed by a government entity. The adaptation to censorship model captures the complexities of a heterogeneous population of citizens and governments (i.e. social actors) adapting to each other's cyber actions (e.g. digital activism, digital censorship) and physical actions (e.g. mass demonstrations, punishment). This initial baseline model relied upon all open-source data to parameterize the model in a templated fashion to allow for future extensions and iterations of the model. Following the SIMCA framework of Van Zomeren et al. (2008), the primary outputs of the model capture how citizen attributes serving as proxies for the antecedents of collective action evolve throughout various simulation run experiments. The experiment presented in this chapter parameterized model runs according to available public data associated with Russia and Turkey. The comparative results showed that overall levels of government perception and technical savviness declined precipitously when both government entities extended the punishment duration of citizens beyond approximately 285 time steps. Furthermore, the model captured potential self-censorship evidence as both the Russian and Turkish populations exhibited relatively stable technical savviness and government perception levels until a growing number of citizens achieved a 0.90 technical savviness threshold to avoid government detection.

There are, of course, primary limitations associated with the presented model. First, this Level-1 model, in the Axtell and Epstein (1994) schema, does not support

validation steps in its current state. This limitation is justifiable in the near-term as the baseline model, while populated with empirical data, is meant for initial exploratory analysis to inform a more sophisticated version of the model. Furthermore, the data used to support the initial parameterization of the model, as demonstrated with the proof of concept experiment, was purposefully limited and restricted to only easily-accessible data sources for the purpose of immediate reproducibility. Future extensions will include additional data sources dependent on the availability for given locales of interest. Such data sources include OSN data (e.g. Twitter, VKontakte), network traffic data (e.g. VPN connections, sensor placements) and additional sources (e.g. race, religion, economic status) that could serve as further proxies for the antecedents of collective action.

Immediate extensions of the model should focus on ascertaining data at finer level of granularity to support the determination of a temporal time step in a simulation run at the daily or hourly level. Such granularity is achievable via timestamps available from harvestable OSN platforms (e.g. Twitter, VKontakte). The use of OSN platform data should be deliberate to account for their well-known biases (Ruths & Pfeffer, 2014; Tufekci, 2014), while also exercising extreme caution to validate OSN participation patterns at the hyper-local level in targeted digital censorship campaigns (Reuter & Szakonyi, 2015). An additional immediate model extension will afford the government entity the capability to increase its technical capability over time by gaining technical competence each time it punishes a citizen with a certain level of technical savviness. Furthermore, the government technical capability can potentially decrease over time as the public will become aware of government techniques over time. This tactical

awareness is similar to the strategic decisions governments contemplate when using cyber weapons as discussed by Edwards et al. (2017).

The adaption to censorship model presented in this chapter is a baseline abstraction of reality, and, as such, is meant to serve as an initial starting point leading to a more complex model. This initial version, however, does highlight the potential for using an ABM to examine such a complex phenomenon and lays the foundation for future extensions as described previously. Understanding the complexity of interactions resulting from digital censorship campaigns is essential in helping aid oppressed citizens that are seeking to access and share information.

CHAPTER 8. CONCLUSION

8.1. Summary of Dissertation Results

Overall, this dissertation examined the range of complexity associated with social interactions that take place in rapidly evolving socio-technical systems. Specifically, this dissertation provided primary, original contributions to the nascent research fields of social bot analysis and digital censorship. The remainder of this section summarizes the observed results from each of this dissertation's chapters.

The methodological framework set forth in Chapters 2 and 3 established the initial foundation for all social bot research in the dissertation. The framework successfully created a reproducible process map for fusing bot detection classification results with harvested OSN conversation data. These processes enabled the application of quantitative analysis techniques that have not previously been conducted in the field of social bot research. Comparative social bot analysis results presented in the first two social bot chapters showed that social bots, while comprising only a small portion of the total tweet corpus author population (using one bot detection platform), attained substantially high centrality rankings and could be inferred as actors of relative social importance in the overall social network. Applying the same methodology to identify social bots within OSN mass shooting conversations, Chapter 4 found that social bots attained similar levels of structural social influence. Finally, Chapter 5, the culmination of the increasingly more

rigorous social bot analysis methods applied in this dissertation, extended the framework to incorporate three separate bot detection platforms. Ultimately, the results of Chapter 5 showed that the separate bot detection platform algorithms detected different types of bots with little overlap in the classification of bot accounts actively participating within the 2018 U.S. midterm election OSN conversation.

Chapters 6 and 7 addressed the second adaptation research focus area of the dissertation, digital censorship. Chapter 6 took advantage of a Twitter data collection harvest focusing on the aftermath of the failed 2016 Turkish coup and captured two emergent periods of digital censorship. The subsequent analysis allowed for the evaluation of the Turkish government's digital censorship attempt to block access to Twitter. The results showed that while the censorship campaign was marginally successful, it did not fully restrict the most active voices from continuing to participate in OSN conversations. The observation of continued digital participation in the face of government-imposed censorship mandates led to the development of the adaptation-to-censorship model presented in Chapter 7. This model parameterized citizen and government social actors with empirical data sources to simulate the decision process citizens undergo to determine if they should continue to engage in online activities. Using attribute proxies representing the collective action antecedents of the SIMCA framework (Van Zomeren et al., 2008), citizens with various technical expertise examine both their physical and cyber environments to determine if they will continue to engage in online activities, while attempting to potentially circumvent digital censorship efforts.

8.2. Contributions of Dissertation

As shown in this dissertation, social bot adaptation and digital censorship in online systems both require accounting for a myriad of social complexities. Given the complex nature of such a task, the overarching research methodology required this dissertation to follow a true computational social science (CSS) multi-disciplinary approach as presented in Figure 41. By following this approach, this dissertation was able to fuse data acquisition and processing, social science theory and computational modeling to present novel contributions back to the respective social bot and digital censorship research areas of interest. Specifically, the array and scale of so many different social bot analysis use-cases have not been conducted in such a simultaneous fashion before. This achievement necessitated the use of extensive data acquisition and processing skills, including scalable cloud deployments to account for the variety and volume of required data. Furthermore, the consolidation of these data enabled the application of social network analysis (SNA) techniques that have not been previously attempted in social bot analysis. This included the incorporation of social science theory such as media framing theory (Chyi & McCombs, 2004) and the SIMCA framework (Van Zomeren et al., 2008) to derive greater understanding of the underlying data results. The resulting applied SNA analyses employed in Chapters 2-5 directly led to a comparative analysis framework capable of identifying social bot pervasiveness and, more importantly, the social bot accounts of relative structural importance (Research Question 1). The multi-detection platform demonstration in Chapter 5 proved that different bot detection algorithms recognize different types of bots, thus answering Research Question 2 of this dissertation.

This feedback to the larger social bot research field as it recognizes the essential requirement to go beyond the use of sole bot detection to account for as many types of bots as possible. Finally, the digital censorship research presented in Chapter 6 showed how computational methods can evaluate the effectiveness of an authoritarian government's digital censorship campaign (Research Question 3) and can serve as the foundation for the creation of a simulation model (Chapter 7) to better infer the decision-making processes of social actors facing digital censorship practices (Research Question 4).

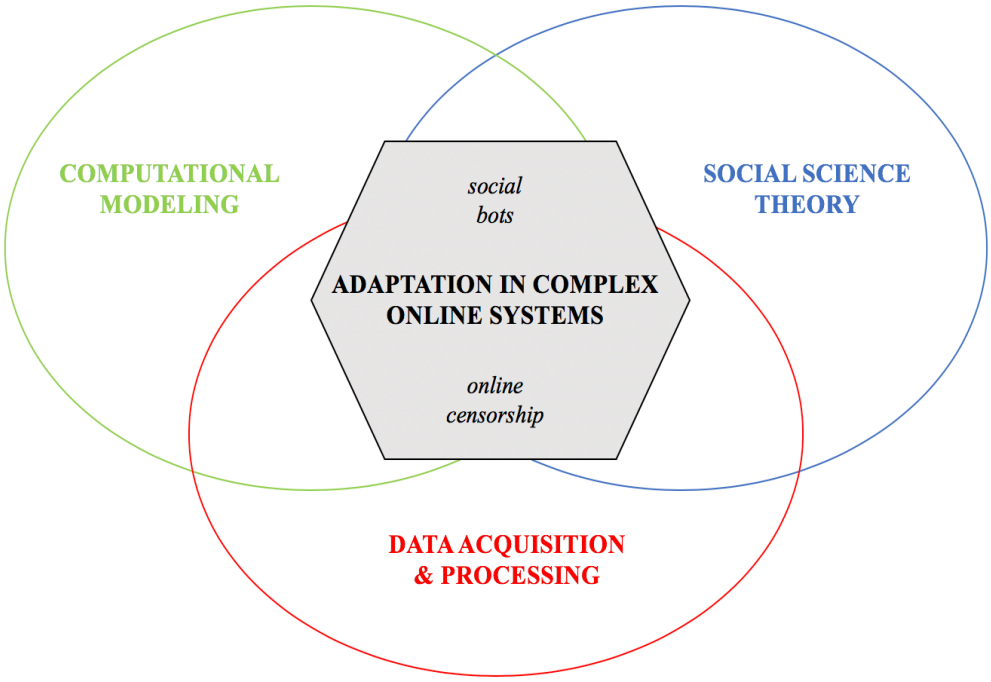


Figure 41: Overview of computational social science multi-disciplinary approach used in this dissertation to research social adaptation in complex online systems.

8.3 Future Work

Beyond the individual extensions mentioned at the conclusion of each chapter, such as the inclusion of additional use-cases viewed on other OSN platforms using a greater variety of bot detection platforms, the immediate comprehensive future work motivation for this dissertation is the consolidation of the social adaptation areas of focus (i.e. social bots, online censorship). Figure 42 visualizes the proposed intent to logically link social bot research and digital censorship research to inform a more robust ABM of digital adaptation. Such an iterative feedback framework could account for the shortcomings that exist for each of these research areas when viewed in isolation. For example, it is not possible to accurately determine the motivation of bots with given detection methodologies, so simulation could provide an avenue to test potential motivational theories. Furthermore, social bot research and digital censorship research are dependent upon available OSN data. Therefore, to proceed past easily accessible data platforms, such as Twitter, and to account for other digital participatory environments (e.g. Facebook, LinkedIn), one could consider using this dissertation's ABM to simulate these potential environments. Ultimately, emergent simulation results could lead to additional investigative ideas within the digital censorship and social bot research areas.

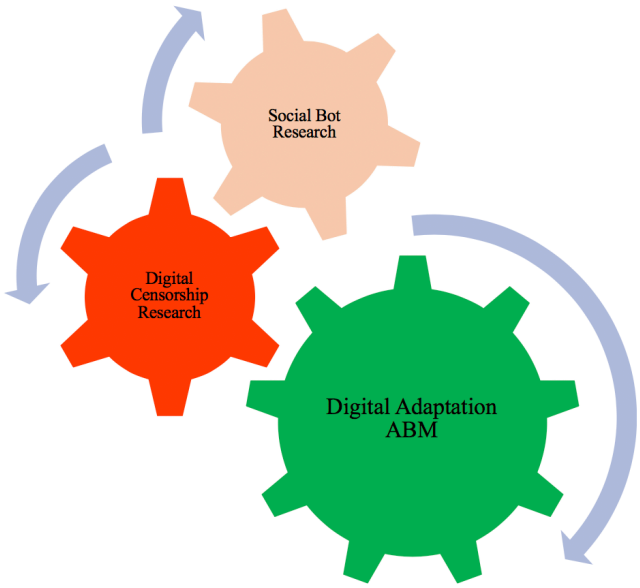


Figure 42: Vision to logically link social bot and digital research through an ABM implementation.

As presented and discussed in Chapters 2-5, social bot research currently focuses on the binary classification of OSN accounts as bot or human. While some bot detection platforms attempt to classify the sophistication of bots via a scoring continuum (e.g. Botometer (Davis et al., 2016), Bot-hunter (Beskow & Carley, 2018)), it is essential to move beyond a binary results approach and to further identify and detect bots according to motivation or intent. A potential set of rules or a bot 'Turing-test' could help inform detection algorithms to provide greater transparency and increase the speed by which bots are currently identified and understood. Media gatekeepers must lead such an endeavor and rectify the ongoing struggle with disruptive technology enabling fake news. Failure to do so could result in the eventual dissolution of trust in any media sources.

The research put forth in this dissertation facilitates immediate greater situational understanding of adaptation in complex online social systems, while enabling a potential

roadmap to account for the ever-evolving and increasing sophistication of future social bots and digital censorship campaigns. To keep pace with the growing sophistication of bots, more advanced bot detection algorithms, such as Cresci et al. (2019b), should immediately be made available as a detection source for the social bot analysis framework described in this dissertation. This would allow for a more active (i.e. real time) detection posture, as opposed to the current open-source detection platforms (e.g. Botometer, DeBot), which are more passive in nature. The incorporation of the latest detection algorithms would overcome the current OSN user verification and account deletion issues that hamper detection platforms, while also posturing the research community to account for more complex media messaging by bots, such as image and video format.

REFERENCES

Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). Dissecting a Social Botnet: Growth, Content and Influence in Twitter. In: *Proc. of 18th ACM CSCW* , 839–851.

Aiello, L. M., Deplano, M., Schifanella, R., & Ruffo, G. (2014). People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks.In: *Proc. of 6th AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 10-17.

Althoff, T., Jindal, P., & Leskovec, J. (2016). Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior. In: *Proc. of the 10th Intl. Conf. on Web Search and Data Mining*, 537-546.

Avvenuti, M., Cresci, S., Vigna, F. D., & Tesconi, M. (2016). Impromptu Crisis Mapping to Prioritize Emergency Response. *Computer*, *49*(5), 28–37.

Axelrod, R. (1993). A Model of the Emergence of New Political Actors. *Santa Fe Institute, Working Papers*.

Axtell, R., & Epstein, J. (1994). Agent-Based Modeling: Understanding Our Creations. *The Bulletin of the Santa Fe Institute*, *9*(4), 28–32.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 65–74.

Bamman, D., O'Connor, B., & Smith, N. (2012). Censorship and deletion practices in Chinese social media. *First Monday*, *17*(3).

Bar-Tal, D. (2017). Self-Censorship as a Socio-Political-Psychological Phenomenon: Conception and Research. *Political Psychology*, *38*(S1), 37–65.

Batty, M., & Torrens, P. M. (2005). Modelling and prediction in a complex world. *Futures*, *37*(7), 745–766.

Behrouzian, G., Nisbet, E. C., Dal, A., & Çarkoğlu, A. (2016). Resisting Censorship: How Citizens Navigate Closed Media Environments. *International Journal of Communication*, *10*(0), 23.

Berger, J. M., & Morgan, J. (2015). The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Project on US Relations with the Islamic World*, *3*(20), 4–1.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43.

Beskow, D. M., & Carley, K. M. (2018). *Bot-hunter: A Tiered Approach to Detecting & Characterizing Automated Activity on Twitter*.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, *21*(11).

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences*, *116*, 69–72.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bode, L., & Dalrymple, K. E. (2016). Politics in 140 Characters or Less: Campaign Communication, Network Interaction, and Political Participation on Twitter. *Journal of Political Marketing*, *15*(4), 311–332.

Bohdanova, T. (2014). Unexpected revolution: the role of social media in Ukraine's Euromaidan uprising. *European View*, *13*(1), 133–142.

Boichak, O., Jackson, S., Hemsley, J., & Tanupabrungsun, S. (2018). Automated Diffusion? Bots and Their Influence During the 2016 U.S. Presidential Election. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Transforming Digital Worlds* (pp. 17–26). Springer International Publishing.

Bolsover, G., & Howard, P. (2017). Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda. *Big Data*, *5*(4), 273–276.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, *29*(4), 555–564.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298.

Borge-Holthoefer, J., Baños, R. A., González-Bailón, S., & Moreno, Y. (2013). Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, *1*(1), 3–24.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The Socialbot Network: When Bots Socialize for Fame and Money. *Proceedings of the 27th Annual Computer Security Applications Conference*, 93–102.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2013). Design and analysis of a social botnet. *Computer Networks*, *57*(2), 556–578.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 7.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, 1–10.

Briggs, A., & Burke, P. (2009). *A Social History of the Media: From Gutenberg to the Internet*. Polity.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., … Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Budak, C., & Watts, D. J. (2015). Dissecting the Spirit of Gezi: Influence vs. Selection in the Occupy Gezi Movement. *Sociological Science*, *2*(18), 370–397.

Calero Valdez, A., Brauner, P., & Ziefle, M. (2016, January 28). *Preparing Production Systems for the Internet of Things The Potential of Socio-Technical Approaches in Dealing with Complexity*. Presented at the COMA 2016.

Casilli, A. A., & Tubaro, P. (2012). Social Media Censorship in Times of Political Unrest - A Social Simulation Experiment with the UK Riots. *Bulletin of Sociological Methodology*, *115*(1), 5–20.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in Twitter : the million follower fallacy. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, 10–17.

Chan, M. (2017). Media use and the social identity model of collective action: Examining the roles of online alternative news and social media news. *Journalism & Mass Communication Quarterly*, *94*(3), 663–681.

Chavoshi, N., Hamooni, H., & Mueen, A. (2016). DeBot: Twitter Bot Detection via Warped Correlation. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 817–822.

Chavoshi, Nikan, Hamooni, H., & Mueen, A. (2017, April 5). *Temporal Patterns in Bot Activities*.

Chavoshi, Nikan, & Mueen, A. (2018). Model Bots, not Humans on Social Media. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 178–185. IEEE.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

Chyi, H. I., & McCombs, M. (2004). Media Salience and the Process of Framing: Coverage of the Columbine School Shootings. *Journalism & Mass Communication Quarterly*, *81*(1), 22–35.

Ciampaglia, G. L. (2018). Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*, *1*(1), 147–153.

Clark, J. D., Faris, R. M., Morrison-Westphal, R. J., Noman, H., Tilton, C. B., & Zittrain, J. L. (2017). *The Shifting Landscape of Global Internet Censorship*. Retrieved from https://dash.harvard.edu/handle/1/33084425

Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. *Fifth International AAAI Conference on Weblogs and Social Media*. Presented at the Fifth International AAAI Conference on Weblogs and Social Media.

Cook, P., & Heilmann, C. (2013). Two Types of Self-Censorship: Public and Private. *Political Studies*, *61*(1), 178–196.

Cormode, G., & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday*, *13*(6).

Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2018). Social Fingerprinting: Detection of Spambot Groups Through DNA-Inspired Behavioral Modeling. *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 561–576.

Cresci, Stefano, Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. *Proceedings of the 26th International Conference on World Wide Web Companion*, 963–972.

Cresci, Stefano, Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). $ FAKE: Evidence of Spam and Bot Activity in Stock Microblogs on Twitter. *Twelfth International AAAI Conference on Web and Social Media*, 580–583.

Cresci, Stefano, Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter. *ACM Trans. Web*, *13*(2), 11:1–11:27.

Cresci, Stefano, Petrocchi, M., Spognardi, A., & Tognazzi, S. (2018). From Reaction to Proaction: Unexplored Ways to the Detection of Evolving Spambots. *WWW (Companion Volume)*, 1469–1470.

Cresci, Stefano, Petrocchi, M., Spognardi, A., & Tognazzi, S. (2019a, April 10). *Better Safe Than Sorry: An Adversarial Approach to Improve Social Bot Detection*. Presented at the ACM Web Science Conference 2019, Boston, MA, USA.

Cresci, Stefano, Petrocchi, M., Spognardi, A., & Tognazzi, S. (2019b). On the capability of evolved spambots to evade detection via genetic engineering. *Online Social Networks and Media*, *9*, 1–16.

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, *53*, 47–64.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147.

Crooks, A., Masad, D., Croitoru, A., Cotnoir, A., Stefanidis, A., & Radzikowski, J. (2014). International Relations. *Social Science Computer Review*, *32*(2), 205–220.

Dahmen, N. S., Abdenour, J., McIntyre, K., & Noga-Styron, K. E. (2018). Covering mass shootings: Journalists' perceptions of coverage and factors influencing attitudes. *Journalism Practice*, *12*(4), 456–476.

Dainotti, A., Squarcella, C., Aben, E., Claffy, K. C., Chiesa, M., Russo, M., & Pescapé, A. (2014). Analysis of Country-Wide Internet Outages Caused by Censorship. *IEEE/ACM Transactions on Networking*, *22*(6), 1964–1977.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. *Proceedings of the 25th International Conference Companion on World Wide Web*, 273–274.

Deibert, R. J., Palfrey, J. G., Rohozinski, R., & Zittrain, J. (2008). *Access Denied: The Practice and Policy of Global Internet Filtering (Information Revolution and Global Politics)*. Cambridge: MIT Press.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2012). The Intergenerational Transmission of Risk and Trust Attitudes. *The Review of Economic Studies*, *79*(2), 645–677.

Duh, A., Slak Rupnik, M., & Korošak, D. (2018). Collective Behavior of Social Bots Is Encoded in Their Temporal Twitter Activity. *Big Data*, *6*(2), 113–123.

Duxbury, S. W., Frizzell, L. C., & Lindsay, S. L. (2018). Mental illness, the media, and the moral politics of mass violence: The role of race in mass shootings coverage. *Journal of Research in Crime and Delinquency*, *55*(6), 766–797.

Edwards, B., Furnas, A., Forrest, S., & Axelrod, R. (2017). Strategic aspects of cyberattack, attribution, and blame. *Proceedings of the National Academy of Sciences*, *114*(11), 2825–2830.

Edwards, F., Howard, P. N., & Joyce, M. (2013). Digital Activism and Non-Violent Conflict. *Available at SSRN 2595115*.

Epstein, J. M. (2002). Modeling civil violence: An agent-based computational approach. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7243–7250.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

Esarey, A., & Xiao, Q. (2011). Digital communication and political change in China. *International Journal of Communication*, *5*, 22.

Fang, C., Huang, L., & Wang, M. (2008). Technology spillover and wage inequality. *Economic Modelling*, *25*(1), 137–147.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, *22*(8).

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The Rise of Social Bots. *Commun. ACM*, *59*(7), 96–104.

Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., & Galstyan, A. (2016). Predicting Online Extremism, Content Adopters, and Interaction Reciprocity. In E. Spiro & Y.-Y. Ahn (Eds.), *Social Informatics* (pp. 22–39). Springer International Publishing.

Florio, A. D., Verde, N. V., Villani, A., Vitali, D., & Mancini, L. V. (2014). Bypassing Censorship: A Proven Tool against the Recent Internet Censorship in Turkey. *2014 IEEE International Symposium on Software Reliability Engineering Workshops*, 389–394.

Forelle, M., Howard, P., Monroy-Hernández, A., & Savage, S. (2015). Political Bots and the Manipulation of Public Opinion in Venezuela. *ArXiv:1507.07109*.

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44.

Fourie, I., Bothma, T. J., & Bitso, C. (2013). Trends in transition from classical censorship to Internet censorship: selected country overviews. *Innovation: Journal of Appropriate Librarianship and Information Work in Southern Africa*, *2013*(46), 166–191.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35–41.

Fuchs, C. (2005). The internet as a self-organizing socio-technological system. *Cybernetics & Human Knowing*, *12*(3), 37–81.

Fuchs, C. (2007). *Internet and society: Social theory in the information age*. Routledge.

Funcke, A., & Franke, U. (2016). Partial participation towards collective action: To stifle or instigate. *Rationality and Society*, *28*(4), 453–467.

Geiger, N., & Swim, J. K. (2016). Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *Journal of Environmental Psychology*, *47*, 79–90.

Geller, A., & Alam, S. J. (2010). A socio-political and-cultural model of the war in Afghanistan. *International Studies Review*, *12*(1), 8–30.

Gibson, R., & Cantijoch, M. (2013). Conceptualizing and Measuring Participation in the Age of the Internet: Is Online Political Engagement Really Different to Offline? *The Journal of Politics*, *75*(3), 701–716.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420–1443.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, *221*(23), 2760–2768.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Guggenheim, L., Jang, S. M., Bae, S. Y., & Neuman, W. R. (2015). The Dynamics of Issue Frame Competition in Traditional and Social Media. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 207–224.

Hagberg, A., Schult, D., & Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11–15.

Hamm, M. P., Newton, A. S., Chisholm, A., Shulhan, J., Milne, A., Sundar, P., … Hartling, L. (2015). Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA Pediatrics*, *169*(8), 770–777.

Hatmaker, T. (2016, December 21). Turkey maintains Tor block, flicks social networks offline for 12 hours. Retrieved April 16, 2018, from TechCrunch website: http://social.techcrunch.com/2016/12/20/turkey-blocks-internet-whatsapp-twitter-assassination/

Hecking, T., Steinert, L., Masias, V. H., & Ulrich Hoppe, H. (2018). Relational Patterns in Cross-Media Information Diffusion Networks. In C. Cherifi, H. Cherifi, M. Karsai, & M. Musolesi (Eds.), *Complex Networks & Their Applications VI* (pp. 1002–1014). Springer International Publishing.

Hegelich, S., & Janetzko, D. (2016). Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. *ICWSM*, 579–582.

Hendler, J. (2009). Web 3.0 Emerging. *Computer*, *42*(1), 111–113.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). *Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?* (SSRN Scholarly Paper No. ID 2595096).

Howard, P. N., & Kollanyi, B. (2016). Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *ArXiv:1606.06356*.

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics*, *15*(2), 81–93.

Iman, Z., Sanner, S., Bouadjenek, M. R., & Xie, L. (2017). A Longitudinal Study of Topic Classification on Twitter. In: *Proceedings of the 11th ICWSM 2017,* 552-555.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 67.

Jiang, M., Cui, P., & Faloutsos, C. (2016). Suspicious Behavior Detection: Current Trends and Future Directions. *IEEE Intelligent Systems*, *31*(1), 31–39.

Johnson, S. L., Faraj, S., & Kudaravalli, S. (2014). Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment. *MIS Q.*, *38*(3), 795–808.

Jovanovic, B., & Rob, R. (1989). The growth and diffusion of knowledge. *The Review of Economic Studies*, *56*(4), 569–582.

Katz, E. (2014). Back to the Street: When Media and Opinion Leave Home. *Mass Communication & Society*, *17*(4), 454–463.

Kawakami, K., & Dion, K. L. (1995). Social Identity and Affect as Determinants of Collective Action: Toward an Integration of Relative Deprivation and Social Identity Theories. *Theory & Psychology*, *5*(4), 551–577.

Khondker, H. H. (2011). Role of the New Media in the Arab Spring. *Globalizations*, *8*(5), 675–679.

King, G., Pan, J., & Roberts, M. E. (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review*, *107*(2), 326–343.

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(3), 484–501.

Kitzie, V. L., Mohammadi, E., & Karami, A. (2018). "Life never matters in the DEMOCRATS MIND": Examining strategies of retweeted social bots during a mass shooting event. *Proceedings of the Association for Information Science and Technology*, *55*(1), 254–263.

Kleinberg, J. (2008). The Convergence of Social and Technological Networks. *Commun. ACM*, *51*(11), 66–72.

Kušen, E., & Strembeck, M. (2018, March 26). *Why so Emotional? An Analysis of Emotional Bot-generated Content on Twitter*.

Kuymulu, M. B. (2013). Reclaiming the right to the city: Reflections on the urban uprisings in Turkey. *City*, *17*(3), 274–278.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a Social Network or a News Media? *Proceedings of the 19th International Conference on World Wide Web*, 591–600.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lemos, C., Coelho, H., & Lopes, R. J. (2013). Agent-based Modeling of Social Conflict, Civil Violence and Revolution: State-of-the-art-review and Further Prospects. *EUMAS*, 124–138.

Levin, J., & Wiest, J. B. (2018). Covering mass murder: An experimental examination of the effect of news focus—killer, victim, or hero—on reader interest. *American Behavioral Scientist*, *62*(2), 181–194.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983–1992.

Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social Media and Suicide: A Public Health Perspective. *American Journal of Public Health*, *102*(S2), S195–S200.

Mahabir, R., Croitoru, A., Crooks, A., Agouris, P., & Stefanidis, A. (2018). News coverage, digital activism, and geographical saliency: A case study of refugee camps and volunteered geographical information. *PLOS ONE*, *13*(11), e0206825.

Makowsky, M. D., & Rubin, J. (2013). An Agent-Based Model of Centralized Institutions, Social Network Technology, and Revolution. *PLOS ONE*, *8*(11), e80380.

Malmberg, A., & Maskell, P. (2002). The Elusive Concept of Localization Economies: Towards a Knowledge-Based Theory of Spatial Clustering. *Environment and Planning A: Economy and Space*, *34*(3), 429–449.

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). *RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter*. Presented at the ACM Web Science 2019, Boston, MA, USA. Retrieved from http://arxiv.org/abs/1902.04506

McGoogan, C. (2016, December 20). Turkey blocks access to Facebook, Twitter and WhatsApp following ambassador's assassination. *The Telegraph*. Retrieved from

https://www.telegraph.co.uk/technology/2016/12/20/turkey-blocks-access-facebook-twitter-whatsapp-following-ambassadors/

Merry, M. K. (2016). Constructing Policy Narratives in 140 Characters or Less: The Case of Gun Policy Organizations. *Policy Studies Journal*, *44*(4), 373–395.

Meserve, S. A., & Pemstein, D. (2017). Google Politics: The Political Determinants of Internet Censorship in Democracies. *Political Science Research and Methods*, 1–19.

Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., & Finn, S. (2015). What Do Retweets Indicate? Results from User Survey and Meta-Review of Research. *Ninth International AAAI Conference on Web and Social Media*.

Metzger, M. M., & Tucker, J. A. (2017). Social Media and EuroMaidan: A Review Essay. *Slavic Review*, *76*(1), 169–191.

Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and Social Psychology*, *53*(2), 298.

Miller, J. H., & Page, S. E. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton, N.J: Princeton University Press.

Mitchell, A. (2018). *Americans Still Prefer Watching to Reading the News - and Mostly Still Through Television*. Retrieved from Pew Research Center website: http://www.journalism.org/wp-content/uploads/sites/8/2018/12/PJ_2018.12.03_read-watch-listen_FINAL1.pdf

Mitchell, M. (2011). *Complexity: A Guided Tour* (1 edition). Oxford: Oxford University Press.

Moffat, B. S. (2019). Medical Response to Mass Shootings. In M. Lynn, H. Lieberman, L. Lynn, G. D. Pust, K. Stahl, D. D. Yeh, & T. Zakrison (Eds.), *Disasters and Mass Casualty Incidents: The Nuts and Bolts of Preparedness and Response to Protracted and Sudden Onset Emergencies* (pp. 71–74).

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLOS ONE*, *12*(9), e0184148.

Moretti, E. (2004). Workers' education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review*, *94*(3), 656–690.

Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., & Liu, H. (2016). A new approach to bot detection: Striking the balance between precision and recall. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 533–540.

Mou, Y., Wu, K., & Atkin, D. (2016). Understanding the use of circumvention tools to bypass online censorship. *New Media & Society*, *18*(5), 837–856.

Mumford, E. (2006). The story of socio-technical design: reflections on its successes, failures and potential. *Information Systems Journal*, *16*(4), 317–342.

Murthy, D., Powell, A. B., Tinati, R., Anstead, N., Carr, L., Halford, S. J., & Weal, M. (2016). Automation, Algorithms, and Politics| Bots and Political Influence: A Sociotechnical Investigation of Social Network Capital. *International Journal of Communication*, *10*(0), 20.

Muschert, G. W., & Carr, D. (2006). Media salience and frame changing across events: Coverage of nine school shootings, 1997–2001. *Journalism & Mass Communication Quarterly*, *83*(4), 747–766.

Nabi, Z. (2014). ~~Resistance~~ censorship is futile. *First Monday*, *19*(11).

Newman, B. J., & Hartman, T. K. (2017). Mass shootings and public support for gun control. *British Journal of Political Science*, 1–27.

Nied, A. C., Stewart, L., Spiro, E., & Starbird, K. (2017). Alternative Narratives of Crisis Events: Communities and Social Botnets Engaged on Social Media. *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 263–266.

Niederer, S., & van Dijck, J. (2010). Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society*, *12*(8), 1368–1387.

Nisbet, E. C., Stoycheff, E., & Pearce, K. E. (2012). Internet use and democratic demands: A multinational, multilevel model of Internet use and citizen attitudes about democracy. *Journal of Communication*, *62*(2), 249–265.

Nunziato, D. C. (2010). How (Not) to Censor: Procedural First Amendment Values and Internet Censorship Worldwide. *Georgetown Journal of International Law*, *42*, 1123–1160.

OECD. (2016). *Skills Matter: Further Results form the Survey of Adult Skills*. Retrieved from OECD website: http://dx.doi.org/10.1787/9789264258051-en

O'Gorman, H. J. (1975). Pluralistic ignorance and White estimates of White support for racial segregation. *Public Opinion Quarterly*, *39*(3), 313–330.

O'Gorman, H. J. (1986). The discovery of pluralistic ignorance: An ironic lesson. *Journal of the History of the Behavioral Sciences*, *22*(4), 333–347.

Onuch, O. (2015). EuroMaidan protests in Ukraine: Social media versus social networks. *Problems of Post-Communism*, *62*(4), 217–235.

O'Reilly, T. (2005, September 30). What Is Web 2.0. Retrieved January 9, 2019, from https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Parks, L., Goodwin, H., & Han, L. (2017). "I Have the Government in My Pocket…": Social Media Users in Turkey, Transmit-Trap Dynamics, and Struggles Over Internet Freedom. *Communication, Culture & Critique*, *10*(4), 574–592.

Pearce, K. E., & Kendzior, S. (2012). Networked authoritarianism and social media in Azerbaijan. *Journal of Communication*, *62*(2), 283–298.

Perna, D., & Tagarelli, A. (2018). Learning to Rank Social Bots. *Proceedings of the 29th on Hypertext and Social Media*, 183–191.

Persily, N. (2017). The 2016 US Election: Can democracy survive the internet? *Journal of Democracy*, *28*(2), 63–76.

Pew Research Center. (2017). *News Use Across Social Media Platforms*. Retrieved from http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Pew Research Center. (2018). *Activism in the Social Media Age*. Retrieved from http://www.pewinternet.org/2018/07/11/activism-in-the-social-media-age/

Piedrahita, P., Borge-Holthoefer, J., Moreno, Y., & González-Bailón, S. (2018). The contagion effects of repeated activation in social networks. *Social Networks*, *54*, 326–335.

Piraveenan, M., Prokopenko, M., & Hossain, L. (2013). Percolation Centrality: Quantifying Graph-Theoretic Impact of Nodes during Percolation in Networks. *PLOS ONE*, *8*(1), e53095.

Pires, B., & Crooks, A. T. (2017). Modeling the emergence of riots: A geosimulation approach. *Computers, Environment and Urban Systems*, *61*, 66–80.

Polletta, F., & Jasper, J. M. (2001). Collective Identity and Social Movements. *Annual Review of Sociology*, *27*(1), 283–305.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*(2), 243–256.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, *7*(1), 13006.

Reuter, O. J., & Szakonyi, D. (2015). Online social media and political awareness in authoritarian regimes. *British Journal of Political Science*, *45*(1), 29–51.

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, *52*(5), 949–975.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Sahm, C. R. (2012). How Much Does Risk Tolerance Change? *The Quarterly Journal of Finance*, *2*(4).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 919–931.

Sayama, H., Pestov, I., Schmidt, J., Bush, B. J., Wong, C., Yamanoi, J., & Gross, T. (2013). Modeling complex systems with adaptive networks. *Computers & Mathematics with Applications*, *65*(10), 1645–1664.

Schildkraut, J., Elsass, H. J., & Meredith, K. (2018). Mass shootings and the media: Why all events are not created equal. *Journal of Crime and Justice*, *41*(3), 223–243.

Schildkraut, J., & Muschert, G. W. (2014). Media salience and the framing of mass murder in schools: A comparison of the Columbine and Sandy Hook massacres. *Homicide Studies*, *18*(1), 23–43.

Scholtes, I., Pfitzner, R., & Schweitzer, F. (2014). The Social Dimension of Information Ranking: A Discussion of Research Challenges and Approaches. In K. Zweig, W. Neuser, V. Pipek, M. Rohde, & I. Scholtes (Eds.), *Socioinformatics - The Social Impact of Interactions between Humans and IT* (pp. 45–61).

Schuchard, R., Crooks, A., Stefanidis, A., & Croitoru, A. (2019). Bots in Nets: Empirical Comparative Analysis of Bot Evidence in Social Networks. In L. M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, & L. M. Rocha (Eds.), *Complex Networks and Their Applications VII* (pp. 424–436). Springer International Publishing.

Schwab, K. (2018). *Global Competitiveness Index: Public trust in politicians*. Retrieved from World Economic Forum: The Global Competitiveness Report website: http://reports.weforum.org/global-competitiveness-report-2018/downloads/

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787.

Shirky, C. (2011). The Political Power of Social Media: Technology, the Public Sphere, and Political Change. *Foreign Affairs*, *90*(1), 28–41.

Silva, J. R., & Capellan, J. A. (2019). The media's coverage of mass public shootings in America: fifty years of newsworthiness. *International Journal of Comparative and Applied Criminal Justice*, *43*(1), 77–97.

Simon, H. A. (1996). *The Sciences of the Artificial - 3rd Edition* (3rd edition). Cambridge, Mass: The MIT Press.

Solomon, E., & Srivastava, M. (2016, December 23). Isis video prompts Turkey to block Twitter, YouTube, Facebook access. Retrieved August 16, 2018, from Financial Times website: https://www.ft.com/content/8ac4bda2-c15e-3269-9339-95df4702c535

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. *Eleventh International AAAI Conference on Web and Social Media*.

Stella, M., Cristoforetti, M., & De Domenico, M. (2018). Influence of augmented humans in online interactions during voting events. *ArXiv:1803.08086*.

Stella, M., Ferrara, E., & Domenico, M. D. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435–12440.

Stets, J. E., & Burke, P. J. (2000). Identity Theory and Social Identity Theory. *Social Psychology Quarterly*, *63*(3), 224–237.

Stöber, R. (2004). What Media Evolution Is: A Theoretical Approach to the History of New Media. *European Journal of Communication*, *19*(4), 483–505.

Strohmaier, M., & Wagner, C. (2014). Computational Social Science for the World Wide Web. *IEEE Intelligent Systems*, *29*(5), 84–88.

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting Bots on Russian Political Twitter. *Big Data*, *5*(4), 310–324.

Suárez-Serrato, P., Roberts, M. E., Davis, C., & Menczer, F. (2016). On the Influence of Social Bots in Online Protests. In E. Spiro & Y.-Y. Ahn (Eds.), *Social Informatics* (pp. 269–278). Springer International Publishing.

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., … Menczer, F. (2016). The DARPA Twitter Bot Challenge. *Computer*, *49*(6), 38–46.

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Tajfel, H. (1979). Individuals and groups in social psychology*. *British Journal of Social and Clinical Psychology*, *18*(2), 183–190.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational Identity: A Reader*, 56–65.

Tanash, R. S., Chen, Z., Thakur, T., Wallach, D. S., & Subramanian, D. (2015). Known Unknowns: An Analysis of Twitter Censorship in Turkey. *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, 11–20.

Tanash, R. S., Chen, Z., Wallach, D. S., & Marschall, M. (2017). The Decline of Social Media Censorship and the Rise of Self-Censorship after the 2016 Failed Turkish Coup. In: *7th USENIX Workshop on Free and Open Communcations.*

Theocharis, Y., & Deth, J. W. van. (2018). The continuous expansion of citizen participation: a new taxonomy. *European Political Science Review*, *10*(1), 139–163.

Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., & Castillo-Chavez, C. (2015). Contagion in Mass Killings and School Shootings. *PLOS ONE*, *10*(7), e0117259.

Trist, E. L., & Bamforth, K. W. (1951). Some Social and Psychological Consequences of the Longwall Method of Coal-Getting: An Examination of the Psychological Situation and Defences of a Work Group in Relation to the Social Structure and Technological Content of the Work System. *Human Relations*, *4*(1), 3–38.

Tufekci, Z. (2014, March 28). *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*. 505–514.

Tufekci, Z., & Wilson, C. (2012). Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. *Journal of Communication*, *62*(2), 363–379.

United Nations. (2018). *UNCTADstat: International Trade in ICT services*. Retrieved from https://unctadstat.unctad.org

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political Expression and Action on Social Media: Exploring the Relationship Between Lower- and Higher-Threshold Political Activities Among Twitter Users in Italy. *Journal of Computer-Mediated Communication*, *20*(2), 221–239.

Valente, T. W., Coronges, K., Lakon, C., & Costenbader, E. (2008). How Correlated Are Network Centrality Measures? *Connections (Toronto, Ont.)*, *28*(1), 16–26.

Van Stekelenburg, J., & Klandermans, B. (2013). The social psychology of protest. *Current Sociology*, *61*(5–6), 886–905.

Van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin*, *134*(4), 504–535.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Eleventh International AAAI Conference on Web and Social Media*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Waldherr, A., & Wijermans, N. (2017). Modelling the Role of Social Media at Street Protests. In W. Jager, R. Verbrugge, A. Flache, G. de Roo, L. Hoogduin, & C. Hemelrijk (Eds.), *Advances in Social Simulation 2015* (pp. 445–449). Springer International Publishing.

Warf, B. (2011). Geographies of global Internet censorship. *GeoJournal*, *76*(1), 1–23.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (1 edition). Cambridge ; New York: Cambridge University Press.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: Finding Topic-sensitive Influential Twitterers. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 261–270.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence*, *14*(1), 11–29.

Wilensky, U. (1999). NetLogo (Version 6.0.4). Retrieved from http://ccl.northwester.edu/netlogo/

Wojcik, S., Messing, S., Smith, A., Rainie, L., & Hitlin, P. (2018). *Twitter Bots: An Analysis of the Links Automated Accounts Share | Pew Research Center*. Retrieved from https://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/

Wolfsfeld, G., Segev, E., & Sheafer, T. (2013). Social Media and the Arab Spring: Politics Comes First. *The International Journal of Press/Politics*, *18*(2), 115–137.

Won, H.-H., Myung, W., Song, G.-Y., Lee, W.-H., Kim, J.-W., Carroll, B. J., & Kim, D. K. (2013). Predicting National Suicide Numbers with Social Media Data. *PLOS ONE*, *8*(4), e61809.

Woolley, S. C., & Howard, P. N. (2016). Automation, Algorithms, and Politics| Political Communication, Computational Propaganda, and Autonomous Agents — Introduction. *International Journal of Communication*, *10*(0), 9.

World Bank. (2017). *Individuals using the Internet (% of population)*. Retrieved from https://data.worldbank.org/indicator/IT.NET.USER.ZS

Xu, B., & Albert, E. (2014). Media censorship in China. *Council on Foreign Relations*, *25*, 243.

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with AI to counter social bots. *Human Behavior and Emerging Technologies*, 48–61.

Yesil, B., & Sözeri, E. K. (2017). Online Surveillance in Turkey: Legislation, Technology and Citizen Involvement. *Surveillance & Society*, *15*(3/4), 543–549.

Zeifman, I. (2017). *Bot Traffic Report 2016*. Retrieved from https://www.incapsula.com/blog/bot-traffic-report-2016.html

Zhang, H., Hill, S., & Rothschild, D. (2018). Addressing Selection Bias in Event Studies with General-Purpose Social Media Panels. *J. Data and Information Quality*, *10*(1), 4:1–4:24.

Zhdanova, M., & Orlova, D. (2017). *Computational Propaganda in Ukraine: Caught Between External Threats and Internal Challenges* (No. Working Paper 2017.9). Retrieved from Project on Computational Propaganda website: http://comprop.oii.ox.ac.uk/publishing/working-papers/computational-propaganda-in-ukraine-caught-between-external-threats-and-internal-challenges/

Zhu, Q. (2017). Citizen-Driven International Networks and Globalization of Social Movements on Twitter. *Social Science Computer Review*, *35*(1), 68–83.

Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., & Wallach, D. S. (2013). The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions. *ArXiv:1303.0597*.

BIOGRAPHY

Ross Schuchard hails from the 'legendary' state of North Dakota where he graduated from Century High School of Bismarck in 2000. He received a congressional appointment to attend the United States Military Academy at West Point. Following four years as a cadet, Ross graduated from West Point in 2004 with a B.S. in Economics and received his commission as an aviation officer in the U.S. Army. Flight school served as his initial military training assignment, which resulted in his certification rating as a UH-60 Blackhawk helicopter pilot in 2005. For the next eight years, Ross served as an aviation officer amassing over 600 combat flight hours in support of overseas combat operations in Iraq and Afghanistan. In 2014, the U.S. Army selected him to serve as an Operations Research and Systems Analysis (ORSA) officer and immediately sent him to obtain his master's degree in Computational Social Science (CSS) from George Mason University (GMU). He graduated from the CSS program in 2015 and returned back to the operational force to establish the first Data Science cell within the new U.S. Army Cyber Command. The Army then sent Ross back to GMU where he eventually was awarded a PhD in CSS in the summer of 2019.