GRASSROOTS TO VOTING BOOTHS: A STUDY OF THE SPATIOTEMPORAL DYNAMICS OF TRADITIONAL MEDIA COVERAGE AND SOCIAL MEDIA IMPACT IN 2016 UNITED STATES PRESIDENTIAL CANDIDATES

by

Scott Heneghan

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

_____	Dr. Anthony Stefanidis, Thesis Director
_____	Dr. Andrew Crooks, Committee Member
_____	Dr. Arie Croitoru, Committee Member
_____	Dr. Anthony Stefanidis, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Spring Semester 2017 George Mason University Fairfax, VA

Grassroots to Voting Booths: A Study of the Spatiotemporal Dynamics of Traditional Media Coverage and Social Media Impact in 2016 United States Presidential Candidates

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Scott Heneghan
Bachelor of Science
University of Maryland, College Park, 2008

Director: Anthony Stefanidis, Professor
Department of Geography

Spring Semester 2017
George Mason University
Fairfax, VA

DEDICATION

This is dedicated my parents, Daniel and Catherine Heneghan, and fiancée, Victoria, for their love and confidence in me.

ACKNOWLEDGEMENTS

I would like to thank my professors at George Mason University for their expertise and tutelage, my fiancée Victoria for her patience while I attained this degree, my family and friends for their encouragement, and my mentor and friend Joe Governski for his professional guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

OSM ..... OpenStreetMap
EPSG ..... European Petroleum Survey Group
GDELT ..... Global Database of Events, Language, and Tone
URL ..... Uniform Resource Locator
CSV ..... Comma Separated Values (data format)
TSV ..... Tab Separated Values (data format)

ABSTRACT

GRASSROOTS TO VOTING BOOTHS: A STUDY OF THE SPATIOTEMPORAL DYNAMICS OF TRADITIONAL MEDIA COVERAGE AND SOCIAL MEDIA IMPACT IN 2016 UNITED STATES PRESIDENTIAL CANDIDATES

Scott Heneghan, M.S.

George Mason University, 2017

Thesis Director: Dr. Anthony Stefanidis

The subject of social media effect on elections has been studied in multiple peer reviewed journal articles, however the effect of social media on traditional media in elections is not as well studied. This thesis reviews the space time patterns of Twitter and how tweets can potentially correlate to future news coverage from local newspapers in different parts of the United States. This study is being done in relation to the campaigns of Bernie Sanders and Donald Trump. The correlation between one week of tweets and newspaper reports was compared against the tweets from a previous week to the newspaper reports. Local trends were reviewed in order to determine if social media is in fact a driver of change using Pearson product moment correlation. Limited correlation was detected between the values as a result of sparse data from local newspapers, which do not contribute a significant number of reports regarding national elections.

CHAPTER ONE: INTRODUCTION

Information Dissemination in the Web 2.0 Era

The Internet has changed the way that people access and digest current events. Prior to the Internet, options for receiving news were limited to traditional media (i.e. broadcast and print) and personal interactions (i.e. word of mouth). The former presented a top-down, one-to-many avenue of curated mass communication, while the latter reflected a non-curated but physically limited (one-to-few) avenue for information dissemination. The introduction of the Internet affected traditional media first by expanding the number of media outlets that persons had access to, from a limited array of local or national news stations and newspapers to a practically endless list of global sources of such information with often diverse and differing viewpoints.

However, these early days of the Internet still primarily relegated creation of content to technology experts and large organizations, as website generation and maintenance is a complicated undertaking. Accordingly, while the consumers of news were the general public, the producers of this news were still large media outlets like CNN or BBC. Since the mid-2000s, however, the emergence of Web 2.0 extended the level of participation of the general public in the Internet, allowing it not only to access content but to contribute as well (O'Reilly, 2005). This led to the emergence of social media platforms, which allow users to communicate and network with each other, establishing alternate routes for information dissemination and opinion formation, often

bypassing large media outlets to establish a grassroots ecosystem (Kaplan and Haenlin, 2010, p.60). One can now receive news and opinions from friends, family, public figures, and corporations with the same ease as from traditional news sources.

This culture shift allows the general public to be more vocal about facts and opinions, which in turn produces the potential for broadcasting and swaying opinion on political views (Hosch-Dayican et al., 2016, p.21). With a steadily rising number of social media users, each election cycle sees a greater role played by such platforms in shaping opinions and, potentially, affecting the outcome of the election process. The 2016 US election served as a prime example of the power of social media (Guynn, 2016)(Greenwood et al., 2016) with the unprecedented use of Twitter by Donald Trump to communicate with his base and shape the debate narrative (Wells et al., 2016, p.675)

The purpose of this thesis is to study how relevant social media has become in the current political discourse. More specifically, I will study whether social media is beginning to sway what traditional media focuses on (driving official news media coverage, as opposed to following it). In order to do so, I will review social media volume as compared to corresponding newspaper coverage throughout several cities in the United States, using the Trump and Sanders 2016 campaigns.

Both of these candidates began on the fringes of their respective parties with neither of them receiving significant backing from their respective national committees to either force a neck-and-neck race for candidate as was the case with Bernie Sanders, or become the party's candidate and President-elect, as was the case for Donald Trump. The research period for this spatiotemporal analysis spans the nearly year-long period from

the first Republican debates (6 August 2015) to the Democratic National Convention (12 July 2016). This time period will be referred to as the "research period" for this thesis.

Both of these candidates leveraged social media to reach a larger audience. Donald Trump had a significant following on social media sites like Twitter due to his position as a well-known businessman and television celebrity. Bernie Sanders, as a senator from Vermont, did not have as significant a following nationwide. For reference, on October 13, 2015 on Twitter Bernie Sanders has 815,541 followers compared to 4,402,138 for Donald Trump (Trackalytics, 2016). During the research period, and up through the conclusion of this research, both candidates accumulated thousands of followers each day on Twitter, with Sanders reaching 2,215,627 followers (a 270% increase) and Trump reaching 9,662,861 followers (a 219% increase) by the end of the research period (Trackalytics, 2016). These figures are reflective of the ability of social media to provide an alternate venue for these candidates to have their voice heard regularly and without filter by a large portion of the electorate.

Social Media

Social Media is a nearly ubiquitous term in current discourse, however it requires definition in the context of this thesis. The origins of social media can be found in the form of weblogs (commonly shortened to "blog") in the late 1990s. Similar to an online diary, these web platforms allowed users to present discrete posts on a personal webpage and allowed others to comment and respond. These communities allowed users to present ideas and topics to a user group while potentially accepting responses in return. These blogs were not widely dispersed considering they were on separate domains and Internet

penetration remained relatively limited considering access at the time to high speed Internet.

In the early 2000s, the introduction of MySpace in 2003 and Facebook in 2004 became popular sites for users to create individual websites similar to blogs (Kaplan and Haenlin, 2010, p.60). These advancements lowered the bar for users to share information and express thoughts on somewhat customized sites with a defined group of friends on the same domain. Facebook was primarily aimed at students in colleges connecting with one another while being able to perform basic interactions, while MySpace was designed as a blogging platform that allowed users to create a somewhat customized website. At the same time, both Web 2.0 and User-Generated Content were emerging as the new trend for the Internet, allowing users to interact with their content (Kaplan and Haenlin, 2010, p.60).

User-generated content refers to increased participation and contribution from Internet users in order to express opinions, rather than static data which was the standard prior to Web 2.0 (van Beuzekom, 2008, p.8). In earlier days of the Internet, content creation was relegated to users who understood how to create websites using web programming languages like HTML and Javascript. Without knowledge of that and the servers to host data, users were solely consumers of the Internet. This shift can be seen in popular websites like Wikipedia, where users generate encyclopedia articles (Wikipedia:About, 2016). In the geospatial realm websites like OpenStreetMap allow users to create crowd sourced vector data through various means and for any user to be able to manipulate that data (OSM About, 2016).

Following the success of Facebook and Myspace, in 2006 Twitter was introduced, making mainstream the concept of microblogging. Twitter allows registered users the ability to post 140 character messages along with embedded images and videos, referred to as a tweet for the remainder of this thesis. Tweets can also reference previously made tweets (called retweets) which allows users to respond to each other or amplify a statement. Further, users can use a hashtag (by using the # symbol) to reference specific topics. In practice, Twitter serves as a social media tool for discussion with friends, corporations, organizations, and public figures. Most relevant to this thesis, however, is that it acts as a medium for dispersion of news, events, and opinions, including those of a political nature (About Twitter, 2016).

Social Media Source

For this thesis, Twitter data is being used as a representative for social media as it is the most readily available data to acquire. The George Mason University Department of Geography is collecting Twitter data on relevant subjects, including data relevant to this study in the form of tweets mentioning Bernie Sanders or Donald Trump.

Tweet data provided by GMU included relevant fields on each individual tweet. In order to demonstrate the wealth of information contained in a single tweet, Table 1 summarizes accompanying metadata extracted from the Twitter API, as fully documented in the George Mason GeoSocial journal article on their collection methodologies (Croitoru et al., 2013).

Table 1 Data fields from social media data available from George Mason University

Name	Mandatory?	Description
Id	Y	Tweet identifier assigned by twitter.
location	Y	Location name. This field may contain name of the location from which the tweet was sent.
country	Y	If location name was processed through a gazetteer, this field contains country name.
state	Y	State name for tweets processed through the gazetteer. May be empty.
zip	Y	Postal code identified by the gazetteer. May be empty.
X	Y	Longitude of a geolocated tweet (X coordinate on X/Y grid)
Y	Y	Latitude of a geolocated tweet (Y coordinate on X/Y grid)
published_at	Y	Time when the tweet was published. The time may be in one of two formats: • ISO – ISO – Date and time formatted according to ISO8601 international standard: "YYYY-MM-DD hh:mm:ss Z", where YYYY is the year, MM is the month number, DD is the day of the month, hh is the hour of the day in 24 format, mm is the minute, ss is the second and Z is the time zone • ArcScene – Date and time formatted according ArcScene requirements: "YYYYMMDD hh:mm:ss", where YYYY is the year, MM is the month number, DD is the day of the month, hh is the hour of the day in 24 format, mm is the minute and ss is the second. Time zone is UTC.
author	Y	Author of the tweet.
coords_from	Y	Information about source of the geographic coordinates in x,y fields: • blank – for tweets without geolocation information. • location – for tweets geolocated using place name from location field in the tweet processed through the geocoder. • twitter – Location information has been determined/guessed by twitter based on available data (IP address, etc). • coords – when tweets was sent with geolocation information (geographic coordinates of the place from where the tweet For analysis on scales finer that a country level, location-based geolocation may be too rough and only tweets with coords_from set to coords should be considered.

Name	Mandatory?	Description
mood	N	Information about mood of the tweet. Negative values indicate negative mood, positive values indicate positive mood, zero means neutral or not computed. Information available only for tweets written in English.
retweeted_id	N	Present only for retweets. The value in this field contains ID of the tweet which has been retweeted.
response_id	N	Present only for tweets which are responses to previous tweets.Contains ID of the tweet to which the current tweets responds.
response_author	N	Present only for tweets which are responses to previous tweets. Contains handle of the author of the tweet to which the current tweets responds.
text	N	Text of the tweet.
links	N	Links extracted from the tweet.
topic_word1 *topic_word2* *topic_word3* *…* *topic_wordn*	N	Flags for topic words. Each column informs whether the text of the tweet contains the word for which the column has been generated. Values 0 – topic word not detected 1 – topic word detected

In total, approximately 25 gigabytes of data was delivered in response to the query for all tweets related to Donald Trump or Bernie Sanders in zipped tab spaced value (TSV) format. These files contained over 229 million tweets about Donald Trump and over 41 million about Bernie Sanders for the entire research period. A full breakdown of the tweet corpus including the source of geocoded tweets is included in table 2.

Table 2: Tweet Corpus for Delivered Data from GMU

	Bernie Sanders	Donald Trump
Total Number of tweets	41,240,350	229,150,400
Geocoded Tweets	21,476,765	112,326,129
location	20,672,039	107,463,558
twitter	675,181	4,161,341

coords	129,545	701,230
Number of Retweets	25,038,720	134,690,100

To isolate only relevant data, the points with no latitude and longitude were expunged from the dataset. Implied locational data from either the place names geocoded to a city center or Twitter determining the location by IP (Internet Protocol) address were considered in addition to precisely geocoded tweets as precise location was not necessarily required for this analytic study. Retweets were included in this study as they represent a second voice making the same statement or opinion.

Traditional Media in the Context of this Thesis

Traditional media, in the context of this thesis, is a catchall description of any media type that existed before the medium of the Internet. These sources include newspapers, television, radio, and magazines. In theory, news reports from traditional media are designed to be factual, and providing opinions (known as editorializing) is discouraged. In practice, there are tonal differences in news media, which can be derived from the editorial staff, reporters, or publishers of the content. This tonal disparity is especially noticeable in the discussion of political candidates (Pew Research Center, 2012).

There is some overlap in these two spaces where traditional media sources like newspapers hold social media accounts. These accounts primarily serve as a dispersion mechanism to send new articles to the followers of that news service. This allows the potential for counting articles twice in my research as both social and traditional media.

In practice, it would be difficult to remove all tweets made by traditional media platforms without a comprehensive whitelist of all those Twitter IDs. Instead, it is a limitation in this research.

As noted, traditional media sources are discouraged from editorializing a topic with one large exception: the editorial section. These sections allow for the editor of the paper as well as write-in readers to provide opinions on news related topics either in general or related to a specific article that the newspaper previous printed. This is another space where social media and traditional media overlap with write-in editorials as a sort of curated user generated content.

Editorials provide an interesting dynamic to this thesis, where opinions are welcome but only the ones that are validated by the paper are printed. These are spaces that can show potential slant in media sources. While outside the scope of this current thesis, editorials can be a potential future research project that would further explore the biases of media sources.

Traditional Media Source

Numerous sources were considered in order to quantify and analyze traditional media for this thesis.

The Global Database of Events, Language, and Tone (GDELT) was considered firstly as a source for analyzing traditional media. GDELT is a worldwide database of events captured from various print and broadcast media, categorized by type and rated by tone. A shortcoming regarding the use of GDELT for this thesis is it strives to quantify events rather than news articles, multiple articles regarding an event are commonly

condensed into a single record. Additionally, due to the international nature of the database, it relies more heavily on national news and wire sources, which are not presented at a small enough spatial specificity to be useful for this thesis (The GDELT Story).

An alternative data source is the Phoenix Data Project. Phoenix, like GDELT, is a worldwide categorized event database; the difference between the two lies in the open source nature of the methods used to create Phoenix's data fields. However, it is at its heart still an analysis of international events and focuses primarily on larger news outlets in the US. Also like GDELT, it strives to isolate events rather than news stories, and analyzing the same news story from multiple sources is at the crux of the goals of this thesis (Phoenix Data Project).

The source chosen for this analysis ultimately was the Newspaper Source Plus database made available by the George Mason University library. This source, from EBSCO, provides a full-text of the major news content from over 1200 newspapers (Newspaper Source Plus). Using search terms for both candidates, all news articles that matched the candidates could be reviewed individually. It also allows for review of smaller news sources as compared to the international sources from Phoenix and GDELT.

The search terms used for acquiring this data were "Donald Trump" and "Bernie Sanders." The search was limited only to US-based newspapers. Large national news sources that do not have specific geographic focuses or had significant national readership were excluded; those sources were the New York Times, the Wall Street

Journal, the Boston Herald, and USA Today. Additionally, in order to analyze only significant temporal trends, only news outlets that had reported on Donald Trump and Bernie Sanders in a significant number of articles were included. The result is 21 newspapers from smaller cities across the country. Table 2 shows the names of these sources as well as their potential slant. These biases were captured primarily from a site allowing readers to vote on the slant of an outlet.

Table 3: Traditional Media Outlets and Projected Slant

Outlet Name	**Projected Slant**
Akron Beacon Journal (OH)	Left-Center (Media-bias/Fact Checking, 2016)
Austin American-Statesman (TX)	Unknown
Blade, The (OH)	Unknown
Buffalo News, The (NY)	Left-Center (Media-bias/Fact Checking, 2016)
Columbus Dispatch, The (OH)	Right-Center (Media-bias/Fact Checking, 2016)
Daily Oklahoman, The (OK)	Right-Center
Dallas Morning News, The (TX)	Right-Center (Media-bias/Fact Checking, 2016)
Dayton Daily News (OH)	Unknown
Gazette, The (Cedar Rapids, IA)	Unknown
Gazette, The (Colorado Springs, CO)	Right-Center (Media-bias/Fact Checking, 2016)
New Hampshire Union Leader (Manchester)	Unknown
Orange County Register, The (Santa Ana, CA)	Right-Center (Media-bias/Fact Checking, 2016)
Oregonian, The (Portland, OR)	Right-Center (Media-bias/Fact Checking, 2016)

Palm Beach Post, The (FL)	Unknown
Philadelphia Inquirer, The (PA)	Left-Center (Media-bias/Fact Checking, 2016)
Pittsburgh Post-Gazette (PA)	Left-Center (Media-bias/Fact Checking, 2016)
Pittsburgh Tribune Review (PA)	Right-Center (Media-bias/Fact Checking, 2016)
San Diego Union-Tribune, The (CA)	Right-Center (Media-bias/Fact Checking, 2016)
St. Louis Post-Dispatch (MO)	Left-Center (Media-bias/Fact Checking, 2016)
Star Tribune (Minneapolis, MN)	Left-Center (Media-bias/Fact Checking, 2016)
Wisconsin State Journal (Madison, WI)	Unknown

These results were geocoded to their city center locations in order to create buffers that would be used to isolate nearby tweets as well as display on graphics used for this thesis. The result of this process can be seen in the Figure 1, a map displaying the chosen outlets with their buffers. While it only appears like there are 20 locations, two of the outlets were from Pittsburgh, Pennsylvania and therefore overlap exactly.

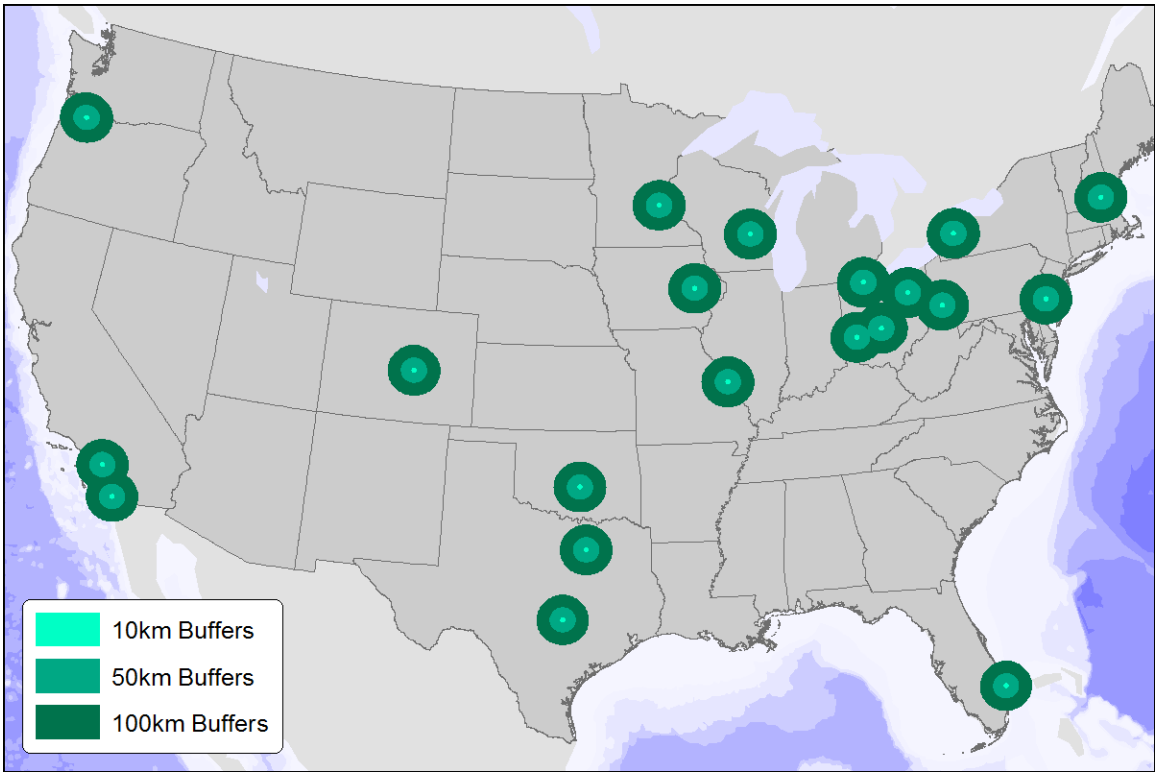

Figure 1: Map of Traditional Media Outlets and Buffers

Initially, the desire was to compare daily data for social and traditional media. However, on review of the data, these smaller news outlets do not have significant coverage of the political candidates in national elections. Most of the news outlets listed above made around 100 references to the candidates, which does not even account for one article per day. This appears to be a result of these smaller news stations not having the reporting staff necessary to cover national news events; a cursory review of the website of some of these outlets show that the majority of the presidential election coverage was from wire services like the Associated Press (Akron Beacon Journal Search Results)(Wisconsin State Journal Search Results).

CHAPTER TWO: BACKGROUND RESEARCH

Numerous studies have been performed on reviewing the effects of social media including several on the effects of election. However, the study of how traditional media is influenced by social media in the context of an election has not been studied in peer reviewed sources.

One of the most widely references studied was research that concluded that an election in Germany can be determined based on analysis of Twitter. The deterministic factor is that the number of tweets referencing a specific political party directly correlated to the number of votes for that political party (Tumasjan et al, 2010, p.183). This was completed by studying 104,000 tweets in the weeks prior to an election in Germany. The conclusion that elections can be predicted using Twitter is dubious; this research only focuses on a small time period prior the election. Further, the spatial trend of voting as compared to tweeting is not explored; without understanding where people are voting and how those users tweeted, it is difficult to determine if you are recognizing a coincidental trend or if it actually matches reality (Tumasjan et al, 2010, p.183). Additionally, social media participation is not necessarily deliberative; it would be interesting if through the article they were able to detect users that were changing their political affiliations as an effect of social media (Tumasjan et al, 2010, p.183). The Tumasjan results are refuted in

a response paper, which shows that the results from Tumasjan are specifically derived from the arbitrary temporal choices made by the authors (Jungherr et al, 2012, p.230).

Owing to the questionable nature of this study, another research that came to a similar conclusion, that volume of tweets makes for a reasonable predictor, needed to be uncovered. Using Twitter as a method to monitor political sentiment was performed during the Irish General Election in 2011 (Bermingham and Smeaton, 2011, p.2). The analysis comes to significantly less lofty conclusions as compared to Tumasjan, where they find a limited error from non-parametric analysis of the data (Bermingham and Smeaton, 2011, p.9). Additionally, they find that volume and sentiment analysis derive roughly the same conclusion about the result of the election. This preponderance of evidence can reasonably be said to allow us to continue with my research based on the volume of social media data as compared to traditional media. A similar study on the Dutch Senate election of 2011 found that the methodology of the collection is most relevant—removing tweets mentioning more than one party name (and in the case of this study, candidates) and multiple tweets from the same user help to make for a better dataset to derive results (Sang and Bos, 2012, p.8).

Using Twitter data in the context of the 2016 national election has also been completed in comparing the tweet sentiment to the polling scores. Analysis found predicted Twitter opinions matched well the New York Times polling average scores (Bovet et al, 2016, p.9). This does predict correctly that Hillary Clinton would win a larger number of votes, however the paper does acknowledge that the popular vote does not determine presidency (Bovet et al, 2016, p.10). This introduces a limitation that may

require further research on state based studies. If a similar study were conducted at the state level for every state, it may have more accurately predicted the outcome of the race.

Social media has been shown to predict outcomes better than traditional standards outside of elections as well. A research paper on comparing the buzz on social media with the profitability of a movie in theaters has proven more effective than a benchmark called the Hollywood Stock Exchange (Asur and Huberman, 2010, p.7). This paper shows that there is usefulness in social media data that can be used as a predictor for outcomes outside of it. My research attempts to do something similar, though the prediction is for how newspaper publishers react as opposed to moviegoers.

With regards to politics and the effect of social media, older studies are available that show the effect of blogs on political discourse. Woodly in 2007 found that blogs are especially effective at mobilizing opinions with regards to politics (Woodly, 2007, p.122). This supports the premise of web blogs being able to change opinions; it is reasonable to assert that if it is changing the mind of readers of the blog it can potentially shift the thinking of a newspaper publisher. As the article is from 2007, Twitter was still too new for this paper to reference microblogs but rather longer blogs. There are no immediately discoverable studies that review the effects of social media on traditional media especially in reference to an election.

While outside the context of political discussion, the effects social media users have on each other is a much better studied phenomena. Because of the overwhelming amount of content to read on Twitter, users are forced to form quick opinions on user credibility, users makes a quick judgment call about the credibility of the tweet and

author based on the text, avatar, and username (Morris et al, 2012, p.9). It is found that this is a poor predictor of how credible the user is, and therefore users are likely to undervalue a tweet when it comes from a source they do not recognize (Morris et al, 2012, p.9). Further, social media has been studied as a platform for persuasion—nearly half of the tweets collected for analyzing politics were geared towards convincing other users to vote similarly (Hosch-Dayican et al, 2016, p.21). These two conclusions indicate that users have the potential to generate echo chambers on their social media if they only follow users of similar opinion.

My study focuses on a method by which less mainstream candidates or even opinions can gain traction on social media. As noted in the introduction, prior to this study period Bernie Sanders and Donald Trump had more than 825,000 and 4.4 million respectively, which is a large number of people who are potentially influenced by an opinion on Twitter (Trackalytic, 2016). If, for instance, factual rebuttals to those tweets were posted by users that the Twitter user was not following, they are unlikely to accept those tweets as credible (Morris et al, 2012, p.9).

Background research has also revealed one notable point in political discourse: the influence of how a candidate's other engagements may affect traditional media's coverage of them (Scharl and Weischselbraun, 2008, p.131). If the candidate has previous engagements that have affected the way that traditional media has treated them in the past, it paints the way in which they are treated moving forward. This is especially relevant in the case of this research as Bernie Sanders and Donald Trump have both had significantly different interactions with traditional media prior to their candidacy for

President. Bernie Sanders has had a much more traditional route towards president, and traditional media outside of Vermont, the state he represents in the US Senate, has rarely reported on him. There was some national coverage of him prior to the campaign owing to his position on numerous committees in the US Senate as well as a 2010 filibuster against the Middle Class Tax Relief Plan (Memoli, 2010). Donald Trump, however, has had a significantly different progression to candidacy that was covered very differently in news media. Donald Trump, as a celebrity businessman, has been the subject of national media sources for a significant time leading up to his run in the 2016 race.

It is also relevant to understand how the users of Twitter vary from the general population and how that portion of the population has shifted since the last presidential election in 2012. Twitter has historically been more popular among younger and non-white users (Duggan and Brenner, 2013). However, in the time since the 2012 election, social media adoption has seen significant growth in white, higher income, and more educated audiences (Greenwood et al., 2016). This shift and the specific shifts can be seen in the demographics in Table 2 below:

Table 4: Percent of Twitter Users by Demographic 2012, 2016

	2012	2016	Percentage Shift
Total	16	24	33.33%
Men	17	24	29.17%
Women	15	25	40.00%
Race/ethnicity*			
White, Non-Hispanic	14	20**	30.00%
Black, Non-Hispanic	26	28**	7.14%
Hispanic	19	28**	32.14%
Age			
18-29	27	36	25.00%

30-49	16	23	30.43%
50-64	10	21	52.38%
65+	2	10	80.00%
Education attainment			
High School Grad or Less	15	20	25.00%
Some College	17	25	32.00%
College+	15	29	48.28%
Household income			
Less than $30,000/yr	16	23	30.43%
$30,000-$49,999	16	18	11.11%
$50,000-$74,999	14	28	50.00%
$75,000+	17	30	43.33%
Urbanity			
Urban	20	26	23.08%
Suburban	14	24	41.67%
Rural	12	24	50.00%

**Pew Research data for 2016 unavailable by race. Instead it is from 2015
Data from Pew Research (Greenwood et al., 2016) (Duggan and Brenner, 2013) (Duggan, 2015)

Firstly, it is important to note from this chart that social media, especially Twitter in this case, does not represent the electorate as a whole but, rather, a partially biased sample of it. And additionally, it is unknown what portion of these users are voters. However, if social media continues at the trend shown in Table 4, the continued uptake will potentially improve analysis going forward.

From Table 4, the spread between demographic breakdowns is relatively even with 24% of total Internet users using Twitter. Among several specific groups, Twitter is more popular: 28% of Black non-Hispanic, 36% of persons between 18 and 29 years old, and 26% urban. And there are several groups which are less represented: users 50+ years

old and suburban and rural users. Both of these represent a trend: the most represented ages are younger and urban. The urban and rural disparity will be interesting considering the analysis in this research is using variable sized buffer rings around city centers (Greenwood et al., 2016) (Duggan and Brenner, 2013). The dichotomy between urban and rural users of social media has been further studied in the way they interact with social media as well including the geographic spread of their friend groups and a higher rate of using private profiles (Gilbert et al., 2012, p.1383).

This trend was similar in both the 2012 and 2016 survey. Though in the four years between surveys, the user group has expanded most significantly with users 30-49, users with more education, and a higher household income. These increases far outpace the total increase from 16-23% that the total population experienced (Greenwood et al., 2016) (Duggan and Brenner, 2013).

Another Pew survey of interest questioned participants if they received their news from social media sites like Twitter. In 2016, it was found that 62% of social media users have used it to get news and 18% do it often. This has been increasing since 2013, indicating that more users are going to social media to get information on current events and potentially form opinions. The Pew survey does not specify if the users are seeking out the Twitter accounts of traditional media outlets however like BBC or CNN, which is just a separate dissemination method. Even so, by following traditional media outlets on social media sites, those stories are then juxtaposed with other potentially less veracious reporting (Gottfried and Shearer, 2016).

CHAPTER THREE: METHODOLOGY

For the purpose of this study, a dataset representing both traditional and social media is required. As discussed in the sources section of this thesis, they are the Newspaper Source Plus for traditional media and Twitter data acquired from George Mason University for social media.

The purpose of the study of these datasets is to determine if there are any significant temporal patterns of the way the candidates were portrayed in social media as compared to traditional media and whether social media has potentially influenced the traditional media coverage of a candidate at the level of small newspaper outlets. The working theory is that social media has allowed these fringe candidates that were not readily accepted by the traditional media as serious candidates to become mainstream because of how their large online following drove discussion on social media. Reacting to that social media increase in communication, traditional media provided additional coverage for these candidates that legitimized their candidacy.

This entire methodology is outlined in Figure 2, a flow chart of the operations performed on the data prior to the statistical analysis detailed in the conclusions section.

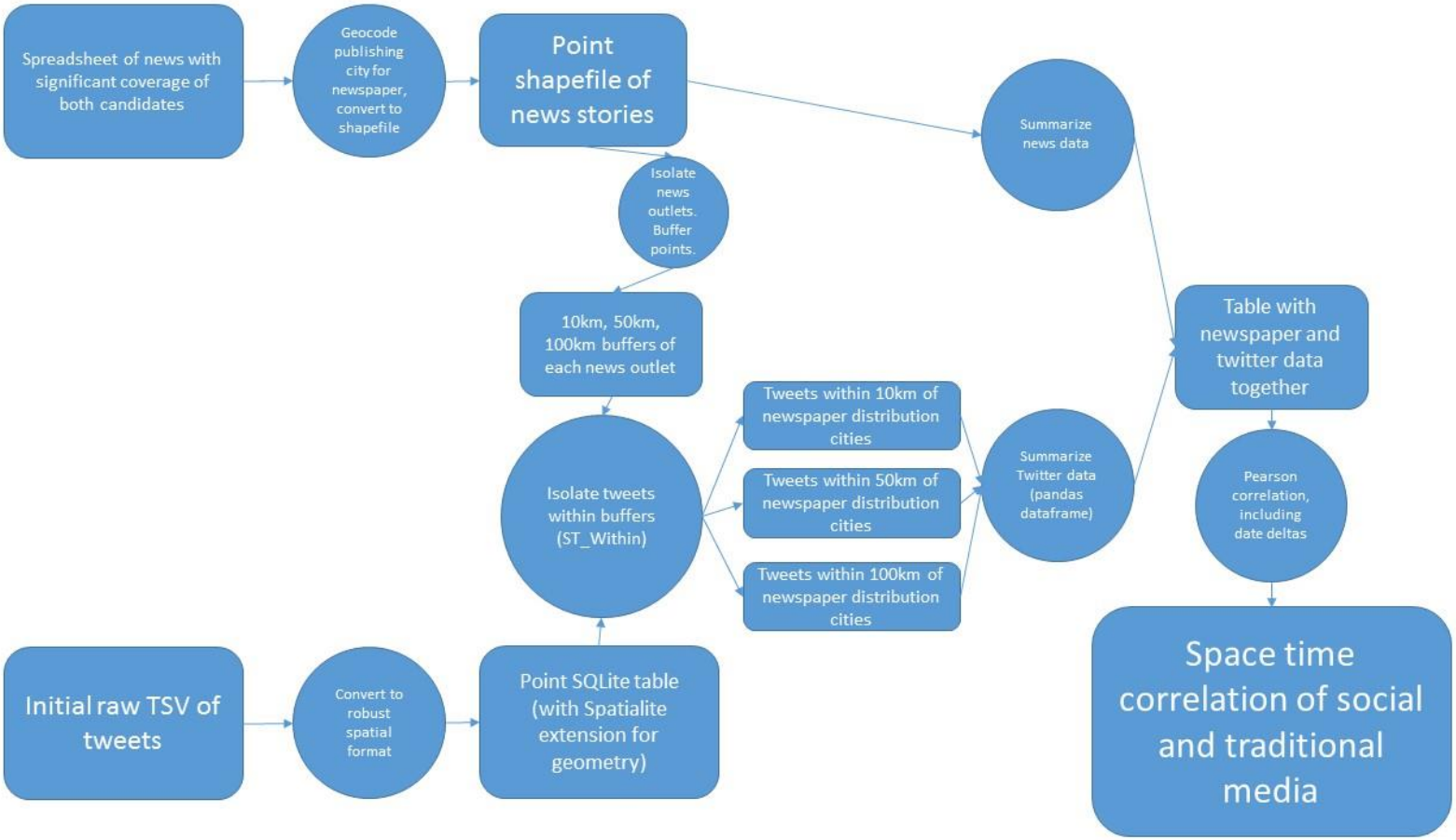

Figure 2: Flowchart of the Analytic Methodology for Traditional and Social Media

In order to study this, I need to quantify the way in which newspaper articles and social media can be compared. To that end, the volume of both the social media and traditional media articles are compared in space and time to compare coverage of Donald Trump or Bernie Sanders.

The traditional media, as it is reasonably sized, was managed using Microsoft Excel. Each article that resulted from the keyword search for each candidate's name was included in a file. The resulting document of all newspaper articles was then reviewed to determine which news outlets had too few mentions and which came from national papers such as the New York Times or Wall Street Journal. Both of these categories of data were removed in order to analyze the trend only at small spatial locations

Once the newspapers were chosen, the coordinates for their city centers were determined in order to buffer them to isolate tweets only relevant to that city. For the buffers, a distance of 10, 50, and 100km was chosen in order to see potential spatial dispersion from city center to see how tweets are varying based on distance from city center. In order to accurately create buffers to query only the relevant tweets, the North America Equidistant Conic projection (EPSG:102010) was used in order to minimize distance based distortion. Following the buffer creations, all further spatial computations were performed in unprojected World Geodetic System 1984 (WGS84, EPSG:4326).

The social media data was significantly larger, as noted in the introduction to the source. Microsoft Excel could not be employed as a method to do research based on this larger dataset. For this part of the research, the data was first converted from the tab separated value format to a Spatialite database in order to speed queries and the spatial

join that would have to occur between the buffers and the tweet point data. Using the ST_Within command, tweets within the buffer ring were made into separate tables and exported as comma separated value (CSV) files.

With CSV files now that only represented the tweets made near cities, the data had to be further distilled in order to determine the total number of tweets made near each city center location. With the data still too large for Microsoft Excel, python with the pandas library was instead leveraged. Using a jupyter notebook, the code shown in below in Figure 3 was run against the CSVs in order to concatenate them, which was relevant in the case of the Donald Trump tweets as they were delivered in multiple CSVs. The code then creates a pivot table from that data and exports it to a new CSV file.

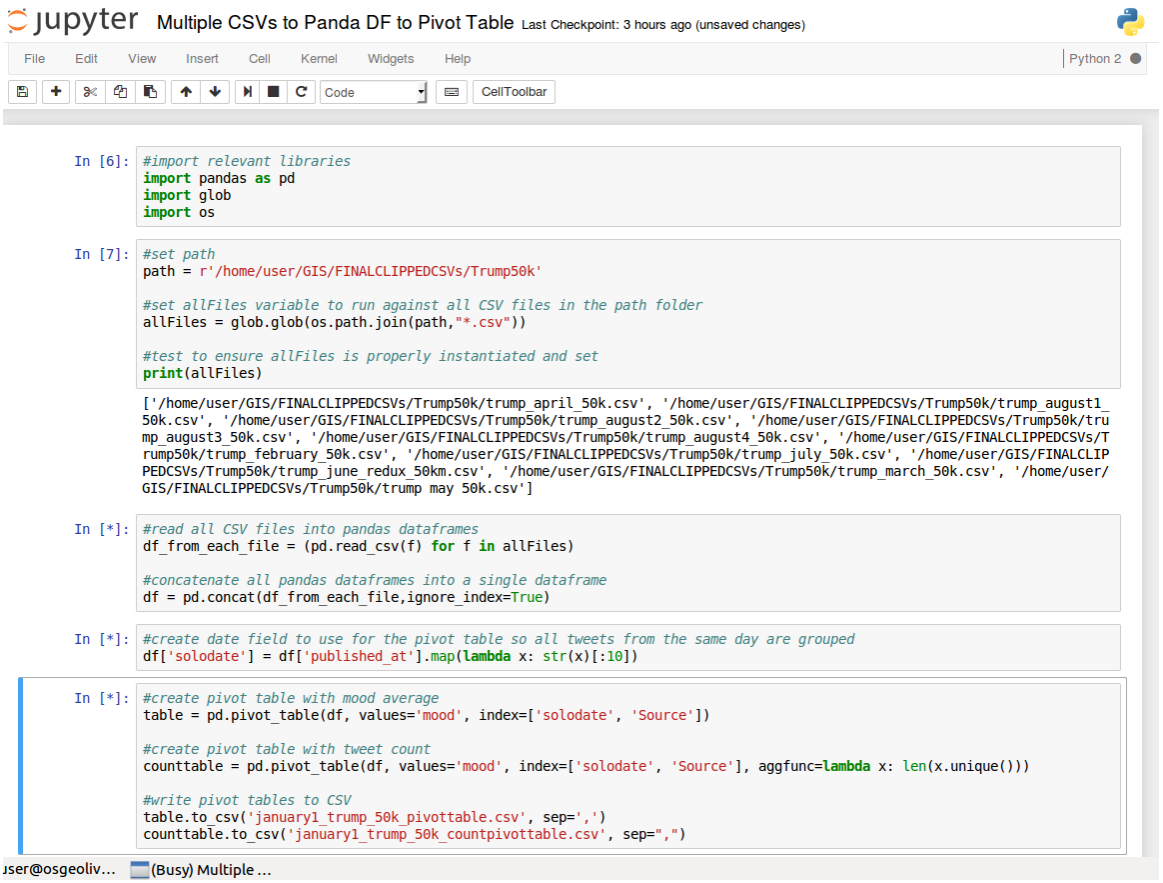

Figure 3: Python Notebook for Processing CSV to Pivot Table

This data from social media and traditional media were then combined together using Microsoft Excel with each record of the spreadsheet showing:

- Date
- Media Outlet
- Total count of tweets about Bernie Sanders within 10, 50, and 100km of the media outlet (three columns)
- Total count of tweets about Donald Trump within 10, 50, and 100km of the media outlet (three columns)

- Number of media mentions of Bernie Sanders by the media outlet
- Number of media mentions of Donald Trump by the media outlet

The final step in the analytics was to determine the correlation. As noted in the traditional media section, significant gaps in the data required aggregation from daily to weekly data. Both the traditional media and social media were aggregated to this level in order to perform the correlations.

The hypothesis of this research is that social media is more readily responsive to news stories compared to traditional media outlets in small town America. To test this, the correlation between newspaper articles and the tweets from the same vicinity will be compared, both at the same week, as well as with a week-long lag between social media traffic and corresponding news stories. If correlation is significantly more positive between traditional media and the lag week tweets as compared to the same week, it indicates that social media discussion precedes traditional media.

For this process, the Pearson product moment correlation test was employed to assess the correlation between traditional and social media. Pearson product moment was ultimately chosen because of its flexibility in dealing with non-monotonic series with the potential for zero values. The correlation test determines a linear equation associated for each data series and then compares the similarity of the two in order to calculate a correlation score. For each newspaper outlet, this test was run six times: once for each week at each buffer distance to determine correlation between the same week news and social media posts made within 10, 50, and 100km from the traditional media location,

and once for the lag week news and social media at the same distances to test the hypothesis.

For the significance test, the null hypothesis is that there is no correlation between the social and traditional media. The alternative hypothesis is that there is significant correlation between the two media types. Further, once these statistical tests are run, the numbers will be considered against each other to determine if there is higher significance from the same week or from the previous week's social media data.

CHAPTER FOUR: RESULTS

As noted in the methodology section, the correlation between traditional and social media is analyzed twice in order to account for both the same week and the lag week with the results then compared against each other. For each media outlet and its buffers from 10, 50, and 100km, tweet volumes were compared against the number of media reports. Firstly, the number of observations in each category is shown in Table 5.

Table 5: Total Number of Observations from Traditional and Social Media

Outlet	News Reports on Sanders	News Reports on Trump	Tweets about Bernie within 10km	Tweets about Bernie within 50km	Tweets about Bernie within 100km	Tweets about Trump within 10km	Tweets about Trump within 50km	Tweets about Trump within 100km
Akron Beacon Journal (OH)	22	71	1978	4099	4347	4225	6830	7361
Austin American-Statesman (TX)	48	160	4765	4848	4911	7086	7277	7486
Blade, The (OH)	58	158	15	4804	5121	4447	5209	7756
Buffalo News, The (NY)	56	97	3590	3992	5007	5894	6408	7943
Columbus Dispatch, The (OH)	91	283	3633	4959	5200	6196	8223	8431
Daily Oklahoman, The (OK)	49	63	588	2371	4001	1876	4681	6474
Dallas Morning News, The (TX)	45	181	4397	5056	5211	7335	8351	8454
Dayton Daily News (OH)	42	95	2613	3031	4607	4786	5682	7508
Gazette, The (Cedar Rapids, IA)	76	56	1702	2715	3731	3281	4674	6535
Gazette, The (Colorado Springs, CO)	20	59	2764	2873	4766	5396	5613	7616
New Hampshire Union Leader (Manchester)	108	97	1615	3480	5478	3456	5663	8394
Orange County Register, The (Santa Ana, CA)	41	69	3516	6348	6714	5738	9719	10188
Oregonian, The (Portland, OR)	72	96	4887	5520	5584	7133	8500	8572
Palm Beach Post, The (FL)	30	214	2655	3749	4852	5710	6838	8018

Philadelphia Inquirer, The (PA)	136	198	4856	5826	6068	7478	9047	9403
Pittsburgh Post-Gazette (PA)	42	79	4301	4551	4751	6904	7380	7742
Pittsburgh Tribune Review (PA)	55	101	4301	4551	4751	6904	7380	7742
San Diego Union-Tribune, The (CA)	56	132	4830	5045	5275	7456	7724	7899
St. Louis Post-Dispatch (MO)	87	154	3894	4215	4256	3684	7003	7074
Star Tribune (Minneapolis, MN)	65	77	3955	4450	4505	6110	6767	6854
Wisconsin State Journal	25	92	57	859	3733	226	2795	5967

The Pearson product moment correlation value for each of the media outlet was assessed using the following function:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Equation 1: Test statistic for Pearson product moment correlation

This function compares the observed value (tweet count) as compared to the expected value (newspaper count). The results of this correlation are -1 if there is an exactly negative correlation between the values, 0 if there is no correlation, and +1 if there is an exactly positive correlation. The result of these calculations comparing the same week is included below in Table 6, with low numbers highlighted in red, high in blue:

Table 6: Pearson Product Moment Correlation Scores, Same Week

Pearson Score Results, same week	News Results to Tweet Count (10k, Bernie Sanders)	News Results to Tweet Count (50k, Bernie Sanders)	News Results to Tweet Count (100k, Bernie Sanders)	News Results to Tweet Count (10k, Donald Trump)	News Results to Tweet Count (50k, Donald Trump)	News Results to Tweet Count (100k, Donald Trump)
Akron Beacon Journal (OH)	0.46715	0.50659	0.49143	0.24883	0.17817	0.15789
Austin American-Statesman (TX)	0.04171	0.04125	0.04295	0.48432	0.47385	0.46115
Blade, The (OH)	-0.3118	0.24079	0.26366	0.50479	0.49662	0.43572
Buffalo News, The (NY)	0.37081	0.35031	0.29666	0.56041	0.52256	0.45162
Columbus Dispatch, The (OH)	0.18106	0.207	0.21431	0.44964	0.44602	0.43923
Daily Oklahoman, The (OK)	0.2003	0.38558	0.39042	0.49596	0.35209	0.34705
Dallas Morning News, The (TX)	0.10563	0.14266	0.12397	0.34153	0.32625	0.32181
Dayton Daily News (OH)	0.6315	0.62142	0.54826	0.55659	0.45587	0.43421
Gazette, The (Cedar Rapids, IA)	0.3466	0.33014	0.26113	0.0129	0.06409	-0.08816
Gazette, The (Colorado Springs, CO)	0.2469	0.29367	0.29536	0.17463	0.22553	0.30814
New Hampshire Union Leader (Manchester)	0.4383	0.21899	0.23357	0.28612	0.18295	0.0979
Orange County Register, The (Santa Ana, CA)	0.10549	0.12245	0.09878	0.28858	0.3214	0.30309
Oregonian, The (Portland, OR)	0.14412	0.1223	0.1219	0.39191	0.37013	0.37037
Palm Beach Post, The (FL)	0.07882	0.04865	0.0331	0.36442	0.42895	0.4599
Philadelphia Inquirer, The (PA)	0.38722	0.36342	0.3517	0.34871	0.31073	0.29507
Pittsburgh Post-Gazette (PA)	0.21324	0.21056	0.20044	0.38737	0.40151	0.41413
Pittsburgh Tribune Review (PA)	0.34616	0.34304	0.34135	0.44921	0.45309	0.4555

San Diego Union-Tribune, The (CA)	0.33556	0.32049	0.32719	0.19147	0.20885	0.19646
St. Louis Post-Dispatch (MO)	0.35441	0.35191	0.34922	0.48325	0.43264	0.43761
Star Tribune (Minneapolis, MN)	0.27646	0.26623	0.27745	0.2963	0.32527	0.31933
Wisconsin State Journal	0.59773	0.85983	0.52103	0.41432	0.5119	0.47449

This data table is also displayed as figures to show spatial trends in the data. See figures 4 and 5 below:

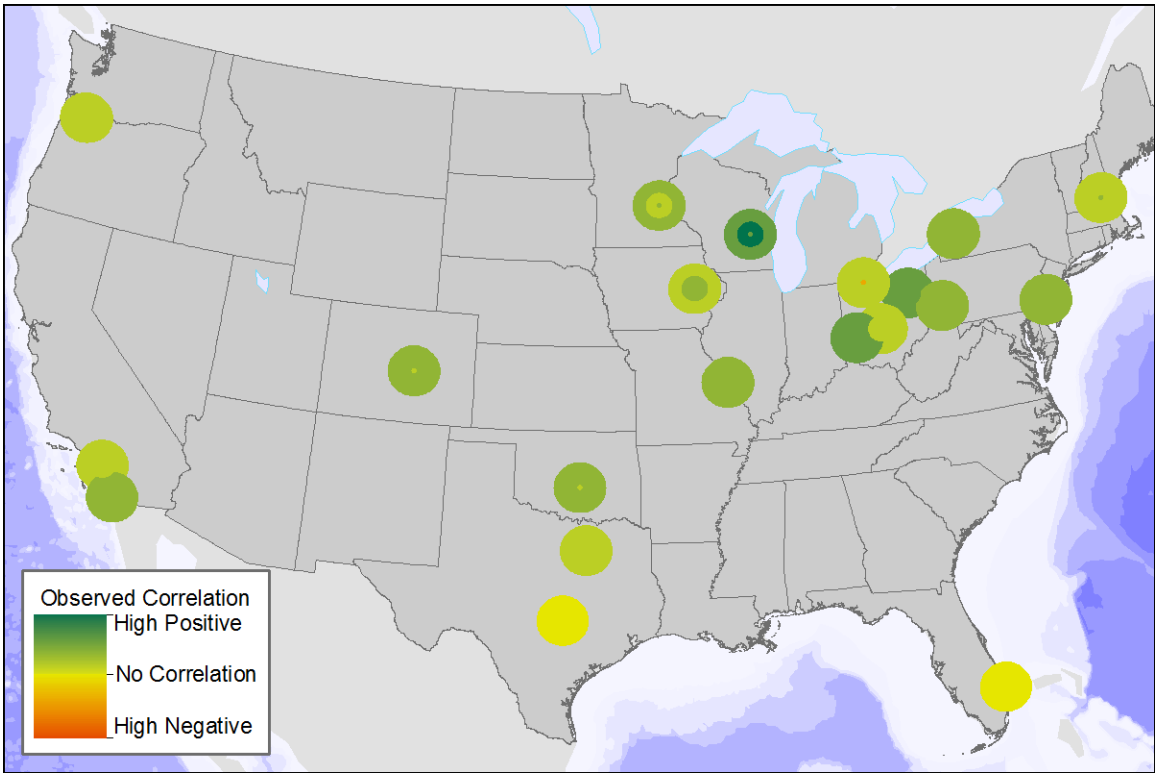

Figure 4: Map of Correlation of News and Social Media, Bernie Sanders, Same Week

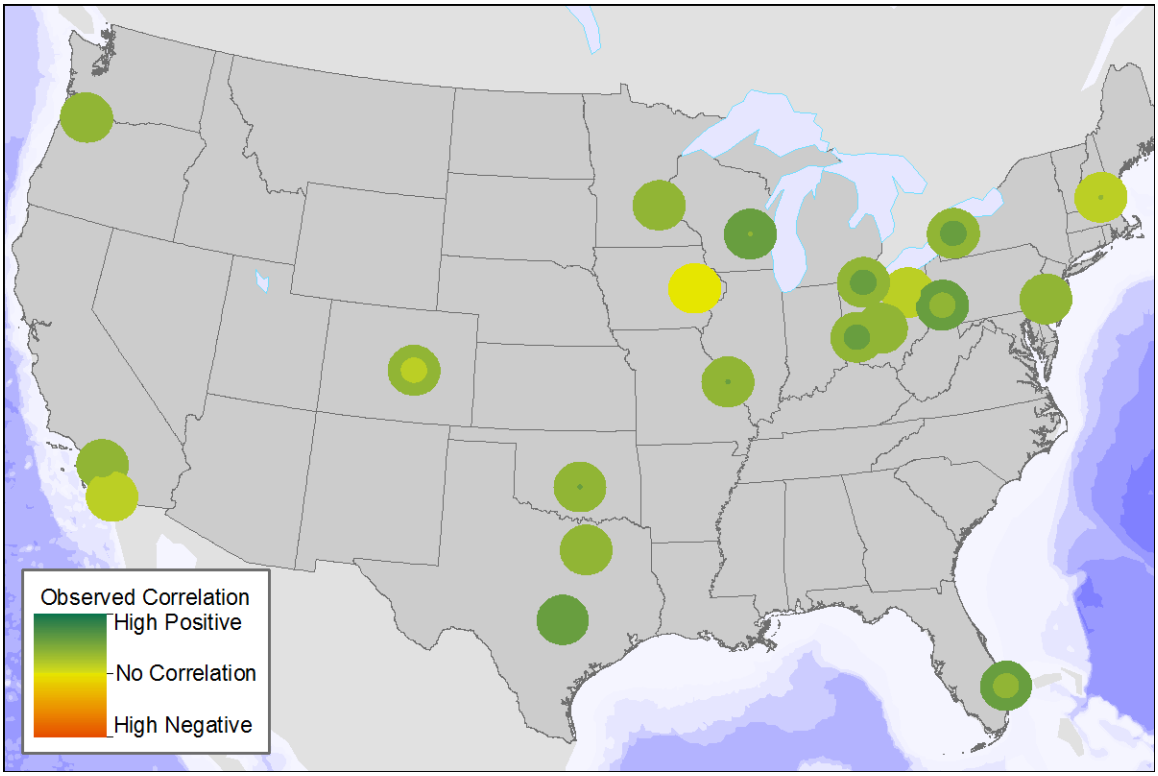

Figure 5: Map of Correlation of News and Social Media, Donald Trump, Same Week

The data table showing the values from the lag week version are included in Table 7 below, again with low numbers highlighted in red, higher in blue:

Table 7: Pearson Product Moment Correlation Scores, Lag Week

	News Results to Tweet Count (10k, Bernie Sanders)	News Results to Tweet Count (50k, Bernie Sanders)	News Results to Tweet Count (100k, Bernie Sanders)	News Results to Tweet Count (10k, Donald Trump)	News Results to Tweet Count (50k, Donald Trump)	News Results to Tweet Count (100k, Donald Trump)
Pearson Score Results, lag week						

Akron Beacon Journal (OH)	0.2466	0.3390	0.2829	0.3808	0.4031	0.4232
Austin American-Statesman (TX)	-0.1263	-0.1346	-0.1441	0.4723	0.4860	0.4866
Blade, The (OH)	-0.2209	0.0664	0.0925	0.5164	0.4603	0.4899
Buffalo News, The (NY)	0.3994	0.3888	0.2948	0.5038	0.5294	0.5193
Columbus Dispatch, The (OH)	0.1316	0.1108	0.1060	0.5399	0.5051	0.5181
Daily Oklahoman, The (OK)	-0.0161	0.2608	0.2151	0.3378	0.3481	0.3617
Dallas Morning News, The (TX)	-0.0506	-0.0715	-0.0803	0.4099	0.4177	0.4184
Dayton Daily News (OH)	0.6216	0.5773	0.5507	0.4152	0.3785	0.4793
Gazette, The (Cedar Rapids, IA)	0.1905	0.2155	0.1778	-0.2428	-0.1351	-0.1217
Gazette, The (Colorado Springs, CO)	0.3618	0.3660	0.2502	0.0713	0.0553	0.1567
New Hampshire Union Leader (Manchester)	0.2861	0.1237	0.1506	-0.0449	0.0575	-0.0421
Orange County Register, The (Santa Ana, CA)	0.2850	0.2351	0.2265	0.1931	0.2459	0.2766
Oregonian, The (Portland, OR)	0.2484	0.2415	0.2374	0.3273	0.3429	0.3442
Palm Beach Post, The (FL)	-0.2569	-0.2094	-0.2377	0.4507	0.5174	0.5365
Philadelphia Inquirer, The (PA)	0.2862	0.2952	0.2872	0.3594	0.3958	0.4070
Pittsburgh Post-Gazette (PA)	0.2863	0.2779	0.2618	0.3785	0.4017	0.4072
Pittsburgh Tribune Review (PA)	0.1510	0.1448	0.1417	0.6497	0.6323	0.6298
San Diego Union-Tribune, The (CA)	0.3167	0.3111	0.3020	0.1666	0.1469	0.1288
St. Louis Post-Dispatch (MO)	0.3249	0.3160	0.3196	0.5449	0.4652	0.4706
Star Tribune (Minneapolis, MN)	0.2658	0.2393	0.2375	0.3664	0.3335	0.3257
Wisconsin State Journal	-0.3865	0.7927	0.3341	0.2330	0.3727	0.3631

This data table is also displayed as figures to show spatial trends in the data. See Figures 6 and 7 below:

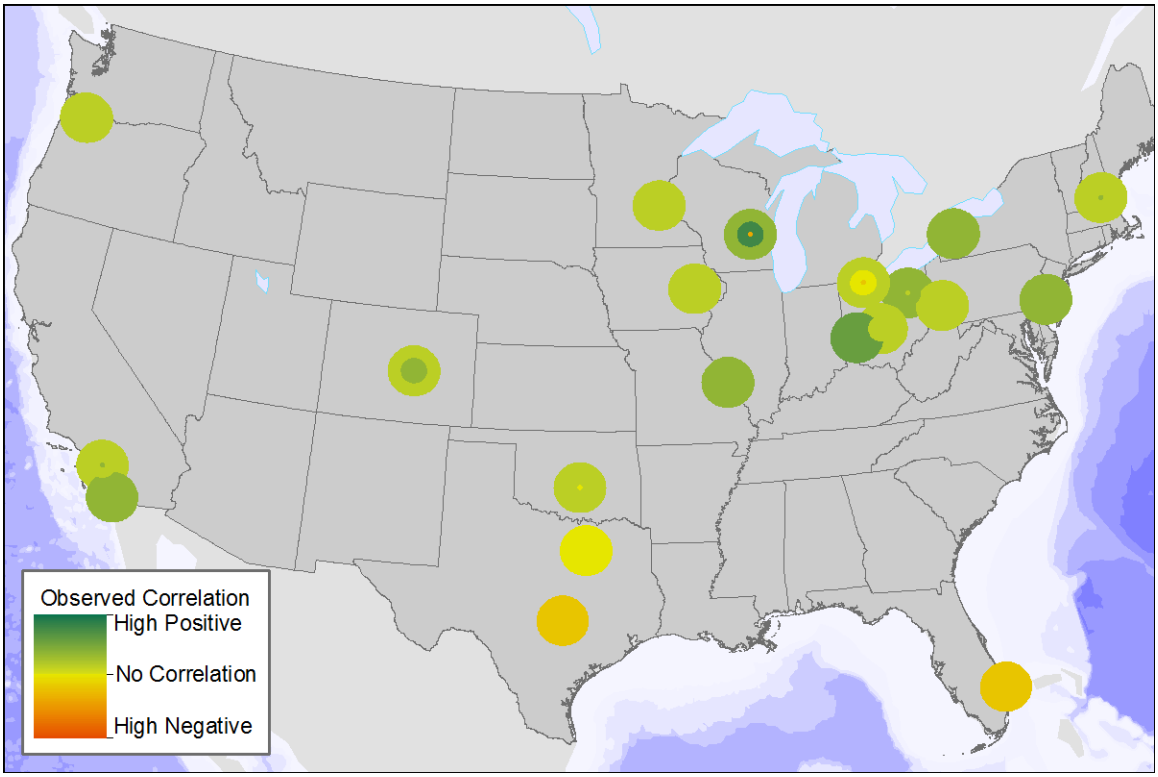

Figure 6: Map of Correlation of News and Social Media, Bernie Sanders, Lag Week

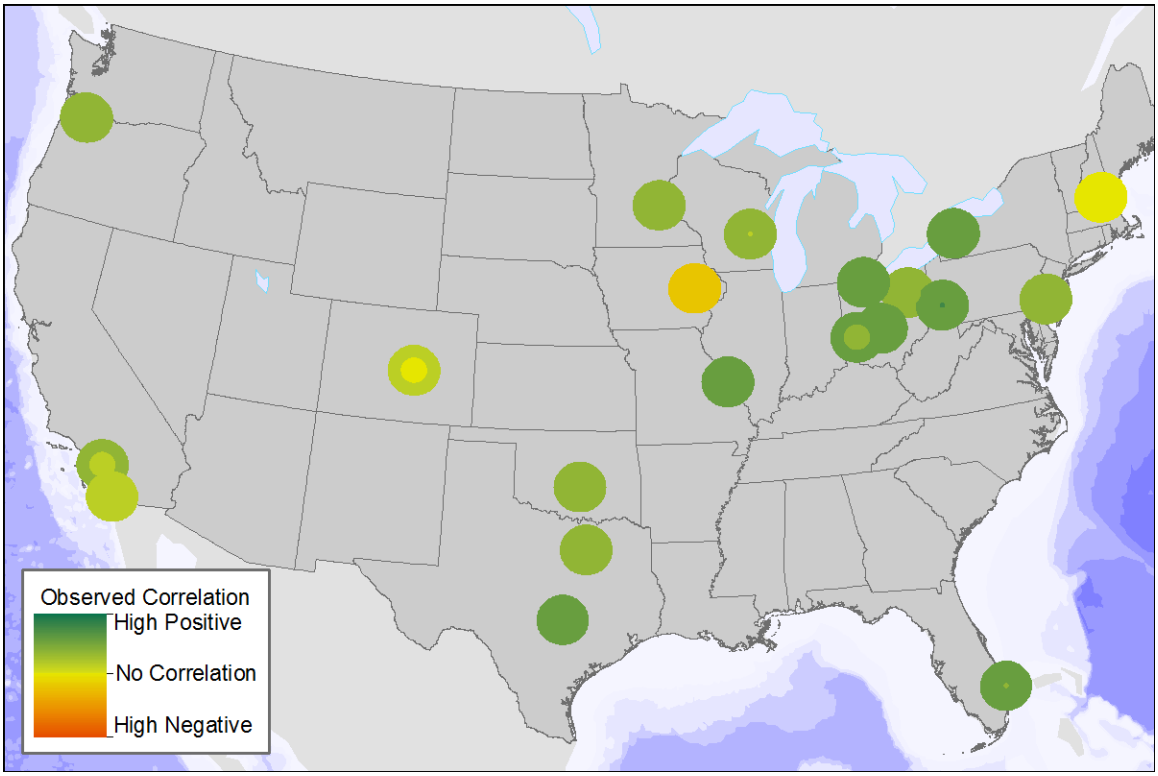

Figure 7: Map of Correlation of News and Social Media, Donald Trump, Lag Week

Finally, Table 8 below shows the difference between these two values, positive numbers indicating that the social media lag displayed increased correlation while negatives showing the social media lag displayed decreased correlation.

Table 8: Pearson Product Moment Correlation Scores, difference between Same Week and Lag Week

Pearson Score Delta	News Results to Tweet Count (10k, Bernie Sanders)	News Results to Tweet Count (50k, Bernie Sanders)	News Results to Tweet Count (100k, Bernie Sanders)	News Results to Tweet Count (10k, Donald Trump)	News Results to Tweet Count (50k, Donald Trump)	News Results to Tweet Count (100k, Donald Trump)
Akron Beacon Journal (OH)	-0.2206	-0.1676	-0.2085	0.1320	0.2250	0.2653
Austin American-Statesman (TX)	-0.1680	-0.1759	-0.1870	-0.0120	0.0122	0.0254
Blade, The (OH)	0.0909	-0.1744	-0.1711	0.0116	-0.0363	0.0542
Buffalo News, The (NY)	0.0286	0.0384	-0.0019	-0.0566	0.0068	0.0676
Columbus Dispatch, The (OH)	-0.0494	-0.0962	-0.1084	0.0903	0.0590	0.0788
Daily Oklahoman, The (OK)	-0.2164	-0.1248	-0.1753	-0.1582	-0.0039	0.0147
Dallas Morning News, The (TX)	-0.1562	-0.2142	-0.2043	0.0683	0.0914	0.0966
Dayton Daily News (OH)	-0.0099	-0.0441	0.0024	-0.1414	-0.0774	0.0450
Gazette, The (Cedar Rapids, IA)	-0.1561	-0.1147	-0.0834	-0.2557	-0.1992	-0.0336
Gazette, The (Colorado Springs, CO)	0.1149	0.0723	-0.0452	-0.1033	-0.1702	-0.1514
New Hampshire Union Leader (Manchester)	-0.1522	-0.0953	-0.0829	-0.3310	-0.1254	-0.1400
Orange County Register, The (Santa Ana, CA)	0.1795	0.1127	0.1277	-0.0954	-0.0755	-0.0265
Oregonian, The (Portland, OR)	0.1043	0.1192	0.1155	-0.0647	-0.0273	-0.0262
Palm Beach Post, The (FL)	-0.3357	-0.2581	-0.2708	0.0863	0.0885	0.0766
Philadelphia Inquirer, The (PA)	-0.1010	-0.0682	-0.0645	0.0107	0.0851	0.1120
Pittsburgh Post-Gazette (PA)	0.0730	0.0674	0.0614	-0.0088	0.0002	-0.0069
Pittsburgh Tribune Review (PA)	-0.1952	-0.1982	-0.1997	0.2005	0.1793	0.1743

San Diego Union-Tribune, The (CA)	-0.0188	-0.0094	-0.0251	-0.0248	-0.0620	-0.0676
St. Louis Post-Dispatch (MO)	-0.0295	-0.0359	-0.0297	0.0616	0.0325	0.0330
Star Tribune (Minneapolis, MN)	-0.0106	-0.0269	-0.0400	0.0701	0.0082	0.0063
Wisconsin State Journal	-0.9842	-0.0671	-0.1869	-0.1813	-0.1392	-0.1113

CHAPTER FIVE: DISCUSSION

As Table 6 shows, there is mostly positive correlation across all the different tests. Most tests also stay similar throughout the different buffer values from 10, 50, and 100km. The most notable result is The Blade in Toledo, Ohio. Within 10km of the city center, there is significant negative correlation between tweets and the news reports on Bernie Sanders. However, after moving further from the city center, the correlation jumps to .24 and .26 for 50km and 100km respectively. Additionally, the Wisconsin State Journal experiencing a positive spike in the 50km test up to .85, whereas 10km and 100km are .59 and .52 respectively.

This data shows an average of .31 correlation between the news media reports and nearby tweets for all records combined. Several outlets, including the Dayton Daily News and the Wisconsin State Journal show an average of greater than .50 correlation for both Donald Trump and Bernie Sanders. This shows that while there is generally positive correlation, it varies based on the news outlet.

Table 7 shows the lag week correlation. The majority of these results show similar patterns to the same week results. Again, there is little variation in the totals when considering the buffer sizes. Notable deviations include once again the Wisconsin State Journal, which shows the same low-high-low pattern in the time lag Pearson as in the same week correlation.

When these charts are compared against each other in Table 8, it becomes clear what the difference between Tables 6 and 7 are. This chart shows a nearly even split between values below zero (indicating more correlation in the same week Pearson score) and above zero (indicating more correlation in the lag week Pearson score). The data leans slightly towards being negative, which indicates that the null hypothesis of this study is accepted. This study cannot prove that there is higher correlation in social media prior to an event as compared to traditional media.

As expected after review of tables 6 and 7, the buffered distance had little effect on the differences shown in table 8. In general, negative trends remained negative and positive trends remained positive notwithstanding of the buffer distance.

What this data shows is statistically inconclusive. In general, the time lag of a week did not have a significant effect on the correlation between the datasets. However, this research does leave several possibilities for future research as well on this election as well as elections in the future.

Deep Dive: Ohio and Pennsylvania

One interesting area for further analysis is the spatially dense area of newspapers in Ohio and western Pennsylvania that reported on both Bernie Sanders and Donald Trump. Western Pennsylvania in the context of this research refers to the areas immediately around Pittsburgh and west. The newspapers that are the focus of this research are: the Akron Beacon Journal, The Blade (Toledo, Ohio), Columbus Dispatch, Dayton Daily News, Pittsburgh Post-Gazette, and Pittsburgh Tribune Review. In order to

review the space time trends in the data, first the general temporal trend must be established. Once a small term of interest is located, the data can be reviewed in a spatiotemporal context, as creating static maps based on space time data is not feasible.

Temporal trends are shown in the figures below. Due to erratic patterns in the daily number of tweets, it was necessary to show a moving average of tweets rather than the count, meaning the values displayed on the figures show a series of averages of the five day period on and around the date. In short, for each date, the value shown in the chart is an average of that day, the two prior days, and the two subsequent days.

As part of this research, I would like to see what the response from traditional and social media is as a result of campaign stops in the state and in a nearby state. To do this, campaign stops in western Pennsylvania and Ohio were found for Bernie Sanders and Donald Trump.

Bernie Sanders had six stops in Ohio and three in western Pennsylvania during the research period, according to a National Journal compiled list of campaign stops. In Ohio, one was in November 2015, and the other five were in early March 2016. In Pennsylvania, one was in late March 2016 and the two were in April 2016 (2016 Travel Tracker, 2016). Figures 8 and 9 show the totals for social media posts and traditional media reports respectively.

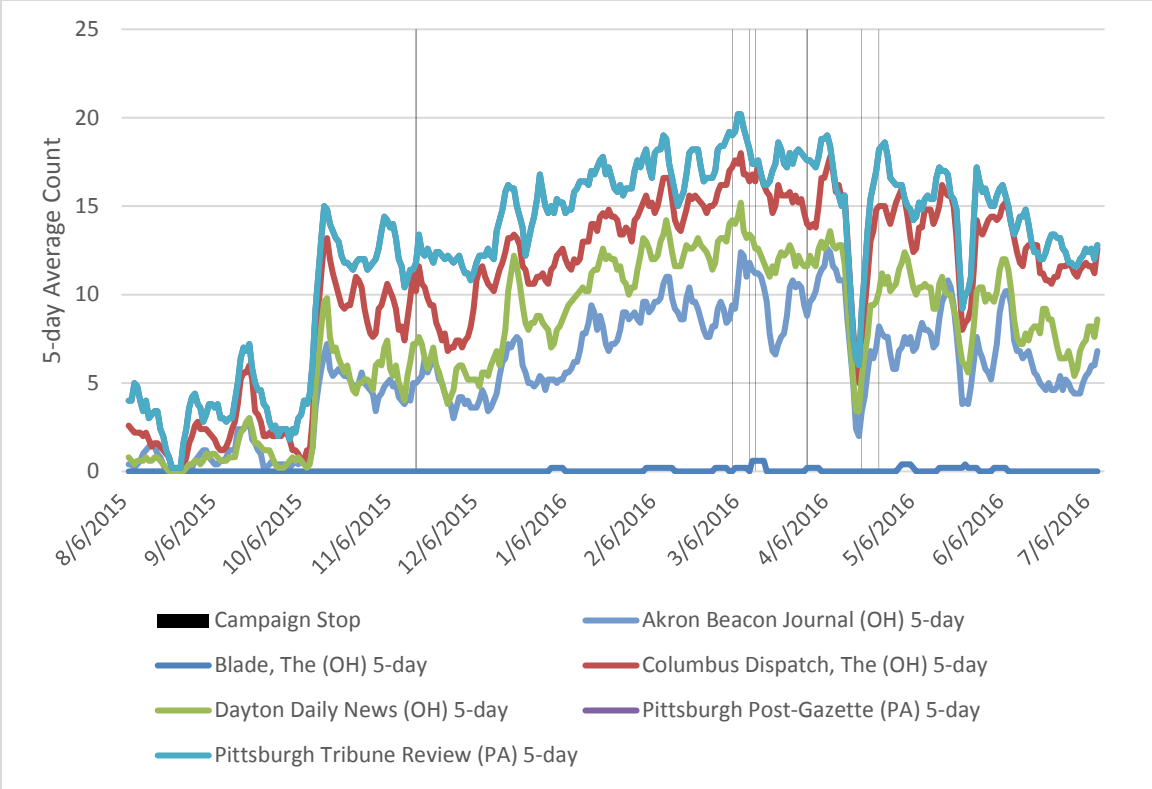

Figure 8: Moving Average of Tweets about Bernie Sanders with Campaign Stops

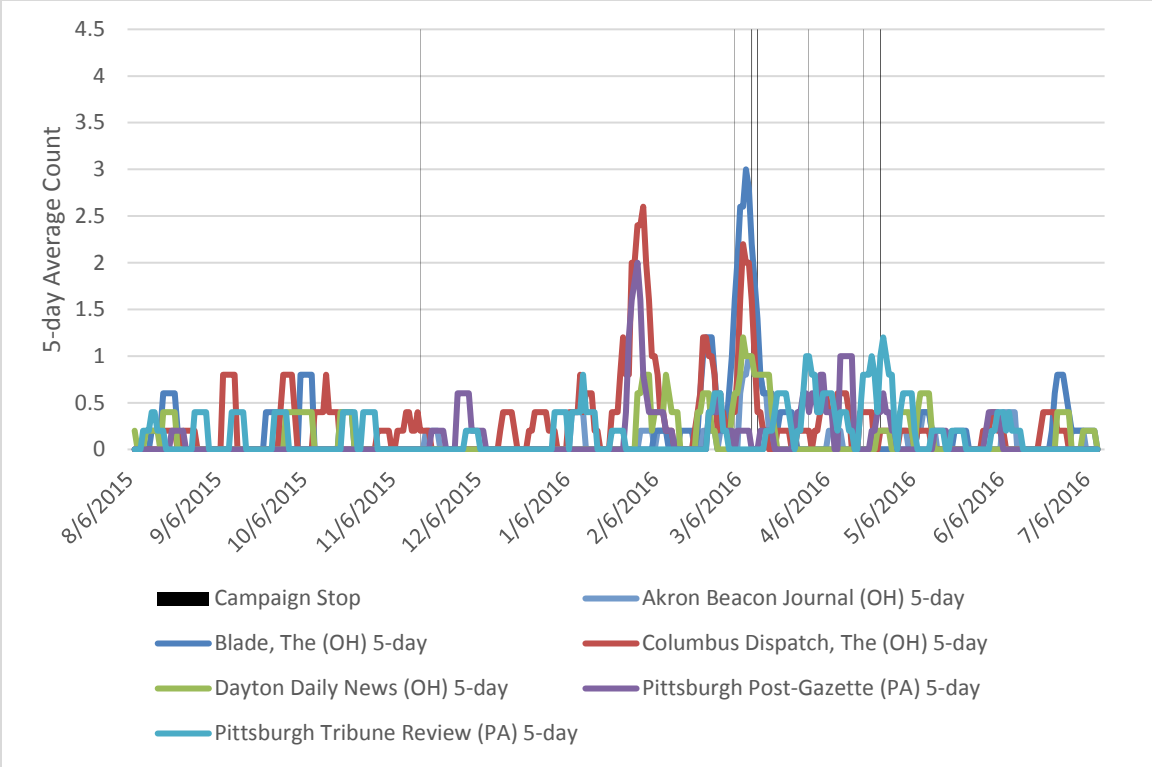

Figure 9: Moving Average of Newspaper Reports about Bernie Sanders with Campaign Stops

Donald Trump had nine stops in Ohio and three in western Pennsylvania during the research period, again according to the National Journal. The stops were primarily during March but with several other events spread out throughout the time frame (2016 Travel Tracker, 2016). Figures 10 and 11 show the totals for social media posts and traditional media reports respectively.

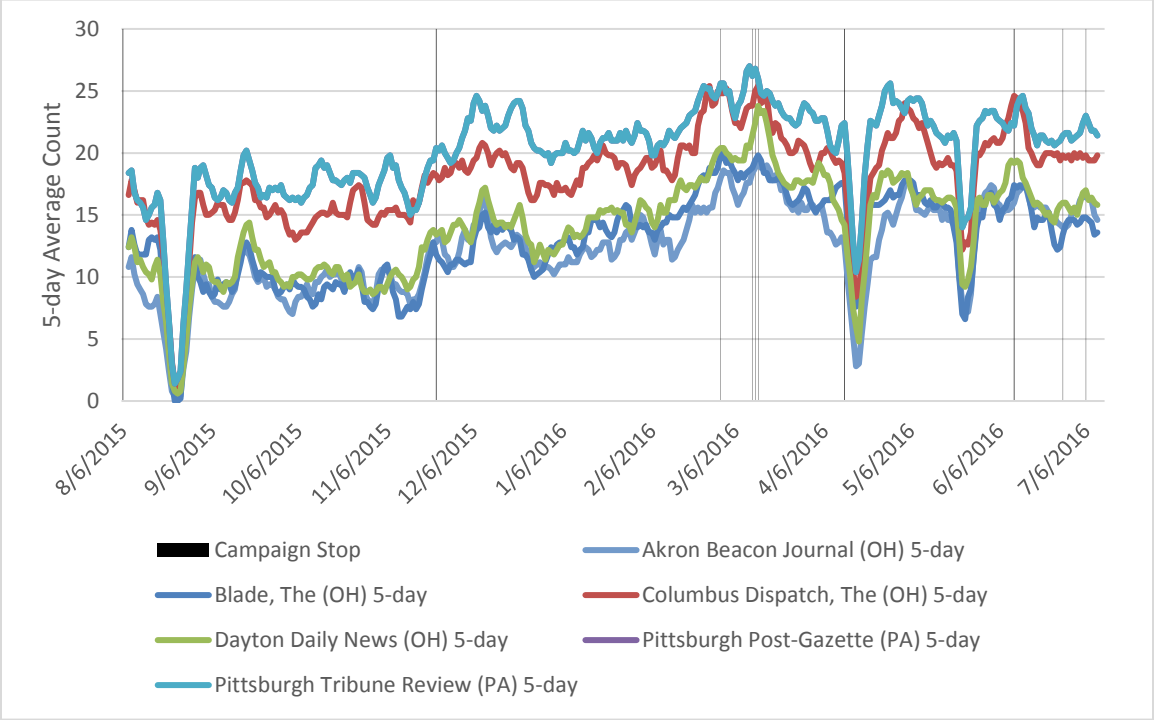

Figure 10: Moving Average of Tweets about Donald Trump with Campaign Stops

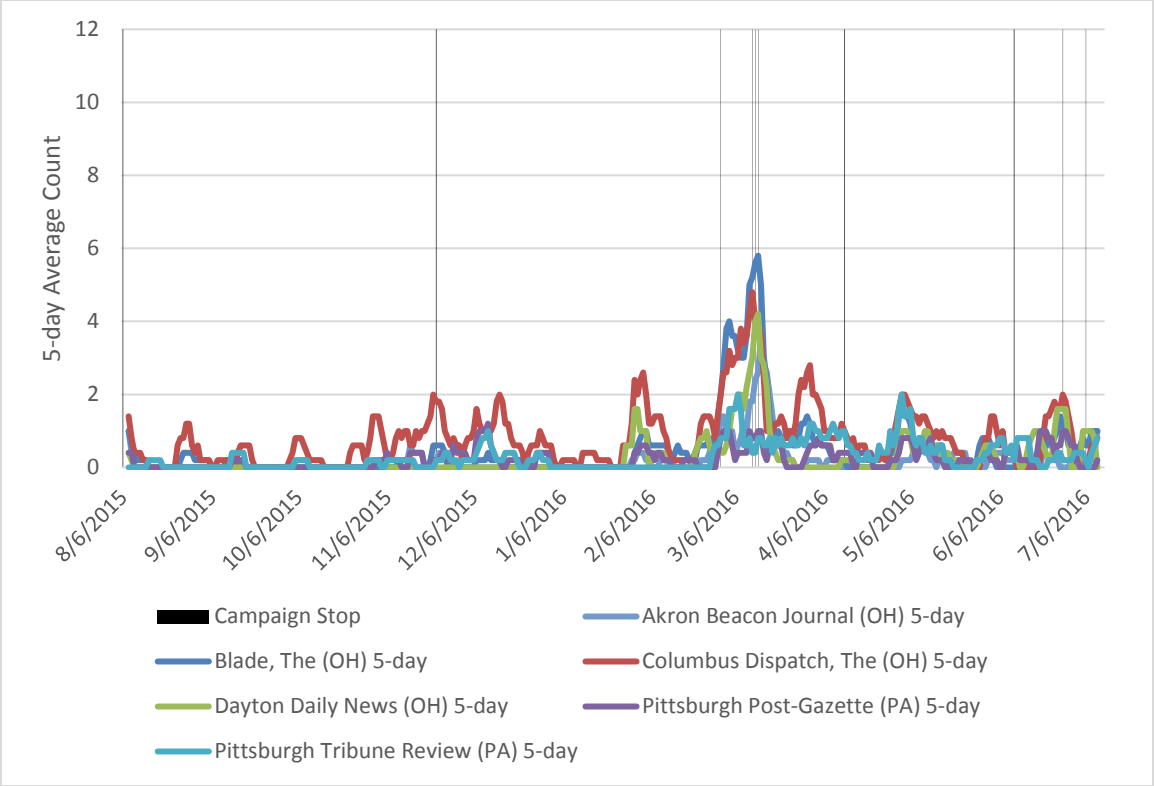

Figure 11: Moving Average of News Reports about Donald Trump with Campaign Stops

In order to put these results into a spatial context and view how the effect of a rally disperses as an effect of distance away from rallies, a specific time period must be chosen. From the figures, mid-March presents a good time to show the effect of rallies, as there were a significant number of rallies as well as increases in tweets relevant to each candidate. The increase centers around the Ohio primaries, where both Democrat and Republican voters went to polling stations to select their party's nominee, on March 15th. A seven day period from March 11th to March 18th was chosen to show the effects immediately leading up to and following both the rallies from both candidates as well as

the primary. Figures 12 and 13 below show the temporal trends in different parts of the state over time as they relate to where rallies occurred for Bernie Sanders and Donald Trump respectively. Vertical lines on the chart represent days where the respective candidate held a rally in Ohio. Points with dates indicate locations with rallies, and green circles are the 10km buffer regions around newspaper outlets.

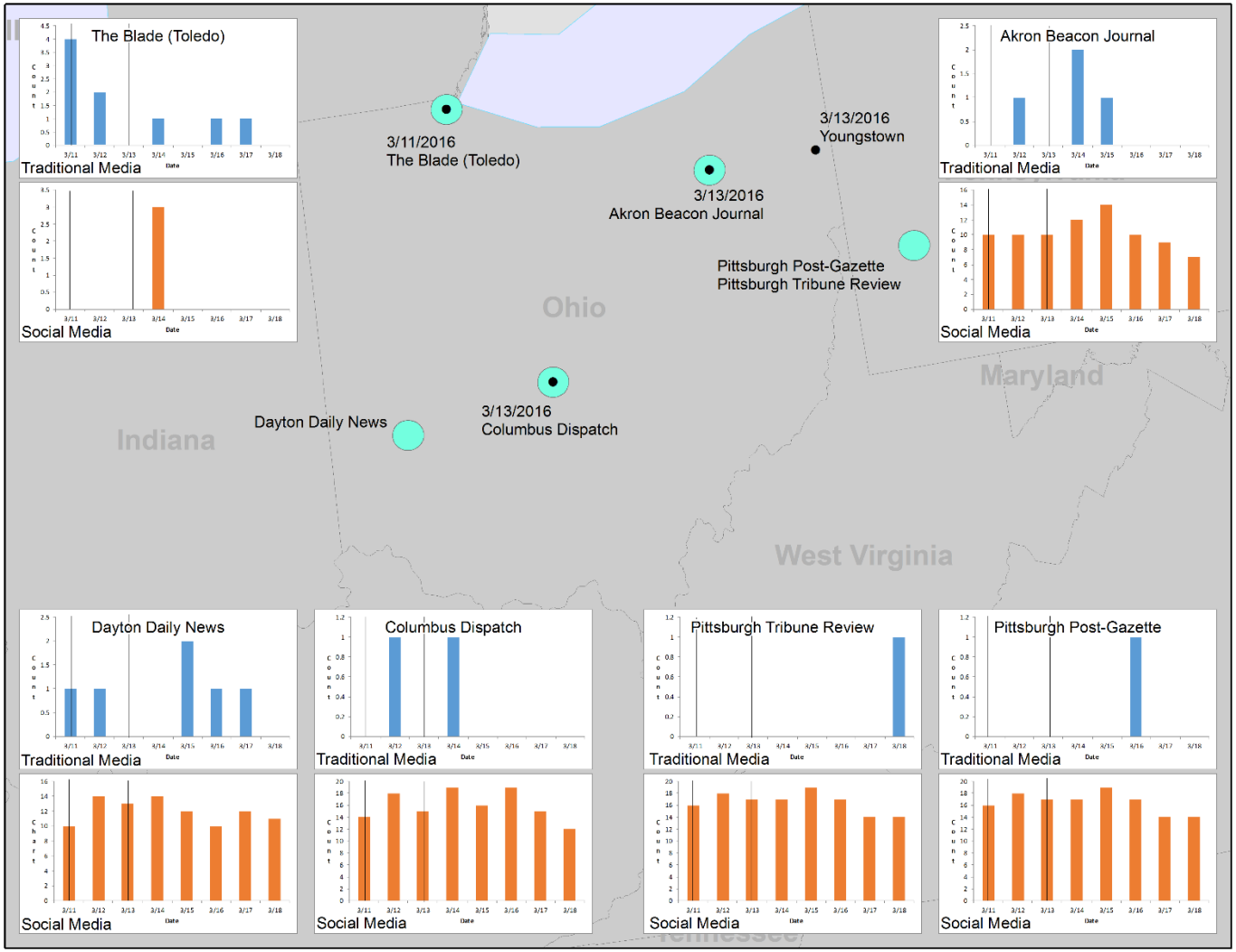

Figure 12: Space-time Trends of Social and Traditional Media related to Bernie Sanders in Ohio and Western Pennsylvania from March 11 through 18, 2016

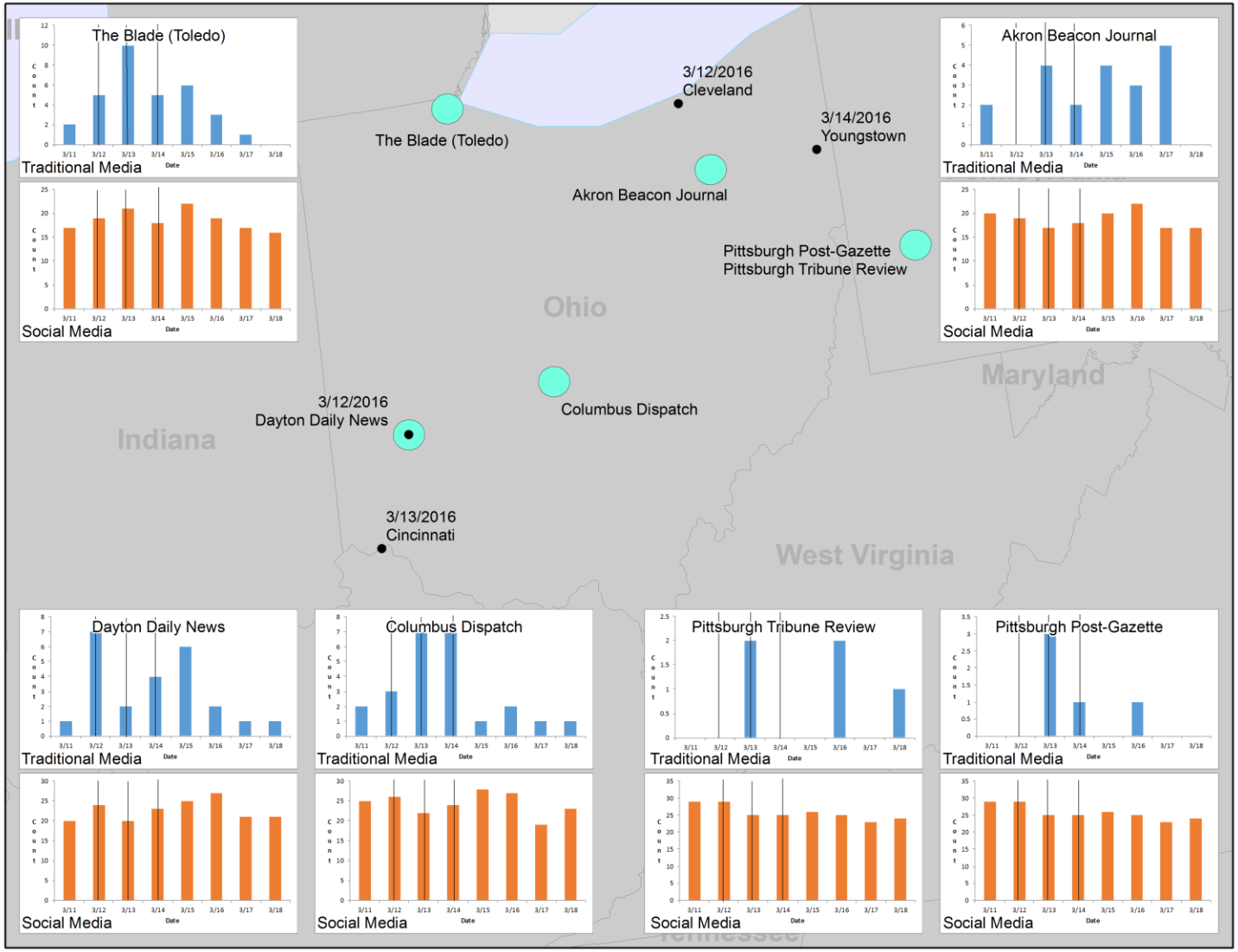

Figure 13: Space-time Trends of Social and Traditional Media related to Donald Trump in Ohio and Western Pennsylvania from March 11 through 18, 2016

From these maps, several trends can be recognized in the data. Firstly, traditional media in most cases does not show significant trends related to the rally events. However, some newspapers show a build up to the number of news articles a few days before the date of the primary. Additionally, it is clear that Donald Trump received a significant amount more coverage compared to Bernie Sanders from the newspaper outlets chosen for this timeframe.

For social media, in the case of Bernie Sanders, the majority of users around the cities showed the same trend: moderate level on the days leading up to and the day of the rally, elevated levels of tweets for a day or two immediately following the rally, and then a return to the moderate level. Notable exceptions to this are The Blade in Toledo, which prior to the rally in Toledo showed no social media, three days later a few tweets came out.

An interesting trend seen from this is that response to events shows a clear drop off at state lines. For instance, despite being near the border, outlets and social media users from Pittsburgh did not show any trends related to the rallies in Ohio even when they were close to the border in Youngstown, Ohio. This clear state divide in how social media users react shows the importance of election related social media analysis to consider political borders.

For social media, in the case of Donald Trump, the Dayton Daily News shows an interesting trend. On the day of a rally in Dayton, the number of tweets is elevated, followed by a drop the next day. And then when Trump travels to the northeast of the state to Youngstown it kicks off three days in a row of increased tweets. The Columbus

Dispatch shows a similar result, though slightly less pronounced. Some of this can be explained by the imminent primary occurring on the 15th, however the same trend is not seen in the other cities in Ohio that are part of this research study.

CHAPTER SIX: LIMITATIONS

There are numerous confounding factors that could have affected the analysis presented in this research paper. Some of these can be potentially mitigated in future research, and some are merely the nature of dealing with data.

Not every voice is at the same volume on social media. More users from the fringes of either political group may be more vocal on social media but less representative of the entire population. Studies have shown that the discussion on social media represents a long tail of the data, where the fringe opinions are more frequently voiced than the moderate opinions (Mustafaraj et al., 2011, p.7). This leads to vocal minorities and silent majorities on social media, so the peaks and valleys in social media data do not necessarily reflect the average opinions in an area. This may have been what affected the Wisconsin State Journal in the results: a more vocal opinion may have dominated at one specific buffer level. This could be mitigated by ensuring each user is only counted at most one per day as opposed to however many times they tweet, however the issue remains that the silent do not tweet at all.

Misspellings or nicknames for the candidate's may confound searches in social media to find relevant tweets. Nicknames like "The Donald" and "Bern" which were commonly used in social media to refer to the candidates are too generic to the candidates, and to include them in the search terms would likely generate a significant

number of false positives in the social media dataset. Further, if instead names are included in photos and video and not explicitly mentioned in the text of the tweet, it will be ignored as well. Having a more complex search based on hash tags related to each candidate or using optical character recognition for finding text in images would be possible remedies to this issue.

On the traditional media side, initial review of the data did not reveal the relatively small amount of data as compared to social media. Small local newspapers, that were the focus of this article, do not contribute a significant number of articles on national elections. This is likely because reporting on it would cost significant money, when wire services like the Associated Press can sell articles to the newspaper at a much better rate with comprehensive reporting. This could be overcome by using a smaller geographic region for this research, as one would expect more research about local elections from the local papers chosen.

This lack of data has the further effect of making daily correlation analysis more difficult. Numerous statistical methods were considered for this thesis, however many were discounted because of the nature of the data. Pearson chi-square would have been an appropriate function; however a significant number of zeroes and low values in the traditional media dataset made this calculation impossible. Instead, I used a more simple linear correlation model with the Pearson product moment correlation, which may not have shown the best fit against this data. This would also likely be remedied by choosing a smaller geographic area.

CHAPTER SEVEN: CONCLUDING REMARKS AND FUTURE RESEARCH

This thesis has yielded limited statistical conclusions in comparing small traditional media outlets with social media traffic in the area nearby. While the correlation for the "same week" and "lag week" study both yielded positive correlation, the lag week correlation was lower, indicating that there was more correlation in the same week study. There are many potential reasons for this from this thesis missing the temporal shift because the data had to be binned by week and because of the limited nature of reporting from traditional media outlets. Traditional media outlets, in general, do not have as much reporting on national elections and rely on wire services for this reporting, leading to significant gaps in the frequency of reporting. However there are worthwhile conclusions drawn from this research especially that can be used in further research studies. The role of state borders had similar effects on social media and traditional media; when a candidate visits a city near the border, the other side of the border does not see significant increases in social media or traditional media traffic.

The effect of borders on spatial trends in elections can be an interesting future topic, or in order to mitigate some of the limitations from this thesis, research can be repeated at a different spatial scale to eliminate relevant political boundaries. The crux of this analysis rested on traditional media and social media showing significant interest in candidates as they run for office. Because this analysis focused on local news reporting of

a national race, there were limited results from media sources. However, at state and county level elections it is more likely that the smaller media sources will play a more significant role in the reporting. Further, the social media discussion regarding candidates on a local scale would most likely be more related to their candidacy rather than in the case of Donald Trump, where tweets could also be about his status as a celebrity or businessman. It would be my recommendation that a future research could use a US Senate race as the topic.

Another avenue of research would be to review this as compared to other similar votes where an initially long shot vote like Donald Trump proved successful. An example of this would be an analysis of traditional and social media in the lead up to the Brexit vote. How the media and social media treated the events should be similar, but it would be interesting to see if it came up with any novel results.

One further interesting analytic venue would be to complete a similar review comparing Hillary Clinton with Donald Trump in the actual presidential election and comparing national media sources with their distribution zones. This would require some method to locate and isolate the distribution zones, which is no simple task considering the Internet has allowed users to view whatever newspaper they choose.

One final further venue of interest is to study the unique way that social media has effected small traditional media outlets in the context of multiple elections including ones in the future. Social media has shown increased penetration over the past several elections, and it is likely that more citizens will engage in social media in the future (Greenwood et al., 2016). Traditional media sources must continue to update their tactics

to remain both relevant and profitable, and it would be interesting to see how spatially small town press outlets have been effected. This research could also further analyzed based on the breakdown of the spatial location of the populations, reviewing population density based on the latest census data, as well as demographically based on past election results in the areas and demographics.

REFERENCES

2016 Travel Tracker. (n.d.). Retrieved January 7, 2017, from http://traveltracker.nationaljournal.com/

Akron Beacon Journal. (2016, October 7). Retrieved from https://mediabiasfactcheck.com/akron-beacon-journal/

Akron Beacon Journal Search Results. (n.d.). Retrieved January 14, 2017, from http://www.ohio.com/searchohiocom/abj-search-7.226

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492–499). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5616710

Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results. Retrieved from http://doras.dcu.ie/16670/

Bovet, A., Morone, F., & Makse, H. A. (2016). Predicting election trends with Twitter: Hillary Clinton versus Donald Trump. *arXiv Preprint arXiv:1610.01587*. Retrieved from https://arxiv.org/abs/1610.01587

Buffalo News. (2016, July 30). Retrieved from https://mediabiasfactcheck.com/buffalo-news/

Colorado Springs Gazette. (2016, July 29). Retrieved from

https://mediabiasfactcheck.com/colorado-springs-gazette/

Columbus Dispatch. (2016, July 29). Retrieved from https://mediabiasfactcheck.com/columbus-dispatch/

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, *27*(12), 2483–2508. https://doi.org/10.1080/13658816.2013.825724

Daily Source Bias Check: Philadelphia Inquirer. (2016, November 21). Retrieved from https://mediabiasfactcheck.com/2016/11/21/daily-source-bias-check-philadelphia-inquirer/

Daily Source Check: Manchester Union Leader. (2016, October 5). Retrieved from https://mediabiasfactcheck.com/2016/10/05/daily-source-check-manchester-union-leader/

Dallas Morning News. (2016, July 14). Retrieved from https://mediabiasfactcheck.com/dallas-morning-news/

Duggan, M. (2015, August 19). The Demographics of Social Media Users. Retrieved from http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

Duggan, M., & Brenner, J. (2013, February 14). Social Networking Site Users. Retrieved from http://www.pewinternet.org/2013/02/14/social-networking-site-users/

Gilbert, E., Karahalios, K., & Sandvig, C. (2010). The Network in the Garden: Designing Social Media for Rural Life. *American Behavioral Scientist*, *53*(9), 1367–1388.

https://doi.org/10.1177/0002764210361690

Gottfried, J., & Shearer, E. (2016, May 26). News Use Across Social Media Platforms 2016. Retrieved from http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

Greenwood, S., Perrin, A., & Duggan, M. (2016, November 11). Social Media Update 2016. Retrieved from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Guynn, J. (2016, November 8). Forget Trump: Election's big winner was Twitter. Retrieved January 7, 2017, from http://www.usatoday.com/story/tech/news/2016/11/08/election-winner-twitter/93509896/

Hosch-Dayican, B., Amrit, C., Aarts, K., & Dassen, A. (2016). How Do Online Citizens Persuade Fellow Voters? Using Twitter During the 2012 Dutch Parliamentary Election Campaign. *Social Science Computer Review*, *34*(2), 135–152. https://doi.org/10.1177/0894439314558200

Journalism, P. R. C., & staff, M. (2012, November 1). Tone of Mainstream Media Coverage. Retrieved from http://www.journalism.org/2012/11/01/tone-mainstream-media-coverage/

Jungherr, A., Jurgens, P., & Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment." *Social Science*

Computer Review, *30*(2), 229–234. https://doi.org/10.1177/0894439311404119

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Wisconsin State Journal Search Results. (n.d.) Retrieved January 14, 2017, from http://host.madison.com/search/wsj/

Memoli, M. A. (2010). Sen. Bernie Sanders ends filibuster. *Los Angeles Times*. Retrieved from http://articles.latimes.com/print/2010/dec/10/news/la-pn-sanders-filibuster-20101211

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 441–450). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2145274

Mustafaraj, E., Finn, S., Whitlock, C., & Metaxas, P. T. (2011). Vocal minority versus silent majority: Discovering the opionions of the long tail. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 103–110). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6113101

Newspaper Source Plus. (n.d.). Retrieved January 14, 2017, from https://www.ebscohost.com/public/newspaper-source-plus

Orange County Register. (2016, July 16). Retrieved from

https://mediabiasfactcheck.com/orange-county-register/

Oregonian. (2016, July 29). Retrieved from https://mediabiasfactcheck.com/oregonian/

O'Reilly, T. (2005, September 30). What Is Web 2.0. Retrieved January 7, 2017, from http://oreilly.com{file}

Phoenix Data Project. (n.d.). Retrieved November 27, 2016, from http://phoenixdata.org/

Pittsburgh Post-Gazette. (2016, July 30). Retrieved from https://mediabiasfactcheck.com/pittsburgh-post-gazette/

Pittsburgh Tribune Review. (2016, July 16). Retrieved from https://mediabiasfactcheck.com/pittsburgh-tribune-review/

San Diego Union Tribune. (2016, July 18). Retrieved from https://mediabiasfactcheck.com/san-diego-union-tribune/

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp. 53–60). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2389976

Scharl, A., & Weichselbraun, A. (2008). An Automated Approach to Investigating the Online Media Coverage of U.S. Presidential Elections. *Journal of Information Technology & Politics*, *5*(1), 121–132. https://doi.org/10.1080/19331680802149582

St. Louis Post-Dispatch. (2016, July 29). Retrieved from https://mediabiasfactcheck.com/st-louis-post-dispatch/

The GDELT Story. (n.d.). Retrieved November 27, 2016, from

http://gdeltproject.org/about.html

Trackalytics - Bernie Sanders. (2016). Retrieved November 6, 2016, from http://www.trackalytics.com/twitter/profile/sensanders/

Trackalytics - Donald Trump. (2016). Retrieved November 6, 2016, from http://www.trackalytics.com/twitter/profile/realdonaldtrump/

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of theFourth International AAAI Conference on Weblogs and Social Media*, 178–185.

van Beuzekom, B. (2008). Measuring User-Created Content: Implications for the ICT Access and Use by Households and Individuals Surveys. OECD Digital Economy Papers, No. 139. *OECD Publishing (NJ1)*. Retrieved from http://eric.ed.gov/?id=ED504074

Wells, C., Shah, D. V., Pevehouse, J. C., Yang, J., Pelled, A., Boehm, F., … Schmidt, J. L. (2016). How Trump Drove Coverage to the Nomination: Hybrid Media Campaigning. *Political Communication*, *33*(4), 669–676. https://doi.org/10.1080/10584609.2016.1224416

Wikipedia:About. (2016, November 26). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=751559726

Woodly, D. (2007). New competencies in democratic communication? Blogs, agenda setting and political participation. *Public Choice*, *134*(1–2), 109–123. https://doi.org/10.1007/s11127-007-9204-7

BIOGRAPHY

Scott Heneghan graduated from Atlantic City High School, Atlantic City, New Jersey, in 2004. He received his Bachelor of Science in Geography with a Specialization in Computer Cartography and GIS in 2008 from the University of Maryland at College Park. He is presently employed as a Senior Geospatial Analyst for The HumanGeo Group for the past year. He has a fish named Ham.