USING TWITTER DATA AS A COMMUNITY POLICING MECHANISM OF CRIMINAL ACTIVITY IN WASHINGTON DC

by

Kevin Marc Glodava
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. Arie Croitoru, Thesis Director
Dr. Andrew Crooks, Committee Member
Dr. Anthony Stefanidis, Committee Member
Dr. Dieter Pfoser, Committee Member
Dr. Anthony Stefanidis, Department Chair
Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
Dr. Peggy Agouris, Dean, College of Science

Date: 2 Dec 2014

Fall Semester 2014
George Mason University
Fairfax, VA
Using Twitter Data as a Community Policing Mechanism of Criminal Activity in Washington DC

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

by

Kevin Marc Glodava
Bachelor of Arts and Science
Metropolitan State University of Denver, 2001 and 2008

Director: Arie Croitoru, Professor
Department of Geoinformatics and Geospatial Intelligence

Fall Semester 2014
George Mason University
Fairfax, VA
This work is licensed under a creative commons attribution-noderivs 3.0 unported license.
DEDICATION

This is dedicated to my amazing wife Trish, my two crazy and energetic children Brighton and Hutton, and my two pups Clyde and the late Jonnie who all lost out on “play time” as I committed 4 ½ years to complete this worthwhile program.
I would like to thank the many friends, relatives, and supporters who have made this happen. My incredible Wife, Trish, whom assisted me in my edits and clarification of ideas and more important made me laugh when I got frustrated by the homework or putting together this beast of a thesis. Dr. Arie Croitoru and Dr. Anthony Stefanidis, and the other members of my committee were of invaluable help. Thanks go out to the Fenwick Library for providing a clean, quiet, and well-equipped repository in which to work. Finally thanks go out to the Denver Broncos for winning multiple games and became a welcome distraction as I researched and wrote this while watching and cheering them to victory. Just forget about the Super Bowl, it didn’t happen.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>7</td>
</tr>
<tr>
<td>List of Figures</td>
<td>8</td>
</tr>
<tr>
<td>Abstract</td>
<td>11</td>
</tr>
<tr>
<td>Chapter One: introduction</td>
<td>13</td>
</tr>
<tr>
<td>1.1 What Can Be Discovered Through Social Media Analysis?</td>
<td>13</td>
</tr>
<tr>
<td>1.1.1 Establishment of Community Policing</td>
<td>15</td>
</tr>
<tr>
<td>1.1.2 Gravitation to Intelligence-Led Policing</td>
<td>18</td>
</tr>
<tr>
<td>1.2 Use of Geographic Information Systems (GIS) for Crime Analysis</td>
<td>19</td>
</tr>
<tr>
<td>1.2.1 Increase in Social Media Participation</td>
<td>21</td>
</tr>
<tr>
<td>1.2.2 Volunteered Geographic Information (VGI)</td>
<td>28</td>
</tr>
<tr>
<td>1.2.3 Ambient Geographic Information (AGI)</td>
<td>29</td>
</tr>
<tr>
<td>1.3 Why Law Enforcement Should Use Social Media</td>
<td>29</td>
</tr>
<tr>
<td>1.4 Thesis Organization</td>
<td>31</td>
</tr>
<tr>
<td>Chapter Two: Background</td>
<td>32</td>
</tr>
<tr>
<td>2.1 Crime Mapping</td>
<td>32</td>
</tr>
<tr>
<td>2.1.1 Social Media and Law Enforcement</td>
<td>35</td>
</tr>
<tr>
<td>2.1.2 Crowdsourcing</td>
<td>36</td>
</tr>
<tr>
<td>2.1.3 Analysis of Social Media Applications</td>
<td>37</td>
</tr>
<tr>
<td>2.2 Identified Intelligence Gaps from Prior Research</td>
<td>40</td>
</tr>
<tr>
<td>Chapter Three: Methodology</td>
<td>43</td>
</tr>
<tr>
<td>3.1 Methodology to Process to Capture, store and Filter Twitter Data</td>
<td>45</td>
</tr>
<tr>
<td>3.1.1 Keywords for Harvesting Data</td>
<td>45</td>
</tr>
<tr>
<td>3.1.2 Open-Source Comparative Analysis</td>
<td>47</td>
</tr>
<tr>
<td>3.1.3 Twitter Data Collection, Storage and Filtering</td>
<td>48</td>
</tr>
<tr>
<td>Chapter Four: Crime Analysis</td>
<td>52</td>
</tr>
<tr>
<td>4.1 Classic Approach to Crime Analysis</td>
<td>52</td>
</tr>
<tr>
<td>4.1.1 Washington DC Crime Data</td>
<td>52</td>
</tr>
<tr>
<td>4.1.2 Crime Mapping/Hot Spot Analysis</td>
<td>54</td>
</tr>
<tr>
<td>4.2 Washington D.C Crime Data Crime Statistics Analysis</td>
<td>68</td>
</tr>
</tbody>
</table>
5.5.1 Normalization of Data ................................................................. 126
Chapter Six: Conclusion and Outlook .................................................... 131
  6.1 Future Work ............................................................................. 133
References ....................................................................................... 135
LIST OF TABLES

Table 2: Counties of 50,000 population or more based on the percent population increase due to commuting: 2006 - 2010 (United States Census Bureau, 2010) ........................................ 27
Table 3: Keywords and Generated Using www.Tagcrowd.com and Phrases Generated from Manual Inspection .......................................................... 47
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1: The First Modern Day Police Uniform (Johnson, 2014)</td>
<td>16</td>
</tr>
<tr>
<td>Figure 2: The Estimated Number of Smartphone Users in the United States from 2010 to 2012 (in Millions) (Malleson &amp; Anderson, 2014)</td>
<td>22</td>
</tr>
<tr>
<td>Figure 3: Twitter Usage Overtime (Smith &amp; Brenner, 2014)</td>
<td>23</td>
</tr>
<tr>
<td>Figure 4: Who Uses Twitter (Smith &amp; Brenner, 2014)</td>
<td>24</td>
</tr>
<tr>
<td>Figure 5: Total Visitation to Washington DC (Destination DC, 2014)</td>
<td>26</td>
</tr>
<tr>
<td>Figure 6: Research Road Map</td>
<td>43</td>
</tr>
<tr>
<td>Figure 7: Word Cloud Generated Using <a href="http://www.Tagcrowd.com">www.Tagcrowd.com</a> on 16 April 2014</td>
<td>46</td>
</tr>
<tr>
<td>Figure 8: Geotagged Tweets Filtered Using ESRI Intersect Tool and applied to a Kernel Density</td>
<td>50</td>
</tr>
<tr>
<td>Figure 9: Tweet per day in Washington DC – June, July, and August 2014</td>
<td>51</td>
</tr>
<tr>
<td>Figure 10: Kernel Density of Overall Crime - Washington DC June, July and August 2013</td>
<td>57</td>
</tr>
<tr>
<td>Figure 11: Kernel Density of Overall Crime - Washington DC June, July and August 2014</td>
<td>57</td>
</tr>
<tr>
<td>Figure 12: Overall Crime 2014 minus Overall Crime 2013</td>
<td>58</td>
</tr>
<tr>
<td>Figure 13: Kernel Density of Arson - Washington DC June, July and August 2013</td>
<td>59</td>
</tr>
<tr>
<td>Figure 14: Kernel Density of Arson - Washington DC June, July and August 2014</td>
<td>59</td>
</tr>
<tr>
<td>Figure 15: Kernel Density of Assault with a Deadly Weapon - Washington DC June, July and August 2013</td>
<td>60</td>
</tr>
<tr>
<td>Figure 16: Kernel Density of Assault with a Deadly Weapon - Washington DC June, July and August 2014</td>
<td>60</td>
</tr>
<tr>
<td>Figure 17: Kernel Density of Burglary - Washington DC June, July and August 2013</td>
<td>61</td>
</tr>
<tr>
<td>Figure 18: Kernel Density of Burglary - Washington DC June, July and August 2014</td>
<td>61</td>
</tr>
<tr>
<td>Figure 19: Kernel Density of Homicide - Washington DC June, July and August 2013</td>
<td>62</td>
</tr>
<tr>
<td>Figure 20: Kernel Density of Homicide - Washington DC June, July and August 2014</td>
<td>62</td>
</tr>
<tr>
<td>Figure 21: Kernel Density of Motor Theft - Washington DC June, July and August 2013</td>
<td>63</td>
</tr>
<tr>
<td>Figure 22: Kernel Density of Motor Theft - Washington DC June, July and August 2014</td>
<td>63</td>
</tr>
<tr>
<td>Figure 23: Kernel Density of Robbery - Washington DC June, July and August 2013</td>
<td>64</td>
</tr>
<tr>
<td>Figure 24: Kernel Density of Robbery - Washington DC June, July and August 2014</td>
<td>64</td>
</tr>
<tr>
<td>Figure 25: Kernel Density of Sexual Assault - Washington DC June, July and August 2013</td>
<td>65</td>
</tr>
<tr>
<td>Figure 26: Kernel Density of Sexual Assault - Washington DC June, July and August 2014</td>
<td>65</td>
</tr>
<tr>
<td>Figure 27: Kernel Density of Theft from Automobile - Washington DC June, July and August 2013</td>
<td>66</td>
</tr>
</tbody>
</table>
ABSTRACT

USING TWITTER DATA AS A COMMUNITY POLICING MECHANISM OF CRIMINAL ACTIVITY IN WASHINGTON DC

Kevin Marc Glodava, M.S.
George Mason University, 2014

Thesis Director: Dr. Arie Croitoru

In recent years social media has emerged as a rich source of information that spans across a wide spectrum of human activities and events. Tapping into this spectrum, law enforcement agencies have recently begun utilizing social media primarily as a communication tool for aiding in criminal investigations as well as a platform for disseminating information and managing public opinion. By leveraging the ability of social media to quickly reach a broad audience, law enforcement agencies can now inform citizens about criminal activities in their area, increase awareness, and promote campaigns. In conjunction with these trends, community policing has gained momentum among the police and community leaders as they search for effective ways to reduce crime and enhance the safety of their communities. In community policing, citizens are not seen only as passive receivers of law enforcement information, but also as active stakeholders in developing solutions to public safety problems and promoting trust in law enforcement agencies. This thesis aims to explore how social media can be used within the community policing paradigm. In particular, we seek to investigate whether geolocated social media feeds can be used to identify and study the impact areas of
specific types of crimes in urban areas, as well as to gauge public awareness to crime. Using the Washington D.C. metropolitan area as a case study, we collected both Twitter data and police records on major crime types in 2013 and 2014. Based on this data, density maps of reported crimes were compared to density maps of geolocated tweets containing crime-related keywords. This comparison revealed a substantial overlap between the hotspots of the two data sources. In addition, we explored the temporal variation of crime activities and related Twitter communication volume. This comparison provided additional information regarding the alignment between crime (and police) and Twitter activity, thus allowing us to gauge public response and awareness to criminal activities. These results support the premise that social media, and in particular Twitter, can be used as a citizen-driven tool for empowering community policing activities, and supporting intelligence-led policing.
CHAPTER ONE: INTRODUCTION

Social media is a relatively recent phenomenon that has engrained itself into the daily lives of people, politicians, celebrities, the media and businesses. Once viewed as a fad, the role that social media continues to play as it evolves has been extensively studied. Analysis of volunteered geographic information (VGI) and ambient geographic information (AGI) has allowed researchers to study geotagged data of information, thoughts, and observations that people post to various social media sites which has created large amounts of data that can be collected and studied (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013). By examining and leveraging the capabilities of social media platforms, we hope to study how law enforcement agencies can now inform citizens about criminal activities in their area, increase awareness, and promote campaigns.

1.1 What Can Be Discovered Through Social Media Analysis?

Social media encompasses a vast range of topics, from daily minutia and daily news, to events of large scale and impact. As social media has become a significant component of the modern lifestyle, big retailers, such as Walmart, Amazon, and Netflix, use data analytics, the science of examining raw data with the purpose of drawing conclusions from that information to predict consumer behavior in order to make better business decisions that affect the company. Health care insurers and providers also rely on data analytics to develop treatment recommendations for patients and predict the
likelihood that individuals will develop certain illnesses (Rouse, 2014). At the same time, social media has also gained increasing interest in law enforcement. For example, law enforcement agencies have begun to collect big data and have begun to develop software that generates regions where a crime is likely to occur, leading to possible predictive analysis (Bachner, 2012).

Today, social media is used by law enforcement and news agencies to increase awareness about news and information, post notices and provide an opportunity to interact with the community. Twitter is now a key reporting information source for those seeking up-to-date information from headline news and provides the user and opportunity to respond to fresh news. According to Twitter, 40% of their users worldwide use Twitter news feeds vice the traditional news sites and sources (Schneider, 2014). Law enforcement agencies are currently using social media for two main purposes; aiding in criminal investigations and using these tools as platforms for disseminating information as well as managing public opinion (Ruddell & Jones, 2013).

However, social media use by law enforcement agencies is not only for investigation and information dissemination but can also be used for increasing public awareness and harnessing the power of community policing for crime prevention and awareness. The Washington DC Police Department currently has a social media account on Twitter, using the handle @DCPoliceDept. They use the platform to communicate and inform the community on a variety of topics ranging from; holiday tips, traffic closures, traffic incidents, public safety announcements, locations of criminal incidents and requesting assistance from the community to identify perpetrators of crimes or
information leading to an arrest. As of November 2014, @DCPoliceDept has posted over 32 thousand tweets and has approximately 50 thousand followers (@DCPoliceDept, 2014).

1.1.1 Establishment of Community Policing

Community led policing can trace its roots back to the early 1829 when Sir Robert Peel established the London Metropolitan Police District in response to the soaring crimes rates in and around London (Johnson, 2014). It was also the birth of the world’s first modern day police force and has served as the traditional police model for British and American law enforcement ever since. Community policing can be defined as bringing police and citizens together to prevent crime, solve problems, with the emphasis on crime prevention rather than responding to crime after it happens. Sir Robert Peel describes community policing best, “… the police are the public and the public are the police” (Bureau of Justice Assistance, 1994).

However the establishment of the modern day police force is not Sir Robert Peel most influential idea. He is most regarded in establishing regular patrol areas, known as police “beats”, which were established in order to periodically monitor areas, and through that create deterrence that would prevent crime. This notion of proactive policing was innovative since prior to 1829, the police, military or civilians only responded to crimes after they were reported (Patterson, 2014) (Bureau of Justice Assistance, 1994).

The British Police became known as the “bobbies” or “peelers,” named after Sir Robert (also called Bob) Peel. They did not wear traditional military uniform, but were dressed in blue tail-coats and a top hat (as shown in figure 1). The bobbies/peelers carried
a wooden truncheon in a long pocket in the tail of their coat, a pair of handcuffs and a wooden rattle which was later replaced by a whistle (Johnson, 2014).

Figure 1: The First Modern Day Police Uniform (Johnson, 2014)

The Bobbies were assigned a specific geographic region in order to learn more about the specific region, but more important for the community to become more familiar with the police and feel more comfortable reporting crimes to the police. The idea was to bring the police and public closer together with the intention that community policing would help prevent crime. In doing so it opened up a two-way communication with the community and created neighborhood subject matter expects within the police force. Community policing has largely been considered a success in crime prevention, but it is not always been embraced by the community and/or police officers and the statistics did not always show community policing actually reduced crime rates and in some instances crime increased (Wilson & Kelling, 1982).
One criticism typically revolves around a typical poor black neighborhood where the beat cops are white males. Other criticisms include limited implementation in cities whom often implement community policing through small, specialized units in well-defined neighborhoods. This approach often leads to the alienation of some officers as well as claims that the police are ignoring other residents (Patterson, 2014).

In 1982 community policing evolved into a widely controversial theory introduced in 1982 by social scientists, James Q. Wilson and George L. Kelling known as Broken Windows Theory. The theory grew out of an evaluation of foot-patrols in Washington DC which studied foot patrols and determined that it did not have an effect on crime. The basic premise of Broken Windows Theory is as follows;

“Consider a building with a few broken windows. If the windows are not repaired, the tendency is for vandals to break a few more windows. Eventually, they may even break into the building, and if it's unoccupied, perhaps become squatters or light fires inside. Or consider a pavement. Some litter accumulates. Soon, more litter accumulates. Eventually, people even start leaving bags of refuse from take-out restaurants there or even break into cars” (Wilson & Kelling, 1982) (Levitt & Dubner, 2005) (Shelden, 2009).

Simply put if the roads, subways and buildings are clean and maintained crime will be reduced. However if rude remarks by loitering youth were left unchallenged, such youth will be under the impression that no one cares about their behavior and will likely result in more serious crimes in the future. In other words, take care of minor problems before things get worse. Controversial but Broken Windows Theory was embraced in the early 1990’s by New York’s City’s Mayor Rudolph Giuliani and New York City’s new commissioner William Bratton.
With Broken Windows Theory fully embraced by New York City, crime rates dropped significantly. Between 1990 and 1998, murder rates decreased by 72 percent and violent crime was reduced by 51% (Wilson & Kelling, 1982). The claim has been questionable if Broken Windows Theory actually reduced crime, most notably by economists Steven D. Levitt and Stephen J. Dubner whom published, the bestselling book, *Freakonomics*. They stated the reduction in crime in the late 1990’s was not a result of Broken Windows Theory, but the time frame coincided with 1973 Supreme Court decision legalizing abortion. Their theory compared the reduction in crime rates with legalized abortion, suggesting that the children who were not born would have been teens in the 1990’s and the circumstances surrounding the decision to abort were likely to have been born into broken homes and become criminals later in life (Levitt & Dubner, 2005).

1.1.2 Gravitation to Intelligence-Led Policing

In 2002, a movie starring Tom Cruise and Colin Farrell hit the movie theatres called “Minority Report” (Internet Movie Database (IMDb), 2014). This futuristic movie takes place in the Northern Virginia/Washington DC area in the year 2054. A key theme in the movie is that the police department has a predictive crime unit called, Pre-Crime. The unit has the ability to predict a crime before it is committed and the Police arrest the future perpetrator prior to the commission of the crime. The movie demonstrates that having the ability to predict crimes before they are committed all but eliminates crime in the movie.
Over the last few decades policing efforts at law enforcement agencies have turned to technology as it has developed speeds were computers can process big data and new approaches have been incorporated into law enforcement. Policing is shifting from a traditional reactive incident-motivated crime to intelligence-led policing. Drastically different from police beats, the approach relies heavily on expertise, efficiency, statistical analysis, and scientifically proven tactics that guided decision makers through evidence; in particular, large volumes of quantitative data (Bachner, 2012).

1.2 Use of Geographic Information Systems (GIS) for Crime Analysis

One particular area in intelligence-led policing that has flourished is the area of crime mapping also known as “hot spot” analysis. Crime mapping first appeared in 1829 when an Italian geographer and French statistician designed the first maps that visualized crime data. These maps depicted three years of property crime data and education information obtained from France’s census. The layers of information revealed that areas with higher levels of education experienced a higher incidence of property crimes (Bachner, 2012).

In the 1920’s social scientists turned to crime mapping in order to understand the distribution of crime across the city of Chicago called the Chicago School. They began by plotting the known residences of offenders and noticed that most the offenders tended to congregate in relationship to each other and gravitated to specific areas of the city. For example, in Minneapolis, Minnesota less than 3% of the addresses accounted for approximately 50% of the predatory crimes (Wellford & Lum, 2014). By identifying “hot
spots” where crime is occurring, law enforcement can focus their efforts into key areas where most of the crimes occur.

Crime mapping is currently used by analysts in law enforcement agencies to map, visualize, and analyze crime incident patterns using Geographic Information Systems (GIS). This allows crime analysts to identify crime hot spots, along with other trends and patterns. Crime mapping has the ability to possibly identify early warning signs and form a proactive approach to policing and crime reduction.

Various techniques such as hot spot and density analysis are effectively implemented by law enforcement to identify areas where the highest rates of crimes are committed which directly affect police resources and patrolling efforts (Groff & LaVinge, Volume 13). Crime mapping has become a key tool for crime analysts to use in order to understand the patterns of crime allowing them to visualize the differences and similarities across time and space (Brantingham & Brantingham, 1997).

Recently police have been turning to social media sites to help communicate with the community by posting incident reports and public service announcements. They have also turned to social media sites as a form of policing to help identify criminals through crowd-sourced information by posting images of criminals they are trying to identify as well as posting as “friends” of users in order to engage potential criminals in a virtual environment (Crump, 2011).

In the past few years social media has been a common virtual meeting place for people to interact, display photos, videos, but more importantly express one’s thoughts and ideas to the world. There are several social networking sites available. According to
Alexa Global Traffic Rank, social media sites such as; Facebook, Twitter, LinkedIn, Pinterest, Google +, Tumblr, Instagram, Flickr, Meetup, and MySpace have been the world’s most popular social networking sites. Facebook was ranked first with 900,000,000 estimated unique monthly visitors and Twitter ranked second with over 310,000,000 estimated unique monthly visitors (Top 15 Most Popular Social Media Sites | May 2014, 2014).

1.2.1 Increase in Social Media Participation

One of the main drivers for the increase of participants in social media is the development and affordability of smartphone technology. Many users are accessing social media websites using smartphone technology. A smartphone is a mobile cellular phone with more advanced computing capability and Wi-Fi connectivity. They typically include the features of a computer with those of another popular consumer device, like a personal digital assistant (PDA), a media player, and a digital camera (Wikipedia, 2014).

According to Statista, there will be an estimated 220 million users in the United States using smartphones by the year 2018 (as shown in figure 2). A 351% increase compared to 2010 (Statista, 2014). The United States Census Bureau estimates that in 2018 there will be an estimated 335,005 million people living in the United States (United States Census Bureau, 2014). Additionally it is estimated that 20% of smartphones owners use Twitter accounts. It can be assumed by 2018 65% of the population (or over 200 million people in the United States) will be a consumer of smartphone technology.
In 2012 the Pew Research Center’s Internet & American Life Project surveyed how Americans are using the internet and identified Twitter use and demographics. They study found that 15% of online adults use Twitter which is a large increase when compared to only 8% in of adult online Twitter users in 2010 (as shown in figure 3) (Smith & Brenner, 2014).
The survey also identified demographics among the internet users using Twitter; African-Americans 28%, young adults (aged 18 – 29) 26% and specifically, young adults (ages 18-24) 31%. They also found that residents of urban and suburban areas are significantly more likely to use Twitter than their rural counterparts (as shown in figure 4) (Smith & Brenner, 2014).
## Who uses Twitter?

*% of internet users within each group who use Twitter*

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>All adult internet users (n=1729)</td>
<td>15%</td>
</tr>
<tr>
<td>Men (n=804)</td>
<td>14%</td>
</tr>
<tr>
<td>Women (n=925)</td>
<td>15%</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>18-29 (n=316)</td>
<td>26**%</td>
</tr>
<tr>
<td>30-49 (n=532)</td>
<td>14%</td>
</tr>
<tr>
<td>50-64 (n=521)</td>
<td>9%</td>
</tr>
<tr>
<td>65+ (n=320)</td>
<td>4%</td>
</tr>
<tr>
<td><strong>Race/ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>White, Non-Hispanic (n=1229)</td>
<td>12%</td>
</tr>
<tr>
<td>Black, Non-Hispanic (n=172)</td>
<td>28**%</td>
</tr>
<tr>
<td>Hispanic (n=184)</td>
<td>14%</td>
</tr>
<tr>
<td><strong>Annual household income</strong></td>
<td></td>
</tr>
<tr>
<td>Less than $30,000/yr (n=390)</td>
<td>19%</td>
</tr>
<tr>
<td>$30,000-$49,999 (n=290)</td>
<td>12%</td>
</tr>
<tr>
<td>$50,000-$74,999 (n=250)</td>
<td>14%</td>
</tr>
<tr>
<td>$75,000+ (n=523)</td>
<td>17%</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
</tr>
<tr>
<td>No high school diploma² (n=108)</td>
<td>22%</td>
</tr>
<tr>
<td>High school grad (n=465)</td>
<td>12%</td>
</tr>
<tr>
<td>Some College (n=447)</td>
<td>14%</td>
</tr>
<tr>
<td>College + (n=698)</td>
<td>17%</td>
</tr>
<tr>
<td><strong>Geographic location</strong></td>
<td></td>
</tr>
<tr>
<td>Urban (n=520)</td>
<td>19**%</td>
</tr>
<tr>
<td>Suburban (n=842)</td>
<td>14**%</td>
</tr>
<tr>
<td>Rural (n=280)</td>
<td>8%</td>
</tr>
</tbody>
</table>

Source: Pew Research Center’s Internet & American Life Project Winter 2012 Tracking Survey, January 20-February 19, 2012. N=2,253 adults age 18 and older, including 901 cell phone interviews. Interviews conducted in English and Spanish. The margin of error is +/-2.7 percentage points for internet users. **Represents significant difference compared with all other rows in group.**

Figure 4: Who Uses Twitter (Smith & Brenner, 2014)
Our original intention for this research was to identify areas in Washington DC where Twitter use would be highest based on the demographics authored by the Pew Research 2012 study. However the city of Washington DC itself carries very unique challenge demographically with its relationship to its tourism and daily commuters into the city. Including this data would have to be used with extreme caution. According to the 2013 Destination DC visitor study, Washington DC welcomed a record setting 19 million visitors, including 1.6 million visitors from oversees in 2013 (as shown in figure 5). Visitation to Washington DC is expected to follow the growing trend and Destination DC expects over 21 million visitors will visit the city by the year 2017 (Destination DC, 2014).
According to the United States Census Bureau 2006-2010 Commuter-Adjusted Population Estimates study, the District of Columbia is the second largest city that experienced a statistically significant gain in population due to commuters from its surrounding states (as shown in table 2). On a daily basis the city can experience a gain in population by 79% (United States Census Bureau, 2010). With the high influx of tourist and commuters visiting from outside of Washington DC on a daily basis we decided not to include Twitter demographics in Washington DC for the study.
Table 2. Counties of 50,000 population or more based on percent population increase due to commuting: 2006-10

<table>
<thead>
<tr>
<th>County and State</th>
<th>Residence Population</th>
<th>Commuter-Adjusted Population</th>
<th>Percent Population Change due to Commuting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 New York County, NY</td>
<td>1,583,345</td>
<td>3,083,102</td>
<td>94.7</td>
</tr>
<tr>
<td>2 District of Columbia, DC</td>
<td>584,400</td>
<td>1,046,038</td>
<td>79.0</td>
</tr>
<tr>
<td>3 Fulton County, GA</td>
<td>866,982</td>
<td>1,256,406</td>
<td>41.6</td>
</tr>
<tr>
<td>4 St. Louis city, MO</td>
<td>318,809</td>
<td>433,778</td>
<td>36.1</td>
</tr>
<tr>
<td>5 Richmond city, VA</td>
<td>201,826</td>
<td>268,594</td>
<td>33.1</td>
</tr>
<tr>
<td>6 Suffolk County, MA</td>
<td>764,460</td>
<td>932,030</td>
<td>32.3</td>
</tr>
<tr>
<td>7 Lynchburg city, VA</td>
<td>73,726</td>
<td>96,824</td>
<td>31.3</td>
</tr>
<tr>
<td>8 Roanoke city, VA</td>
<td>95,793</td>
<td>124,032</td>
<td>29.5</td>
</tr>
<tr>
<td>9 San Juan Municipio, PR</td>
<td>404,748</td>
<td>522,144</td>
<td>29.0</td>
</tr>
<tr>
<td>10 Norfolk city, VA</td>
<td>242,143</td>
<td>309,689</td>
<td>27.9</td>
</tr>
<tr>
<td>11 Christian County, KY</td>
<td>72,678</td>
<td>92,510</td>
<td>27.3</td>
</tr>
<tr>
<td>12 Denver County, CO</td>
<td>578,087</td>
<td>734,217</td>
<td>27.0</td>
</tr>
<tr>
<td>13 Arlington County, VA</td>
<td>197,467</td>
<td>249,979</td>
<td>26.6</td>
</tr>
<tr>
<td>14 Albany County, NY</td>
<td>304,032</td>
<td>378,209</td>
<td>24.4</td>
</tr>
<tr>
<td>15 Orleans Parish, LA</td>
<td>295,285</td>
<td>358,772</td>
<td>21.5</td>
</tr>
<tr>
<td>16 San Francisco County, CA</td>
<td>769,172</td>
<td>961,627</td>
<td>20.6</td>
</tr>
<tr>
<td>17 Durham County, NC</td>
<td>258,578</td>
<td>310,654</td>
<td>20.1</td>
</tr>
<tr>
<td>18 Potter County, TX</td>
<td>120,124</td>
<td>143,625</td>
<td>19.6</td>
</tr>
<tr>
<td>19 Hennepin County, MN</td>
<td>1,136,522</td>
<td>1,356,471</td>
<td>19.4</td>
</tr>
<tr>
<td>20 Cole County, MO</td>
<td>74,787</td>
<td>89,308</td>
<td>19.4</td>
</tr>
<tr>
<td>21 Lee County, MS</td>
<td>81,446</td>
<td>96,828</td>
<td>18.9</td>
</tr>
<tr>
<td>22 Davidson County, TN</td>
<td>612,884</td>
<td>723,432</td>
<td>18.0</td>
</tr>
<tr>
<td>23 Pulaski County, AR</td>
<td>377,060</td>
<td>444,943</td>
<td>18.0</td>
</tr>
<tr>
<td>24 Dauphin County, PA</td>
<td>264,823</td>
<td>312,545</td>
<td>18.0</td>
</tr>
<tr>
<td>25 Anderson County, TN</td>
<td>74,257</td>
<td>86,994</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Source: American Community Survey, 2006-10.
For more information on the ACS, see www.census.gov/acs.

Table 1: Counties of 50,000 population or more based on the percent population increase due to commuting: 2006 - 2010 (United States Census Bureau, 2010)
1.2.2 Volunteered Geographic Information (VGI)

Within the last few years smartphones have become equipped with global positioning system (GPS), thus enabling users to provide locational data when posting/uploading information to various social media sites (Sui & Goodchild, 2011). When the users post/upload the information, they volunteer information by providing location data (known as a geotag) and as well as their thoughts to the world. The capacity to study social media can be tied directly to the rise of Web 2.0 technologies. Web 2.0 technologies are web applications that allow for an efficient exchange of information that are centered on user design and input, and are hosted on the internet. Social media sites like, Wikipedia, Blog EBay, OpenSource Street Map and Blog sites are good examples. Web 2.0 is defined by its main characteristics, which include user-generated content, crowdsourcing, big data, networking and transparency of data. Essentially anyone with an internet connection can add a topic and provide a description, including links and other sources (Goodchild, 2007).

This type of information has been characterized as volunteered geographic information (VGI) and is a readily accessible way to harvest information about people, events and networks that can be visualized through mapping (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013) (Stefanidis, Crooks, & Radzikowski, 2012). The information allows the business community to peer into consumers head and customize marketing campaigns, driving more customers to their business. However, it can also be used to map and display where various events that are taking place. Research has shown that Twitter has been used to in emergency response situations allowing for the
identification and localization of areas impacted by natural disasters (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013).

VGI is not a new phenomenon, citizen-driven data collection efforts have used in the past. In the 1930s and 1940s Stamps land-use survey of Britain, the main data collectors were volunteer teachers and students (Elwood, Goodchild, & Sui, 2012).

1.2.3 Ambient Geographic Information (AGI)

As social media began to gain a handle into society, social media sites began to add a geolocation feature (or tags), known as ambient geographic information (AGI), as their customer posted information to their perspective sites. AGI is a way for people to contribute geographical information but must be harvested and analyzed using a social media site application programming interface (API). Where VGI can provide first responders a probable specific area impacted by natural disasters, AGI can provide the actual context. For example, through data analysis of AGI, it may be determined that multiple injuries occurred in one area and a bridge collapsed in another area allowing first responders the ability to direct specific resources to areas in need (Stefanidis, Crooks, & Radzikowski, 2012).

1.3 Why Law Enforcement Should Use Social Media

Research indicates that 66% of adult’s online use some form of social media (Malleson & Anderson, 2014). Social media streams offers a new source of large amounts of information allowing law enforcement a new ability to peer into the perceptions, opinions, actions, feelings and tensions expressed by individuals and
neighborhoods. Analysis using Twitter's API can be useful for discovering criminal activity or identifying area affected by natural/artificial disasters in near real-time (Alkutkar, Sam, Tambe, & Ainapure, 2012). By harvesting Twitter data and mining the text using keywords we can draw conclusions of public activity and identify potential hot spots where specific activity is taking place.

Over time community policing has gained and lost momentum among the police and community leaders as they search for effective ways to reduce crime and enhance their communities. In order to be effective, community policing requires the active participation of local government, civic and business leaders, public and private agencies, area residents, churches, schools and hospitals (Bureau of Justice Assistance, 1994). Social media has the potential to act as a bridge, allowing for two-way communication with law enforcement and the community that has the ability to connect all parties concerned with their communities.

**Purpose of the Study**

In view of the increased interest in community policing on one hand and the proliferation of social media on the other hand, this thesis aims to explore how social media can be used within the community policing paradigm. In particular, we seek to investigate whether precisely geolocated social media feeds can be used to identify and study the impact areas of specific types of crimes in urban areas, as well as to gauge public awareness to crime. What we hope to accomplish with our study is develop a method in which social media, specifically Twitter, can be studied to identify how law
enforcement agencies can use a citizen-driven tool for empowering community policing activities, and supporting intelligence-led policing.

1.4 Thesis Organization

The structure for the remainder of this paper will be organized as follows. Chapter Two identifies previous studies surrounding crime mapping, community policing and analysis of social media applications. Chapter Three describes the methodology used for our research. Chapter Four presents a crime analysis of our study area. Chapter Five presents our analysis of crime and Twitter. Chapter Six discusses our findings from our study and potential future work this study may generate.
CHAPTER TWO: BACKGROUND

Crime analysis has been extensively studied to understand criminal behavior with an overall goal of reducing or eliminating crime all together. The definition of crime analysis is; the qualitative and quantitative study of crime and law to apprehend criminals, prevent crime, reduce disorder, and evaluate organizational procedures (Boba, 2001). This chapter will discuss previous studies regarding crime analysis by examining crime mapping methods, repeat victimization theory, social media and law enforcement use, crowdsourcing and finally we identify any intelligence gaps we identified from previous research.

2.1 Crime Mapping

Crime mapping is a well-established technique employed by most law enforcement agencies which relies heavily on the capturing of previous crime data (Brantingham & Brantingham, 1997). Repeat victimization of individuals, businesses and residences is a well-established. By definition, repeat victimization occurs when the same type of crime incident is experienced by the same (or virtually the same) victim or target within a specified period of time such as a year. The victim may be an individual, a dwelling unit, a business or even a chain of businesses at multiple locations. Motor vehicles can also be victims of repeat victimization (Weisel, 2005).

By capturing information such as the type of offense, time of day, day of the week, month and where it took place can enable analysts the ability to generate
predictions about time of the day crimes will occur, the day of the week but more importantly where the locations of future crimes will be committed (Brantingham & Brantingham, 1997).

In 1995 when New York City was embracing Broken Windows Theory they also made a drastic change to move to towards the reliance on statistics and automated mapping, termed CompStat. Crime data was collected and analyzed using GIS in order to improve accountability and resource allocation. By mapping the distribution of criminal activity over specified areas, the police can deploy officers to high-crime areas and track changes over time. Since its inception this philosophy has transformed policing and has since been adopted by nearly every law enforcement agency in the country. However CompStat is fundamentally reactive, the goal of predictive policing is proactive to prevent crime from occurring in the first place (Bachner, 2012).

In 1997 Patricia L. Brantingham and Paul J. Brantingham published Mapping Crime for Analytic Location Quotients, Counts, and Rates. They understood that crime mapping had become a key tool for crime analysis to understand and visualize patterns across time and space. They looked for ways to improve crime mapping by mapping three different crime measures – crime count, crime rate and the crime location quotient (LQC) (Brantingham & Brantingham, 1997).

They theorized that crime counts were being used to identify hot spot locations. Crime rates were used to access areas where specific demographics were targets of specific locations. The LQC can be used for comparison that helps identify whether a specific crime pattern is disproportionately high or low in a particular place or location. It
is a tool that can be used at different levels of spatial and temporal resolution. LQC was a new approach to study crime, there is no need to obtain a count of the number of targets as is necessary in calculating a crime rate. For example, the LQC for robbery would be based on counts of robberies and all crimes not population or number of target businesses (Brantingham & Brantingham, 1997).

Their research would lay the foundation for the future of predictive policing. In July 2011, the Santa Cruz Police Department became the first to deploy a new experimental predictive policing software, known as PredPol. The program was developed at the University of California, Los Angeles and Santa Clara University and used Brantingham and Brantingham theory that you need only three pieces of data to make predictions for crime – type of crime, place of crime, and time of crime. Using advanced mathematics and computer learning, PredPol’s algorithms analyzes the data and assigns a 500 foot by 500 foot square where a crime will likely take place that day. Prior to their shifts, officers are briefed on the locations and are encouraged to devote extra time to monitoring these areas. The algorithm theoretically predicts many types of crime, including property crimes, drug incidents, gang activity, and gun violence as well as traffic accidents. By July 2012 the program moved for experimental to fully operational and is now in multiple cities, to include Kent Country, England, Seattle, Washington, Atlanta, Georgia, and Los Angeles, California (Bachner, 2012) (PrepPol, 2014) (Brantingham & Brantingham, 1997).
2.1.1 Social Media and Law Enforcement

In 2008 Police Forces in the United Kingdom began experimenting with social media in an effort to engage the public, build trust with the public and instill confidence in the police in an effort to reduce crime. They started by posting incidents on Twitter and by 2010 they went from 3,000 followers to 17,000, a 567% gain (Crump, 2011). It is estimated today that more than 2,800 law enforcement agencies in the United States have a social media account, a number that is still growing (Davis, Alves, & Sklansky, 2014).

Law enforcement agencies have been using social media in a variety of applications. Social media is being used to enhance citizens' input in police investigations, to strengthen the public image of police departments, to control crowds, to issue public service announcements, to obtain better input in policy-making processes and to attract new police officers (Meijer & Thaens, 2013).

Many police departments are using various forms of social media for research and monitor user’s posts. In 2012 a study by the Lexis Nexis Risk Solutions found that 80% of officers were using social media in their investigations. The police were turning to social media sites to identify people and locations, discover criminal activity and gather evidence (Williams, et al., 2013). In 2009 the Metropolitan Police Department in Washington DC (MPDC) began using YouTube for investigation purposes. Ironically in response to the 2010 earthquake in Washington DC, MPDC realized the potential that Twitter could be used to aid in getting information out to their citizens (Meijer & Thaens, 2013).
2.1.2 Crowdsourcing

One particular area that law enforcement has found success with social media has been through the use of crowdsourcing. Crowdsourcing is closely related to the concept of VGI, where information obtained from a crowd of many observers is likely to be closer to the truth than information from one observer (Goodchild & Glennon, Crowdsourcing geographic information for disaster response: a research frontier, 2010). Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community. The process of crowdsourcing is often used to subdivide tedious work and has occurred successfully offline. It combines the efforts of numerous self-identified volunteers or part-time workers, where each contributor of their own initiative adds a small portion to the greater result (Wikipedia, 2014). For example, within 10 minutes of the Boston Marathon bombing, the Boston Police Department turned to Twitter and started posted messages and keep the public informed as the situation developed. They kept the public and media informed through Twitter about road closures, news conferences and police activates (Davis, Alves, & Sklansky, 2014).

However the Boston Police Department went a step further and opened up two-way communication process and started requesting photos and videos captured on digital cameras and smartphones from the finish line in order for the police to investigate and possibly discover possible suspects involved in the bombing. Through crowdsourcing the Boston Police Department were able to identify potential suspects that eventually lead to the capture and death of the perpetrators.
Another area where social media text is analyzed is the area were sentimental analysis determined from the text in social media content. A “social media tension-monitoring engine” was developed in 2010 by the Cardiff Online Social Media Observatory (COSMOS). This engine monitors social media data streams for signs of high tension that will be evaluated to find deviations from the ‘norm’ (levels of cohesion/low tension) (Williams, et al., 2013). When areas are identified, appropriate measures can be taken to investigate the area further, from the use of closed caption television (CCTV) or police presence.

In 2011 the city of Melbourne, Australia proposed an interactive mobile platform called Transafe, which captures and analyses public perceptions of safety by using crowdsourcing collection efforts. Their goal was to use an interactive social media application that visualizes the community’s persecutions of safety with a goal to understand how the public’s perceptions of safety and develop policies and strategies. Though crowdsourcing the public could get a better idea to the safety of a particular neighborhood, or public transportation route. In addition, they have the ability to influence others based on their safety experience in Melbourne (Hamilton, Salim, Cheng, & Choy, 2011).

2.1.3 Analysis of Social Media Applications

Over the last few years the public sector has been gaining traction on how they can use Twitter to gain a competitive advantage over their competitors. They have come to understand that the average citizen is now a walking eye on the world, a citizen journalist. A citizen who is able to take a photo, add a caption or a short story and upload
it to the Internet for all their friends, and usually everyone else, to see (Williams, et al., 2013). The private sector has come to realize that by analyzing the text in Twitter they can delve into the raw thoughts the public is expressing.

In 2010 Sitaram Asur and Bernardo Huberman demonstrated how social media content can be used to predict real-world outcomes. They theorized that the movies that will be talked about the most in Twitter will directly lead to which movies will be watched in theaters. They extracted Tweets using Twitter's API over a period of three months where the context revolved around 24 different new movies released in theaters during their collection period. Almost 3 million Tweets were collected and analyzed by mining the tweet text, against the timestamp and author. Their results validated that chatter among the Twitter community can be used to make quantitative predictions when compared to actual movies sales. Movies that are talked about the most will be the most watched in movie theaters (Asur & Huberman, 2010).

A similar study by Hillary Bliss mined the text of Tweets using Twitter's API over a period of a few days in relation to the 2013 movie pop culture sensation Sharknado. She demonstrated that using a word cloud generator can show the most common entries within Tweet text. The information that can be extracted from Twitter is a valuable nearly free focus group that the private and public sectors can benefit from. Hillary Bliss work shows that a company can mine Twitter that mentioned their brand name, as well as their competition, to get a general idea of what people are talking about in relation to their product (Bliss, 2014).
In an effort to see how the public perceives its brand, companies can understand how social media drive marketing decisions. Researchers Wu He, Shenshua Zha, and Lin Li data mined the text content on Facebook and Twitter searching for chatter surrounding the nation’s three largest pizza chains; Pizza Hut, Domino’s Pizza and Papa John’s Pizza. They also concluded that analysis of social media content can allow a business a competitive advantage for its business and its competitors. Through analysis of social media text a business can drive traffic, increase customer loyalty, increase sales and revenue, increase customer satisfaction, create brand awareness and build a reputation (He, Zha, & Li, 2012).

One area that has been researched is data mining of social media text to predict the outcomes of political elections. In 2014, researchers Kazem Jahanbakhsh and Yumi Moon analyzed 32 million tweets in relation to the United States presidential election by using a combination of machine learning techniques. Their analysis was later compared to the actual election results. One technique they introduced was the data mining of geotagged Tweets to discover popularity of candidates by geographical location. Their analysis revealed that data mining of geotagged Tweets can predict the actual outcomes by a surprising 76% success rate (Jahanbakhsh & Moon, 2014).

Many of the previous crime studies surrounding predictive analysis have concentrated on past events such as type of crime, place of crime, and time of crime. This data led to the development and visualizing of crime mapping applications. However applying social content geospatially has though social media has not been researched extensively. One study proposed an empirical analysis approach to analyze the
relationship between online social user interaction and crime incidents. They collected Twitter Data using Twitter's API and recorded crimes in a fine-grained manner concerning both dimensions time and location according to an area from within the city of San Francisco. They demonstrated that Twitter data using geotagged tweets can discover patterns and can strengthen the explanation of criminal activity in urban areas and provided evidence Twitter data can improve prediction of crime (Johannes, Tobia, Sebatain, & Dirk, 2014).

As we have discussed, social media can have impact on how law enforcement distributes information, how the public can contribute to crime solving through crowdsourcing and how data mining of social media text can have a direct effect on communities. In 2012 researchers developed a java based application called an “Area Danger Check” using Googles Android technology. They developed an algorithm that data mines geotagged Tweets using Twitter's API by searching for associated keywords that will analyzed by the algorithm and determined if it a positive tweet or a negative tweet, along with a severity raking. They found that their application but be analyzed for social behavioral or ethnographical studies on public safety perceptions and help to inform the stakeholders of the city for future policy-making and policing strategies for safety perception management. (Alkutkar, Sam, Tambe, & Ainapure, 2012).

2.2 Identified Intelligence Gaps from Prior Research

Although social media may provide a powerful tool for law enforcement, analysis of Twitter come with concerns using Twitter API. For example, Twitter's API has several services including the Spritzer, a free randomly sampled containing approximately 1% of
all tweets, the Gardenhose, a paid randomly sampled containing 10% of all tweets and the Firehose, a paid service containing all public tweets. Most of the previous research have concentrated on use of Twitter's free API service. Most free and commercial software programs available that facilitate API access are subject to ‘rate limitations’ making missing data a potential problem (Williams, et al., 2013).

Our study will also utilize Twitter's free API and will collect a random sample of 1%. Yet our study plans to concentrate only on precise geotagged tweets. Previous research has demonstrated the percentage of precisely geotagged tweets range anywhere from 0.5% to 3% of the 1% depending on the area and underlying conditions (Croitoru, et al., 2014). We have to consider the small sample size and must be careful about making assumptions and generalizations based on a small sample.

Demographic information is also limited, with most APIs provide only the account name of social media users. If social media networks are to be utilized for the monitoring we will have to account that the ambient population may be misrepresented (Williams, et al., 2013).

Finally the drivers of VGI information operate without any kind of standardized or community accepted controls. The information this is generated does not typically come from subject matter experts and does not have to undergo quality controls measure from experts in the field (Elwood, Goodchild, & Sui, 2012).

Although we understand intelligence gaps will exist, we believe the sample size of data used in this study will be adequate to draw conclusions. We hope to build from
the methods described from previous works and develop an approach for how Twitter can be used as a law enforcement tool.
CHAPTER THREE: METHODOLOGY

Using the Washington D.C. metropolitan area as a case study, we collected both Twitter data and utilized police records data categorized by crime types in 2013 and 2014. Using this data, density maps of reported crimes were compared to density maps of geolocated tweets containing crime-related keywords. This comparison provided additional information regarding the alignment between crime (and police) and Twitter activity, thus allowing us to gauge public response and awareness to criminal activities.

Figure 6: Research Road Map
In this section we present three methods used in our study. The first approach develop a method to data mine and search Twitter text using keywords and phrases to locate areas where the keywords and phrases are mentioned (as shown in figure 6) (Described in section 3). The second approach will utilize a classic approach to crime though the use of crime or “hot spot” mapping and statistical analysis. We then map and analyze criminal statistics for all of types of crimes that occurred in Washington DC in both the summers of 2013 and 2014. Our research will also study community policing though field work with Washington DCs District 3 police patrols. We will draw upon the expertise of the patrol officers and map areas they identify as high crime areas and compare them with actual reported crimes (as shown in figure 6) (Described in Section 4).

Our third approach will examine the summer of 2014 Twitter data using the keyword and phrase search method and map the hot spots using the same mapping techniques we used from crime data. The Twitter data was also studied temporally to identify high Tweet days and cross referenced with open source data to validate probable reasons (as shown in figure 6) (Described in Section 5). Finally we will investigate the geospatial relationship of keywords from Tweets and crime (as shown in figure 6) (Described in Section 6).
3.1 Methodology to Process to Capture, store and Filter Twitter Data

3.1.1 Keywords for Harvesting Data

The data was harvested from the Twitter streaming API using different keywords that may be tweeted to express various crimes. The keywords were chosen by conducting a word cloud analysis from 16 to 29 April 2014 using Tagxedo – Creator, (Lang, 2014) a web service that pulls from Twitter’s API to generate a word cloud of what was talked about most, revolving around “crime in Washington DC”. The Tagxedo word cloud generator uses the word size as a depiction of what words were most tweeted (as shown in figure 7). Other word cloud generators, such as those found on Wordle.net and tagcrowd.com offer the ability to customize the words clouds, meaning not all word clouds should be interpreted the same (Feinberg, 2014). The generator was ran daily at approximately 1700 to maintain a consistency and compared to open source data to filter out high frequency words. We also manually inspected the tweets to gleam out the eight key phrases.
Figure 7: Word Cloud Generated Using [www.Tagcrowd.com](http://www.Tagcrowd.com) on 16 April 2014

Analysis of the word clouds identified ten primary words and eight phases that will be used as keywords to filter possible crime related data (as shown in table 3). Each primary word and phase will include singular and plural meaning and variants, i.e. theft/thefts/thief.

<table>
<thead>
<tr>
<th>Primary Words</th>
<th>Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theft</td>
<td>Gun shot</td>
</tr>
<tr>
<td>Burglary</td>
<td>Fatal stab</td>
</tr>
<tr>
<td>Robbery</td>
<td>Found dead</td>
</tr>
<tr>
<td>Murder</td>
<td>Police Officer</td>
</tr>
<tr>
<td>Rape</td>
<td>Fatal Shooting</td>
</tr>
<tr>
<td>Kill</td>
<td>Arrested for</td>
</tr>
<tr>
<td>Assault</td>
<td>Sexual assault</td>
</tr>
<tr>
<td>Evacuated</td>
<td>Body found</td>
</tr>
</tbody>
</table>
3.1.2 Open-Source Comparative Analysis

In order to account for non-crime related tweets that may skew the word cloud, open source data was reviewed during the collection period to account for the numerous “hot words” the word cloud identified. Analysis of open source data identified that a United States Cyber Crime Conference occurred on 27 – 28 April 2014 offered by the SANS Institute (SANS Institute, 2014).

A high profile trial in Washington DC occurred involving celebrity, Chris Brown which generated a lot of Tweets and the word cloud generator identified the high levels of traffic. Singer Chris Brown, who was trial for a misdemeanor assault charge, trial was delayed from 21 April to 23 April 2014 (Gresko, 2014). Another event occurred on 21 April 2014 when the Washington DC National Zoo was evacuated in response to a shooting outside the gate of the Zoo, on Connecticut Avenue (Stieber, 2014). In addition, the words “Tours” and “Museums” were tweeted regularly during the collection period. The Tweets are most likely due to a popular attraction in downtown Washington DC known as the Crime Museum. The museum is a popular tourist attraction that provides;

“Guests of all ages with memorable insight into our nation’s history of crime and its consequences, law enforcement, forensic science, and crime scene investigation through a captivating, interactive, entertaining, and educational experience” (Crime Museum, 2014).
Open-source comparative analysis will again be ran to identify possible events that occurred in the Washington DC area during the summer of 2014 that again may skew the data and may not be actual crime events.

3.1.3 Twitter Data Collection, Storage and Filtering

George Mason University’s (GMU) Center for Geospatial Intelligence has created a vigorous process for “Harvesting ambient geospatial information from social media feeds.” This system collects and stores Twitter data for analysis (Croitoru, Crooks, Radzikowski, & Stefanidis, Geosocial gauge: a system prototype for knowledge discovery from social media, 2013). The GMU collection system uses a multi-stage process that collects the Twitter data utilizing Twitter’s API to process and filter geospatial enabled tweets using a bounding box method, and finally storing the data in a dedicated database that can be exported for analysis. This system allows for the discovery of valuable geospatial information from social media.

The data can be exported in a variety of formats; javascript object notation (JSON), comma-separated values (CSV) and tab-separated values (TSV). A TSV file was selected based on the stability it provides, where a CSV file may lose stability were the commas included in the text of the actual tweet will cause problems when ingested into a database. The TSV file will be converted into a text file which will be injected into a Microsoft Access database (MDB). The MDB file will be accessed using Environmental Systems Research Institute (ESRIs) ArcCatalog by using an Object Linking and Embedding, Database (OLE-DB) connection. With this connection we can import and
convert the data into a table within a file geodatabase. The table will then be converted into a feature class and stored in a geodatabase within ArcGIS. A geodatabase in ArcGIS was chosen for its ability to store and filter data using a structured query language (SQL), a series of relational functions and operators that can be used to create, modify, and query tables and their data elements (ArcGIS Resources, 2014). SQL operations will then be applied using the keywords previously identified.

The Twitter data was collected using a bounding box method (Upper Left = 38.99 latitude -76.90 longitude and Lower Right = 38.79 latitude - 77.20 longitude) between the dates of 1 June and August 31 2014. A total of 2,129,412 geospatial enabled tweets were collected using the bounding box method. The Twitter data was then filtered using ESRI intersect tool to identify geotagged tweets that occurred within the Washington DC boundary. Using the intersect tool 783,129 tweets were eliminated leaving 1,346,283 tweets within the city of Washington DC. From there a Kernel Density was produced that identified the areas surrounding the White House and United States Capital had the highest instances for Tweets (as shown in figure 8). A kernel density is a common crime mapping method used by law enforcement agencies to analyze subtle trends in criminal movement and to predict the movement of repeat offenders (Manepalli, Bham, & Kandada, 2011).

During the collection period an error occurred while harvesting the Twitter data. There was no collection of Tweets from 6/14/2014 11:30 UTC to 6/23/2014 14:31 UTC and again from 7/72014 23:34 UTC to 7/9/2014 23:14 UTC. The Tweet Histogram
highlights the Tweets per day and a large and small gap can be recognized (as shown in figure 9).

Figure 8: Geotagged Tweets Filtered Using ESRI Intersect Tool and applied to a Kernel Density
The data collected and filtered using utilizing Twitter’s API and filtered through keywords will then be compared and analyzed to the classic crime approach of crime mapping. The Washington DC crime database classifies crimes into specified categories (i.e. robbery, homicide) and the keywords will be categorized to replicate a similar classification method.
CHAPTER FOUR: CRIME ANALYSIS

4.1 Classic Approach to Crime Analysis

We present two approaches for crime analysis for this study. First we will use the classic approach to crime analysis by collect previous crime data and identifying areas, utilizing density mapping techniques to identify where high amounts of crime are occurring. We will also study the crime statistics to study crime rate changes. The second approach will engage subject matter experts from Washington DCs District 3 (the patrol officers) to how they interact with the community in their patrol areas.

4.1.1 Washington DC Crime Data

The District of Columbia government website contains a wealth of information for its residents, ranging anywhere from how to obtain a driver’s license to the street paving schedule (Data Catalog, 2014). Within the site exists a Data Catalog that provides public access to operational data from 498 datasets from multiple agencies and organizations. The data comes in the form of XML, Text/CSV, KML or ESRI Shapefile formats. All of the data can be visualized and mapped to provide geospatial context, which includes crime. The crime data that will be used is reported from nine different types of crimes; assault with a dangerous weapon, arson, burglary, homicide, robbery, sex abuse, stolen auto, theft, and theft from auto. Below are the categories/definitions of each crime (Data Catalog, 2014).

• **Assault with a Dangerous Weapon** - Knowingly or purposely causing serious bodily injury to another person, or threatening to do so; or under circumstances
manifesting extreme indifference to human life, knowingly engaging in conduct that creates a grave risk of serious bodily injury to another person, and thereby causes serious bodily injury. Weapons include, but are not limited to, firearms, knives, other objects, hands and feet.

- **Arson** - The malicious burning or attempt to burn any dwelling, house, barn, or stable adjoining thereto, or any store, barn, or outhouse, or any shop, office, stable, store, warehouse, or any other building, or any steamboat, vessel, canal boat, or other watercraft, or any railroad car, the property, in whole or in part, of another person, or any church, meetinghouse, schoolhouse, or any of the public buildings in the District, belonging to the United States or to the District of Columbia.

- **Burglary** - Breaking and entering, or entering without breaking, any dwelling, bank, store, warehouse, shop, stable, or other building or any apartment or room, whether at the time occupied or not, or any steamboat, canal boat, vessel, other watercraft, railroad car, or any yard where any lumber, coal, or other goods or chattels are deposited and kept for the purpose of trade, with intent to break and carry away any part thereof or any fixture or other thing attached to or connected with the same.

- **Homicide** - Killing of another purposely, in perpetrating or attempting to perpetrate an offense punishable by imprisonment, or otherwise with malice aforethought.

- **Robbery** - The taking from another person, or immediate actual possession of another, anything of value, by force or violence, whether against resistance or by sudden or stealthy seizure or snatching, or by putting in fear. This category includes carjacking’s.
• **Sex Abuse** - One of many sexual acts against another, either forcibly or without his/her permission, and/or against a child or someone who is otherwise incapable of communicating unwillingness. The severity ranges from forcible rape to other forms of sexual contact.

• **Stolen auto** - Theft of a motor vehicle (any automobile, self-propelled mobile home, motorcycle, truck, truck tractor, truck tractor with semi-trailer or trailer, or bus)

• **Theft** - This includes conduct previously known as larceny, larceny by trick, larceny by trust, embezzlement, theft of services and false pretenses. The Theft/Other category excludes theft of items from a motor vehicle or the motor vehicle itself, which are captured under other categories.

• **Theft from Auto** - Theft of items from within a vehicle, excluding motor vehicle parts and accessories.

### 4.1.2 Crime Mapping/Hot Spot Analysis

Employing crime mapping/hot spot analysis allows law enforcement agencies a visual representation on what areas are affected by high crime rates or what areas are not affected. The Washington DC crime metadata provided details about location, neighborhood, ward and shift where the crime was committed. Using the locational data, a kernel density was conducted using ESRI ArcGIS 10.1 in order to visualize the areas where the crimes were committed with regularity.

A kernel density tool calculates the density of features in a neighborhood around those features. Many uses include finding density of houses, crime reports, or roads or utility lines influencing a town or wildlife habitat (How Kernel Density Works, 2014). As
previously mentioned, kernel density mapping is one of the most common crime mapping methods utilized by law enforcement agencies (Manepalli, Bham, & Kandada, 2011). The kernel density displays areas where the highest amount of crimes (overall and crime specific) is displayed in red and lower amounts in green.

Comparison of the 2013 and 2104 crime density maps showed that overall crime is most likely subjected to the well-established phenomenon repeat victimization or individuals, businesses and residences previously mentioned in this study (Brantingham & Brantingham, 1997) (Weisel, 2005). Most of the crime overall falls in the areas north of the Washington DC National Mall, the United Stated Capital, the White House, but crime was most prevalent in District 3 (as shown in figure’s 10 and 11).

A difference map was created comparing the summer of 2014 overall crime density with the summer of 2013 overall crime density (as shown in figure 12). This was created in ArcGIS 10.1 utilizing the raster calculator function. The positive numbers show that in increase in crime occurred in the areas (depicted in light green, yellow, orange, and red) and the negative numbers depict no change in crime occurred (depicted by dark green). Examination of the difference map revels that areas where high crime rates occurred 2013 were also areas were crimes occurred again in 2014. Yet these areas also saw an increase in crime. This indicates the areas are subjected to repeat victimization and the community may respond to high crime rates by engaging more with the police.

When we separated the crimes by type we were able to visualize how different areas are affected by different types of crimes. Geospatially we identified the following
types of crimes crime rate had little to no change when compared between the summers of 2013 and 2014; assault with a deadly weapon, burglary, robbery and theft. The following crimes density maps did show a pattern of movement between summers; arson, homicide, motor theft, sexual assault and theft from an automobile. Below are the kernel density maps generated to reflect all crimes and crimes by type from both the summers of 2013 and 2014 (as shown in figure’s 10 – 30).
Figure 10: Kernel Density of Overall Crime - Washington DC June, July and August 2013

Figure 11: Kernel Density of Overall Crime - Washington DC June, July and August 2014
Figure 12: Overall Crime 2014 minus Overall Crime 2013
Figure 13: Kernel Density of Arson - Washington DC June, July and August 2013

Figure 14: Kernel Density of Arson - Washington DC June, July and August 2014
Figure 15: Kernel Density of Assault with a Deadly Weapon - Washington DC June, July and August 2013

Figure 16: Kernel Density of Assault with a Deadly Weapon - Washington DC June, July and August 2014
Figure 17: Kernel Density of Burglary - Washington DC June, July and August 2013

Figure 18: Kernel Density of Burglary - Washington DC June, July and August 2014
Figure 19: Kernel Density of Homicide - Washington DC June, July and August 2013

Figure 20: Kernel Density of Homicide - Washington DC June, July and August 2014
Figure 21: Kernel Density of Motor Theft - Washington DC June, July and August 2013

Figure 22: Kernel Density of Motor Theft - Washington DC June, July and August 2014
Figure 23: Kernel Density of Robbery - Washington DC June, July and August 2013

Figure 24: Kernel Density of Robbery - Washington DC June, July and August 2014
Figure 25: Kernel Density of Sexual Assault - Washington DC June, July and August 2013

Figure 26: Kernel Density of Sexual Assault - Washington DC June, July and August 2014
Figure 27: Kernel Density of Theft from Automobile - Washington DC June, July and August 2013

Figure 28: Kernel Density of Theft from Automobile - Washington DC June, July and August 2014
Figure 29: Kernel Density of Theft - Washington DC June, July and August 2013

Figure 30: Kernel Density of Theft - Washington DC June, July and August 2014
4.2 Washington D.C Crime Data Crime Statistics Analysis

In addition to the visual representations that crime mapping creates, statistical analysis can aid in identifying crime rates, crimes by type that will directly affect how police patrol throughout the year. Overall, analysis of the crime data in June, July and August in 2014 showed there was an increase in crime by 113% when compared to the same time frame in 2013. Below, we describe the statistical characteristics of crimes during the case study period.

4.2.1 Statistical Analysis of Washington DC Crime 2013

According to the District of Columbia there were 35,821 reported crimes in the District of Columbia for all of 2013, resulting in a crime rate of approximately 98 crimes per day (as shown in figure 31) (Data Catalog, 2014). However the report shows that 92% of the crimes committed revolved around the taking of one’s property – burglary, robbery, theft, motor theft, and theft from an automobile. Consequently, on an average day, approximately 90% of the crimes that are committed involved the taking of one’s property.
Figure 31: Washington DC Reported Crime by Type in 2013

Figure 32: Washington DC Reported Crime by Month in 2013
Overall, the data also shows that crime increases during the months of May to October when compared to October to April (as shown in figure 32). This study chose to concentrate on crimes that were committed during the summer months (June, July and August) in 2014 based on the statistical analysis of Washington DCs crime for all of 2013 since during these months there is an overall increase in crime activity, thus improving our ability to observe the relationships between crime and social media traffic.

Specifically, according to the District of Columbia crime data, in the months of June, July and August of 2013, 9662 crimes were reported which accounts for 26.9% of the reported crime for the year (as shown in figure 33) (Data Catalog, 2014). Approximately 105 crimes were committed daily during the summer months representing a 107% increase compared to the yearly daily average.

As we identified with the yearly average, reporting shows that 92% of the crimes committed during the summer months revolved around the taking of one’s property – burglary, robbery, theft, motor theft, and theft from an automobile. In addition, 10924 crimes were reported during the summer months in 2014 (as shown in figure 34). Analysis of the crime data in June, July and August in 2014 showed there was an increase in crime by 113% when compared to the same time period in 2013 (as shown in figure 35).
Figure 33: Washington DC Reported Crime in June, July and August in 2013

Figure 34: Washington DC Reported Crime in June, July and August in 2014
4.3 Washington DC Field Work

In conjunction with crime and Twitter data, interviews with officers from District 3’s patrol were carried out in order to validate any District 3’s community policing efforts. Interviews with patrol officers revealed their deep understanding where crimes occurred based on their experience in the area, interacting with the community and experience in police work. Although high crime areas are known and patrolled, these areas are not crime-free.

4.2.1 Washington DC Ride-Along Program

The Washington DC Metropolitan Police Department provides numerous public education programs in order to give the public a better understanding of what police work entails. One of these programs is the Police Ride-Along Program which allows residents...
to accompany officers during their tour of duty in a police vehicles in various districts in
the District (DC.gov, 2014). Field work was conducted in Washington DC in order to
interview police officers and gain a better perspective on possible causation of why
criimes committed in Washington DC geospatially.

4.2.2 District 3

Washington DCs District 3 was selected based on the crime mapping and
statistical analysis for all of 2013, which identified District 3 as a hot spot where a
majority of crime is committed in the city. Analysis of crime in the summer months in
2014 also showed that District 3 had the highest amount of crime during the study period
(as shown in figure 36).
Field work was conducted in Washington DCs District 3 on 19 - 20 July 2014 between the hours of 2300 and 0300. Through field work we were able leverage District 3’s front line subject matter experience and identified hot spots thought the eyes of the Patrol Officer. We conducted multiple interviews which identified the following areas, intersections and streets as hot spots for crime in Washington DCs District 3 from the patrol officers perspective; Adams Morgan, U Street, 14th Street, 18th Street, Sherman Street and Georgia, Sherman Street and Barry Street. A kernel density map was produced
to map areas where the highest amount of crimes were committed in District 3 (as shown in figure 37).

The density map shows that the District 3 patrol officers are familiar with their area and have a good understanding on where crimes are committed. The kernel density shows that the areas, intersections and streets acknowledged as hot spots by law enforcement held the highest incidents for crime according to the crime map. One particular area, Adams Morgan was identified as one of the hot spots, and the kernel density shows a majority of crimes committed in District 3 occurred in the Adams
Morgan area. This is the same area where the highest amount of crime occurred in Washington DC in 2013.

The Adams Morgan area is considered to be a culturally diverse neighborhood in Washington, DC’s District 3, centered at the intersection of 18th Street and Columbia Road. Adams Morgan is considered the center of Washington's Hispanic immigrant community, and is a major hub for night life with many bars and restaurants, particularly along 18th Street (the primary commercial district) and Columbia Road. Much of the neighborhood is composed of 19th- and early 20th-century row houses and apartment buildings (Wikipedia, 2014).

However it is also considered to be one of Washington DCs highest crime areas. District 3 has also identified this area as a high crime risk area and has a dedicated Adams Morgan Team composed of 20 – 30 Police Officers per night that patrol the area on foot and bike for crime reduction purposes.

Another the question was asked, “Why are there so many crimes in District 3?” Without hesitation every beat cop’s answer was gentrification. The U.S. Department of Housing and Urban Development defines gentrification as the process where a “neighborhood occupied by lower-income households undergoes revitalization or reinvestment through the arrival of upper-income household.” The Adam Morgan area has been identified as an area where gentrification is taking place and has been studies in depth (Lee, 2010).

While on patrol it was clear the patrol officers had a handle on why crime occurred with frequency in their district. The district had a combination of risky facilities,
like colleges, bars and shopping areas that routinely attract or generate a disproportionate amount of crime. For example, lots where students routinely park may generate more larcenies from vehicles because the vehicles of students may routinely contain desirable electronic equipment (Weisel, 2005).
CHAPTER FIVE: CRIME AND TWITTER RESULTS

For our third approach our study examined the summer of 2014 Twitter data using the keyword and phrase search method and mapped the hot spots using the same mapping techniques we used for mapping crime hot spots. The Twitter data was also studied temporally to identify high Tweet days and cross referenced with open source data to validate probable reasons. We utilized a word cloud generator to aid in filtering high Tweeted words. In this chapter we present the results of this analysis.

5.1 Filtering Twitter Data by Keywords

The Tweets were filtered using ESRI’s “Select by Attribute” tool. Each keyword was searched applying various spelling to account for capitalization, plural and past tense spelling. For the keyword phrases a combination of capitalizations was used. In order to account for variations in words, various methods were attempted to explore what spelling search created the best results (as shown in figure 38). A misspelling word generator was used from the website http://tools.seobook.com/spelling/keywords-typos.cgi. According to the website the tool can be used to generate a list of typos and common keyword misspellings that were identified from Overture (now Yahoo! Search Marketing), Microsoft adCenter (now Bing Ads) and Google AdWords (SEOBOOK, 2014). The misspelling word generator was reviewed and generated mixed results. The misspelling word generator was not used in this study.
There were 158 theft Tweets collected during the collection period. Analysis of the tweets using the keyword “theft” showed that a majority of the geotagged tweets were tweeted near the White House, National Mall and United States Capital (as shown in figure 39). A look at the word cloud shows the Tweets revolved around the words; wagetheft and identity (as shown in figure 41).

A temporal look revels there were three dates when tweets mentioning theft elevated, June 3, July 14 and August 6 (as shown in figure 40). When compared to open-source media major events occurred in Washington DC in regards to theft. On June 3, 2014 the Washington DC Council voted on legislation known as the Wage Theft

Figure 39: Theft Kernel Density
5.1.2 Burglary

Only 17 burglary related Tweets were collected during the study period. Analysis of the Twitter density also showed that the geotagged tweets were located near the White
House (as shown in figure 42). The temporal examination did not reveal a large peak in chatter, but an inspection of the Tweets shows a majority of the Tweets were from the Twitter handle WashingtonCP (as shown in figure’s 43 and 44) (Washington City and Press). The Tweets revolved around crowdsourcing asking to public for help in the identification of suspects connected with local robberies.

Figure 42: Burglary Kernel Density
Figure 43: Burglary Temporal Analysis

Figure 44: Burglary Tweets Word Cloud Generated Using Tagcrowd.com
5.1.3 Robbery

A total of 152 Tweets were collected using the keywords Robbery. Examination of the Twitter Density shows a majority of the Tweets centered near the White House and United States Capital (as shown in figure 45). Temporally there was one significant spike in chatter. On August 15, 2014 the spike revolved around the Ferguson Police releasing a video of Michael Brown, who was shot and killed, robbing a conscience store (as shown in figure’s 46 and 47) (Berman & Lowery, 2014).

Figure 45: Robbery Kernel Density
A total of 785 Tweets were collected using the keywords murder. The Density shows that a majority of the geotagged Tweets occurred near the United States Capital and National Mall (as shown in figure 48). Temporal Analysis shows a spike on July 26,
2014 (as shown in figure 49). However 64 of the 65 Tweets were authored by Twitter Handle James_Rogers_II sending Tweets to the CIA, FBI and President of the United States. Investigation of the word cloud also shows that James_Rogers_II Tweets dominated, yet again there is quite a bit of chatter surrounding the Michael Brown fatal shooting in Ferguson, Missouri on August 9, 2014 (as shown in figure 50) (Wikipedia, 2014). The temporal analysis also shows a spike in chatter after the Ferguson shooting took place.

![Figure 48: Murder Kernel Density](image-url)
5.1.5 Rape

A total of 893 Tweets were collected during the study period using the keyword rape. The density revealed a majority occurred near the United States Capital and National Mall area, but also had three additional hot spots around the Washington Navy Yard and near North and South DC (as shown in figure 51). Temporal analysis shows
that the Tweets occurred on a semi-regular basis (as shown in figure 52). Examination of
the word cloud does not show specific topics (as shown in figure 53).

Figure 51: Rape Kernel Density
5.1.6 Kill

A total of 9226 Tweets were collected using the keywords kill. The density shows that a majority of the Tweets occurred around the United States Capital, but hot spots were rampant throughout central and southeast DC (as shown in figure 54). Temporal
analysis shows a spike in chatter on June 30, 2014 and August 9, 2014 forward (as shown in figure 55). Inspection of the June 30 Tweets revealed many of the Tweets were in response to the 2014 BET Awards Show and the spike in August was in response to the Michael Brown Ferguson fatal shooting (Raspers France, 2014) (Wikipedia, 2014). The word cloud analysis did not show any specific topics and were vague (as shown in figure 56).

Figure 54: Kill Kernel Density
There were 340 Tweets collected during the study period using the keyword assault. The density shows two hot spots, one is near the White House and the second near the United States Capital (as shown in figure 57). Temporal analysis shows that the Tweets occurred on a semi-regular basis (as shown in figure 58). Examination of the word cloud shows that chatter surrounding assault were sexual in nature (as shown in figure 59).
Figure 57: Assault Kernel Density

Figure 58: Assault Temporal Analysis
5.1.8 Evacuated

A total of 76 Tweets were collected using the keywords evacuated. The density shows that a majority of the Tweets occurred around the White House and the United States Capital (as shown in figure 60). Temporal analysis shows a spike in chatter on June 6, 2014 (as shown in figure 61). Open-source comparison of the June 6 Tweets revealed many of the Tweets were in response to the aircraft in the area violated restricted airspace and communications with the plane could not be made. The United States Capital and surrounding office buildings were temporarily evacuated (Associated Press (AP), 2014). The word cloud analysis shows chatter about the airspace violation, but also shows chatter in response to hurricanes that we active during the summer months.
(as shown in figure 62).

Figure 60: Evacuated Kernel Density
Figure 61: Evacuated Kernel Density and Temporal Analysis

Figure 62: Evacuated Tweets Word Cloud Generated Using Tagerowd.com

5.1.9 Shot

There were 3,585 Tweets collected using the keywords shot. The density shows that a majority of the Tweets occurred around the White House, National Mall and the
United States Capital (as shown in figure 63). Temporal analysis shows a spike in chatter on July 1, 2014, July 17, 2014 and began to spike August 9, 2014 moving forward. The August 9 continuous chatter was in response to the Michael Brown fatal Ferguson shooting (Wikipedia, 2014). Chatter surrounding July 1 was surrounding the 2014 World Cup Soccer match where the United States lost 2-1 to Belgium in the Round of 16 match (as shown in figure’s 64 and 65) (FIFA, 2014). And the July 17 spike was in response to a commercial flight that was shot down over eastern with almost 300 people onboard (Cohen, 2014).

Figure 63: Shot Kernel Density
5.1.10 Arrested

There were 619 Tweets collected during the study period filtered using the keyword arrested. The density shows that a majority of the Tweets occurred around the White House, National Mall and the United States Capital (as shown in figure 66). Temporal analysis shows a significant spike in chatter on June 5, 2014 and again August 14, 2014 (as shown in figure 67). Open-source comparison shows the chatter after August
14, 2014 was in response to the protests surrounding the Michael Brown fatal Ferguson shooting. We could not make a connection through open-source surrounding the June 5 Tweet spike. Analysis of the word cloud shows multiple protests occurring in the DC area and many people Tweeting possible arrest for the protest and/or activism (as shown in figure 68).

Figure 66: Arrested Kernel Density
Figure 67: Arrested Temporal Analysis

Figure 68: Arrested Tweets Word Cloud Generated Using Tagcrowd.com
5.2 Filter by Key Phrases

5.2.1 Gun Shot

There were 25 Tweets collected during the study period filtered using the keyword phrase “gun shot.” The density shows the highest amount of Tweets occurred in the east portion of Washington DC, known as the Capital Heights area (as shown in figure 69). A large spike occurred on July 4, 2014 in response to the fireworks occurring during the various firework shows and/or neighborhood participants. However the mini spikes in Tweet activity indicated that the author possibly heard gun shots in the area (as shown in figure 70). Analysis of the word cloud also shows many Tweets surrounding the fourth of July festivities, but also shows people Tweeting about hearing gun shots, either real gun shots of fireworks that sound like gun shots (as shown in figure 71).
Figure 69: Gun Shot Kernel Density

Figure 70: Gun Shot Temporal Analysis
5.2.2 Fatal Stab

No data was collected filtering for the search phrase “Fatal Stab.”

5.2.3 Found Dead

There were 28 Tweets collected during the study period filtered using the keyword phrase “found dead.” The density shows that most of the Tweets were located near the White House and United States Capital (as shown in figure 72). Analysis of the temporal graph and examination of the Tweets shows a large spike occurred on August 11, 2014, this spike is in reference to the Oscar-Winning Comedian, Robin Williams’s suicide (Itzkoff, 2014). The small spikes are from various news organizations like, the previously mentioned WashingtonCP, authoring Tweets indicating a crime has occurred and bodies have been found (as shown in figure 73). Analysis of the word cloud also
indicates high amount of Tweets in reference to Robin Williams’s death, but also information from the various news agencies (as shown in figure 74).
Figure 73: Found Dead Temporal Analysis

Figure 74: Found Dead Tweets Word Cloud Generated Using Tagcrowd.com
5.2.4 Police Officer

There were 164 Tweets collected during the study period filtered using the keyword phrase “police officer.” The density shows that a majority of the Tweets came from the areas near the White House, the United States Capital and the Mount Vernon Square area (as shown in figure 75). Analysis of the temporal graph shows a large and continued spike and surge of Tweets, which in a direct response to multiple people Tweeting about the Michael Brown fatal Ferguson shooting involving a police officer (as shown in figure 76) (Wikipedia, 2014). Inspection of the smaller spikes shows a majority of the Tweets were in reference to various community issues involving have with the police and the word cloud analysis seems to indicate nothing specific is talked about other than the Ferguson shooting (as shown in figure 77).
Figure 75: Police Officer Kernel Density

Figure 76: Police Officer Temporal Analysis
5.2.5 Fatal Shooting

There were only 7 Tweets collected during the study period filtered using the keyword phrase “fatal shooting.” The density shows that most occurred near the White House area (as shown in figure 78). Analysis of the temporal graph and word cloud analysis shows that most of the Tweets were from various news organizations asking for crowdsourcing help to identify suspects in some fatal shootings as well and a few Tweets about the Michael Broun Ferguson shooting (as shown in figure’s 79 and 80).
Figure 78: Fatal Shooting Kernel Density

Figure 79: Fatal Shooting Temporal Analysis
5.2.6 Arrested For

There were 58 Tweets collected during the study period filtered using the keyword phrase “arrested for.” The density shows a majority of the Tweets occurred near the White House and United States Capital area (as shown in figure 81). Analysis of the temporal graph and word cloud shows that a majority of the tweets were from news agencies tweeting about various events in addition to citizens expressing opinions to their disagreement with laws in relation to marijuana. Conversely chatter was limited surrounding the Michael Brown Ferguson shooting (as shown in figure’s 82 and 83).
Figure 81: Arrested For Kernel Density

Figure 82: Arrested Temporal Analysis
5.2.7 Sexual Assault

There were 111 Tweets collected during the study period filtered using the keyword phrase “sexual assault.” The density shows a majority of the Tweets were around the White House, National Mall, United States Capital and DuPont Circle areas (as shown in figure 84). Inspection of the temporal graph, word cloud and Tweets did not indicate a specific subject, with the exception of the July 30, 2014 spike which was from chatter surrounding a bill introduced in the Senate looking to crack down on sexual assault on college campuses (as shown in figure’s 85 and 86) (Kingkade & Fang, 2014). However a majority of the Tweets were talking and creating awareness of sexual assault in society.
Figure 84: Sexual Assault Kernel Density

Figure 85: Sexual Assault Temporal Analysis
5.2.8 Body Found

There were 25 Tweets collected during the study period filtered using the keyword phrase “body found.” The density shows that most of the Tweets occurred near the White House and United States Capital area but also in northwest DC near the Forest Hills park area (as shown in figure 87). Analysis of the temporal graph, word cloud and Tweets showed that most of the Tweets were from news organizations alerting the public to crimes in the area (as shown in figure’s 88 and 89).
Figure 87: Body Found Kernel Density

Figure 88: Body Found Temporal Analysis
5.3 Temporal Analysis of Crime and Tweets

We used the temporal analysis that we generated from the keywords and phrases from the Tweets and conducted a comparison of the temporal analysis of the different types of crimes (as shown in figure’s 90 – 98). Temporal analysis allows us to explore the variation of crime activities as it relates Twitter communication volume. In community policing strategies, a two-way communication must be opened between law enforcement and the community. As we have already established law enforcement agencies use social media to engage the community as a communication tool for aiding in criminal investigations as well as a platform for disseminating information.

However, the community also uses social media platforms to disseminate information to increase awareness, and promote public campaigns. For example, during
our collection period a social media phenomenon known as the ice bucket challenge occurred. This challenge used social media to promote awareness and raise money for Amyotrophic Lateral Sclerosis (ALS) also known as Lou Gehrig's disease. ALS is a progressive neurodegenerative disease that affects nerve cells in the brain and the spinal cord. Through this social media community awareness program, the Ice Bucket Challenge raised over $115 million dollars since July 2014 (ALS Association, 2014).

Our goal for this section was to determine if Twitter chatter about specific crimes or community awareness campaigns from the public and law enforcement has an effect on crime rates.

Although temporal analysis was inconclusive with many of the crimes we did see evidence that Tweets about rape and sexual assault may lead to a temporary decrease in reported sexual assaults in Washington DC. Many of the spikes in Tweets were followed by a temporary decrease in sexual assault crimes (as shown in figure 95). This temporary decrease leads use to conclude that the community raising awareness coinciding with law enforcement informing the community about sexual related crimes has a direct effect on crime.

With law enforcement informing the community about sexual assault crimes coupled with the public raising awareness about sexual assaults, it may have led to a reduction in sexual assaults crimes. As previously mentioned, during our collection period a bill introduced in the Senate looking to crack down on sexual assault on college campuses (Kingkade & Fang, 2014). Through community awareness campaigns like the
ALS Ice Bucket Challenge, social media has the power to influence community involvement and possibly affect public policy.

Figure 90: Comparison of Tweets and Reported Crime for Assault with a Deadly Weapon

Figure 91: Comparison of Tweets and Reported Crime for Burglary
Figure 92: Comparison of Tweets and Reported Crime for All Keyword Tweets Related to Homicide

Figure 93: Comparison of Tweets and Reported Crime for Phrase Tweets Related to Homicide
Figure 94: Comparison of Tweets and Reported Crime for Robbery

Figure 95: Comparison of Tweets and Reported Crime for All Tweets Related to Sex Abuse
Figure 96: Comparison of Tweets and Reported Crime for Stolen Automobiles

Figure 97: Comparison of Tweets and Reported Crime for Theft from an Automobile
Finally our study compared the kernel densities of crimes by type and Tweets using keywords and phrases that occurred in Washington DC in the summer of 2014. We compared the densities of burglaries, robberies, rape and sexual assault. We chose these densities based on the variations that we found in the Tweets. For example, inspection of the Tweets showed that shots, had multiple connotations. The World Cup was played over the summer and many of the Tweets mentioned “take a shot” or in relation to night life, “who wants to take a shot.”

The density overlaps were created using ESRI’s ArcGIS 10.1. The process we used was to first reclassifying the kernel densities into five classes. We did this to account for the differences in counts the densities would encounter without reclassification. For
example, the burglary crimes hot spots represented 42 burglaries and Twitter represented 12. We then used a raster calculator and added the densities together and re-symbolized the results.

The results show that areas where the densities overlap are the areas where specific crime type and related Twitter chatter were identified (as shown in figure’s 99 – 102). Areas where the densities did not overlap would not be displayed.

5.4.1 Burglary

Figure 99: Geospatial Comparison of Crime and Twitter Related to Burglary Densities
5.4.2 Robbery

Figure 100: Geospatial Comparison of Crime and Twitter Related to Robbery Densities
5.4.3 Sexual Assault

Figure 101: Geospatial Comparison of Crime and Twitter Related to Sexual Assault Densities
5.4.4 Rape

5.5 Discussion

Geospatial analysis of the kernel densities of burglaries, robberies, rape and sexual assault revealed that areas where specific crimes are committed are also areas communicated in Twitter where specific crime occur. This leads us to believe that specific areas in the community (yet not all areas) may be engaged in efforts to combat specific crimes through law enforcement public engagement or community awareness programs. For example, one of the areas in Washington DC that had a high sexual assault
crime rate was the southern tip of the city. This area was also one of the areas where the Twitter densities were also high (as shown in figure 101).

Even though we did not use the 2010 United States Census data in this study this area has been identified by the 2010 census as a “black neighborhood” (United States Census Bureau, 2014). This study choose not to use the 2010 census data based on the population variations unique to the city from high tourism and outside commuter employment into the District (Destination DC, 2014) (United States Census Bureau, 2010). The southern tip of Washington DC is not a tourist destination. Yet the 2012 Pew Research Center’s Internet & American Life Project revealed that 28% of Twitter participants where African American.

According our Twitter collected data and the 2014 Washington DC Data Catalog, the southern tip of the District had a moderate amount of overall Twitter chatter (as shown in figure 8) and a moderate amount of overall reported crimes (as shown in figure 11) (Data Catalog, 2014). Based on the variations in population the city experiences, this area of the city may act to normalize the data.

This leads us to believe that specific areas within a large urban environment may have a higher susceptibility to community policing efforts. Since the southern area of the District was not subjected to tourism and high volume of commuter inflation.

5.5.1 Normalization of Data

Our data showed promise that areas where crimes occurred, are also areas where people were possibly tweeting about criminal activity or promoting social awareness programs. However, many of our Twitter densities hot spots were focused around the
White House, National Mall and United States Capital. This may be due to a number of factors, such as Tweets from politicians, political aids, or news agencies.

To further study crime and Twitter, we combined all the reported theft crimes that occurred in June, July and August 2014. These crimes included; theft, robbery, burglary, motor theft, theft from a vehicle, and other theft. As previously mentioned, theft crimes accounted for 92% of the crime in Washington DC (Data Catalog, 2014). Theft crimes also had a higher rate for repeat victimization compared to homicide, arson and assault. We created a density of the all the theft crimes (as shown in figure 103).

Figure 103: Combined Crimes Related to Theft Kernel Density
The theft crimes were then normalized by dividing the theft density (as shown in figure 103) by the density we created showing all the Tweets from the summer of 2014 in Washington DC (as shown in figure 8). Our results show that areas where crimes occurred are also areas where people are using precisely geolocated Tweets (as shown in figure 104).

![Normalized "Theft Crimes Density" vs "Twitter Density"](image)

**Figure 104: Combined Theft Crimes vs Normalized Using Twitter Density**

Additionally we then combined the normalized density data with the ground truth “all thefts” crime density data. We used the same process we used previously by reclassifying each kernel density into five classes and then adding the densities together.
by means of the raster calculator (as shown in figure 105). Our results further illustrated that high crimes rate areas are areas where there is high amounts precisely geolocated Tweets.

![Normalized Ground Truth Density Overlaps](image)

**Figure 105: Normalized Ground Truth Density Overlaps**

Although our study mined the text by filtering keywords and phrases, by normalizing the “all theft” and Twitter density data we can show that areas where crimes occurred are also areas where people are using precisely geolocated Twitter. Since our study mined specific keywords and phrases, we could not account for the surplus of additional keywords and phrases people may use to express criminal activity or communicate thoughts and feeling about crime and law enforcement. Through
normalization of the data we can get a better understanding of crime and the correlation of crime statistics.
CHAPTER SIX: CONCLUSION AND OUTLOOK

The use of social media by the law enforcement as well as the public is a device that can be used for community policing to reduce crime as well as change public policy. Our results support the premise that social media, is a viable platform for law enforcement and the public to raise awareness about criminal activity in their communities. Social media allows for a broader decimation of information from law enforcement agencies to the public and creates a virtual environment for the public to express their views about the police and/or crime. Our study demonstrated that Twitter, can be used a as a citizen-driven tool for empowering community policing activities, and supporting intelligence-led policing.

As first demonstrated in 1829 by Sir Robert Peel who established the London Metropolitan Police District, the western world’s first known Police district, community policing will continue to be a key strategy used by law enforcement agencies. With the use of smartphone technology expected to increase coupled with the increase of user participation using social media it is imperative that law enforcement agencies engage with the community utilizing social media.

In our thesis we were able to demonstrate how law enforcement agencies have been communicating with the public by posting information and updates, alerting the community to matters of public safety. We also demonstrated how law enforcement has used VGI and crowdsourcing techniques to engage the public to aid in solving crimes, indicating a possible paradigm shift in how police utilize the community for crime
enforcement. Through social media, law enforcement agencies are able to leverage the power of employing hundreds to thousands of volunteers and sensors to aid in identifying criminals and solve crime.

We also demonstrated that by filtering the text in Twitter using keywords and phrases we were able to demonstrate that not only are the law enforcement agencies using social media to engage with the community, the community itself is talking about specific crimes and social issues that are affecting their daily lives.

The Michael Brown fatal shooting in Ferguson, Missouri that occurred on August 9, 2014 is a prime example. Our analysis showed that five of the ten keywords and three of the eight phrases we identified a spike in chatter in regards to shooting. In each instance we saw an elevation in Twitter chatter from the community surrounding the keywords and phrases; murder, kill, robbery, shot, arrested, police officer, fatal shooting and arrested for. Many of the Tweets were outraged by what transpired, but law enforcement agencies must be aware of what social media is portraying the police as professional. Chatter on Twitter surrounding high profile crimes can indicate an overall community feeling, yet law enforcement agencies must embrace what is talked about on social media and possibly alter policing strategies.

Our thesis revealed that Twitter chatter on social issues likely has an effect on crime and public policy. Upon examination of the keywords and phrases, rape, assault and sexual assault, we perceived continual chatter to raise awareness to the community in regards to sexual assault, specifically on college campuses. Our results showed the hot spots where the crimes regarding sexual assault occurred also had areas where high
Tweets talking about the same subject. This may indicate that localized areas where sexual assaults are occurring are also areas where chatter in social media is also prevalent. This leads us to conclude that communities may respond to the sexual assault crimes by creating awareness through social media that may temporally have an effect on actual crimes.

Social media presents a paradigm shift in community policing practices. Law enforcement must use social media platforms to engage their communities. The community must also use social media platforms to engage with law enforcement and the public to promote awareness about crime.

6.1 Future Work

Our thesis took a macro approach to looking all crimes that occurred in Washington DC and compared it to specific keywords and phrases that we previously identified. Our crime density maps showed that the specific crimes; assault with a deadly weapon, burglary, robbery and theft had little to no change over time geospatially, indicating the areas are prevalent to repeat victimization. Yet we were able to show that the following crimes; arson, homicide, motor theft, sexual assault and theft from an automobile did change over time geospatially. We would like to investigate this phenomenon further.

We would also like to take a micro approach to Washington DCs Districts 6 and 7. Both districts are not subjected to the same high amounts of tourism or large increases in population from outside the District commuting.
This thesis also showed there is potential to using Twitter to filter keywords and phrases to identify possible crime hot spots. We would be interested to see how applying the same techniques used in this study to a different study area, like Santa Clara, California. Santa Clara located in the center of Silicon Valley and is home to the headquarters of several high-tech companies, including PredPol.

Finally we would like to study the effects of social media and the seasonal crime phenomena.
REFERENCES


139


BIOGRAPHY

Kevin Marc Glodava is proud to be a Colorado Boy who graduated from Arvada West High School, Arvada, Colorado in 1994. He received his Bachelor of Arts in Human Performance and Sports and Exercise Science from Metropolitan State University of Denver in 2001 and later received a Bachelor of Science in Criminal Justice and Criminology in 2008 also from Metropolitan State University of Denver. He is a Commissioned Intelligence Officer in the United States Naval Reserve and served on active duty as an enlisted Intelligence Specialist conducting on-scene forensic investigations of improvised explosive devices (IEDs) in Iraq where he was awarded the Bronze Star Medal. Mr. Glodava is also a world class athlete, he competed on Team USA’s World Championship Duathlon (run-bike-run) Team in 2007 and again in 2008.