

MINING SPATIAL ASPECTS OF CYBERPHYSICAL COMMUNITIES

by

Xu Lu
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Science

Committee:

_____	Dr. Anthony Stefanidis, Dissertation Director
_____	Dr. Peggy Agouris, Committee Member
_____	Dr. Arie Croitoru, Committee Member
_____	Dr. Andrew Crooks, Committee Member
_____	Dr. Anthony Stefanidis, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Interim Dean, College of Science
Date: _____	Fall Semester 2014 George Mason University Fairfax, VA

Mining Spatial Aspects of Cyberphysical Communities

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

By

Xu Lu
Master of Science
George Mason University, 2012
Bachelor of Science
Peking University, 2008

Director: Anthony Stefanidis, Professor
Department of Geography and Geoinformation Science

Fall Semester 2014
George Mason University
Fairfax, VA

Copyright: Xu Lu, 2014
All Rights Reserved

Table of Contents

List of Tables	v
List of Figures	vi
Abstract	viii
Chapter 1. Introduction	1
1.1 Social Media and GeoSocial Networks	4
1.2 Communities in Cyber and Physical Spaces	6
1.3 Problems and Objectives.....	8
Chapter 2. Literature Review	11
2.1 Physical Community and Cyber Community	11
2.1.1 The Concept of Physical Community	11
2.1.2 The Concept of Cyber Community	12
2.1.3 Definition of “Community” in this Dissertation	13
2.2 Geospatial Analysis on Physical Community	14
2.3 Geospatial Analysis on Cyber Community.....	15
2.4 Spatial Factor on Physical and Cyber Spaces	18
2.4.1 Distance in Physical Space	19
2.4.2 Distance in Networks	19
2.4.3 Distance in Other Fields.....	20
2.4.4 Distance in Cyberphysical Communities	21
Chapter 3. Corresponding Communities in Cyber and Physical Space.....	25
3.1 Datasets	25
3.1.1 Physical Space Community: Alumni Dataset	26
3.1.2 Cyber Space Community: Twitter Dataset	27
3.2 Level of Spatial Aggregation	29
3.3 Distance from Point of Interest	35
3.4 Goodness of Fit	38
3.5 Temporal Analysis and Event Detection	41
Chapter 4. Cyber Community in Physical Space.....	45

4.1	Retweet Network	46
4.2	Retweet Network in Physical Space	47
4.2.1	Retweet Distance of Mason Network	53
4.2.2	Retweet Distance for Each State	56
4.3	State Network.....	60
4.3.1	State Network Structure	61
4.3.2	Originality	63
4.3.3	Self-retweet Percentage.....	66
4.4	Random Location Network.....	68
4.5	Another Case Study	69
Chapter 5.	Physical Community in Physical Space	72
5.1	Location Based Social Network Dataset.....	72
5.1.1	Check-ins.....	73
5.1.2	Friendship.....	74
5.1.3	Home Location of Each User	76
5.1.4	Human Mobility	79
5.2	Friendship Distance	83
5.3	Travel Distance	87
5.4	City Distance.....	90
Chapter 6.	Conclusion.....	93
References	99

List of Tables

Table 1. Individual, friendship and community of physical and cyber society in physical space	7
Table 2. Top 10 most active zip codes in the dataset.....	33
Table 3. Degree statistics for retweet network.....	50
Table 4. Statistics of retweet distance in Mason network (kilometers)	55
Table 5. Top 10 states in centrality	63
Table 6. Top 10 states in originality	66
Table 7. Top 10 states in self-retweet percentage.....	67
Table 8. Mean and median of human travel mobility	83
Table 9. Distance of different communities (kilometers)	96

List of Figures

Figure 1. Online content generated per minute as of 2013	6
Figure 2. Distribution of GMU alumni address in the United States.....	27
Figure 3. Spatial distribution of tweets discussing GMU in the period 8/12 – 7/13. 28	
Figure 4. A log-log plot of alumni (horizontal axis) versus Twitter traffic (vertical axis) spatially aggregated at the zip code level.....	30
Figure 5. A log-log plot of alumni (horizontal axis) versus Twitter traffic (vertical axis) spatially aggregated at the state level.....	30
Figure 6. A log-log plot of alumni (horizontal axis) versus Twitter traffic (vertical axis) aggregated at the zip code level, using only tweets with precise coordinates	34
Figure 7. Upper: Alumni number (vertical axis) versus distance from Fairfax (horizontal axis). Lower: Number of geolocated tweets (vertical axis) as a function of distance from Fairfax (horizontal axis)	36
Figure 8. A log-log graph of tweets per alumnus (vertical axis) as a function of distance from Fairfax (horizontal axis).....	38
Figure 9. Tweets per alumnus by region.....	39
Figure 10. Daily volume of tweets traffic.....	41
Figure 11. Word clouds of Oct 5, 2012 (left) and Mar 25, 2013 (right).....	42
Figure 12. Sentiment analysis for all tweets	43
Figure 13. Events identified by tweets volume and sentiment	44
Figure 14. Tweet network, color represents: connected component (left); degree(right).....	47
Figure 15. Mason retweet network	48
Figure 16. Mason retweet network in physical space	49
Figure 18. Top three largest groups detected in the network.....	52
Figure 19. Histogram of retweet distance in kilometers, upper: 0~8000 kilometers, bin is 100 kilometers; lower: 0~4000 kilometers, bin is 50 kilometers.....	54
Figure 20. Histogram of retweet distance in the network (left) and the one without 22030 (right)	56
Figure 21. Retweet distance in each state: (a) minimum (b) mean (c) median (d) maximum	58

Figure 22. Maximum and minimum retweet distance for each state	60
Figure 23. Retweet network of states in Fruchterman-Reingold layout	61
Figure 24. Original posts vs. retweets for each state (a) every state included, (b) remove VA, (c) logarithmic axis, (d) labeled with states	65
Figure 25. Histogram of retweet distance in Mason retweet network (left) and in random location network (right), bin is 200 kilometers	68
Figure 26. Retweet distance parameters following the Boston bombing event of April 2013 (during the first 13 hours after the event)	70
Figure 27. Retweet distance of Mason network (left) and Boston network (right) ..	71
Figure 28. Histogram of all users' check-in data	73
Figure 29. Percentage of all users with different check-in times	74
Figure 30. Histogram of numbers of friends for all users	75
Figure 31. Percentage of all users with different number of friends	76
Figure 32. Check-in data of one user	77
Figure 33. Left: hierarchal tree for user ID=0; Right: clusters (only check-ins in the US)	78
Figure 34. Active users' home locations	79
Figure 35. Count and percentage of users with different travel percentage	81
Figure 36. Histogram of short-distance travel and long-distance travel for users with different travel percentages	82
Figure 37. User ID=0 and his/her friends	84
Figure 38. Histogram of friendship distance (bin is 200 kilometers)	84
Figure 39. Histogram of friend distance in the interval of 0~100 kilometers (bin is 10 kilometers)	86
Figure 40. Histogram of friend distance in the interval of 100~500 kilometers (bin is 10 kilometers)	87
Figure 41. Histogram of check-in data for users in the United States (bin is 50 kilometers)	88
Figure 42. Histogram of travel distances for all users in the United States (bin is 200 kilometers)	90
Figure 43. Major cities in the United States with the population greater than 100,000	91
Figure 44. Histogram of distance between major cities (bin is 200 kilometers)	92
Figure 45. Comparison of distance histogram for cyber communities (left) and physical communities (right), bin is 200 kilometers	94
Figure 46. Comparison of distance in different communities	97

Abstract

MINING SPATIAL ASPECTS OF CYBERPHYSICAL COMMUNITIES

Xu Lu, PhD

George Mason University, 2014

Dissertation Director: Dr. Anthony Stefanidis

This dissertation studies the newfound concept of cyberphysical communities, focusing in particular on their manifestation through social media activities, and on the role of distance in communities formed in physical and cyber spaces. We use the term *cyberphysical* in the context of this dissertation to refer to communities that comprise physical members (i.e. individuals) whose communications are observed in the cyber space (and in particular through social media). The main objective of this dissertation is to investigate whether the well-accepted Tobler's First Law in the physical space is also applicable to these cyberphysical communities. Simply put, Tobler's First Law states that while everything is related to everything else on the Earth's surface, near things are more related than distant things.

While this has long been an accepted concept for activities and processes manifesting themselves in the physical space, newfound capabilities to interact through cyber space have introduced the potential to interact and form communities regardless of the cost of physical distance. Cyber interactions offer the potential to bypass the physical distance, and allow people to interact with anyone, anywhere in the world. The question still remains: does physical distance limit or drive the spatial distribution of cyberphysical communities, making them subject to Tobler's Law? While there have been some partial studies of issues related to this question, there is still a lack of a thorough study of the spatial characteristics of cyberphysical communities, and this is the gap that this dissertation is attempting to bridge. By comparing the distance distributions for three cyber communities and three physical communities, we show how distance plays an important role in cyberphysical communities.

Chapter 1. Introduction

Fostered by Web 2.0 and corresponding technological advancements, social media and social networks have become massively popular during the last decade. An increasingly sizeable portion of such content is geolocated, which enables new types of data mining processes for a variety of applications. The rapidly emerging and quickly prevailing social media and resulting networking activities present us the opportunity to rethink and redefine the social concept of *community*, and further the concept of the *human landscape*. In the context of this dissertation we use the term human landscape to refer to the spatial distribution of human population characteristics. These novel cyber-interaction systems provide us the ability to access massive amounts of streaming data, whose volumes, variety, and rates are comparable to the standards used to characterize big data applications (Vatsavai *et al.*, 2012; Croitoru *et al.*, 2014a). These data provide information about a variety of human activities, which can be mined to derive human landscape content at spatial and temporal resolution unheard up to now.

However, the analysis of such data also bring forth substantial challenges, for example, how to transform social media observations into real world knowledge, or assessing the

extent to which the cyber communities can serve as a proxy to corresponding physical communities. This dissertation makes a contribution to science by addressing such challenges, aiming to advance our ability to understand social dynamics from the perspective of community, especially focusing on the spatial distribution of such communities, and the factors that bring them together or keep them apart.

This dissertation studies the emerging *cyberphysical communities*, focusing in particular on their manifestation through social media activities, and on the role of physical distance on the formation of such communities across the physical and cyber spaces. We use the term cyberphysical in the context of this dissertation to refer to communities that comprise members from the physical space (i.e. individuals) whose communications are observed in cyber space (and in particular through social media). A main premise of this dissertation is the investigation of the Toblerian nature of such cyberphysical communities, i.e. assessing whether Tobler's First Law of geography is also applicable to these cyberphysical communities. Simply put, Tobler's First Law states that while everything is related to everything else on the Earth's surface, near things are more related than distant things (Tobler, 1970). While this has long been an accepted concept for activities and processes manifesting themselves in the physical space, interaction through cyber space is not thus limited: one can interact and form communities with any users, regardless of physical proximity. But while one *can*, does one *do*?

While at their early onset these newfound capabilities gave rise to critical views that led even to extreme statements pronouncing the death of distance^{1,2} in that context, these views regarding the cyber space's power to vanquish geography have been widely retracted³ recently. The question still remains: does physical proximity limit or drive the spatial distribution of cyberphysical communities, and are these communities Toblerian in their nature? While some relevant issues have been partially studied in the past, there is still a lack of a thorough study of the spatial characteristics of corresponding communities in the cyber and physical spaces, and this is the gap that this dissertation is attempting to bridge. More specifically, this dissertation aims to identify the role of physical distance in cyberphysical spaces from the perspective of communities, as they are formed through cyber interactions. As part of this effort, we will be comparing the spatial characteristics of cyberphysical communities to corresponding physical communities.

The potential practical benefits from this challenging scientific topic are manifold, spanning the whole range from improving our response to natural disasters to enhancing marketing and information dissemination operations. By providing a geospatial approach to this problem, we will also advance human geography by gaining additional

¹ <http://hbswk.hbs.edu/archive/2234.html>

² <http://www.economist.com/node/598895>

³ <http://politi.co/1qBRXcJ>

understanding of the spatial characteristics of communities, and the factors that bring them together or keep them apart in today's era of cyber-communications.

1.1 Social Media and GeoSocial Networks

Since its emergence, social media has been altering human interactions by providing a novel platform for community formation. As Web 2.0 altered the scope of the Internet, evolving it from an information access platform to an information dissemination and human interaction environment, community formation has expanded beyond its traditional physical and technological limitations. While people had always been taking advantage of technology to interact and communicate beyond the limits that physical distance would impose (e.g. using networks to transmit thought-provoking philosophical words in the renaissance, or the telegraph and telephone over the past two centuries, and email more recently) the emergence of social media represents a watershed moment in community formation. For the first time people can form instantaneously globe-spanning communities focused on a wide variety of thematic topics (see e.g. Stefanidis *et al.*, 2013a; Crooks *et al.*, 2014), or act as hybrid sociocultural sensors and contribute news and views during natural or man-made disasters (Goodchild, 2007; Stefanidis *et al.*, 2013b).

Through social media, people not only *post information* online, but they *act socially*: they interact with each other, update their status, share opinions, and form social networks.

Participation in such activities has moved well beyond being the niche of a technologically versed minority to become a global phenomenon. Facebook, for example, reached 1.3 billion users at the time that this dissertation is written (July 2014)⁴, to tie China as the world's most populous community.

At the same time, Location-Based Social Networks (LBSNs), such as Foursquare, Gowalla and Brightkite, are also becoming popular. In LBSN, people share their current location with a broader on-line social network by “checking in” at various locations. But even beyond LBSNs, numerous social media platforms support the publishing of geotagged content, so that people can upload their pictures (Flickr), express their opinions (Twitter), post their updates (Facebook) with a location tag (Valli and Hannay, 2010). Accordingly, social media content is becoming increasingly geospatial in nature, and mining such data is the subject of geospatial analysis. Furthermore, as these nuggets of information are communicated at unparalleled rates, with massive amounts of digital data generated and accessed every single second (Figure 1), we are encountering a big geosocial data paradigm (Croitoru *et al.*, 2014a), that is bringing forth both opportunities and challenges.

⁴ <http://www.insidefacebook.com/>

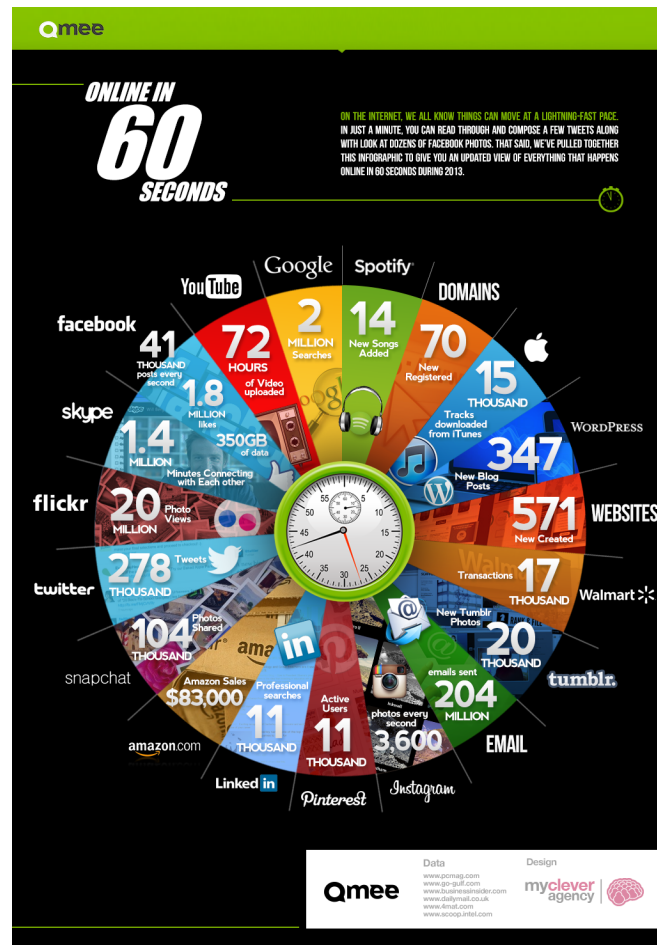


Figure 1. Online content generated per minute as of 2013⁵


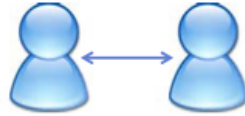

1.2 Communities in Cyber and Physical Spaces

As the Internet has become globally accessible, participation in it is leading to the emergence of a cyber culture and cyber societies (Deuze, 2006). These cyber societies are built upon the same structural elements as physical societies as shown in Table 1. Each person has physical identity (or identities), and interactions among individuals lead to the

⁵ Source: <http://blog.qmee.com/qmee-online-in-60-seconds/> last access: 10/16/2013

establishments of links and, potentially, friendships. As more people connect with each other they gradually form communities. Similar to physical societies, a person can have one or more identities in cyber space. Through online interaction, people can become cyber acquaintances, and form cyber communities.

Table 1. Individual, friendship and community of physical and cyber society in physical space

			
Physical Society	individual	friendship	community
Cyber Society	cyber identity	online interaction	cyber community
Physical Space	location	rendezvous	distribution

Besides sharing similar structural elements, physical and cyber communities also share one more issue that is of particular importance to this dissertation: they both interact *with* and *within* the physical space (Wellman, 2001; Herrera, 2007). Such communities are often formed around specific events that have a certain physical footprint, or extend their mode of interaction from cyber to physical. Conversely, in physical space, people function across locations, where they act and interact.

As physical and cyber communities are spread across the physical space, it is essential to gain a better understanding of the spatial particularities of corresponding cyber and physical communities that are formed around the same topic, and this is a focus of this dissertation. Gaining such an insight is an essential step towards advancing our understanding of how information is disseminated and used during marketing campaigns (Kaplan and Haenlein, 2010), or at moments of crisis (see e.g, Sutton *et al.*, 2008; Croitoru *et al*, 2014b) to gaining a better understanding to the threat posed to established societies by decentralized, leaderless, geographically dispersed groups of terrorists, subversives, extremists, and dissidents (Ressler, 2006). Therefore, it is critical to understand the community structures of the social network and to find out the crucial nodes in the network in order to prevent terrorist attacks and reinforce security.

1.3 Problems and Objectives

This dissertation studies the newfound concept of cyberphysical communities, focusing in particular on their manifestation through social media activities, and on the role of distance in communities formed in physical and cyber spaces. We use the term *cyberphysical* in the context of this dissertation to refer to communities that comprise physical members (i.e. individuals) whose communications are observed in the cyber space (and in particular through social media). The main objective of this dissertation is to investigate whether the well-accepted Tobler's First Law in the physical space is also

applicable to these cyberphysical communities. Simply put, Tobler's First Law states that while everything is related to everything else on the Earth's surface, near things are more related than distant things (Tobler, 1970). While this has long been an accepted concept for activities and processes manifesting themselves in the physical space, newfound capabilities to interact through cyber space have introduced the potential to interact and form communities regardless of the cost of physical distance. If physical and cyber communities are to follow Tobler's Law, their members would tend to gravitate towards their neighbors, leading to a propensity of local connections compared to distant, and to heavily localized membership clusters in the physical space. Cyber interactions offer the potential to bypass the physical distance, and allow people to interact with anyone, anywhere in the world (Cairncross, 1997; Ratti *et al.*, 2010). The question still remains: does physical distance limit or drive the spatial distribution of cyberphysical communities, making them subject to Tobler's Law? While there have been some partial studies of issues related to this question (Hecht and Moxley, 2009), there is still a lack of a thorough study of the spatial characteristics of cyberphysical communities, and this is the gap that this dissertation is attempting to bridge.

More specifically, this dissertation addresses the following research questions:

- What is the spatial distribution of interaction groups formed through social media?
- How to measure distance in cyberphysical space?

- What is the meaning of distance in cyber and physical communities?
- Does Tobler's First Law affect cyber and physical communities?

Solving these intriguing problems will facilitate the development of open source geospatial intelligence (Caverlee *et al.*, 2013; Stefanidis *et al.*, 2014) and will bring a unique understanding of the human landscape, its structure and organization, and its evolution over time.

Chapter 2. Literature Review

The literature review will cover the previous research on the geospatial analysis on physical and cyber communities, and the distance factor in cyberphysical spaces. Firstly we will review and discuss the definition of “community” in physical and cyber spaces. Secondly, we will review of geospatial analysis on physical community, mainly about human mobility patterns of group of people. After that, we will review the geospatial analysis on cyber community, which is basically about the Location-Based Social Network (LBSN). Next, we will review the previous research on distance in physical space, networks and many other fields. Finally, a restatement and discussion of research questions on distance of physical and cyber communities will be given.

2.1 Physical Community and Cyber Community

2.1.1 The Concept of Physical Community

The concept and definition of community have been discussed over many years, yet there is no single widely accepted or adopted definition. In fact, Hillery (1955) stated ninety-four difference definitions of community. Researchers tend to give their own

definition based on their primary research interests. Therefore, the term "community" refers to different things, depending upon who is using it and upon the context in which it is used (Nelson, Ramsey, Verner 1960).

At first, the concept of community was locally constrained, almost synonymous with the term neighborhood. Wellman and Leighton (1979) stated that the traditional definition of community is a spatially compact set of people with a high frequency of interaction, interconnections, and a sense of solidarity. Benedict Anderson (1983) stated about Imagined Communities: “the members of even the smallest nation will never know most of their fellow-members, meet them, or even hear of them, yet in the minds of each lives the image of their communion.”

As the rise of the Internet and development of information technology, the concept of community is also evolving. Social links between long distance relationships led scholars to extend their view on non-local communities. Also, the social psychological aspect of community was emphasized. Wellman (1979, 2001) stated that communities are based on sociable and supportive social relations, and not on physical locality. In this sense, Gruzd *et al.* (2011) stated community as “a set of people who share sociability, support, and a sense of identity.”

2.1.2 The Concept of Cyber Community

The concept of “virtual community” was raised, but its definition was as troublesome as

community itself. The term virtual community was considered by many to be synonymous with a class of group computer-mediated communication (Johns, 1997). Hagel and Armstrong (1997) defined virtual communities as computer-mediated space where there is an integration of content and communication with an emphasis on member-generated content. Some did not agree with the use of “virtual community” as it was completely different from traditionally defined community (Weinreich, 1997). Still, there are many researchers trying to give a definition of virtual communities: Smith (1992) defined virtual community as “a set of on-going many-sided interactions that occur predominantly in and through computers linked via telecommunications networks”. Erickson (1997) stated that virtual communities are "long term, computer-mediated conversations". Porter (2004) defined virtual community as an aggregation of individuals or business partners who interact around a shared interest, where the interaction is at least partially supported and/or mediated by technology and guided by some protocols or norms. He also delimited a typology for the virtual community based on this definition.

2.1.3 Definition of “Community” in this Dissertation

The term community has a dynamic meaning, and its definition is evolving over time. It is case-sensitive and dependent upon what context it is in. As we will focus on the characteristics of cyber and physical communities and the relationship between the two, we would emphasize on this aspect. Our definition of community is an aggregation of

individuals who *interact* with each other among at least one shared interest for a certain period of time. In our opinion, the necessary conditions of a community, according to our definition, are:

- (1) Members. They are the main subjects that form the community.
- (2) At least one shared interest among members. Community members may communicate on various topics, but there should be one topic/interest in common that brought them together.
- (3) Interactions continue for at least a certain period. If interactions among members only happens for very short period and then die out, the community is considered as dead as well.

2.2 Geospatial Analysis on Physical Community

When applying geospatial analysis on physical community, lots of researches have been done to explore human mobility patterns. Human mobility research observes, simulates and predicts people's movement patterns in the physical space. It is a branch of social science, and it plays an important role in human geography, especially in urban planning, traffic engineering, emergency evacuation and disaster response. Human mobility analysis can be conducted at multiple scales, large to the whole country or world and small to a single room. For country or even world scale, social scientists have been examining the spatial pattern of migration for years. For region or city level, human

mobility analysis plays a significant role in urban modeling and urban planning. For building or room level, we can extract implicit scene structures (Lu *et al.*, 2011) by analyzing human mobility for better utilizing the indoor space.

When considering the geospatial factors with the physical communities, most researchers aim to construct mobility models to simulate and predict human mobility patterns for group of people. Musolesi and Mascolo (2006) proposed a mobility model based on social network theory, which allows individuals to be grouped together based on social relationships. In their model, the grouping is mapped into a topographical space with movements influenced by the strength of social ties. Zignani *et al.* (2012) defined a concept called the “geo-community”, which merged both “location” and “community” together. They analyzed real GPS datasets of human mobility traces, quantitatively described geo-communities and inferred the probability distribution of all the features of human behavior. Wu and Wang (2013) studied human mobility among different types of communities and described the communities with the most probable activities. Through the analysis of digital footprint data, they found that there existed a strong correlation in daily activity patterns within the group of people who live in the same type of communities.

2.3 Geospatial Analysis on Cyber Community

When combining geospatial analysis with cyber community, there are large amounts of

researches dealing with Location Based Social Network (LBSN). Human mobility patterns have also been studied based on LBSN. Noulas *et al.* (2011) presented a large-scale study of user behavior on an LBSN (Foursquare). They analyzed user check-in dynamics over space and time, discussed how different behaviors succeeded each other. Their research demonstrated how LBSN could reveal meaningful spatial-temporal patterns and offer the opportunity to study both user mobility and urban spaces. Cho *et al.* (2011) studied the human mobility patterns through cell phone location data and two LBSN data. They found that “humans experience a combination of periodic movement that is geographically limited and seemingly random jumps correlated with their social networks.” Periodic behavior can explain 50%~70% of all human movements, while social relations can explain 10%~30%. They also found that the friendship probabilities declined as the distance between two people increased, though there were some interesting minor pits. They even developed a human mobility model based on their findings, which can predict location and dynamics of human movements.

Recently, Roick and Heuser (2013) reviewed the current research activities of LBSN. In this review paper, they presented a comprehensive definition of LBSN, provided an overview of research activities, and also concluded a research agenda from existing research directions. They pointed out that only a few LBSN have been studied thoroughly; the exploitation of geographic information and information retrieval from LBSN has just

started. Few studies have been conducted in applying LBSN data with natural hazard response and monitoring diseases.

Other research activities focus on how to infer social relationship and detect events based on geospatial locations of people (Ritterman *et al.*, 2009; Culotta, 2010; Sugumaran and Voss, 2012). Many researches have been done on the topic of friendship prediction and recommendation on LBSN. Crandall *et al.* (2010) investigated the extent to which social ties between people can be inferred from co-occurrence in time and space. They developed a probabilistic model and applied it to data from a social media site (Flickr). They found that even a very small number of co-occurrences could result in a high empirical likelihood of a social tie. Cranshaw *et al.* (2010) improved the friendship prediction model as they did not only consider the number of co-locations, but also took other properties into consideration. They examined the location traces of 489 users of location sharing social network for relationships between the users' mobility patterns and structural properties of their underlying social network, and they developed 67 different features to describe the social network and location properties. Using these features, they provided a model for predicting friendship between two users by analyzing their location trails. Their work showed a potential future direction: the relationship between online and offline social behavior. Sadilek *et al.* (2012) explored the interplay between people's location, interactions, and their social ties within a large real-world dataset. They

proposed a system to predict friendship based on friendship formation patterns, content of messages and user locations. Their system could also predict user's location based on a scalable probabilistic model of human mobility.

There have also been lots of researches focusing on detecting cyber communities using geospatial location information. Qi *et al.* (2013) designed online algorithms for online community detection using location data collected in real time through social sensing mobile applications. Gennip *et al.* (2013) proposed a spectral clustering model to identify social communities among gang members in Los Angeles based on social interactions and geographic locations of the individuals. They discussed different ways of encoding the geosocial information using graph structures and corresponding influence on the clustering results. Hannigan *et al.* (2013) argued that current community detection approaches often ignore the spatial location of people; thus they proposed a novel method to detect spatially-near communities. They introduced a new metric to measure the quality of a community partition in a geolocated social network, called "spatially-near modularity". Based on their method, the ideal community will be partitioned with respect to both social ties and geographic location.

2.4 Spatial Factor on Physical and Cyber Spaces

The key spatial factor in this dissertation is distance. The topic of distance has been studied for many years by scholars from different backgrounds, whereas nowadays there

are still many research papers dealing with distance. In this section, the distance measurement in physical space, networks and other fields will be reviewed. In the end, the dissertation topic on distance in cyberphysical communities will be briefly introduced and discussed.

2.4.1 Distance in Physical Space

Distance in geography is traditionally viewed as a geometric concept. The most common measurement of distance is Euclidean distance. For two points on the Euclidean plane $(x_1, y_1), (x_2, y_2)$, the Euclidean distance d_E is defined as:

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Another measurement of distance is called Manhattan distance, also known as city block distance. It is measured by the sum of the absolute differences of two points' coordinates. For two points on the Euclidean plane $(x_1, y_1), (x_2, y_2)$, the Manhattan distance d_M is defined as:

$$d_M = |x_1 - x_2| + |y_1 - y_2|$$

There are some other measurements of distance that have been used in other scenarios, such as Chebyshev distance, Mahalanobis distance, etc.

2.4.2 Distance in Networks

Distance in networks is traditionally defined in terms of network connectivity, path

lengths or node centrality, etc. (Newman, 2009). In graph theory, distance of two vertices is defined as the number of edges in a shortest path connecting them. Similarly in network analysis, distance of two nodes is defined as the number of minimum steps from one node to another.

2.4.3 Distance in Other Fields

Besides the mathematical definition of distance, there are many kinds of distance measurements that have extensive practical applications. For example, the weighted distance (cost distance) is widely used to estimate the “cost” from an origination to a destination. The weight (cost) might be complicated and can be calculated in multiple ways, such as spatial length, traveling time, toll cost, etc. It is very useful in transportation management, travel planning, logistic routing and supply chains.

Not only do geographers care about distance, scholars in other science fields also study on it. In business and economy, the cross-national distance is crucial to company and industry. Ghemawat *et al.* (2001) argued that we should consider not only geographic distance, but also three other dimensions of distance: cultural distance, administrative or political distance, and economic distance. Based on these four dimensions of distance, they performed a case study on expanding market for a company. Berry *et al.* (2010) extended the dimensions of distance into nine dimensions, which were economic, financial, political, administrative, cultural, demographic, knowledge, global

connectedness and geographic distance. They provided a method to measure such distances and discussed their impacts on business in detail. There is also a branch of socio-psychology called proxemics, to study the relationship between spatial distance and social relationship of individuals. For example, Matthews and Matlock (2011) conducted three experiments on spatial distance and social relationships, and they found that people tends to get close with friends and try to avoid strangers, which implies that social distance and spatial distance are conceptually linked. Such research shows how important distance is, and it is very inspiring to study the role of distance in cyberphysical communities.

2.4.4 Distance in Cyberphysical Communities

Although there have been lots of studies on distance from different aspects, only a few studied distance from the perspective of community in cyberphysical space. The findings of previous research indicated that spatial proximity increase the probability that two individuals establish an online connection, and geographic closeness could influence the formation of cyber communities (Brown *et al.*, 2012).

Kaltenbrunner *et al.* (2012) studied the social interactions and geographic locations on users from an LBSN, and they found that the geographic distance strongly affects how social links are created, however, social interactions are only weakly affected by distance. Brown *et al.* (2012) studied the spatial and social properties of two online communities

with location-sharing features. They found that community structure in social networks may arise from both social and spatial factors. Therefore, it is beneficial to extract the information about places where people go to facilitate new community detection methods and community evolution models. They further studied the structural properties of an LBSN (Brown *et al.*, 2013). They reconstructed the same LBSN into social community and spatial community (local-focused community), and they discovered that these two communities had very different graph structural properties. They found that local-focused communities could be more valuable for friend suggestion and place recommendation. As for the community evolution, social communities are more stable while local-focused communities can be very transient or very stable.

Researches on social and spatial distance of communities have many practical applications. For instance, Rothenberg *et al.* (2005) studied the spatial and social distance of an urban group of people in Colorado Spring, Colorado, of their risk for the sexually transmitted disease, HIV. They found that the HIV-positive subgroup people are tightly spatial clustered, and they are all within six steps away in the social network. They also found that social connections reflect geographic proximity, which could be a crucial element in the dynamics of disease transmission among groups. Tayebi *et al.* (2012) examined the relation of spatial and social distances in a crime network in Los Angeles. They found a strong correlation between social and spatial distance of crime offenders,

which indicated that offenders who are socially close are also spatially close. As shown above, such research is very helpful in understanding how disease transmits through group, and what crime network structures are in the physical world. Despite the importance of linking spatial and social distance in communities, the amount of such research is very limited.

With the process of globalization, it is very easy and common for people to connect with other entities from anywhere of the world, whereas people still live locally and need local resources. In fact, geographic proximity might be more important with the Internet revolution, according to Goldenberg and Levy (2009). To explain such phenomenon, Robertson (1995) popularized an idea called “Glocalization”, for which he explained as people think globally while act locally. Therefore, people may “exist” in multiple times and places: physically in one time and one place, virtually link to many other places and time zones (Mok *et al.*, 2010).

When considering the cluster of people – groups and communities, we expect similar patterns. Although communities might be associated with a particular location (e.g. headquarter, populated place, place of interest), they are virtually connected to many other places by their social links. They have a level of affinity between them. So we can compare communities in terms of discussion topics, keywords etc., and find affinity distance metrics that express, for example, New York City, NY is closer to Los Angeles,

CA than Austin, TX. This can be used to compare distant communities (e.g. an affinity distance between NY and LA) and also to identify the dispersion of a single community (as a metrics of homogeneity). There are many challenges when dealing with this issue, such as how to compare two places through different interested communities, how to link communities in cyberphysical spaces, how to define and calculate affinity distance metrics. This dissertation will shed lights on such problems.

Chapter 3. Corresponding Communities in Cyber and Physical Space

In this section, we will focus on the spatial factors of corresponding communities in cyber and physical spaces. Our objective is to identify a community in the physical space and compare it to its closest cyberphysical counter-part. Firstly, we will introduce our datasets for the communities. Secondly, we will discuss the correlation between the corresponding communities and explain how level of spatial aggregation affects the correlation. After that, we will examine the distance from point of interest and the participation index. Then, we will test the hypothesis for goodness of fit for the two communities to see whether they follow a certain distribution. Finally, we will perform temporal analysis on the cyberphysical community dataset and show how we can use these data to detect interest events.

3.1 Datasets

In order to select a case study that is both sufficiently large and spatially distributed to support meaningful analysis, and sufficiently distinct so that it can be identified in social

media content, we chose the community of interest formed around a large higher education institution, namely George Mason University (GMU).

Furthermore, as studies have indicated that there exists a strong relationship between the alumni community and the branding presence of a university (McAlexander *et al.*, 2006) we can reasonably argue that the alumni community in physical space can serve as a close counterpart to the cyber community for the study. Accordingly, we choose the alumni community of GMU as a representative sample of the spatial distribution of that community in physical space. As its cyberspace counterpart, we chose the online community formed in Twitter as part of a discussion about the same university. Given these datasets, the task is to compare these two communities in terms of their spatial patterns in order to gain insight on how observed patterns in geolocated social media compare to the corresponding community in physical space.

3.1.1 Physical Space Community: Alumni Dataset

In order to establish the spatial footprint of the physical community, we collect anonymized address data for GMU alumni. Accordingly, for the 154,140 alumni that have graduated from the university since 1961, we collect their current home address zip codes, which are spread across 9,822 different zip code areas throughout the United States. As the United States Census Bureau lists approximately 43,000 different zip codes in the United States, our alumni are spread across 22.8% of all these zip code areas. The

spatial distribution of these alumni is shown in Figure 2. The clusters of states in the background (shown in different colors) correspond to the ten Federal Standard Regions.

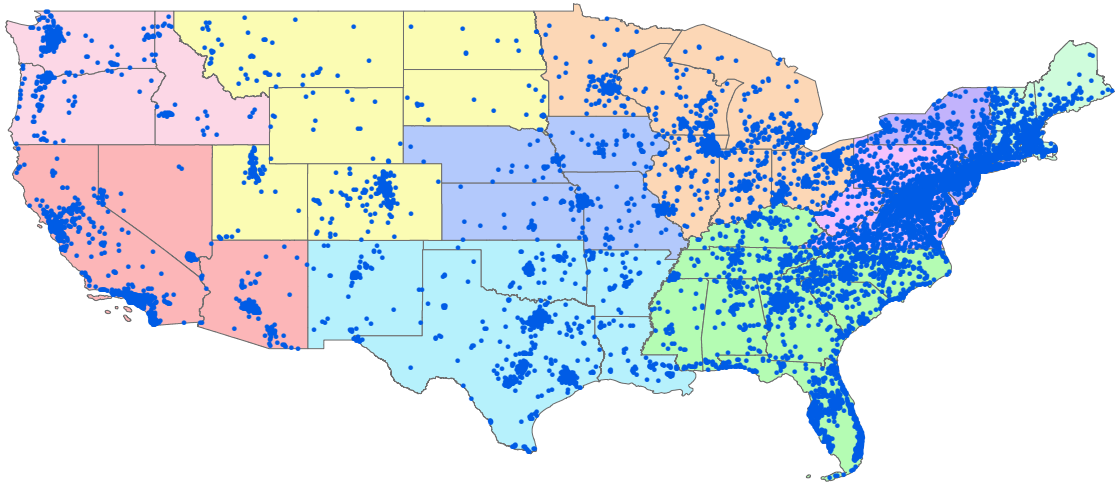


Figure 2. Distribution of GMU alumni address in the United States

3.1.2 Cyber Space Community: Twitter Dataset

The corresponding cyber community is generated by collecting tweets that discussing GMU over a period of eleven months (August 1st, 2012 to July 3rd 2013). The data are collected and processed using the GeoSocial Gauge system prototype (Croitoru *et al.*, 2012; 2013). By accessing Twitter's Application Programming Interface (API) and its

one percent streaming content, we collect 151,900 tweets discussing GMU during that period. From among them, 70,600 are geolocated within the United States. Location information is obtained either by harvesting tweets tagged with precise coordinates (as is usually the case when tweets are posted from a GPS-enabled mobile device) or by identifying descriptive toponyms in tweets, which can be geolocated using a standard Gazetteer (Yahoo! Geocoder in this case). The spatial distribution of these geolocated tweets is shown in Figure 3.

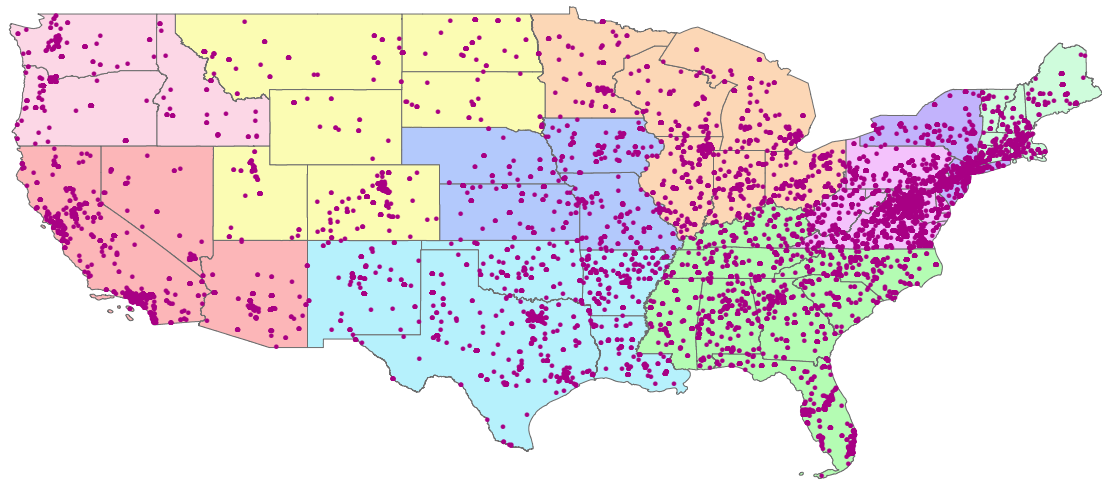


Figure 3. Spatial distribution of tweets discussing GMU in the period 8/12 – 7/13

A visual comparison of these two maps (Figure 2 and Figure 3) suggests that these two communities have comparable spatial distributions. In the following section, we will present specific metrics comparing the spatial characteristics of these two communities. The key findings of the analysis relate to the selection of an appropriate level of spatial aggregation for analyzing social media contents, and on the effect of distance from the point of interest in participation. They are presented in the following sections.

3.2 Level of Spatial Aggregation

A key issue when analyzing the effect of the spatial distribution is to select the appropriate level of spatial aggregation in order to infer meaningful information from the dataset (for example, see the modifiable areal unit problem efforts (Openshaw, 1983)). In order to explore this problem, we have aggregated the data at two different levels: zip code level and state level. In Figure 4 we show a log-log scatter plot of alumni versus tweets spatially aggregated at the zip code level, while we show a similar scatter plot aggregated at the state level in Figure 5.

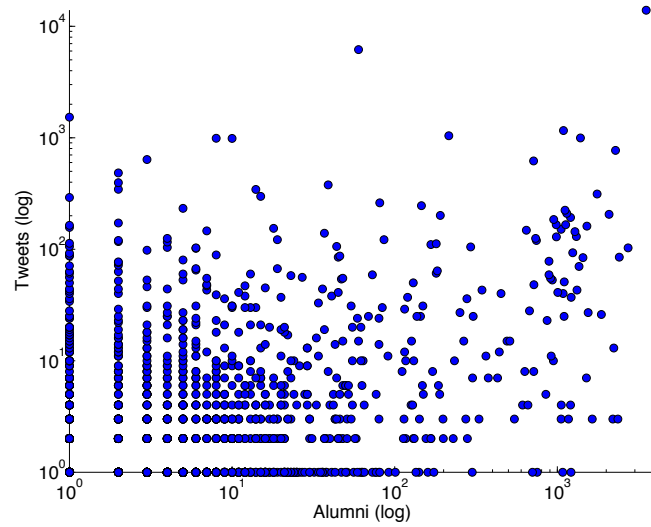


Figure 4. A log-log plot of alumni (horizontal axis) versus Twitter traffic (vertical axis) spatially aggregated at the zip code level

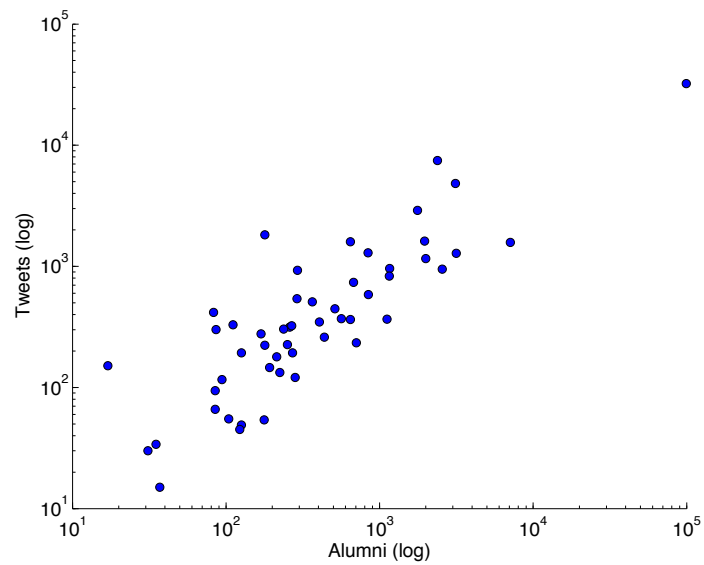


Figure 5. A log-log plot of alumni (horizontal axis) versus Twitter traffic (vertical axis) spatially aggregated at the state level

It is easy to observe that Figure 4 does not exhibit any discernible pattern of correlation, whereas Figure 5 does. More specifically, the correlation coefficient for the dataset that is analyzed at the zip code level (Figure 4) is only 0.34, whereas the same analysis at the state level (Figure 5) is 0.97, indicating a very strong correlation between cyber community activity and physical community population.

This would suggest that although the data can be collected at very fine resolutions, even the individual location level (e.g. precise coordinates), their analysis is better performed at a coarser spatial resolution (e.g. state in this case) in order to better correlate these observations to their real world equivalent. Potentially, this can be attributed to two primary types of artifacts resulting from the geolocation process:

- Toponym-induced artifacts: geolocating tweets whose spatial footprint is available as a toponym (rather than precise coordinates) may result into two different types of errors. Firstly, the geolocation assigned to a toponym from the gazetteer is typically the geographic center of the area associated with this toponym and not necessarily the correct zip code. This tends to inflate the presence of certain zip codes (especially ones at the geographic center of populous geographical areas) in our datasets. Secondly, as toponyms are often non-unique, and this may lead to errors when attempting to resolve a location given an ambiguous toponym.

- User-induced artifacts: as user activity is not regulated by any means, it is not uncommon to have users that have a disproportionate level of participation in the datasets, generating a massive number of contributions. This may bias our analysis by inflating the presence of the corresponding zip codes relative to the rest.

The effects of these artifacts are shown in Table 2, where we list the top 10 most active zip codes for our dataset. We have highlighted zip codes 10007, 46123 and 94930 as they reflect these types of artifacts. For zip code 10007, we have a toponym-induced artifact, as it is the zip code corresponding to the geographical center of New York, NY. For 94930, we have a classic gazetteer-induced error, as Fairfax is associated with Fairfax, CA rather than Fairfax, VA, which is the location of GMU. For zip code 46123, we have a particular user who is contributing enormous numbers of tweets for GMU, thus inflating that zip code’s presence in our datasets.

Table 2. Top 10 most active zip codes in the dataset

Zip Code	Alumni	Tweets	Tweets per alumni	Location
10007	1	1531	1531	New York, NY
08505	1	291	291	Bordentown, NJ
33133	2	484	242	Miami, FL
46123	3	638	212.67	Avon, IN
94930	2	396	198	Fairfax, CA
11767	2	345	172.5	Nesconset, NY
60602	1	164	164	Chicago, IL
33122	1	157	157	Miami, FL
23219	8	989	123.63	Richmond, VA
77002	1	113	113	Houston, TX

The aggregation at a state level minimizes the effects of these artifacts, thus allowing more meaningful patterns to emerge when analyzing the corresponding datasets. In order to further assess the effects of these artifacts, we also plotted in Figure 6 the equivalent of Figure 4 but using only the tweets that had precise geolocation information in the form of coordinates. We see that even in this case, when toponyms are not used and the source for potential artifacts is eliminated, there is no discernible correlation (the correlation coefficient is only 0.41, just slightly improved compared to the earlier 0.34). This is further suggesting that the emergence of patterns at the state level analysis not be due to toponym-induced artifacts, but rather due to the nature of the data, which renders an

aggregation at the state level more suitable for their study. We further studied the behavior of our data when aggregating their spatial resolution to the Federal Standard Region level as they were identified in Figure 2 and Figure 3. The correlation coefficient improved only slightly (0.99, compared to 0.97), which suggests that the state-level analysis offers sufficiently meaningful results while still preserving some reasonable level of spatial granularity.

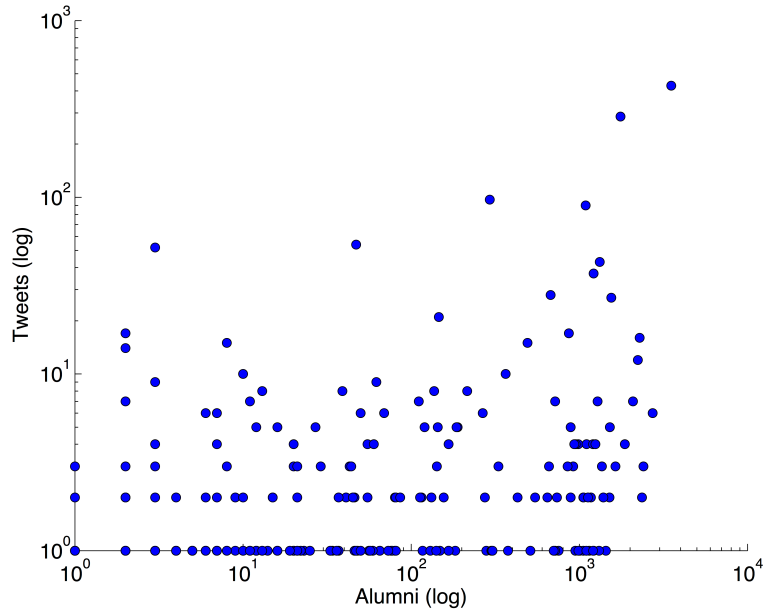


Figure 6. A log-log plot of alumni (horizontal axis) versus Twitter traffic (vertical axis) aggregated at the zip code level, using only tweets with precise coordinates

3.3 Distance from Point of Interest

When we analyze the distributions of alumni and tweets as a function of distance from Fairfax, VA (the location of GMU) we see that they follow comparable patterns (Figure 7). The small bump at a distance of approximately 2,000 miles in both tweets and alumni is corresponding to the transition from the Rocky Mountain region to the Pacific States. The patterns in Figure 7 indicate a distance decay effect, with the numbers dropping quickly and steadily as we move away from the point of interest. However, when we analyze the level of participation of different communities as a function of distance, we get a different picture.

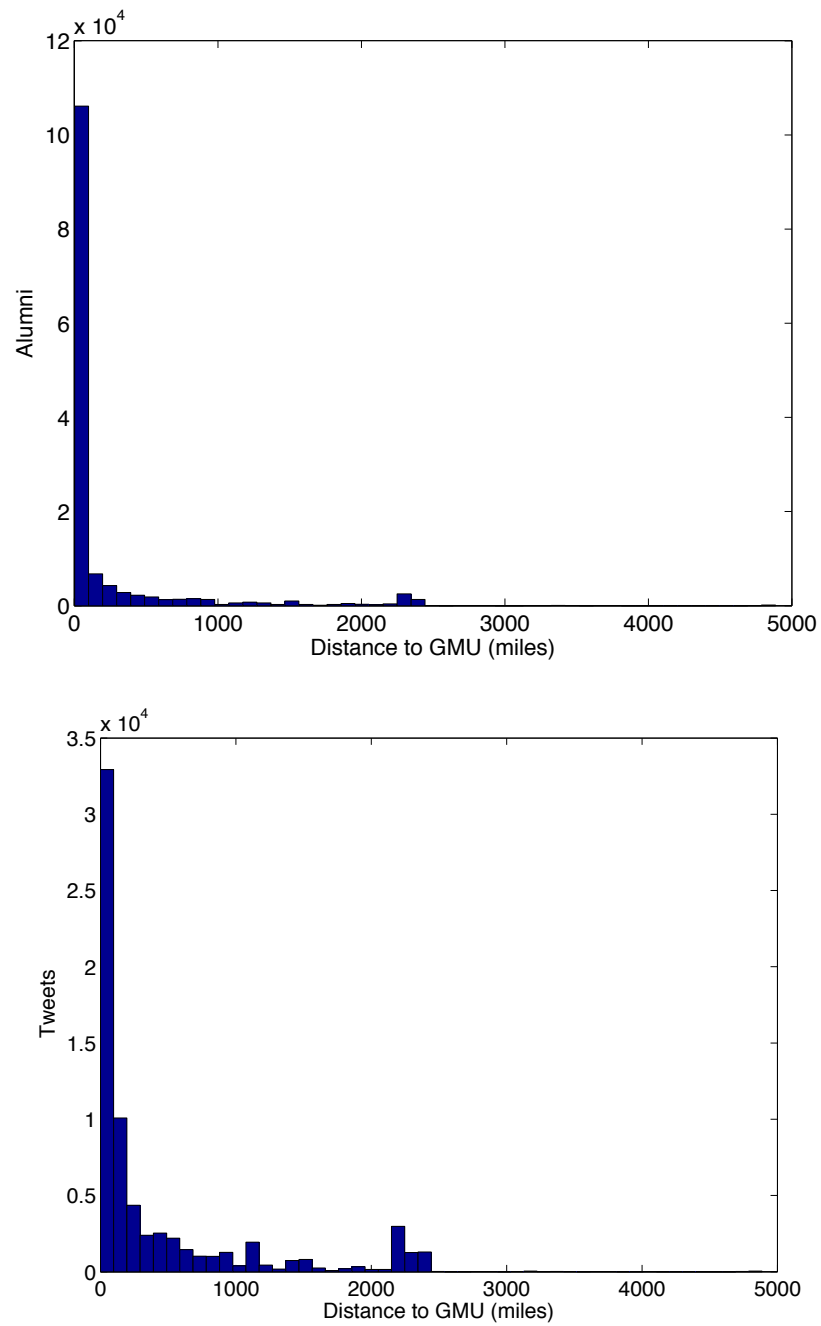


Figure 7. Upper: Alumni number (vertical axis) versus distance from Fairfax (horizontal axis). Lower: Number of geolocated tweets (vertical axis) as a function of distance from Fairfax (horizontal axis)

We define the level of participation as the ratio of tweets per alumnus at different locations. In Table 2, for example, in addition to the numbers of tweets and alumni we have also listed this ratio. This allows us for example to see that Miami, FL has a larger participation index (291 tweets per alumnus) than Houston, TX (113). When we plot this participation index as a function of distance from Fairfax, VA in Figure 8, we observe a very different pattern than the distance decay pattern observed in Figure 7. As a matter of fact, the graph suggests that there exists a reverse participation decay function as participation is overall increasing with distance (instead of decreasing). This is a very interesting finding, as it not only indicates that Tobler's first law does not apply to participation index, but that its reversed one does: enthusiasm (and participation) is higher as we move away from the actual point of interest.

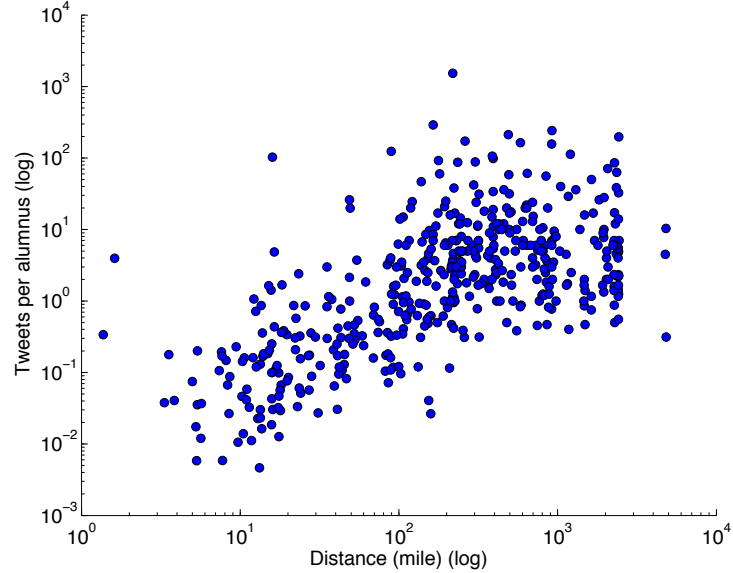


Figure 8. A log-log graph of tweets per alumnus (vertical axis) as a function of distance from Fairfax (horizontal axis)

3.4 Goodness of Fit

In order to reveal whether the cyber footprint can reflect and represent the physical communities, we take a test of goodness of fit for the measure of tweets per alumnus. Several states have few alumni and few tweets, but the ratio of the two turns out to be large. For the purpose of avoiding overemphasize small things, the states are grouped into ten standard federal regions as shown in Figure 9.

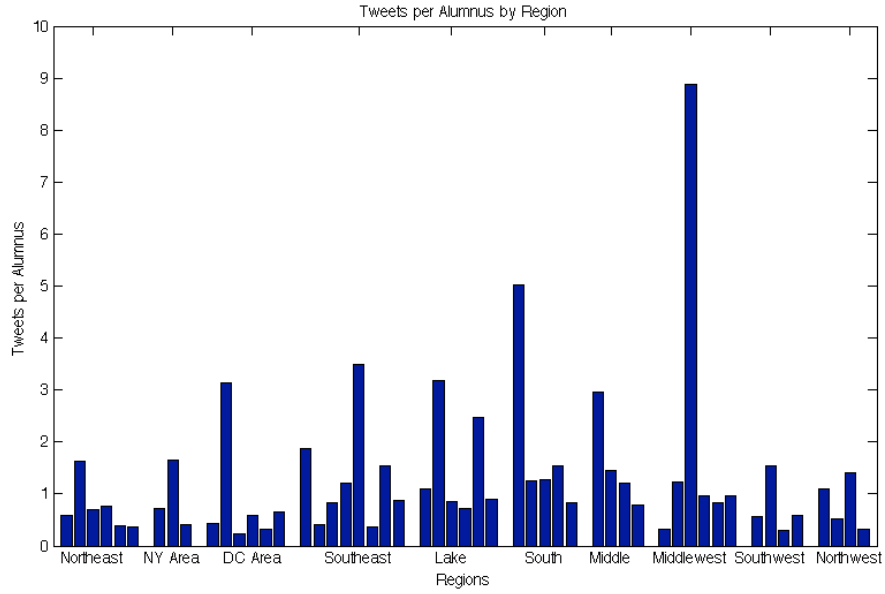


Figure 9. Tweets per alumnus by region

Our assumption is that the measurement of tweets per alumnus rate (ρ) in regions follows the Pearson distribution with mean equal to the national mean (u). Therefore, $\rho \sim P(u)$. In that case, we want to examine whether our observation of tweets is consistent with the expectation.

First, we calculate the nationwide mean of tweets per alumnus:

$$u = T/A$$

where T is the total number of tweets, A is the total number of alumni. As a result, u is equal to 0.49.

Next we get the expectation of tweets (E) in each region:

$$E_i = u * A_i$$

where A_i is the number of alumni in region i.

Now we compare our observation (T_i) with the expectation (E_i), and take a Pearson's

Chi-square test:

$$\chi^2 = \sum_i \frac{(T_i - E_i)^2}{\sigma^2}$$

σ^2 is the variance of expected tweets.

For tweets per alumnus rate, the variance is:

$$var(\rho) = \frac{1}{n} \sum (\rho_i - u)^2$$

Now for the variance of expected tweets,

$$\sigma^2 = \frac{1}{n} \sum (\rho_i * A_i - u * A_i)^2 = var(\rho) * A_i^2$$

As we know, in Pearson distribution, the variance is equal to mean, therefore,

$$var(\rho) = u$$

and

$$\sigma^2 = var(\rho) * A_i^2 = u * A_i^2$$

As a result, χ^2 is equal to 7.7865. The degree of freedom is n-1, which is 9 in our case, and the p value is 0.55, which is far beyond the significant threshold of 0.05. Therefore, we cannot reject the hypothesis, and our observation is consistent with our expectation.

Accordingly, we can conclude that the tweets distribution can reveal the distribution of alumni population.

3.5 Temporal Analysis and Event Detection

As shown Figure 10, we plot the total volume of Twitter traffic daily for the research period. The patterns are corresponding to the university activities, as twitter volume drops during the winter and summer vacations, and builds up through the semesters. It reaches several peaks when there are special events on campus. The daily average tweets about GMU are about 209 tweets per day.

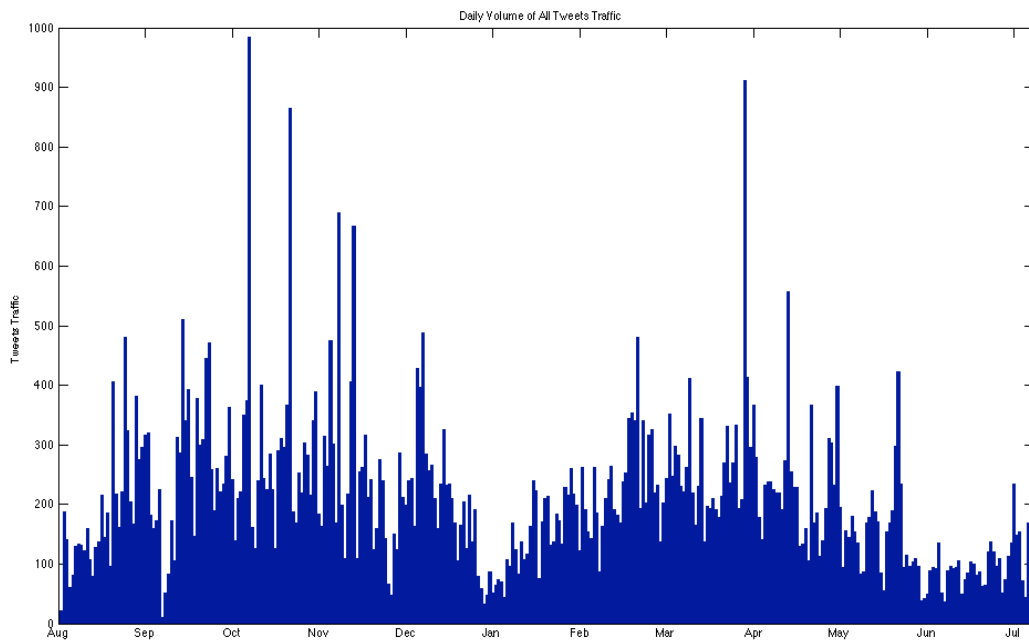


Figure 10. Daily volume of tweets traffic

The daily tweet traffic shows high peaks on particular days. We can further inspect the reason that triggers such a boost in tweet volume. Gathering all tweets in the particular days that have a very high volume, we extracted the most frequent words that have been used on that day, and we generated word clouds as shown in Figure 11. It shows the two word clouds of two days that have the two highest volumes over the research period. Based on the keywords in the word clouds, we can detect the events happened or hot topics discussed on those days (e.g. the President visiting; sports alliance).

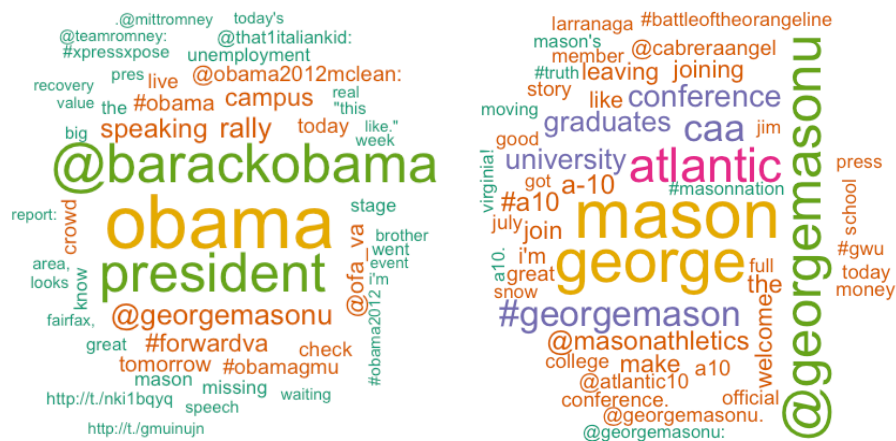


Figure 11. Word clouds of Oct 5, 2012 (left) and Mar 25, 2013 (right)

For each tweet, we applied sentiment analysis and assigned a value indicating the mood of the tweet. The neutral ones are assigned zero; positive ones are assigned a positive value, depending on how “happy” the tweets are, vice versa. The following two figures show the distribution of sentiments and the daily volume of positive and negative tweets for the research period.

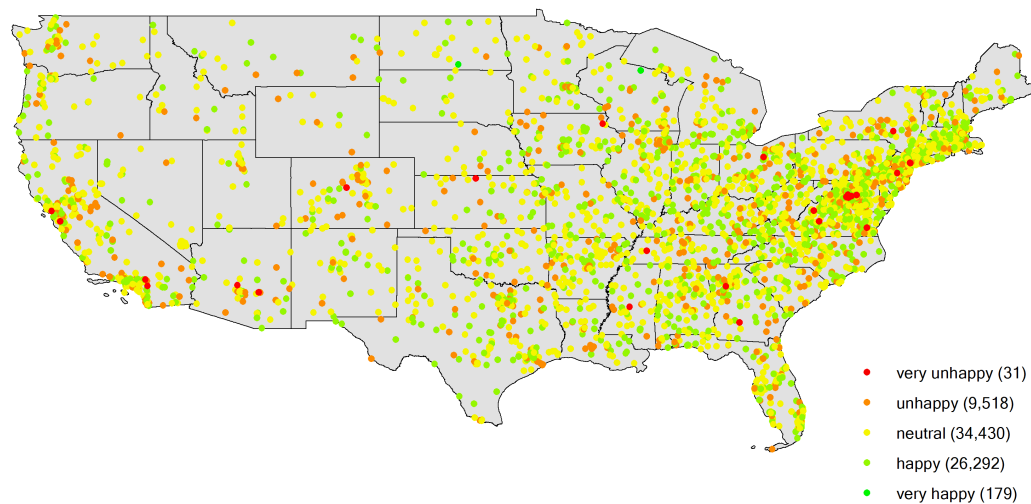


Figure 12. Sentiment analysis for all tweets

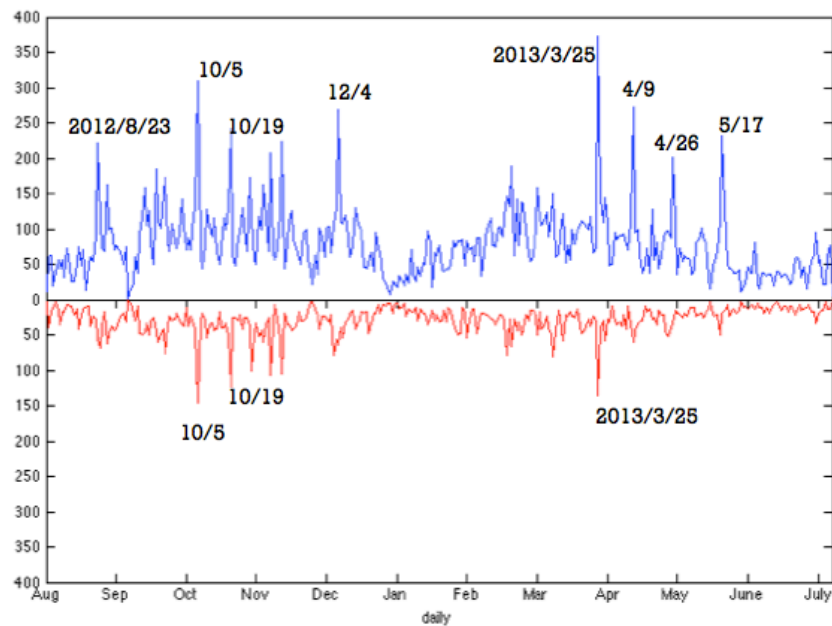


Figure 13. Events identified by tweets volume and sentiment

To summary, social media data provides us opportunities to detect important events/topics, to analyze sentiments/mood of users, and to monitor the trend of the cyber interactions. It is critical to do so, especially with emphasis on geospatial locations. For example, only a few negative tweets might not be a big problem if they distributes sparsely across the space over a long period. But if there is any spatial aggregations (hot spots) of lots of negative tweets in a very short period, it would be a strong indicator that something is going wrong at that specific location.

Chapter 4. Cyber Community in Physical Space

In this section, we will mainly focus on the structures and characteristics of cyber communities and how they distribute in the physical space. In Section 4.1, the network of Twitter data is constructed and described. In Section 4.2, the retweet network of the Mason Twitter data is introduced and examined in the physical space. The retweet distance, which is the most important factor, is examined and evaluated. In Section 4.3, we aggregate the retweet network in a coarser level by state and examine the structure of the state network. The following two sections focus on comparing the retweet network with other networks. In section 4.4, the Mason retweet network is compared with a random network. The random network preserves the network structure of the Mason network, while the location of each user is randomized in the territory of the United States. In section 4.5, another study case is conducted for comparison. The dataset is about a public threat event. Although it is also location-focused, it has stimulated more border responses than the Mason network.

4.1 Retweet Network

When user retweets a retweet post from another user, s/he distributes the original post into his/her own network. Based on the retweet relationship of the tweets, we build a retweet network on all the retweets and original posts. In this network, each node represents a user, and the link between users represents their retweet relationship. The network is directed, as the link points from the user who posts original post (source) to the user who retweets (target), symbolizing the information flows from the source to the target. The following figures illustrate the network. The color in the left figure represents different connected components. There is a huge connected component in the network, which centered with the university's official Twitter account. Besides, there are also many small components, but they are not comparable to the huge component. The color on the right represents the degree of the node. The higher the degree, the more involved or active the user is in participating discussion about the university.

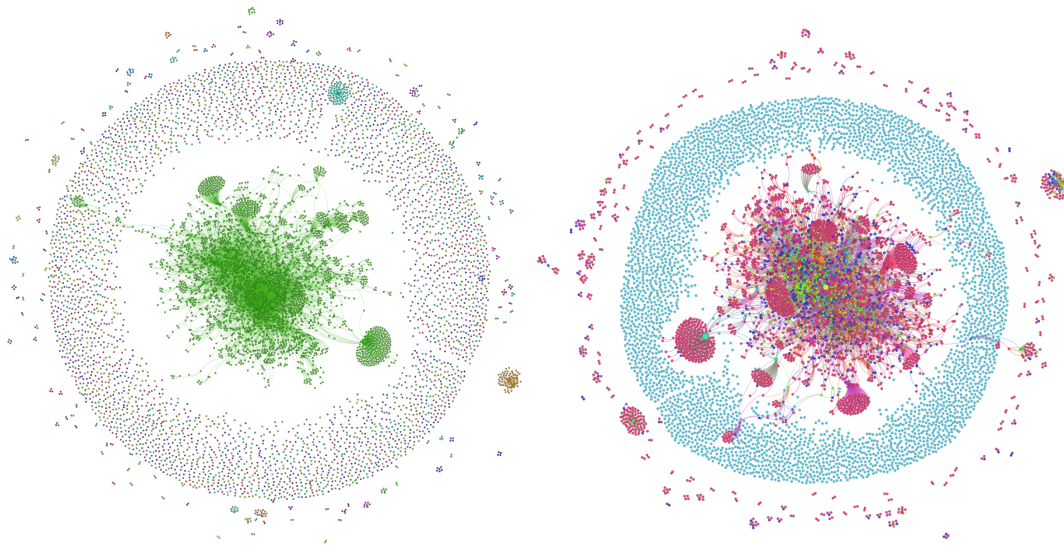


Figure 14. Tweet network, color represents: connected component (left); degree(right)

4.2 Retweet Network in Physical Space

The former network contains some nodes (users) that are unknown in physical location and some isolated nodes that have no retweet at all. Therefore, we generate a subset of the network with only users that have known locations and the retweet links among them.

The network is called Mason retweet network, as shown in Figure 15.

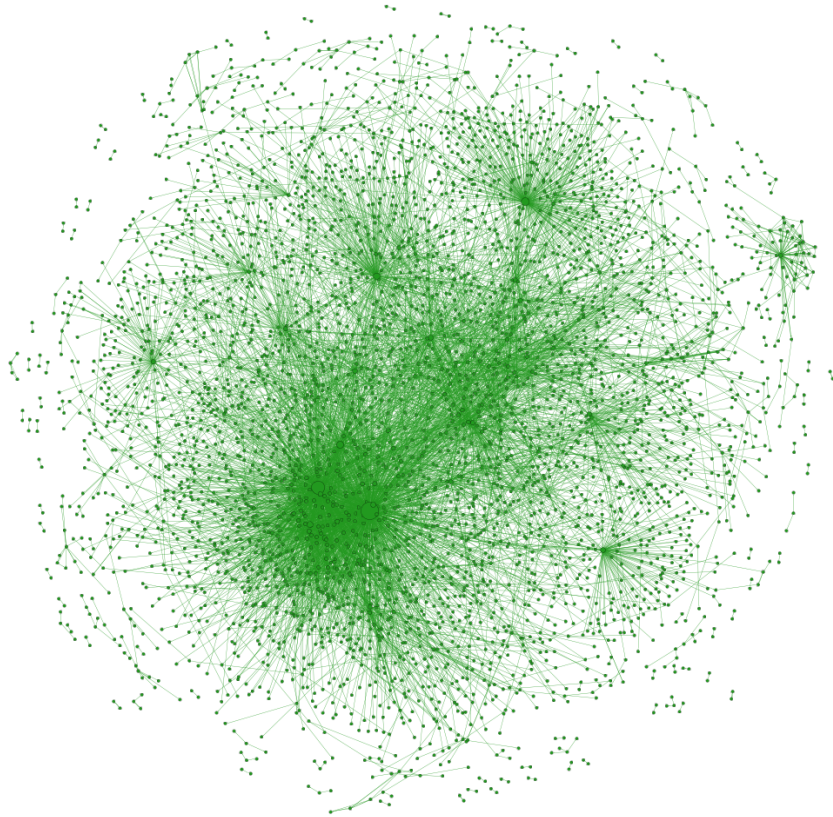


Figure 15. Mason retweet network

We can get some vague impressions about the network from Figure 15, for example, there are several hubs in the network with dense links to other users. There are also many isolated retweet links that only link two users together. Luckily, we have the advantage of knowing the location of each user, and we can map them into the geospatial space, as shown in Figure 16.

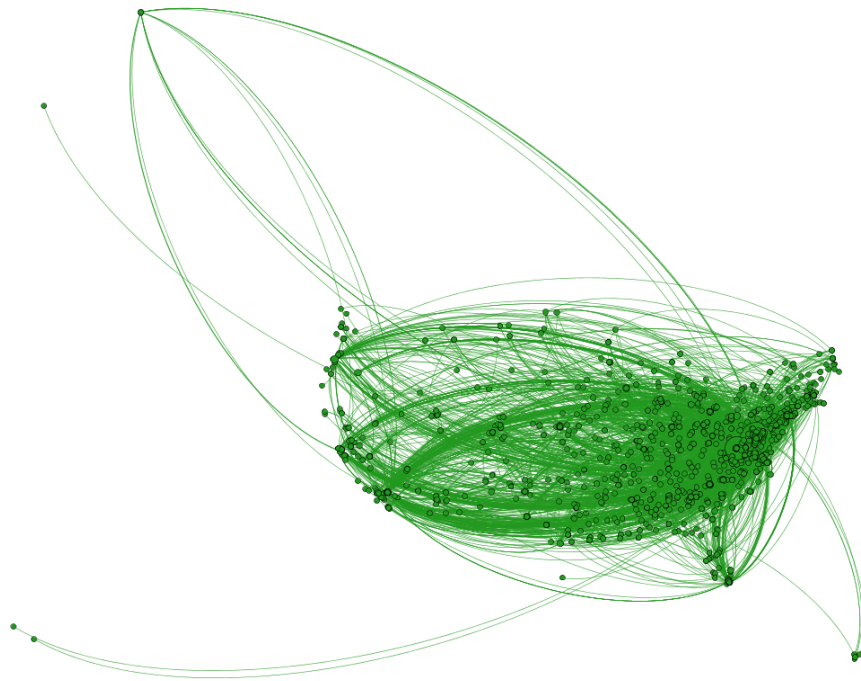


Figure 16. Mason retweet network in physical space

Figure 16 shows the Mason retweet network in geospatial layout. It is very interesting to see that the figure delineates the silhouette of the United States without using a base map. The major cities and populated urban area are revealed clearly on the map. The east coast is denser than the west coast because the point of interest is in the east coast.

There are 4013 nodes and 7422 edges in this subset retweet network. The average indegree and outdegree of the nodes are 1.849, which means for each user, s/he retweets 1.849 users on average (indegree), and there are 1.849 users retweet from him/her

(outdegree). The maximum outdegree is 522, which means that there are 522 users all retweeted from one user, and this user would be the most influencing user in this network. The maximum indegree is 169, which means that there is one user who retweets from 169 different users, and this user would be the most influenced user in the network, and s/he received information from the most “diverse” sources. Interestingly, with a further examination, the most influencing and most influenced user is the same one, which is George Mason University official Twitter account (@GeorgeMasonU).

Table 3. Degree statistics for retweet network

	Minimum	Median	Mean	Maximum
Indegree	0	1	1.849	169
Outdegree	0	0	1.849	522
Degree	1	1	3.699	691

There are 165 different weakly connected components and 3313 strongly connected components in the network. For weakly connected components, there is a giant component with 3381 nodes (84.25%) and 6850 edges (92.29%). It encompasses a significant fraction of the whole network, and it is centered of the point of interest (the

university). The three largest weakly connected components are shown in Figure 17. Surprisingly, the second and third largest components are not centered on the point of interest; instead, it is centered on Manhattan, New York City.

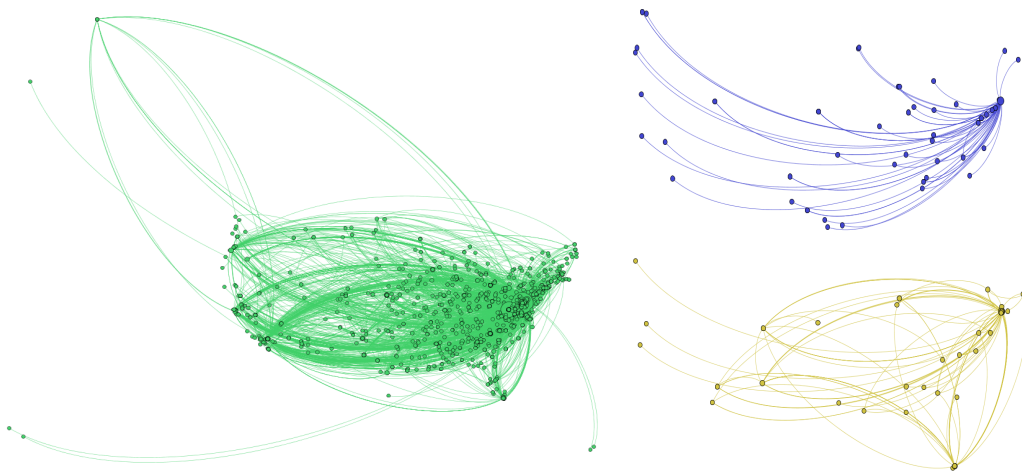


Figure 17. The three largest connected components in the network

The average path length is 4.462, which means on average it takes 4.462 steps for one user to reach another in the network. Also, the network diameter is 13, which is the longest path between two users. It means through retweet relationship, it takes more than 4 retweets to reach another user in the network, and the longest retweet step is 13.

The average clustering coefficient is 0.068, which is relatively low. This means the

neighbors of a node are not likely to be neighbors themselves. This might be attributed to the nature of the network. As the dataset is focused on the university, therefore users are more likely to retweet from the university other than retweet from each other.

There are 185 different groups (sub-communities) detected in the network using modularity separation, and the modularity metric is 0.654. The three largest groups are shown in Figure 18. They cover 9.79%, 9.47% and 7.8% of the total nodes, respectively.

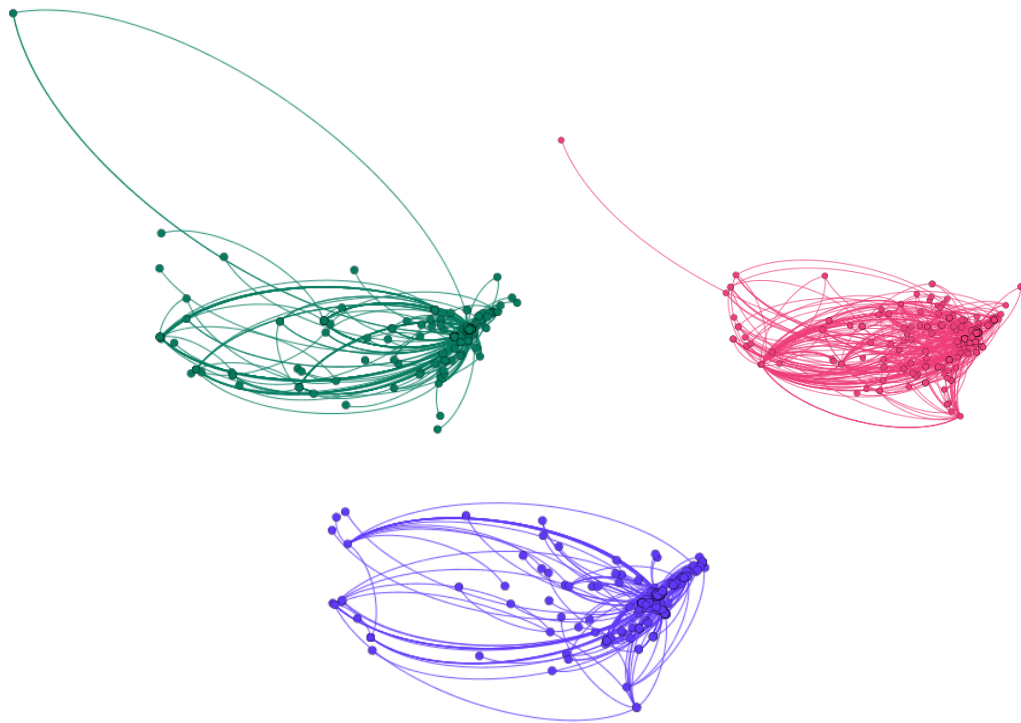


Figure 18. Top three largest groups detected in the network

4.2.1 Retweet Distance of Mason Network

Now that we have the spatial retweet network, we can calculate the retweet distance of each retweet. The following figures show the histogram of retweet distance in kilometers.

The upper figure is the whole histogram with bin equal to 100 kilometers, and the lower figure shows the distance in the range of 2000 kilometers with bin of 50 kilometers.

The histograms show that there exists distance decay effect in the retweet distance of the dataset. Most retweets are within close distance (100 kilometers) to the original post. In fact, if we take a close look at the distance range to 2000 kilometers, we can see that most of the retweets are within 50 kilometers (about 31 miles) of the original post. In contrast, very few retweets are spatially far away from the original post, although we do find a slight boom in the distance of 4000 kilometers (which is approximately the distance from the east coast to the west coast), and the largest distance is 7774 kilometers in our dataset.

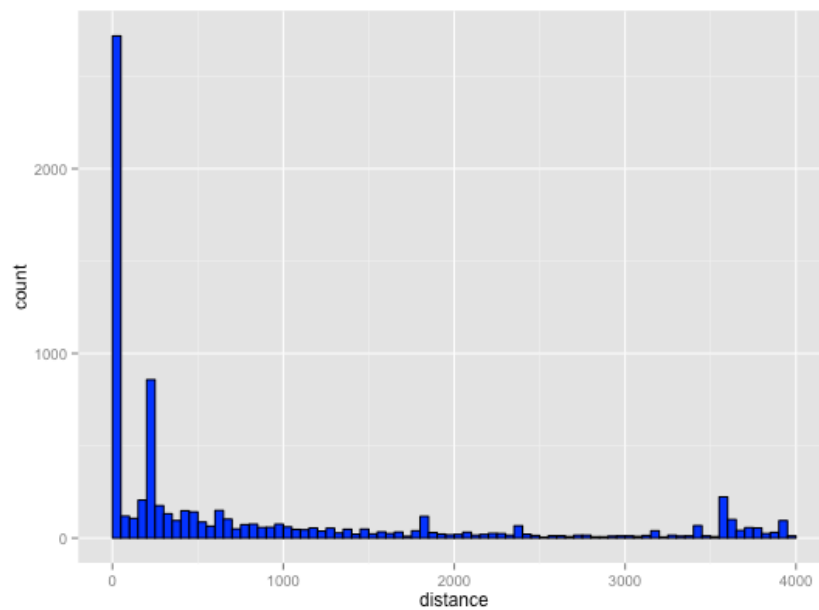
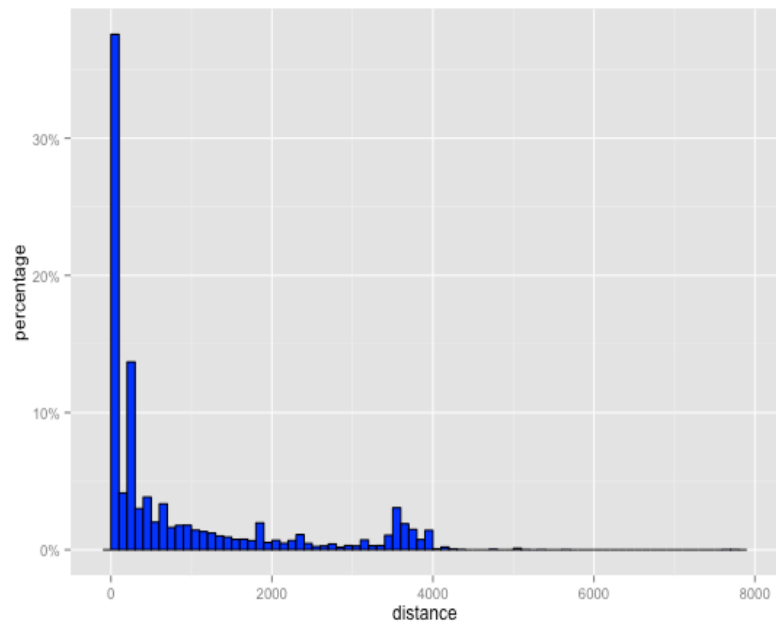


Figure 19. Histogram of retweet distance in kilometers, upper: 0~8000 kilometers, bin is 100 kilometers; lower: 0~4000 kilometers, bin is 50 kilometers.

It is noticed that there is a pit between 50 kilometers to 250 kilometers (31 to 155 mile) and a peak at 250 kilometers. One explanation to this phenomenon is that this distance is likely to relate to the average distance between two nearest major cities. A nature of our dataset is that we target a dataset with explicit location focus. We collect all the tweets about George Mason University, whereas the university locates in Fairfax, northern Virginia. The dataset tends to be centered on the university itself, thus results in shortening the retweet distance. In order to balance such effect, we remove the retweet link that both originates and retweets in the same zip code zone as the university (22030), and calculated the retweet distance again. The statistics of retweet distance are shown in Table 4, and the histograms are shown in Figure 20.

Table 4. Statistics of retweet distance in Mason network (kilometers)

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Retweet Distance	0	17.34	222	844.8	1152	7774
Retweet Distance without 22030	0	28.53	323.3	969	1452	7774

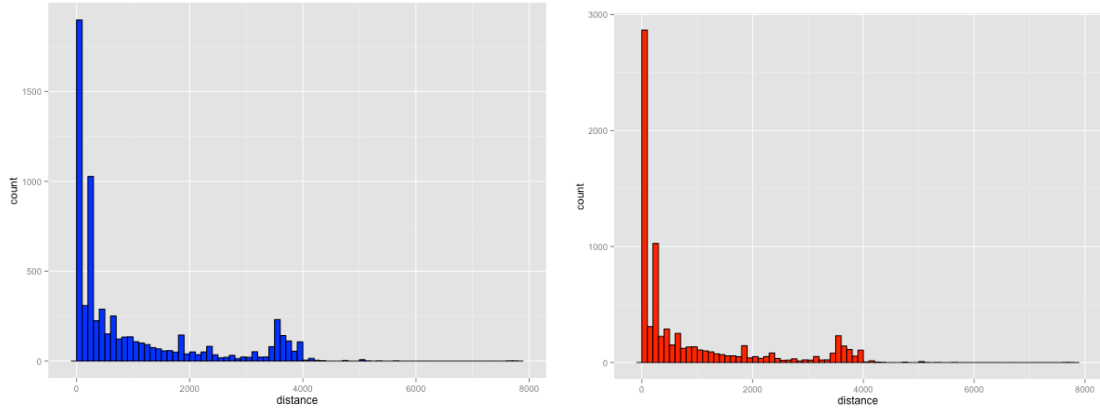


Figure 20. Histogram of retweet distance in the network (left) and the one without 22030 (right)

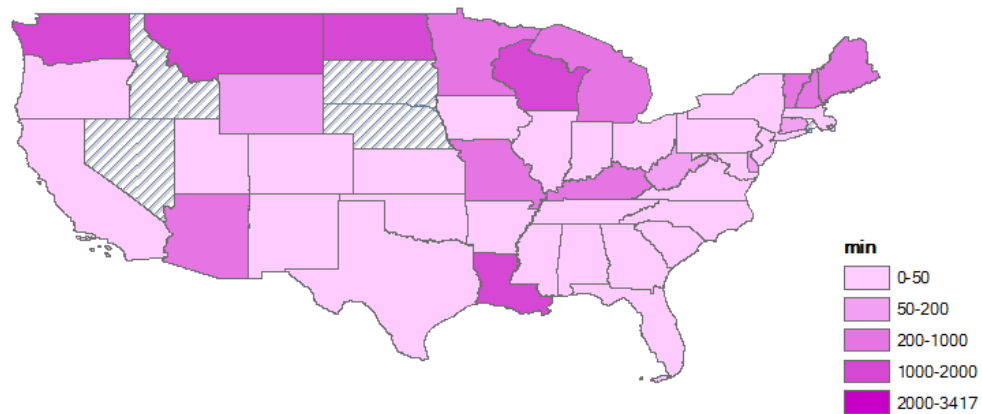
After removing almost 1000 retweets in 22030, we find that the retweet distance does increase, but the increase is very limited. As for the histogram, it is extremely similar as the former one. The distance decay effect may not result from spatial-centered nature of the dataset. Furthermore, another test case is also applied for comparison in Section 4.5.

4.2.2 Retweet Distance for Each State

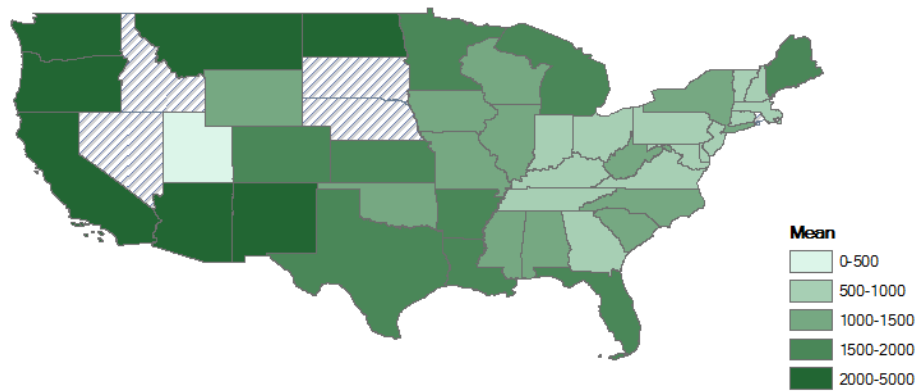
In order to thoroughly study the retweet distance, we calculate and visualize the retweet distance for each state, as shown in Figure 21 (a)~(d). Figure 21(a) shows the minimum retweet distance in each state. The majority of states have a very short minimum retweet distance (less than 50 kilometers), as the users in the states retweets from within the states.

Several states have very long minimum retweet distances (e.g. Montana, North Dakota, Louisiana, etc.). These states are less populated, and the users in the states are more likely to retweet from other states. The Washington state is also among the ones with large minimum distance. It is probably because of its distance to the point of interest.

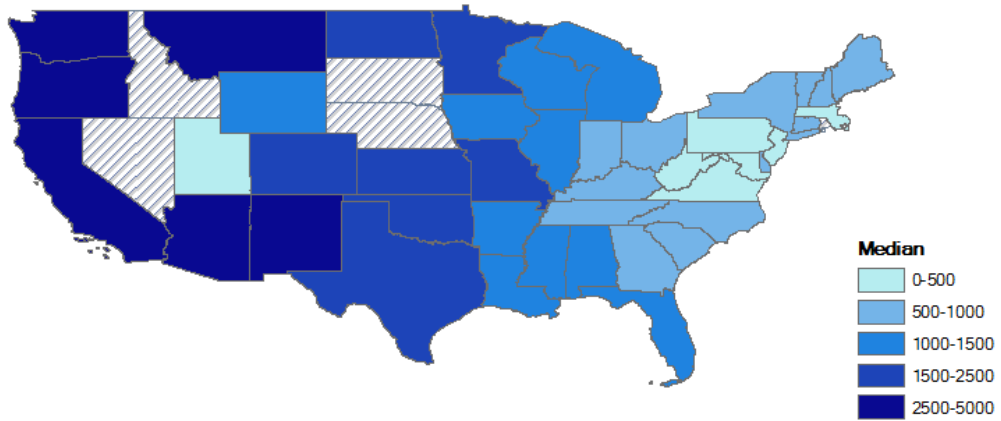
The mean and median distances of each state are comparable as shown in Figure 21 (b) and (c). Figure 21(c) shows very clearly that how the median retweet distance increases when moving away from the point of interest. Concretely, the states that are close to the point of interest have short median/mean retweet distance, while the states that are far away from point of interest have long median/mean retweet distance.



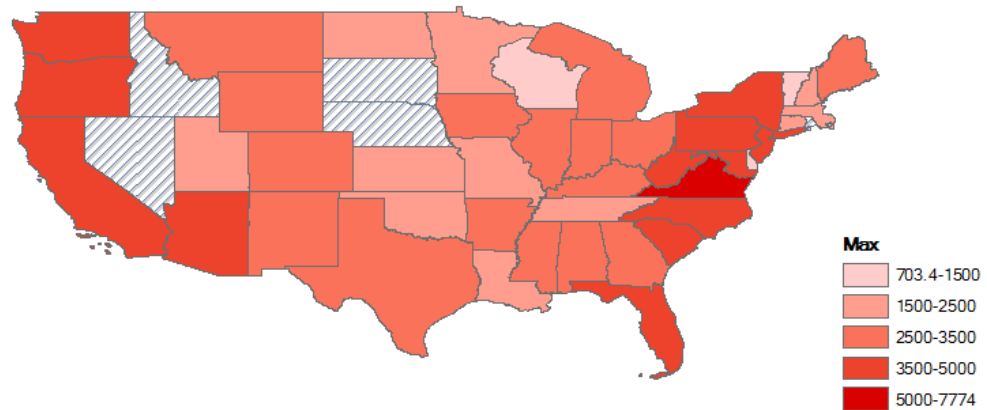
(a)



(b)



(c)



(d)

Figure 21. Retweet distance in each state: (a) minimum (b) mean (c) median (d) maximum

Figure 21(d) shows the maximum retweet distance in each state. It is observed that the states in the middle of the nation have the smaller maximum retweet distance, which is reasonable as they are in the middle, so they have the shorter distance to both east and west coasts. Whereas the states in the east and west coasts have large maximum retweet distance as they have longer distance to the other side of the coasts. There are also some outliers (Delaware, some of New England States) in the coast area, so the users in these states are more likely to retweet from users close to them.

To summary, the maximum retweet distance represents the distance between the state and the furthest location it receives information from, the minimum retweet distance represents how closes each state received information from. Figure 22 shows the scatter plot of minimum versus maximum distance in each state. It is observed that most states have minimum retweet distance of 0, while the states that have very long minimum distance are considered as less populated rural area.



Figure 22. Maximum and minimum retweet distance for each state

4.3 State Network

In order to examine the retweet network in a coarser level, we aggregate the origination and destination of the retweets to the state it belongs to, and generate the state retweet network (Figure 23). The Fruchterman-Reingold layout tends to show the most important hubs in the center and less important nodes in the periphery. As it shown, Virginia, as the home location of the point of interest, is the hub in the network with most connections (retweet links) to other states. It is tightly connected with Washington D.C., Maryland, which are spatially close to Virginia, and also it is connected with California, Texas, New

York, which are highly populated states.

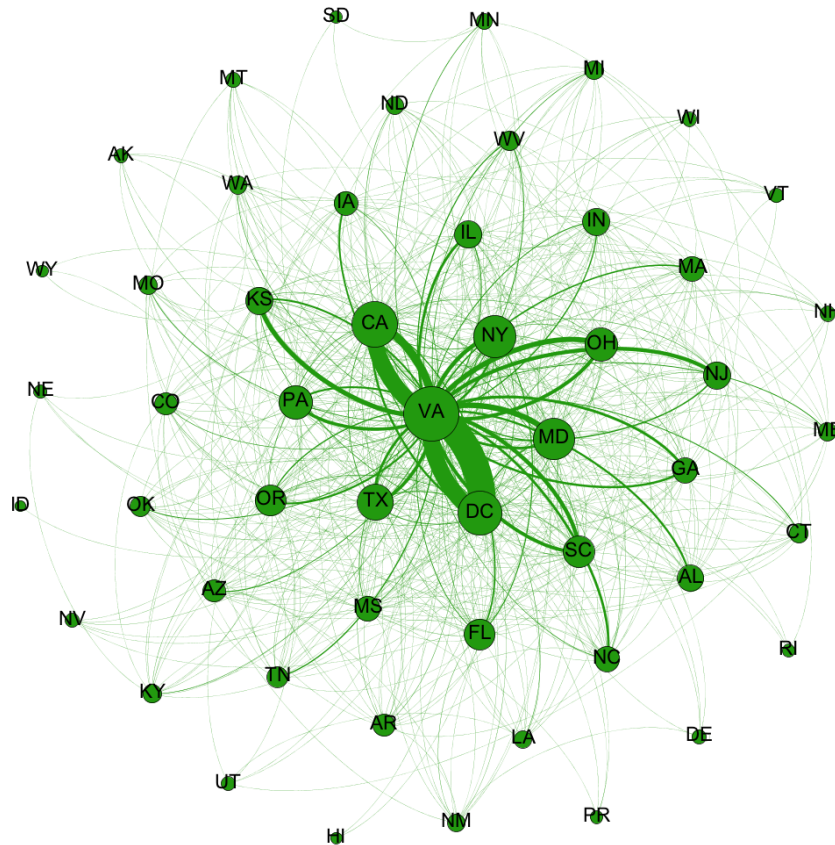


Figure 23. Retweet network of states in Fruchterman-Reingold layout

4.3.1 State Network Structure

The state network has 52 nodes and 654 weighted edges. The 52 nodes are corresponding

to the 50 states plus Washington D.C. and Puerto Rico. The average degree is 12.58, and the weighted average degree is 145.17. The network diameter is 4, and the average path length is 1.78. The network has one weakly connected component and eight strongly connected components. Table 5 shows the top 10 states in centrality. Centrality of a node measures its relative importance and influence within the network. Closeness Centrality of a node measures how reachable the node is to every other node in the network. Betweenness Centrality evaluates the importance of how a node serves as a bridge in the network. Eigenvector Centrality measures the influence of a node in the network. It is based on the concept that connecting to a critical node will enhance the importance of the node itself. Therefore, the top ten states in the table shows the top ten important states in the retweet network. Although locating in the west coast and being far away from the point of interest, California ranks the second in the centrality, exceeds closer states as Maryland and Washington D.C.

Table 5. Top 10 states in centrality

State	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
VA	1.06	531.28	1.00
CA	1.27	229.36	0.86
DC	1.22	188.63	0.75
MD	1.29	174.24	0.73
NY	1.27	116.11	0.78
IL	1.65	76.44	0.46
TX	1.45	74.22	0.68
KS	1.65	61.06	0.50
OR	1.47	44.39	0.43
PA	1.43	43.54	0.54

4.3.2 Originality

For each retweet link, we find the state of the original post, and the state of the retweet. In this way, we can determine the number of original posts and number of retweets in each state. Because of the nature of the retweet network, the tweets in this network are either retweets or original posts that being retweeted. There is no isolated post that is neither retweet nor be retweeted in this network. Figure 24(a) shows the scatter plot of original posts (source) versus retweets (targets). Each state is represented by a point on the figure.

The x axis is the number of original posts; the y axis is the number of retweets. It is shown that the points roughly stay in the diagonal line of $y=x$, which means original posts and retweets are roughly of the same number in each state. Virginia as the home location of the university, has both high numbers in original posts and retweets, and it may skew the result of the fitting line, therefore, we examine the scatterplot and regression fitting line after removing the data in Virginia, as shown in Figure 24(b). It shows that the result is similar to Figure 24(a), and the regression line does not change much. Figure 24(c) shows the same scatterplot in the logarithmic axes, and Figure 24(d) is labeled with states. It is more clearly to see that for most states, the two counts are very similar, while the number of retweets is slightly higher than the original posts.

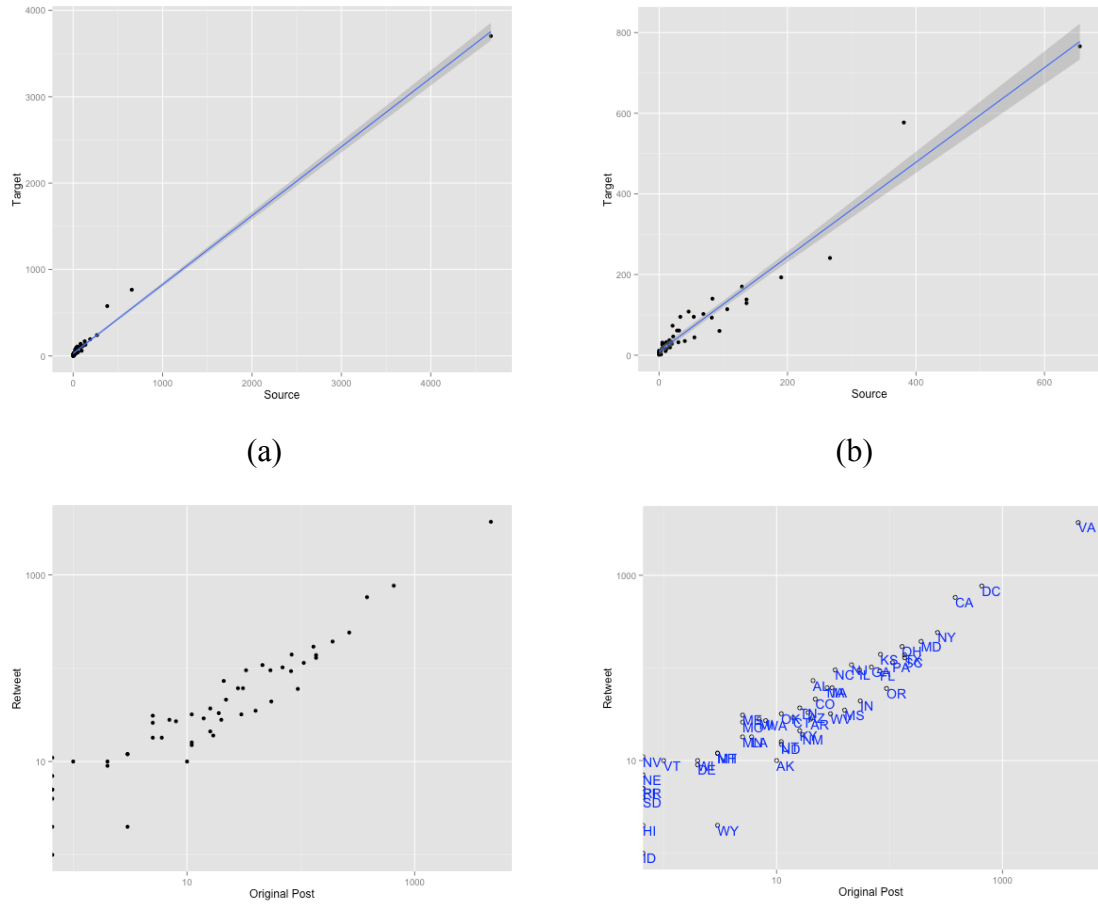


Figure 24. Original posts vs. retweets for each state (a) every state included, (b) remove VA, (c) logarithmic axis, (d) labeled with states

With the statistics of original posts and retweets, we can obtain the originality of each state. Originality is measured as the percentage of the number of original posts that being retweeted to the total number of tweets. Table 6 shows the top ten states of originality in the network. As highlighted, Alaska, ranked the eighth, has the equal number of original posts and retweets, and all other states have more retweets than original posts, thus results

in originality lower than 50%. Virginia, where the point of interest locates, ranks the third in originality after Oregon and Wyoming.

Table 6. Top 10 states in originality

Rank	State	Retweets	Original Post	Originality
1	Oregon	60	94	61.0%
2	Wyoming	2	3	60.0%
3	Virginia	3702	4672	55.8%
4	Indiana	44	55	55.6%
5	Mississippi	35	40	53.3%
6	New York	241	266	52.5%
7	South Carolina	129	136	51.3%
8	Alaska	10	10	50.0%
9	Texas	138	136	49.6%
10	Maryland	193	190	49.6%

4.3.3 Self-retweet Percentage

The retweet link that originates and terminates in the same state is defined as self-retweet.

For each state, we evaluate the self-retweet percentage, which is defined as the ratio of the number of self-retweets to the total number of retweets. This percentage shows whether the major source of information for a state is from interior or exterior. The higher

the self-retweet percentage is, the more interior the state is. The results of top ten states are shown in Table 7. As the point of interest, Virginia has the highest self-retweet percentage. Only Virginia, Utah and Wyoming have self-retweet percentage larger than 50%. The major information resources of these three states are from themselves, and the major information resources for other states are from exterior.

Table 7. Top 10 states in self-retweet percentage

State	Original Posts	Retweets	Self-Retweets	Self-Retweet%
VA	4669	3701	2637	71.25%
UT	11	16	9	56.25%
WY	3	2	1	50.00%
NY	266	241	67	27.80%
MA	28	61	12	19.67%
DC	654	766	129	16.84%
CA	381	574	66	11.50%
DE	2	9	1	11.11%
CO	22	46	5	10.87%
MD	190	193	15	7.77%

4.4 Random Location Network

The Mason retweet network is compared with a random network to further study the pattern of retweet distance. The random location network is generated by randomizing the locations of the users across the United States. The network structure of retweet links is preserved, so the users are getting the same number of retweets, from the same users. As shown in Figure 25, the retweet distance of random location network has a distribution that is very different from Mason retweet network. It shows similar pattern as Poisson distribution. There is a peak at the retweet distance of approximately 1200 kilometers and then decays after the peak with a flat tail up to 10000 kilometers.

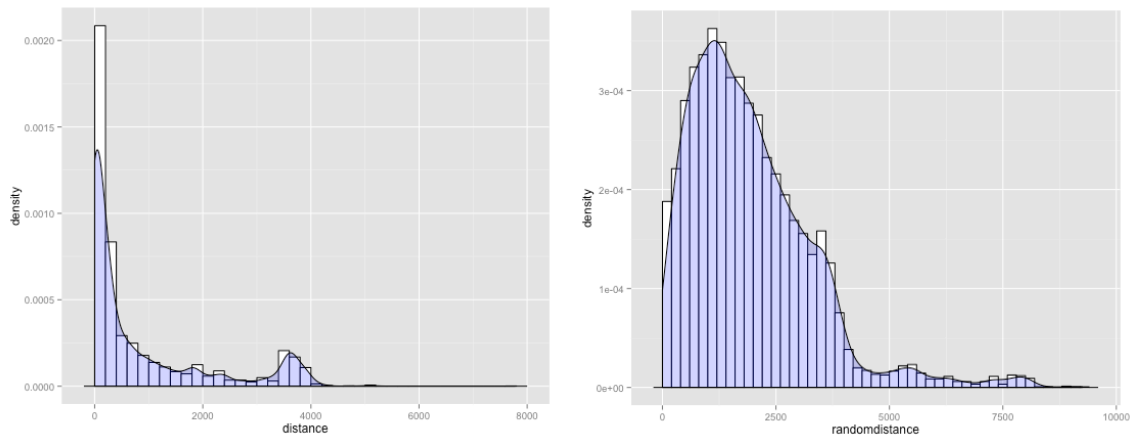


Figure 25. Histogram of retweet distance in Mason retweet network (left) and in random location network (right), bin is 200 kilometers

4.5 Another Case Study

On April 15th 2013, two bombs exploded on the finishing line of Boston Marathon at 2:49pm, and 280 people were injured, five people dead. The news of the tragedy quickly spread out all over the world. We collect Twitter data for the first 10 minutes of every hour after the event for twelve hours, and we calculate the retweet distance for each 10 minutes time slot. Figure 26 illustrates the minimum, maximum, mean and median retweet distance of each time slot. The minimum distance remains zero and unchanged over the period, as there are lots of local discussions about the tragedy. The median and mean distance show that there is a boost in spreading after 4 hours and 11 hours of the event.

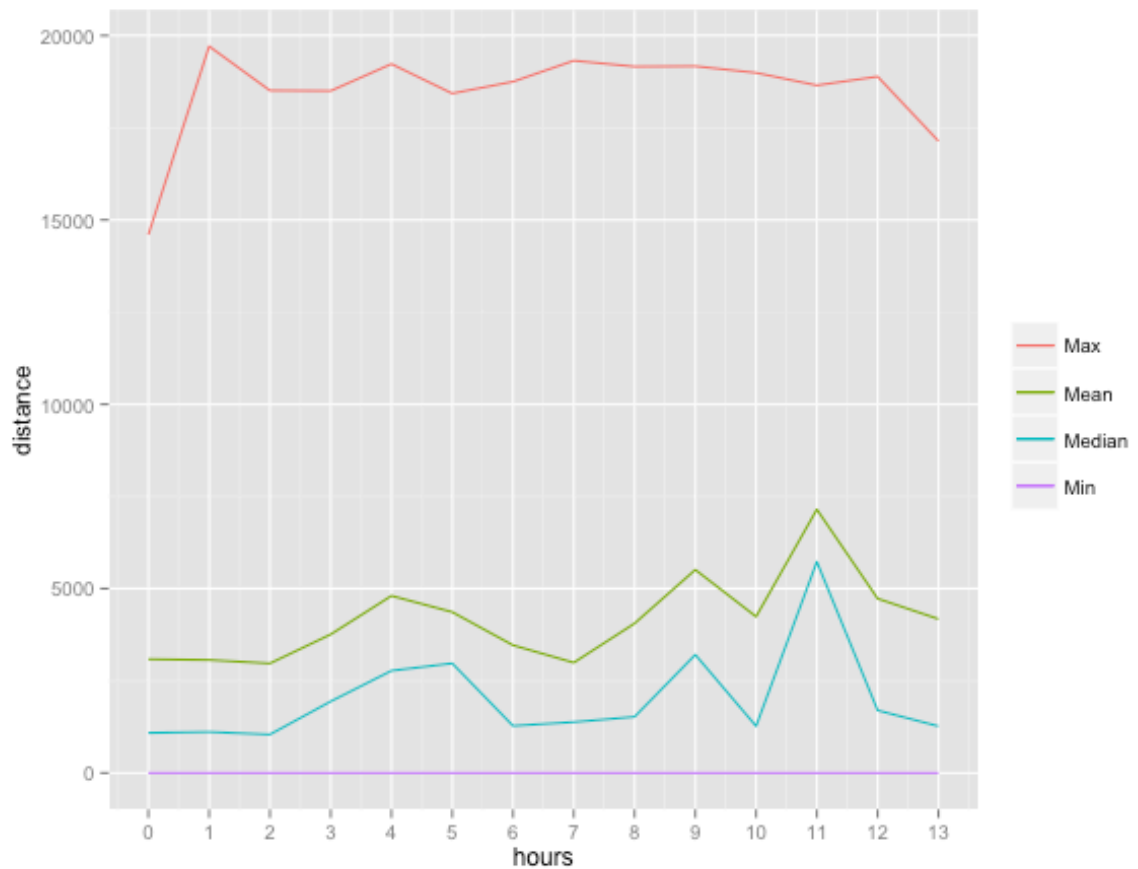


Figure 26. Retweet distance parameters following the Boston bombing event of April 2013 (during the first 13 hours after the event)

When comparing to the Mason Retweet Network, Figure 27 shows the two histograms. Both distributions show a clear distance decay effect. It is clear to see that the retweet distance is much longer in Boston Bombing data, as this event attracts much more border discussion all over the world.

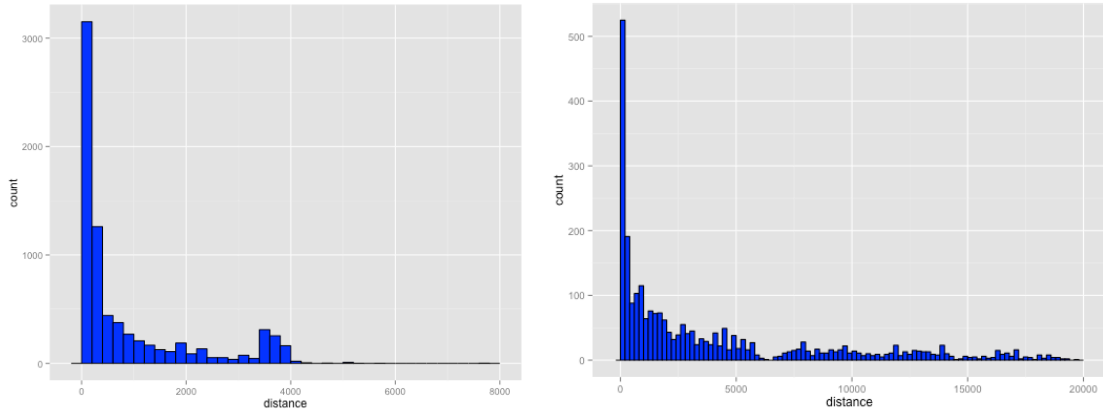


Figure 27. Retweet distance of Mason network (left) and Boston network (right)

Chapter 5. Physical Community in Physical Space

In this section, we study on the physical communities and how they distribute in the physical space. In Section 5.1, the dataset of the physical community is introduced, as well as characteristics of the dataset. It is a Location-Based Social Network called Brightkite. Further, we examine how the spatial factor influences two physical communities based on this dataset: a friendship network (Section 5.2) and a travel network (Section 5.3). In the last section, we study on the distance between major populated cities in the United States.

5.1 Location Based Social Network Dataset

Data from a Location-Based Social Network, Brightkite, is applied. On Brightkite, registered users can check in, post texts or photos, and send messages to their friends. The dataset contains all the public check-ins from April 2008 to October 2010, which amount to 4.7 million records (Cho *et al.*, 2011). The friendship network includes 58,228 users and 428,156 friendship connections.

5.1.1 Check-ins

To further study the nature of the dataset, Figure 28 shows the feature of user check-in times. Majority of the users have only a few check-ins (less than ten times), and very few users check in more than 100 times.

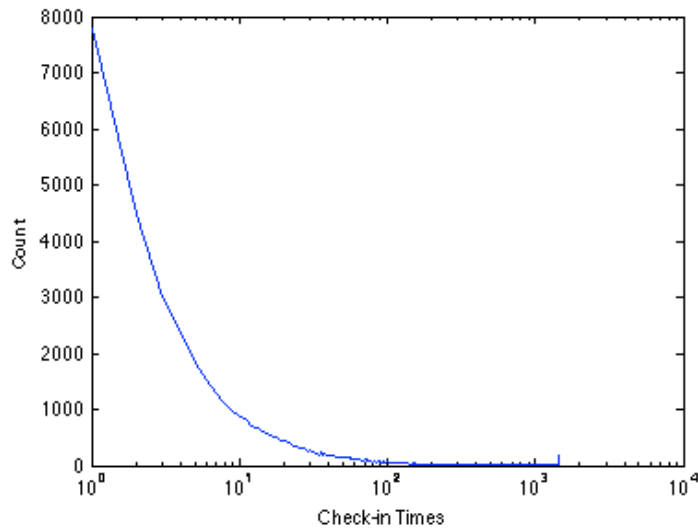


Figure 28. Histogram of all users' check-in data

The percentages of users with different check-in times are shown in Figure 29. Statistically, 12.95% users did not check in at all; 41.75% users checked in at least once but less than ten times. Only 45% of all the users are active users who checked in more than ten times. The further studies will be conducted only on the active users.

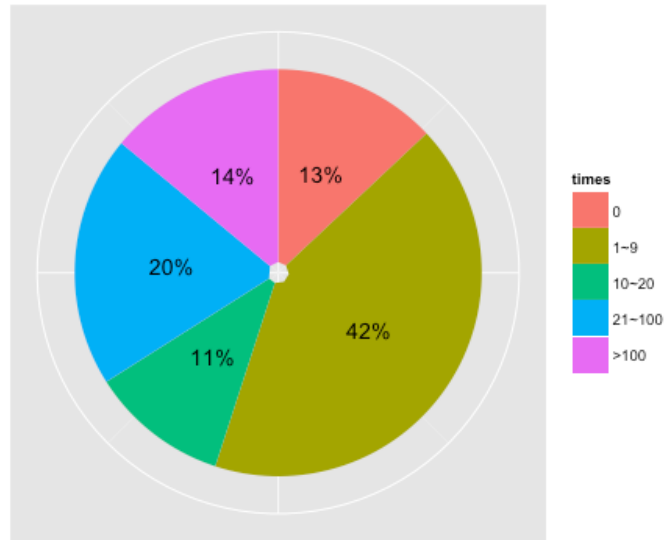


Figure 29. Percentage of all users with different check-in times

5.1.2 Friendship

Similar with the check-in data, the histogram of the number of friends is shown in Figure 30. It is noticed that the histogram line shows a similar pattern as the check-in data. It declines dramatically when the number of friends increases.

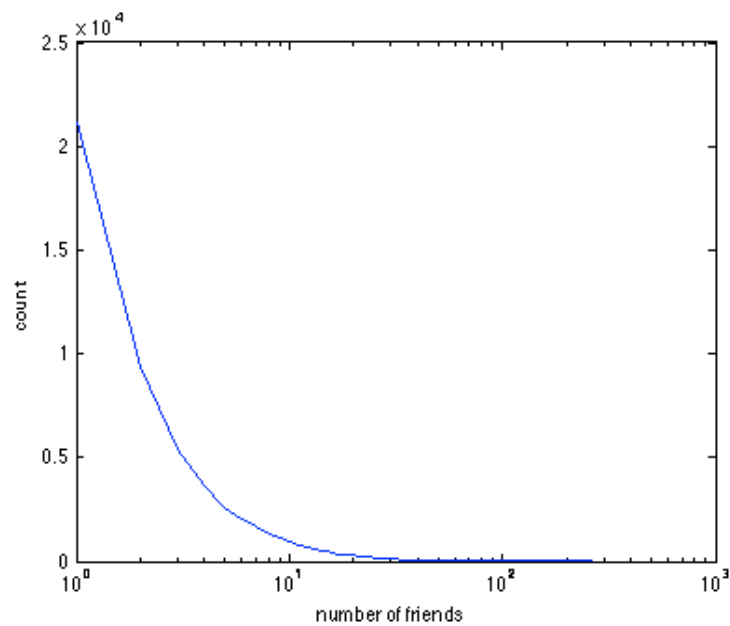


Figure 30. Histogram of numbers of friends for all users

The pie chart shows the percentage of users with different number of friends: although the average number of friends is around seven in Brightkite social network, 68% users have less than five friends.

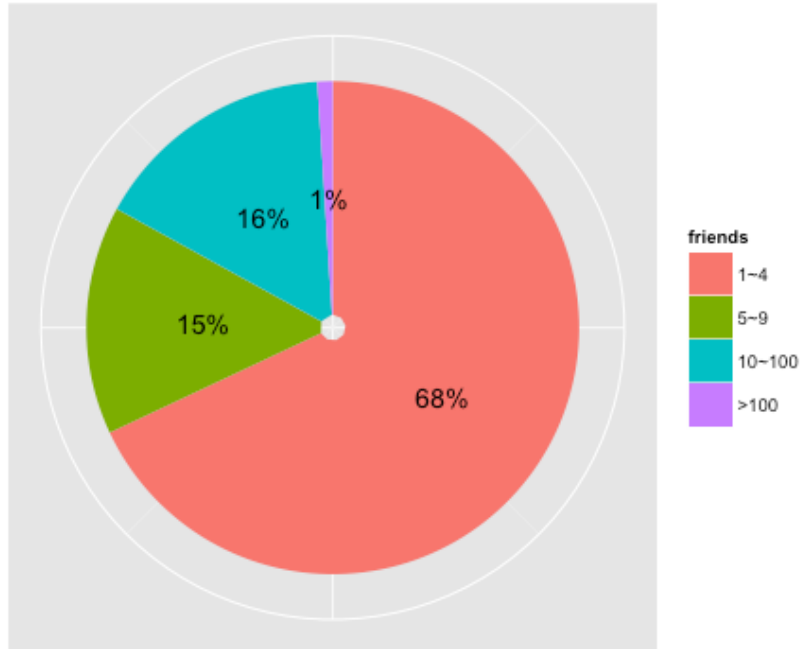


Figure 31. Percentage of all users with different number of friends

5.1.3 Home Location of Each User

For each user, the home location is defined as his/her most frequent check-in place. Take the check-in data of the first user as an example. Figure 32 shows the distribution of the check-in data in different scales. This user has 2085 check-ins during the data-collection period: most check-ins locate in the United States, and several in Europe (Figure 32(a)). When zooming into all the check-ins in the United States (Figure 32(b)), we found that most check-ins in the United States are in Colorado (Figure 32(c)). When zooming into

the densest area in Colorado (Figure 32(d)), a centroid can be found in the densest area, which will be the home location of this user.

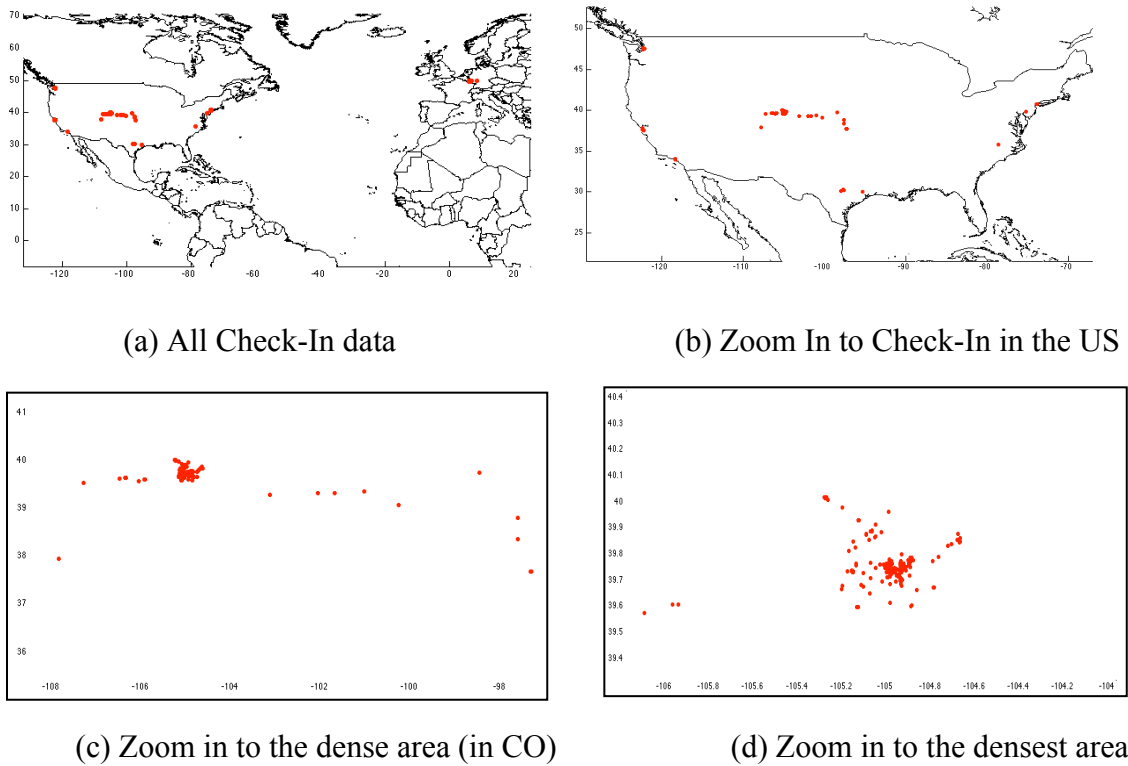


Figure 32. Check-in data of one user

Hierarchical clustering method is applied to find out the home location of each user. A prerequisite of clustering is that there should be some certain number of data to be clustered. Otherwise the result might not be meaningful. Therefore, only active users who

have at least ten check-ins are included in the further study.

To build the hierarchal tree, single linkage (minimum distance) is applied, and thresholds are set to cut the tree at the proper location. The cluster with the largest number of objects is the “home” cluster of the user. So the centroid of the largest cluster is the “home location” of the user. For example, for a specific user (user ID=0), the hierarchal tree is shown in Figure 33. To cut the tree, set a threshold d as the distance between two places (cities, towns, etc.). If one checks in farther than d kilometers from her/his last check-in, s/he has moved to a new place. The threshold is set as 0.5 degree, which is about 30miles or 50 kilometers in medium latitude area.

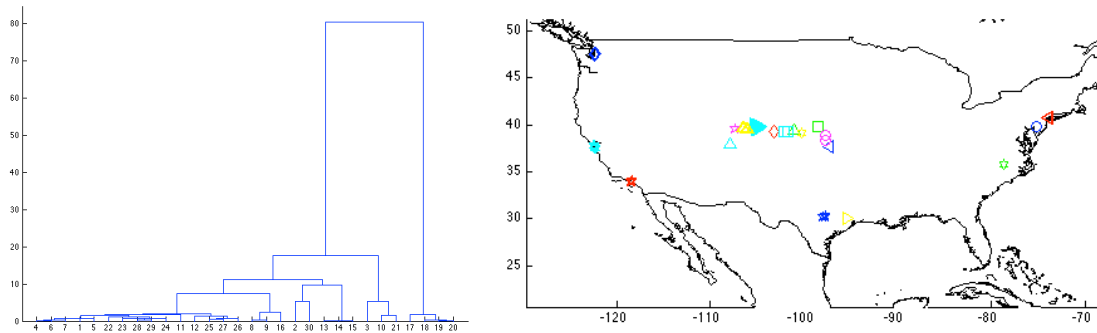


Figure 33. Left: hierarchal tree for user ID=0; Right: clusters (only check-ins in the US)

For each user, the home location is set from the procedure described above. Figure 34

shows the home location of all the active users. In order for comparison with other dataset, only the users who are based in the United States are studied in the Section 5.2 and 5.3.

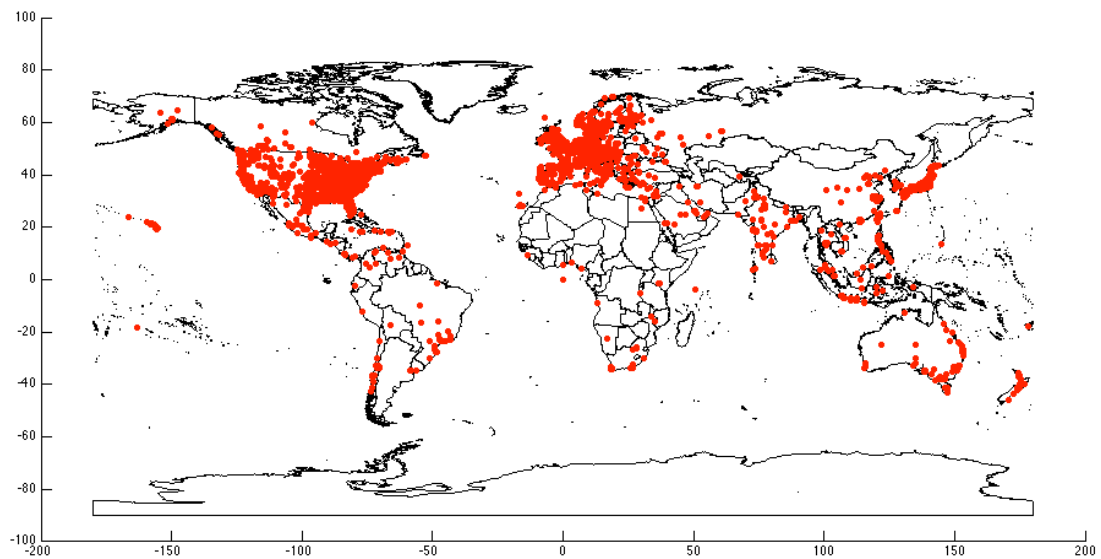


Figure 34. Active users' home locations

5.1.4 Human Mobility

In order to evaluate the mobility of human activities, we classify the check-ins for each user into two categories: “home check-in” and “travel check-in”. Home check-ins are check-ins that are within 50 miles (80 kilometers) of the home location. For travel

check-ins, we also differentiate the travel check-ins into short distance check-ins (more than 50 miles (80 kilometers) and less than 300 miles (500 kilometers) from the home location) and long distance check-ins (farther than 300 miles).

Once the check-ins are categorized, we calculate the travel/home percentage and the percentage of the short and long distance travel check-ins for each user. For each user, the percentages are calculated as:

$$\begin{aligned}
 P(home) &= \frac{\sum home\ check - in}{\sum check - in} \\
 P(travel) &= \frac{\sum travel\ check - in}{\sum check - in} \\
 P(short - distance\ travel) &= \frac{\sum short - distance\ travel\ check - in}{\sum check - in} \\
 P(long - distance\ travel) &= \frac{\sum long - distance\ travel\ check - in}{\sum check - in}
 \end{aligned}$$

These percentages reveal the pattern of human mobility. Figure 35 shows the histogram and percentage of travel of all active users. The x axis shows the percentage of travel; the left y axis shows the count of users, and the right y axis shows the percentage of users in that category. It is obvious that most population has small percentage of travel, as there are more than half population has a travel percentage less than 10%. Very few people have large percentage of travel check-ins.

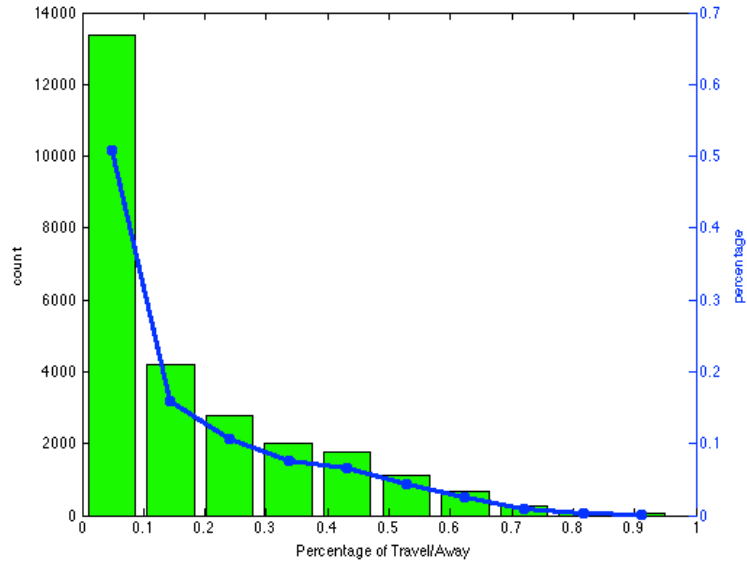


Figure 35. Count and percentage of users with different travel percentage

Figure 36 shows the histograms of short distance travel and long distance travel for users with different travel percentage. For users in the three categories (more than 40% travel, less than 10% travel, 10%~40% travel), the percentage of short distance and long distance travel are very close to each other. Therefore, short distance travel and long distance travel takes up similar percentage of human mobility.

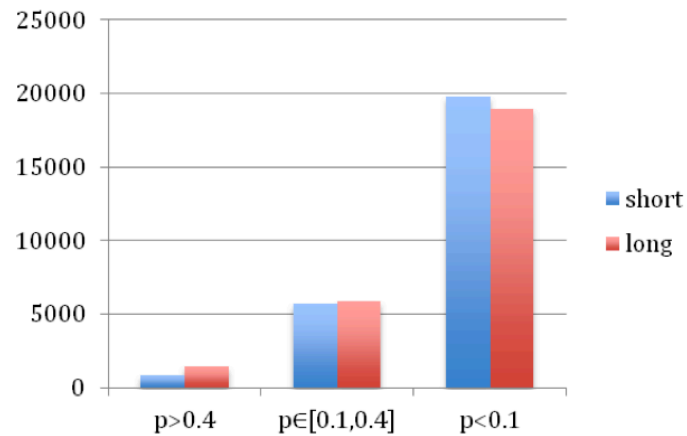


Figure 36. Histogram of short-distance travel and long-distance travel for users with different travel percentages

For human mobility, Table 8 shows that the mean travel percentage is 16.24% (median is 9.13%). Combined with the histogram shows Figure 35, it is obvious that most people stay around their homes, and travelling takes only a small part of their lives. By comparing the short distance and long distance travel, we observe that they are of the similar percentage, which means generally people travel long distance and short distance of the similar times.

Table 8. Mean and median of human travel mobility

	Mean	Median
Travel	16.24%	9.13%
Short-distance travel	7.39%	1.76%
Long-distance travel	8.85%	0.84%

5.2 Friendship Distance

In this section, the spatial distance in the Brightkite friendship network is examined. The dataset contains the friendship connections between users. For example, the friend connection of user ID=0 is shown in Figure 37. In order for comparison with other dataset, only users with home location in the United States will be included.

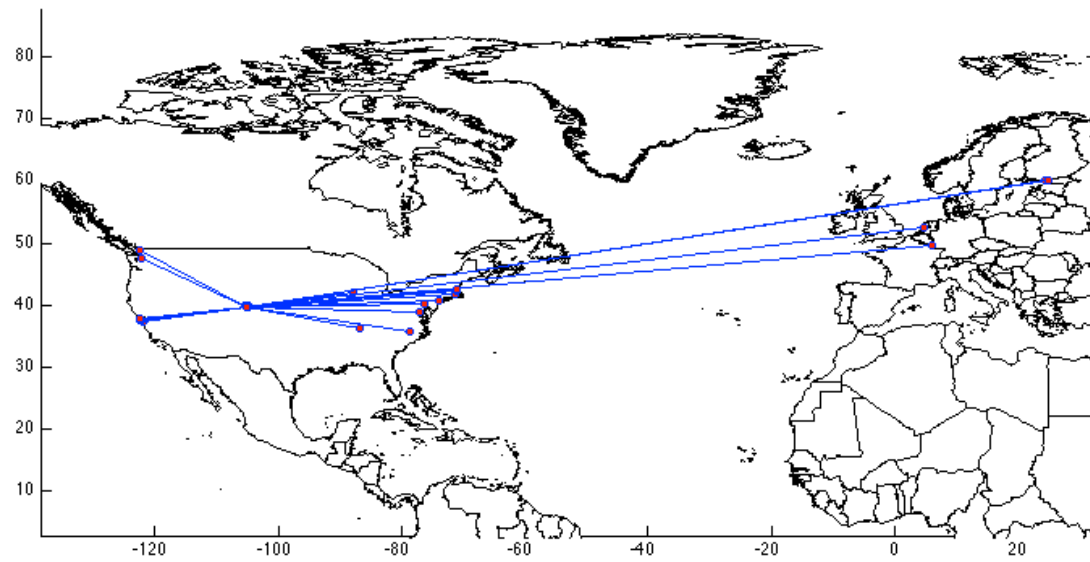


Figure 37. User ID=0 and his/her friends

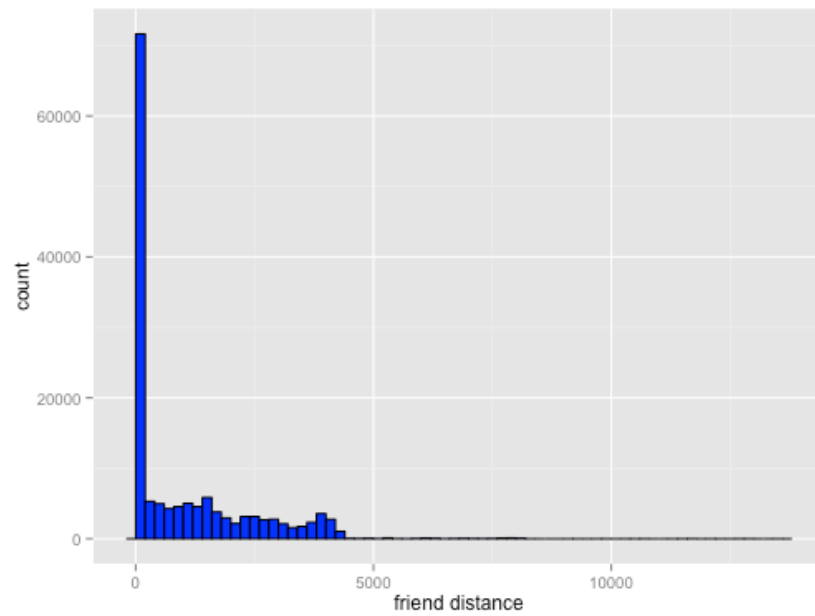


Figure 38. Histogram of friendship distance (bin is 200 kilometers)

For each friendship pair in the United States, we calculate the distance between their home locations and plot the histogram of distances in Figure 38. It is noticed that the histogram of the friendship distance is highly skewed with a high peak of distance less than 200 kilometers and a long tail for distance more than 5000 kilometers. Generally speaking, the number of friendship decays as distance increases. Concretely, majority of the friend pairs live very closely within 200 kilometers, and few friends live very far from each other. Therefore, people are most likely to have friends close to them than far away from them.

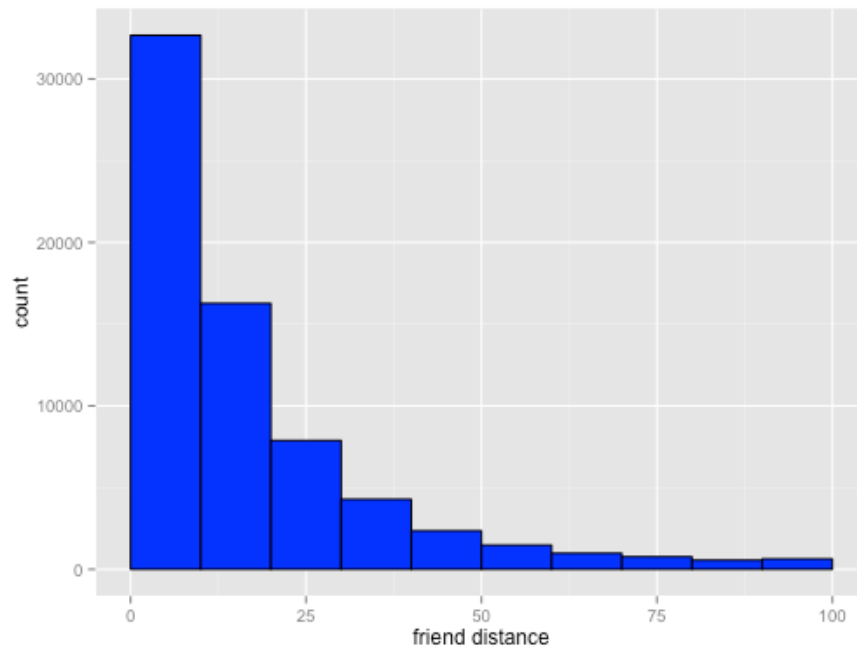


Figure 39. Histogram of friend distance in the interval of 0~100 kilometers (bin is 10 kilometers)

Take a close look at the distance interval of 0 to 100 kilometers (Figure 39), the number of friendships decays monotonically as distance increases. However, when zooming into the distance interval of 100 to 500 kilometers, the results are different. As shown in Figure 40, the number of friend pairs does not decrease monotonically as there are several peaks at distance of 150 kilometers and 300 kilometers. The reason for this phenomenon might be that the friends within 150 kilometers are local friends. As distance increases, the probability to have a local friend decreases. The friends beyond 150 kilometers might be online friends or former local friends who have moved away, so

the probability increases a little bit. In Cho *et al.* (2011), they explained that the non-uniform population density might be the reason, and we will further examine this factor in Section 0.

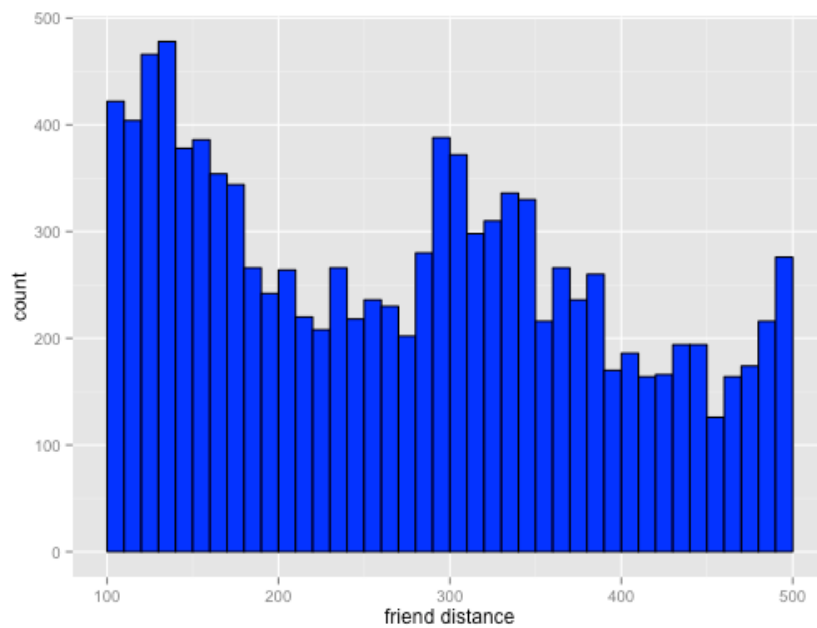


Figure 40. Histogram of friend distance in the interval of 100~500 kilometers (bin is 10 kilometers)

5.3 Travel Distance

In this section, we examine the travel distance based on Brightkite check-in data. Similar

as friend distance, only users with home location in the United States are included, and the dataset of check-ins are also subsetting accordingly. There are 16,179 active users and 2,835,325 check-in records in the United States. The average check-in times for active users in the United States are 175, and the median check-in times are 50. Figure 41 shows the histogram of check-in times for active users in the United States. It is highly skewed with skewness of 3.71 and standard deviation of 347.71. Interestingly, there is a very slight bump at the check-in times of 2000 times, indicating a small group of highly active users.

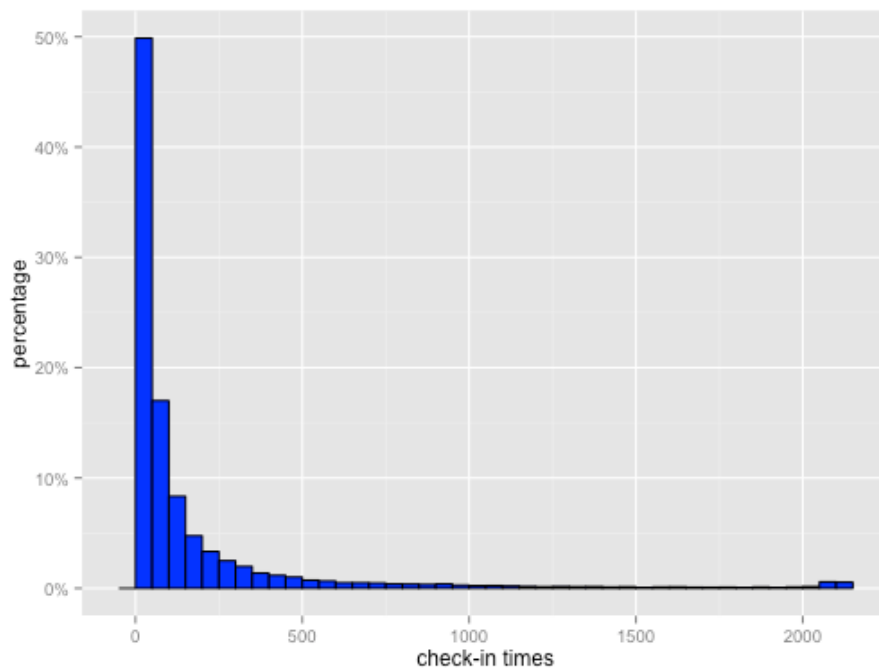


Figure 41. Histogram of check-in data for users in the United States (bin is 50 kilometers)

As introduced in the section of human mobility, each check-in record of each user is categorized into either home check-in or travel check-in according to its distance to the home locations, and the threshold is set as 80 kilometers. For each travel check-in, we calculated the distance from the check-in location to the user's home location, and we obtain the dataset of travel distance of all the Brightkite users in the United States. As shown in Figure 42, the histogram of travel distance has the highest peak within 200 kilometers. It drops dramatically at about 5000 kilometers and becomes flat after that. Therefore, people are more likely to travel to locations close to their homes rather than far away from their homes. The travels that spans than 5000 kilometers are very few.

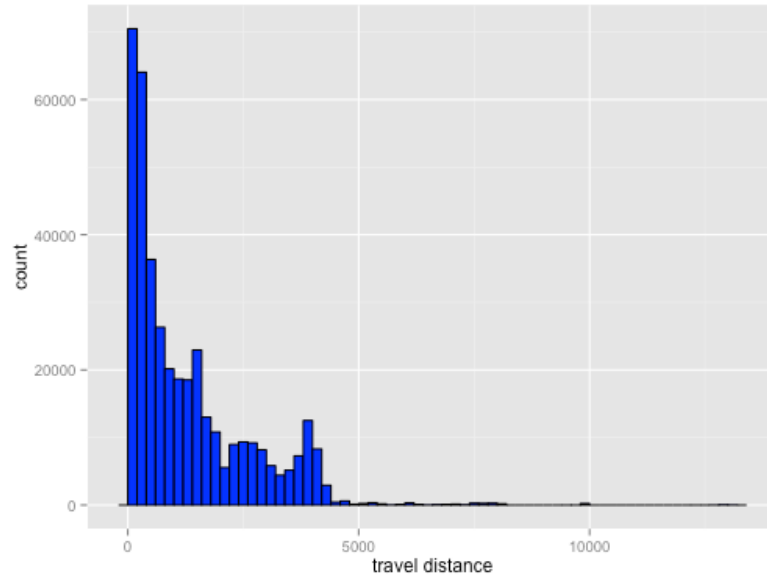


Figure 42. Histogram of travel distances for all users in the United States (bin is 200 kilometers)

5.4 City Distance

In Cho *et al.* (2011), they explain that the non-monotonic decreasing of friendship distance might be attributed to uneven distribution of population. In this section, we examine the distance distribution of major populated cities with the population greater than 100,000. There are 245 cities in the United States as shown in Figure 43.

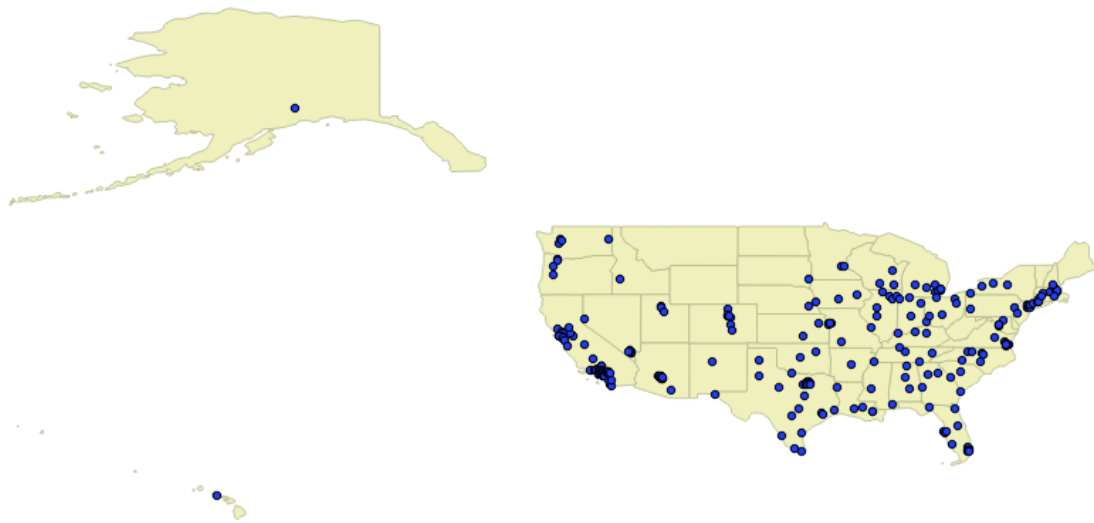


Figure 43. Major cities in the United States with the population greater than 100,000

For each city, the distances to the rest of the cities are calculated. The result is a symmetric distance matrix with 245 rows and 245 columns. The elements in the upper triangle without the diagonal line of the matrix are the dataset for the city distances. As shown in Figure 44, the histogram shows a peak at the distance of 1500 kilometers, but generally speaking, the distance between populated cities distribute roughly evenly less than 4000 kilometers. While the 4000 kilometers are approximately the distance from the east coast to the west coast, so the distances further than 4000 kilometers are corresponding to the cities outside the continent of the United States. When comparing to the friendship distance, the histograms do not show the similar patterns, therefore the distribution of friendship distance might be due to other factors.

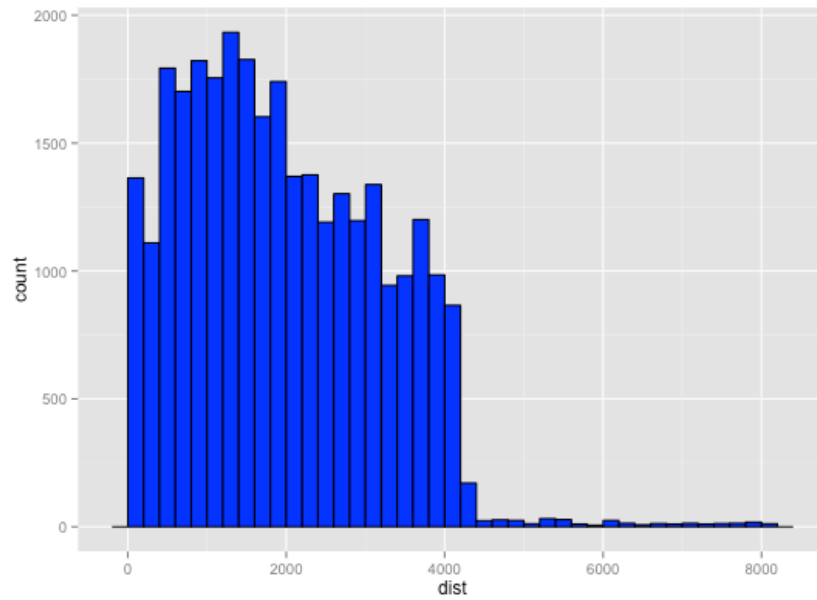
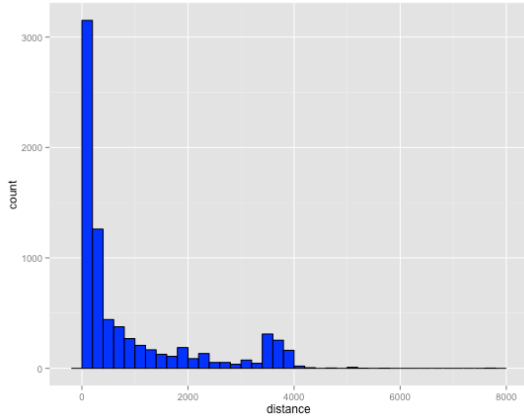


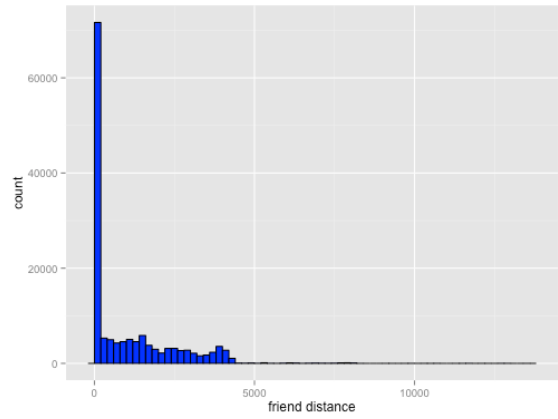
Figure 44. Histogram of distance between major cities (bin is 200 kilometers)

Chapter 6. Conclusion

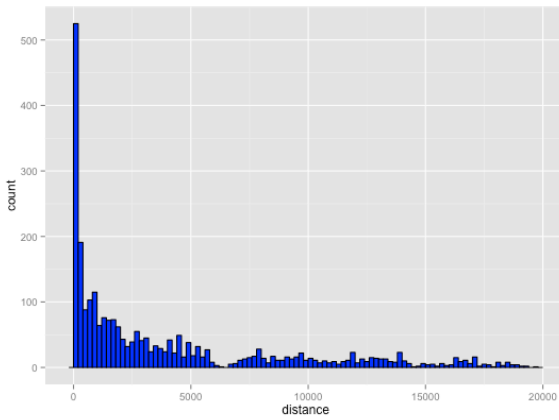
In this section, we will compare the distance in the three cyber communities and three physical communities. As shown in Figure 45, the left three distance histograms are corresponding to the cyber communities: Mason retweet community, Boston retweet community, and random location retweet community. The right three distance histograms are corresponding to the physical communities: Brightkite friendship community, Brightkite travel community, and the major cities in the United States. The distance histograms show a high peak where the distance is short in most communities. It is shown that geospatial distance still plays a significant role as most communications happen in short distance rather than long distance.



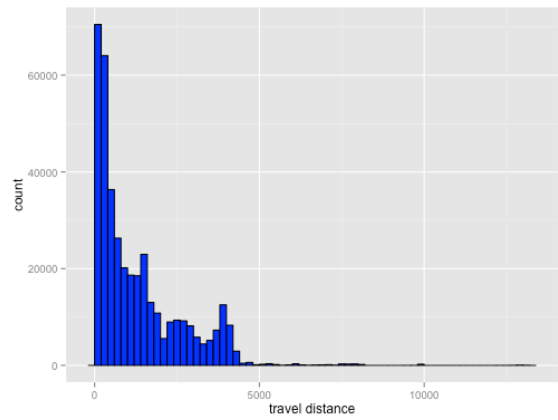
Mason Retweet Network



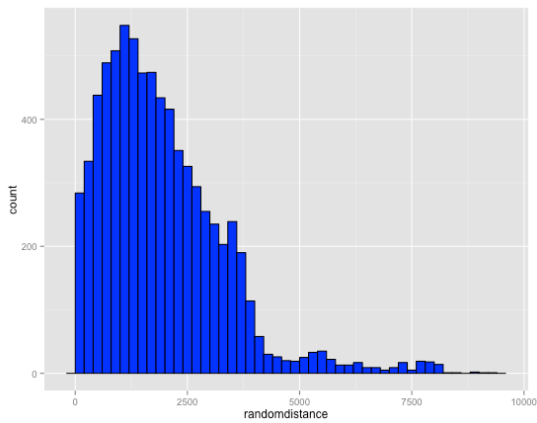
Brightkite Friendship Network



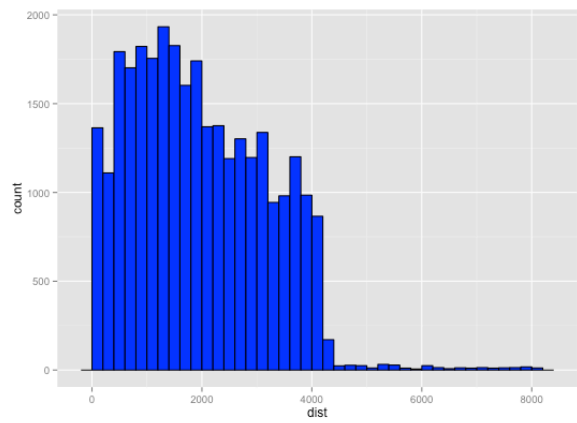
Boston Retweet Network



Brightkite Travel Network



Random Location Network



Major Cities Network

Figure 45. Comparison of distance histogram for cyber communities (left) and physical communities (right), bin is 200 kilometers

We ran Kolmogorov–Smirnov tests between each pair of the distance distributions. The result suggests that they are all of different distributions. The reason might be that the datasets have the different amount of samples. Statistically, the distance of the six communities are shown in Table 9 and Figure 46. When comparing the statistics of the distance distribution of the six communities, we can see that the Mason retweet community has the similar statistics as the Brightkite friendship community. It indicates that the Mason retweet network is a friendship-based community. It not only appears online in the cyber space, but also has bonds in the physical space.

The distance distribution of the random location retweet community and the urban community are very different from the other four datasets. They are similar to each other with most samples in short distance and very few samples for long distance. The random location retweet community is an artificial community consisted by randomizing the location of Mason retweets, and the randomization procedure is also based on the territory of the United States. It might be the reason that the random location community has similar distance distribution with the major cities in the United States.

When comparing the distance distribution of the Brightkite friendship communities and the populated cities, both communities reveal the distribution of population to some

extent, but the distributions are very different from each other. Brightkite friendship community is limited to the people who use Brightkite actively. Although Cho *et al.* (2011) attribute the friendship distance distribution to the non-uniform distribution of the population over the physical space, it might not be the sole reason based on our comparison.

Table 9. Distance of different communities (kilometers)

Communities	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Mason Retweet	0	17.34	222	844.8	1152	7774
Random Location	0	942.5	1677	1929	2648	9248
Boston Retweet	0	332.7	1806	3990	5574	19710
Brightkite Friendship	0	11.36	199.6	1010	1641	13460
Brightkite Travel	80	316.9	979.1	1330	1879	13160
Major Cities	5.395	966.1	1806	1962	2875	8176

The Boston retweet community has the longest distance than other communities, as other communities are US-focused while the Boston retweets reach out of the United States.

This community has very different distance distribution from the Mason retweet community and the Brightkite friendship community. It reveals that the Boston retweet community is not based on friendship. It is event-focused, so it is very transient and it dies out quickly. Such event-based news-focused cyber community does not have a solid physical basis so it did not last long.

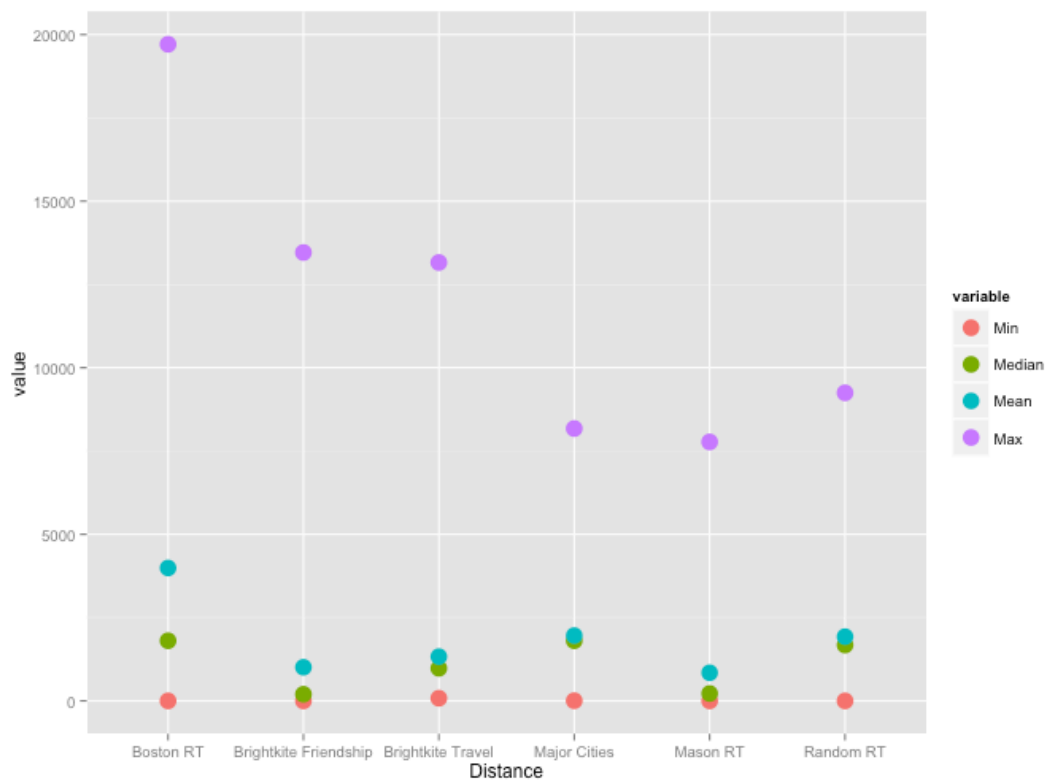


Figure 46. Comparison of distance in different communities

In conclusion, geospatial distance still plays a vital role in cyberphysical communities. Distance decay effect is found in all cyberphysical communities. By studying the distance distribution of different communities, we can gain insights of the nature of the communities, like whether it is friend-based or event-focused, whether it has physical connections to maintain its existence for a long period. It is intriguing to study more cyberphysical communities in the future to find typical distribution of different type of communities. Such work would greatly facilitate the research on cyberphysical communities.

References

- Andris, C. (2011). *Metrics and methods for social distance* (Doctoral dissertation, Massachusetts Institute of Technology).
- Bagrow, J., Wang, D., & Barabasi, A. (2011). Collective response of human populations to large-scale emergencies. *PloS one*, 6(3), 1–8. doi:10.1371
- Berry, H., Guillén, M. F., & Zhou, N. (2010). An institutional approach to cross-national distance. *Journal of International Business Studies*, 41(9), 1460–1480. doi:10.1057/jibs.2010.28
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Brown, C., Nicosia, V., Scellato, S., Noulas, A., & Mascolo, C. (2012). Where Online Friends Meet: Social Communities in Location-Based Networks. *ICWSM*, 415–418.
- Brown, C., Nicosia, V., Scellato, S., Noulas, A., & Mascolo, C. (2013). Social and place-focused communities in location-based online social networks. *arXiv preprint arXiv:1303.6460*.
- Bruns, A., & Burgess, J. (2011). # ausvotes: How Twitter covered the 2010 Australian federal election. *Communication, Politics and Culture*, 44, 37–56.
- Cairncross, F. (2001). *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press.
- Caverlee, J., Cheng, Z., Sui, D. Z., & Kamath, K. Y. (2013). Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media. *IEEE Data Eng. Bull.*, 36(3), 33–41.
- Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge*

- discovery and data mining* (pp. 1082–1090). New York, NY, USA: ACM.
doi:10.1145/2020408.2020579
- Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6), 066111.
- Culotta, A. (2010, July). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115-122). ACM.
- Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring Social Ties from Geographic Coincidences. *Proceedings of the National Academy of Sciences*, 107(52), 22436–22441. doi:10.1073/pnas.1006155107
- Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010). Bridging the gap between physical location and online social networks (pp. 119–128). ACM. doi:10.1145/1864349.1864380
- Croitoru, A., Stefanidis, A., Radzikowski, J., Crooks, A., Stahl, J., & Wayant, N. (2012). Towards a collaborative geosocial analysis workbench. In *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications* (p. 18). ACM.
- Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, 27(12), 2483-2508.
- Croitoru, A., Crooks, A., Radzikowski, J., Stefanidis, A., Vatsavai, R., & Wayant, N. (2014a). Geoinformatics and Social Media: A New Big Data Challenge. In *Big Data Techniques and Technologies in Geoinformatics* (H. Karimi, editor), CRC Press, pp. 207-232
- Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & A. Stefanidis, A. (2014b). Linking Cyber and Physical Spaces through Community Detection and Clustering in Social Media Feeds, *Computers, Environment & Urban Systems*. (in press)
- Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17(1), 124–147. doi:10.1111/j.1467-9671.2012.01359.x
- Crooks, A., D. Masad, D., Croitoru, A., Cotnoir, A., Stefanidis, A. & Radzikowski, J. (2014). International Relations: State-Driven and Citizen-Driven Networks, *Social Science Computer Review*. 32(2): 205-220.

- Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M., & Baum, S. (2012). Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social Networks*, 34(1), 6–17. doi:10.1016/j.socnet.2010.12.001
- Deuze, M. (2006). Participation, remediation, bricolage: Considering principal components of a digital culture. *The information society*, 22(2), 63-75.
- Elwood, S. (2009). Geographic information science: emerging research on the societal implications of the geospatial web. *Progress in Human Geography*, 34(3), 349–357. doi:10.1177/0309132509340711
- Elwood, S. (2010). Geographic Information Science: Visualization, visual methods, and the geoweb. *Progress in Human Geography*, 35(3), 401–408. doi:10.1177/0309132510374250
- Elwood, Sarah, & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, 42(1), 6–15. doi:10.1016/j.geoforum.2010.08.003
- Friedland, G., & Sommer, R. (2010, August). Cybercasing the Joint: On the Privacy Implications of Geo-Tagging. In *HotSec*.
- Gennip, Y., van Hunter, B., Ahn, R., Elliott, P., Luh, K., Halvorson, M., & Brantingham, P. J. (2013). Community detection using spectral clustering on sparse geosocial data. *SIAM Journal on Applied Mathematics*, 73(1), 67-83.
- Ghemawat, P. (2001). Distance still matters. *Harvard Business Review*, 79(8), 137–147.
- Goldenberg, J., & Levy, M. (2009). Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*.
- Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221.
- Gruzd, a., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an Imagined Community. *American Behavioral Scientist*, 55(10), 1294–1318. doi:10.1177/0002764211409378
- Hannigan, J., Hernandez, G., Medina, R. M., Roos, P., & Shakarian, P. (2013). Mining for Spatially-Near Communities in Geo-Located Social Networks. *arXiv preprint arXiv:1309.2900*.
- Hecht, B., & Moxley, E. (2009). Terabytes of Tobler: Evaluating the first law in a massive, domain-neutral representation of world knowledge. In *Spatial information theory* (pp. 88-105). Springer Berlin Heidelberg.

- Herrera, G. L. (2007). Cyberspace and Sovereignty: Thoughts on physical space and digital space. Power and security in the information age: Investigating the role of the state in cyberspace, 67-94.
- Jones, Q. (1997). Virtual-Communities, Virtual Settlements & Cyber-Archaeology: A Theoretical Outline. *Journal of Computer-Mediated Communication*, 3(3), 0-0.
- Kaltenbrunner, A., Scellato, S., Volkovich, Y., Laniado, D., Currie, D., Jutemar, E. J., & Mascolo, C. (2012). Far from the eyes, close on the Web: impact of geographic distance on online social interactions. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks* (pp. 19-24). ACM.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), 59-68.
- Kwan, M. (2007). Mobile Communications, Social Networks, and Urban Travel: Hypertext as a New Metaphor for Conceptualizing Spatial Interaction*. *The Professional Geographer*, 59(April), 434–446.
- Lu, X., Wang, C., Karamzadeh, N., Croitoru, A., & Stefanidis, A. (2011). Deriving implicit indoor scene structure with path analysis. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness* (pp. 43–50). New York, NY, USA: ACM.
doi:10.1145/2077357.2077367
- MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J., & Mitra, P. (2011, July). Geo-twitter analytics: Applications in crisis management. In *Proceedings, 25th International Cartographic Conference, Paris, France*.
- Matthews, J. L., & Matlock, T. (2011). Understanding the Link Between Spatial Distance and Social Distance. *Social Psychology*, 42(3), 185–192. doi:10.1027/1864-9335/a000062
- McAlexander, J. H., Koenig, H. F., & Schouten, J. W. (2006). Building relationships of brand community in higher education: a strategic framework for university advancement. *International Journal of Educational Advancement*, 6(2), 107-118.
- Mok, D., Wellman, B., & Carrasco, J. (2010). Does distance matter in the age of the Internet?. *Urban Studies*, 47(13), 2747-2783.
- Musolesi, M., & Mascolo, C. (2006). A Community Based Mobility Model for Ad Hoc Network Research Categories and Subject Descriptors, 31–38.

- Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An Empirical Study of Geographic User Activity Patterns in Foursquare. *ICWSM, 11*, 70-573.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America, 103*(23), 8577–82. doi:10.1073/pnas.0601602103
- Newman, M. (2009). *Networks: an introduction*. Oxford University Press.
- Openshaw, S. (1983). *The modifiable areal unit problem* (Vol. 38). Norwich: Geo books.
- Porter, C. E. (2004). A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. *Journal of Computer-Mediated Communication, 10*(1), 00–00. doi:10.1111/j.1083-6101.2004.tb00228.x
- Qi, G. J., Aggarwal, C. C., & Huang, T. S. (2013, February). Online community detection in social sensing. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 617-626). ACM.
- Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., ... & Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PloS one, 5*(12), e14248.
- Ressler, S. (2006). Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs, 2*(2), 1-10.
- Ritterman, J., Osborne, M., & Klein, E. (2009, November). Using prediction markets and Twitter to predict a swine flu pandemic. In *1st international workshop on mining social media* (Vol. 9).
- Robertson, R. (1995). Glocalization: Time-space and homogeneity-heterogeneity. *Global modernities, 25-44*.
- Rothenberg, R., Muth, S. Q., Malone, S., Potterat, J. J., & Woodhouse, D. E. (2005). Social and Geographic Distance in HIV Risk: *Sexually Transmitted Diseases, 32*(8), 506–512. doi:10.1097/01.olq.0000161191.12026.ca
- Sadilek, A., Kautz, H., & Bigam, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 723–732). New York, NY, USA: ACM. doi:10.1145/2124295.2124380

- Scellato, S., Mascolo, C., Musolesi, M., & Latora, V. (2010). Distance matters: geo-social metrics for online social networks. In *Proceedings of the 3rd conference on Online social networks* (pp. 8-8). USENIX Association.
- Scellato, S., & Mascolo, C. (2011). Measuring user activity on an online location-based social network. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on* (pp. 918–923). doi:10.1109/INFCOMW.2011.5928943
- Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-Spatial Properties of Online Location-Based Social Networks. *ICWSM, 11*, 329-336.
- Stefanidis, A., Cotnoir, A., Croitoru, A., Crooks, A., Rice, M., & Radzikowski, J. (2013a). Demarcating New Boundaries: Mapping Virtual Polycentric Communities through Social Media Content, *Cartography and Geographic Information Science*, 40(2): 116-129.
- Stefanidis, A., Crooks, A., & Radzikowski, J. (2013b). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2): 319-338
- Sugumaran, R., & Voss, J. (2012, July). Real-time spatio-temporal analysis of west nile virus using twitter data. In *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications* (p. 39). ACM.
- Sutton, J., Palen, L., & Shklovski, I. (2008). Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires. In *Proceedings of the 5th International ISCRAM Conference* (pp. 624-632). Washington, DC.
- Takhteyev, Y., Gruz, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, 34(1), 73–81. doi:10.1016/j.socnet.2011.05.006
- Tayebi, M. A., Frank, R., & Glässer, U. (2012, November). Understanding the link between social and spatial distance in the crime world. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* (pp. 550-553). ACM.
- Tobler W., (1970) "A computer movie simulating urban growth in the Detroit region". *Economic Geography*, 46(2): 234-240.
- Torrens, P. M. (2010). Geography and computational social science. *GeoJournal*, 75(2), 133–148. doi:10.1007/s10708-010-9361-y

- Tumasjan, a., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*, 29(4), 402–418. doi:10.1177/0894439310386557
- Valli, C., & Hannay, P. (2010). Geotagging Where Cyberspace Comes to Your Place. In *Security and Management* (pp. 627-632).
- Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (pp. 1-10). ACM.
- Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1079-1088). ACM.
- Volkovich, Y., Scellato, S., Laniado, D., Mascolo, C., & Kaltenbrunner, A. (2012). The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions. In *ICWSM*.
- Wellman, B. (2001). Physical place and cyberplace: The rise of personalized networking. *International journal of urban and regional research*, 25(2), 227-252.
- Wu, H., & Wang, W. (2013, June). Identifying the Daily Activity Pattern of Community Dynamics Using Digital Footprint. In *Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on* (pp. 782-785). IEEE.
- Zignani, M., Gaito, S., & Rossi, G. (2012). Extracting human mobility and social behavior from location-aware traces. *Wireless Communications and Mobile Computing*.
- Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2(2), 6–32. doi:10.2202/1948-4682.1069

BIOGRAPHY

Xu Lu graduated from Beijing No.2 Middle School, Beijing, China, in 2004. She received her Bachelor of Science from Peking University in 2008. She received her Master of Science in Cartographic and Geographic Sciences from George Mason University in 2012.