

To the Graduate Council:

I am submitting herewith a thesis written by Sumang Liu entitled “Online Social Network Friends and Spatio-temporal Proximity of Their Geotagged Photos – A Case Study of Flickr Data.” I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Geography.

Shih-Lung Shaw, Major Professor

We have read this thesis
and recommend its acceptance:

Bruce Ralston

Liem Tran

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

**Online Social Network Friends and Spatio-
temporal Proximity of Their Geotagged Photos –
A Case Study of Flickr Data**

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Sumang Liu
December 2012

Copyright © 2012 by Sumang Liu.
All rights reserved.

ACKNOWLEDGEMENTS

I am grateful to the dedicated instructions from Dr. Shih-Lung Shaw. During the past two years, Dr. Shaw gave me many valuable instructions and ideas on my thesis project. He did not only inspire my interests in online social network analysis, but also led me to the right track of academic research. He also helped me with thesis writing. Although this work is still not perfect, I have accumulated valuable experience from it.

Many thanks go to Dr. Bruce Ralston and Dr. Liem Tran. During the past two years, I received a lot of help from them. Every time I walked into their office with my strange questions, they welcomed me and answered these questions very patiently. These instructions helped me go through my projects smoothly.

I also appreciate my beloved girl friend, Kris Zhang, for her support during my hardest time.

ABSTRACT

This empirical study aims to analyze relationships between online social network (OSN) friends and spatio-temporal proximity of their geotagged photos, using Flickr data as a case study. First, this study analyzes whether Flickr friends tend to post geotagged photos that are closer to each other compared to Flickr non-friends in space and time. Second, this study investigates whether the number of geotagged photos posted by users is related to the distance and time difference between their geotagged photos. Third, this study examines the spatial distributions of geotagged photos of Flickr friends within specific distance intervals to further understand the geographic meanings of Flickr user's geotagging activities. Findings of this study can improve our understanding of the relationship between users' virtual friendships and their physical activities. These understandings can support future research, including location-based services, location-based OSN searches, and location-based online marketing.

Keywords: online social network, spatio-temporal proximity, geotagged photos, online friendship

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION	1
CHAPTER 2 LITERATURE REVIEW	4
CHAPTER 3 RESEARCH OBJECTIVES AND RESEARCH QUESTIONS	10
3.1 Research Objectives	10
3.2 Research Questions	10
CHAPTER 4 METHODOLOGY	15
4.1 Data acquisition	15
4.2 Data processing	17
4.3 Frequency analysis on the spatio-temporal proximity of geotagged photos	19
4.4 Visualize the spatial distribution of geotagged photos	21
CHAPTER 5 FINDINGS AND INTERPRETATIONS	24
5.1 General spatial distribution of Flickr geotagged photos	24
5.2 Relationships between online friendships and the spatio-temporal proximity of their geotagged photos	26
5.3 Spatial visualization of geotagged photos: where are they and what happened?	49
CHAPTER 6 CONCLUSION AND FUTURE RESEARCH	65
6.1 Conclusions	65
6.2 Limitations and Future Research	67
LIST OF REFERENCES	69
APPENDIX	74
VITA	76

LIST OF TABLES

Table 1. The four main date granularities of Flickr geotagged photos	16
Table 2. Summaries of the two-sample Kolmogorov-Smirnov test.....	27
Table 3. Hypothesis test summary of the two-sample Kolmogorov-Smirnov test	28
Table 4. Comparison of the closest pair distances between Flickr friends and non-friends.....	28
Table 5. Summaries of the two-sample Kolmogorov-Smirnov test.....	30
Table 6. Hypothesis test summary of the two-sample Kolmogorov-Smirnov test	31
Table 7. Comparison of the nearest neighbor distances between Flickr friends and non-friends.....	31
Table 8. Summaries of the two-sample Kolmogorov-Smirnov test.....	34
Table 9. Hypothesis test summary of the two-sample Kolmogorov-Smirnov test	35
Table 10. The number of closest photo pairs within different distance ranges and one day of Flickr friends.....	35
Table 11. The number of closest photo pairs within different distance ranges and one day of Flickr non-friends.....	36
Table 12. Time difference between the closest photo pairs of Flickr friends.....	38
Table 13. Time difference between the closest photo pairs of Flickr non-friends	38
Table 14. The proportion of the closest photo pairs within some typical spatial and temporal thresholds	40
Table 15. Frequency distributions of the four kinds of distances	41
Table 16. Spatial distribution of the closest photo pairs between Flickr friends between 4,200 km and 5,000km.....	42
Table 17. Correlation analysis between the closest pair distance and the higher photo number.....	47
Table 18. Correlation analysis between the closest pair distance and the lower photo number.....	48
Table 19. The ten highest ratios between the frequencies of the closest photo pairs within 10 km of Flickr friends and non-friends in urban areas of the lower 48 states.....	54
Table 20. The ten highest frequencies of the closest photo pairs within 10 km in urban areas.....	55
Table 21. The ten highest frequencies of the closest photo pairs within 10 km in national parks	61

LIST OF FIGURES

Figure 1a. Co-location in time	Figure 1b. Co-location in space	Figure 1c. Co-existence.....	8
Figure 2. Geotagged photo points of two Flickr users.....			11
Figure 3. The nearest neighbor photo distances between a pair of Flickr friends			12
Figure 4. The closest two geotagged photos between a pair of Flickr users.....			13
Figure 5. Existing and new geotagged photos from two Flickr users			14
Figure 6. Date granularity of downloaded Flickr geotagged photos			16
Figure 7. Geotagged photo records of user “38795929@N00”			17
Figure 8. A sample visualization of the closest photo pairs between 300 km to 350 km in the lower 48 states of the U.S.			22
Figure 9. Histogram of the numbers of Flickr users’ geotagged photos			24
Figure 10. Spatial frequency of Flickr geotagged photos around the world			25
Figure 11. Cumulative proportion of the closest pair distances between Flickr friends.....			26
Figure 12. Cumulative proportion of the closest pair distances between Flickr non-friends.....			27
Figure 13. Cumulative proportion of the nearest neighbor distances between Flickr friends			29
Figure 14. Cumulative proportion of the nearest neighbor distances between Flickr non-friends			30
Figure 15. Time differences between the closest photo pairs of Flickr friends....			33
Figure 16. Time differences between the closest photo pairs of Flickr non-friends			33
Figure 17. Time differences between the closest photo pairs of Flickr friends and non-friends.....			34
Figure 18. The closest photo pairs between 4,200 km and 4,600 km of Flickr friends at the West Coast of Europe			43
Figure 19. The closest photo pairs between 4,200 km and 4,600 km of Flickr friends at the East Coast of the U.S.....			44
Figure 20. The closest photo pairs between 4,200 km and 4,600 km of Flickr non-friends at the West Coast of Europe			44
Figure 21. The closest photo pairs between 4,200 km and 4,600 km of Flickr non-friends at the East coast of the U.S.			45
Figure 22. The closest photo pairs within 10 km in the lower 48 states			50
Figure 23. Frequency map of the closest photo pairs within 10 km of Flickr friends in urban areas of the lower 48 states			51
Figure 24. Frequency of the closest photo pairs within 10 km of Flickr non-friends in urban areas of the lower 48 states.....			52
Figure 25. The ratios between the frequencies of the closest photo pairs within 10 km of Flickr friends and non-friends for urban areas in the lower 48 states			53
Figure 26. The closest photo pairs within 10 km in San Francisco, CA			57

Figure 27. The closest photo pairs within 10 km and the population density in San Francisco, CA.....	58
Figure 28. The closest photo pairs within 10 km in Manhattan, NY	59
Figure 29. The closest photo pairs within 10 km and the population density in Manhattan, NY	60
Figure 30. The closest photo pairs between 300 km and 550 km of Flickr friends in the U.S.	63
Figure 31. The closest photo pairs between 300 km and 550 km of Flickr non-friends in the U.S.	64

CHAPTER 1

INTRODUCTION

The rapid growth of Online Social Networks (OSN) has attracted public attention to the burgeoning online communities built upon information and communication technology. By the end of 2011, the registered user accounts of Facebook, Twitter, and Flickr were 810 million (<http://www.facebook.com>), 510 million (<https://twitter.com>), and 51 million (<http://www.flickr.com>), respectively. OSN users communicate with each other by posting, commenting, and messaging. The large population of OSN provides the opportunity to obtain individual-based data with unprecedented depth and scale. With the development of location-based services and the popularity of ubiquitous devices, OSN users are now able to share their physical locations online. There are several ways in which users may do so. The registered location in a user's profile page releases basic information about where he/she works, lives, or studies. However, due to privacy issues the location information on users' profile pages is more often than not unavailable to the public. Geotagged photo provides another way to share user's spatio-temporal activities online. Some of these data are publicly available, and therefore attract many researchers.

The potential relationship between OSN friendships and spatio-temporal proximity of their geotagged photos is a widely concerned topic among researchers. For classic social networks, it is believed that "geography and social relationships are inextricably intertwined" (Backstrom et al., 2010, p. 61). Previously, sociologists and geographers found that geographic proximity had a powerful influence on the formation of social ties (Milgram, 1967; Killworth and Bernard, 1978; Dodds et al., 2003; Lewis et al., 2008). For OSN, however, there is an absence of knowledge about relationships between users' virtual friendship and their physical activities in space and time. Building an understanding of whether OSN friends tend to geotag their photos that are closer to each other in space and time compared to OSN non-friends can significantly benefit this area of research. In a study of OSN privacy, Backstrom et al. (2010) argued that the knowledge of the relationship between users' friendship and their geotagged posts can

help “infer” the structure of OSN. In a heuristic geographic search through OSN, Adamic and Adar (2008) studied the location-based search through “Club Nexus”, a small online student network at Stanford University. Their search tried to set up the “acquaintance chain” between Club Nexus users based on the geographic proximity between user-uploaded location data. However, their search did not work effectively due to a lack of knowledge of the relationship between students’ friendship and their locations. In a friendship inference experiment, Crandall et al. (2009) used “spatio-temporal co-occurrence”, which refers to two persons existing at approximately the same location and the same time, to “infer” the friendship between Flickr users. However, only a very small proportion of Flickr users have enough “spatio-temporal co-occurrences” revealed by their geotagged photos to make a convincing inference. This lack of “spatio-temporal co-occurrences” between Flickr friends raises the concern of whether OSN friendships and the spatio-temporal proximity of their geotagged photos are actually related.

The aforementioned projects reveal the research potentials based on the relationships between online friends and spatio-temporal proximity of their geotagged photos. However, some distinct features of geotagged photos and online friendships challenge the researchers to further understand these relationships. For geotagged photos, a user may geotag many photos at different locations, but none of them can be directly interpreted as users’ residential location. A user may geotag more photos where he/she travels than where he/she lives. Thus, it is possible for users who live further away to geotag photos closer in space. Moreover, some users are more enthusiastic in geotagging photos than others. Users may have different numbers of geotagged photos at different frequency which record different aspects of their lives. Furthermore, very few OSN users record their daily routine with geotagged photos. Hence, large volumes of geotagged photos from many OSN users do not imply the completeness of any single user’s spatio-temporal activities. These quality issues of geotagged photos challenge the feasibility of geotagged photos as an appropriate data source for human activity studies. For online friendships, Boyd and Crawford (2011) suggested that OSN could be characterized as an “articulated social network” or a “behavior social network”. The concept of an articulated social network meant that friendship was explicitly filtered and specified by users (Lewis et al., 2008). The concept of a behavior social network meant

that friendship was revealed by social interactions, such as wall posting and status commenting. However, neither of these friendship networks compromised users' complete social connections. For example, users may be colleagues, classmates, or relatives but do not list each other as friends on OSN for any number of reasons. Consequently, researchers should be careful when dealing with any "missing connection" between OSN users. Furthermore, though the populations of mainstream OSNs are large, they can hardly represent the general population of the world. For example, Crandall et al. (2010) admitted that in using Flickr as a dataset, they had access "by definition only to the behavior to its users, who are a small and not necessarily representative sample of broader population" (p. 22440).

Given the benefits and challenges of OSN data, fascinating network analysis still awaits researchers (Boyd and Crawford, 2011). In the GIS community, Sui and Goodchild (2011) were optimistic about integrating GIS into the analysis of OSN data. They encouraged researchers to explore new ways in which "the fusion of GIS with social media could be deployed to promote the human-as-sensor paradigm in spatial-data generation."

Geotagged photos from OSN users may not imply completeness of single user's spatio-temporal activities. Friendship connections on OSN may not comprise users' complete social connections. However, there may still be some relationships between OSN users' friendship and their geotagging activities which could benefit many fields of research. In this case, an empirical study with a large volume of geotagged photos and online friendships from a mainstream OSN website can help consolidate our understanding of these relationships. This study addresses the necessity of such a kind of empirical analysis. Using Flickr as a case study, it applies a data-intensive analysis to explore the relationships between online friendships and the spatio-temporal proximity of their geotagged photos.

CHAPTER 2

LITERATURE REVIEW

Publicly available data from OSN services have emerged as a milestone of “Big Data Era” (Boyd and Crawford, 2010, p. 02), where large volumes of digital traces from individuals are disclosed and deposited by themselves. “Big Data,” also called “data avalanche” (Miller, 2010, p. 181) or “exaflood” (Sui and Goodchild, 2011, p. 1742), is a relatively broad concept related to most computational intensive studies. OSN data, composed of a large number of OSN users, is one source of “Big Data.” Though a single user’s activities on Facebook, Twitter, Flickr or other OSN services may not contain strong clues of collective importance, a large set of activities extracted from a ‘crowd’ may indicate strong collective knowledge which is worth directing to interested users (Caverlee, 2010). Therefore, OSN data act as an integral part of the prospective web which broadcasts signals at both individual and societal levels (Sui, 2010).

Geotagged OSN data demonstrate some distinct spatio-temporal characteristics of social networks/interactions in the age of Web 2.0 (Elwood, 2010). On one hand, they reveal a new spatial turn in social media, which reflects Tobler’s first law of geography that everything is connected to everything else (Sui and Goodchild, 2011). On the other hand, they stress the importance of the temporal aspect in social interactions. To address these spatio-temporal features in social media, Adams (2009) and Sui (2010) introduced an analytical framework that consists of perspectives of space and place, coding and representation, and spatial organization. Based on spatial and temporal features of OSN data, many studies have been carried out from different perspectives.

Spatio-temporal data of OSN have been applied to analyze geographic meanings of human activities. User-uploaded spatio-temporal activities are conducted within specific geographic contexts. Analyzing the patterns of users’ spatio-temporal activities can therefore help us evaluate the geographic contexts behind them. For example, Crandall (2008) studied the spatial distribution of Flickr geotagged photos to define a relational structure between popular places on Flickr. Ratti et al. (2007) and Girardin et al. (2008)

visualized the digital footprints of tourists using their geotagged photos to illustrate spatio-temporal tourist flows. Their findings helped define the tourism hot-spots and cluster the tourist routes. Forsyth (2010) investigated spatio-temporal distribution of OSN geotagged photos and found that region boundaries could be reshaped dynamically by different OSN communities. Ahern et al. (2007) also claimed that the geographic boundaries derived from user-uploaded geotagged data are imprecise. Hollenstein and Purves (2010) tested these hypotheses by analyzing how large numbers of Flickr users name the city cores through eight million geotagged photos. Their findings provided new evidences that geographies of Flickr users' geotagged photos are not often captured by administrative representations. These projects reveal some specific geographic meanings behind the spatio-temporal activities derived from OSN users' geotagged data. However, most of these projects do not take online friendship into consideration.

Spatio-temporal data of OSN can improve location-awareness service. As claimed by Backstrom et al. (2010), "geography has a number of compelling applications within Internet technology, and accurately predicting a user's location can significantly improve a user's experience" (p. 61). Although location-awareness functions of OSN are eye-catching, the majority of OSN users adopt them very slowly and hesitantly. For example, in a test over 1 million Twitter users, the percentage of users who geotagged at least one Tweet at the city level was only 26%, and the percentage of Tweets which were geotagged was only 0.42% (Caverlee, 2010). Caverlee (2010) referred to this as a location sparsity problem. To address it, he tried to automatically estimate a user's location by analyzing the publicly-available spatio-temporal data from OSN users. He found that the location estimates converged quickly, placing 51% of Twitter users within 100 miles of their actual location. Findings from his studies are expected to lead to broader innovations in many fields, such as emergency management and infectious diseases control.

The aforementioned research demonstrates the great potential in focusing on spatio-temporal features of OSN data. Integrating spatio-temporal features of OSN with social features of OSN expands the horizon of the spatio-temporal analysis of OSN data. For

example, the relationship between geography and online friendship has concerned many scholars (Gilbert et al., 2008; Backstrom, et al., 2010; Liben-Nowell, et al., 2005). Gilbert et al. (2008) categorized MySpace users as rural or urban users according to their residence locations and pointed out that urban users tended to have friends that were more scattered throughout the country. Backstrom et al. (2010) studied the locations of Facebook users from their profiles and observed an inverse relationship between distance and friendship at medium to long-range distances. For shorter distance ranges, they did not observe a strong impact of distance on friendship. In contrast, Liben-Nowell et al. (2005) studied the geographic and social proximity of OSN users to find a baseline probability of geographic independent relations between the likelihood of friendship and the extremely long distances.

The aforementioned projects provide a general view of the relationships between online friends and their geographic locations. Their findings, however, are limited: First, they evaluated geographic proximity between OSN users through residence locations reported in user-profiles. Most of time a user chooses only one place as his/her residence on his/her profile page. Therefore, the location of each user in these projects is fixed. Since user profiles are usually protected by privacy restrictions, most researchers are unable to access them. In other words, these conclusions are less useful for most researchers who base their studies on publicly available spatio-temporal data, such as geotagged posts or geotagged photos. Second, most of these projects did not address time, an important aspect of human activity. The amount of temporal data obtained from user profiles is very limited. As a result, researchers have limited temporal information to conduct effective temporal or spatio-temporal analyses of online friendship. In comparison, geotagged photos provide more temporal information. It is therefore important to analyze the geotagged posts to establish a better understanding of the relationships between online friendships and their spatio-temporal proximity.

The knowledge of relationships between OSN friends and the spatio-temporal proximity of their geotagged posts can benefit many research areas. For example, in the studies of location-based search through OSN many researchers addressed the question of how OSN strangers were able to find short paths to connect each other using only

geographic information about their immediate contacts (Adar and Adamic, 2008; Kleinberg, 2000; Watts et al., 2002). They assumed that online friendship and the spatial proximity of user-uploaded geotagged data are related and based their geographic search on that assumption. However, this assumption needs further verification. There are embedded weaknesses of geotagged posts and online friendships. Objectivity, accuracy, accessibility, equity, and ethicality of OSN data are all venerable areas which have been questioned by many researchers (Boyd and Crawford, 2002; Sui and Goodchild, 2011). These deeply entangled challenges exposed in existing research await further exploration. For example, Crandall et al. (2009) found a high correlation between Flickr friendships and the spatio-temporal proximity of their geotagged photos. They used the phrase “spatio-temporal co-occurrence” to refer to two persons existing at approximately the same location and approximately the same time. They observed that if two Flickr users took photos within 24 hours and 100 kilometers on at least five occasions and at five distinct geographic locations, there was a 59.8% chance that they were Flickr contacts.

However, the methodologies and the findings of this study are still limited. First, they divided the world into a grid to detect spatio-temporal co-occurrences and adjusted their spatio-temporal thresholds in an arbitrary way. Though they obtained a relatively high rate of friendship (59.8%) within a specific spatio-temporal threshold (100 km), their particular choices of spatio-temporal thresholds were not strongly justified. Second, in their study, most Flickr friends have few “spatio-temporal co-occurrence” of geotagged photos. For example, only 1.5% of all friendships in their analysis had at least one co-occurrence in a $100 \times 100 \text{ km}^2$ area within one day and only 0.03% of all friendships had three such co-occurrences. They concluded that “most friendships did not reveal themselves through a pattern of repeated spatio-temporal co-occurrences” (Crandall, et al., 2010, p. 22440). In this case, more empirical studies based on large volumes of geotagged photos and online friendships are needed to further investigate how the spatio-temporal proximity of geotagged photos relates to online friendships.

The framework of time geography provides a potential perspective to analyze the relationship between online friendships and the spatio-temporal proximity of their

geotagged photos. Time geography was introduced by Hägerstrand (1970) to analyze human activities under different types of constraints. Space-time path and space-time prism are two useful tools in time geography (Hägerstrand, 1970). Space-time path connects an individual's activities at different locations according to their temporal order, while space-time prism delimits the possible locations that an individual can visit within specific space-time constraints (Hägerstrand, 1970; Lenntorp, 1976; Shaw, 2010). To illustrate the relationships between activities of different individuals, Yu and Shaw (2006) summarized three classic relationships of space-time paths: co-location in time, co-location in space and co-existence. Co-location in time represents activities in different space-time paths that interact with each other within a common time window (see Figure 1a). Co-location in space occurs when activities in different space-time paths occupy the same location in different time windows (see Figure 1b). Co-existence describes the cases when activities take place at the same location and within a common time window (see Figure 1c). The spatio-temporal proximity between geotagged photos may also follow these three typical relationships. The “spatio-temporal co-occurrence”, as discussed above, is one example of “co-existence” of individuals reflected by their geotagged photos.

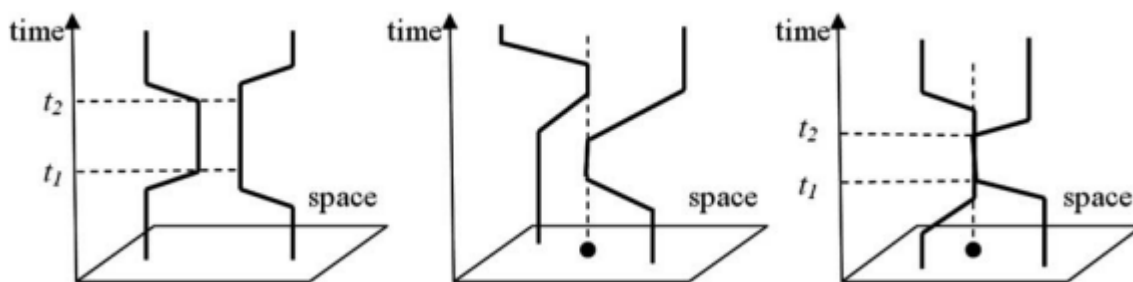


Figure 1a. Co-location in time Figure 1b. Co-location in space Figure 1c. Co-existence

Time geography concepts, together with time geography analytical tools, have been implemented in geographic information systems (GIS) to manage activity and travel diary data (Shaw and Wang, 2000; Wang and Cheng, 2001; Fridiha et al., 2002, 2004; Buliung and Kanaroglou, 2004). However, with the development of information and communication technology, time geographers also noticed the existing gap between classical time-geographic framework and the virtual activities and interactions

conducted via information and communication technologies (Shaw, 2010; Miller 2005). To further extend classic time geography to accommodate the needs of representing and analyzing activities and interactions in a hybrid physical-virtual space, Shaw and Yu (2009) presented a space-time GIS design that was capable of organizing complex activity and interaction data as spatio-temporal processes in an integrated space-time environment. This design helped researchers manage, analyze, and visualize individual activities and interactions in both physical and virtual spaces (Shaw et al. 2008; Shaw and Yu 2009, Yu and Shaw 2008). However, our understanding of the relationships between physical space and virtual space remains limited. For example, Adams (1995) questioned the value of mapping population distribution inside a city due to the potential existence of virtual space linking people from different cities. Additional empirical studies are needed to examine the potential interactions between physical and virtual activities (Shaw, 2010). The research addressed in this paper, which aims to investigate the relationships between online friends and the spatio-temporal proximity of their geotagged photos, is one such empirical study.

CHAPTER 3

RESEARCH OBJECTIVES AND RESEARCH QUESTIONS

3.1 Research Objectives

In focusing on the relationships between online social network (OSN) friends and spatio-temporal proximity of their geotagged photos, the objectives of this empirical study are unfolded through four steps: First, it analyzes whether Flickr friends tend to post geotagged photos that are closer to each other in space than Flickr non-friends do. Second, it investigates the temporal relationships between geotagged photos of Flickr friends. Third, it examines the potential relationship between spatio-temporal proximity of users' geotagged photos and the number of geotagged photos posted by them. Fourth, it visualizes the geographical distributions of geotagged photos of Flickr friends and Flickr non-friends to further evaluate the corresponding geographical meanings. Flickr, an online photo sharing service with social network functions, is chosen for this case study. Flickr enables users to geotag their photos at various spatio-temporal precision levels. Each geotagged photo has geographic coordinates and a time stamp. Additional features of Flickr data will be discussed in detail in the next chapter.

3.2 Research Questions

3.2.1 Do Flickr friends tend to post geotagged photos that are closer to each other in space compared to Flickr non-friends do?

This research question focuses on the spatial proximity of Flickr geotagged photos. Distance between geotagged photos is an indicator of their spatial proximity. In order to investigate whether the spatial proximity of geotagged photos is influenced by online friendship, distances between geotagged photos of Flickr friends and Flickr non-friends are examined respectively. Though only around 20% of Flickr users use geotagging functions, most of them have more than one geotagged photo. In most cases, there is a collection of distances between the geotagged photos of a pair of Flickr users who use

geotagging functions. Figure 2 provides a hypothetical example of geotagged photo points of two Flickr users.

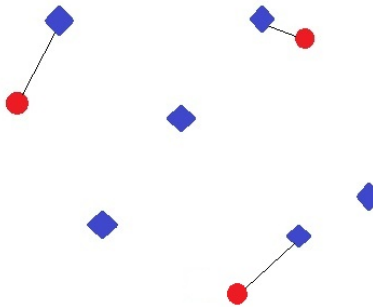


Figure 2. Geotagged photo points of two Flickr users

In Figure 2, the three red dots refer to geotagged photos of User A, while the six blue squares refer to geotagged photos of User B. In order to investigate how proximate the geotagged photos of User A are to the geotagged photos of User B, three pairs of nearest neighbors are identified from point set A (red dots) to point set B (blue squares). The nearest neighbors identified from point set A to point set B can be different from the nearest neighbors identified from point set B to point set A. Distances between all pairs of nearest neighbors provide an overall view of the spatial proximity between geotagged photos of a pair of Flickr users¹. It is termed “**overall proximity**”. The distribution of distances in Figure 3 provides an example of the overall proximity between a pair of Flickr friends (user “113775914@N02” and user “97458541@N03”).

¹Distances are calculated as Great Circle Distance. See the Appendix for details.

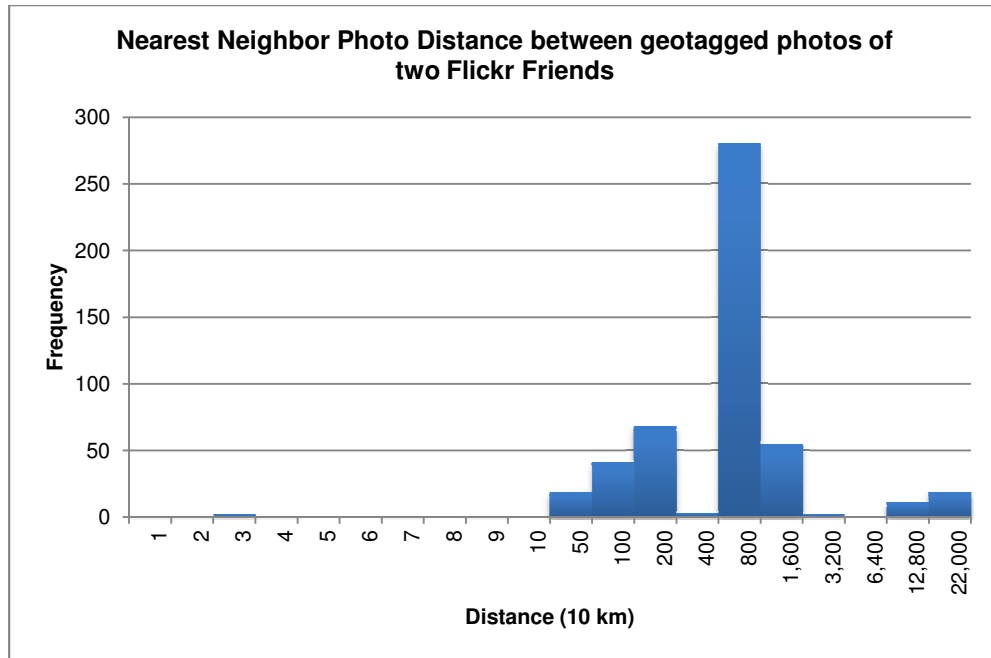


Figure 3. The nearest neighbor photo distances between a pair of Flickr friends

In Figure 3, there are 487 samples of the nearest neighbor photo distances from user “113775914@N02” to user “97458541@N03”. The highest frequency appears at around 800 km. To get a more comprehensive understanding of the overall proximity of Flickr friends, the nearest neighbor photo distances from many pairs of Flickr friends are compared with those of Flickr non-friends.

The overall proximity provides us a general view about how geotagged photos of two Flickr users are close to each other in space. Another informative view to evaluate the spatial proximity between a pair of Flickr users through their geotagged photos involves the closest two geotagged photos between them, as shown in Figure 4.

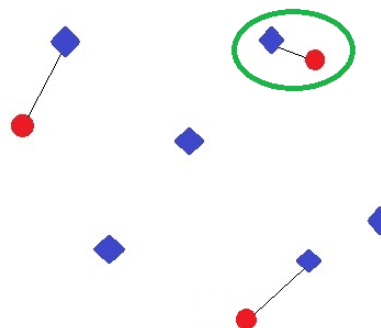


Figure 4. The closest two geotagged photos between a pair of Flickr users

In Figure 4, the red dots refer to geotagged photos of user A, while the blue squares refer to geotagged photos of user B. Three pairs of nearest neighbor photos between User A and User B are connected with black lines. Among them, the two photo points highlighted by green circle are the closest two photo points between User A and User B. The distance between these two points is termed the “**closest pair distance**”. The spatial proximity evaluated by the closest pair distance is termed the “**the closest pair proximity**”. The corresponding question is whether the closest pair distances of Flickr friends are shorter than those of Flickr non-friends.

3.2.2 Is the time difference between the closest photo pair of Flickr friends shorter than that of Flickr non-friends?

In the previous research question, the closest photo pairs are analyzed to evaluate the spatial proximity between Flickr friends. Adding temporal analysis of these photo pairs can further establish a view of how proximate the geotagged photos posted by Flickr friends can be in both space and time. Time differences can assist in investigating the temporal relationship between geotagged photos and online friendship. The corresponding research question is whether the time difference between the nearest neighbor photo pairs or the closest photo pairs of Flickr friends is shorter than that of Flickr non-friends.

3.2.3 Is the closest pair distance between Flickr friends related to the number of geotagged photos posted by them?

Different Flickr users have different numbers of geotagged photos. Figure 5 shows the geotagged photos from Flickr users A and B. The red dots represent the photos which user A has already geotagged. The blue squares represent the photos which user B has already geotagged. The orange dots represent the new photos which user A will geotag. The green circle points out the closest pair distance between the red dots and the blue squares.

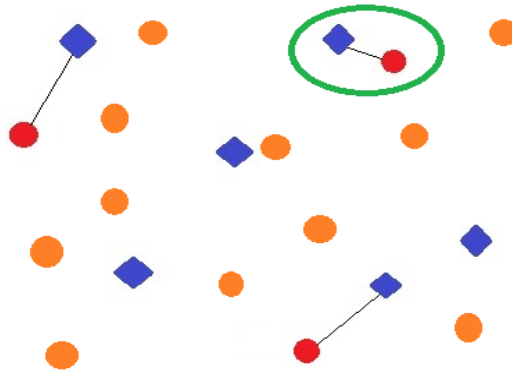


Figure 5. Existing and new geotagged photos from two Flickr users

When user A geotags more photos (the orange dots), the closest pair distance between A and B will either decrease or stay the same. In other words, geotagging more photo points creates opportunities to shorten the closest pair distance between Flickr users. This is a frequent concern from many scholars who question the credibility of using geotagged photos to study people's spatio-temporal activities. If the number of geotagged photos posted by Flickr users influences the spatio-temporal proximity of their geotagged photos, it is then necessary to differentiate the users with different numbers of geotagged photos in many analyses. This study therefore investigates whether the distance between users' geotagged photos are related to the number of geotagged photos posted by them in response to the concerns mentioned above.

3.2. 4 How are the closest photo pairs and the nearest neighbor photo pairs of Flickr friends and non-friends distributed geographically?

Flickr geotagged photos are unevenly distributed around the world. This draws our attention to the geographic meaning underneath the spatial distribution of Flickr geotagged photos. However, since geotagged photos are an incomplete recording of Flickr users' spatio-temporal activities, simply locating them on a map may lead to many biases. To address this problem, we compare the spatial distributions of geotagged photos between Flickr friends with those of non-friends. This comparison helps us focus on the effect of online friendship by subtracting other variables for both Flickr friends and non-friends.

CHAPTER 4

METHODOLOGY

4.1 Data acquisition

Due to the large quantity of information present on Flickr, data acquisition is computationally demanding. It is therefore more practical to download a sample network of users and their geotagged photos. There are many ways to sample a network such as random sampling. However, this method cuts off many connections among users. Snow ball sampling, on the other hand, works better to keep network connections. Since the friendship connection is the point of emphasis in this study, snow ball sampling is adopted. A program using Flickr API successfully obtained around 46,844,044 public geotagged photo records from 1.1 million users. Additionally, the program downloaded the friendship network of these 1.1 million users.

The downloaded data have following features:

First, there is a ten-year archive of geotagged photos in the downloaded dataset. Flickr has enabled users to geotag their photos online since 2002. Its geotagging function predates comparable functions in Facebook and Twitter. Consequently it enables us to trace users' spatio-temporal activities over a longer time span.

Second, most downloaded photos are geotagged at relatively high spatial precision levels. When users manually geotag their photos on Flickr World Map, they can zoom among 16 different scales. A higher scale represents a more detailed map. For example, the 11th scale is designed to demonstrate the geographic features at the city level. Among the 46,844,044 pieces of geotagged photos downloaded, 94.7% of them are geotagged at the city or more detailed levels.

Third, the downloaded photos have time stamps. The time stamp of a geotagged photo is automatically recorded in an EXIF file by digital camera. EXIF is the abbreviation of Exchangeable Image File Format. It is a standard that "specifies the formats for images,

sound, and ancillary tags used by digital cameras (including smart phones), scanners and other systems handling image and sound files recorded by digital cameras” (http://en.wikipedia.org/wiki/Exchangeable_image_file_format). Flickr keeps the EXIF file for each photo. Flickr uses *Date Granularity* to refer to temporal precision of geotagged photos. There are four main date granularities:

Table 1. The four main date granularities of Flickr geotagged photos

0	Y-m-d H:i:s
4	Y-m
6	Y
8	Circa...

For the time stamps of 46,844,044 geotagged photos downloaded, 46,276,737 of them have the highest granularity level (see Figure 6).

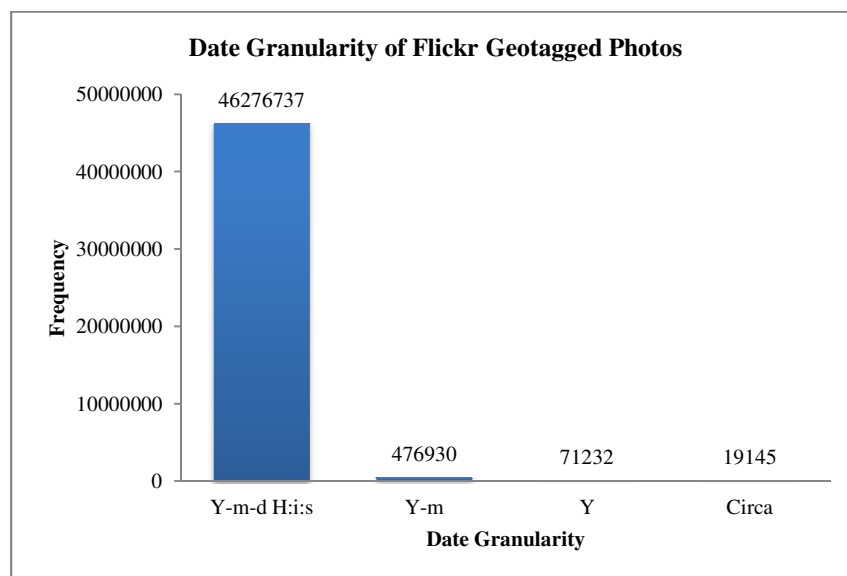


Figure 6. Date granularity of downloaded Flickr geotagged photos

In summary, every geotagged photo record downloaded from Flickr has a time stamp, geographic coordinates, a user id, a date granularity, and a spatial precision level.

4.2 Data processing

Data processing addresses data error and data redundancy. As shown in Figure 7, the geographic coordinates of certain consecutive photo records of user “38795929@N00” are the same. “Batch geotagging” serves as a potential reason. This method can be applied to locate all photos at the same point on the Flickr World Map. In addition, these consecutive photo records are taken within 24 hours. In other words, these geotagged photos refer to a similar situation that user “38795929@N00” photos at that specific geographic location within 24 hours. Retaining these records leads to redundant computation when performing the nearest neighbor analysis. Therefore, only one record is kept when two or more photos of the same user are taken within 24 hours and have the same geographic coordinates. Ten million out of the fifty million geotagged photos remain after this data reduction.

FID	Shape	OBJECTID	Photoid	UserID	Username	Longitude	Latitude	DateTaken	Dateupload	Dategranul	Accuracy
4721	Point	41825176	4625360357	38795929@N00	shaneandrut	-83.938609	35.944172	5/6/2010	127440990	0	16
4722	Point	41825177	4625985564	38795929@N00	shaneandrut	-83.938534	35.944168	5/6/2010	127440996	0	16
4723	Point	41825178	4625984352	38795929@N00	shaneandrut	-83.938609	35.944172	5/6/2010	127440994	0	16
4724	Point	41825179	4625983840	38795929@N00	shaneandrut	-83.938609	35.944172	5/6/2010	127440982	0	16
4725	Point	41825180	4625377635	38795929@N00	shaneandrut	-83.938609	35.944172	5/6/2010	127440981	0	16
4726	Point	41825181	4625982214	38795929@N00	shaneandrut	-83.938609	35.944172	5/6/2010	127440977	0	16
4727	Point	41825182	4625375885	38795929@N00	shaneandrut	-83.938609	35.944172	5/6/2010	127440976	0	16
4728	Point	41825183	3102458905	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911124	0	16
4729	Point	41825184	3102329264	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911121	0	16
4730	Point	41825185	3102458129	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911118	0	16
4731	Point	41825186	31023291612	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911117	0	16
4732	Point	41825187	3102456639	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911113	0	16
4733	Point	41825188	3102455971	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911111	0	16
4734	Point	41825189	3102455111	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911109	0	16
4735	Point	41825190	3102328274	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911106	0	16
4736	Point	41825191	3102327566	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911104	0	16
4737	Point	41825192	3102452659	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911100	0	16
4738	Point	41825193	3102452221	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911099	0	16
4739	Point	41825194	3102451853	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911097	0	16
4740	Point	41825195	3102451285	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911096	0	16
4741	Point	41825196	3102328458	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911094	0	16
4742	Point	41825197	3102450177	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911092	0	16
4743	Point	41825198	3102328356	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911090	0	16
4744	Point	41825199	3102328296	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911086	0	16
4745	Point	41825200	3102447245	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911083	0	16
4746	Point	41825201	3102328272	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911080	0	16
4747	Point	41825202	3102445863	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911079	0	16
4748	Point	41825203	3102445285	38795929@N00	shaneandrut	-81.687662	41.468811	12/6/2008	122911077	0	16

Figure 7. Geotagged photo records of user “38795929@N00”

For friendship connections, the friendship between user A and user B is saved in a friendship table as one record of “user A, user B”. There are four million records in the friendship table. However, the snow ball sampling method does not cover all friendship connections among these 1.1 million Flickr users. During data acquisition, snow ball

sampling was executed step by step. In each step a new layer of users (outer users) was downloaded. They were connected to the users which had already been downloaded in previous steps (inner users). In this case, we knew how outer users were connected to inner users. We did not know how outer users were connected among themselves. This lack of friendship knowledge affects how non-friend pairs are generated. To get many pairs of non-friends, random users who are not recorded as friends in the friendship table are matched up as non-friend pairs. However, it is only practical to match an inner user with another inner user or to match an inner user with an outer user. An outer user cannot be matched with another outer user because their friendship relation is unknown from the friendship table.

Though geotagging functions of mainstream OSN are becoming more popular, it is noteworthy that only a small proportion of OSN users actually geotag their photos. As mentioned before, in a test over one million Twitter users, the percentage of users who geotag at least one Tweet at the city level is only 26% (Caverlee, 2010). In our Flickr dataset, only 205,120 out of 1.1 million users have geotagged photos. The maximum number of a single user's geotagged photos is 52,551. Since this study aims to analyze the distance between geotagged photos, it only includes those users with at least one geotagged photo. Furthermore, after checking the users with more than 1,000 geotagged photos, it is found that most of them are organizations, events or institutions. Therefore, this study includes only those users with 1,000 or fewer geotagged photos. In addition, although there are around 4,000,000 pairs of Flickr friends in the dataset, it takes too much computational time to perform analyses on all of them. Therefore, every tenth friend pairs is selected to be included in this study. This results in 400,000 pairs of the original 4,000,000 pairs of Flickr friends, among which 92,525 pairs have at least one but fewer than 1,000 geotagged photos. These 92,525 pairs of Flickr friends are then used in this study.

4.3 Frequency analysis on the spatio-temporal proximity of geotagged photos

Tasks in this section address the first three research questions in Chapter 3. In order to evaluate whether Flickr friends tend to geotag photos that are closer to each other in space than Flickr non-friends (research question 3.2.1), frequency analyses are applied to the distances of geotagged photos. To evaluate the overall proximity, the nearest neighbor distances between the geotagged photos of the 92,525 pairs of Flickr friends mentioned above are calculated. Since the nearest neighbors from point set A to point set B can be different from those from point set B to point set A, the “from” user is assigned as “the first user” and the “to” user is assigned as “the second user.” For each photo point of the first user, its nearest neighbor photo point from the second user was identified and the distance between them is calculated. The number of nearest neighbor pairs between the two users is equal to the number of geotagged photos of the first user. A similar experiment is conducted on Flickr non-friends. To generate non-friend pairs, the inner users are matched with random non-friend inner users or random non-friend outer users. 92,525 pairs of non-friends who have at least one and fewer than 1,000 geotagged photos are then generated. The nearest neighbor distances between the geotagged photos of the 92,525 pairs of Flickr non-friends are calculated. A two-sample Kolmogorov-Smirnov test is then applied to test whether the nearest neighbor photo distance is related to Flickr friendship. Kolmogorov-Smirnov test is a nonparametric test algorithm. For the nearest neighbor photo distance, the null hypothesis is that the two samples are from the populations with the same distribution function. For each of the two samples, the data are sorted into ascending order, from $X_{[1]}$ to $X_{[n]}$. The empirical cumulative distribution function ($\hat{F}_i(X)$) for group i is computed as:

$$\hat{F}_i(X) = \begin{cases} 0 & -\infty < X < X_{[1]} \\ j/n_i & X_{[j]} \leq X < X_{[j+1]} \\ 1 & X_{[n_i]} \leq X < \infty \end{cases}$$

For all X_j values in the two groups, the difference between the two groups(D_j) is

$$D_j = \widehat{F}_1(X_j) - \widehat{F}_2(X_j)$$

Where $\widehat{F}_1(X_j)$ is the cumulative distribution function for the group with the larger sample size.

The test statistic (Z value) is:

$$Z = \max_j |D_j| \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

For the closest pair distance, a similar frequency analysis is conducted. There are two differences: First, since there is only one pair of the closest geotagged photos between two users, only one distance is recorded for each pair of Flickr users. Second, since there is no “from” user and “to” user, the numbers of geotagged photos of both users are recorded. In order to maintain consistency, the same 92,525 pairs of Flickr friends and non-friends are used.

To answer whether the time difference between the closest photo pair is related to Flickr friendship, the frequency analysis is applied on the time difference between the closest photo pairs. A two-sample Kolmogorov-Smirnov is then applied to compare the two distributions.

To answer whether the closest pair distance between Flickr friends is related to the numbers of geotagged photos posted by them, two Pearson’s correlation analyses are conducted. For the closest pair distance, there is neither a “from” user nor a “to” user. As a result, each distance is related to the photo numbers of two users. In most cases the numbers of geotagged photos of the two users are different. The two photo numbers are then differentiated as the higher photo number and the lower photo number. As discussed previously, geotagging more photos may reduce the closest pair distance between a pair of Flickr friends. However, it is not clear whether the closest pair distance is related to the higher photo numbers or the lower photo numbers. The first

correlation analysis is between the closest pair distance of Flickr friends and the higher photo numbers. The second is between the closest pair distance of Flickr friends and the lower photo numbers.

4.4 Visualize the spatial distribution of geotagged photos

Tasks in this section address the fourth research question in Chapter 3. Since geographic distributions of Flickr users' geotagged photos can help us understand the geographic meaning of their geotagging activities, this study leverages the strength of GIS to visualize the geographic distribution of Flickr geotagged photos. To illustrate the pairwise relationship, the nearest neighbor photo pairs and the closest photo pairs are connected with a line and located on the map. For the closest photo pairs, the same color is used to represent all point pairs (see Figure 8). For the nearest neighbor photo pairs, red is used to represent the photo point of the first user and green is used to represent the photo point of the second user.

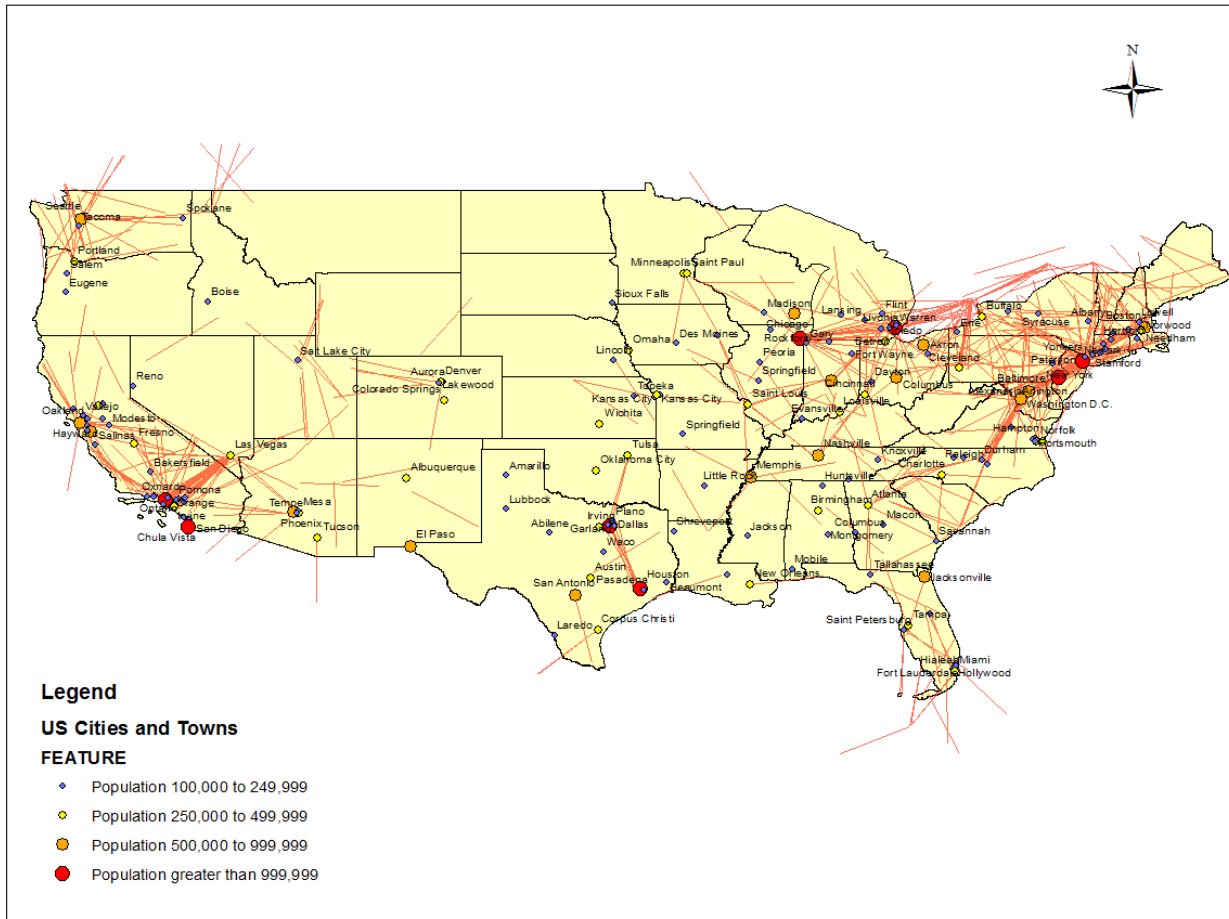


Figure 8. A sample visualization of the closest photo pairs between 300 km to 350 km in the lower 48 states of the U.S.

Visualizing the geographic distribution of geotagged photos of Flickr friends reveals some interesting details on the relationship between Flickr friendship and their geotagged photos. However, as discussed above, Flickr geotagged photos have some embedded biases which challenge the credibility of these findings. Merely exploring the geographic distributions of Flickr friends can hardly make any solid conclusions. To address this problem, a similar visualization is applied to the nearest neighbor photo points of Flickr non-friends. A comparison between these two spatial distributions further distills the analysis to the variable of online friendship.

In summary, to investigate whether Flickr friends tend to geotag photos that are closer in space and time than Flickr non-friends, frequency analyses are applied on a large volume of geotagged photos and online friendships. To analyze the relationship

between the spatial proximity of geotagged photos and the number of geotagged photos posted by Flickr users, correlation analyses are applied. To further investigate how Flickr friendship influences Flickr users' geotagging activity from geographical views, spatial visualization is applied. Most of the analyses in this study are data-intensive and call for much computational effort.

CHAPTER 5

FINDINGS AND INTERPRETATIONS

This project samples 1.1 million Flickr users using the snowball sampling method. 250,120 of the sampled Flickr users have geotagged photos. 99.7% of the 250,120 users have less than 1,000 geotagged photos (see Figure 9). In total, fifty million geotagged photo records from these 250,120 users are downloaded, most of which were taken between 2000 and 2010.

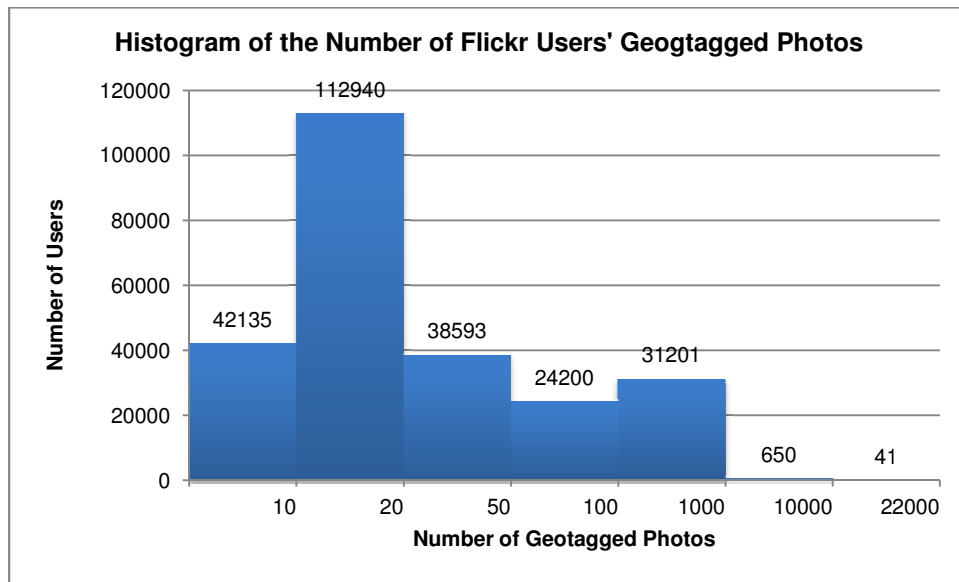


Figure 9. Histogram of the numbers of Flickr users' geotagged photos

To study the relationship between online friendship and spatio-temporal proximity of their geotagged photos under controllable computational time, 92,525 pairs of Flickr friends and 92,525 pairs of Flickr non-friends are selected. The number of geotagged photos for each user is between 1 and 1,000. Among the 92,525 pairs of Flickr friends, 7,884,291 pairs of nearest neighbor photos are found. Among the 92,525 pairs of Flickr non-friends, 7,763,305 pairs of nearest neighbor photos are found.

5.1 General spatial distribution of Flickr geotagged photos

Though Flickr geotagged photos are taken all over the world, their geographic distribution is uneven. In Figure 10, a world map is divided into a grid of 30x30 km² cells.

The number of geotagged photos is counted in each cell. The photo densities are classified into four categories according to the quantile values. Few green cells are in the ocean area. By checking the images and comments of the geotagged photos in these cells, it is found that many of them are taken on islands or cruises. Generally speaking, there are more geotagged photos in the east and west coasts of the United States, the United Kingdom, and some other European countries than in other regions. These spatial distribution patterns of Flickr geotagged photos direct our attention to the geographical meanings of Flickr users' geotagging activities in the following analysis.

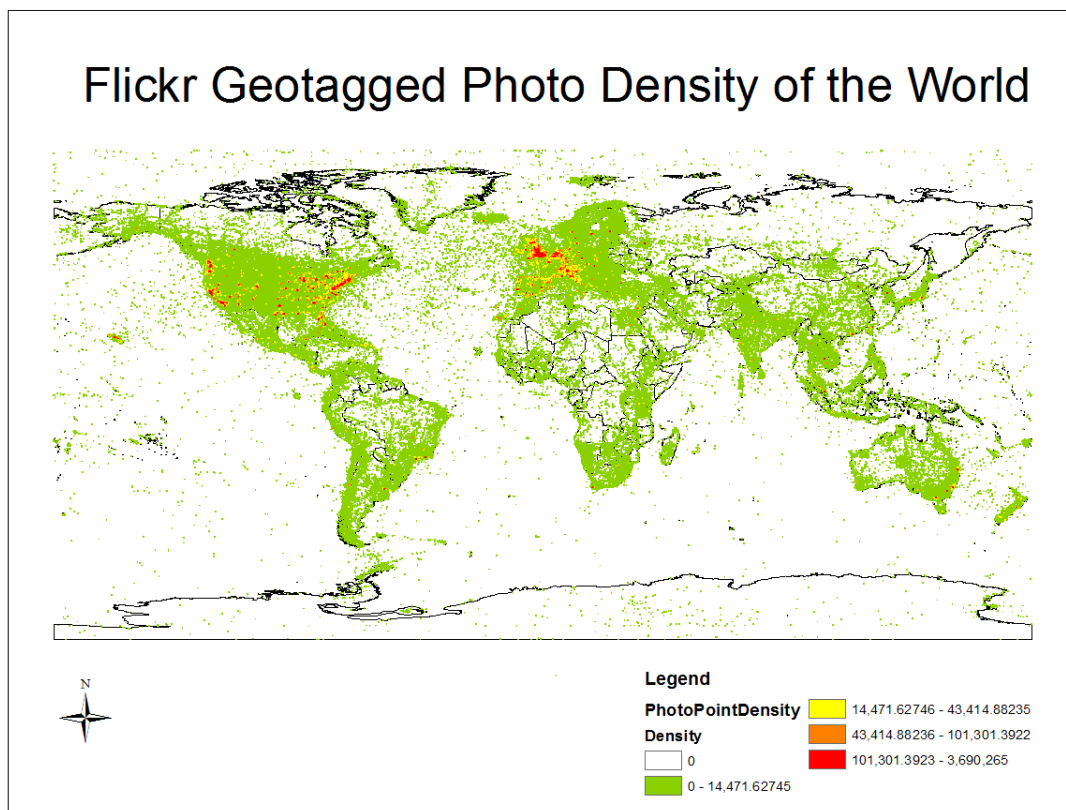


Figure 10. Spatial frequency of Flickr geotagged photos around the world

5.2 Relationships between online friendships and the spatio-temporal proximity of their geotagged photos

5.2.1 Do Flickr friends tend to geotag their photos closer to each other in space than Flickr non-friends do?

To answer this question, the frequency analyses are applied to the closest pair distance and the overall distance respectively. For 92,525 pairs of Flickr friends, 22,909 pieces of closest pair distances are within 10 km, which accounts 24.8% of 92,525. 28,073 of them are within 50 km, which accounts 30% of 92,525. 46,367 of them are within 530 km, which accounts 50% of 92,525. The cumulative proportion of the closest pair distances between Flickr friends are shown in Figure 11.

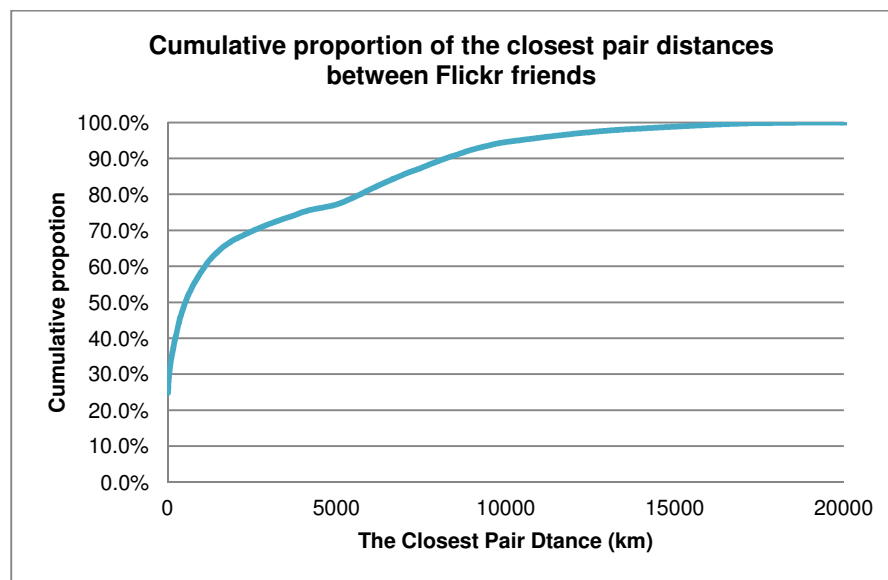


Figure 11. Cumulative proportion of the closest pair distances between Flickr friends

For 92,525 pairs of Flickr non-friends, 4,686 pieces of the closest pair distances are within 10 km, which accounts only 5.06% of 92,525. 6,492 of them are within 50 km, which accounts only 7.02% of 92,525. 20,630 of them are within 530 km, which accounts 22.3% of 92,525. The cumulative proportion of the closest pair distances between Flickr non-friends are shown in Figure 12.

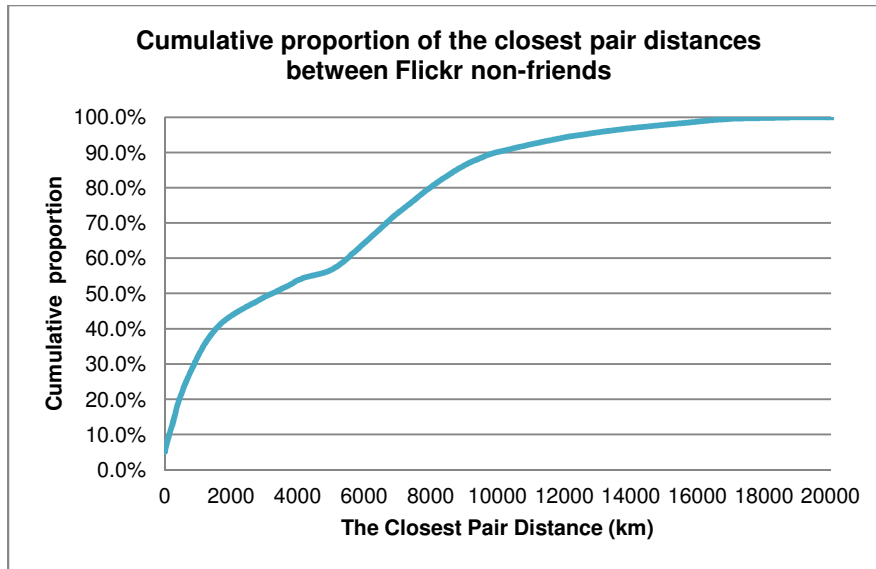


Figure 12. Cumulative proportion of the closest pair distances between Flickr non-friends

To test whether the closest pair distance is related to the friendship of Flickr users, a two-sample Kolmogorov-Smirnov test is applied. The null hypothesis is the two samples come from the same distribution.

The test is run by SPSS and the result of the hypothesis test is shown as follows:

Table 2. Summaries of the two-sample Kolmogorov-Smirnov test

Frequencies		
	Friendship	N
The closest pair distance	Friends	92525
	Non-friends	92525
	Total	185050

Table 3. Hypothesis test summary of the two-sample Kolmogorov-Smirnov test

Test Statistics ^a		
		The Closest Pair Distance
Most Extreme Differences	Absolute	.279
	Positive	.279
	Negative	.000
Kolmogorov-Smirnov Z		60.094
Asymp. Sig. (2-tailed)		.000

a. Grouping Variable: Friendship

According to the Kolmogorov-Smirnov test, the null hypothesis that two samples come from the same distribution is rejected at the significance level of 0.05. In other words, this analysis shows that the closest pair distance is related to Flickr friendship.

A comparison of the closest pair distances between Flickr friends and non-friends at some typical distance ranges are shown to further illustrate these differences:

Table 4. Comparison of the closest pair distances between Flickr friends and non-friends

Distance Range (km)	Proportion (Friends)	Proportion (Non-friends)
0-10	24.8%	5.1%
0-50	30.3%	7.0%
0-100	33.8%	8.7%
0-500	49.4%	21.5%
0-900	57.1%	30.3%

The table shows that more Flickr friends have their closest photo pairs within a shorter distance than Flickr non-friends do. In particular, the proportion of the closest pair

distances within 10 km of Flickr friends is 4.86 times that of Flickr non-friends. From this perspective Flickr friends tend to geotag photos that are closer than Flickr non-friends do.

Furthermore, from the perspective of overall proximity, the nearest neighbor distance is studied. The cumulative proportions of the nearest neighbor distances of Flickr friends and Flickr non-friends are shown in Figure 13 and 14, respectively.

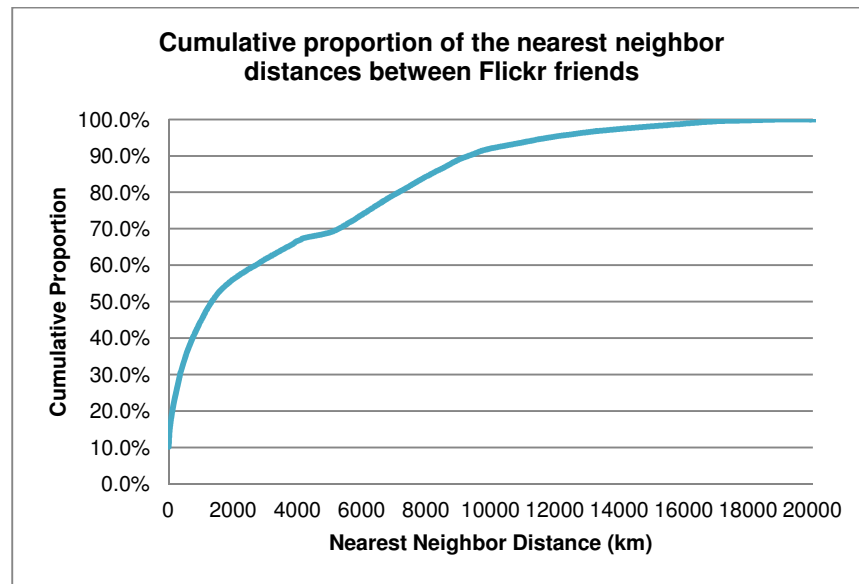


Figure 13. Cumulative proportion of the nearest neighbor distances between Flickr friends

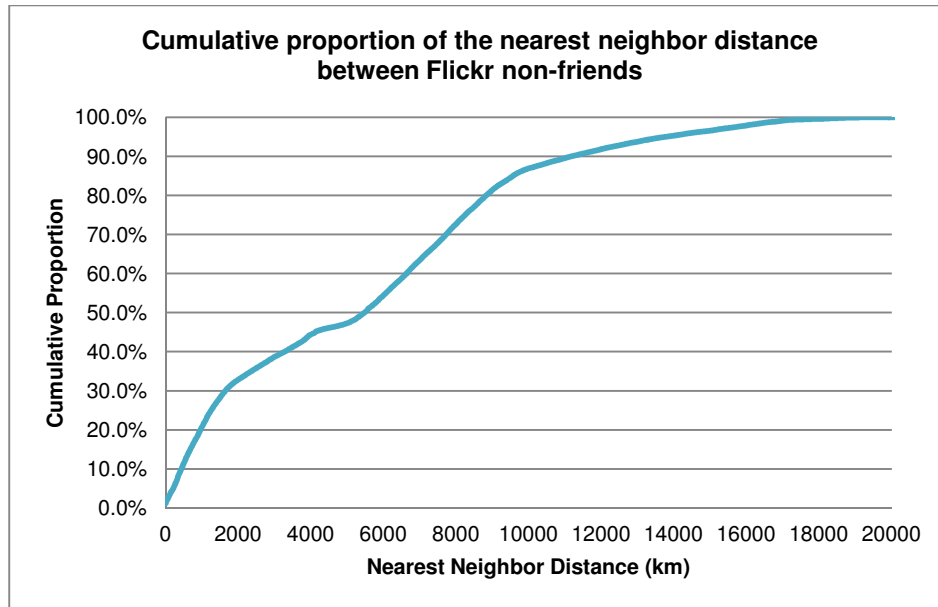


Figure 14. Cumulative proportion of the nearest neighbor distances between Flickr non-friends

To test whether the nearest neighbor distance is related to the friendship of Flickr users, a two-sample Kolmogorov-Smirnov test is applied. The null hypothesis is the two samples come from the same distribution.

The test is run by SPSS and the result of hypothesis test is shown as follows:

Table 5. Summaries of the two-sample Kolmogorov-Smirnov test

Frequencies		
	Friendship	N
The nearest neighbor distance	Friends	7884291
	Non-friends	7763305
	Total	15647596

Table 6. Hypothesis test summary of the two-sample Kolmogorov-Smirnov test

Test Statistics ^a		The Nearest Neighbor Distance
Most Extreme Differences	Absolute	.244
	Positive	.244
	Negative	.000
Kolmogorov-Smirnov Z		482.098
Asymp. Sig. (2-tailed)		.000

a. Grouping Variable: friendship

According to the Kolmogorov-Smirnov test, the null hypothesis that two distributions come from the same distribution is rejected at the significance level of 0.05. In other words, this analysis shows that the nearest neighbor photo distance is related to Flickr friendship.

A comparison of the nearest neighbor distances between geotagged photos of Flickr friends and non-friends at some typical distance ranges are also shown to further illustrate their differences:

Table 7. Comparison of the nearest neighbor distances between Flickr friends and non-friends

Distance Range (km)	Proportion (Friends)	Proportion (Non-friends)
0-10	10.2%	1.5%
0-50	16.1%	2.2%
0-100	19.3%	3.2%
0-350	30.0%	8.3%
0-1330	50.0%	26.1%

The table shows that Flickr friends have more nearest neighbor photo pairs within shorter distance ranges than Flickr non-friends do. In particular, the proportion of the nearest neighbor photo distances within 10 km of Flickr friends is 6.8 times that of Flickr non-friends.

In summary, the two frequency analyses above are conducted from two different perspectives. However, both of them show a similar pattern: The spatial proximity between geotagged photos is related to Flickr friendship. Generally, Flickr friends tend to geotag photos that are closer to each other in space than Flickr non-friends do.

5.2.2 Is the time difference between the closest photo pairs of Flickr friends shorter than that of Flickr non-friends?

Time is considered as an important feature of people's activity. In existing research, some scholars assumed that OSN friends tend to geotag photos which are close in both time and space. Some projects attempt to infer the friendship between OSN users based on this assumption (Crandall et al., 2010). To further verify this assumption, three analyses are conducted. In the first analysis, the time differences between the closest photo pairs of Flickr friends and non-friends are calculated respectively. The distribution of the time differences (within 1,000 days) between the closest photo pairs of Flickr friends is shown as follows:

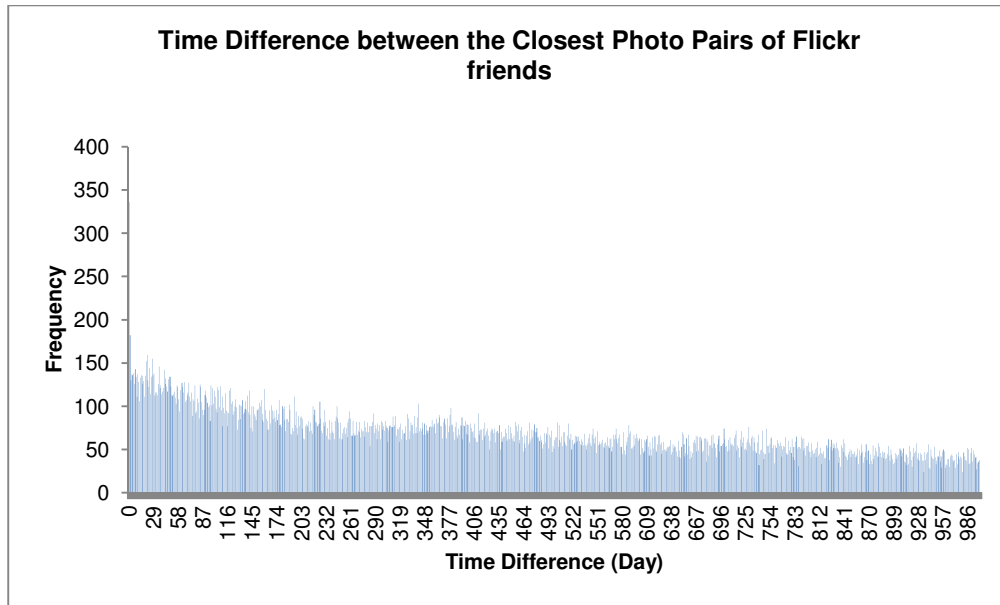


Figure 15. Time differences between the closest photo pairs of Flickr friends

In Figure 15, the horizontal axis represents the time differences between the closest photo pairs of Flickr friends. One unit represents 1 day. The longest time difference is 1,000 days. The frequency decreases gradually. In comparison, the distribution of the time differences between the closest photo pairs of Flickr non-friends is shown as follows:

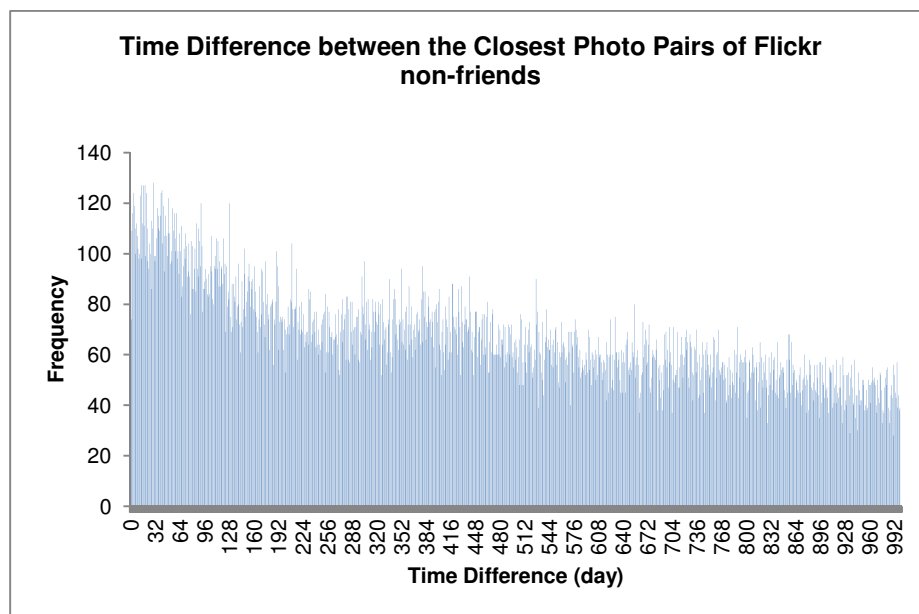


Figure 16. Time differences between the closest photo pairs of Flickr non-friends

In Figure 16, the horizontal axis represents the time difference between the closest photo pairs of Flickr non-friends. One unit represents 1 day. The longest time difference is also 1,000 days. To better compare the two distributions, both of them are illustrated in the same figure as follows:

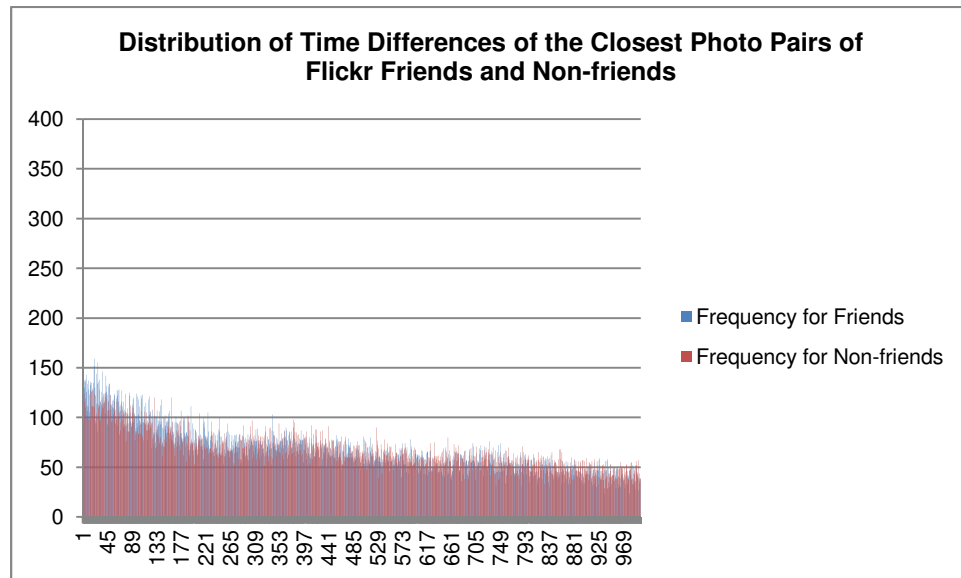


Figure 17. Time differences between the closest photo pairs of Flickr friends and non-friends

To test whether the time difference between the closest photo pairs is related to the friendship of Flickr users, a two-sample Kolmogorov-Smirnov test is applied. The null hypothesis is the two samples come from the same distribution.

The test is run by SPSS and the result of the hypothesis test is shown as follows:

Table 8. Summaries of the two-sample Kolmogorov-Smirnov test
Frequencies

	Friendship	N
Time_Difference	Friends	69083
	Non-friends	66865
	Total	135948

Table 9. Hypothesis test summary of the two-sample Kolmogorov-Smirnov test

Test Statistics ^a		Time_Difference
Most Extreme Differences	Absolute	.039
	Positive	.039
	Negative	.000
Kolmogorov-Smirnov Z		7.126
Asymp. Sig. (2-tailed)		.000

a. Grouping Variable: Friendship

According to the Kolmogorov-Smirnov test, the null hypothesis that two distributions come from the same distribution is rejected at the significance level of 0.05. In other words, this analysis shows that the time difference between the closest photo pairs is related to Flickr friendship.

To further investigate this relationship, a second analysis is applied. The closest photo pairs of Flickr friends within specific spatial distance are selected. Then, the time differences between these photo pairs are calculated:

Table 10. The number of closest photo pairs within different distance ranges and one day of Flickr friends

Spatial distance range (km)	Number of the closest photo pairs within the given spatial distance range	Number of the closest photo pairs within the given spatial distance range and one day	Proportion of the closest photo pairs within the given spatial distance range and one day out of those within the given spatial distance range
0-10	22,090	216	0.98%
10-20	2,929	5	0.2%
20-30	1,306	1	0.1%
30-40	930	2	0.2%

Table 10. Continued

Spatial distance range (km)	Number of the closest photo pairs within the given spatial distance range	Number of the closest photo pairs within the given spatial distance range and one day	Proportion of the closest photo pairs within the given spatial distance range and one day out of those within the given spatial distance range
40-50	818	1	0.1%

Table 11. The number of closest photo pairs within different distance ranges and one day of Flickr non-friends

Spatial distance range (km)	Number of the closest photo pairs within the given spatial distance range	Number of the closest photo pairs within the given spatial distance range and one day	Proportion of the closest photo pairs within the given spatial distance range and one day out of those within the given spatial distance range
0-10	4,686	6	0.12%
10-20	661	1	0.2%
20-30	442	0	0%
30-40	364	1	0.3%
40-50	339	0	0%

In Table 10, the proportion for the closest photo pairs within one day and ten kilometers is obviously higher than others. For Flickr non-friends, the corresponding statistics in Table 10 are obviously lower. However, it is noteworthy that the proportions in the fourth columns of both Table 10 and Table 11 are very low. In other words, for both Friends and non-friends, most of their closest photo pairs are not taken within relatively short time span (e.g. one day).

In the third analysis, the importance of time is better stressed. The closest photo pairs of Flickr friends within specific time differences are selected. Then, the spatial distances between these photo pairs are calculated. The result is as follows:

Table 12. Time difference between the closest photo pairs of Flickr friends

Time difference (day)	Number of the closest photo pairs within the given time difference	Number of the closest photo pairs within the given time difference and 10 km	Proportion of the closest photo pairs within the given time difference and 10 km out of those within the given time difference
0-1	348	216	62%
1-2	182	70	38%
2-3	130	44	33%
3-7	541	162	30%
7-30	3,041	779	26%

In Table 12, the proportions in the fourth column decrease gradually. For Flickr non-friends, the corresponding statistics are obviously lower (see Table 13). For the closest photo pairs taken within one day, only 5% are within 10 km, which is only 1/12 of that of Flickr friends.

Table 13. Time difference between the closest photo pairs of Flickr non-friends

Time difference (day)	Number of the closest photo pairs within the given time difference	Number of the closest photo pairs within the given time difference and 10 km	Proportion of the closest photo pairs within the given time difference and 10 km out of those within the given time difference
0-1	120	6	5%
1-2	101	5	5%
2-3	146	0	0%

Table 13. Continued

Time difference (day)	Number of the closest photo pairs within the given time difference	Number of the closest photo pairs within the given time difference and 10 km	Proportion of the closest photo pairs within the given time difference and 10 km out of those within the given time difference
3-7	83	17	20%
7-30	2961	108	3%

When the closest photo pairs of Flickr friends are taken within one day, 62% of them are within 10 km. It is 12.4 times higher than that of Flickr non-friends. In other words, the closest photo pairs from Flickr friends within one day are more likely to be within a short distance range.

However, findings from Table 12 and Table 13 are not enough to conclude that time is one of the most influential factors on Flickr friends' geotagging activities. Further investigation reveals some limitations. A key concern is that very few of the closest photo pairs are within a short time span (e.g. one day) and a short distance range (e.g. 10km). For example, only 348 of the 92,525 closest photo pairs of Flickr friends are within 1 day and 10 km. To further illustrate it, Table 14 shows the proportion of the closest photo pairs of Flickr friends within typical spatial and temporal thresholds.

Table 14. The proportion of the closest photo pairs within some typical spatial and temporal thresholds

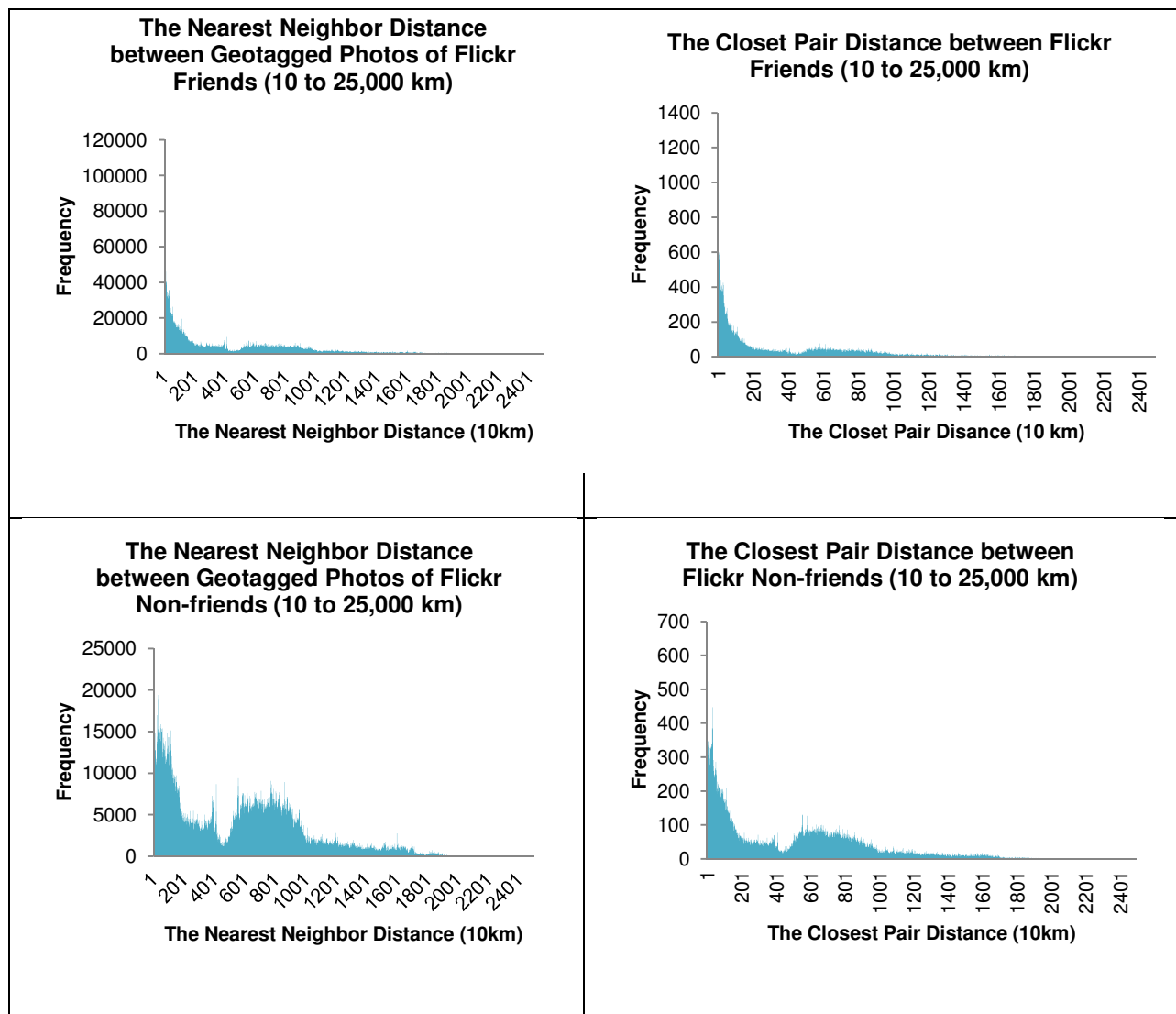
Distance (km) \ Time Difference (hour)	10	20	50	100	250	500
0-12	0.18%	0.18%	0.18%	0.19%	0.19%	0.20%
0-24	0.23%	0.24%	0.24%	0.25%	0.26%	0.28%
0-168	0.53%	0.56%	0.60%	0.64%	0.72%	0.80%
0-720	1.37%	1.48%	1.64%	1.77%	2.08%	2.76%

In Table 14, only 0.18% of the 92,525 closest photo pairs are within 12 hours and 10 km. From the perspective of time geography, 10 km is still a relatively coarse scale to define “co-location in space” while 12 hours is still a relatively coarse scale to define “co-location in time”. When the spatial threshold is 500 km and the temporal threshold is 720 hours (30 days), the proportion is still only 2.76%. Hence, most Flickr friends do not have many geotagged photos which are close in both space and time between them. The cases of “co-existence” between Flickr friends revealed by their geotagged photos are very limited. In the literature review, some scholars tried to “infer online friendships” based on the “spatio-temporal co-occurrence” between geotagged photos of Flickr users. According to the findings of this analysis, it is reasonable to question the effectiveness of this friendship inference since the proportion of the “spatio-temporal co-occurrence” photo pairs is so limited.

5.2.3 The low frequency point in distance distributions: influence of physical boundary.

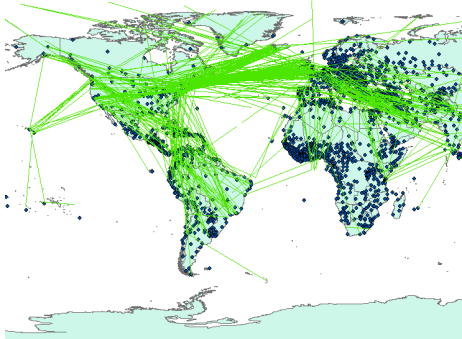
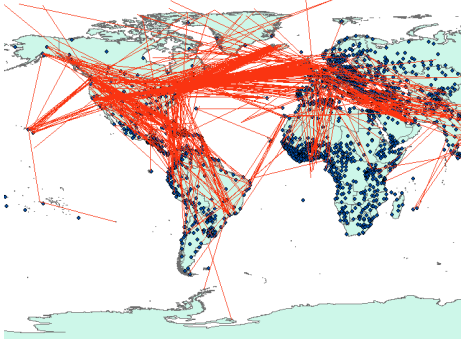
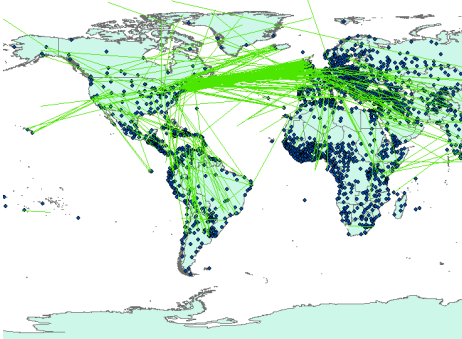
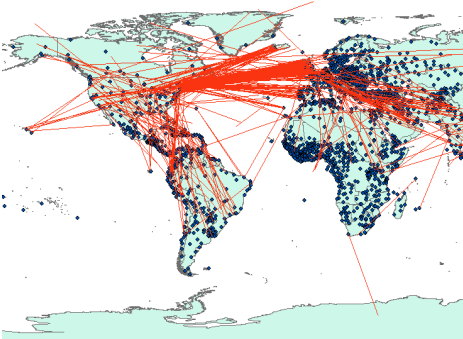
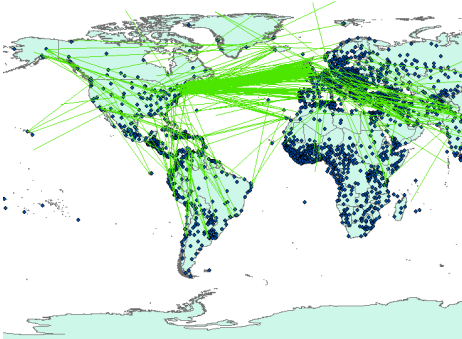
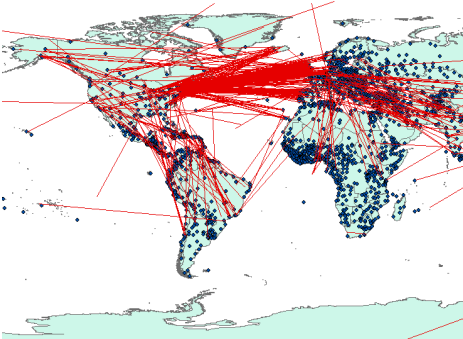
The frequency distributions of both nearest neighbor distance and the closest pair distance display a low point at around 4,000 km for both Flickr friends and non-friends. It reminds us of the potential boundary effect (see Table 15).

Table 15. Frequency distributions of the four kinds of distances



To further investigate this phenomenon, the geographic distribution of the closest photo pairs at round 4,000 km is illustrated in Table 16.

Table 16. Spatial distribution of the closest photo pairs between Flickr friends between 4,200 km and 5,000km

Distance Range (km)	Flickr Friends	Flickr Non-Friends
4200-4600		
4600-4800		
4800-5000		

In Table 16, the closest photo pairs are connected with green or red lines. The blue dots represent world cities. Some distinct geographic patterns are revealed. First, a large

proportion of the closest photo pairs within 4,000 km and 5,000 km are between the Northeast Coast of the U.S. and the West Coast of Europe. Between 4,200 km and 4,600 km, some of them connect Iceland and the Northeast Coast of the U.S. Generally speaking, the distance between the Northeast Coast of the U.S. and the West Coast of Europe (e.g. the distance between New York City and Lisbon) is around 5,000 km. Photo pairs within shorter distance between the East Coast of the U.S. and the West Coast of Europe may reach the ocean area, where there are few geotagged photos. This may be a potential explanation for the low frequency of the closest photo pairs within this distance range.

To better illustrate the distribution in these areas, the closest photo pairs of Flickr friends and non-friends within 4,200 km and 4,600 km between the West Coast of Europe and the East Coast of the U.S. are shown in Figure 18, 19, 20, and 21.

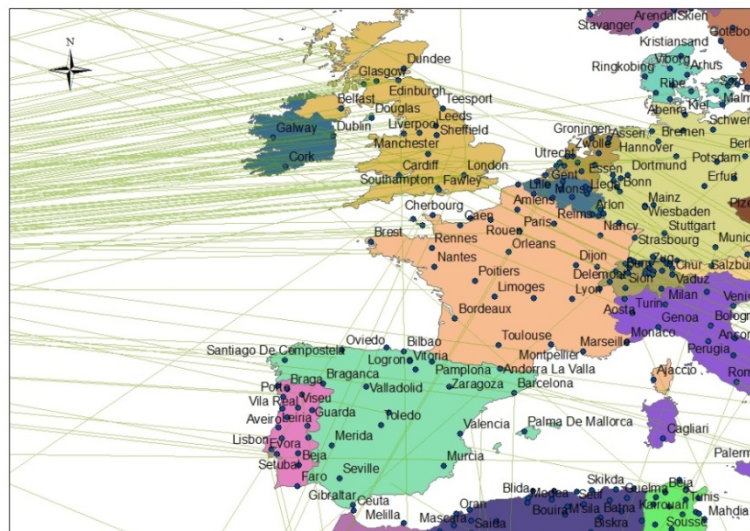


Figure 18. The closest photo pairs between 4,200 km and 4,600 km of Flickr friends at the West Coast of Europe

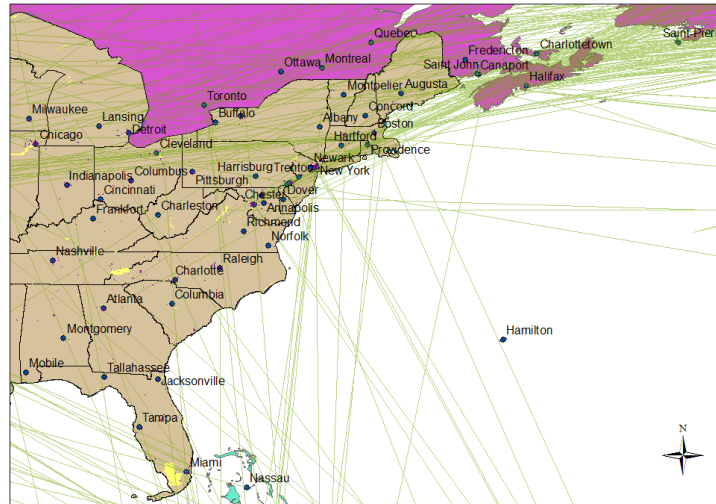


Figure 19. The closest photo pairs between 4,200 km and 4,600 km of Flickr friends at the East Coast of the U.S.

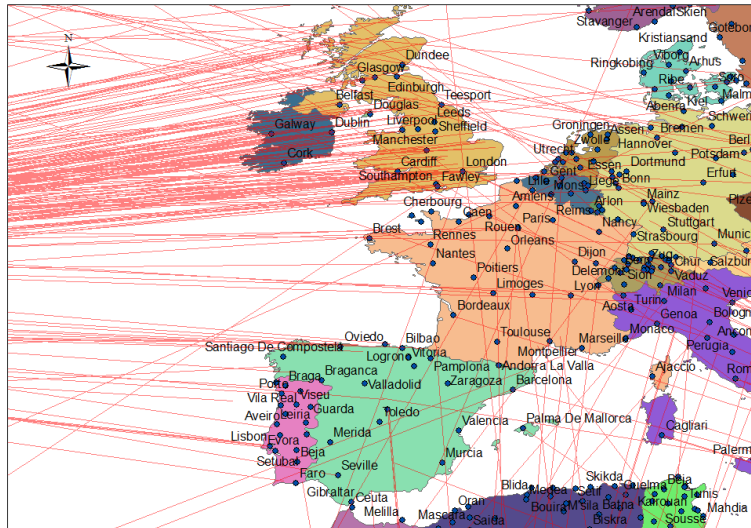


Figure 20. The closest photo pairs between 4,200 km and 4,600 km of Flickr non-friends at the West Coast of Europe

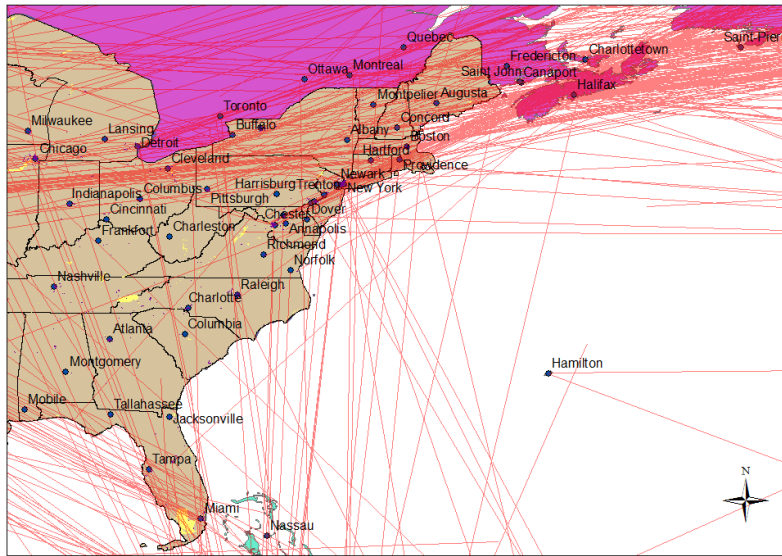


Figure 21. The closest photo pairs between 4,200 km and 4,600 km of Flickr non-friends at the East coast of the U.S.

Figures 18 and 20 reveal that Spain and Britain are highly covered by closest photo pairs at around 4,000 km. For the West Coast of Spain, which is popular among tourists, tourism may be one of the key factors for these connections. In the United Kingdom, big cities such as London and Edinburg are connected more by green or red lines. Population, infrastructure and tourism may be the potential explanations for this phenomenon. In addition, Figures19 and 21 show that big cities near the northeast coast of the U.S. and the big cities near the east coast of Canada are more covered by the closest photo pairs.

In addition to the connections mentioned above, there are two other typical patterns in Figures19 and 21. First, the connections between the Northwest of the U.S. and some tourism areas (e.g. the Caribbean, Hawaii, etc.) are highlighted. Shorter lines following these connections can only reach the places in oceans. Second, the connections between the West Coast of the U.S. (e.g. Los Angeles, CA) and the East Coast of the U.S. are also highlighted. Longer lines following these connections can only reach the

places in oceans. It is therefore reasonable to assume that the presence of an ocean leads to the low frequency of the closest photo pairs at this distance range.

5.2.4 The relationship between the spatial proximity of geotagged photos and the number of geotagged photos posted by Flickr friends.

For the closest photo pair, there is neither a “from” user nor a “to” user. As a result, each distance is related to the photo numbers of two users. In most cases the numbers of geotagged photos of the two users are different. The two photo numbers are then differentiated as the higher photo number and the lower photo number. As discussed before, geotagging more photos may reduce the closest pair distance between a pair of Flickr friends. However, it is not clear whether the closest pair distance is more related to the higher photo numbers or the lower photo numbers. In order to investigate these issues, two Pearson’s correlation analyses are applied. One is between the closest pair distance of Flickr friends and the higher photo numbers. The other is between the closest pair distance of Flickr friends and the lower photo numbers. The results are illustrated as follows:

Table 17. Correlation analysis between the closest pair distance and the higher photo number

		The closest pair distance	Higher Photo Number
The closest pair distance	Pearson Correlation	1	-.170**
	Sig. (2-tailed)		.000
	N	92525	92525
Higher Photo Number	Pearson Correlation	-.170**	1
	Sig. (2-tailed)	.000	
	N	92525	92525

** . Correlation is significant at the 0.01 level (2-tailed).

Table 18. Correlation analysis between the closest pair distance and the lower photo number

		The closest pair distance	Lower Photo Number
The closest pair distance	Pearson Correlation	1	-.152**
	Sig. (2-tailed)		.000
	N	92525	92525
Lower Photo Number	Pearson Correlation	-.152**	1
	Sig. (2-tailed)	.000	
	N	92525	92525

** . Correlation is significant at the 0.01 level (2-tailed).

In Table 17, a Student's t-test is applied to test the null hypothesis that the Pearson's correlation coefficient equals to zero. The result of the t-test rejects the null hypothesis at the significant level of 0.01. In other words, the correlation coefficient between the closest pair distance and the higher photo numbers does not equal to zero. However, it is also noteworthy that the sample size in this test is relatively large (92525) and the correlation coefficient value (-0.170) is very close to zero. It implies that though the correlation coefficient does not equal to zero, the strength of the correlation is not strong either.

In Table 18, the null hypothesis that the correlation coefficient between the closest pair distance and the lower photo number equals to zero is rejected at the significant level of 0.01. In other words, the correlation coefficient between the closest pair distance and the lower photo numbers does not equal to zero. However, it is also noteworthy that the

sample size in this test is relatively large (92525) and the correlation coefficient value (-0.152) is very close to zero. It implies that though the correlation coefficient does not equal to zero, the strength of the correlation is not strong either.

5.3 Spatial visualization of geotagged photos: where are they and what happened?

The previous frequency analyses show that the closest pair distance between Flickr friends is closer than that of Flickr non-friends. In order to further explore this difference between Flickr friends and non-friends from geographical views, spatial distributions of the closest photo pairs are visualized.

Most of the closest photo pairs in the lower 48 states of the United States demonstrate the following geographic patterns:

- a) Concentrated in urban areas

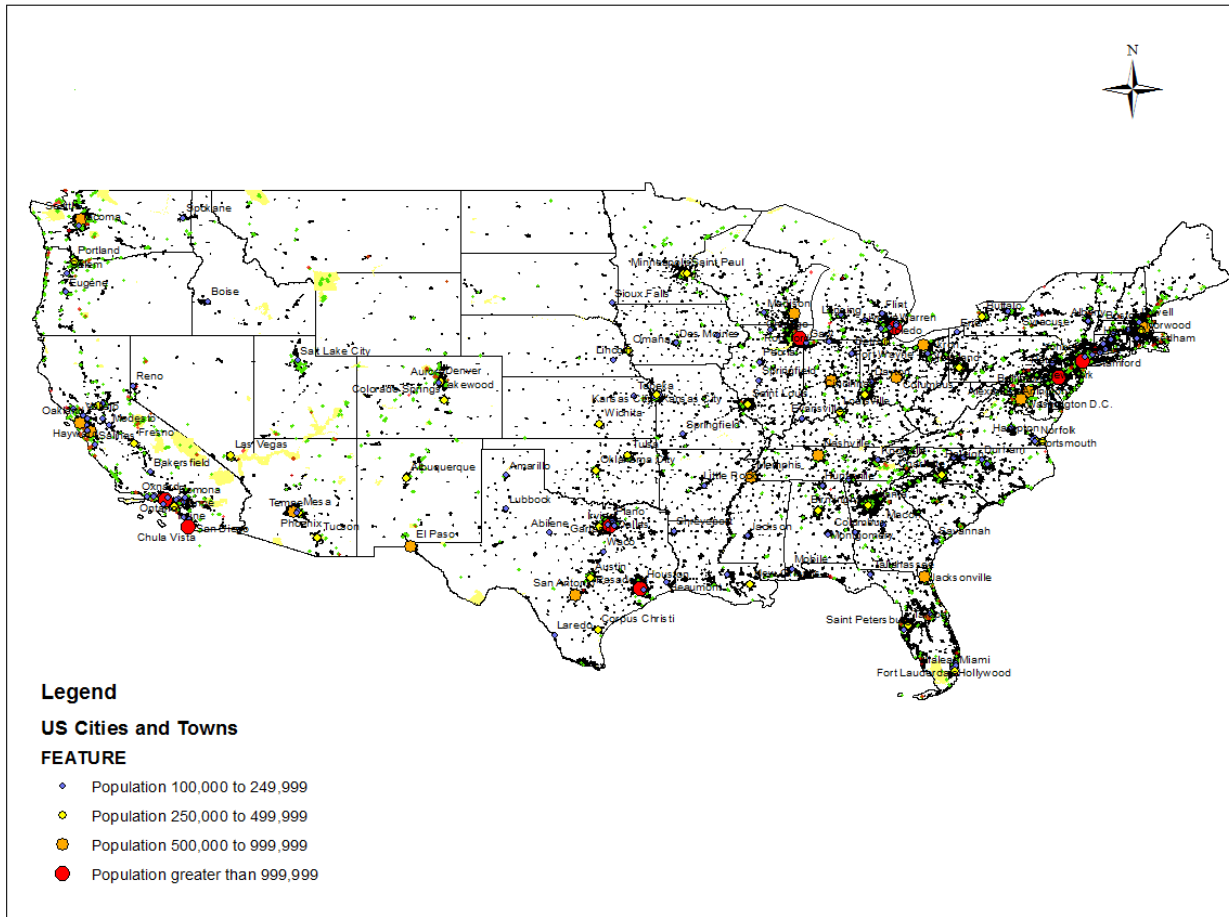


Figure 22. The closest photo pairs within 10 km in the lower 48 states

Take a short distance range (0 km to 10 km) for example. In Figure 22, the green lines represent the closest photo pairs within 10 km of Flickr friends, while the red lines represent those of Flickr non-friends. The blue points represent the main cities (population based) of the U.S. The yellow polygons represent national parks. The black polygons represent the U.S. urban areas. The urban area data are downloaded from the U.S. Census Bureau Tiger\Line dataset. On this map, most of the closest photo pairs are located in urban areas.

In the frequency analysis, 24.8% of the closest photo pairs of Flickr friends are within 10 km. In comparison, only 5% of the closest photo pairs of Flickr non-friends are within 10 km. In the lower 48 states of the U.S., there are 11,464 closest photo pairs within 10 km for Flickr friends and 2,457 pairs for Flickr non-friends. An overlay analysis shows that 9,873 of the 11,464 the closest photo pairs of Flickr friends (86%) are located in urban

areas. In comparison, 2,284 of the 2,457 pairs of Flickr non-friends (92%) are located in urban areas. The ratio between the numbers of the closest photo pairs within 10 km of Flickr friends and non-friends in the urban areas of the lower 48 states is 4.39. Figure 23 is the frequency map of the closest photo pairs within 10 km of Flickr friends in urban areas of the lower 48 states. In comparison, Figure 24 is the frequency map of the closest photo pairs within 10 km of Flickr non-friends in urban areas of the lower 48 states. These color maps are classified into 5 categories according to the *Natural Breaks (Jenks)* method in ArcGIS 10.

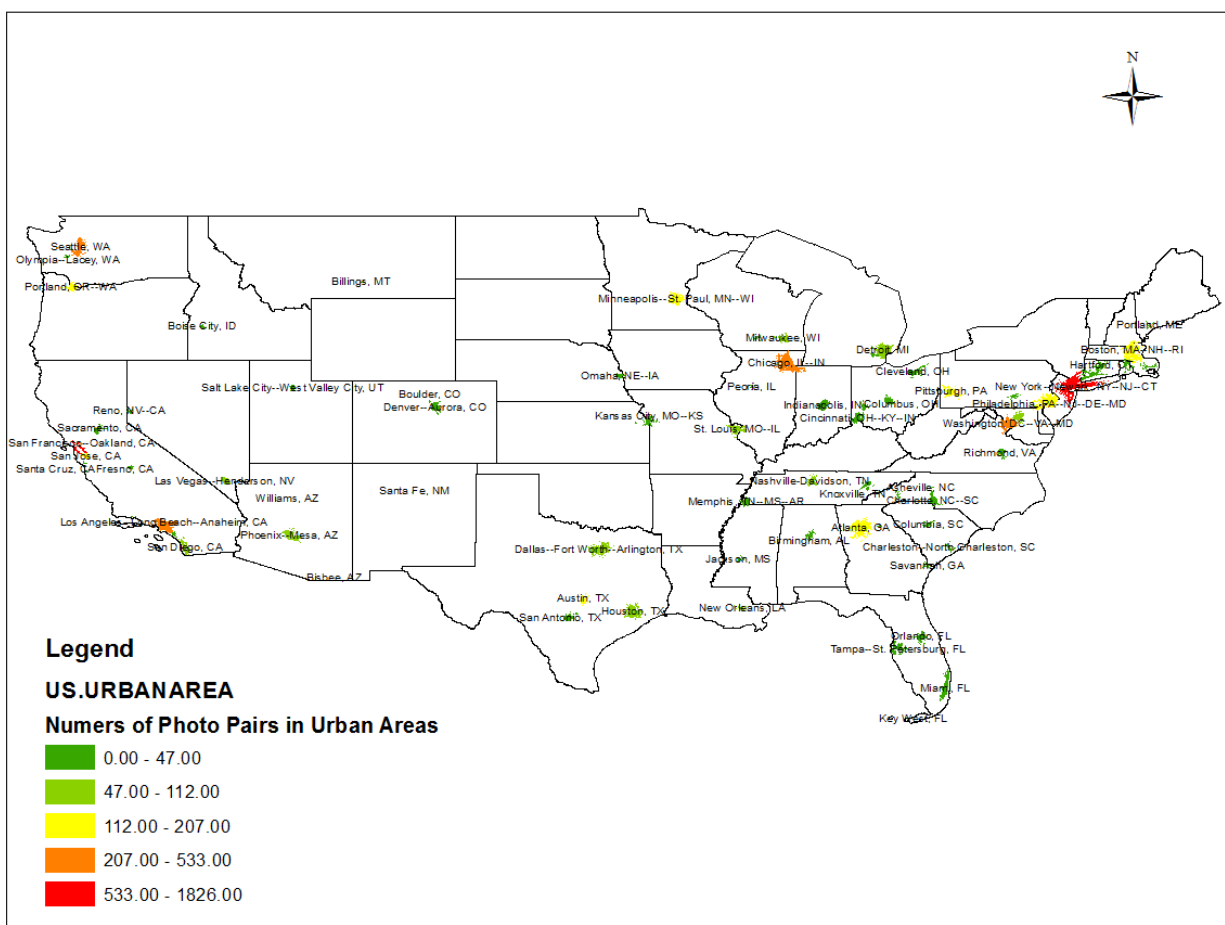


Figure 23. Frequency map of the closest photo pairs within 10 km of Flickr friends in urban areas of the lower 48 states

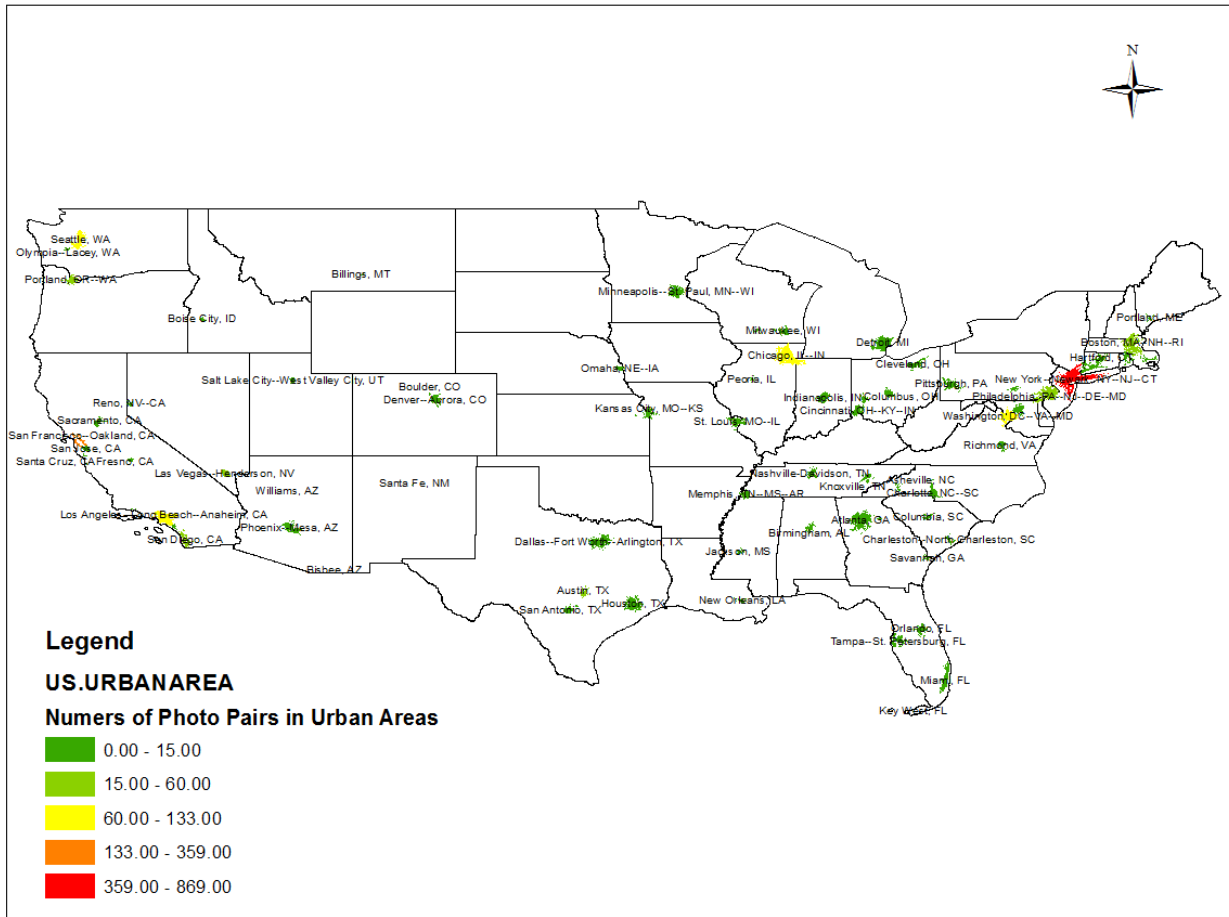


Figure 24. Frequency of the closest photo pairs within 10 km of Flickr non-friends in urban areas of the lower 48 states

In Figures 23 and 24 some urban areas have high frequencies for both Flickr friends and non-friends (e.g. New York City, San Francisco, etc.), while some urban areas are classified into different categories for Flickr friends and non-friends. To better compare these two distribution maps, the ratios between the frequencies of the closest photo pairs within 10 km of Flickr friends and non-friends for different urban areas of the lower 48 states are calculated and illustrated in Figure 25. The equation is as follows:

Since the frequency of non-friends is the denominator, areas with zero values for non-friends are excluded from this calculation. Moreover, since the average ratio for all the

urban areas in the lower 48 states is 4.39, ratios are classified manually to highlight their relationship with this average value.

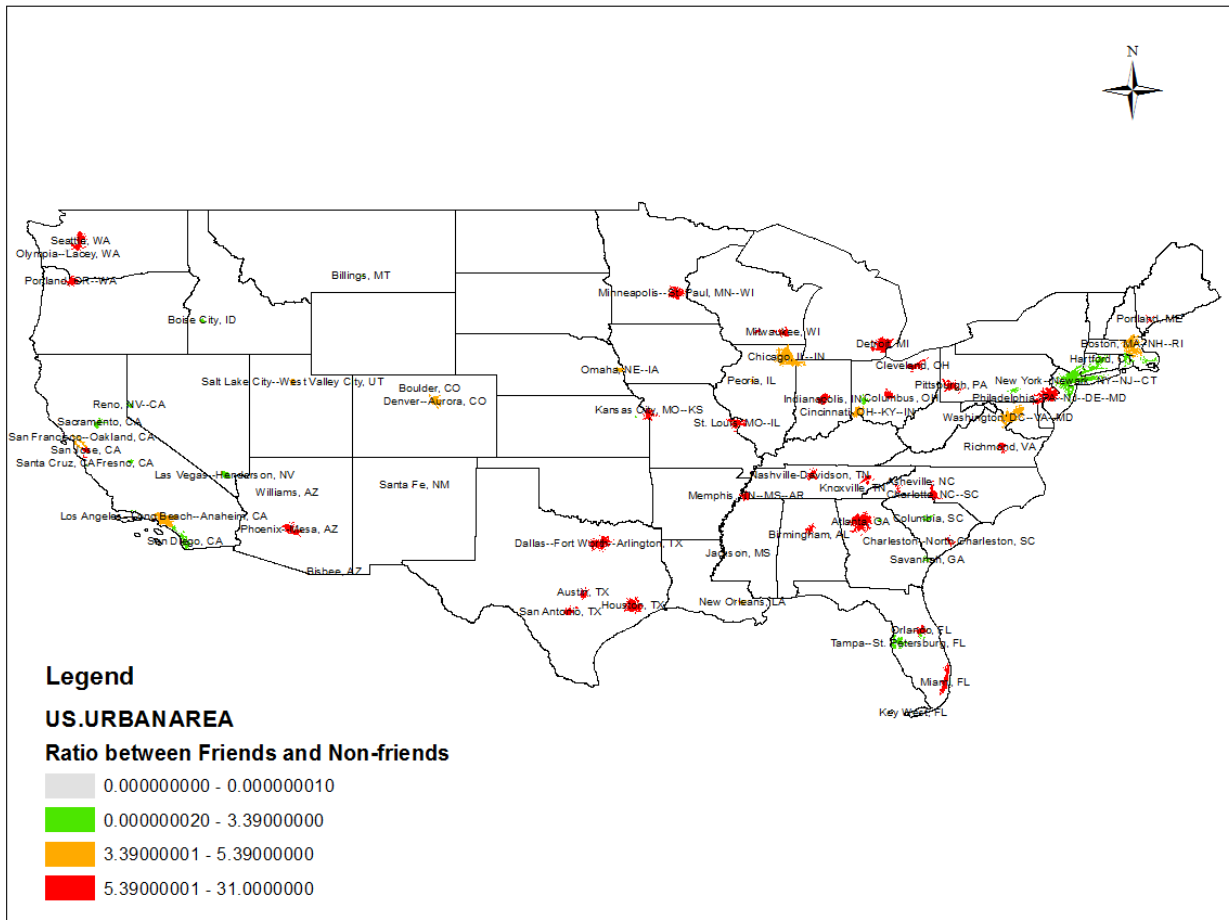


Figure 25. The ratios between the frequencies of the closest photo pairs within 10 km of Flickr friends and non-friends for urban areas in the lower 48 states

In Figure 25, the red color represents the high ratio (5.39 to 31.00). In these areas, the ratios between Flickr friends and non-friends are above the average (4.39). To better illustrate this comparison, the ten highest ratios are illustrated as follows:

Table 19. The ten highest ratios between the frequencies of the closest photo pairs within 10 km of Flickr friends and non-friends in urban areas of the lower 48 states

Name	Frequencies of the closest photo pairs within 10 km of Friends	Frequencies of the closest photo pairs within 10 km of Non-friends	Ratio
Detroit, MI Urbanized Area	93	3	31
Madison, WI Urbanized Area	45	2	22.50
Ann Arbor, MI Urbanized Area	43	2	21.50
Minneapolis—St. Paul, MN—WI Urbanized Area	122	6	20.33
Nashville-Davidson, TN Urbanized Area	112	6	18.67
Phoenix—Mesa, AZ Urbanized Area	106	4	17.67
Pittsburgh, PA Urbanized Area	141	8	17.63
Milwaukee, WI Urbanized Area	70	4	17.5
Charlotte, NC Urbanized Area	33	2	16.5
Columbus, OH Urbanized Area	30	2	15

Though the ratios in Table 19 are high, it is noteworthy that the frequencies, especially the frequencies for non-friends, are very low. This may be the main reason that the ratios for these areas are much higher than the average value (4.39) of the lower 48 states. Therefore, areas with high frequencies may be more informative to illustrate the difference the closest photo pairs within 10 km between Flickr friends and non-friends. Consequently, the ten highest frequencies of the closest photo pairs within 10 km in urban areas are illustrated as follows:

Table 20. The ten highest frequencies of the closest photo pairs within 10 km in urban areas

Flickr Friends		Flickr Non-friends	
Urbanized Area	Frequency	Urbanized Area	Frequency
New York City--Newark, NY-- NJ—CT	1826	New York City--Newark, NY-- NJ—CT	869
San Francisco--Oakland, CA	1538	San Francisco--Oakland, CA	359
Chicago, IL—IN	533	Washington, DC--VA—MD	133
Washington, DC--VA--MD	474	Chicago, IL—IN	108
Seattle, WA	470	Los Angeles--Long Beach— Anaheim, CA	103
Los Angeles--Long Beach-- Anaheim, CA	417	Seattle, WA	72
Boston, MA--NH—RI	207	Boston, MA--NH—RI	60

Table 20. Continued

Flickr Friends		Flickr Non-friends	
Urbanized Area	Frequency	Urbanized Area	Frequency
Portland, OR—WA	204	Las Vegas-- Henderson, NV	44
Austin, TX	167	San Diego, CA	33
San Jose, CA	154	Portland, OR—WA	30

In Table 20, the frequencies of “New York City--Newark, NY--NJ--CT” and “San Francisco--Oakland, CA” are higher than other urban regions. The ratio of frequencies between Flickr friends and non-friends in “New York City--Newark, NY--NJ--CT” is 2.10. In comparison, the corresponding ratio for San Francisco is 4.28.

To better understand the differences of geotagging activities inside these regions, the spatial distributions of the closest photo pairs inside New York City and San Francisco are provided as follows:

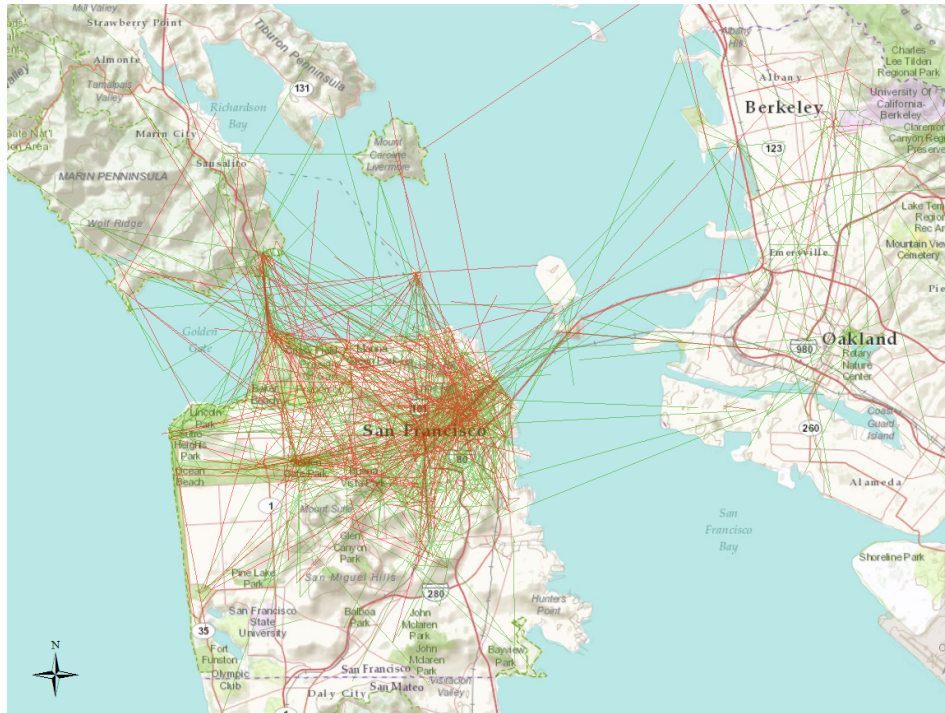


Figure 26. The closest photo pairs within 10 km in San Francisco, CA

Figure 26 shows the spatial distribution of the closest photo pairs within 10 km in San Francisco, CA. The red lines represent the photo pairs of Flickr non-friends while the green lines represent those of Flickr friends. On this map, the spatial distributions of the green lines and the red lines are very similar. A large proportion of them are concentrated in northeastern San Francisco, where downtown and many tourist attractions (e.g. Fisherman's Wharf, Pier 39, etc.) are located. Figure 27 overlaps the closest photo pairs with the population density map of San Francisco. In order to differentiate from the population density map, the closest photo pairs of Flickr non-friends are connected with blue lines. At this specific scale, there are several highly populated block groups highlighted in dark red. However, the closest photo pairs cover only part of these highly populated block groups. For example, in southern San Francisco, several highly populated block groups are barely covered by green/blue lines. From this perspective, population may not be the only factor in deciding the spatial distribution of the closest photo pairs within 10 km inside the urban areas of San Francisco.

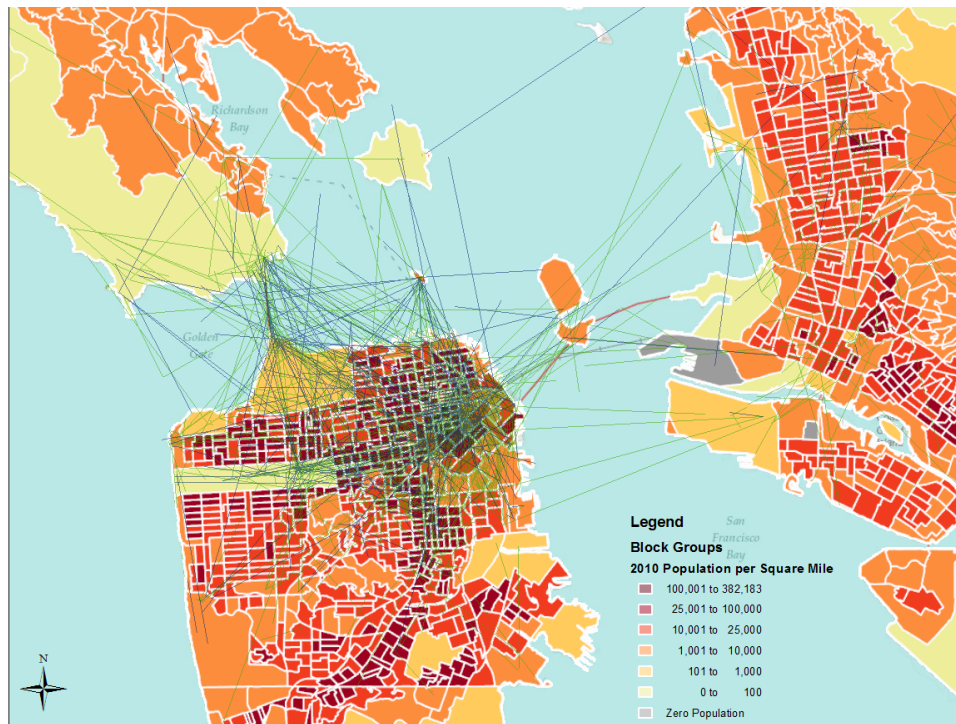


Figure 27. The closest photo pairs within 10 km and the population density in San Francisco, CA

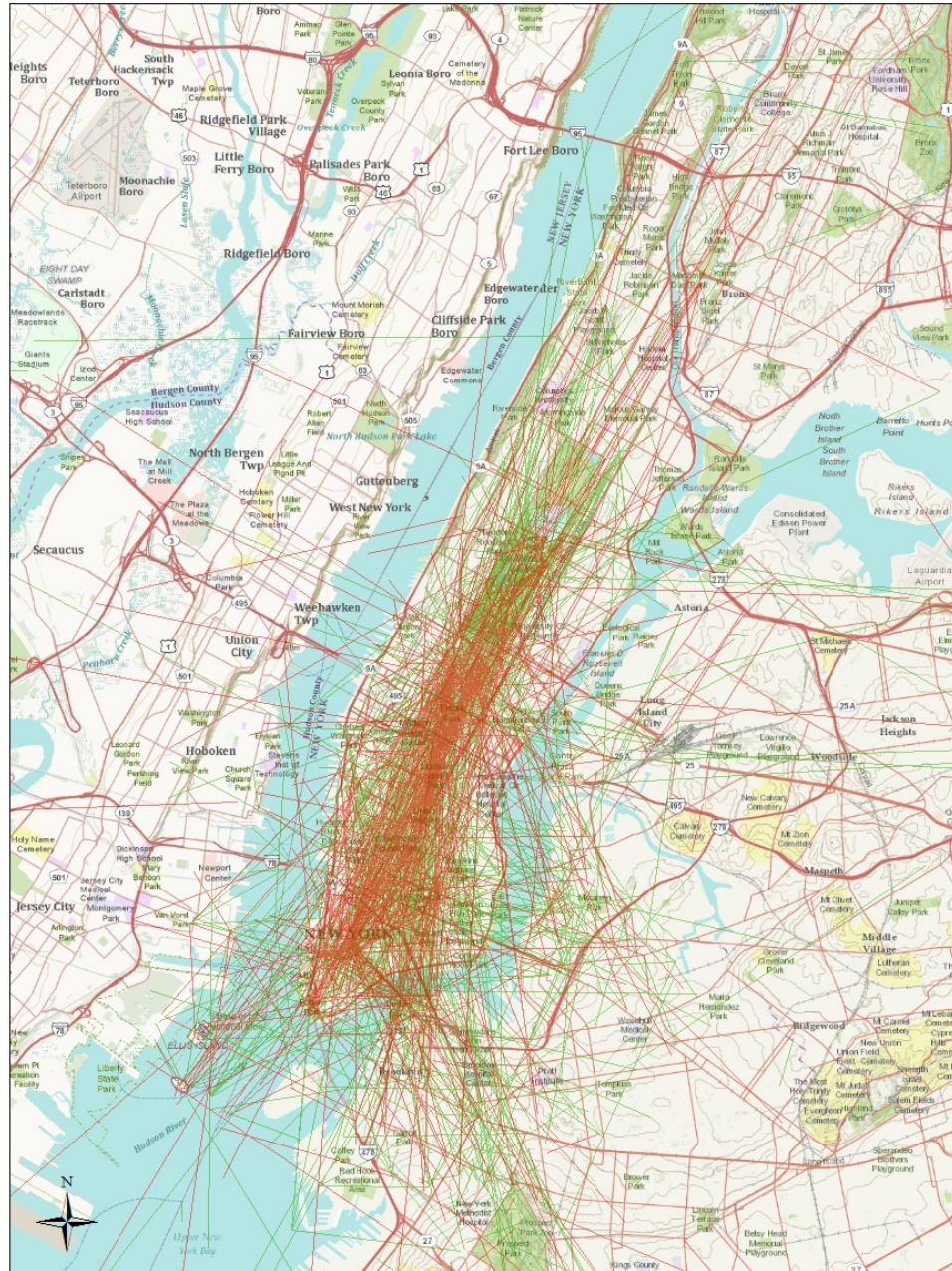


Figure 28. The closest photo pairs within 10 km in Manhattan, NY

In New York City, most of the closest photo pairs within 10 km of Flickr friends and non-friends are concentrated in Manhattan (see Figure 28). Inside Manhattan most of the tourist attractions are located in Midtown or Downtown. In comparison, Uptown Manhattan has fewer tourist attractions. From this perspective, regions with more tourist attractions are more geotagged than others inside Manhattan. Additionally, it is noteworthy that Uptown is mostly covered by green lines. In other words, Uptown

seems to be geotagged more by friends than non-friends. A potential explanation is that the closest photo pairs in Uptown are less likely to be taken by tourists.

Figure 29 shows the population density and the closest photo pairs within 10 km inside Manhattan. At this specific scale most Manhattan areas, including Uptown, are highly populated. This further supports our conclusion that population is not the only factor in deciding the spatial distribution of the closest photo pairs inside urban areas.

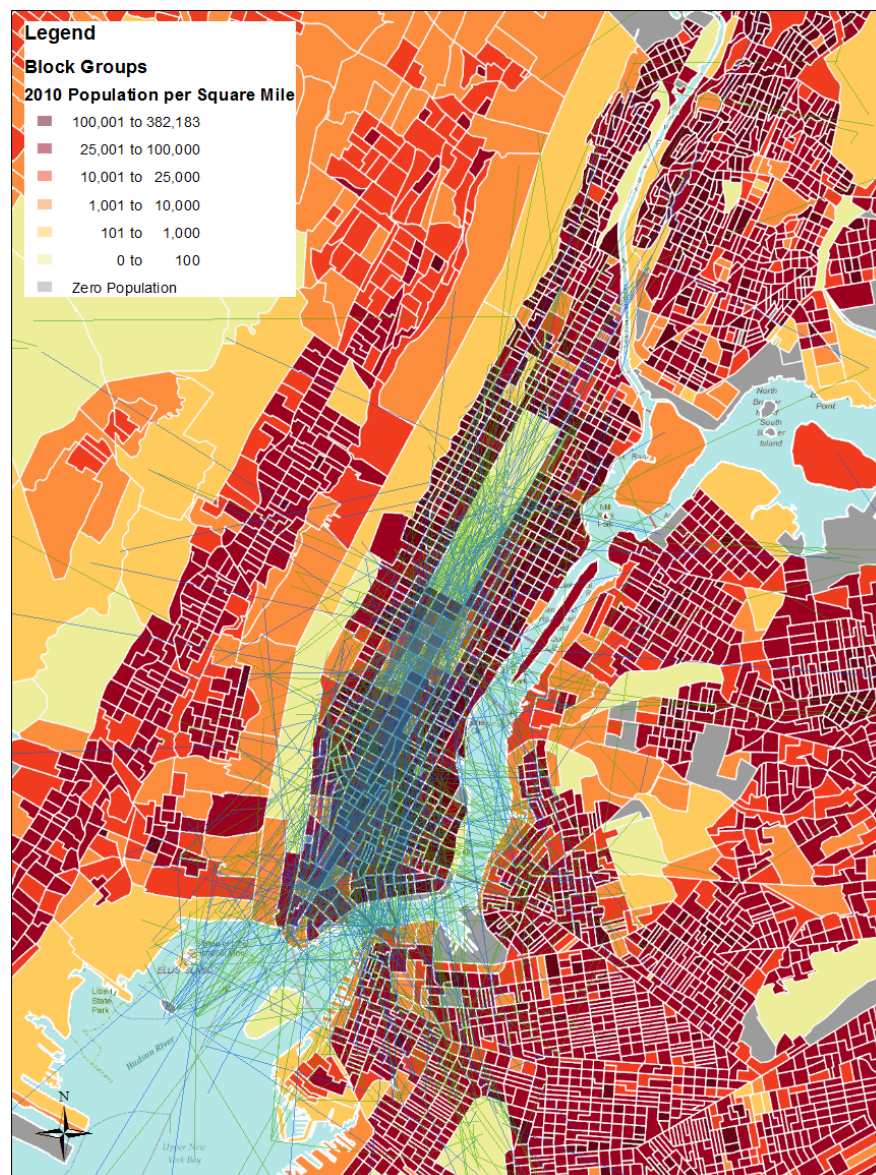


Figure 29. The closest photo pairs within 10 km and the population density in Manhattan, NY

In summary, for the closest photo pairs within 0 km and 10 km, most of them are concentrated in highly populated urban areas. However, population may not be the only factor in deciding the spatial distribution of these photo pairs inside urban areas.

b) Scattered in national parks

In the frequency analysis, 24.8% of the closest photo pairs of Flickr friends are within 10 km. In comparison, 5% of those of Flickr non-friends are within 10 km. In the lower 48 states of the U.S., there are 11,464 closest photo pairs within 10 km for Flickr friends and 2,457 pairs for Flickr non-friends. An overlay analysis shows that 368 out of the 11,464 pairs from Flickr friends (3.2%) are located in national parks. In comparison, 64 out of the 2,457 pairs from Flickr non-friends (2.6%) are located in national parks. The ten highest frequencies of the closest photo pairs within 10 km in national parks are illustrated as follows:

Table 21. The ten highest frequencies of the closest photo pairs within 10 km in national parks

Flickr Friends		Flickr Non-friends	
Park Name	Frequency	Park Name	Frequency
Golden Gate	59	West Potomac Park	15
Yosemite	36	Yosemite	8
West Potomac Park	27	Grand Canyon	6
Yellowstone	18	Golden Gate	4
Mississippi	15	Washington Monument and Grounds	4
Washington Monument and Grounds	13	Santa Monica Mountains	3
Grand Canyon	12	Crater Lake	2

Table 21. Continued

Flickr Friends		Flickr Non-friends	
Urbanized Area	Frequency	Urbanized Area	
Grand Teton	10	Fort Point	2
Mount Rainier	10	Rocky Mountain	2
Presidio of San Francisco	10	The Mall, Seaton Park	2

In Table 21, the highest frequency is only 59. In other words, national parks are not as attractive as urban areas for the closest photo pairs within 10 km.

c) Connect two cities

At longer distance ranges, a large number of the closest photo pairs connect two different cities. Take the distance range between 300 km and 550 km for example. Figures 30 and 31 show the spatial distributions of the “best of best” photo pairs within this distance range in the lower 48 states. The points are the U.S. cities and towns with 100,000 populations or higher. Some distinct patterns are shown. First, most of the closest photo pairs connect two different cities. The connections demonstrate a “hub and spoke” pattern where cities with higher populations tend to be at the center. Second, the Northeastern and the Southwestern U.S. are covered by more photo connections. According to the patterns in Figures 30 and 31, population size seems to be very influential on the spatial distributions of the closest photo pairs between cities.

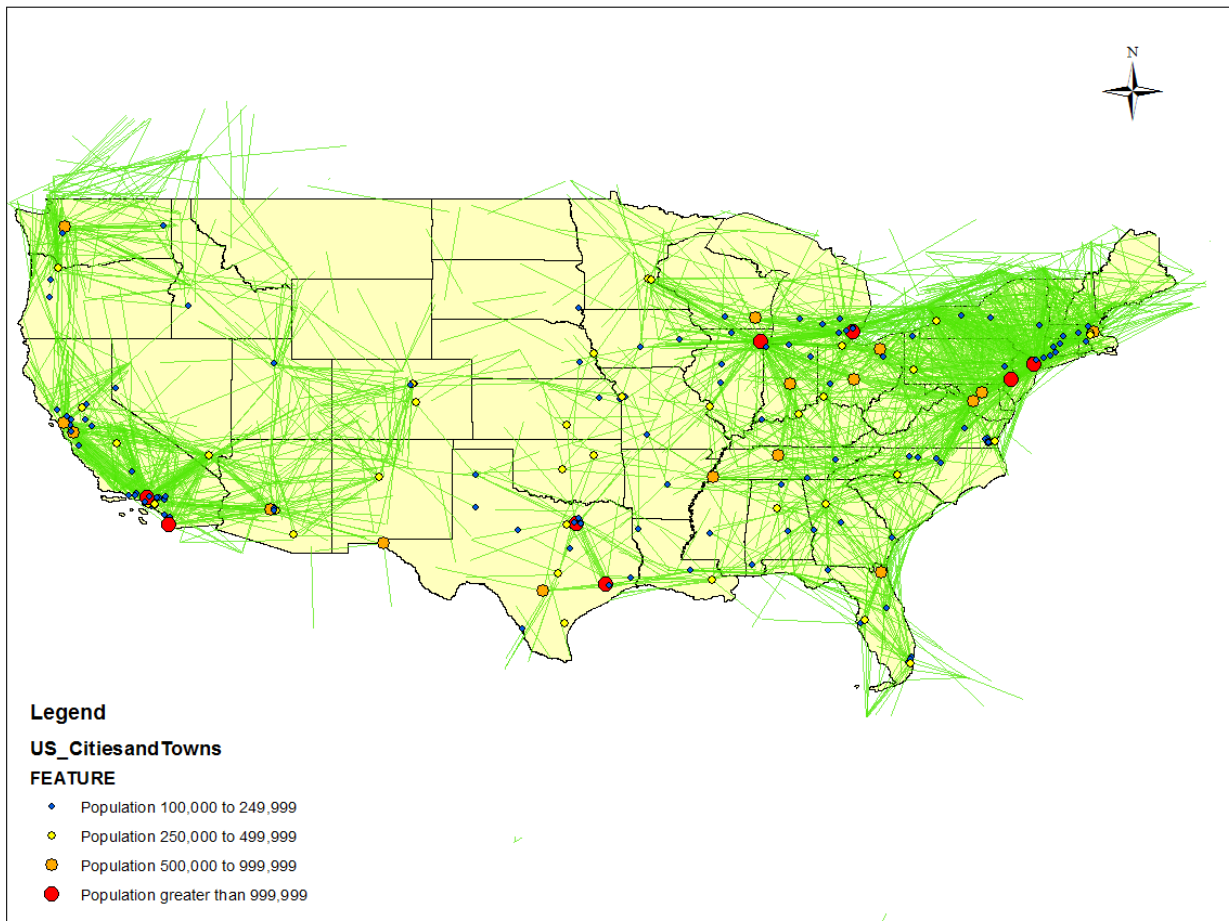


Figure 30. The closest photo pairs between 300 km and 550 km of Flickr friends in the U.S.

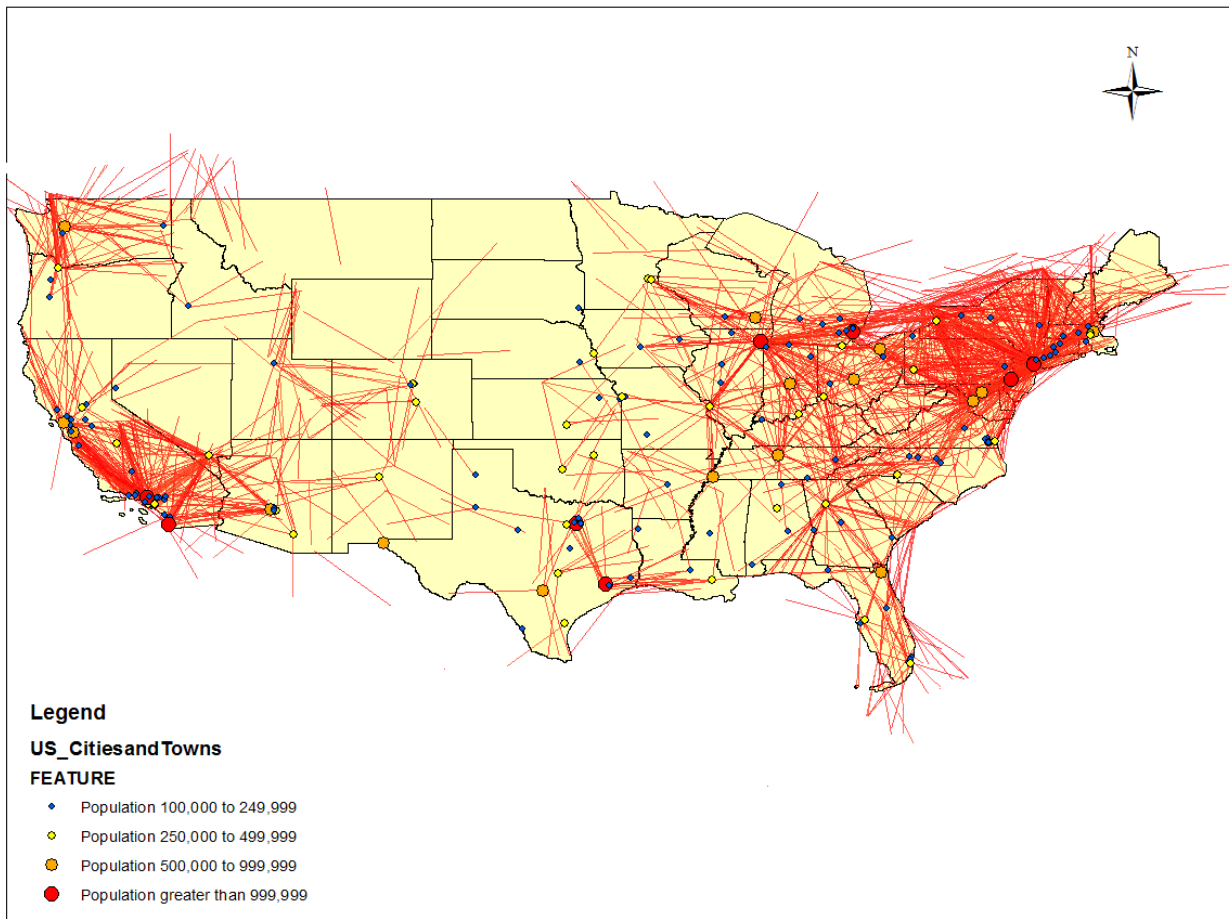


Figure 31. The closest photo pairs between 300 km and 550 km of Flickr non-friends in the U.S.

Figures 30 and 31 also reveal the potential relationship between the physical distance and virtual connections. The “hub and spoke” patterns between cities at this distance range provide some clues that may explain how virtual connection and physical distance interact with each other. At this distance range, most of the closest photo pairs are located in different cities. Very few of the closest photo pairs are located in rural areas. In other words, at this specific distance range the physical proximity between Flickr users reflected by their closest photo pairs is more frequently identified in cities than in rural areas.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

6.1 Conclusions

This empirical study investigates the relationships between geotagging activities of Flickr users and their online friendship from spatial and temporal perspectives. It not only explores these relationships from a statistical perspective, but also leverages the visualization capability of geographic information systems to investigate their geographic meanings. By calculating, comparing and analyzing different kinds of distances between the geotagged photos of Flickr friends and non-friends, it is shown that spatial distances between Flickr geotagged photos are related to Flickr friendship. In the analysis of the closest photo pairs between Flickr users, more Flickr friends tend to have their closest photo pairs within a shorter distance range than Flickr non-friends do. In the analysis of the nearest neighbor photo pairs between Flickr users, a similar pattern is also found.

The importance of the relationships above is supported by another finding. Statistical tests on the closest pair distance and the number of geotagged photos posted by Flickr users show that these two features are not strongly related. This finding clarifies a widely addressed issue of whether it is meaningful to study the distance between geotagged photos of OSN users since different users post different numbers of geotagged photos. These findings further stress the possibility and feasibility to leverage the spatial information of geotagged photos in the analysis of online social network users.

In order to study the geographic meanings under the distribution of the closest photo pairs between Flickr users, this study incorporates the visualization capacity of GIS. The spatial distributions of the closest photo pairs within various distance ranges are visualized. The distributions for friends and non-friends are compared with each other. In this paper, we select two typical spatial thresholds to demonstrate some spatial patterns. Between 300 km and 550 km, the distributions of the closest photo pairs

demonstrate a “hub and spoke” pattern among the cities of the U.S. Since most hubs are cities with large populations, population is considered as an important factor in determining the spatial distributions of the closest photo pairs within this distance range. Between 0 km and 10 km, most of the closest photo pairs are located in urban areas. However, by overlapping the block group population map with the closest photo pair distribution map in New York City and San Francisco, it shows that many highly populated block groups are barely covered by any photo pairs within 10 km. In this case, population may not be the only factor in determining the spatial distributions of the closest photo pairs inside urban areas. In fact, most tourist attractions in these two cities are intensively covered by many photo pairs. This reminds us about the potential roles of tourist in determining Flickr users’ geotagging activities.

Statistics of different kinds of distances between geotagged photos also reveal some boundary effects on Flickr users’ geotagging activities. All four frequency distributions in Table 15 demonstrate low points at around 4,000 km. By visualizing the closest photo pairs at this distance, we find that physical boundaries (e.g., presence of oceans) are the main causes for this phenomenon. Though online social network is supposed to weaken the role of physical distance by connecting people from different places, some specific physical constraints still influence the geotagging activities on OSN.

In addition to the analyses of spatial features of Flickr users’ geotagging activities, this project also studies the temporal features of geotagged photos. The statistical tests show that the time difference between the closest photo pair is related to Flickr friendship. The importance of this relationship depends on how we explore the spatial and temporal features of geotagged photos. On one hand, it is found that most Flickr friends do not reveal their online friendship by geotagging photos taken within short distance and short time interval. Very few of the closest photo pairs between Flickr friends are within a relatively small spatio-temporal threshold (e.g. ten kilometers and one day). In other words, the “co-existences” of Flickr users revealed by their geotagged photos are quite limited. This finding therefore questions the effectiveness of friendship inferences based on “spatio-temporal co-occurrence” of geotagged photos between Flickr users.

On the other hand, analyses show that temporal features of geotagged photos may be more sensitive to Flickr friendship. For the closest photo pairs of Flickr friends within 10 km, only 0.98% of them are within one day. For the closest photo pairs of Flickr friends within one day, 64% of them are within 10 km. Since the strict “co-existences” of Flickr users revealed by their geotagged photos are quite limited, more attention may be addressed to the cases of “co-location in space” or “co-location in time”. These findings demonstrate the potential of temporal features in the study of online friendship and may direct more focus to the temporal perspective of Flickr users’ geotagging activities.

6.2 Limitations and Future Research

Though this empirical study answers some questions with a large geotagged photo dataset from a well-known OSN, some embedded limitations in our analytical process reveal the potential for improvement in future research.

First, though this project visualizes the closest photo pairs between Flickr users at many different distance ranges, only two distance ranges (0 km to 10 km and 300 to 500 km) are selected to demonstrate some typical distribution patterns of the closest photo pairs. This selection is arbitrary. In addition, one day and ten kilometers have been used as small spatial and temporal intervals in this study. However, in many contexts, one day can be a relatively long time interval and ten kilometers can be a relatively long distance range. To avoid arbitrary selected spatial or temporal thresholds, future research needs to compare geotagged photos under different spatial and temporal thresholds and identify more clearly defined patterns based on these comparisons.

Second, due to computational limitation, only a small part of our downloaded geotagged photos are included in this project. The larger datasets have not been fully explored. For the statistical test on the relationship between the spatial distance of geotagged photos and Flickr friendship, the size of our sample data is large enough. However, in the visualization analysis, the number of the closest photo pairs for Flickr non-friends is quite limited in some urban areas. This leads to a very high ratio between the closest photo pair frequencies of Flickr friends and non-friends in these areas. Without enough samples, this ratio can be quite misleading. In the future research, more data should be

included to better explore the distributions of geotagged photos from geographical perspectives.

Third, this study analyzes the population and its influence on the spatial distribution of the closest photo pairs between Flickr users. However, according to our findings population is not the only feature which influences Flickr users' geotagging activities. Tourism, infrastructure, transportation and other geographic contexts may also be influential in determining the spatial distribution of Flickr geotagged photos. Future research should take various geographic contexts into consideration. During this process, the scale issue needs to be carefully addressed. Currently we study the relationship between population and the spatial distribution of the closest photo pairs at the block group level. However, the modifiable areal unit problems may influence the results from the analyses at different scales. Future research needs to address these issues by analyzing the influence of geographic contexts on Flickr users' geotagging activities under different scales.

LIST OF REFERENCES

- Adams, P., 1995. A reconsideration of personal boundaries in space-time. *Annals of the Association of American Geographers* 85(2), 267-285.
- Adams, P., 2009. *Geographies of Media and Communication*. New York: Wiley-Blackwell.
- Adamic, L., Adar, E., 2008. How to search a social network. *Social Networks* 27 (3), 187-203.
- Ahern, S., Naaman, M., Nair, R., Yang, J., 2007. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1-10.
- Backstrom, L., Sun, E., Cameron, M., 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. *WWW'10 Proceedings of the 19th International Conference World Wide Web*, 61-70.
- Bernard, H. R., Killworth, P. D., McCarty, C., 1982. Index: An informant-defined experiment in social structure. *Social Forces* 61 (1), 99-133.
- Boyd, D., Crawford, K., 2011. Six provocations for big data. *A decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, September 2011. Available at SSRN: <http://ssrn.com/abstract=1926431> or <http://dx.doi.org/10.2139/ssrn.1926431>
- Buliung, R., Kanaroglou, P., 2004. On design and implementation of an object-relational spatial database for activity/travel behavior research. *Journal of Geographical Systems* 6, 237-262.
- Caverlee, J., 2010. Towards web-scale geo-semantic crowd discovery. *2010 Specialist Meeting – Spatio-Temporal Constraints on Social Networks*. Available online: <http://www.ncgia.ucsb.edu/projects/spatio-temporal/docs/Caverlee-position.pdf>

- Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg J., 2009. Mapping the world's photos. WWW'09 Proceedings of the 18th International Conference on World Wide Web, 761-770.
- Crandall, D., Bakcktrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinburg, J., 2010. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Science, Dec. 28, 2010 107 (52), 22436-22441.
- Dodds, P.S., Muhamad, R., Watts, D. J., 2003. An experimental study of search in global social networks. Science 301, 827-829.
- Elwood, S., 2010. Spatiality, temporality, and contexts: Geosocial data as evidence of social interactions and networks. 2010 Specialist Meeting – Spatio-Temporal Constraints on Social Networks. Available on line:
<http://www.ncgia.ucsb.edu/projects/spatio-temporal/docs/Elwood-position.pdf>
- Forsyth, D., 2010. The power of dynamic spatial and temporal characterization in social networks. 2010 Specialist Meeting – Spatio-Temporal Constraints on Social Networks. Available online: <http://www.ncgia.ucsb.edu/projects/spatio-temporal/docs/Forsyth-position.pdf>
- Girardin, F., Blat, J., Calabrese, F., Dal Fiore, F., Ratti, C., 2008. Digital footprinting: uncovering tourists with user-generated content. Pervasive Computing, IEEE 7.4, 36-43.
- Gilbert, E., Karahalios, K., Sandvig, C., 2008. The network in the garden: an empirical analysis of social media in rural life. Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, 1603-1612.
- Hägerstrand, T., 1970. What about people in regional science? Papers in Regional Science 24 (1), 6-21.
- Hollenstein, L., Purves, R.S., 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. Journal of Spatial Information Science 1, 21-48.

- Killworth, P., Bernard, H., 1978. Reverse small world experiment. *Social Networks* 1, 159-192.
- Kleinberg, J.M., 2000. Navigation in a small world. *Nature* 406, 845.
- Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N., 2008. Tastes, ties, and time: A newsocial network dataset using Facebook.com. *Social Networks* 30(4), 330-342.
- Lenntorp, B., 1977. Paths in space-time environments: a time geographic study of movement possibilities of individuals. *Environment and Planning A* 9(8), 961-972.
- Liben-Nowell, D., Novak J., Kumar R., Raghavan, P., Tomkins, A., 2005. Geographic routing in social networks. *Proceedings of the National Academy of Science* 102(33), 11623.
- Milgram, S., 1967. The small-world problem. *Psychology Today* 1, 62-67.
- Miller, H.J., 2005. Necessary space-time conditions for human interaction. *Environment and Planning B* 32, 381-401.
- Miller, H.J., 2010. The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science* 50(1), 181-201.
- Ratti, C., Girardin, F., Blat, J., Dal Fiore, F., 2008. Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. *Journal of Location Based Services* 2(1), 41-56.
- Shaw, S.-L., Wang, D., 2000. Handling disaggregate spatio-temporal travel data in GIS. *Geoinformatica* 4(2), 161-178.
- Shaw, S.-L., Yu, H., Bombom, L., 2008. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12(4), 425-441.

- Shaw, S.-L., Yu, H., 2009. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical-virtual space. *Journal of Transport Geography* 17 (2), 141-149.
- Shaw, S.-L., 2010. Relevance of time geography to spatio-temporal constraints on social networks. 2010 Specialist Meeting – Spatio-Temporal Constraints on Social Networks. Available online: <http://www.ncgia.ucsb.edu/projects/spatio-temporal/docs/Shaw-position.pdf>
- Wang, D., Cheng, T., 2001. A spatio-temporal data model for activity-based transport demand modeling. *International Journal of Geographical Information Science* 15(6), 561-585.
- Sui, D., 2010. A geographic conceptual framework for understanding the spatio-temporal constraints on social networks. 2010 Specialist Meeting – Spatio-Temporal Constraints on Social Networks. Available online: <http://www.ncgia.ucsb.edu/projects/spatio-temporal/docs/Sui-position.pdf>
- Sui, D., Goodchild, M., 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science* 25(11), 1737-1748.
- Watts, D., Dodds, P., Newman M., 2002. Identity and search in social networks. *Science* 296 (5571), 1302-1305.
- Yu, H., Shaw, S., 2008. Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *International Journal of Geographic Information Science* 22(4), 409-430.

APPENDIX

The spatial distance between geotagged photos calculated in this project is the great circle distance. The codes to calculate the great circle distance between the geographic coordinates between two photos are as follows:

```
public double distance(double lat1, double lng1, double lat2, double lng2)
{
    double EARTH_RADIUS = 6378.137;
    double radLat1 = rad(lat1);
    double radLat2 = rad(lat2);

    double radLng1 = rad(lng1);
    double radLng2 = rad(lng2);

    double a = radLat1 - radLat2;
    double b = radLng1 - radLng2;
    double s = 2 * Math.Asin(Math.Sqrt(Math.Pow(Math.Sin(a / 2), 2) +
    Math.Cos(radLat1) * Math.Cos(radLat2) * Math.Pow(Math.Sin(b / 2), 2)));
    s = s * EARTH_RADIUS;
    s = Math.Round(s * 10000) / 10000;
    return s;
}
```

VITA

Sumang Liu was born in Hubei, China. He received his Bachelor of Science degree from Wuhan University and entered the University of Tennessee, Knoxville in 2010 to pursue a Master's Degree in the Department of Geography.