# ANALYZING THE DYNAMICS BETWEEN THE USER-SENSED DATA AND THE REAL WORLD

Haipeng Zhang

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing,

Indiana University

September 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

---

Assistant Professor David J. Crandall, PhD, Committee Chair

---

Assistant Professor Yong-Yeol (YY) Ahn, PhD

---

Associate Professor Johan Bollen, PhD

---

Professor David Leake, PhD

August 22, 2014

# ACKNOWLEDGMENTS

I am very grateful to my parents for their unconditional love and support through all these years. The longer time that I am away from home, the more I understand how hard it was to raise me up. We are culturally too reserved to speak out our emotions. But I do feel your love in each and every aspect of life. My achievements always have your contribution and dedication. Thank you, my first teachers and life advisors. Here is a fun piece of the past. When I was born, my head was big. My grandfather said: Haipeng has a PhD head. Now my earliest self-suggestion comes true. I also realize that a PhD degree takes more than a big head and it is not the result of my solo effort. There is my family that always supports me. Tomorrow is the Mid-Autumn Festival. I am eating a moon cake while writing this and missing all of you.

I also want to thank my advisor, Prof. David Crandall. David, I really appreciate your guidance and patience. I still remember the metaphor that you used when we talked about the advisor-advisee relationship in 2010. I think it is very precise. I remember how you patiently guided me through when I first dabbled in research, how we brainstormed ideas, how we worked together on deadlines and how you helped me practice all these presentations. You taught me the research methods as well as the academic attitude that has a lifelong impact. I remember the dilemma that I was in and you helped me make the correct decision. Here is another fun piece of memory. It was in 2011. You, Mohammed and I were working intensely on a paper deadline in the evening. You came to the lab and told us that you had to leave for a while to let your dog out. I always smile when I think about this. I guess the takeaway message is: work hard, but still take it easy! The research went on

smoothly and I got the chances to see the world. Thanks, David. I would also like to thank other professors on my research committee: Prof. Yong-Yeol (YY) Ahn, Prof. Johan Bollen and Prof. David Leake. Thank you for all these helpful suggestions and insightful discussions.

My mentors and colleagues at various internships also helped me grow. Dr. Xiaohan Yun was my first manager and I always feel fortunate about getting all these study advice and career advice from him. Jacky Zhu, Prof. Takasu Atsuhiro, Prof. Masada Tomonari, Dr. Zeqian Shen, Dr. Lucas Joppa, Dr. Zhixian Yan, Dr. Jun Yang, Dr. Emmanuel Munguia Tapia, Nish Parikh, Gyanit Singh and Dr Neel Sundaresan: thank you for being mentors and friends that widened my world.

I would also like to thank the people who shared my joys and tears. Kun, Jingya, Mohammed, Sven, Stefan, Jerome: thanks for being the great lab mates and friends! Kun, if you were not here, my life would have been much less exciting – I wish you success in your new endeavors. Mohammed, thanks for guiding me into research – I still remember our long talk in the library in 2010. Sven, please keep on the inspiring drawings! Stefan, thank you for opening up the gym world to me! Yin, I always learn a lot from you, even though we did not see each often these years. Distance is never a problem. Vahid, do you remember how we worked on the projects in Lindley? These are tough good old days! Xiaoyong, I hope I have learned your tennis skills – I was just never competent to play against you! Jared, Qing, Huina, Xin, Zheng, Mingjie, Yangyi, Luyi, Peng, Liang, Yongan, Linger, Erkang, Qieyun, Min, Jianfeng, Can, Zhenghao, Rong, Simo, Jingru, Yajia, Guangchen, Yue, Yifan, Zhou, Qiqi, Junjie, Kunpeng, Tao, Chenren, Jiayuan, Xiaohui, Donghua, Jinjie, Lei, Tong, Xiaolu, Feifei, Fei, Jiawei... There are too many of you that I want to thank and there are so many warm memories.

Thank you for helping me reach a new chapter of my life.

Haipeng Zhang

ANALYZING THE DYNAMICS BETWEEN THE USER-SENSED DATA AND THE REAL WORLD

Modern technologies like smartphones and social media websites have created new ways for people to communicate and interact. In the process, these Web and mobile users are also creating collections of fine-grained and large-scale digital observational data about the world. For example, many users publicly share photos and status updates that can serve as visual or textual records of the state of the world and people across time and space. Meanwhile, many mobile users have apps that monitor their GPS trajectories and physical activity from onboard accelerometers and gyroscopes, for recording exercise and other purposes. Combined together, this web and phone-sourced 'user-sensed data' is potentially useful in observing and predicting the real world and people's behavior from new perspectives, overcoming traditional observational tools' restrictions such as expense and coverage. But user-sensed datasets are usually large, noisy and biased, given that they are generated by millions of users in uncontrolled conditions, making it unclear what types of useful and reliable observations can actually be extracted.

In this thesis we design techniques for analyzing and characterizing geo-temporal properties of these datasets, investigating how they could give us credible observations about the real world.

Using this framework, we investigate applications in four main threads of research. First, we analyze the geo-tags, timestamps, text tags, and visual content of 150 million online photos to measure and quantify the occurrence of ecological phenomena, comparing against ground truth at a continental scale. Second, we investigate and visualize the geo-temporal relationships between photo text tags, showing that these reveal connections between real world concepts. Third, we study the relationship between social media and e-commerce websites, showing that signals from social media can predict aggregate consumer behavior. Finally, we study how simple behavioral statistics of mobile users over time can characterize and even uniquely identify them.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 VISION

Ever since the dawn of our kind, humans have been trying to observe, understand and influence the physical environment. The desire to explain and even manipulate various phenomena has driven the progress of science for thousands of years. Thanks to the rapid development in science, humans are witnessing the world not solely by eyes, but also with the aid of advanced instruments such as thermometers, cameras, satellites and telescopes. Still these tools are not perfect — their expenses, areas of deployment, types of phenomena being detected and physical conditions all restrict the availability, quantity and quality of the observations.

The Internet, together with vastly available smart mobile devices, links billions of people, allows them to communicate anytime and anywhere and creates a gigantic virtual world. For instance, Twitter has 230 million active users posting over 200 million microblogs (tweets) per day, 76% of whom log in via mobile devices [109, 115]. Meanwhile over 1 million new images are uploaded to Flickr daily, with the smart mobile phone iPhone being the most popular camera [36, 63]. Like the physical world, this world has been studied in order to model the web, understand the user society, and improve its services. A natural question would be: from this virtual world, can we learn about

the real world?

The virtual web of real people inevitably reflects the real world. Many mobile devices are equipped with GPS modules allowing users to stamp geographic locations on the content that they share on social media websites. Geo-stamped and time stamped content can potentially serve as subjective or objective observations of the surrounding world — in other words, the users, with the help of their mobile devices, are sensing their environment intentionally or unintentionally. For instance, each 'tweet' on Twitter is a textual expression of the state of a person and his or her environment, while each photo serves as a visual snapshot of what the world looked like at a particular point in time and space. Besides geolocation information, other sensors on the mobile devices record other properties of the environment. For example, accelerometers measure the movements of the mobile devices, the compasses measure the directions and light sensors tell us about the intensity of light.

Users and their mobile devices act as a widely distributed web of sensors that capture the world, contributing to a collection of fine grained and large-scale observational data. In this thesis, we use the term 'user-sensed data' to refer to data about the real world captured by users or users' devices, including publicly-shared web data and data collected directly from mobile devices. Due to its collective nature, the data has the potential to aid our monitoring of the real world from new angles overcoming the long-existing geographic, temporal and labor resource restrictions. By unlocking this power, we hope to monitor and even investigate how the world (humans and various natural phenomena) changes. We look forward to making exciting innovations possible, including predicting people's behaviors, instrumenting natural phenomena and discovering knowledge.

## 1.2 RECENT INNOVATIONS AND CHALLENGES

In the last few years, researchers have begun mining these 'user-sensed' datasets to estimate and predict properties of the real world, including monitoring the spread of disease [38], predicting product adoption rates and election outcomes [53], estimating aggregate public mood [12, 75] and

inferring users' social patterns and relationships from data collected on device [31]. Some of these estimations look for specific keyword phrases related to the phenomena or sentiments of interest in the corpus and compare with ground truth or proxies of ground truth. Some use data from mobile device sensors as evidence of physical encounters of surrounding devices which reflect real world interactions.

But biases and noise in user-sensed data make these data mining tasks challenging. For web sharing data, the biases and noise lie in the user distributions and user behaviors. Social web users are not distributed evenly across the entire human population for cultural, economical and political reasons. For instance, in the US, social media users are younger than average and tend to live in large cities [30]. User behaviors create noise and biases. Besides, their photographing behaviors are biased. They do not photograph the Earth evenly, but instead take more photos in cities and tourist attractions than in rural areas. On Flickr, northeast US has a photo density almost nine times the average of North America and results from Crandall *et al* [23] show the seven most photographed landmarks are all located in big cities. Web sharing data also often has noisy or inaccurate metadata; for example, GPS units might fail to find fixes and photo timestamps can be incorrect as people forget to set the clock.

Moreover, the content of social media posts can be an imperfect and biased representation of the world. For example, photo tags can be noisy. One reason is that they are created under uncontrolled conditions by various users for various purposes including: indicating geographic locations, descriptions of actions and events, identities of objects, people, and groups, and so on [99]. Visual content can also be misleading. A specific example showing the errors caused by misleading textual and visual content would be a snow detection task using Flickr photos (detailed in Chapter 5). We try to detect the existence of natural snow in a certain time and space indicated by the photo's geo-tag and time stamp, by utilizing text-tags and visual content. However, the tag "snow" on an image might refer to a snow lily or a snowy owl, while snow appearing in an image might be artificial

(ski slope with artificial snow). In the aforementioned Flickr dataset, 1 in 700 people take a photo containing evidence of snow (tagged with snow related terms) at a non-snowy place.

## 1.3 USER-SENSED DATA IN THIS THESIS

In this thesis, the term 'user-sensed data' refers to observational data about the real world captured by the users or the users' devices, including the web data and the data collected directly from mobile devices. In this section, we describe the types and volume of the data with a focus on the data analyzed in this thesis. We also discuss the biases and sources of noise inherent in these user-sensed datasets.

### 1.3.1 TYPES OF DATA

The web data is generated either implicitly or explicitly by the users. Some users share content such as photos, microblogs (short textual content), social ties, music listening history, location check-ins and review articles publicly on various websites with social sharing elements. These websites sometimes provide access to the aforementioned data via Application Programming Interfaces (APIs) where historical and/or live streaming data can be obtained for non-commercial research purposes. For instance, in this thesis, we get Flickr data through its API. Rich metadata consists of the non-visual information such as exposure settings and timestamps recorded by the cameras as well as the information made available during the social sharing process, including text tags, comments, and ratings. The geolocation information (latitude-longitude coordinates) embedded in the photos taken by the cameras with GPS (e.g. some smartphones) or simply indicated by users when uploading helps us mapping these photos as well as the phenomena they captured.

Apart from these publicly available social sharing datasets, users are generating behavioral data when using services such as search engines and e-commerce platforms. For example, an e-commerce website (eBay) user might issue a query and click on several resulting items, then bid

on some of them or can also purchase some price-fixed items right away. In this thesis, from the recorded user behaviors, we obtain the timestamped search query logs, each of which has a text query, a user ID of the person issuing the query, the number of resulting items the user clicks on, the number of bids the resulting items receive from the user, the total number of price-fixed item the user purchases and the total prices the user pays on these items.

Mobile devices also observe their users and their surrounding environments through the interactions with them. Users' phone usage data such as app usage, web browsing history, SMS, phone calls and recharges can be captured while the various hardware sensors can also collect many sorts of data from the environment. We list a few hardware sensors that are being equipped on some latest mobile devices and enable us to understand the users as well as the surrounding environments from multiple angles: GPS, Wifi, Bluetooth, light sensor, accelerometer, touch sensor, compass, etc. In this thesis, we collect and utilize the following timestamped data from mobile devices: (1) the surrounding Wifi/Bluetooth device IDs and cell tower IDs; (2) the apps being used; (3) the geolocations of the mobile devices.

### 1.3.2 VOLUME OF DATA

Web data is usually of large scale, consisting of content contributed by millions of users through their interactions with the web service providers while the mobile data collected from the devices, on the other hand, can be rich in dimensions, as fine grained user behaviors and readings from various sensors are tracked. As a demonstration of this, we give a sense of the volume of data used in this thesis. The two Flickr datasets analyzed in this thesis contain tens of millions of geo-tagged photos (one with 150 million photos and other with 80 million), together with the associated information as aforementioned in this section. Given the fact that Flickr hosts more than 6 billion photos in 2011 [63] and the estimate that less than 8% of the photos are geo-tagged [69], the 150 million geo-tagged photos approximately take up 30% of all geo-tagged Flickr photos. For the Twitter dataset,

we get a 1% sample of the total tweets posted on the Twitter website as provided by the public API. The dataset has over 2 million tweets daily and the analysis spans over 3 months of data. On eBay, in the same 3-month timespan, there are over 250 million daily search queries together with other user behaviors related to viewing and purchasing. Our eBay dataset is a 25% sample of all these behavioral logs.

For the mobile dataset, 22 volunteers' mobile phone activities including the usage and hardware sensor readings are tracked either every 5 or 10 minutes. In spite of the limited number of users that could be recruited, various dimensions of mobile data from Wifi module readings to app usage are recorded nonstop for about 2 months. As a result, the data is far more detailed with regard to a specific user, recording the user's daily patterns as well as the environments, compared with the web data which is more related with specific usage purposes such as photographing, blogging and purchasing without various sorts of interactions.

## 1.4 SCOPE AND MOTIVATION

To confront the challenges in dealing with the user-sensed data, researchers have used a variety of ad-hoc methods based on experiences and the specific tasks. However, there lacks a general framework that describes the shared core steps and organizes the corresponding commonly used methods. In this thesis, we describe a straightforward framework with a focus on geo-temporal properties. We hope it could serve as a reference or a guide for data mining tasks on similar datasets.

Using user-sensed data and the tools and techniques from this framework, we explore the connection between user-sensed data. We ask our research question: **to what extent can we mine the virtual world to understand the real world?** As this is a relatively large topic, we approach it by breaking it down into four manageable sub-domains and addressing one representation problem in each sub-domain with novel solutions, as shown in Figure 1.1. First of all, we divide the real world into halves: one is the physical world including things such as natural phenomena and real world

entities/concepts; the other is the human world that consists of individual human users. Questions in these two sub-domains can usually be studied with two sets of methods. One set of methods measures in absolute terms with common tasks being estimating or predicting the exact values for natural phenomena or human behavior while the other set of methods measures in relative terms, studying the connections between humans or things in the physical world. These two sets of methods further divide the two sub-domains into quadrants. As a first step, we want to prove that the virtual world reflects the real world. Therefore, in the quadrant for studying the connections in the physical world, we explore the geo-temporal relationships between online photo tags, showing that these reveal connections between real world concepts. In this **tagged data exploration** project, the results are evaluated by a panel of human judges. We then focus on **estimating natural phenomena** where we can better evaluate our estimation techniques in crowd-sourced data. We accurately measure and quantify the occurrence of ecological phenomena including snowfall and vegetation cover, with the results compared against fine grained large-scale ground truth. Apart from the physical world being studied with the two projects, we are also interested in the human world as all these user-sensed datasets are generated by users and we want to mine these datasets to understand the users and eventually benefit them. Considering that mobile users contribute a lot to social webs as well as these user-sensed datasets, we choose to characterize them by studying the connections between them – e.g. users with similar behaviors may share similar hobbies or interests. Previous approaches usually require fine-grained data such as locations, app usage and communication logs. We show very simple behavioral statistics can **characterize mobile users** and even identify them. This tells us about individual users while the study on large group of users gives us a bigger picture. We study the temporal user behavior to discover signals from social media for **predicting aggregate consumer behaviors**.

We explicitly list the four corresponding questions that we address in this thesis here:

1. Can we reveal geo-temporal connections between real world concepts from these datasets?

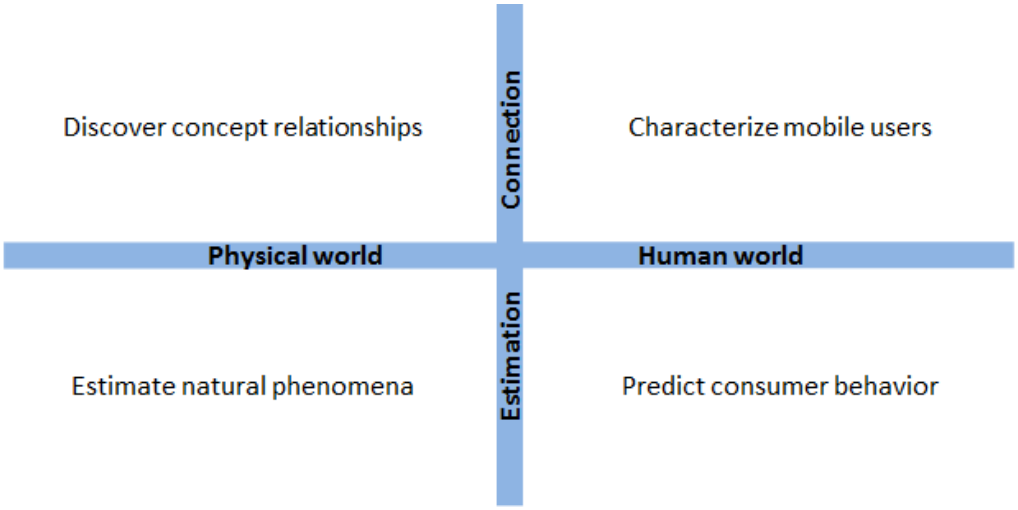



Figure 1.1: Quadrants of real world problems, each with a representative problem.

2. Can we mine these datasets to detect real world phenomena?
3. Can simple behavioral statistics of mobile users characterize and even uniquely identify them?
4. Can signals from social media predict aggregate consumer behavior?

We build four novel applications to answer these questions as well as to demonstrate and test the tools and techniques in the framework. We motivate them in the following subsections.

### 1.4.1 TAGGED DATA EXPLORATION FOR REAL WORLD KNOWLEDGE

As a first step in answering our meta research question, we are interested in seeing whether the user-sensed data can reflect the real world. We start with exploring text tagged web data. As mentioned in Sections 1.2 and 1.3, text tagged data is a common form of web data where the entries are labeled with keywords by human users to better organize and describe content. Existing approaches mainly rely on tag co-occurrences as a tag relationship measure to explore the tag semantics, understand the corpus and build tag suggestion systems. We take a different route here. Given the abundance of geo-temporal signals provided by geo-tags and timestamps, we show how to find connections between tags by comparing their distributions over time and space, discovering tags with similar geographic and temporal patterns of use. We also give a sense for why these tags are connected at

a geo-temporal usage level, instead of merely claiming that they co-occur a lot. In Chapter 4, we extract and represent geospatial, temporal and geo-temporal distributions of tag usage as vectors which can then be compared and clustered. Using a dataset of 80 million geo-tagged Flickr photos, we show that we can cluster Flickr photo tags based on their geographic and temporal usage patterns, and we evaluate the results both qualitatively and quantitatively using a panel of human judges. We visualize the temporal and geographic distributions of tag clusters. As suggested in a case study, the visualizations help humans recognize subtle semantic relationships in the real world. All these suggest that the discovered tag connections reveal the geo-temporal connections between concepts in the real world. Besides the immediate application in the tagged image datasets, this approach to finding and visualizing similar tags is potentially useful for exploring any geographic and temporal data.

As discussed in Sections 1.2 and 1.3, the tags, provide a rich (albeit noisy, incomplete and inconsistent) source of information about the semantic content of photos that has been a popular corpus for the data mining community. Through the analysis on tags, the properties of online photo collections have been studied, including identifying temporal bursts of photographic activity corresponding to important events [91], finding geospatial peaks of activity corresponding to important landmarks [79], selecting iconic images to represent particular places [23], and even predicting product adoption rates by monitoring the popularity of product photos [53].

Much of this existing work has been based on *tag co-occurrences*, which assumes tags that frequently appear with one another on the same photos are related semantically. Suggestion systems have been built upon tag co-occurrence which prompt new tags based on the existing tags of a photo [37, 70, 99]. Clustering techniques have also been applied on tags in order to discover semantically-related concepts represented by the groups of tags that frequently co-occur [7, 97]. While these approaches often yield reasonable results, they assume that tags are related only if they often appear on same photos. Some other factors that relate the tags have been overlooked. For

example, the Statue of Liberty and Central Park are clearly related as both are major New York City landmarks. However, the tags *statueofliberty* and *centralpark* do not co-occur much because it is nearly impossible to take a photo that includes both. Similarly, though both Mexico and Puerto Rico are popular winter vacation destinations, the tags *mexico* and *puertorico* do not have many co-occurrences. Moreover, co-occurrences give little information about the nature of the relationship – i.e. *why* two tags are related.

We explore the more specific connections between tags as well as the connections between the concepts that the tags represent, by comparing the spatial and temporal usage distributions of tags. The intuition is that some related concepts that do not often co-occur on same photos, do have similar geo-temporal distributions because they co-occur a lot in time and space which is not necessarily reflected by the textual co-occurrences on photos. The rich metadata that includes the generated time and geolocations of the contents make our analysis possible. geospatial and temporal properties of tags have been studied in existing work (e.g. [53, 79, 91]), but we are not aware of work that has used these properties to quantify connections between tags.

We detail our aforementioned approach with visualization and evaluation in Chapter 4.

### 1.4.2 REAL WORLD PHENOMENON DETECTION

In the last subsection, we motivated geo-temporal knowledge exploration and extraction where the ground truth comes from human judgments. The ground truth obtained in this way is usually subjective at relatively smaller scales. Thus, it is not suitable for evaluating estimates of large-scale real world phenomena. Addressing this, we choose to estimate natural phenomena for which precise and large-scale ground truth data is available such that estimation techniques can be better evaluated. With this effort, we hope to assist biology and ecology research in need of observational data. Besides this direct impact, the techniques can be potentially generalized for producing estimates in other domains such as product distributions and political views.

Recent work has begun to make sense of passively-collected data from social networking and microblogging websites to make estimates and predictions about world events, including tracking the spread of disease [38], monitoring for fires and emergencies [24], predicting product adoption rates, election outcomes [53] and news events [89, 90], and estimating aggregate public mood [12, 83]. However, these previous studies either lack the ground truth available to judge the quality of the estimates and predictions or have to use indirect proxies as ground truth (e.g. since no aggregate public mood data exists, [83] evaluates against opinion polls, while [12] compares to stock market indices). While these studies have generated promising results, it is not clear when crowd-sourcing data from social media sites can yield reliable estimates, or how to deal with the substantial noise and bias in these datasets. Moreover, these studies mostly relied on textual feature without taking advantage of the vast amount of visual content online.

We study the particular problem of estimating geo-temporal distributions of ecological phenomena using geo-tagged, time-stamped Flickr photos. The motivations are three-fold. First, biological and ecological phenomena frequently appear in images, both because photographers take photos of them purposely (e.g. close-ups of plants and animals) or incidentally (a bird in the background of a family portrait, or the snow in the action shot of children sledding). Second, for the two phenomena we study here, snowfall and vegetation cover, large-scale and fine grained (albeit imperfect) ground truth is available publicly from satellites and ground-based weather stations. Thus we can explicitly evaluate the accuracy of various techniques for extracting semantic information from large-scale social media collections as a concrete test and application of the proposed framework.

Third, while ground truth is available for these particular phenomena, for other important ecological phenomena (like the geo-temporal distribution of plants and animals) no such data is available, and social media could help cater this need. Specifically, the scientists who study climate change are in great need of real-time, global-scale monitoring data. Recent work shows that global climate change is impacting a variety of flora and fauna at local, regional and continental scales:

for example, species of high-elevation and cold-weather mammals have moved northward, some species of butterflies have become extinct, waterfowl are losing coastal wetland habitats as oceans rise, and certain fish populations are rapidly declining [86]. However, to monitor these changes is not an easy task: plot-based studies involving direct observation of small patches of land yield high-quality data but are costly and possible only at very small scales, while aerial surveillance gives data over large land areas but cloud cover, forests, atmospheric conditions and mountain shadows can interfere with the observations, and only certain types of ecological information can be collected from the air. To understand how biological phenomena are responding to both landscape changes and global climate change, an efficient system for ground-based fine grained data collection at a global scale is desired. A tantalizing alternative possibility for creating such ground-level, continental-scale datasets is to use passive data-mining of the huge number of visual observations produced by millions of users worldwide, in the form of digital images uploaded to photo-sharing websites.

There are two challenges if we want to distill credible ecological estimates from these social sharing datasets. The first is how to discover the ecological phenomena of interest appearing in photos and how to map these observations to specific places and times. Fortunately, as discussed in Section 1.3.1, online photos include the ingredients necessary to produce geo-temporal data about the world, including information about content (images, tags and comments), and when (timestamp) and where (geo-tag) each photo was taken.

The second challenge is how to deal with the biases and noise inherent in online data. As mentioned earlier in this chapter, user distributions and behaviors are biased and in the particular case of Flickr dataset, GPS modules sometimes malfunction and the tags can be wrong or semantically misleading.

We study how to mine data from photo-sharing websites to produce credible crowd-sourced observations of ecological phenomena. We start with studying two types of phenomena – ground snow cover and vegetation cover ("green-up") data, as a first step towards the long-term goal of mining

for many other types of phenomena. On one hand, they are both critical features for ecologists monitoring the earth's ecosystems; on the other hand, these two phenomena have accurate fine-grained ground truth available at a continental scale in the form of aerial instruments like NASA's Terra earth-observing satellites [46, 66] and networks of ground-based observing stations like the U.S. NOAA National Weather Service. With this data, we are able to evaluate the performance of our data mining techniques against large-scale fine grained ground truth data across an entire continent spanning over thousands of days. Using a dataset of nearly 150 million geo-tagged Flickr photos, we study whether this data can potentially be a reliable resource for scientific research. An example comparing ground truth snow cover data with the estimates produced by our Flickr analysis on one particular day (December 21, 2009) is shown in Figure 1.2. Note that the Flickr analysis is sparse in places with few photographs, while the satellite data is missing in areas with cloud cover, but they agree well in areas where both observations are present. This and the more extensive experimental results from this project suggests that Flickr analysis may produce useful observations either on its own or as a complement other observational sources.

The main contributions of this project include: (1) introducing the novel idea of mining photo-sharing sites for geo-temporal information about ecological phenomena, (2) applying several techniques in the proposed framework for deriving crowd-sourced observations about real world phenomena from noisy, biased data using both visual and textual tag analysis, and (3) evaluating the ability of these techniques to accurately measure these phenomena, using dense large-scale ground truth. Details of this project are elaborated in Chapter 5.

### 1.4.3 CHARACTERIZING MOBILE USERS

After motivating studying the physical world in Sections 1.4.1 and 1.4.2, we move onto the study of the human world. These user-sensed datasets are generated by the users and it naturally makes sense to mine them in order to understand the users and build services for the users. Among these

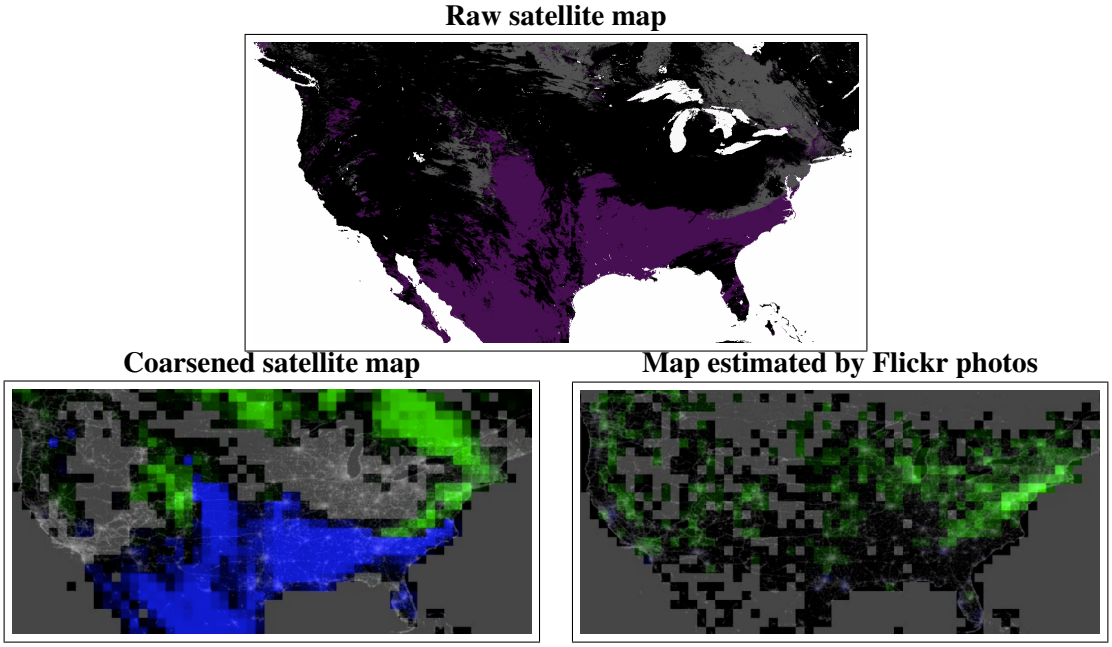



Figure 1.2: Comparing MODIS satellite snow coverage data for North America on Dec 21, 2009 with estimates produced by analyzing Flickr tags (best viewed in color). *Top:* Original MODIS snow data, where white corresponds with water, black is missing data because of cloud cover, grey indicates snow cover, and purple indicates no significant snow cover. *Left:* Satellite data coarsened into 1 degree bins, where green indicates snow cover, blue indicates no snow, and grey indicates missing data. *Right:* Estimates produced by the Flickr photo analysis proposed in this project, where green indicates high probability of snow cover, and grey and black indicate low-confidence areas (with few photos or ambiguous evidence).

users, mobile users take up a large portion. Researchers study the connections between mobile users to characterize them. For example, users with similar usage patterns may have similar hobbies or interests.

As mentioned in Section 1.1, modern smartphones and tablets are equipped with various sensors. Apart from that, the devices have also advanced significantly in terms of computational capacity, memory and storage. All these allow fine grained rich mobile device usage data and information about the surrounding environment to be collected through the interaction with the users as well as the environment from multiple perspectives. This has stimulated people-centric mobile applications, ranging from inferring and sharing real-time contexts such as location and activities [15, 65] to identifying heterogeneous social behaviors of mobile users [25, 68, 87, 120]. These studies focus

on using phone data to infer physical and social contexts for a particular user at a specific point in time while additional studies have concentrated on analyzing long-term data from mobile devices to monitor trends and to establish predictive models of location [28] and app usage [98]. However, all these previous studies tend to rely on very fine-grained information. This leads us to think whether this information can be reduced to still be able to achieve the same goals. We show that mobile users can be characterized with far less information. We use simple behavioral statistics to characterize mobile users. We analyze multimodal mobile usage data to extract simple statistics that can uniquely represent mobile users effectively. Our experiments show its descriptive power in characterizing users and identifying them uniquely. As future work, recommender systems can be built by associating such simple features with user attributes such as social and life patterns.

It is non-trivial to model users based on high level statistics due to the following challenges: (1) Accurately characterizing users with simple and indirect statistics is difficult. What distinguishes the users is usually obvious and carries fine-grained information. (2) Mobile data is generated under complex real-life settings, introducing significant noise that demands robustness in processing. (3) Different from most existing offline data analysis approaches deployed on the server side, energy-efficient design considerations are required for on-device applications.

To address these challenges, in Chapter 6, we present a novel approach with the following contributions: (1) By designing a discriminative set of high-level statistical features, we are able to identify users effectively, demonstrating the descriptive power of these features; (2) The algorithms are designed to be lightweight and energy-efficient for on-device processing.

### 1.4.4 TEMPORAL USER BEHAVIOR STUDY

After studying individual users, we are curious about how these datasets reflect aggregate behavior for large group of users. Specifically, we study the interplay between users' behaviors on social media and e-commerce platforms, trying to discover signals from social media for predicting large-

scale consumer behaviors as one of the first efforts to bridge the two domains. For example, if someone, something or some event suddenly gains popularity on social media, are people also interested in purchasing relevant merchandise on e-commerce platforms? If there is such a correlation, is there also a lag indicating that social media is faster than e-commerce? If there is such a lag, can we detect signals from social media to predict e-commerce user behaviors? One of the potential applications could be an instant recommendation system that monitors social media and tells the e-commerce sellers and buyers what is going to be popular such that sellers can list the relevant items well ahead of time and buyers can be more attracted and engaged. With such a system, the e-commerce platform can benefit from the increased transactions.

The explosion of social media has enabled the spread of public opinion, intentions, and observations, sometimes even at a faster pace than traditional news media [64, 84, 92]. Compared with social media where news, thoughts and observations are shared and spread, users on e-commerce websites are more likely to take real actions (eg. selling and buying). Apart from the various study on mining social media to generate real world insights in previous subsections, there is extensive work in understanding user sessions, behaviors, and activities on search and e-commerce sites which provide a closer delegate of users' real world actions. Such analysis sometimes aims at unlocking commercial value through understanding user intentions. Recent work on this includes quantifying the influence of users' domain knowledge on their search behaviors [114], comparing users' search behaviors across different devices [56], understanding user behaviors when their queries return no result [101] and rewriting these queries to increase recall [100]. User's bidding behaviors have also been studied [4, 49, 82, 96]. However, these standalone studies have not involved social components which might bring new possibilities in understanding the users' intentions.

We believe that understanding the behavioral characteristics of both social media and e-commerce domains and their impact on each other would create opportunities of both social and economic value. A step taken in this direction studies the price of trust by analyzing the internal social net-

work of a Chinese e-commerce website, Taobao, to estimate how much a buyer is willing to pay for a transaction with a trusted seller [44]. The interplay between Facebook mouth-to-mouth recommendations and a special form of e-commerce, daily deal (Groupon), have been studied [14] where the researchers find a correlation between Facebook likes and the size of the corresponding deal. Though these two papers reveal interesting aspects of the interaction between social network and e-commerce, the correlations between the temporal trends from both domains which reflect the interest and demand of the general public still remain unexplored.

As one of the first efforts, we study the interplay of user behavior on a global social media site (Twitter) and a global e-commerce site (eBay) at a large scale and at a fine grain by analyzing the timestamped tweets, and e-commerce search logs and transaction data. To demonstrate the motivation, we make up a hypothesized and simplified scenario involving the two platforms: on Twitter, you see people talking about a recent victory of your favorite soccer club, Chelsea, after which you go to eBay, search for and eventually purchase a Chelsea T-shirt. We answer questions including: What is the correlation between the sudden surge in popularity on social media and e-commerce behaviors? Does one platform lead or trail in the burst in activity when such events occur? Our findings suggest that about 5% of general eBay query streams have strong positive correlations with the corresponding Twitter mention streams and for trending queries [85], the percentage goes up to around 25%. Queries from certain eBay categories such as 'Sports' show more obvious correlations. We also discover evidence that eBay lags Twitter for correlated pairs of streams. By monitoring the popularities of a list of celebrities, we find that their Twitter popularities correlate with related eBay search and sales. These correlations and lags can be useful for predictive tasks. For example, a system monitoring social media can potentially make instant decisions whether a burst is going to lead a sales drive on e-commerce platforms. As a result, relevant sellers and potential buyers will be notified in advance such that the transactions can be stimulated. As a step toward this, we adapt a burst detection algorithm to better monitor the two streams.

In Chapter 7, we further demonstrate the methods from the framework to quantify the correlation as well as the lags between the two streams and apply them to the time series extracted from hundreds of millions of eBay search queries and tweets as a test on real world data. We then describe an adapted peak detection algorithm in preparation for a real-time monitoring and recommendation system. We also conduct a case study to demonstrate the characteristics of the two trends that we observe.

## 1.5 CONTRIBUTIONS

We study the observational datasets generated by the users or their smart devices to distill credible insights about the real world. Some of our contributions include: (1) describing a framework that assembles the techniques for dealing with these large 'user-sensed' datasets while addressing the inherent biases and noise as well as the geo-temporal characteristics; (2) building applications that on one hand showcase the techniques in the framework and on the other hand answer the meta research question which is: "to what extent can we mine the virtual world to understand the real world?", by addressing representative problems in each quadrant of all possible problems, with regard to collective knowledge, natural phenomena, individual user behavior and large-scale user behavior respectively.

To be specific about the four applications: (1) we investigate and visualize the geo-temporal relationships between photo text tags, showing that these reveal connections between real world concepts; (2) we analyze the geo-tags, timestamps, text tags, and visual content of 150 million online photos to measure and quantify the occurrence of ecological phenomena, comparing against ground truth at a continental scale; (3) we study the relationship between social media and e-commerce websites, showing that signals from social media can predict aggregate consumer behavior; (4) we study how simple behavioral statistics of mobile users over time can characterize and even uniquely identify them.

We hope the novelty as well as the framework's techniques contextualized in the applications can inspire and guide future data mining tasks on such datasets to reveal real world value in various domains.

## 1.6 OUTLINE OF THE THESIS

We organize the remainder of this thesis as follows. In Chapter 2, we briefly survey the related work and organize them with regard to our four applications. In Chapter 3, we describe the lightweight framework for dealing with geo-temporal user-sensed datasets. In Chapters 4, 5, 6 and 7, we test and further demonstrate the techniques in the framework with four data mining applications while addressing research questions with regard to exploring and extracting geo-temporal knowledge, estimating natural phenomena, characterizing mobile users and predicting aggregate consumer behaviors. We conclude this thesis in Chapter 8.

**Our published papers and their corresponding chapters:**

- Chapter 4
  - Haipeng Zhang, Mohammed Korayem, Erkang You and David J. Crandall, Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from geospatial and Temporal Similarities, in *Web Search and Data Mining (WSDM) 2012* [123].
- Chapter 5
  - Haipeng Zhang, Mohammed Korayem, David J. Crandall and Gretchen Lebuhn, Mining Photo-sharing Websites to Study Ecological Phenomena, in *World Wide Web (WWW) 2012* [122].
- Chapter 6
  - Haipeng Zhang, Zhixian Yan, Jun Yang, Emmanuel Munguia Tapia and David J. Crandall, mFingerprint: Privacy-Preserving User Modeling with Multimodal Mobile Device Footprints, in *Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP) 2014* [125].

- Chapter 7
  - — Haipeng Zhang, Nish Parikh, Gyanit Singh and Neel Sundaresan, Chelsea Won, and You Bought a T-shirt: Characterizing the Interplay Between Twitter and E-Commerce, in *Advances in Social Networks Analysis and Mining (ASONAM) 2013* [124].

# CHAPTER 2

# RELATED WORK

First, at a higher level, we briefly survey some representative guiding work that studies the dynamics between the web and mobile data and the real world (Section 2.1). We then organize the specific related work according to the different focuses of the four proposed applications in four areas: *tagged data exploring and organizing* (Section 2.2), *event detection and prediction* (Section 2.3), *cross-platform user behaviors* (Section 2.4), and *mobile user inference* (Section 2.5).

## 2.1 GUIDING WORK

There have been various studies that guide and inspire our research. Berners-Lee *et al* [9] foreseeingly address that Web science is not only about modeling the web itself. They encourage research on understanding the society that uses the Web, reusing information in innovative ways, and creating beneficial new systems. In fields including economics and psychology, Surowiecki in his book *'Wisdom of the Crowd'* [106] argues that aggregate decisions from large groups of people (collective intelligence) are often more intelligent than the ones from individual members. Can Web users and mobile users contribute to intelligent decisions or insights about the real world, if we aggregate the vast amount of information from them wisely?

Google Flu Trends[1] by Ginsberg *et al* [38] demonstrates that the 'collective intelligence' can be used to estimate real world events and benefit humankind. It aggregates geo-temporal distributions of Google search queries to provide up-to-date national and regional estimates of influenza-like illness activity ahead of official data that lags the estimates by 1-2 weeks. They notice that the usage frequency of certain search queries strongly correlates with the presence of flu-like illness indicated by conventional data from health agencies. They discover and aggregate the top correlated queries to fit a linear regression model to accurately estimate the flu activity. However, critics have pointed out recently that the service overestimated peak flu levels from August 2011 to September 2013 [13,67]. They suggest that the possible overfitting problem results from selecting best matches among 50 million queries to fit 1152 data points and that algorithm changes behind the search engine itself might also contribute to the mistakes. To avoid this, it requires the mining approach to be more robust and adaptive. Suggestions from the critics include incorporating other near-real-time data sources and calibrating the service frequently [67].

There is work on understanding fundamental human social patterns and structures, by analyzing the web data and mobile phone data. Crandall *et al* [22] mine a large social photo sharing data set to quantify the probability that two people have a social tie given their co-occurrences in time and space. They show that a very small number of co-occurrences can result in a high probability of a social tie, with a probabilistic model accounting for this effect. Eagle *et al* [32] mine the everyday mobile phone communication, location and proximity data collected from 94 people to infer the friendship structure comparing with self-reported data, with high accuracy. As shown in these two studies, large scale web data can potentially reveal macroscopic phenomena while mobile phone data usually comes in very high resolution with various dimensions that can depict user's daily life patterns at a finer grain.

Researchers also study social media websites as channels where large amounts of information is instantly shared and spread across the globe. Kwak *et al* [64] study the topology of Twitter

[1] http://www.google.org/flutrends/

and its information diffusion, suggesting that Twitter is a medium for breaking news as if its users are monitoring the world. Novel applications have been developed along this direction. Sakaki *et al* [92] propose that tweets with geo-temporal attributes can been seen as sensory information, from which real world events such as earthquakes and typhoons can be detected. These lead us to think whether real world events can be further quantified at large scales and at finer grains by leveraging other features and ground truth data with high resolution and coverage.

Along the temporal dimension, social user behaviors are studied to understand how their characteristics change over time. Mislove *et al* [75] estimate the temporal mood variations across the U.S. by examining the tweets. Further efforts are made to associate them with real world human behaviors. Golder and Macy [40] suggest this temporal mood varies with work, sleep, and daylength patterns while Bollen *et al* [12] perform causality analysis on Twitter mood trends and stock market indices which result from societies' collective decision making. These all inspire our intuition of examining social media user behaviors and the e-commerce consumer behaviors from the temporal perspective.

As these annotated (tagged) datasets grow to large scales, it becomes important to organize the pieces. During this process, patterns and knowledge are usually revealed. Crandall *et al* [23] and Kennedy *et al* [61] discover popular landmarks and their representative images by incorporating features extracted from timestamps, geotags, textual tags and images to organize online photo sharing datasets. Sigurbjörnsson and Van Zwol [99] create a tag suggestion system based on tag co-occurrence to help users better annotate and organize the content. We continue the efforts in organizing and exploring the tagged data by applying novel tag similarity measurements from geo-temporal perspectives other than tag co-occurrence.

## 2.2 TAGGED DATA EXPLORING AND ORGANIZING

Tags on social sharing systems have been studied extensively. Here we review the work most relevant to our project described in Chapter 4 that studies tag semantics and relationships in photo collections.

***Clustering tags based on co-occurrences.*** Much work has been based on photo tag co-occurrences, mostly in the context of tag suggestion systems. Garg and Weber [37] use tag co-occurrences to suggest additional tags for a new image. Sigurbjörnsson and Van Zwol [99], as discussed in Section 2.1, take a similar approach but also conduct a study of tagging behavior on Flickr. Liu *et al* [70] rank tags by performing Pagerank-like random walks on tag graphs where the edge weights are frequency of co-occurrence. Shepitsen *et al* [97] use TF-IDF to build trees of del.icio.us and Last.Fm tags, and then partition them into homogeneous clusters. Begelman *et al* [7] partition tag graphs using spectral clustering. Markines *et al* [74] measure the semantic relationships among users, resources and tags based on co-occurrence. These papers inspired our idea of clustering tags, but our work differs in that we use features other than co-occurrence, instead looking for tags with second-order connections like similarities in spatial and/or temporal distributions.

***Temporal and geospatial properties of tags.*** Timestamps and geo-tags of photos have been used to study temporal and geospatial distributions of individual tags. Jin *et al* [53] measure the usage of hand-selected Flickr tags over time to make predictions such as the future sales of products and the outcome of elections. Serdyukov *et al* [95] predict where a photo was taken given its tags. Rattenbury *et al* [91] use burst detection techniques to find tags with significant peaks in time and space, while Moxley *et al* [79] build on this work by using entropy analysis on a quadtree data structure to improve performance. Chen *el al* [17] model tag occurrences as points in a 3D geo-temporal space, and then use wavelet transform-based techniques to find tags with bursts in both temporal and spatial distributions, and then cluster these tags using DBSCAN. Ahern *et al* [3]

cluster photos to find dense areas based on geo-tags and find representative tags for the areas using TF-IDF, while Moxley *et al* [78] use a similar approach to rank local tags. As mentioned in Section 2.1, Crandall *et al* [23] and Kennedy *et al* [61] find both distinctive tags and images for clusters of photos found based on geo-tags. While these papers analyze geospatial and temporal attributes of photos to study individual tags or similarity between "event" tags with significance in geospatial and temporal distributions which consist of a small portion of all the tags, we instead use these features to compare general tags with one another.

***Studies of query logs, tweets and news articles.*** Temporal and geospatial patterns have been studied to discover concept relationships in other domains, including tweets, news articles, and queries in search engine logs. Vlachos *et al* [112] use frequency-space analysis of search query time series to identify bursts and semantically-similar queries. Chien and Immorlica [18] use similar techniques but perform an experimental evaluation, and find that for 70% of the queries, at least three of the top ten keywords identified by temporal similarity are semantically related. The geospatial distributions of search engine queries were studied by Backstrom *et al* [5], who estimate geographic centers and dispersions of queries. Vadrevu *et al* [110] use the co-occurrence of a query term with place names in a region to determine whether the query is related to this region. Very recently, Mohebbi *et al* [76] developed a tool which takes a web query from the user and finds the queries with correlated temporal or spatial distributions. They quantize the weekly time series usage data and the state-by-state series usage data of individual queries into vectors to represent temporal and spatial distributions, and then apply K-means clustering and an approximate nearest neighbor algorithm to look up similar vectors efficiently. Yang and Leskovec [121] cluster Twitter hashtags and short phrases in news documents to identify their temporal patterns and the dynamics of human attention they receive. Radinsky *et al* [88] extend Explicit Semantic Analysis to represent concepts as time series of word occurrences in the New York Times archive. While search queries, tweets and news articles have different properties than Flickr tags and are thus not directly relevant to our work, we

use these papers as inspiration, and in particular borrow the idea of using clustering to reveal the shared topics or semantics of groups of tags.

***Spatial clustering and co-location pattern mining.*** Spatial clustering, as discussed by Ng and Han [81] and Sander *et al* [93] groups together similar spatial data points based on their locations. It differs from our work as it clusters spatial data points while ours clusters high dimensional feature vectors extracted from the geographical distributions of the text tags. Spatial co-location pattern mining [33, 34, 50, 117] shares a more similar task to ours which is to "find spatial features frequently located together in spatial proximity" [50], where the spatial features in our case are text tags. Clustering is not used by Xiao *et al* [117], while Huang and Zhang [50] and Estivill-Castro *et al* [33,34] take different clustering approaches. Our work differs in clustering different geographical feature vectors extracted from vast user-generated content, showing that with the help of a straightforward clustering method, high quality semantics can be effectively generated from the wisdom of the crowd.

***Visualizing tag clusters.*** The usual method to visualize tags is to draw tag clouds [59], while Dubinko *et al* [29] visualize interesting Flickr tags that evolve over time through animations. In our project, we introduce visualizations of the temporal semantics of tag clusters by plotting time series representing their usage along time, and the geospatial semantics of tag clusters in a 3-D space over a map to represent their usage across space. We also show that the visualizations can help humans understand subtle semantic relationships.

***Connections between tags.*** Perhaps the closest related work to ours in spirit is that of Wu *et al* [116], who compute the "Flickr distance" between a pair of tags by computing the visual similarity of photos having those tags, and then cluster tags based on this score. Our work is similar in that it defines connections between tags using a property other than co-occurrence, but is complementary in that we define similarity metrics based on metadata like timestamps and geo-tags instead of the visual content of images. This both allows us to discover connections that may not be apparent

from visual features, and also allows our techniques to scale to much larger datasets (having tens of millions of photographs) because processing metadata is much more efficient than visual analysis.

## 2.3 EVENT DETECTION AND PREDICTION

A variety of recent work relevant to our project described in Chapter 5 has studied how to apply computational techniques to analyze online social datasets in order to monitor and predict the real world events.

***Crowd-sourced observational data.*** Some studies have shown the power of social networking sites as a source of observational data about the world itself. Like us, Jin *et al* [53] use Flickr as a source of data for prediction, but they estimate the adoption rate of consumer photos by monitoring the frequency of tag use over time. They find that the volume of Flickr tags is correlates with with sales of two products, Macs and iPods. They also estimate geo-temporal distributions of these sales over time but do not compare to ground truth, so it is unclear how accurate these estimates are. Xu *et al* [118] estimate time, locations and counts of roadkills using tweets. However, the lack of quantitative ground truth also makes it hard to evaluate the estimation performance. In contrast, we evaluate our techniques against a large ground truth dataset, where the task is to accurately predict the distribution of a phenomenon (e.g. snow) across an entire continent each day for several years.

***Crowd-sourced geo-temporal data.*** Other work has used online data to predict geo-temporal distributions, but again in domains other than ecology. Perhaps the most striking is the aforementioned work of Ginsberg *et al* [38], who estimate the spread of flu-like illness. DeLongueville *et al* [24] study tweets related to a major fire in France, but their analysis is at a very small scale (a few dozen tweets) and their focus is more on human reactions to the fire as opposed to using these tweets to estimate the fire's position and severity. In perhaps the most related existing work to ours, Singh *et al* [102] create geospatial heat maps (dubbed "social pixels") of various tags, including snow and

greenery, but their focus is on developing a formal database-style algebra for describing queries on these systems and for creating visualizations. They do not consider how to produce accurate predictions from these visualizations, nor do they compare to any ground truth.

## 2.4 CROSS-PLATFORM USER BEHAVIORS

Recent research has studied user behaviors on e-commerce websites and social media websites, which is relevant to our project described in Chapter 7. We introduce the work on e-commerce user behaviors as well as the work on the connection between the two kinds of platforms.

***User behaviors on e-commerce websites.*** User behaviors on e-commerce websites alone have been analyzed in order to build better services for users and increase the profits for the websites. Much of the work has been based on understanding the search query logs. Parikh and Sundaresan [85] develop a near real-time burst detection system to suggest trending queries for the buyers and sellers. Singh *et al* [100,101] study user behaviors in a situation where their initial queries return no matches and build a system to rewrite these queries to increase recall. User behaviors with regard to online auctions have been another theme of research, including cross-bidding [4], last-minute bidding [82], shilling [96] and effect of seller reputation on price [49]. Instead of only focusing on the e-commerce domain itself, we also examine the social media domain to quantify the correlations and explore what drives the sales on e-commerce websites.

***Connection between social media and e-commerce.*** The studies at the intersection of social media and e-commerce bring interesting findings. Guo *et al* [44] analyze the internal social network of buyers and sellers on a Chinese e-commerce website, to model how much a buyer will pay for a transaction with a trusted seller. The focus is on the structure of the internal social network and the price of trust, rather than analyzing trends and user behaviors on an open social media platform and a global e-commerce website as we do in this paper. Byers *et al* [14] study the daily deal websites

including Groupon and their ties with social media websites such as Facebook and Yelp. On the non-social side of their research, they analyze the incentives for users to purchase and use the parameters of the deal to predict the sales. On the social side, they examine how daily deals affects merchants' reputation on Yelp and they suggest that word-of-mouth recommendations on Facebook benefit daily deal sites. The deals that they study are a special form of e-commerce offering the consumers with localized discounted merchandise in a relatively short period of time while we conduct a study of more general user behaviors in both domains at a large scale and in a long run. Moreover, instead of studying information diffusion in social media, we focus on analyzing the temporal attributes of the information, which reflect users' instant interest and demand.

## 2.5 MOBILE USER INFERENCE

There has been work on inferring mobile users' context and characteristics using data collected from the phone, which is relevant to our project described in Chapter 6.

***Inferring user's physical context.*** There are studies on inferring users' physical context and characteristics, such as identifying physical activities from accelerometer data [65], estimating user's environmental contexts such as crowdedness from Bluetooth [113] and noise-level using the microphone [71]. These studies however focus on analyzing only unimodal physical sensors to infer specific types of daily contexts and activities. We analyze multimodal heterogeneous sensor data including WiFi, GPS, cell tower, and Bluetooth collected in non-laboratory settings.

***Inferring user's behaviors and social context.*** On the social and behavioral analysis side of mobile data mining, a recent focus is on continuously monitoring user's daily social contexts, such as characterizing individual human mobility patterns with cell tower registering information [41], inferring emotions from audio [87], detecting mood from communication history and application usage patterns [68], predicting user's personality using various phone usage data [19, 25] and esti-

mating users' profile and sociability level using Bluetooth crowdedness [120].

***Privacy implications.*** The inferred context and characteristics of mobile users, as well as the information required for the inference, could raise privacy concerns. Communication logs and web browsing histories are intuitively personal and sensitive [45,57] while there is other information that users may not want to share even though the purpose is as simple as discovering the users' interests for recommendation systems. For example, Barkhuus and Dey [6] demonstrate that though people are not overly concerned on average about location-based services, a moderate portion of the participants (5 out of 16) are still 'concerned' or 'highly concerned'. Some concerns are reasonable: according to Yan *et al* [119], semantic locations such as home and office can be detected given users' GPS locations. Geolocations could further reveal mobile users' identities that are sensitive in nature [10]. We hope our work on using simple behavioral statistics to identify mobile users can draw some interest from the privacy research community to discuss the privacy implications here.

## 2.6 SUMMARY

In this chapter, we first briefly discussed the work on utilizing the Web and mobile data to study the real world that guides and inspires our work on a higher level. We then looked into the specific related work with regard to our four data mining applications. Section 2.2 surveyed the line of work in *tagged data exploring and organizing* that is relevant to our project that investigates and visualizes the geo-temporal relationships between photo tags to be described in Chapter 4. In Section 2.3, we reviewed literature on *event detection and prediction* and we describe our project that measures and quantifies the occurrence of ecological phenomena by analyzing online photos in Chapter 5. In Section 2.4, we discussed the previous work on *cross-platform user behaviors* that is relevant to our project on discovering signals from social media websites for predicting aggregate consumer behavior on e-commerce platforms as to be described in Chapter 7. In Section 2.5, we reviewed related work on *mobile user inference* and correspondingly, we describe our work on characterizing

and identifying mobile users using simple behavioral statistics in Chapter 6.

# CHAPTER 3

# A FRAMEWORK FOR ANALYZING USER-SENSED DATA

Various ad-hoc approaches have been applied in different data mining tasks on geo-temporal user-sensed datasets. We pick the ones from existing literature that are of general significance and organize them into a lightweight framework with application scenarios. This framework assembles essential components that are shared across many such applications. These components include: geo-temporal pattern extraction, comparison, clustering, visualization and external phenomenon detection. They will be contextualized in following chapters when referred to. This chapter also extends and complements the previous chapter where we review literature relevant to the application areas while in this chapter we are presenting relevant work to the technical approaches in the framework. We hope this framework, as a possible guide or reference, can facilitate fast prototyping of data mining projects on user-sensed data with geo-temporal attributes to yield accurate insights related to the real world.

## 3.1 GEO-TEMPORAL PATTERN EXTRACTION

Many datasets including crowd-sourced user generated content and user behavioral data have time stamps and geo-tags. Extracting geo-temporal patterns with simplicity and efficiency is usually an initial step for many research projects on such datasets. We will describe binning approaches that fit into the three research scenarios in Section 4, 5 and 7 and the results prove the feasibility and effectiveness of the approaches in return.

### 3.1.1 QUANTIZATION OF GEO-TAGS

***Binning.*** We are interested in geographical distributions of geo-referenced user sensed data as the user activities and observations can be mapped to locations. Because the distribution of social media users/mobile users over the world is highly non-uniform, with most of their activities concentrated in cities, many areas of the world have very few activities. Thus, one of the efficient and straightforward ways of extracting the distributions is to aggregate the geo-referenced entries (photos, tweets, etc) together into coarse geospatial buckets instead of clustering the entries directly. (The quantization process also helps reduce the impact of noise in the geo-tags.) To do this, we divide the world into $n$ bins, each with $s$ degrees of latitude by $s$ degrees of longitude. We assign each of the bins a unique index number in the range $[1, n]$, and define a quantization function $q_G(g)$ that maps a latitude-longitude coordinate $g$ into the index number of the corresponding geo-bin. Each bin can then record the occurrences of phenomenon of interest after which a vector is produced. These phenomena might have different overall frequencies of occurrence. For tasks where we are interested in the distributions instead of the overall frequencies of occurrence, normalization is necessary. For example, we can divide the value in each bin by the sum across all bins (L1 norm) such that each bin records the percentage of occurrence it takes up among all bins. Figure 3.1 shows the heatmap of the geospatial distributions of Flickr tags “corn” and “coconut” over North America from a dataset of 44 million geo-tagged Flickr photos. Greater intensity in the geo-bin color indicates more users

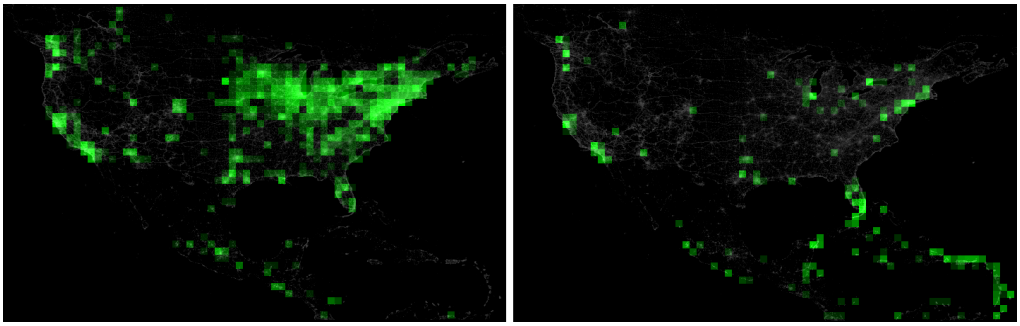


Figure 3.1: Geographic distributions for tag “corn” (left) and “coconut” (right).

applied the tag to photos taken in the very bin. Note that the bins do not have the same surface area because degrees of longitude become closer together near the poles; this tends not to be a problem in practice because the vast majority of user sensed datasets are closely associated with human activities near the middle latitudes. An equal-area partitioning of the globe [42] could address this issue.

***Hierarchical binning.*** Predefined geo-bin size affects the geospatial distributions being captured, e.g., New York City and Los Angeles might both fall into the North America bin at a continental bin size granularity while they would be different bins at a state granularity. In order to achieve hierarchical and efficient indexing of the bins, the quadtree can potentially be applied. It is a tree data structure where each internal node has four child nodes. It is used to partition a 2D space by recursively dividing it into four regions where the data points can be geo-tags. If a region is above its max capacity, it splits. In a special case, each split generates four equal-sized subregions. Figure 3.2 shows a quadtree with a max capacity of 2 and equal splits. A denser region has nodes of higher levels. Moxley *et al* [79] use quadtree to find Flickr places tags which have a tight distribution on a location by computing a sum of entropy over multiple levels. We describe this method here as we believe it can be a possible extension to the regular binning method described above and used in Chapter 4.

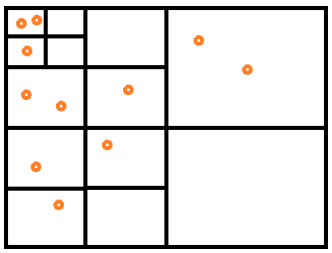


Figure 3.2: A quadtree with a max capacity of 2.

### 3.1.2 QUANTIZATION OF TIMESTAMPS

As with the quantization of geo-tags, we can aggregate the timestamps into coarse temporal bins. Let $q_T(\tau)$ be a quantization function that maps a timestamp $\tau$ into one of $m$ temporal bins for a stream that begins at timestamp $0$ and ends at timestamp $N$, returning a bin index in the range $[1, m]$. This quantization function could be designed to operate at different levels of granularity, for example mapping timestamps to hours of the day, days of the week, months of the year, etc. Similar to hierarchical geo-binning, we can use a binary tree to partition the one dimensional space, by recursively dividing it into 2 (equal-sized) subspaces. Here the data points are timestamps.

## 3.2 PATTERN COMPARISON AND CLUSTERING

One common yet important step in research relevant to geo-temporal datasets is to compare the geo-temporal patterns and quantify the pairwise relationships. Following that, we can perform clustering algorithms to find interesting groups with shared characteristics.

***Similarity measurement.*** When we have extracted the vectors representing the geo-temporal patterns, we can compute their pairwise similarity using simple metrics such as Manhattan distance, Euclidean distance, histogram intersection distance and chi-squared distance [26]. They quantify how similar two patterns are. However, we are sometimes interested in the relationships between two vectors: how do changes in one vector correspond to changes in another? For this purpose, Pearson product-moment correlation coefficient measures the linear correlation between two ran-

dom variables. Unlike the aforementioned similarity measurements, the coefficient is between +1 and -1, with 1 being total positive correlation, 0 being no correlation and -1 being total negative correlation. The Pearson product-moment correlation coefficient for vector $X$ and $Y$ is defined as the covariance of $X$ and $Y$ divided by the product of their standard deviation:

$$P(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

where $X = x_1, x_2, ..., x_n$, $\bar{x}$ is mean of $X$, $cov(X,Y)$ calculates the covariance and $\sigma_X$ is the standard deviation of $X$. When Pearson coefficient is applied to measuring the similarity between time series, it can be modified as a function of a temporal shift — the two time series can be similar in shape but with shifts along time. We define it here as also suggested in [88]. We assume streams $\mathcal{X}$ and $\mathcal{Y}$ begin at timestamp 0 and ends at timestamp $N$. First, let $U_{\mathcal{X}}(t_s, t_e)$ be a function that extracts a vector $X$ for $\mathcal{X}$, from starting timestamp $t_s$ to ending timestamp $t_e$. Similarly, $U_{\mathcal{Y}}(t_s, t_e)$ extracts a vector $Y$ for $\mathcal{Y}$. The cross-correlation is defined as:

$$f(\mathcal{X}, \mathcal{Y}, \Delta t) = \begin{cases} P(U_{\mathcal{X}}(\Delta t, N), U_{\mathcal{Y}}(0, N - \Delta t)), \Delta t \geq 0 \\ P(U_{\mathcal{X}}(0, N - |\Delta t|), U_{\mathcal{Y}}(|\Delta t|, N)), \Delta t < 0 \end{cases} \tag{3.2}$$

If $\Delta t \geq 0$, the starting point of $X$ is shifted to a later time stamp and meanwhile the end point of $Y$ is shifted to an earlier time stamp to ensure that the two resulting vectors have the same dimension. If $\Delta t < 0$, the starting point of $Y$ is shifted to a later time stamp and the end point of $X$ is shifted to an earlier time stamp. We can further calculate the lag (shift) in a temporal shift range ($[-\mathcal{S}_1, \mathcal{S}_2]$) that best aligns the two time series, or in other words, maximizes the cross-correlation. The lag $l_{\mathcal{X},\mathcal{Y}}$ is computed as:

$$l_{\mathcal{X},\mathcal{Y}} = \underset{\Delta t \in [-\mathcal{S}_1, \mathcal{S}_2]}{\operatorname{argmax}} f(\mathcal{X}, \mathcal{Y}, \Delta t), \tag{3.3}$$

where positive $l_{\mathcal{X},\mathcal{Y}}$ suggests $\mathcal{X}$ lags $\mathcal{Y}$ by $l_{\mathcal{X},\mathcal{Y}}$ while negative $l_{\mathcal{X},\mathcal{Y}}$ suggests $\mathcal{Y}$ lags $\mathcal{X}$ by $|l_{\mathcal{X},\mathcal{Y}}|$.

Two time series being compared may sometimes have different scales. In this case, the Dynamic Time Warping algorithm implemented with a dynamic programming approach [8] can measure their similarity by finding a minimal optimal match between the two time series.

***Clustering.*** With the geo-temporal vectors extracted and the similarity metrics chosen, K-means and spectral clustering techniques can cluster these vectors to reveal geo-temporal patterns and knowledge [121, 123]. For datasets with geolocations, the data mining tasks sometimes involve clustering the geolocations to find the places of interest (e.g. the most photographed places or the user's home/work locations). Clustering techniques such as K-means used in [3], DBSCAN in Chapter 6 and mean shift used in [23] are usually applied for this purpose. While K-means tends to discover geographical clusters of convex shapes, DBSCAN, as a density-based clustering method, can discover clusters with more arbitrary shapes.

## 3.3 VISUALIZATION

We describe visualization approaches that make the geo-temporal tagged data understood more easily.

***geospatial visualization.*** After quantization of geo-tags, we often want to interpret the process geospatial data. One common practice is to make the quantity in each geo-bin better understood visually, by utilizing colors and shapes overlaid on a 2D or a 3D map to create "heatmaps". In Figure 3.1, two heatmaps show the geographic distributions for tag "corn" and "coconut" respectively, where the intensity of green color is a function of the value in the corresponding geo-bin. In this case, the value is the number of unique users who applied this tag in this bin. Similarly, heatmaps in 3D can be produced. Figure 3.3 displays a schematic map of North America, with colored peaks in 3D space, indicating the usage of tags related to the Toronto and Niagara area. The heights of peaks, as well as the intensity of the red color, correspond to the tag usage in the area underneath.

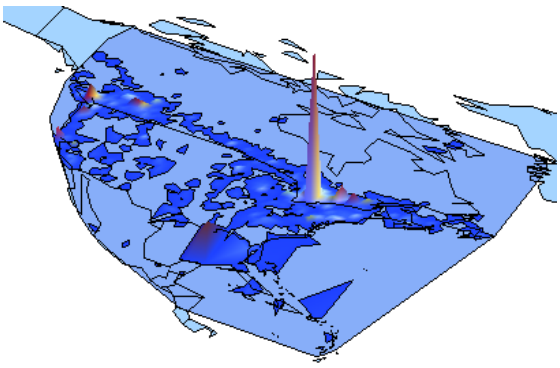


Figure 3.3: Tag usage for “toronto”, “niagara”, “niagarafalls”, “cntower”, “falls”, “ontario”, “canadian”, “canada”, “streetcar”, visualized in a 3D heatmap. Best viewed in color.

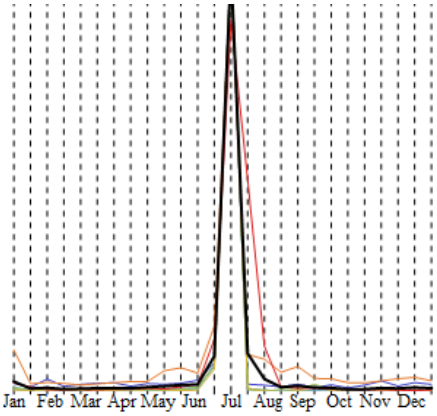



Figure 3.4: Tag usage for “4th”, “fourthofjuly”, “4thofjuly”, “independenceday”, “july4th”, “fireworks”, “july”, visualized as time series. Best viewed in color.

This redundancy in the representation may help delivering the information.

***Temporal visualization.*** The time series extracted by the simple temporal binning method can be visualized by plotting the values along the temporal dimension, where curves for multiple time series can be displayed within one graph showing the trends. Figure 3.4 shows the temporal usage of a set of tags that appear relevant to United States’ Independence Day, the usage of which peaks on July 4th. In the graph, the X-axis corresponds to time while the Y-axis corresponds to the tag usage.

***Tag clouds.*** Besides this, drawing tag clouds is the usual method to visualize tags, their frequencies and co-occurrences [59]. This method can potentially be applied to visualize other weighted graphs.

## 3.4 EXTERNAL PHENOMENON DETECTION AND ESTIMATION

There has been a lot of work on utilizing the signals from user-sensed data to detect the phenomena of interest and estimate their magnitude. We explain the mechanism of phenomenon detection from a Bayesian point of view and describe how the confidence scores calculated by applying a simple probabilistic model can be fit into a regression model to estimate the magnitude of the phenomena.

***A simple probabilistic model.*** We introduce a simple probabilistic model and use it to derive a statistical test that can deal with some noise and bias in the user-sensed data. Any piece of observational data (e.g. a Tweet, a photo or a search query) either contains the evidence of the phenomenon of interest (event $e$) or does not contain any evidence of the phenomenon of interest (event $\bar{e}$). We assume that the piece of data generated at a time and place with the phenomenon has a fixed probability $P(e|phen)$ of containing the evidence of the phenomenon. We also assume that observational data generated at a time and place without the phenomenon have some non-zero probability $P(e|\overline{phen})$ of containing evidence of the phenomenon; this incorporates various scenarios including incorrect timestamps or geo-tags and misleading visual or textual evidence.

Let $m$ be the number of data pieces (e.g. tweets, Flickr photos, search queries) containing evidence of the phenomenon (event $e$), and $n$ be the number of data entries without evidence of the phenomenon (event $\bar{e}$), generated at a place and time of interest. Assuming that each data piece is captured independently, we can use Bayes' Law to derive the probability that a given place has the phenomenon of interest given its number of data pieces with the evidence or without the evidence,

$$P(phen|e^m, \bar{e}^n) = \frac{P(e^m, \bar{e}^n|phen)P(phen)}{P(e^m, \bar{e}^n)} \tag{3.4}$$

$$= \frac{\binom{m+n}{m} p^m (1-p)^n P(phen)}{P(e^m, \bar{e}^n)}, \tag{3.5}$$

where $e^m$ and $\bar{e}^n$ denote $m$ occurrences of evidence $e$ and $n$ occurrences of evidence $\bar{e}$ respectively, and where $p = P(e|phen)$ and $P(phen)$ is the prior probability of the phenomenon of interest. A

similar derivation gives the posterior probability that the time and place (temporal geo-bin) does not contain the phenomenon,

$$P(\overline{phen}|e^m, \bar{e}^n) = \frac{\binom{m+n}{m} q^m (1-q)^n P(\overline{phen})}{P(e^m, \bar{e}^n)}, \tag{3.6}$$

where $q = P(e|\overline{phen})$. Taking the ratio between these two posterior probabilities yields a likelihood ratio,

$$\frac{P(phen|e^m, \bar{e}^n)}{P(\overline{phen}|e^m, \bar{e}^n)} = \frac{P(phen)}{P(\overline{phen})} \left(\frac{p}{q}\right)^m \left(\frac{1-p}{1-q}\right)^n. \tag{3.7}$$

This ratio measures the confidence that the phenomenon of interest actually happened in a given time and place, given pieces of observational data. We may classify a data piece into an event of positive evidence $e$ or an event of negative evidence $\bar{e}$ by leveraging textual or visual features, depending on the specific data being dealt with. The confidence score can then be thresholded to decide the binary existence of the phenomenon.

The above derivation is based on the assumption that the data pieces are generated independently of one another, which is generally not true in reality. One particular source of dependency is that data (tweets, Flickr photos) from the same user are highly correlated with one another. In order to prevent active users from spamming the datasets, instead of counting $m$ and $n$ as numbers of data pieces, we let $m$ be the number of *users* having at least one data piece with evidence of the phenomenon, while $n$ is the numbers of users who did not generate any data with evidence of the phenomenon.

***Regression-based methods.*** The confidence scores or other values estimated from the data sometimes not only indicate the presence or absence of the phenomenon of interest, but also linearly correlate with the magnitude of the phenomenon (e.g. degree or quantity of natural phenomena, box-office revenue of movies). As a result, we can fit linear regression models with the confidence

scores as predictors and one simple example can be written as:

$$Y = \beta X + \epsilon, \tag{3.8}$$

where $X$ is the predictor and $Y$ is the magnitude of the phenomenon of interest. $\beta$ and $\epsilon$ are the regression coefficient and the error term that can be estimated using measured data. It can be further extended to a multiple linear regression model if there is more predictors:

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon. \tag{3.9}$$

Regression-based methods have various applications in phenomenon detections and predictions that leverage user-sensed data. In Chapter 5, we will detail some related techniques applied to the detection and estimation of ecological phenomena. There is work on predicting future reports that reflect current status. Ginsberg *et al* [38] observe that the usage of certain search queries is highly correlated with the presence of influenza-like illnesses reported in official data. They fit a multiple linear regression model with the usage of top correlated queries as predictors and use it to estimate the regional presence of these illnesses one or two weeks ahead of the official data. In [53], the authors extend an autoregressive model with index estimated from Flickr photos to predict product sales. In order to predict box-office revenue of movies and Billboard song rankings, Goel *et al* [39] fit multiple linear regression models on publicly available data as well as Yahoo search counts, while in [20] researchers predict near-term values of economic indicators such as automobile sales and unemployment claims.

***Other classifier-based methods.*** Other than fitting the regression models on a few estimated values such as the confidence scores, we may incorporate more information to build labeled feature vectors for classifiers such as Support Vector Machines, Naive Bayes, and Decision Trees to detect or quantify the phenomenon of interest, depending on the application scenarios. For example, as described

in Section 5, in order to detect the natural phenomena snow using tagged Flickr data, we can build a feature vector based on the tag usage distribution for a day and place, where the label is 'snow' or 'no snow' judged from satellite ground truth data. Aforementioned classifiers can be trained using such vectors to make binary predictions about the existence of snow.

***Case-based methods.*** There is a line of work on learning from news archives to predict future events by Radinsky *et al* [89, 90]. They extract past event chains from news articles and further generalize the event entities (e.g. actors, actions, locations) using world knowledge ontologies. From these past event chains, conditioned probabilities are learned. Future events are then generalized and their outcomes are predicted based on the learned probabilities.

# CHAPTER 4

# TAGGED DATA EXPLORATION FOR REAL WORLD KNOWLEDGE

Tagged crowd-sourcing datasets have grown to large scales, opening up opportunities to monitor the real world geo-temporally. Given the abundance of geo-temporal signals provided by the geo-tags and timestamps, can we discover geo-temporal relationships between photo tags that reveal connections between real world concepts? As motivated in Section 1.4.1, this chapter gives an answer to this question. By answering this question, we will also show that the virtual world reflects the real world, as a first step towards answering our meta research question.

Researchers study the relationships between keyword tags on social sharing websites to improve tag suggestion systems and to discover the connections between the concepts the tags represent at a semantic level. Existing approaches mainly rely on tag co-occurrences. In this project, we show how to find connections between tags by comparing their distributions over time and space, discovering tags with similar geographic and temporal patterns of use. Geospatial, temporal and geo-temporal distributions of tags are extracted and represented as vectors which can then be compared and clustered. Using a dataset of tens of millions of geo-tagged Flickr photos, we show that we can cluster Flickr photo tags based on their geographic and temporal patterns, and we evaluate the results both qualitatively and quantitatively using a panel of human judges. A case study suggests that our

visualizations of temporal and geographic distributions help humans recognize subtle semantic relationships between tags. This approach to finding and visualizing similar tags is potentially useful for exploring any data having geographic and temporal annotations.

## 4.1 DISCOVERING TAG RELATIONSHIPS

We assume that we have a dataset of online objects (e.g. photos), each of which has a user id of the person generating the object (e.g. the photographer), a timestamp specifying when the object was created, a geo-tag specifying latitude-longitude coordinates for the object, and a set of zero or more text tags. To define this formally, it is useful to think of tagging *events* – individual acts of a user tagging a photo with a text tag. The information associated with a particular event includes the tag that was applied, the photo to which the tag was applied, the user who uploaded the tagged photo, the geographic location of the tagged photo, and the timestamp indicating when the photo was taken. Letting $\mathcal{T}$ be the set of all possible text tags, then the set of tagging actions $\mathcal{A} = \{a_1, a_2, ..., a_q\}$ can be defined as a set of tuples of the form $a_i = (u_i, t_i, \tau_i, g_i)$, where $u_i$ is a user, $t_i \in \mathcal{T}$ is a tag, $\tau_i$ is a timestamp, and $g_i \in \mathcal{R} \times \mathcal{R}$ is a geo-tag (latitude-longitude coordinate).

Given a collection of tagged objects, our goal is to cluster the tags based on their geospatial and temporal properties. To do this, we first extract a geospatial or temporal signature for each tag and represent it with a corresponding feature vector, and then we cluster the feature vectors using an unsupervised algorithm like $k$-means [72]. In addition to finding the geospatial and temporal distributions of each tag, we also compute a feature vector based on the cross-product of these two attributes, which allows us to represent a tag's joint geo-temporal distribution – i.e. the "motion" of how a tag's spatial distribution changes over time. The following three subsections explain the geospatial, temporal, and motion feature vectors in turn.

### 4.1.1 GEOSPATIAL FEATURE VECTORS

Since different types of tags have different geographical distributions, our first feature aims to characterize the geographical distribution of a tag. As is discussed in Section 3.1, the geographic distribution of the world's photos is highly non-uniform, with hot spots of the photographic activity, usually in cities. It is thus useful to aggregate photos together into coarse geospatial buckets instead of clustering using raw geo-tags. We apply the geospatial binning technique described in the aforementioned section. To show how it fits into this application scenario and to be coherent, we elaborate it here. We first divide the world into $n$ bins and each one has the size of $s$ degrees of latitude by $s$ degrees of longitude. We assign each bin a unique index number in the range $[1, n]$, and define a quantization function $q_G(g)$ that maps a latitude-longitude coordinate $g$ into the index number of the corresponding geo-bin. In the results presented in this thesis, we use $s = 1$ degree, which corresponds to grid cells of roughly 100 km $\times$ 100 km at the middle latitudes.

To compute a geospatial feature vector for tag $t$, we first count the number of unique users who have used that tag in each geo-bin $g$,

$$U_G(g,t) = || \{u_i | (u_i, t_i, \tau_i, g_i) \in \mathcal{A}, t_i = t, g = q_G(g_i)\} ||. \tag{4.1}$$

We count the number of *users* who applied a tag within a geographic area instead of the number of photos in order to prevent high-activity users from biasing the distribution [3]. (This can be thought of as giving each user a single "vote" for whether or not a tag applies to a given geographic area.) Then we normalize the vector to get the geospatial feature $v^G(t)$ of tag $t$,

$$v_i^G(t) = \frac{U_G(i,t)}{\sqrt{\sum_{j=1}^{n} U_G^2(j,t)}}. \tag{4.2}$$

Normalization is necessary since tags sharing similar geospatial distributions might have different overall frequencies of occurrence. While other work has found that L1 normalization works better

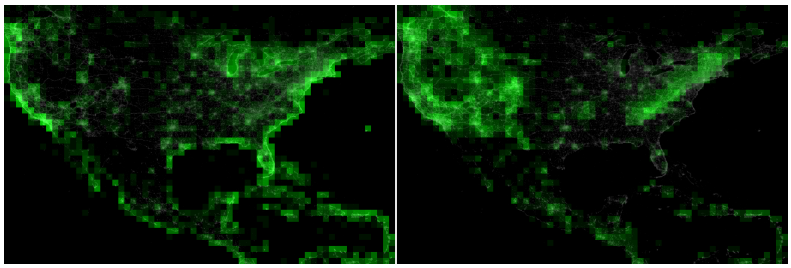


Figure 4.1: Geographic distributions for tag “beach” (left) and “mountains” (right).

in high dimensional spaces [27, 80], we found that L2 normalization works better in our context. (For example, for the clustering results presented in Section 5.2, we found that L2 norm generates clusters that have more uniform and moderate sizes. When clustering 2000 tags into 50 clusters, L1 norm produces 17 singletons (clusters that contain only one tag) while L2 norm produces no singletons. L1 norm cluster sizes also have much greater variation: their standard deviation is 127.8 while the standard deviation of L2 norm is 54.1.)

As an example, Figures 4.1 visualizes the normalized matrices for tags “beach” and “mountains” over North America, in which greater intensity indicates that more users applied the tag to photos in a given geo-bin. Notice that the Appalachian and Rocky Mountain ranges are immediately apparent in the “mountains” map, while the “beach” map highlights the coastline of North America.

### 4.1.2 TEMPORAL FEATURE VECTORS

Tags also have different temporal distributions, because some tags (and semantic concepts) are much more popular at certain times than others — for example, we might expect “beach” to be used more often during the summer than in the winter, while “restaurant” might occur more often during the meal times of the day. As with the geographic feature vectors described above, with temporal features it is also useful to aggregate photos together into coarse temporal bins. Let $q_T(\tau)$ be the quantization function described in Section 3.1.2. It maps a timestamp $\tau$ into one of $m$ temporal bins and returns a bin index in the range $[1, m]$. For any tag $t$, we then build an $m$-dimensional vector,

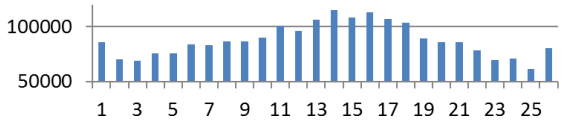



Figure 4.2: Number of unique Flickr users active in North America, for each 2-week period of the year.

again counting the number of unique users who have used the tag in each temporal period $p$,

$$U_T(p,t) = ||\{u_i|(u_i,t_i,\tau_i,g_i) \in \mathcal{A}, t_i = t, p = q_T(\tau_i)\}||, \tag{4.3}$$

and then normalize to produce an $m$-dimensional temporal feature vector $v^T(t)$,

$$v_i^T(t) = \frac{U_T(i,t)}{\sqrt{\sum_{j=1}^{m} U_T^2(j,t)}}. \tag{4.4}$$

In this project we primarily use a quantization function that maps timestamps into one of 26 two-week periods of the year: January 1-14, January 15-28, etc. We disregard the specific year and as a result, all the data is merged together into a single year – for example, photos taken on January 1, 2008 and January 1, 2009 will be mapped to the same cell. In addition to the 26-dimensional 2-week vectors, we also create 7-dimensional day-of-week vectors and 24-dimensional hour-of-day vectors.

Flickr users are significantly more active at certain times of the year than others, as illustrated in Figure 4.2: note that nearly twice as many users take photos during the first two weeks of July (period 14) than in early February (period 3). To correct for this effect, in practice we normalize the temporal counts for tag $t$ and period $p$, $U_T(p,t)$, by the total number of photos taken in North America during $p$, before applying L2 normalization to produce $v^T$. Figure 4.3 shows the process of obtaining temporal feature vectors.

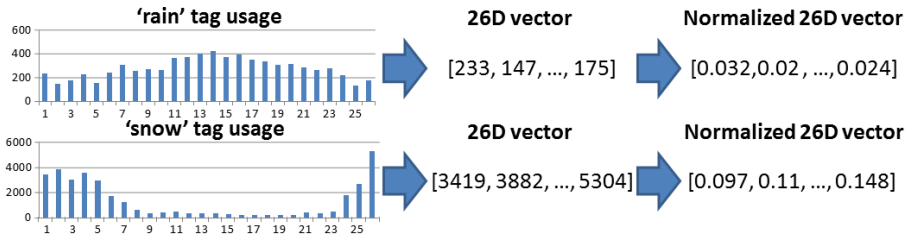



Figure 4.3: Computing temporal feature vectors.

### 4.1.3 GEO-TEMPORAL (MOTION) FEATURES

Finally, we also want to produce a signature for a tag based on its joint geo-temporal distribution – that is, how the geospatial distribution of the tag varies over the course of a year (or equivalently, how the temporal distribution varies with spatial location). We call these geo-temporal signatures our "motion" features. Given geospatial and temporal quantization functions $q_G$ and $q_T$ (described above) that map geo-tags into one of $n$ bins and timestamps into one of $m$ bins, a motion feature vector has one bin per entry in the cross product of these two sets of indices, or $mn$ dimensions total. More precisely, we define a motion quantization function that maps a geo-tag $g_i$ and timestamp $\tau_i$ to a bin index in $[1, mn]$,

$$q_M(g_i, \tau_i) = m \times (q_G(g_i) - 1) + q_T(\tau_i), \tag{4.5}$$

then count the number of unique users who used a given tag $t$ in each geo-temporal bin,

$$U_M(m, t) = ||\{u_i | (u_i, t_i, \tau_i, g_i) \in \mathcal{A}, t_i = t, m = q_M(g_i, \tau_i)\}||, \tag{4.6}$$

and take the L2 norm (as above) to define a final motion feature vector, $v^M$. For the experiments in this project, this vector has $mn = 124,800$ dimensions. As with the geospatial feature vectors, we remove empty dimensions (geo-temporal cells having no photos) as an optimization.

### 4.1.4 CO-OCCURRENCE FEATURES

For comparison purposes, we also define similarity metrics using using two more traditional techniques. First, the pairwise co-occurrence between two tags $t_1$ and $t_2$, co_occur$(t_1, t_2)$, is computed by simply counting the number of photos that are tagged with both $t_1$ and $t_2$.

A disadvantage of this simple co-occurrence measure is that it favors pairs of tags that occur very often, since very frequent tags will co-occur more often than infrequent tags even if they are unrelated. Thus we include a second baseline feature, mutual information, which overcomes this problem by normalizing the co-occurrence measures by the overall frequency of the tags [7],

$$\text{mutual\_info}(t_1, t_2) = \log\left(\frac{\text{co\_occur}(t_1, t_2)}{\text{occur}(t_1) \times \text{occur}(t_2)}\right) \tag{4.7}$$

where occur$(t)$ is the total number of photos having tag $t$. This score can be thought of as a measure of the independence of the two tags: it is minimized if the two tags are completely independent (never co-occur), and is maximized if the tags are strongly correlated (always co-occur).

## 4.2 EXPERIMENTS AND VISUALIZATIONS

To test our techniques for characterizing tags based on geospatial and temporal signatures, we used a dataset of geo-tagged, time-stamped photos downloaded from Flickr through the site's public API interface, using a crawling technique similar to that described in [23]. We collected the following information for each photo: the geo-tag (latitude and longitude) of where the photo was taken, the timestamp of when it was taken, and the set of textual tags (if any) associated with the photo. From this collection of nearly 80 million photos, we selected only the photos in North America (which we defined to be a rectangular region spanning from 10 degrees north, -130 degrees west to 70 degrees north, -50 degrees west). Other details about the data can be found in Section 1.3.

We then computed the top 2000 most frequent text tags (ranked by the number of different users

who used the tag) in North America. (We chose this relatively small number of tags so that our geo-temporal distributions would have substantial mass not dominated by noise (each of these tags has been used by at least 1,200 unique Flickr users), and to make human evaluation tractable. Note that the majority (66.7%) of photos on Flickr are tagged with at least one of these 2,000 tags since Flickr tag frequency follows a long tailed distribution [99].) For each of these tags, we extracted the geospatial feature vectors, temporal feature vectors and motion vectors described in Section 5.1.1. In preparation for geospatial feature vector extraction, we filtered the data by removing photos with geotag precision less than about city-scale (according to the precision reported by Flickr), resulting in a datset with about 44 million photos. For the temporal feature vectors, we removed photos with inaccurate or suspicious timestamps (including photos supposedly taken in the future or distant past), resulting in about 41 million photos; for the motion feature vectors, both of these filters were applied, yielding about 39 million photos.

### 4.2.1 TAG RELATIONSHIPS

We can use the similarity metrics defined in Section 5.1.1 to find pairs of similar tags. Given a tag $t'$, we can find a list of related tags using each of the distances defined above, including geospatial, temporal, and geo-temporal. To do this, we compute the feature vectors $v^G(t)$, $v^T(t)$, and $v^M(t)$ for each tag $t$, and then compute the pairwise Euclidean distances between these vectors and those of $t'$. The tags are ranked according to their distances to $t'$ in ascending order, and the $k$ tags with lowest distance are found. For the co-occurrence and mutual information features, we compute the similarity for each tag $t$ using the co_occur$(t, t')$ and mutual_info$(t, t')$ functions, and then rank the tags in increasing order of similarity.

As an example, Table 4.1 shows the lists of tags that are most similar to the tag "cherryblossoms" under the various measures of similarity. The first column shows the 20 most similar tags according to the geospatial similarity metric. Most of these tags are strongly related to Washington, DC

Table 4.1: Top 20 most similar tags to "cherryblossoms" using different similarity metrics. The columns rank the tags according to (from left): geospatial, temporal, motion (geo-temporal), co-occurrence, and mutual information.

| | geospatial | | Temporal | | Motion | | Co-occurrence | | Mutual information | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | president | 0.321 | cherry | 0.451 | cherry | 0.291 | washingtondc | 4568 | blossoms | -10.577 |
| 2. | whitehouse | 0.321 | blossoms | 0.505 | blossoms | 0.417 | dc | 3443 | cherry | -11.004 |
| 3. | monument | 0.323 | blossom | 0.612 | blossom | 0.546 | spring | 2319 | washingtonmonument | -11.355 |
| 4. | smithsonian | 0.324 | easter | 0.636 | jefferson | 0.673 | washington | 2089 | buds | -12.289 |
| 5. | memorial | 0.324 | spring | 0.638 | kite | 0.868 | flowers | 1969 | washingtondc | -12.456 |
| 6. | georgetown | 0.325 | april | 0.654 | washingtonmonument | 0.908 | blossoms | 1367 | dc | -12.580 |
| 7. | washingtonmonument | 0.327 | magnolia | 0.735 | monument | 0.990 | pink | 1007 | hearts | -12.622 |
| 8. | dc | 0.327 | buds | 0.758 | magnolia | 0.998 | trees | 979 | jefferson | -12.657 |
| 9. | lincolnmemorial | 0.328 | washingtonmonument | 0.813 | spring | 1.026 | canon | 754 | blossom | -12.659 |
| 10. | wwii | 0.331 | tulip | 0.822 | bloom | 1.042 | cherry | 753 | spring | -12.760 |
| 11. | washingtondc | 0.332 | jefferson | 0.837 | memorial | 1.057 | tree | 703 | lions | -12.785 |
| 12. | jefferson | 0.367 | egg | 0.858 | lincolnmemorial | 1.085 | usa | 610 | pink | -12.845 |
| 13. | arlington | 0.370 | tulips | 0.862 | washingtondc | 1.086 | flower | 560 | bloom | -12.884 |
| 14. | lincoln | 0.372 | bloom | 0.868 | dc | 1.091 | washingtonmonument | 552 | petals | -12.955 |
| 15. | mall | 0.392 | break | 0.869 | whitehouse | 1.092 | water | 498 | poppy | -13.051 |
| 16. | capitol | 0.401 | poppy | 0.870 | mall | 1.096 | 2007 | 415 | flowers | -13.263 |
| 17. | soldier | 0.407 | eggs | 0.883 | festival | 1.103 | unitedstates | 412 | drops | -13.283 |
| 18. | war | 0.421 | bud | 0.895 | tulip | 1.106 | festival | 379 | branches | -13.327 |
| 19. | cherry | 0.429 | kite | 0.922 | government | 1.121 | nature | 377 | japan | -13.381 |
| 20. | capital | 0.448 | olympics | 0.924 | capital | 1.122 | brooklyn | 347 | shell | -13.385 |

(which is of course famous for its annual cherry blossom festival in April), including "president," "whitehouse," "smithsonian," and "lincolnmemorial," among others. The second column shows tags having high temporal similarity, including "easter," "spring," "april," and "magnolia". The list of tags under motion similarity appear to be a mixture of geographically similar tags and temporally similar tags. In contrast, the co-occurrence list has arguably much lower quality: "canon," "water," and "usa" are popular tags that also co-occur with many other tags, and are not particularly relevant to "cherryblossoms." Mutual information gives more meaningful results compared to raw co-occurrence, but some tags such as "lions" and "drops" cannot be interpreted easily and do not have clear geospatial or temporal relationships to chery blossoms. Moreover, the mutual information list missed tags like "whitehouse," "april", and "kite" which were picked up by the temporal and geospatial analyses. These tags do not frequently co-occur with "cherryblossoms" on the same photos, but do share similar geospatial and/or temporal patterns.

### 4.2.2 CLUSTERING TAGS

The analysis in the last section can be used to compute the similarity between any arbitrary pair of tags, but it is difficult to visualize or quantify the performance of these results directly because

there are so many possible pairs. In past work, tag similarity results have been summarized by grouping tags into a small number of similar clusters, typically using co-occurrence information (e.g. [18]). We follow a similar strategy and cluster Flickr tags according to each of our three types of distance metrics (temporal, geospatial, geo-temporal) as well as traditional co-occurrence and mutual information measures.

For each feature vector type, we clustered the 2000 tags using $k$-means [72]. Squared Euclidean distance was used to measure distances between vectors. Since $k$-means clustering is sensitive to the initial choice of centroids, we ran $k$-means five times with different random initial cluster centers and chose the best result (in this case, choosing the clustering with the minimum total vector-to-centroid distance). Of course, this clustering algorithm requires an *a priori* choice of the number of clusters ($k$). For the purposes of this project, where our primary focus is on presenting techniques for comparing tags based on their geospatial and temporal distributions and not on presenting an end-to-end system for tag analysis, we simply set $k$ to a value (50) that gave reasonable results in our subjective judgment. Various techniques exist for selecting $k$ automatically based on properties of a dataset (see e.g. [105]) and these techniques could easily be applied to our work.

As we show in the next few sections, the overall "shape" of the geospatial and temporal distributions varies dramatically from tag to tag and cluster to cluster: some clusters contain tags that are diffuse across space and time (like "canon", "geotagged", "blackandwhite", etc.), while other distributions are very "peaky" ("newyorkcity", "washingtondc", etc.), and others are somewhere in between ("usa", "newengland", etc.). It is thus useful to compute a statistical measure of the peakiness of a distribution, in order to compactly characterize its overall "shape". We measure the peakiness of a vector $v$ by computing its second moment,

$$\text{second\_moment}(v) = v \cdot v = \sum_{i=1}^{n} v_i^2, \tag{4.8}$$

and measure the peakiness of a cluster of tags $C$ as the average second moment of the vectors that

it contains,

$$\frac{\sum_{v \in C} \text{second\_moment}(v)}{|C|}. \tag{4.9}$$

Peaky distributions will have higher second moment values, while distributions close to uniform will have low second moment. (Note that the second moment of a discrete probability distribution is the likelihood of sampling twice from the distribution and drawing the same value both times.) The second moment gives a statistic by which to rank clusters, as clusters with high average peakiness usually have bursts in temporal distributions or geospatial distributions which indicate particularly interesting clusters.

We present sample results and visualizations for several sample tag clusters in the following sections. Due to space limitations we do not show all the tag clusters generated from the three different perspectives, but these and other detailed results are available at `http://vision.soic.indiana.edu/tagclusters`.

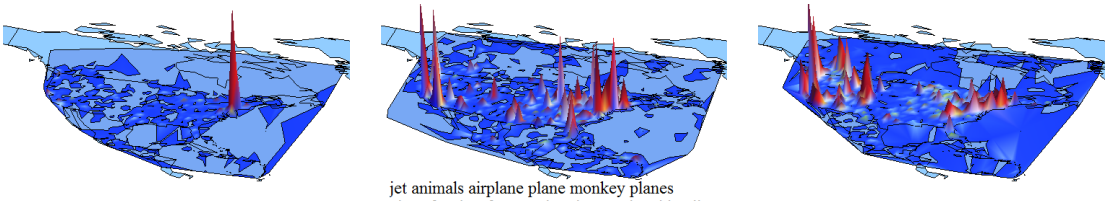
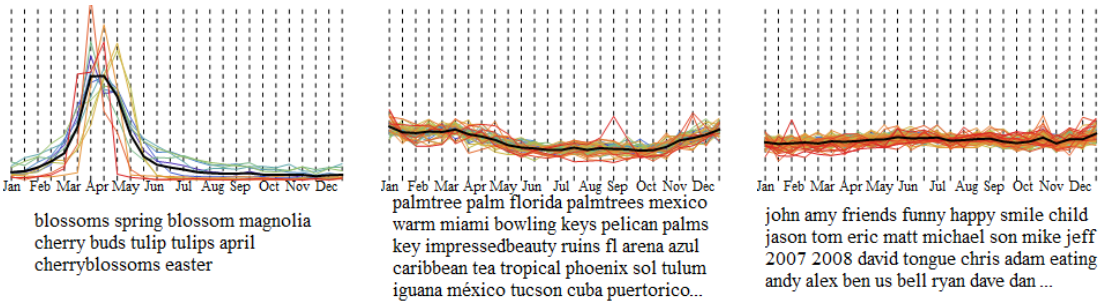



Figure 4.4: Visualizations of sample clusters produced by analyzing similarity of geospatial distributions. The clusters seem to correspond to (a) tags related to New York City, (b) tags related to cities with popular zoos and airports, and (c) tags related to national parks and outdoor areas. Best viewed in color.

## GEOGRAPHICAL CLUSTERS

Figure 4.4 shows visualizations of several tag clusters produced by analyzing geospatial distributions. The visualizations were created by taking a cluster centroid and converting it back to a two-




(a) (b) (c)

Figure 4.5: Visualizations of three clusters produced by analyzing similarity of temporal distributions: (a) tags related to spring, (b) tags related to winter, (c) tags related to gatherings of friends.

dimensional matrix: i.e. for each geo-bin, we find the corresponding latitude-longtiude coordinates for the bin center and plot them together with their values in a 3D space over a schematic map of North America. The result is a topographical visual effect with the heights of the peaks as well as the intensity of red color indicating the usage of the tags in the cluster at corresponding locations underneath. The figure shows three sample clusters. Figure 4.4(a) consists of tags from the New York City area. Some not very obvious tags are: "soho" which is a shopping area in New York City, "bull" which is the Wall Street Bull and "ground" which relates to Ground Zero referring to World Trade Center site. This cluster is ranked 10th out of 50 clusters by second moment. Figure 4.4(b) visualizes the cluster in which most tags are related to zoos and animals and others are related to airports. As a result, the visualization peaks at major US cities with famous zoos and airports. It is ranked 37th. Figure 4.4(c) displays the cluster of tags that occur predominantly in national parks. This cluster is ranked 27th.

Images that record the visualizations of all the 50 geospatial clusters are available at the above website. The top ranked clusters by second moment are more concentrated geographically. From these top clusters, we see state clusters, city clusters, zoo clusters, park clusters, northern city clusters, coastal area clusters and so on. For lower ranked clusters, the tags are more geographically distributed, such as a cluster of rural regions and a cluster of urban regions.

**Top 10 temporal clusters**

| | Tags in cluster | # tags | 2nd moment |
|---|---|---|---|
| 1 | **4th fourthofjuly 4thofjuly independenceday july4th** | 7 | 0.578 |
| 2 | **january newyearseve** | 2 | 0.5 |
| 3 | **turkey thanksgiving november** | 3 | 0.355 |
| 4 | **august** | 1 | 0.2785 |
| 5 | iris may dandelion graduation memorialday | 5 | 0.269 |
| 6 | **costume costumes halloween** | 3 | 0.223 |
| 7 | **christmastree christmaslights christmas ornament holidays** | 9 | 0.215 |
| 8 | pride june | 2 | 0.206 |
| 9 | **fallcolors pumpkins autumn fall foliage** | 7 | 0.190 |
| 10 | irish march | 2 | 0.144 |

**Top 10 geographical clusters**

| | Tags in cluster | # tags | 2nd moment |
|---|---|---|---|
| 1 | ***toronto niagara niagarafalls cntower falls*** | 9 | 0.415 |
| 2 | ***golden cablecar francisco sanfrancisco sf*** | 27 | 0.402 |
| 3 | ***los angeles santamonica la losangeles*** | 8 | 0.397 |
| 4 | ***broadway brooklyn empire cab empirestatebuilding*** | 34 | 0.394 |
| 5 | ***strip paris vegas las lasvegas*** | 10 | 0.379 |
| 6 | ***seattle needle pugetsound spaceneedle wa*** | 8 | 0.374 |
| 7 | ***chicago bean searstower illinois il*** | 7 | 0.366 |
| 8 | ***ma massachusetts boston cambridge newengland*** | 6 | 0.332 |
| 9 | prairie pennsylvania pa philadelphia philly | 58 | 0.287 |
| 10 | ***911 liberty chelsea wtc worldtradecenter*** | 14 | 0.276 |

**Top 10 motion (geo-temporal) clusters**

| | Tags in cluster | # tags | 2nd moment |
|---|---|---|---|
| 1 | ***losangeles angeles santamonica los la*** | 7 | 0.021 |
| 2 | ***taxi broadway empirestatebuilding brooklyn empire*** | 38 | 0.01693 |
| 3 | ***tx texas austin houston dallas*** | 5 | 0.01691 |
| 4 | ***chicago searstower bean illinois il*** | 7 | 0.01679 |
| 5 | ***vegas lasvegas las bellagio strip*** | 10 | 0.01671 |
| 6 | ***alberta calgary banff*** | 3 | 0.01606 |
| 7 | ***francisco sanfrancisco goldengatebridge goldengate berkeley*** | 30 | 0.0158 |
| 8 | ***pa philadelphia philly pennsylvania*** | 4 | 0.0157 |
| 9 | ***statueofliberty liberty newjersey jersey nj*** | 13 | 0.0148 |
| 10 | ski skiing snowboarding tahoe fdsflickrtoys | 73 | 0.0138 |

**10 randomly-chosen co-occurrence clusters**

| | Tags in cluster | # tags |
|---|---|---|
| 1 | ***sea ocean beach boat island*** | 41 |
| 2 | ***coast waves sun shore pier*** | 41 |
| 3 | ***washington statue museum sculpture washingtondc*** | 41 |
| 4 | people model female face hair | 43 |
| 5 | ***vacation travel trip desert arizona*** | 41 |
| 6 | ***water canada winter sky nature*** | 41 |
| 7 | ***trees mountains mountain hiking hike*** | 41 |
| 8 | party wedding friends love dance | 40 |
| 9 | light city night bed sleep | 39 |
| 10 | geotagged us building architecture canon | 41 |

**10 randomly-chosen mutual info. clusters**

| | Tags in cluster | # tags |
|---|---|---|
| 1 | rails rail railway train railroad | 36 |
| 2 | **independenceday july4th fourthofjuly 4th weird** | 39 |
| 3 | ***marriage rings groom love couple*** | 38 |
| 4 | plane jet aviation aircraft planes | 38 |
| 5 | furry sleepy pet kitten fur | 38 |
| 6 | jeans jacket socks shoes feet | 39 |
| 7 | furniture toilet sink seat couch | 34 |
| 8 | ***tide waves surf ocean wave*** | 41 |
| 9 | ***rockies rockymountains peak glacier summit*** | 38 |
| 10 | ***sail port harbor docks sailing*** | 37 |

Figure 4.6: Comparison of clusters produced by different similarity metrics: temporal (top left), geospatial (top right), and geo-temporal (center). Clusters judged to be temporally significant by human judges are printed in **blue boldface**, while clusters judged to be geographically related are printed in ***red boldface italics***. Clusters are sorted in decreasing order of second moment. For comparison, also shown are clusters produced by co-occurrence (bottom left) and mutual information (bottom right). For each cluster, up to 5 top ranked tags are displayed. Relevancy was judged by users without visualizations being shown.

## TEMPORAL CLUSTERS

Figure 4.5 shows visualizations of three of the clusters produced by the temporal similarity metric. For each temporal cluster, we plot the cluster centroid as well as the distributions of the individual tags within the cluster, with each point representing the usage in the corresponding two-week period. Figure 4.5(a) displays the cluster with a strong peak during spring. The thick black curve cooresponds to the cluster centroid while the other curves correspond to the signals for individual tags. This cluster is ranked 11th out of 50 clusters by second moment. Figure 4.5(b) visualizes a

cluster with a shallow peak during the winter season. Most tags are related to winter vacations in warm locales. This cluster is ranked 34th. The cluster in Figure 4.5(c) (ranked 48th by second moment) seems to correspond to family gatherings, with slight temporal peaks around Thanksgiving and Christmas. There are also some year tags (e.g. "2008", "2009", etc.) which appear frequently around New Year's Day. Visualizations of all 50 temporal clusters are available at the website above. We see that top ranked clusters of tags have sharp bursts in smaller time windows and lower ranked clusters have more general seasonal patterns.

We also clustered the 7-dimensional day-of-week vectors and 24-dimensional hours-of-day vectors. Due to space constraints we do not present detailed results, but instead mention a few interesting findings. For day-of-week vectors, we are able to see weekday clusters such as "work office desk students commute" and weekend clusters such as "live sushi gallery concert macys highschool moma". For hours-of-day clusters, we see clusters peak at different time of the day, such as a morning cluster, "early sunrise dawn morning," and a nighttime cluster, "lightning concert longexposure campfire nighttime nightphotography exposure live".

**GEO-TEMPORAL CLUSTERS**

Tags within a motion cluster typically share either geospatial and temporal similiarites, or both. Cluster "vegas lasvegas las bellagio strip paris casino nevada fountains flamingo" captures Las Vegas and its hotels and casinos which has a counterpart in geographical clusters. It is ranked 5th out of 50. Cluster "christmas holiday xmas holidays christmastree christmaslights december decorations ornament decoration cookies gift santa fireplace" captures Christmas which has a counterpart in temporal clusters and is ranked 31st. We observe that top ranked clusters are more likely to have obvious geographic connections while the clusters having temporal patterns are ranked in the middle. All the motion clusters can also be found at the website above.

### 4.2.3 EVALUATION

We evaluated the idea of finding similar tags using geospatial and temporal features by comparing the clustering results from our proposed methods with those of the co-occurrence based techniques. Because it is difficult to define the quality of a tag cluster objectively, we involved humans in our experiments to judge the geospatial and temporal relevance of the clusters we found. We then used the human judgment as ground truth to compute the precisions and recalls when the task is to retrieve semantically meaningful clusters in time and/or space by thresholding the average second moment values. The goal of this evaluation was to test whether our techniques produce coherent tag clusters that correspond to intuitive geospatial and temporal concepts, and how these clusters compare to traditional techniques that use co-occurrence.

#### CLUSTERING BASED ON TAG CO-OCCURRENCES

As a baseline we used the method based on tag co-occurrence described in [7] to cluster the top 2000 tags into 50 clusters for a fair comparison. In particular, we constructed an undirected graph of tags, weighted the edges between tags by metrics of their co-occurrences and removed weak edges by thresholding the weights. We then applied a graph partitioning program, KMETIS [58], to partition the graph into clusters. We tried two different methods to weight the edges, one by raw co-occurrence counts and the other by mutual information (defined in Section 5.1.1). As a result, we generated two sets of 50 clusters: co-occurrence clusters and mutual information clusters. For each cluster, we ranked its tags by the numbers of edges inside the cluster: tags with more edges are considered to be more representative.

#### GROUND TRUTH FROM HUMAN JUDGMENT

To conduct the human judgment study at a large scale, we used Amazon's Mechanical Turk service, asking users to judge the geospatial coherence and temporal coherence of clusters produced by

the various similarity metrics that we propose. To improve the quality of the human judgment, we required users to be in the United States (so that they would be familiar with North American geography and cultural events) and have a good (>=95%) historical approval rate. For each cluster discovered by our methods, we selected its ten top ranked tags to present to the user.

For each geospatial, motion, co-occurrence and mutual information cluster, we asked the users to judge its geographical relevance among the following three options: "more than 50% of its tag representatives represents a specific geographic area such as NYC", "more than 50% of its tag representatives represents an abstract geographic concept such as ocean," or "not geographically relevant." Similarly, for each cluster we also asked the users to judge its temporal relevance according to the scale: "more than 50% of its tag representatives represents a specific temporal event such as Thanksgiving", "more than 50% of its tag representatives represents a broad temporal indication such as spring", "not temporally relevant". For both the geospatial and temporal relevance questions, if at least 80% of the users choose the first or second option for a cluster, we consider it to be geospatially or temporally relevant respectively. We put the clusters in 16 batches of 25 clusters per assignment and each batch was assigned to up to 20 Mechanical Turk users. On average, each cluster was judged by 19.9 users. Users were shown only the top ten tags associated with each cluster. We conducted two independent sets of experiments, one in which the 20 users were not shown the visualization graphs described above, and another in which a separate group of 20 users were shown the visualizations, to quantify how useful the visualizations might be in practice.

## EVALUATION RESULTS AND A CASE STUDY

The evaluation found that geospatial and motion clustering were more effective in finding geographically relevant clusters than the other techniques: 29 (58%) of the geographical clusters were found to be geographically relevant by the human judges, compared to 30 (60%) of motion clusters, 11 (22%) of the co-occurrence clusters, and 11 (22%) of the mutual information clusters. A case study

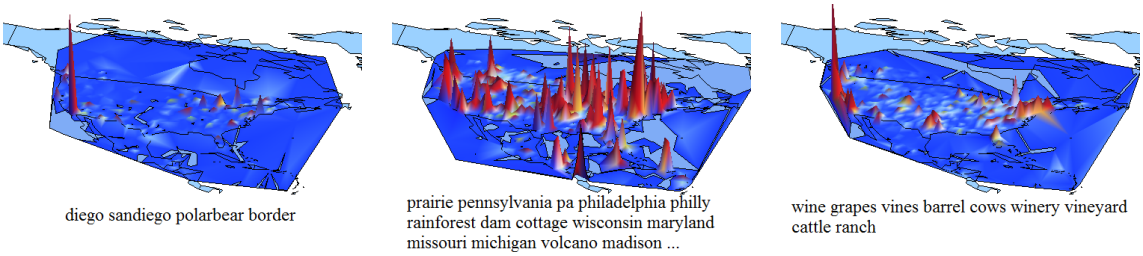



Figure 4.7: Sample geographical clusters judged to be not geographically relevant with high average second moment by users who were not shown the visualizations. The clusters correspond to (a) tags related to San Diego, (b) tags related to cities and states, and (c) tags related to Northern California Wine Country. Best viewed in color.

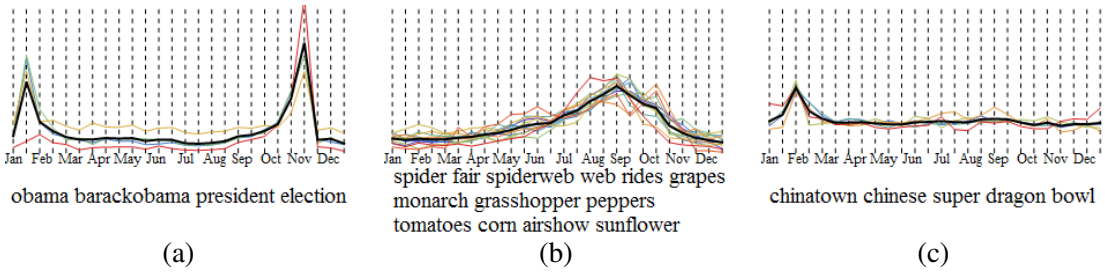



Figure 4.8: Sample temporal clusters judged to be not temporal relevant with high average second moment by users who were not shown the visualizations. The clusters correspond to (a) tags related to the Presidential election, (b) tags related to autumn, (c) tags related to Jan and Feb.

of some geo-temporal clusters judged not to be geo or temporally relevant showed the visualizations gave hints to people and helped understand the not-so-obvious semantics behind the clusters. For the geographical clusters, among the top 31 clusters ranked by average second moment, 28 were judged to be geographically relevant. We examined the 3 clusters that were judged to be "not geographically relevant," and even these appeared to have interesting geographical semantics that were likely not obvious to the human judges. Visualizations for these 3 clusters are shown in Figure 4.7. The cluster in Figure 4.7(a) has an obvious peak in San Diego. The terms "polarbear" and "border" may not be immediately associated with San Diego, but in fact they refer to the San Diego Zoo's famous polar bears while the tag "border" refers to the Mexican border which is just a few miles away. In Figure 4.7(b), most tags are state or city names. They are in one cluster as their geographical distributions are very concentrated resulting in very peaky geo vectors which are not far from

each other as measured by Euclidean distances. In Figure 4.7(c), there is a peak in Northern California and the tags are related to wine. Northern California is famous for its wine industry (Wine Country). Some other lower ranked clusters judged to be "not geographically relevant" also show some geographical signal, such as the cluster displayed in Figure 4.4(b), which highlights zoos and airports.

Temporal clustering also found more temporally-relevant clusters than other techniques: 13 (26%) of the clusters produced by the temporal similarity metric were found to be temporally relevant, versus 5 (10%) of motion clusters and only 1 (2%) of the co-occurrence clusters and 6 (12%) of the mutual information clusters. We examined the temporal clusters judged not to be temporally relevant and found that some did have temporal patterns that were hard to observe when only the text tags were presented to users. We present three such clusters in Figure 4.8. The cluster in Figure 4.8(a) has bursts around important dates for the Presidential election. The cluster in Figure 4.8(b) has an autumn pattern and includes mostly insects and plants that are active or increasing in autumn. Finally, the cluster in Figure 4.8(c) has a January and February pattern, with tags related to the Chinese New Year and the Super Bowl. Some other such examples can be found in Figure 4.5(b) and Figure 4.5(c) which were also judged to be "not temporally relevant".

In all of these cases, the visualizations of geographical and temporal clusters were helpful for us to discover the hidden semantics behind the tag clusters and the results from the control group where the users were presented with these visualizations verified the expected improved understanding of the semantics. The results only differed in a way that some of the clusters that were previously judged to be "not geographically relevant" or "not temporally relevant" were judge to be "geographically relevant" or "temporally relevant". For geographical clusters, 2 more high-ranked clusters (ranked 11 and 31 by second moment) mentioned above and visualized in Figure 4.7(a) and Figure 4.7(c) were judged to be "geographically relevant", which gave in total 62% of the 50 clusters judged to be "geographically relevant". For temporal clusters, 6 more clusters (ranked in

top 21) were judged to be "temporally relevant", which gave in total 38% of the clusters judged to be "temporally relevant" and 18 out the 22 top ranked clustered were "temporally relevant". Clusters displayed in Figures 4.8(a) and (b) and Figure 4.5(b) and (c) were examples of the 6 clusters. Though the cluster in Figure 4.8(c) was still judged to be "not temporally relevant", 65% of the users judged it to be "temporally relevant", comparing with the previous 40%. On average, for each geographical cluster, 66.7% of the users judged it to be "geographically relevant", comparing with the previous 64.4%; for each temporal cluster, 56.9% of the users judged it to be "temporally relevant", comparing with the previous 49.7%.

Motion clustering found both geographically and temporally relevant clusters. However, no motion clusters were judged to be both geographically *and* temporally relevant. Mutual information clustering and co-occurrence clustering found the same number of geographically relevant clusters, but mutual information clustering found 5 more temporally relevant clusters. Using this measurement, mutual information clustering performed better in finding temporally relevant clusters than co-occurrence clustering.

A subset of the evaluation results are shown in more detail in Figure 4.6. The figure shows the top 10 clusters produced by the geospatial, temporal, motion, co-occurrence, and mutual information analyses, and indicates which of these clusters were found to be geographically or temporally significant by the panel of human judges without visualizations being shown.

## GEOGRAPHICALLY AND TEMPORALLY RELEVANT CLUSTER RETRIEVAL

We observed that the average second moment of the geospatial, temporal, and motion clusters appears to be a good indicator of whether a cluster will be judged to be geographically or temporally relevant. We quantified this relevance by studying a retrieval problem, in which the task is to find relevant clusters using different average second moment thresholds. Clusters are considered to be retrieved if their average second moment is equal to or above a certain threshold. We can then sum-

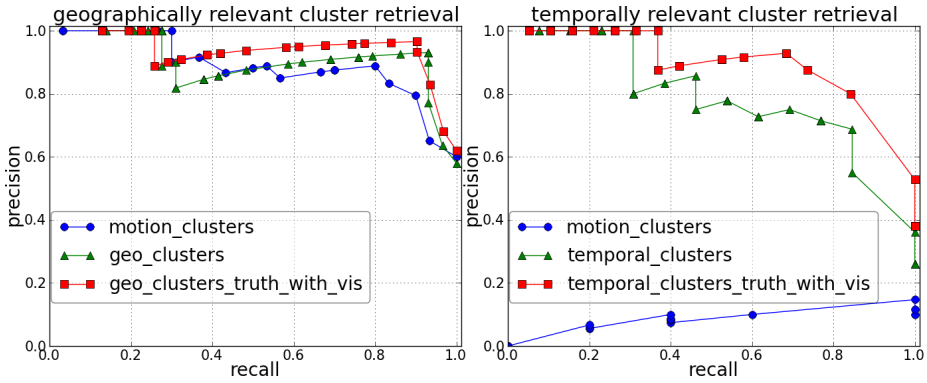



Figure 4.9: Precision-recall curves for retrieving geographically (left) and temporally (right) relevant clusters.

marize the results in terms of standard precision and recall statistics. The precision and recall for geographically relevant cluster retrieval is computed as:

$$\text{precision} = \frac{|R \cap G|}{|R|} \quad \text{recall} = \frac{|R \cap G|}{|G|} \tag{4.10}$$

where $R$ is the set of retrieved clusters and $G$ is the set of clusters judged to be geographically relevant. The precision and recall for temporally relevant cluster retrieval are computed in a similar way.

Figure 4.9(left) shows the precision-recall curves for retrieving geographically relevant clusters, in which the average second moment threshold decreases from left to right on each curve. For example, for geographical clusters in geographically relevant cluster retrieval, when the average second moment threshold is 0.04, both the precision and recall are 93.1%. Motion clusters performed slightly worse at high recalls. When geographical relevancy was judged by users shown the visualizations for geographical clusters, the retrieval has the best performance. When the recall is 90.3%, the precision reaches 96.6%.

Figure 4.9(right) shows the precision-recall curves for retrieving temporally relevant clusters. The precisions and recalls for temporal clusters are worse than geographically relevant cluster retrieval. When the average second moment threshold is 0.07, the precision is 71.4% and recall is

76.9%. Motion clusters performed much worse, because (as we discussed above) for motion clusters the average second moment does not have strong correlation with temporal relevance. In the ground truth, only 5 clusters were judged to be temporally relevant and their average second moment ranks ranged from 13 to 33. As future work, it would be interesting to study alternative statistics other than second moment that may perform better for motion clusters. Similarly with the geographically relevant cluster retrieval, when relevancy was judged by users shown the visualizations, the retrieval improves for temporal clusters. When the recall is 68.4%, the precision reaches 92.9%.

## 4.3 CONCLUSION

In this project, we proposed techniques to measure the semantic similarity of tags by comparing geospatial, temporal, and geo-temporal patterns of use. We used these techniques on a large dataset from Flickr to cluster tags using geo-temporal distributions and proposed novel methods to visualize the resulting clusters. An evaluation and case study showed the overall high quality of the semantics mined by our approach, and that the second moment served as a simple filtering measurement that achieved promising performance in selecting geographically- and temporally-relevant clusters. A case study suggests that our visualizations of tag semantics can help people understand subtle geo-temporal relationships between tags. These show that the geo-temporal connections between tags reveal the connections between real world concepts.

There are many possible improvements and future directions for this research. Currently, we are using only North American data and clustering the top 2000 most used tags into 50 clusters. It would be interesting to apply our approach within a more flexible framework, deciding the number of tags and the number of clusters in an automatic way. As suggested in Section 3.1, bin sizes affect the geospatial distributions being captured. One extreme example is that New York City and Los Angeles both fall into the North America bin at a continental granularity while they are in different bins at a state granularity. The hierarchical binning described in Section 3.1 allows flexible bin

sizes as well as hierarchical and efficient indexing of the bins. We believe it can address the bin size issue. Besides, it would also be interesting to build a tag recommendation system that integrates our techniques, using multiple kinds of tag similarity metrics to improve results and give corresponding visualizations to the user. Finally, our approach could be applied to other collections of objects with geographical and temporal attributes, such as data from Wikipedia or Twitter.

# CHAPTER 5

# REAL WORLD PHENOMENON DETECTION

As discussed in Chapter 1, latent in user-sensed datasets are observations of the world. This was further demonstrated in Chapter 4, where the latent helped us reveal connections between real world concepts and show that the virtual world reflects the real world. Aggregating these observations across millions of social sharing users could lead to new techniques for large-scale monitoring of the state of the world and how it is changing over time. However, the estimation techniques sometimes cannot be well evaluated without large-scale fine grained ground truth data. We want to design new techniques for monitoring the world which can also be well evaluated.

In this project we step towards that goal, showing that by analyzing the tags and image features of geo-tagged, time-stamped photos we can measure and quantify the occurrence of ecological phenomena including ground snow cover, snow fall and vegetation density as motivated in Section 1.4.2. We compare several techniques for dealing with the large degree of noise in the dataset, and show how machine learning can be used to reduce errors caused by misleading tags and ambiguous visual content. We evaluate the accuracy of these techniques by comparing to ground truth data collected both by surface stations and by Earthobserving satellites. Besides the immediate application to ecology, our study gives insight into how to accurately crowd-source other types of information from large, noisy social sharing datasets.

## 5.1 OUR APPROACH

We use a sample of nearly 150 million geo-tagged, timestamped Flickr photos as our source of user-contributed observational data about the world. We collected this data using the public Flickr API, by repeatedly searching for photos within random time periods and geospatial regions, until the entire globe and all days between January 1, 2007 and December 31, 2010 had been covered. We applied filters to remove blatantly inaccurate metadata, in particular removing photos with geo-tag precision less than about city-scale (as reported by Flickr), and photos whose upload timestamp is the same as the EXIF camera timestamp (which usually means that the camera timestamp was missing). Other information about this data can be found in Section 1.3.

For ground truth we use large-scale data originating from two independent sources: ground-based weather stations, and aerial observations from satellites. For the ground-based observations, we use publicly-available daily snowfall and snow depth observations from the U.S. National Oceanic and Atmospheric Administration (NOAA) Global Climate Observing System Surface Network (GSN) [1]. This data provides highly accurate daily data, but only at sites that have surface observing stations. For denser, more global coverage, we also use data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard NASA's Terra satellite. The satellite is in a polar orbit so that it scans the entire surface of the earth every day. The MODIS instrument measures spectral emissions at various wavelengths, and then post-processing uses these measurements to estimate ground cover. In this project we use two datasets: the daily snow cover maps [46] and the two-week vegetation averages [66]. Both of these sets of data including an estimate of the percentage of snow or vegetation ground cover at each point on earth, along with a quality score indicating the confidence in the estimate. Low confidence is caused primarily by cloud cover (which changes the spectral emissions and prevents accurate ground cover from being estimated), but also by technical problems with the satellite. As an example, Figure 1.2 shows raw satellite snow data from one particular day.

### 5.1.1 ESTIMATION TECHNIQUES

Our goal is to estimate the presence or absence of a given ecological phenomenon (like a species of plant or flower, or a meteorological feature like snow) on a given day and at a given place, using only the geo-tagged, time-stamped photos from Flickr. One way of viewing this problem is that every time a user takes a photo of a phenomenon of interest, they are casting a "vote" that the phenomenon actually occurred in a given geospatial region. We could simply look for tags indicating the presence of a feature – i.e. count the number of photos with the tag "snow" – but as discussed in Section 1.2, sources of noise and bias make this task challenging, including:

— *Sparse sampling:* The geospatial distribution of photos is highly non-uniform. A lack of photos of a phenomenon in a region does not necessarily mean that it was not there.

— *Observer bias:* Social media users are younger and wealthier than average, and most live in North America and Europe.

— *Incorrect, incomplete and misleading tags:* Photographers may use incorrect or ambiguous tags — e.g. the tag "snow" may refer to a snowy owl or interference on a TV screen.

— *Measurement errors:* Geo-tags and timestamps are often incorrect (e.g. because people forget to set their camera clocks).

***A statistical test.*** As described in Section 3.4, we apply the simple probabilistic model as well as the statistical test derived from it which can deal with some such sources of noise and bias. We adapt the general model and use the particular case of snow here. Any given photo either contains evidence of snow (event $s$) or does not contain evidence of snow (event $\bar{s}$). We assume that a given photo taken at a time and place with snow has a fixed probability $P(s|snow)$ of containing evidence of snow; this probability is less than 1.0 because many photos are taken indoors, and outdoor photos might be composed in such a way that no snow is visible. We also assume that photos taken at a time and place without snow have some non-zero probability $P(s|\overline{snow})$ of containing evidence of snow; this incorporates various scenarios including incorrect timestamps or geo-tags and misleading

visual evidence (e.g. man-made snow).

Here $m$ is the number of snow photos (event $s$) and $n$ is the number of non-snow photos (event $\bar{s}$) taken at a place and time of interest. With the deriving steps detailed in Section 3.4, we have the ratio that can be thought of as a measure of the confidence that a given time and place actually had snow, given photos from Flickr:

$$\frac{P(snow|s^m, \bar{s}^n)}{P(\overline{snow}|s^m, \bar{s}^n)} = \frac{P(snow)}{P(\overline{snow})} \left(\frac{p}{q}\right)^m \left(\frac{1-p}{1-q}\right)^n. \tag{5.1}$$

A simple way of classifying a photo into a positive event $s$ or a negative event $\bar{s}$ is to use text tags. We identify a hand-selected set $\mathcal{S}$ of tags related to a phenomenon of interest. Any photo tagged with at least one tag in $\mathcal{S}$ is declared to be a positive event $s$, and otherwise it is considered a negative event $\bar{s}$. For the snow detection task, we use the set $\mathcal{S}$=$\{$snow, snowy, snowing, snowstorm$\}$, which we selected by hand.

The above derivation assumes that photos are taken independently of one another, which is generally not true in reality. One particular source of dependency is that photos from the same user are highly correlated with one another. To mitigate this problem, instead of counting $m$ and $n$ as numbers of *photos*, we instead let $m$ be the number of *photographers* having at least one photo with evidence of snow, while $n$ is the numbers of photographers who did not upload any photos with evidence of snow.

The probability parameters in the likelihood ratio of equation (5.1) can be directly estimated from training data and ground truth. For example, for the snow cover results presented in Section 5.2, the learned parameters are: $p = p(s|snow) = 17.12\%$, $q = p(s|\overline{snow}) = 0.14\%$. In other words, almost 1 of 5 people at a snowy place take a photo containing snow, whereas about 1 in 700 people take a photo containing evidence of snow at a non-snowy place.

Figure 1.2 shows a visualization of the likelihood ratio values for the U.S. on one particular day using this simple technique with $\mathcal{S}$=$\{$snow, snowy, snowing, snowstorm$\}$. High likelihood ratio

values are plotted in green, indicating a high confidence of snow in a geospatial bin, while low values are shown in blue and indicate high confidence of no snow. Black areas indicate a likelihood ratio near 1, showing little conference either way, and grey areas lack data entirely (having no Flickr photos in that bin on that day).

### 5.1.2 LEARNING FEATURES AUTOMATICALLY

The confidence score in the last section has a number of limitations, including requiring that a set of tags related to the phenomenon of interest be selected by hand. Moreover, it makes no attempt to incorporate visual evidence or negative textual evidence — e.g., that a photo tagged "snowy owl" probably contains a bird and no actual snow. We use machine learning techniques to address these weaknesses, both to automatically identify specific tags and tag combinations that are correlated with the presence of a phenomenon of interest, and to incorporate visual evidence into the prediction techniques.

***Learning tags.*** We consider two learning paradigms. The first is to produce a single exemplar for each bin in time and space consisting of the set of all tags used by all users. For each of these exemplars, the NASA and/or NOAA ground truth data gives a label (snow or non-snow). We then use standard machine learning algorithms like Support Vector Machines and decision trees to identify the most discriminative tags and tag combinations. In the second paradigm, our goal instead is to classify individual *photos* as containing snow or not, and then use these classifier outputs to compute the number of positive and non-positive photos in each bin (i.e., to compute $m$ and $n$ in the likelihood ratio described in the last section).

***Learning visual features.*** We also wish to incorporate visual evidence from the photos themselves. There is decades of work in the computer vision community on object and scene classification (see [107] for a recent survey), although most of that work has not considered the large, noisy photo collections we work with here. We tried a number of approaches, and found that a classifier using a

simplified version of GIST augmented with color features [48, 108] gave a good trade-off between accuracy and tractability.

Given an image $I$, we partition the image into a $4 \times 4$ grid of 16 equally-sized rectangular regions. In each region we compute the average pixel values in each of the red, green, and blue color planes, and then convert this color triple from sRGB space to the CIELAB color space [51]. CIELAB has a number of advantages, including separating greyscale intensity from the color channels and having greater perceptual uniformity (so that Euclidean distances between two CIELAB color triples are approximately proportional to the human perception of difference between the colors). For each region $R$ we also compute the total gradient energy $E(R)$ within the grayscale plane $I_g$ of the image,

$$E(R) = \sum_{(x,y)\in R} ||\nabla I_g(x,y)|| \tag{5.2}$$

$$= \sum_{(x,y)\in R} \sqrt{I_x(x,y)^2 + I_y(x,y)^2}, \tag{5.3}$$

where $I_x(x,y)$ and $I_y(x,y)$ are the partial derivatives in the $x$ and $y$ directions evaluated at point $(x,y)$, approximated as,

$$I_x(x,y) = I_g(x+1,y) - I_g(x-1,y), \tag{5.4}$$

$$I_y(x,y) = I_g(x,y+1) - I_g(x,y-1). \tag{5.5}$$

For each image we concatenate the gradient energy in each of the 16 bins, followed by the 48 color features (average L, a, and b values for each of the 16 bins), to produce a 64-dimensional feature vector. We then learn a Support Vector Machine (SVM) classifier from a labeled training image set.

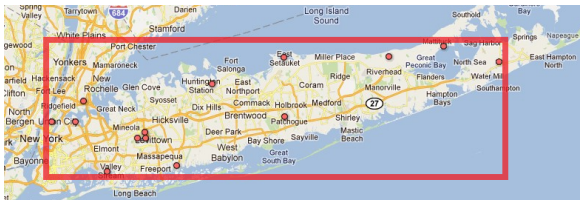
## 5.2 EXPERIMENTS AND RESULTS

We now turn to presenting experimental results for estimating the geo-temporal distributions of two ecological phenomena: snow and vegetation cover. In addition to the likelihood ratio-based score described in Section 5.2 and machine learning approaches, we also compare to two simpler techniques: *voting,* in which we simply count the number of users that use one of a set $S$ of tags related to the phenomenon of interest at a given time and place, and *percentage,* in which we calculate the ratio of users that use one of the tags in $S$ over the total number of users who took a photo in that place on that day.

### 5.2.1 SNOW PREDICTION IN CITIES

We first test how well the Flickr data can predict snowfall at a local level, and in particular for cities in which high-quality surface-based snowfall observations exist and for which photo density is high. We choose 4 U.S. metropolitan areas, New York City, Boston, Chicago and Philadelphia, and try to predict both daily snow presence as well as the quantity of snowfall. For each city, we define a corresponding geospatial bounding box and select the NOAA ground observation stations in that area. For example, Figure 5.1 shows the the stations and the bounding box for New York City. We calculate the ground truth daily snow quantity for a city as the average of the valid snowfall values from its stations. We call any day with a non-zero snowfall or snowcover to be a snow day, and any other day to be a non-snow day. Figure 5.1 also presents some basic statistics for these 4 cities. All of our experiments involve 4 years (1461 days) of data from January 2007 through December 2010; we reserve the first two years for training and validation, and the second two years for testing.

***Daily snow classification for 4 cities.*** Figure 5.2(a) presents ROC curves for this daily snow versus non-snow classification task on New York City. The figure compares the likelihood ratio confidence score from equation (5.1) to the baseline approaches (voting and percentage), using the tag set $\mathcal{S}$=$\{$snow, snowy, snowing, snowstorm$\}$. The area under the ROC curve (AUC) statistics are

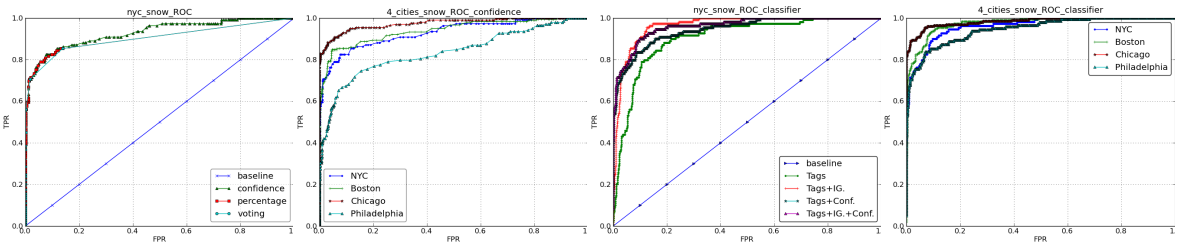



| | NYC | Chicago | Boston | Philadelphia |
|---|---|---|---|---|
| Mean active Flickr users / day | 65.6 | 94.9 | 59.7 | 43.7 |
| Approx. city area ($km^2$) | 3,712 | 11,584 | 11,456 | 9,472 |
| User density (avg users/unit area) | 112.4 | 52.5 | 33.5 | 29.6 |
| Mean daily snow (inches) | 0.28 | 0.82 | 0.70 | 0.35 |
| Snow days (snow¿0 inches) | 185 | 418 | 373 | 280 |
| Number of obs. stations | 14 | 20 | 41 | 26 |

Figure 5.1: *Top:* New York City geospatial bounding box used to select Flickr photos, and locations of NOAA observation stations. *Bottom:* Statistics about spatial area, photo density, and ground truth for each of the 4 cities.

0.929, 0.905, and 0.903 for confidence, percentage, and voting, respectively, and the improvement of the confidence method is statistically significant with $p = 0.0713$ according to the statistical test of [111]. The confidence method also outperforms other methods for the other three cities (not shown due to space constraints). ROC curves for all 4 cities using the likelihood scores are shown in Figure 5.2(b). Chicago has the best performance and Philadelphia has the worst; a possible explanation is that Chicago has the most active Flickr users per day (94.9) while Philadelphia has the least (43.7).

These methods based on presence or absence of tags are simple and very fast, but they have a number of disadvantages, including that the tag set must be manually chosen and that negative correlations between tags and phenomena are not considered. We thus tried training a classifier to learn these relationships automatically. For each day in each city, we produce a single binary feature vector indicating whether or not a given tag was used on that day. We also tried a feature selection step by computing information gain and rejecting features below a threshold, as well as adding the likelihood score from equation (5.1) as an additional feature. For all experiments we used feature vectors from 2007 and 2008 for training and tested on data from 2009 and 2010, and used a LibLinear classifier with L2-regularized logistic regression [35]. Table 5.1 presents the results,




(a) (b) (c) (d)

Figure 5.2: ROC curves for binary snow predictions: (a) ROC curves for New York City, comparing likelihood ratio confidence score to voting and percentage approaches, (b) ROC curves for 4 cities using the likelihood scores, (c) ROC curves from SVM classifiers with different features for New York City, and (d) ROC curves for 4 cities using the logistic regression (LibLinear) classifier with tags, information gain and confidence features. (Best viewed in color.)

showing that information gain (IG) and confidence scores (Conf) improve the results for all cities, and that the classifier built with both IG and Conf generally outperforms other classifiers, except for Boston. Figure 5.2(c) shows ROC curves from different classifiers for NYC and Figure 5.2(d) compares ROC curves for the 4 cities using the classifier using both feature selection and confidence. Note that the machine learning-based techniques substantially outperform the simple likelihood ratio approach (compare Figures 5.2(b) and (d)).

Table 5.1: **Daily snow classification results for a 2 year period (2009-2010) for four major metropolitan areas.**

| Features | Accuracy | Precision | Recall | F-Measure | Baseline |
|---|---|---|---|---|---|
| **NYC** | | | | | |
| **Tags** | 0.859 | 0.851 | 0.859 | 0.805 | 0.85 |
| **Tags+Conf.** | 0.926 | 0.927 | 0.926 | 0.917 | 0.85 |
| **Tags+IG** | 0.91 | 0.906 | 0.91 | 0.898 | 0.85 |
| **Tags+IG+Conf.** | **0.93** | **0.93** | **0.93** | **0.923** | **0.85** |
| **Boston** | | | | | |
| **Tags** | 0.899 | 0.897 | 0.899 | 0.894 | 0.756 |
| **Tags+Conf.** | **0.93** | **0.929** | **0.93** | **0.929** | **0.756** |
| **Tags+IG** | 0.91 | 0.911 | 0.91 | 0.91 | 0.756 |
| **Tags+IG+Conf.** | 0.923 | 0.923 | 0.923 | 0.923 | 0.756 |
| **Chicago** | | | | | |
| **Tags** | 0.937 | 0.938 | 0.937 | 0.935 | 0.728 |
| **Tags+Conf.** | 0.949 | 0.952 | 0.949 | 0.948 | 0.728 |
| **Tags+IG** | 0.938 | 0.938 | 0.938 | 0.938 | 0.728 |
| **Tags+IG+Conf.** | **0.953** | **0.954** | **0.953** | **0.953** | **0.728** |
| **Philadelphia** | | | | | |
| **Tags** | 0.849 | 0.851 | 0.849 | 0.815 | 0.805 |
| **Tags+Conf.** | 0.912 | 0.917 | 0.912 | 0.903 | 0.805 |
| **Tags+IG** | 0.903 | 0.899 | 0.903 | 0.897 | 0.805 |
| **Tags+IG+Conf.** | **0.927** | **0.926** | **0.927** | **0.924** | **0.805** |

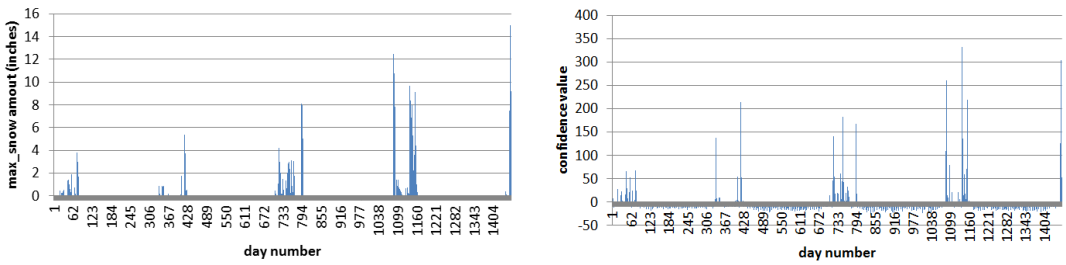



Figure 5.3: Time series of actual daily snow (left) and score estimated from Flickr (right) for New York City, 2007-2010.

***Predicting snow quantities.*** In addition to predicting simple presence or absence of a phenomenon, it may be possible to predict the degree or quantity of that phenomenon. Here we try one particular approach, using our observation that the numerical likelihood score of equation (5.1) is somewhat correlated with depth of snow ($R^2$=0.2972) — i.e., that people take more photos of more severe storms (see Figure 5.3).

Because snow cover is temporally correlated, we fit a multiple linear regression model in which the confidence scores of the last several days are incorporated. The prediction on day $t$ is then given by,

$$\begin{cases} \sum_{i=0}^{T} \alpha_i \log(conf_{t-i}) + \beta & \text{if } conf_t \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

where $conf_t$ represents the likelihood ratio from equation (5.1) on day $t$, $T$ is the size of the temporal window, and the $\alpha$ and $\beta$ parameters are learned from the training data. We found that increasing the spatial threshold $T$ generally improves performance on the 4 cities, but that no additional improvement occurred with $T > 3$. We can measure the error of our predictions with the root-mean-squared error between the time series of our predictions and the actual snow data (following [53]). We achieve an RMS error of between about 1 and 1.5 inches across the 4 cities; Philadelphia has the largest error (1.44), followed by Boston (1.26), New York (1.15), and Chicago (1.06). As an example, Figure 5.4 presents a visual comparison of the prediction time series versus the actual snow

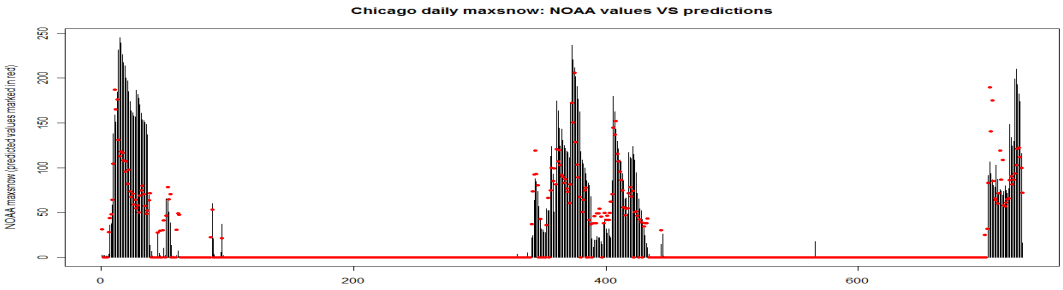



Figure 5.4: Actual daily amount of snow compared to prediction, for Chicago and $T = 3$. Predicted values are shown as red dots, and actual values are vertical lines.

time series for Chicago.

An alternative way of evaluating the snow quantity estimates is to view it as a multi-way classification tasks. We follow an existing snowfall impact scale [103] and quantize daily snow quantity into 7 buckets: no snow, 0-1 inches, 1-4 inches, 4-10 inches, 10-20 inches, 20-30 inches, or more than 30 inches. We then build a classification model to predict the snow ranges for the four cities using number of snow users. We also include the numbers of snow users from the previous three days as extra features. We use Naive Bayesian classifier [54] (which performed best on this task). These multi-way classification results are better than a majority class baseline, with 7-way correct classification rates at 87.5% for Philadelphia, 87.9% for New York, 84.0% for Boston, and 83.7% for Chicago (versus baselines of 80.5%, 85.1%, 75.6%, and 72.9%, respectively).

### 5.2.2 CONTINENTAL-SCALE SNOW PREDICTION

Predicting snow for individual cities is of limited practical use because accurate meteorological data already exists for these highly populated areas. In this section we ask whether phenomena can be monitored at a continental scale, a task for which existing data sources are less complete and accurate. We use the photo data and ground truth described in Section 5.2, although for the experiments presented here we restrict our dataset to North America (which we defined to be a rectangular region spanning from 10 degrees north, -130 degrees west to 70 degrees north, -50

degrees west). (We did this because Flickr is a dominant photo-sharing site in North America, while other regions have other popular sites — e.g. Fotolog in Latin America and Renren in China.)

The spatial resolution of the NASA satellite ground truth datasets is 0.05 degrees latitude by 0.05 degrees longitude, or about $5 \times 5km^2$ at the equator. (Note that the surface area of these bins is non-uniform because lines of longitude get closer together near the poles.) However, because the number of photos uploaded to Flickr on any particular day and at any given spatial location is relatively low, and because of imprecision in Flickr geo-tags, we produce estimates at a coarser resolution of 1 degree square as suggested in Section 3.1, or roughly $100 \times 100km^2$. To make the NASA maps comparable, we downsample them to this same resolution by averaging the high confidence observations within the coarser bin. We then threshold the confidence and snow cover percentages to annotate each bin with one of three ground truth labels:

— Snow bin, if confidence is above 90 and coverage above 80,

— Non-snow bin, if confidence is above 90 and coverage is 0,

— Unknown bin, otherwise.

Our goal is to predict whether or not each geospatial bin had snowcover on each day, given the photos from Flickr.

***Retrieving snow or non-snow bins.*** In many real applications, ecologists would be satisfied in finding bins for which the phenomenon is present, rather than actually classifying all bins. It is thus useful to view this problem as a retrieval task, in which the goal is to identify bins likely to contain the phenomenon, or likely not to contain it. We thus turn to evaluating the performance of our estimation techniques using precision-recall curves with the same notations as in equation 4.10,

$$\text{precision} = \frac{|R \cap G|}{|R|} \qquad \text{recall} = \frac{|R \cap G|}{|G|},$$

where $R$ is the set of retrieved bins and $G$ is the set of correct bins according to the ground truth. Precision-recall curves are also easier to interpret in situations where the classification baselines are

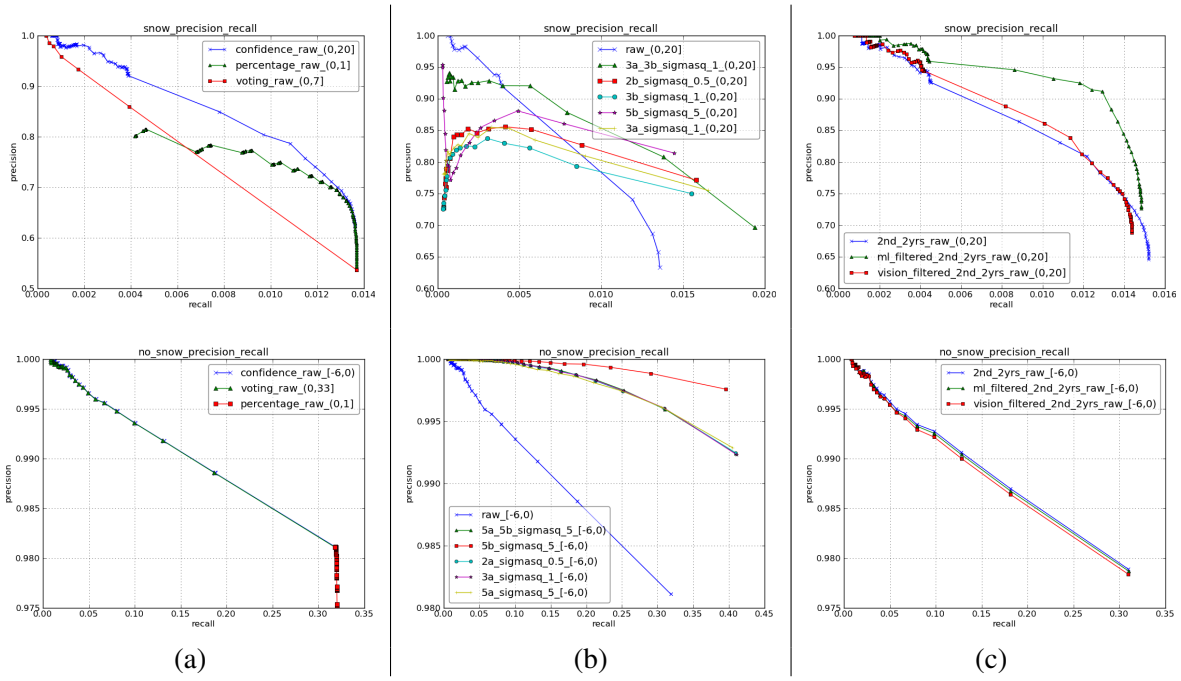



Figure 5.5: Precision and recall curves for retrieving snow (top) and non-snow (bottom) instances, where an instance is a single geospatial bin on a single day, using different techniques: *(a)* comparing the voting, percentage, and statistical confidence estimation techniques, *(b)* comparing different temporal smoothing strategies, *(c)* using classifiers to reject falsely-tagged snow images using visual and textual features.

so high, as in our case.

Figure 5.5(a) shows precision-recall curves for retrieving bins and days containing snow (top) and those not containing snow (bottom). In total, these curves involve classifying 701,280 exemplars (each of which is a single geospatial bin on a single day), of which 11.0% have ground truth. 82.2% of the bins with ground truth are no-snow bins, while snow bins account for 17.8%. We observe that the confidence method performs significantly better than the other two methods for retrieving snow bins, achieving about 98% precision at 0.2% recall, and about 80% precision at 1% recall. For retrieving non-snow bins the three techniques are almost the same, and all three perform better than the random baseline.

While the precisions in these curves are high, the recall values are alarming low. The main reason for this is that large areas of North America, particularly most of Canada and Alaska, have

sparse populations resulting in a very limited number of photos uploaded in these areas. We showed in the last section that accurate snow estimates can be inferred for highly populated cities; the low recalls here are because of low photographic density in much of the continent. Restricting to specific subsets significantly increases the density of observations: for example, the average number of photos per bin over our four years of data is nearly ten times larger for the northeast US compared to all of North America (70,398 vs 8,134). The performance is significantly better in these more densely populated areas; for example, in the Northeast US the precision is 96.3% at a recall of 19.5% for snow retrieval, and 99.9% precision at 9.1% recall for non-snow retrieval. Moreover, recall would naturally improve as our dataset grows; our sample of 150 million images is less than 3% of the photos on Flickr, and thus the recall would improve significantly if we had access to the entire dataset.

***Temporal smoothing.*** For many phenomena (including snow), the existence of an event on one day is strongly correlated with its existence on the next day. Thus one way of addressing the sparsity of Flickr photos in some locations is to propagate evidence forward and backward in time. To do this, we apply a Gaussian filter on the Flickr confidence values for each bin in an attempt to achieve better recalls. We vary the degree of smoothing by using Gaussians with different variance values. We tried smoothing with many different parameters, including smoothing both forward and backwards in time, or in only one direction. Figure 5.5(b) shows curves for several of the best combinations that we found, including the raw confidence score (blue X's), 3 days before and after with variance 1.0 (brown triangles), 2 days before with variance 0.5 (red squares), 3 days before with variance 1.0 (blue circles), 5 days before with variance 5.0 (purple stars), and 3 days after with variance 1.0 (yellow +'s). We find that temporal smoothing three days before and after with variance 1.0 significantly improves performance for both snow and non-snow retrieval, increasing snow retrieval precision by about 7 percentage points at 1% recall.

***Voting.*** Voting performs worse than the statistical confidence given by the Bayesian likelihood

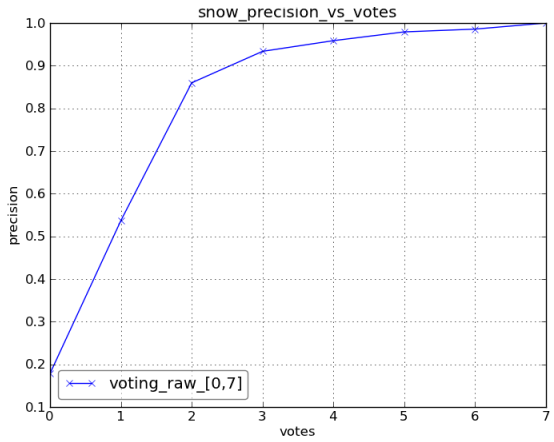



Figure 5.6: Precision vs number of votes for snow predictions using the voting method.

ratio, but it is an interesting technique to study in more detail because of its simplicity. Voting simply counts the number of users who have annotated at least one photo in a given bin and day with a snow-related tag. Figure 5.6 plots precision versus the number of votes for snow retrieval. The shape of these curve illustrates why crowd-sourced observations of the world can be reliable, if enough people are involved: as the number of votes for snow increases, it becomes progressively less likely that these independent observations are coincidental, and more likely that they are caused by the presence or absence of an actual phenomenon. It is interesting to notice that when there are 7 or more snow voters, snow prediction precision becomes 100%, while the same is true for non-snow prediction when the number of non-snow voters reaches 33 if there are no snow voters in the bin.

***Case study of false positives.*** To understand the failure modes of estimating attributes about the world from Flickr photos, we performed a case study of false positives — bins and days in which our Flickr mining predicted the presence of snow, but the NASA ground truth indicated that there was no snow cover. In particular, we studied snow false positives at the operating point at which the likelihood ratio method gives a precision of 74.1% and a recall of 1.2% (i.e. when the threshold is 4). At this operating point, 34,323 total predictions are made (each corresponding to a single geospatial bin on a single day), 2,208 of which have valid ground truth. Of these 2,208 bins, 1,636

Table 5.2: **Taxonomy of manually-labeled false-positive photos (which have at least one snow-related tag despite being taken at a snowless time and place according to the ground truth).**

| Class | Description | # Photos | Percentage |
|---|---|---|---|
| little or distant | photos with trace amount of snow or snow in the distance | 585 | 33.0% |
| man made | photos with snow made by humans, (e.g. at a ski slope) | 152 | 8.6% |
| no snow | photos without snow | 737 | 41.5% |
| snow | photos with significant snow | 279 | 15.7% |
| not sure | other photos | 21 | 1.2% |

(74.1%) are correctly classified, while the 572 false positive bins have a total of 1,855 photos tagged with one of the snow terms (despite the fact that they were taken at places and times in which the NASA satellite did not record snow). We manually examined these 1,855 false positive photos and classified them into 5 different classes according to their visual content, as shown in Table 5.2. Nearly 60% of these photos do actually appear to contain some snow; of these, 33% either show trace amount of snow or snow in the distance (usually on a distant mountain peak), and 8.6% have man-made snow that would not show up on the NASA maps (like in a zoo or ski slope), while only about 16% include a significant amount of natural snow. About 40% of the photos tagged with a snow-related term do not appear to contain any snow at all; these are caused by mis-tagged images or snow-related tags that are used to describe something else (like the interference on a TV screen). Figure 5.7 shows some sample false positives from each class.

(a) (b) (c) (d)

Figure 5.7: Sample photos that were not taken at a place and time with snow according to the ground truth, but that were uploaded with a snow-related tag: (a) photo with trace amounts of snow, (b) photo with distant snow, (c) photo with man made snow, and (d) photo with no snow (but with a "snowy egret").

For images that seem to contain natural snow, there are several possible explanations for why the ground truth does not indicate snow cover at that time and place. One is that the satellite passes over

at an unknown time of day, so it is possible that snowfall occurred after the satellite's observation was taken. Another cause are photos with incorrect time stamps or geo-locations; we assume that such errors occur frequently, although it is hard to quantify the frequency just by looking at the photos. Other photos clearly contain snow, but the amount is so little that it might not be visible from the satellite (e.g. Figure 5.7(a)), or the snow is so far in the distance that it is in a different geospatial bin (e.g. Figure 5.7(b)).

There are some cases where the Flickr evidence for snow is overwhelming, but the NASA ground truth does not indicate snow. This could be caused by the timing issue described above, or by satellite resolution and confidence issues. For example, on February 21, 2008, 5 Flickr users reported snowfall in New York. This bin is marked as a no-snow bin in the ground truth because the vast majority of it has zero snow coverage according to the satellite, but there is a small area within the bin that has low confidence (due to cloud cover) and probably corresponds to a snow squall.

***Machine learning for tag selection.*** Many of the above error modes can be addressed by training classifiers on textual tag and visual images features. As discussed in Section 5.1.1, we are interested in two learning paradigms: the first is to learn combinations of tags that classify geospatial bins well according to the NASA ground truth, while the second task is to reduce false positives by rejecting photos that are tagged with a snow term but do not actually contain snow.

In the first task, we want to learn to classify whether a given bin contains snow on a given day, based on a binary feature vector encoding the set of tags used by all users in that bin on that day. We tried four different classifiers to address this problem: REPTree, a fast decision tree learner which builds a decision tree using information gain and variance and prunes it using reduced-error pruning [47], Support Vector Machines (SVMs) [16], Discriminative Multinomial Naive Bayes (DMNB) [104] and LibLinear classifier with L2-regularized logistic regression [35]. To reduce the large number of features (a total of 404324 tags), we compute information gain and keep all features (13442 tags) with information gain greater than zero. Figure 5.8 presents ROC curves for this task,

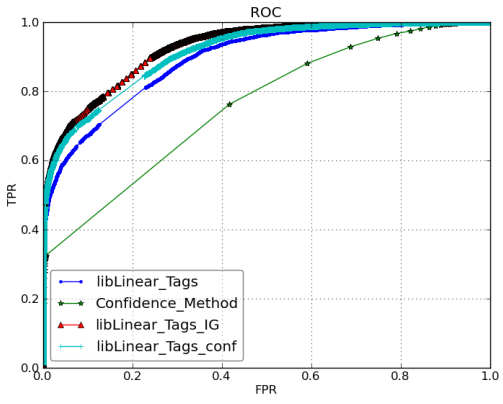



Figure 5.8: ROC curves for classifying whether a geo-bin has snow on a given day, comparing the LibLinear classifier with various tag features to the confidence method using hand-selected tags.

showing that the learned classifier outperforms the likelihood ratio from equation (5.1), and that feature selection with information gain and using the confidence ratio as an additional feature all improve performance.

Next we try the second learning paradigm, in which our goal is to examine photos that have a snow-related tag, and use the other tags as well as visual features to decide whether or not they actually contain snow. For example, the classifier might learn that a photo with "snowy" should be discarded if it also contains the tag "egret," since that photo is likely of a bird and not of actual snow. For training these classifiers, we had a human judge evaluate 1,855 images and to annotate them as to whether or not they actually contain evidence of snow.

We used decision trees for this task because it is easy to understand and interpret what features the classifier is using. In initial experimentation, we found that many of the most discriminative features were place names, like "sandiego" or "canada." These geographic tags are understandably strongly correlated with snowfall, but we would like our classifier to base its decisions on the content of an image (because, for example, climate change might cause snowfall in San Diego some day, and we would like our classifier to be able to detect this). To avoid selecting these tags, we first divide North America into four regions (northeast, northwest, southeast, southwest) and get the

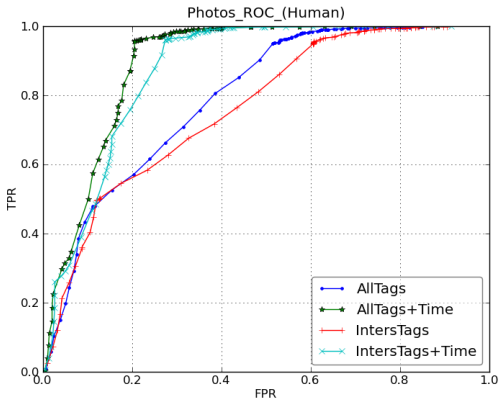



Figure 5.9: ROC curve for classifying whether photos contain snow, using decision trees with various features: AllTags includes all tags, IntersTags excludes tags corresponding to specific geographic areas, and AllTags+Time and IntersTags+Time include the month of the year as an additional feature.

intersection of the sets of tags used in these four regions. We then use only this set of intersected tags ("IntersTags") for building the decision tree. Besides tags, we also tried including the photo's timestamp month as an additional feature.

ROC curves are presented in Figure 5.9. We see that the time feature helps in improve the results, as does using all tags instead of just the spatially-intersected ones. The baseline (majority class) is 86.3%. It is interesting to examine the top few levels of the trained decision tree, to get a sense for which tags are most discriminative. The top decision node is "summer:" if this tag is present, then the photo is classified as not snow. If summer is not present, then the next few layers look at tags like "mountain," "clouds," "ski," "geese," and "egret."

***Machine learning to suppress false positives.*** Finally, we consider using the photo classifier as a filter while computing the likelihood ratios of Section 5.1.1, in order to reject photos that are marked with a snow tag but do not contain snow, using both visual and textual features. For the textual features, we use the decision tree classifier just described. For visual features, we trained an SVM using the GIST-like visual features described in Section 5.1.1, on the same hand-labeled dataset of about 2,000 images explained above. As with all other experiments, the training and testing sets

were kept separate by training on data from 2007-2008 and testing on data from 2009-2010. For the photos in these latter two years, we use our decision tree to try to filter out false positives (photos tagged "snow" but not containing snow), and then re-compute the likelihood ratio confidence score. We find that using a classifier to reject false positives based on tags increased precision by nearly 10 percentage points, as shown in Figure 5.5(f): at 1% recall, precision increased from about 84% to to about 93% for snow retrieval. For the visual features, we find a significant but more modest improvement, from about 84% to 86% at this level of recall.

### 5.2.3 ESTIMATING VEGETATION COVER

Another important measure of the ecological state of the planet is vegetation cover. We perform greenery versus no greenery predictions similarly to snow and no snow predictions using the Flickr confidence threshold method discussed in Section 5.1.1. As with snow, the ground truth is obtained from down-sampling and thresholding the NASA MODIS greenery data which has the same resolution as the snow cover data with similar coverage and quality (confidence) values. The Flickr greenery confidence values of bins are obtained in a similar way as with snow, except that we use a different set of target tags, including "tree," "trees," "leaf," "leaves", and "grass." One important difference between the NASA greenery and snow datasets is that the greenery data is an average of daily observations spanning 16 days. Thus our goal is to predict the geospatial distribution of greenery for each 16-day period of the year.

We require a bin to have no less than 50% greenery coverage and above middle quality to be considered as a ground truth bin. We report experiments using two different definitions of non-green bins: those having less than 1% coverage, and those having less than 5% coverage. For the 50% and 1% threshold combination, 25.6% of the bins with ground truth are greenery bins, while for the 50% and 5% threshold combination, 15.8% of the bin with ground truth are greenery bins. As shown in Figure 5.10, both curves outperform a random baseline for greenery prediction, but the

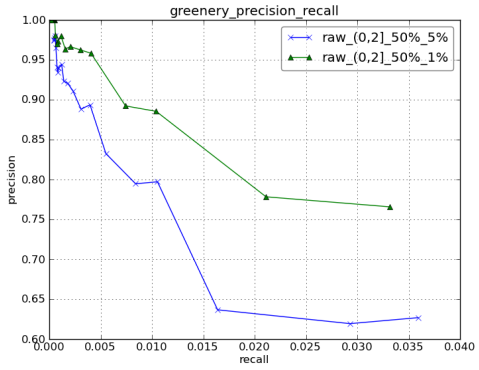



Figure 5.10: Greenery precision-recall curves using two different ground truth thresholds.

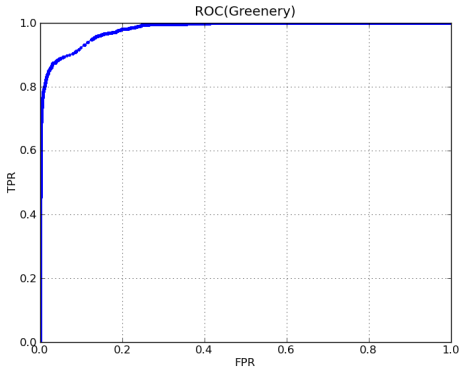



Figure 5.11: ROC curve for classifying greenery of bins, using tag features and LibLinear classifier.

estimates are not as accurate as those observed during in snow predictions. There seem to be several reasons for this drop in performance. One is that the boundary between greenery and no greenery seems more vague than the snow/no-snow boundary. Moreover, the greenery ground truth data has a much coarser temporal resolution (16 days). Finally, it's less clear which tags should be used to estimate greenery; using color analysis of the visual content of images may be a better approach, which we leave for future work. We also tried a learned classifier to predict greenery/non-greenery bins based on the set of tags used by all users in each bin and on each day. We used the LibLinear classifier [35] because it performed well in case of snow classification. Figure 5.11 presents the ROC curve for this classification task, showing an equal-error rate of about 91.6%.

## 5.3 CONCLUSION AND FUTURE WORK

In this project, we propose using the massive collections of user-generated photos uploaded to social sharing websites as a source of observational evidence about the world, and in particular as a way of estimating the presence of ecological phenomena. As a first step towards this long-term goal, we used a collection of 150 million geo-tagged, timestamped photos from Flickr to estimate snow cover and greenery, and compared these estimates to fine-grained ground truth collected by earth-observing satellites and ground stations. We compared several techniques for performing the estimation from noisy, biased data, including simple voting mechanisms and a Bayesian likelihood ratio. We also tested several possible improvements to these basic methods, including using temporal smoothing and machine learning to improve the accuracy of estimates. We found that while the recall is relatively low due to the sparsity of photos on any given day, the precision can be quite high, suggesting that mining from photo sharing websites could be a reliable source of observational data for ecological and other scientific research. In future work, we plan to study additional features including using more sophisticated computer vision techniques to analyze visual content. Also we plan to study a variety of other ecological phenomena, including those for which high quality ground truth is not available, such as migration patterns of wildlife and the distributions of blooming flowers. To solve or partially solve the data sparsity problem, we plan to incorporate data from multiple sources including: (1) other social media platforms such as Twitter, Instagram and Facebook, (2) surveillance camera datasets collected by researchers, such as the Archive of Many Outdoor Scenes (AMOS) dataset [52]. The former source provides data with comparable attributes to the Flickr dataset and may naturally contributes to a denser dataset. For example, tweets have timestamps and sometimes geolocations. Besides the textual content from which we can extract keywords or even obtain the keyword hashtag directly, for some tweets the images are also available. However, this may not solve the whole problem, as these platforms may also have biased user distributions. The latter source, which consists of web cameras located around the world may help fill in the gap in the

data from social media platforms. As most of these cameras are stable, they provide observations of fixed locations over time as opposed to photos from social media users that come and go. This consistency may help us get coherent estimates for these locations.

# CHAPTER 6

# MODELING MOBILE USERS

We studied the physical world in the last two chapters, by addressing two representative problems. We now study the human world as it is the users that generates all these user-sensed datasets and it is natural to try to understand the users and even benefit them. Many of these users are from mobile devices [36, 109] and we choose to study mobile users in this project. There have been studies characterizing the mobile users by finding the connections between them – e.g. users who have similar behaviors may share similar hobbies or interests. These user attributes are usually inferred using fine-grained and direct information such as time, location, app usage and web browsing history. This leads us to think whether the inference can still be accomplished with far less information. As motivated in Section 1.4.3, we ask this question: can we use simple and indirect behavioral statistics to model mobile users? We demonstrate that statistics such as entropy calculated from WiFi, Bluetooth, app usage, cell tower readings collected from user's phone, can be useful in uniquely identifying the user. With this descriptive power, personalized service can be built upon simple user behavioral statistics. We also discuss its privacy implications.

## 6.1 FEATURE COMPUTATION AND USER IDENTIFICATION

We focus on designing simple frequency and entropy-based statistical features from users' mobile footprints and evaluate the features' performance in user identification. We show the feature vectors have the descriptive power to captures the characteristics of a users that differentiate this user from other users.

### 6.1.1 FREQUENCY BASED FEATURES

The number of devices and cell towers that are observed by a phone throughout the day provides information about the owner's environment. For example, a phone in a busy public place will likely observe many wireless devices while a phone in a moving car observes different cell towers over time. Meanwhile, a user's app usage patterns throughout the day tell us something about his or her daily routine. We thus propose simple frequency-based features that measure how much activity of four different types (WiFi, Cell towers, Bluetooth, and App usage) is observed at different time intervals throughout the day. More specifically, we divide time into $T$-minute time periods, and make observations about the phone's state every $M$ minutes, with $M < T$ so that there are multiple observations per time period. In the $i$-th observation of time period $t$, we record: (1) the number of WiFi devices that are observed ($W^{t,i}$), (2) the number of cell phone towers that the phone is connected to ($C^{t,i}$), (3) the number of bluetooth devices that are seen ($B^{t,i}$), and the number of unique apps that have been used over the last $m$ minutes ($A^{t,i}$). We then aggregate each of these observation types to produce four features in each time period:

$$F_W^t = \sum_i W^{t,i}, \qquad F_C^t = \sum_i C^{t,i}, \qquad F_B^t = \sum_i B^{t,i}, \qquad \text{and} \qquad F_A^t = \sum_i A^{t,i}. \tag{6.1}$$

A feature incorporating all of these features is simply the vector $F^t = [F_W^t \ F_C^t \ F_B^t \ F_A^t]$.

### 6.1.2 ENTROPY BASED FEATURES

While the simple frequency features above give some insight into the environment of the phone, they ignore important evidence like the distribution of this activity. For example, in some environments a phone may see the same WiFi hotspot repeatedly through the day, while other environments may have an ever-changing set of nearby WiFi networks. To illustrate this, Figure 6.1 compares observed frequency versus anonymized device IDs for two users across each of the four observation types, for a period of 10 days. We can observe that User 2 is less active in WiFi and cell mobility compared to User 1, but has more Bluetooth encounters and uses more diverse apps.

We thus propose using entropy of these distributions as an additional feature of our user vectors. The entropy feature summarizes the distribution over device IDs in a coarse way. For WiFi, let $W_j^t$ denote the number of times we observe wifi hotspot $j$ during time period $t$. Then we define the WiFi entropy during time period $t$ as,

$$E_W^t = -\sum_j \frac{W_j^t}{F_W^t} \log \frac{W_j^t}{F_W^t}. \tag{6.2}$$

Entropy features for Cell towers, Bluetooth, and Apps ($E_C^t$, $E_B^t$, and $E_A^t$, respectively) are computed in the same way, and we define a multimodal entropy feature vector $E^t = [E_W^t \; E_C^t \; E_B^t \; E_A^t]$, which

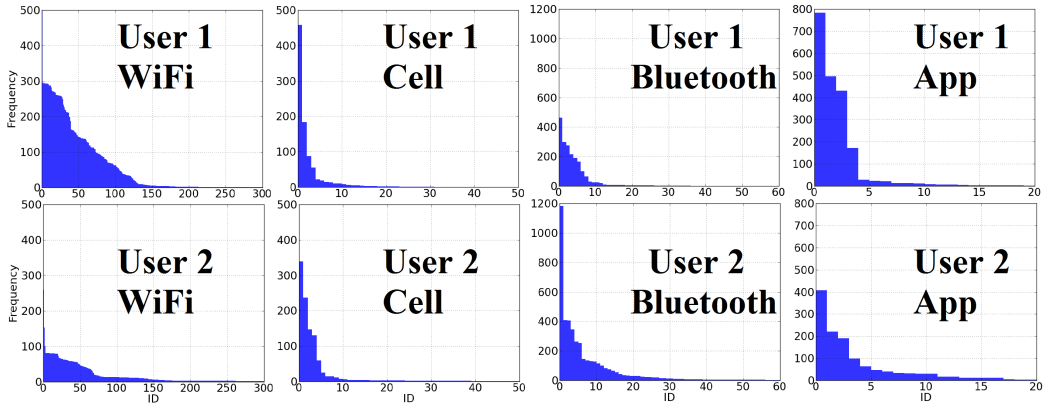



Figure 6.1: Comparison of activity histograms for 2 users over 10 days. Y-axes are frequencies; X-axes are WiFi, Cell, Bluetooth, and App IDs.

incorporates all four perspectives.

### 6.1.3 CONDITIONAL ENTROPY AND FREQUENCY BASED FEATURES

In our system, we calculate the above entropy and frequency features conditioned on time and location. Intuitively at different times and at different locations, users have different patterns of application usage and surrounding devices (Bluetooth, WiFi, cell, etc.). For example, two users might have similar overall apps usage but one user always uses apps in the mornings, while the other uses them only in the afternoon. Two other users might have similar overall Bluetooth entropies but one might have more surrounding devices at work while the other observes the variety at a coffeeshop. Conditioning on time and space is thus useful to better differentiate users.

**Conditional features on time.** For the frequencies and entropies conditioned on time, we differentiate on time of a day and day of a week. Currently we distinguish between three fixed daily time intervals, mornings (0:00 - 8:59), working hours (9:00 - 17:59) and evenings (18:00 - 23:59), and two types of days, weekdays (Mon through Fri) and weekends (Sat and Sun). This gives five time periods over which we compute the conditional features. Future work might explore adaptive intervals instead.

**Conditional features on location.** We also compute frequency and entropy features conditioned on location. For each user, we filter and cluster their geo-locations in order to identify the top-$k$ significant locations. From data collected at these $k$ locations, we compute the conditional entropies and frequencies. There are two steps in finding significant locations: *Segmentation* and *Clustering*. In the segmentation step, we find periods of time when the phone appears to be stationary, by looking for time intervals when the IDs of surrounding devices are stable. In particular, we divide the data streams into 10-minute time frames and for each time frame, we record the IDs of the WiFi, Bluetooth and Cell towers. For adjacent time frames, we compute Jaccard similarity of the corresponding sets of IDs, $J(S_1, S_2) = |S_1 \cap S_2|/|S_1 \cup S_2|$, where $S_1$ and $S_2$ are sets of

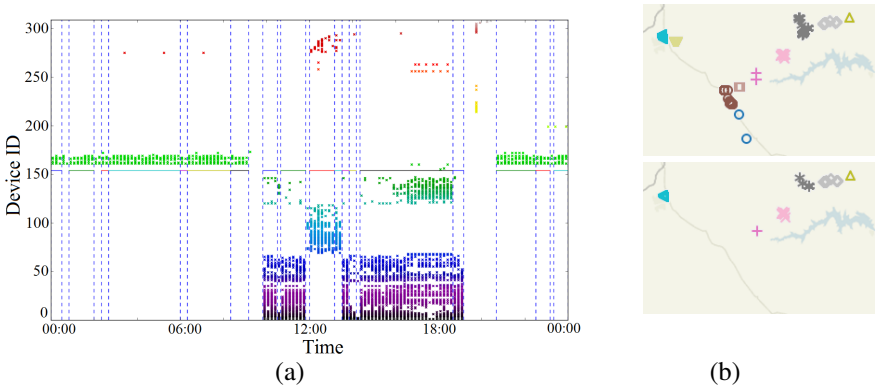



Figure 6.2: Time segmentation and location clustering: (a) Finding stationary times using WiFi. (b) Location clustering using all (top) and only stationary (bottom) data.

device IDs. If the similarity is larger than a threshold, we say that the device is stationary during the two time frames. Figure 6.2(a) illustrates finding stationary and non-stationary periods according to WiFi readings. Similarly, we perform this on Bluetooth and Cell readings and take the union of all stationary time periods.

We then apply the DBSCAN clustering algorithm to the stationary segments in order to identify important locations. Figure 6.2(b) shows the clustering results on the same data with and without non-stationary points. We see that noisy signals (e.g., location points moving along highways) have been removed by keeping only stationary data, which generates better and fewer clusters. We choose $k = 2$ and statistics from the $i$th location will be compared with that from the location of the same rank across users.

## 6.2 EVALUATION

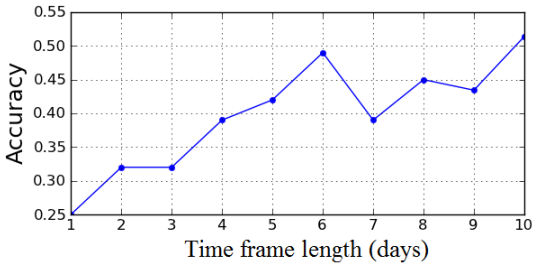
We define a user identification problem in order to test whether the features can uniquely characterize users. We pose this as a classification problem: *can we build a multi-class classifier trained on the entropy and frequency feature vectors labeled with user IDs, such that it tells which user a certain vector belongs to?*

### 6.2.1 DATA COLLECTION AND EXPERIMENTAL SETTINGS

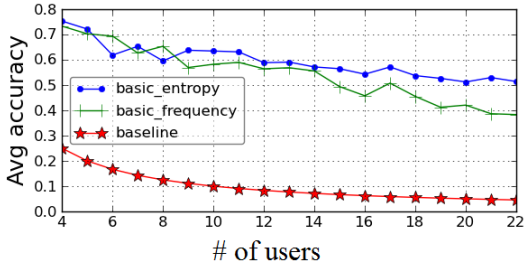
To collect data for our evaluation, we deployed an Android app called EasyTrack based on the Funf Open Sensing Framework [2]. This app has a customizable configuration with 17 data types, including WiFi, Bluetooth, cell tower, GPS, call log, app usage etc. We successfully recruited 22 users to install EasyTrack and collected their mobile footprints for about 2 months with some variation across users. Details about the data and its collection can be found in Section 1.3.

We first test the initial user identification performance using different time frame lengths. We uniformly sample the instances to make sure that the same number of instances is used to build the classifier for each time frame length. As shown in Figure 6.3a, when the length of time frame increases, the classification accuracy generally improves, despite possible variations caused by weekday/weekend patterns. This suggests that in this range, longer time frames better capture the uniqueness. Since the data collection time span is about two months, longer time frames decrease the total number of time periods and thus there are fewer features for training the classifiers. In the following experiments, we fix the time period at 10 days.

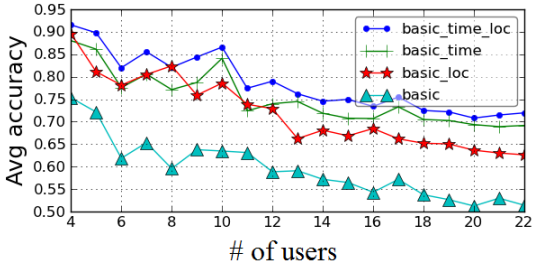
With a 10-day time frame, we reach 107 time frames in total from 22 users. On average, each user has 4.86 time frames. The range is [2, 8] and the standard deviation is 2.2. In total, we have 64 features. We test combinations of features (multi-modal entropies/frequencies, conditional entropies/frequencies) on multiple classifiers including Naive Bayes, decision tree, SVM and Multilayer Perceptron. We report the results from the Multilayer Perceptron, which were best. The learning rate is 0.3, momentum is 0.2, the number of hidden layers is set to $\frac{\#features+\#classes}{2}$ where $\#features$ is 64 and $\#classes$ is 22, which is the number of users. The number of epochs is set to 500. We use 10-fold cross validation for training and testing. We use accuracy as a performance measurement which is defined as $\frac{\#of\ correct\ predictions}{\#of\ instances}$.



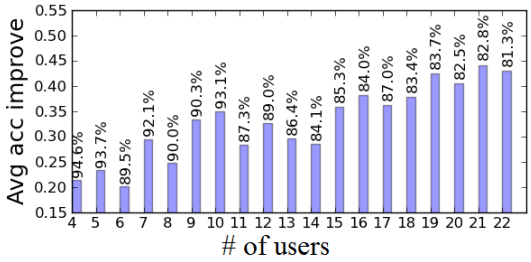

(a) Accuracy for 22 users with different time frame lengths, with the basic entropy features.




(b) Classification accuracy with different # of users for basic frequency and entropy features.




(c) Average classification accuracy with different number of users for different entropy feature combinations.




(d) Average accuracy improvement for all features, compared to baseline frequency. Absolute average accuracy is marked on the top of each bar.

Figure 6.3: User identification classification results.

### 6.2.2 USER IDENTIFICATION PERFORMANCE

We tested on varying numbers of users from 4 to 22, and observe that as the number of users increases, the average accuracy drops but is still significantly better than random guessing (see Figure 6.3b and Figure 6.3c). In these experiments, we randomly sample 10 from the $\binom{22}{n}$ possible combinations (where $n$ is the number of users being tested), apply 10 fold cross validation on each of the samples, calculate the accuracy and then get the average over the 10 samples. On average, each sample group has 63 instances.

**Performance of standalone frequency and entropy features.** For basic multimodal frequency and entropy features, each vector has 4 dimensions, corresponding to WiFi, cell tower, Bluetooth, and apps, respectively. We compare their classification results as shown in Figure 6.3b. Both frequency and entropy features outperform the random baseline significantly. Entropies have better performance for large numbers of users compared to frequencies: mean accuracy with entropies drops 22 percentage points from 4 users to 22 users, versus a 35 point drop using basic frequencies.

**Performance of conditional features.** We also compared various types of conditional features (Figure 6.3c). When features including basic multi-modal entropies, location entropies and time entropies are all combined, the performance is the best. Time features perform slightly better than location features, perhaps because there are only 2 location-conditioned features but 5 time-conditioned features. With all three kinds of features combined, the accuracy is 71.96% for 22 users versus 91.54% for 4 users and 86.59% for 10 users. Though more features improve accuracy, more computation is required as well, especially for the clustering required by location conditioning.

**Performance of all features.** Finally, we compute the performance with all features including basic frequencies, entropies, and conditioned features. Figure 6.3d shows the performance improvement against the basic frequency feature results shown in Figure 6.3b, with the absolute accuracy marked on the top of each bar: 94.68% for 4 users, 93.14% for 10 users and remains 81.30% for 22 users.

## 6.3 DISCUSSION AND CONCLUSION

With regard to design, the lightweight and energy-efficient computation of features can be accomplished on mobile devices. The simple features (frequencies and entropies) show their effectiveness in the user identification task. On one hand, this suggests their descriptive power – they capture the characteristics of a user that differentiate him or her from other users. Future work includes associating simple behavioral statistical features with user attributes such as social patterns and habits by finding the connections between users in order to build personalization applications. On the other hand, as an identification task, it naturally brings in privacy implications as discussed in Section 2.5. Throughout this thesis, we talk about collecting and analyzing user-sensed data for observing the world. Users may be concerned about sharing sensitive data such as locations, app usage, web browsing history and communication logs. In contrast, users may be less concerned about sharing the simple statistics as described in this project. This data can be rich enough to produce useful

information while its richness may still reveal too much information about the users that is deemed to be sensitive and private. To make these features useful yet with privacy preserved, we propose a 'mFingerprint' architecture in which users control whether to reveal the data or not. Figure 6.4 shows the overview of this four-layer architecture. The bottom layer collects various sensor readings using hardware sensors that trace location (e.g., GPS, WiFi, and cell tower), proximity (Bluetooth and microphone). Furthermore, soft sensor data such as application usage and web browser history are also recorded. The second layer computes a set of statistical features from these digital footprints. In this project, we particularly designed frequency and entropy based statistical features to capture mobile device usage patterns. The first two layers are deployed on users' mobile devices. A device-side privacy control module warns the user about the potential privacy risks if these features are revealed. The user then can decide whether to share these features with the third layer for user learning which might happen outside of the device. The user learning layer includes modules for building user models via the feature vectors, identifying users via classification methods, and grouping users into meaningful clusters via unsupervised learning. The forth layer is the application layer where various applications can be established, such as inferring user profile, logging mobile behaviors, and creating personalized services and user interfaces. This layer may either be on the device or on the server side. We hope the descriptive power of these simple features as well as this proposed architecture can be of interest to the privacy research community and promote research on improving the personalized services with privacy better preserved.

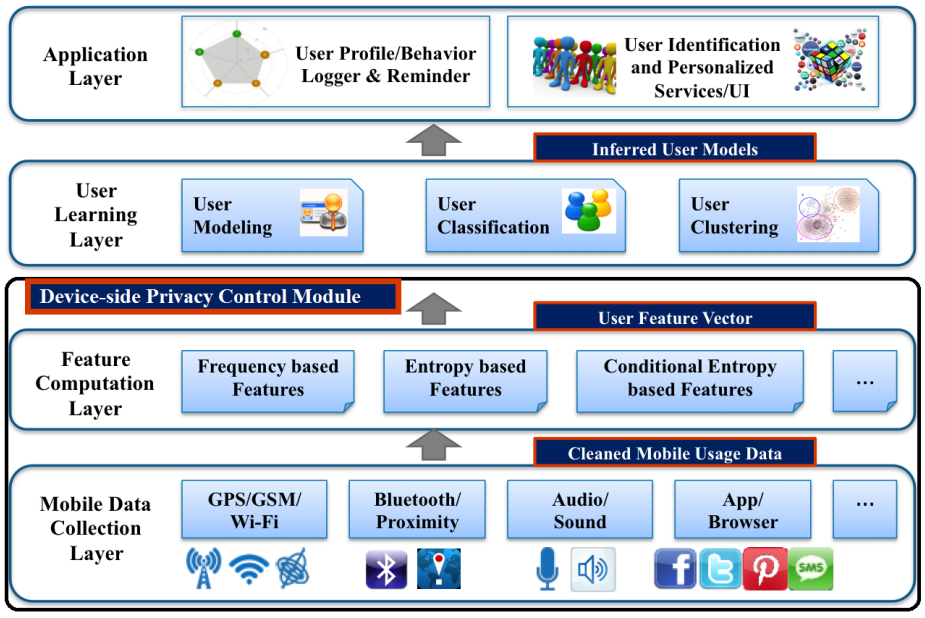



Figure 6.4: The four-layer 'mFingerprint' architecture with privacy control.

# CHAPTER 7

# TEMPORAL USER BEHAVIOR STUDY

In Chapter 6, we studied individual mobile users. In this chapter, we switch our focus onto large-scale online human behaviors.

Most research on online behaviors, up until recently, has focused on social media behaviors and e-commerce behaviors independently. In our study, we choose a particular global e-commerce platform (eBay) and a particular global social media platform (Twitter). As motivated in Section 1.4.4, we ask this question: do user-sensed datasets contain signals for predicting large-scale consumer behaviors? We quantify the characteristics of the two individual trends as well as the correlations between them. We provide evidence that about 5% of general eBay query streams show strong positive correlations with the corresponding Twitter mention streams, while the percentage jumps to around 25% for trending eBay query streams. Some categories of eBay queries, such as 'Video Games' and 'Sports', are more likely to have strong correlations. We also discover that eBay trend lags Twitter for correlated pairs and the lag differs across categories. We show evidences that celebrities' popularities on Twitter correlate well with their relevant search and sales on eBay. The correlations and lags provide predictive insights for future applications that might lead to instant merchandising opportunities for both sellers and e-commerce platforms. In preparation for future study, we adapt a peak detection algorithm to better monitor the two streams and conduct a case study to further understand the complex interplay.

## 7.1 QUANTIFYING CORRELATIONS

We start with two datasets – Twitter tweets and eBay search query logs. The Twitter dataset contains a sample of tweets, each of which has a short text of the Tweet itself, a user id of the person composing the Tweet and a timestamp indicating when the Tweet was posted. The eBay dataset contains a sample of eBay search query logs, each of which has a text query, a user id of the person issuing the query, the number of resulting items the user clicks on, the number of bids the resulting items receive from the user, the total number of price-fixed 'Buy It Now' (BIN) item the user purchases and the total prices the user pays on these BIN items. The details about the datasets can be found in Section 1.3.

We focus on keyword phrases appearing in tweets and eBay queries and for a keyword phrase, we extract two time series of its mentions as a delegate of its popularity over time on Twitter and eBay, respectively. For a keyword phrase such as 'Barack Obama' during a certain period of time, in each time unit, we count the number of unique Twitter users with tweets containing this keyword phrase as well as the number of unique eBay users ('SEARCH') with queries containing this keyword phrase. For eBay dataset, besides the number of unique users in each time unit, we can also compute the total number of clicks ('VIEW') and bids ('BID') on resulting items, the total number of purchased BIN ('BIN_count') items of the resulting item, the total BIN ('BIN_total') and average BIN prices ('BIN_avg') of these items.

As suggested in [94] as well as in Section 3.2, we calculate the Pearson Correlation coefficients between the normalized pairs of time series as well as the 2-tailed p-value for testing the null hypothesis that the true correlation coefficient is equal to 0. In the following subsections, we will describe the time series extraction and correlation calculation in detail.

### 7.1.1 EXTRACTION OF TIME SERIES

We extract time series which represent the keyword trends from the two datasets.

A simplified scenario of a buyer using eBay is: the user issues a query and clicks on several resulting items, then the user has the option to bid on some of them and she can also choose to purchase some price-fixed items right away using 'Buy It Now' (BIN). For the experiment, we define the set of all eBay user actions $\mathcal{E} = \{a_1, a_2, ..., a_q\}$ as a set of tuples of the form $a_i = (u_i, q_i, t_i, v_i, bid_i, bin_i, p_i)$, where $u_i$ is a user, $q_i$ is a query, $t_i$ is a timestamp, $v_i$ is the number of unique items the user clicks on among the resulting items after issuing the query, $bid_i$ is the number of bids the user places on the resulting items, $bin_i$ is the number of BIN items the user purchases among the resulting items and $p_i$ is the total amount of money the user pays for all the BIN items that she purchases.

In order to get the time series of search mentions for a keyword phrase $k$ from time $t_s$ to $t_e$ on eBay, we divide this period of time into $m$ coarse temporal bins. For each bin, we count the number of unique users who issued queries containing the keyword phrase. We iterate over the eBay user action set $\mathcal{E}$ and the resulting eBay vector is denoted as $U_{\mathcal{E}}$. Let $q_T(t_i, t_s, o)$ be the quantization function described in Section 3.1.2. It maps a timestamp $t_i$ into one of $m$ temporal bins with the starting time $t_s$ and temporal bin size $o$, returning a bin index in the range $[1, m]$. Function $contain(q_i, k)$ returns $true$ if $k$ is a substring of $q_i$ or otherwise $false$. (Please notice that we reprocess $q_i$ and $k$ by replacing consecutive non-alphanumeric characters with one space.) For any keyword $k$, we then build an $m$-dimensional vector, counting the number of unique users who issued search queries containing $k$ in each temporal period $b$:

$$U_{\mathcal{E}}(b, k, t_s, t_e, o) = ||\{u_i | (u_i, q_i, t_i, v_i, bid_i, bin_i, p_i) \in \mathcal{E},$$
$$t_i \leq t_e, t_i \geq t_s, contain(q_i, k), b = q_T(t_i, t_s, o)\}||, \tag{7.1}$$

where $b$ is the bin index. Using this approach, we can also extract time series for 'VIEW', 'BID', 'BIN_count', 'BIN_total' and 'BIN_avg' as mentioned above. We denote their quantization functions as $U_{\mathcal{E}}^{view}$, $U_{\mathcal{E}}^{bid}$, $U_{\mathcal{E}}^{bin\_c}$, $U_{\mathcal{E}}^{bin\_t}$ and $U_{\mathcal{E}}^{bin\_a}$ respectively.

Similarly, we extract the time series of mentions for a keyword phrase $k$ on Twitter. We build an $m$-dimensional vector $T$ from the Twitter user action set $\mathcal{T}$ using the quantization function $U_{\mathcal{T}}(b, k, t_s, t_e, o)$ which counts the number of unique users who posted tweets containing $k$ in each temporal period $b$ from $t_s$ to $t_e$.

As $k$ and $o$ are fixed for each comparison between eBay and Twitter, we drop them as inputs for the quantization functions in this particular case. We also omit the index $b$. $U_{\mathcal{E}}(t_s, t_e)$ extracts a vector $E$ for eBay user action set $\mathcal{E}$, from starting timestamp $t_s$ to ending timestamp $t_e$, where $E_i = U_{\mathcal{E}i}(t_s, t_e)$. Similarly, $U_{\mathcal{T}}(t_s, t_e)$ extracts a vector $T$ for Twitter user action set $\mathcal{T}$. For all the vectors extracted, we perform $L1$ normalization.

### 7.1.2 PEARSON'S CORRELATION CO-EFFICIENT AND A $t$-TEST

We compute the Pearson's correlation co-efficient $r = P(E, T)$ between the eBay vector $E$ and the corresponding Twitter vector $T$ for each keyword phrase. As mentioned in Section 3.2, it measures the linear dependence between two datasets and gives a co-efficient in $[-1, 1]$ with +1 indicating exact positive linear relationship, -1 indicating exact negative linear relationship and 0 indicating no correlation.

We perform Student's $t$-test on Pearson's $r$ of the pair of vectors and the null hypothesis is that they are uncorrelated (the actual $r$ is 0). If the underlying variables follow a bivariate normal distribution, the sampling distribution of Pearson's $r$ would have a Student's t-distribution with degrees of freedom $n = 2$. According to [60], this holds approximately for sample sizes that are not small, even when the observed values are not normal. Therefore, even though $E$ and $T$ might not have a bivariate normal distribution, we can still use the two-tailed p-value from the t-test as an approximation of confidence of correlation.

## 7.2 COMPUTING LAG

We quantify the evidence that one stream lags its counterpart in another domain by applying a moving window method as suggested in Section 3.2. It finds a shift that maximizes the Pearson's $r$ between the pair of streams (action sets). Though the evidence from one pair of streams is not strong, when we compute the lags for a massive amount of keyword phrases, there would be enough signals. We choose this method instead of the popular Granger causality test [43] as suggested in [12], because fixed temporal bins are not suitable to capture and quantify the instant sudden changes in popularity at a fine grain. For example, with the bin size fixed to one day, it might not be capable of capturing the changes happen in one day, if one stream lags another by minutes or hours; if the bin size is set to a couple of minutes or hours, there might not be enough counts in each bin; the lag might also vary for different pairs of streams and different types of events which might only be captured by windows moving at finer grains.

For each pair of streams, we shift one stream by a period of time $\Delta t$ to compute the vector and calculate the Pearson correlation between the shifted vector and the other vector which is not shifted. The $\Delta t$ which maximizes the Pearson's $r$ is defined to be the lag. We adapt the definitions in Section 3.2 here. Before shifting, we want to compare the vectors that starts at time $0$ and ends at time $N$. The Pearson correlation co-efficient between the extracted eBay vector and Twitter vector is calculated as:

$$f(\mathcal{E}, \mathcal{T}, \Delta t) = \begin{cases} P(U_{\mathcal{E}}(\Delta t, N), U_{\mathcal{T}}(0, N - \Delta t)), \Delta \geq 0 \\ P(U_{\mathcal{E}}(0, N - |\Delta t|), U_{\mathcal{T}}(|\Delta t|, N)), \Delta t < 0 \end{cases} \tag{7.2}$$

where one of the streams is shifted by $\Delta t$.

With this adaptation, lag $l_{\mathcal{E},\mathcal{T}}$ is computed as:

$$l_{\mathcal{E},\mathcal{T}} = \underset{\Delta t \in [-\mathcal{S}_1, \mathcal{S}_2], \Delta t \in \mathbb{Z}}{\operatorname{argmax}} f(\mathcal{E}, \mathcal{T}, \Delta t), \tag{7.3}$$

where positive $l_{\mathcal{E},\mathcal{T}}$ suggests eBay lags Twitter by $l_{\mathcal{E},\mathcal{T}}$ while negative $l_{\mathcal{E},\mathcal{T}}$ suggests Twitter lags eBay by $|l_{\mathcal{E},\mathcal{T}}|$.

## 7.3 EXPERIMENTS AND RESULTS

### 7.3.1 GENERAL CORRELATIONS

We select the top queries from the long tail distribution of eBay queries as general keyword phrases that we monitor. For the top 150,000 queries on eBay (on July 11, 2012), we monitor the period of time from January 1 2012 to March 31 2012 on eBay and Twitter. In this experiment, the eBay dataset contains a large representative randomized sample of the queries logs and the Twitter dataset includes 1% of the tweets posted on Twitter. We set thresholds on both daily query counts and daily Twitter mentions to ensure the density of the vectors that we extract. For average eBay daily query counts, we require the numbers to be no less than 20 and for average Twitter mentions, we set the threshold to 5 per day. This gives us a list of 16,099 keyword phrases that can represent general queries on eBay. For the convenience of later references, we name it as GeneralQueryList. For each phrase in this list, we get two time series of daily mentions from both eBay and Twitter. For each pair of time series, we compute the Pearson correlation coefficient $r$ and the p-value of the t-test with the null hypothesis that the two time series are generated from two uncorrelated systems.

The statistics in Table 7.1 suggest that around 5.25% of these queries can be considered as having significant positive correlations with their counterparts on Twitter while only around 0.65% show significant negative correlations at $p = 0.0005$. Here we consider several confidence levels from 0.0001 to 0.01 and calculate the fractions of positively and negatively correlated keyword phrase pairs. Notice that when we increase the confidence level, the ratio between positively correlated pairs and negatively correlated pairs increases, suggesting the possible noises are removed.

In order to examine the correlations in behaviors across the two domains for different classes of

Table 7.1: Fractions of correlated pairs of keyword phrases at different confidence levels. The ratio between positively correlated pairs and negatively correlated pairs increases as the confidence level increases, suggesting the possible noises are removed.

| p-value | pos_corr | neg_corr | pos/neg |
|---|---|---|---|
| 0.01 | 8.52% | 2.49% | 3.42 |
| 0.005 | 7.34% | 1.77% | 4.14 |
| 0.001 | 5.75% | 0.80% | 7.18 |
| 0.0005 | 5.25% | 0.65% | 8.07 |
| 0.0001 | 4.35% | 0.32% | 13.59 |

Table 7.2: Top 5 eBay meta categories ranked by portions of queries with strong correlations at the confidence level of 0.0005.

| Category | pos_corr |
|---|---|
| Video Games | 21.28% |
| DVDs & Movies | 14.20% |
| Entertainment Memorabilia | 13.47% |
| Sports Mem, Cards & Fan Shop | 13.45% |
| Tickets | 13.04% |

queries, we break these queries into the 36 eBay meta categories[1] and some categories have higher portions of strongly correlated queries at the confidence level of 0.0005. As we may expect, the categories such as 'Video Game' and 'Sports Mem, Cards & Fan Shop' which are more likely to have sales driving news events such as a championship match and releasing of a video game appear in the top 5 categories are shown in Table 7.2, as apposed to categories like 'Home & Garden' and 'Pottery & Glass'. Among all the general queries, 'Sports Mem, Cards & Fan Shop' takes up 7.34% which is a relatively large portion.

We then select the trending queries detected by the system described in [85] from the General-QueryList mentioned above, resulting in 730 queries which we name as TrendQueryList. As shown in Table 7.3, 24.93% of the trending keyword phrase pairs have very strong correlations, compared to 5.25% for the general pairs. At 0.01, a relaxed level of confidence, almost 30% of them are correlated. At the same level, the ratio between positively correlated pairs and negatively correlated pairs is much higher than that in Table 7.1. When we break these trending queries into the eBay meta categories, we find that some categories demonstrate much more correlations. 69.23% of the

[1]There are 36 meta categories at the top level of the eBay merchandise ontology composed by domain experts. The list is available on `www.ebay.com`. A query is assigned to a category according to what the majority of users issuing the same query clicked on and purchased historically.

Table 7.3: For trending keyword pairs, fractions of correlated pairs of keywords at different confidence levels. At the same level, the ratio between positively correlated pairs and negtively correlated pairs is much higher than that in Table 7.1.

| p-value | pos_corr | neg_corr | pos/neg |
|---|---|---|---|
| 0.01 | 29.58% | 1.64% | 18.03 |
| 0.005 | 28.21% | 1.09% | 25.88 |
| 0.001 | 25.61% | 0.13% | 197 |
| 0.0005 | 24.93% | 0.13% | 191.76 |
| 0.0001 | 22.46% | 0% | INF |

Table 7.4: For trending keyword pairs, top 5 eBay meta categories ranked by portions of queries with strong correlations at the confidence level of 0.0005.

| Category | pos_corr |
|---|---|
| Sports Mem, Cards & Fan Shop | 69.23% |
| Video Games & Movies | 64.28% |
| Cell Phones & PDAs | 52.94% |
| Entertainment Memorabilia | 37.50% |
| Music | 36.66% |

queries in 'Sports Mem, Cards & Fan Shop' are strongly correlated at the confidence level of 0.0005 as shown in Table 7.4. Even for the 'Music' category which ranks at 5th, its percentage is 15% higher than that of the top category in general queries.

### 7.3.2 LAG BETWEEN TWO STREAMS

From the GeneralQueryList, we get a list of pairs which are strongly correlated using a threshold of Pearson's $r$ at 0.4 and that results in 690 queries which we name as CorrelatedQueryList. For these queries, we calculate the lags in the shift range of $[-5000, 5000]$ minutes with positive values indicating eBay lags Twitter and vice versa.

In order to reduce noise, we require the increase in Pearson's r to be at least 5% to be considered that there is a lag. We then plot the histogram of the lags as shown in Figure 7.1. The average lag value is 290 minutes (4.83 hours) and 61.30% of the 690 queries have positive lag values as opposed to 50% if the lags are normally distributed around a mean value of 0, suggesting that for keyword pairs correlated on a daily basis, eBay is lagging Twitter. For the pairs where eBay lags Twitter, the average lag is 1214 minutes (20.2 hours) and the third quantile is 1441 minutes (24.0 hours).

Among the 36 eBay meta categories, 'Clothing, Shoes & Accessories' ('Clothing') and 'Sports

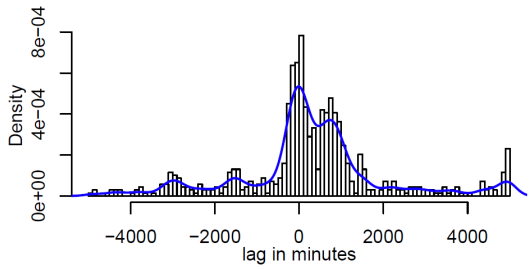



Figure 7.1: Histogram of the lags with the density curve for CorrelatedQueryList. The average lag time of eBay from Twitter is 4.83 hours and 61.30% of the 690 queries have positive lag values. Most of the lags are distributed on the positive half of the x-axis.

Mem, Cards & Fan Shop' ('Sports') have more than 100 keywords out of the 690 keywords and we choose to observe their lag patterns as shown in Figure 7.2 and Figure 7.3. For 'Clothing', only 45.6% have positive lag values, while for 'Sports' the percentage is 70.14%. This suggests that for certain categories, the signal is stronger that Twitter is leading eBay.

Again we pick the trending queries from the CorrelatedQueryList and it gives us 164 keyword phrases. The average lag value is 662 minutes (11.03 hours) and 76.82% of them have positive lag values compared to 290 minutes and 61.30% for general queries. The histogram is shown in Figure 7.4. Noticing that 10 keyword phrases have lags of over 4000 minutes, we plot two of them here with shifted eBay curves in red marked with asterisks: one is a general keyword phrase 'air conditioner' with a lag of 4950 mins (3.43 days) and the other is a more specific product keyword phrase 'droid 4' with a lag of 4981 mins (3.45 days) as shown in Figure 7.5 and Figure 7.6. After shifting, the Pearson's $r$ between Twitter and eBay for 'air conditioner' goes from 0.42 to 0.48 and for 'droid 4', the value goes from 0.45 to 0.57 which suggests a margin of over three days for eBay to respond to Twitter for these two keyword phrases. 'droid 4' is a Motorola cellphone released on Feb 10, 2012 but 3 days before that, there were tweets and news articles about Verizon confirming the releasing date which contributed to the Twitter burst on Feb 7.

In histograms in Figure 7.1, 7.2 and 7.4, the gaps between neighboring local maximas are usually around a day (1440 minutes), which might be related with the daily repetitions of users' activity pattern.

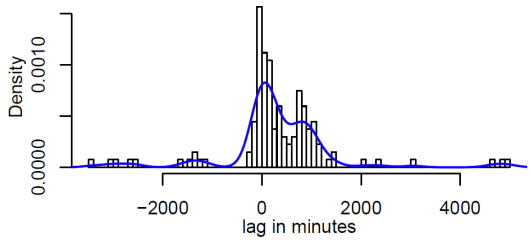



Figure 7.2: Histogram of the lags in Sports from CorrelatedQueryList with the density curve. 74% have positive lag values and as a result, the majority of lags is distributed on the positive side of the x-axis.

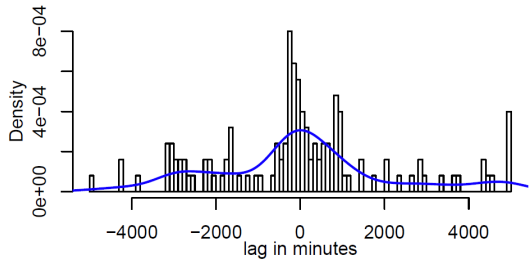



Figure 7.3: Histogram of the lags in Clothing from CorrelatedQueryList with the density curve. 45.6% have positive lag values and the lags are distributed more evenly on both sides of 0.

### 7.3.3 CELEBRITY WATCHING

Celebrity news often trends in Twitter [64] and we want to know when celebrity news – deaths, movies, arrests, games, etc. trends in one network what influence it has on the other network. Some possible scenarios include: the news of a celebrity death triggers a book release, an Oscar nomination is followed by sudden increases in sales of movies in which the celebrity starred, and a celebrity retirement stimulates fans' passion for memorabilia. Here we explore the relationships between popularity of celebrities on Twitter and their related searches and sales on eBay. We make

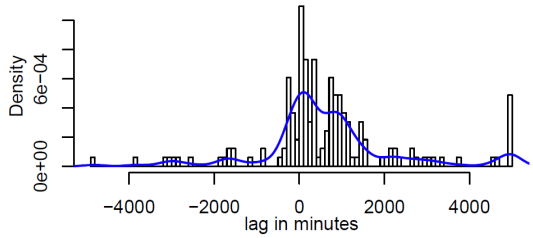



Figure 7.4: Histogram of the lags for trending keywords with the density curve. 76.82% of them have positive lag values, the majority of lags is also distributed on the positive side of the x-axis.

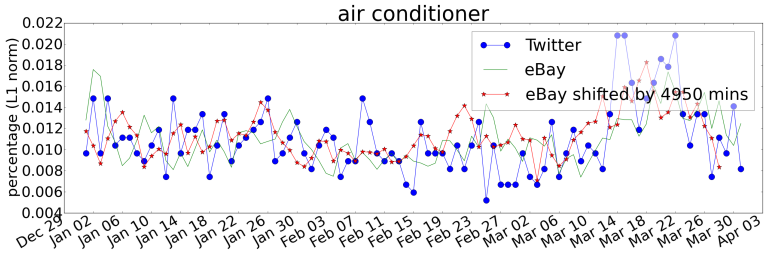



Figure 7.5: Twitter trend, eBay trend and shifted eBay trend for 'air conditioner'. The lag (shift) is 3.43 days.

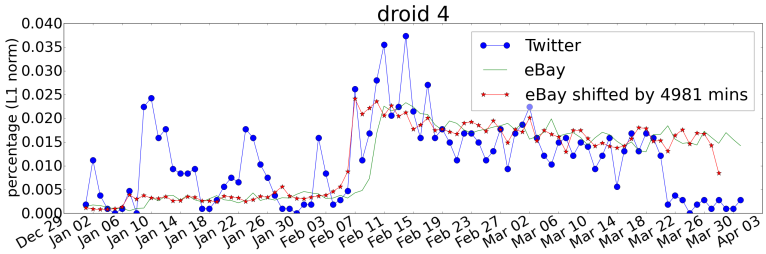



Figure 7.6: Twitter trend, eBay trend and shifted eBay trend for 'droid 4'. The lag (shift) is 3.45 days.

use of Forbes 'The Celebrity 100' list [2] which includes celebrities based on media visibility and entertainment-related earnings.

From Nov 1, 2011 to Jun 27, 2012 (excluding 34 missing days), for each celebrity we compute the daily statistics including number of searches (SEARCH), number of views (VIEW), number of bids (BID), number of BINs (BIN_count), total BIN price (BIN_total) and average BIN price (BIN_avg) such that for each celebrity on the list, there will be 6 time series from eBay. For each of the eBay time series, we compute the Pearson's $r$ between itself and its corresponding Twitter

[2] `www.forbes.com/celebrities/list/`, May 2012

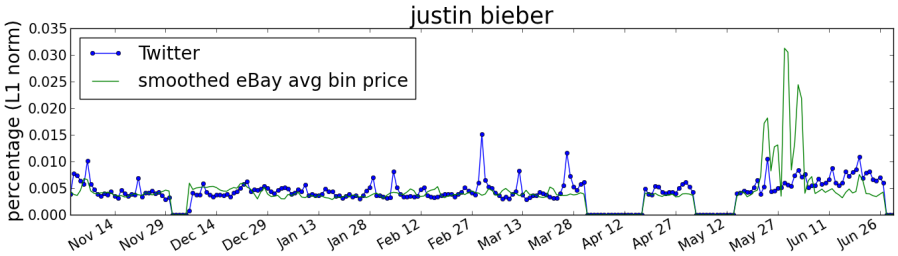



Figure 7.7: Twitter trend and smoothed eBay average BIN price for 'justin bieber', with a smoothing window size of 2. Comparing with no smoothing, the Pearson's $r$ goes from 0.183 to 0.233 with a p-value of 0.0007. (Zero values are due to missing data.)

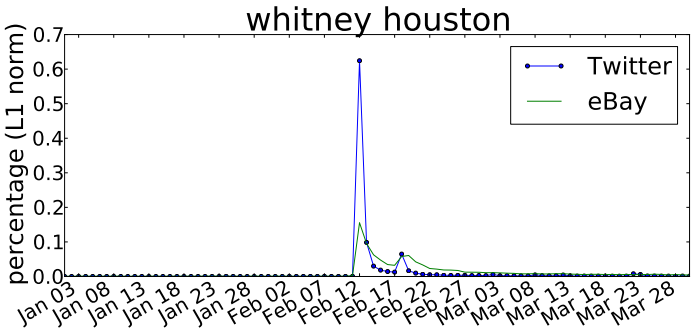



Figure 7.8: Twitter trend and eBay trend for 'whitney houston'. The bursts on both streams correlate well on the day she passed away and the day for her funeral. Users' interest drops slower on eBay which suggests a different pattern of attention from Twitter.

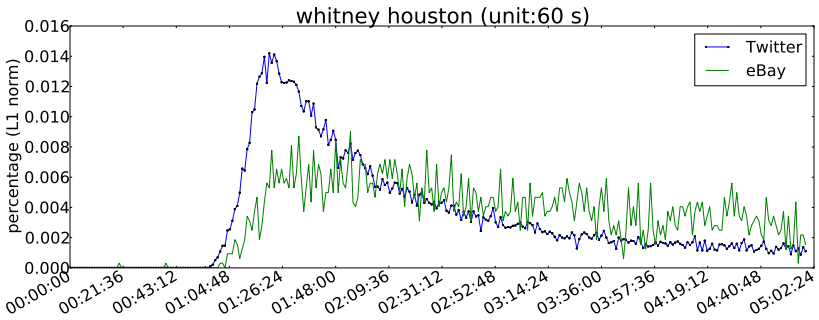



Figure 7.9: Twitter trend and eBay trend for 'whitney houston' with unit set to 60 seconds. Twitter burst rises faster than eBay and peaks earlier as well. Users' interest still drops quicker on Twitter after the peak.

mention time series as well as the p-value of the t-test. We threshold on different confidence levels to examine the percentage of positively correlated pairs and that of negatively correlated pairs and the results are shown in Table 7.5.

It shows that only for BIN_count at lower confidence levels, there are negatively correlated pairs. It also suggests that for SEARCH and VIEW, the pairs are best correlated among the 6 statistics. BID, BIN_count and BIN_total demonstrate moderate correlations and a possible explanation for this is that sometimes users just want to check what is on eBay but will not necessarily make up their minds purchasing.

The average BIN price (BIN_avg) is relatively less correlated – only 5% are correlated at 0.01 confidence level. A possible reason for this is that it takes time for the price to reflect the changes in

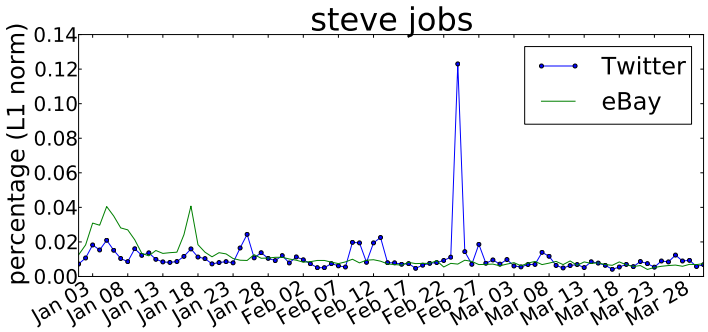



Figure 7.10: Twitter trend and eBay trend for 'steve jobs'. The first two big bursts on eBay are related with the production of Steve Jobs action figures while the huge burst on Twitter is related with his birthday. The relatively weak correlation suggests the interplay between two streams is sometimes profound.

Table 7.5: For 100 celebrity keywords, fractions of correlated pairs of keywords at different confidence levels. For SEARCH and VIEW, the pairs are best correlated among the 6 statistics. BID, BIN_count and BIN_total demonstrate moderate correlations.

| | SEARCH | | VIEW | | BID | | BIN_count | | BIN_total | | BIN_avg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p-value | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr |
| 0.01 | 46% | 0% | 44% | 0% | 14% | 0% | 16% | 2% | 12% | 0% | 5% | 0% |
| 0.005 | 45% | 0% | 43% | 0% | 11% | 0% | 14% | 1% | 11% | 0% | 3% | 0% |
| 0.001 | 39% | 0% | 38% | 0% | 10% | 0% | 11% | 1% | 9% | 0% | 1% | 0% |
| 0.0005 | 37% | 0% | 30% | 0% | 8% | 0% | 7% | 1% | 7% | 0% | 0% | 0% |
| 0.0001 | 37% | 0% | 30% | 0% | 8% | 0% | 7% | 0% | 7% | 0% | 0% | 0% |

popularity (the prices for BIN items are set when the sellers list the items) and the prices might be insensitive in nature. In order to further understand this, we calculate the BIN_avg on a certain day as the average BIN price of $n$ days in the future (BIN_total/BIN_count), together with the current day. We show the average Pearson's $r$ in Table 7.6 where $n$ ranges from 0 to 14 and the portions of correlated pairs at different confidence levels for $n \in [0, 4]$ in Table 7.7. From these two tables, the average Pearson's $r$ peaks when $n = 2$ and at the same time, greater portion (9%) of pairs are correlated at the confidence level of 0.01. It suggests that celebrities' popularity impacts the average prices of their related items in a 3-day window. Here we show one example for Justin Bieber, with the window size 2 in Figure 7.7. Comparing with no smoothing, the Pearson's $r$ goes from 0.183 to 0.233 with a p-value of 0.0007.

Table 7.6: Average Pearson's $r$ between Twitter mention time series and eBay averaged BIN price time series in a $(n+1)$-day window. The average Pearson's $r$ peaks when $n = 2$.

| n | 0 | 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| avg_r | 0.018 | 0.015 | **0.020** | 0.018 | 0.012 | 0.009 | 0.010 | 0.013 | 0.014 | 0.012 | 0.012 | 0.013 | 0.011 | 0.007 | 0.005 |

Table 7.7: For $n \in [0, 4]$, portions of correlated pairs of keywords at different confidence levels. When $n = 2$, a greater portion (9%) of pairs are correlated at the confidence level of 0.01.

| | n=0 | | n=1 | | **n=2** | | n=3 | | n=4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| p-value | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr | p_corr | n_corr |
| 0.01 | 5% | 0 | 6% | 0 | **9%** | 0 | 8% | 1% | 7% | 0 |
| 0.005 | 3% | 0 | 4% | 0 | 8% | 0 | 8% | 0 | 7% | 0 |
| 0.001 | 1% | 0 | 4% | 0 | 4% | 0 | 4% | 0 | 4% | 0 |
| 0.0005 | 0 | 0 | 1% | 0 | 3% | 0 | 4% | 0 | 2% | 0 |
| 0.0001 | 0 | 0 | 1% | 0 | 3% | 0 | 4% | 0 | 2% | 0 |

### 7.3.4 PEAKINESS OF TWO STREAMS

The peakiness of a usage stream from a website reflects its users' attention as well as the nature of the platform itself. We measure the general peakiness of the eBay stream and the Twitter stream by computing their average second moment as suggested in Chapter 4 and defined in equation 4.8.

For the GeneralQueryList mentioned in Section 7.3.1, we compute the average second moment of their time series of mentions on eBay and Twitter. The value for eBay is 0.011 while the value for Twitter is 0.016, suggesting that Twitter streams are more peaky than eBay streams. The reason might be that as a form of news media, Twitter's user attention spikes and drops more quickly.

### 7.3.5 PEAK DETECTION

In order to better track the two streams, we test an on-line peak detection algorithm described in [73] on our datasets to monitor both live streams. Figure 7.11 shows the burst detection results on Twitter and eBay time series for the keyword phrase 'Chicago Bulls' on a daily scale. The huge Twitter peak on April 28 marked in yellow related to the injury of its star player Derrick Rose. Though the corresponding eBay peak does not seem obvious as the injury does not necessarily drive the increase in sales, the algorithm is still able to detect the subtle rise in popularity.

Peak detection serves as a step toward our future goal of understanding the complex interplay

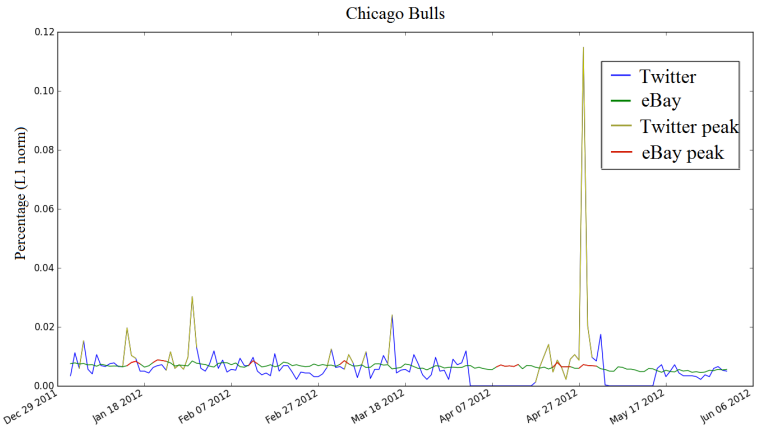



Figure 7.11: Peak detection results for the keyword phrase 'Chicago Bulls' in Twitter and eBay streams. (Best view in color. Zero values are due to missing data.)

between the two streams. For example, with the peaks automatically captured, textual features can be extracted from the tweets or eBay queries within the peak windows as suggested in [73]. These textual features can potentially help us align the peaks across a pair of streams. Then we may further classify the Twitter peaks into sales-driving peaks and non-sales-driving peaks.

## 7.4 CASE STUDIES

We select some cases to demonstrate the characteristics of the two trends that we observe. Two celebrity keyword phrases, 'Whitney Houston' and 'Steve Jobs', are selected as they are both popular and easy to be mapped to real world news events. We also study the trends for American football teams – New England Patriots and New York Giants during the 2011-2012 season as well as the final game.

The 'Whitney Houston' streams shown in Figure 7.8 appear to have a very strong correlation (Pearson's $r = 0.794$, p-value = 5.44e-21). The bursts on both streams correlate well on the day Whitney Houston passed away and the day for her funeral. But there are some differences in users' pattern of attentions. Twitter is more bursty as a news media but after the burst, users have the information and stop talking about it, while on eBay, users still remain interested in purchasing.

We then examine the period from 00:00 to 05:00, Feb 12, 2012 during which public got to know Whitney Houston passed away. As shown in Figure 7.9 where the unit is set to 60 seconds, we can see how queries and tweets arrived at a finer grain and it suggests that Twitter burst rises faster than eBay and peaks earlier as well. At this scale, we are still able to observe that users' interest drops quicker on Twitter after the peak suggesting that users remained interested in purchasing on eBay after they knew the news.

For the 'Steve Jobs' streams shown in Figure 7.10, the correlation is not strong (Pearson's $r = 0.118$, p-value = 0.264). We observe that there are two big peaks in January for eBay while the corresponding peaks on Twitter are relatively weaker which is not so usual as Twitter is often more bursty than eBay. We manually look at the 240 tweets containing 'Steve Jobs' on January 3 corresponding to the first big burst on eBay. Over a third of them are about a Chinese factory planning to make unauthorized lifelike replica of Steve Jobs[3] and some of the tweets express the willingness to purchase. Meanwhile, we check the eBay queries containing 'Steve Jobs' on the same day, about 33% of them were related to action figures or bobble heads while the ratio for that on Jan 1 is only 5%. For the second burst on eBay, we check the related tweets and eBay queries on Jan 17. A third of the tweets were about the Chinese factory canceling the production of the action figures and 50% of the eBay queries were related to the action figures. A news article[4] pointed out that people were still selling and buying the action figures on eBay and the highest price reached $2500.

For the huge burst on Twitter on February 24, there is no obvious corresponding burst on eBay. But when we check the tweets on that day, it turns out that it was Steve Jobs birthday and a lot of people were memorializing him by Tweeting. All these indicate that the interplay between user behavior on Twitter and that on eBay sometimes can be profound and this kind of correlations/non-

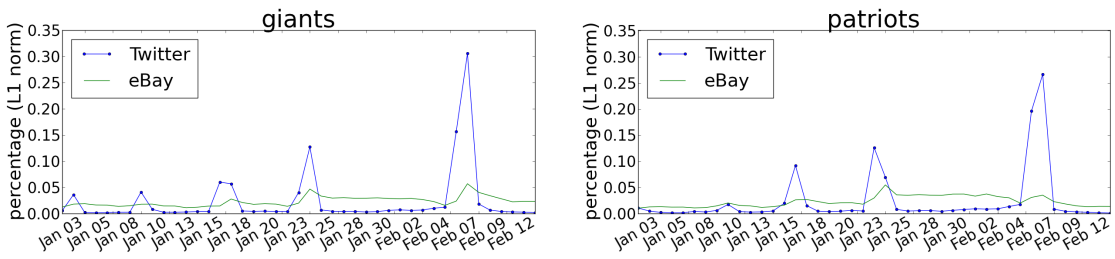
[3] `www.foxnews.com/tech/2012/01/03/steve-jobs-action-figure-planned-for-february`, Fox News

[4] `www.pcworld.com/article/248238/maker_of_steve_jobs_action_figure_kills_project.html`, PCWorld

correlations could offer important insight on discovering the factors that drive sales on e-commerce platforms.

We track the popularity of the American footballs teams New York Giants and New England Patriots on eBay and Twitter. For these two American football teams and corresponding sports events, the pattern seems to have more regularity. As shown in Figure 7.12, their daily popularity during the football season (we examine from Jan 1 to Feb 12 2012) correlates with significance and outcomes of games. The Giants won three rounds of playoff games on Jan 8, Jan 15, and Jan 22, respectively and proceed to the major championship game – Super Bowl XLVI on Feb 5, against the Patriots. For the Patriots, they won playoff games on Jan 14 and Jan 22 and met the Giants in the final game. The games become more and more important as the playoff season proceeds, thus later games receive more attention as it is reflected in the significance of the corresponding Twitter and eBay peaks. The Patriots lost to the Giants in Super Bowl game on Feb 5. We notice that the Patriots receive smaller Twitter and eBay peaks on that day and their eBay popularity drops quickly to less than its popularity before the final game while the Giants' eBay popularity drops slower after the game. This suggests that eBay users remain interested in purchasing the items related to the champion team after the event. Overall, the Giants receive more attention than the Patriots during this 40-day period – they have about twice the number of mentions on both platforms comparing with the Patriots.

As shown in Figure 7.13, we track the two teams' popularity during the Super Bowl game at a finer grain. We notice that touchdowns usually cause increases in popularity for the corresponding teams on both platforms with eBay lags Twitter. The winning team enjoys much higher 'end of game' Twitter and eBay peaks comparing with the losing team, again with eBay lags Twitter. During this period of time, the Giants' overall popularity is more than twice that of the Patriots on both platforms.




(a) Five correlated bursts are related to five winning games for the Giants in 2012, including the Super Bowl game against the Patriots on Feb 5.

(b) After two peaks corresponding to two winning games, the Patriots lost the final game, resulting in a smaller final peak comparing the Giants' final peak.

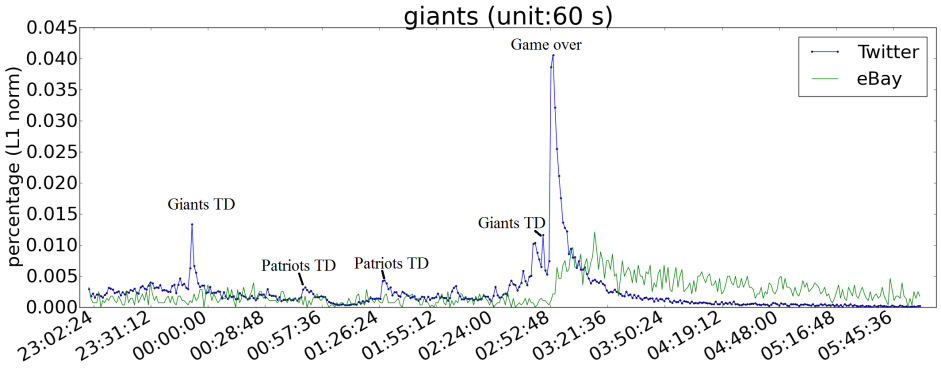
Figure 7.12: Twitter and eBay trends for the keyword phrases 'giants' and 'patriots' corresponding to two American football teams from Jan 1 2012 to Feb 12 2012.

## 7.5 CONCLUSIONS AND FUTURE WORK

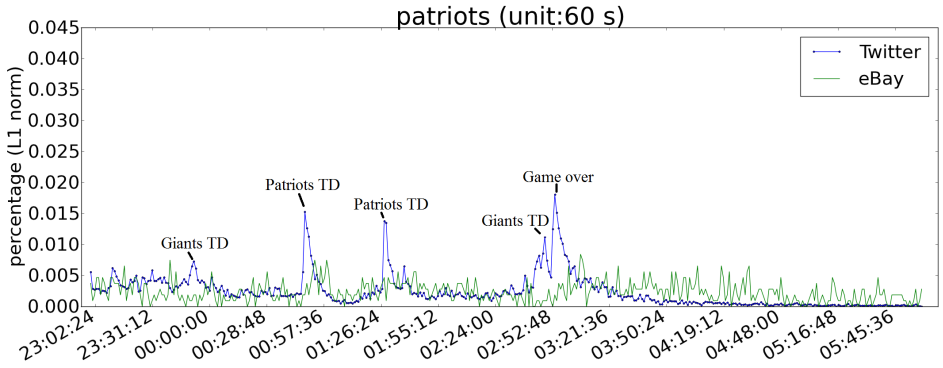
In this project, we proposed techniques to quantify the correlations and lags between the trend on social media and that on e-commerce. We also examined the individual characteristics of the two streams through a case study and by measuring their peakiness. We discover evidence that:

- About 5% of the eBay query streams have strong positive correlations with their corresponding Twitter streams. For trending queries, the percentage jumps to around 25%.
- Some categories of general queries are more likely to have such correlations. For example, for the 'Video Games' category 21.28% of the query trends are strongly correlated and the percentage is 14.20% for 'DVDs & Movies'. It is also more significant for trending queries. For example, the percentage is about 70% for queries from 'Sports Mem, Cards & Fan Shop'.
- For correlated pairs of streams, eBay stream lags Twitter stream and for trending queries and queries in categories such as 'Sports', the lag is more obvious and potentially useful for predictive tasks.
- Celebrities' popularities on Twitter correlate their search and sale trends on eBay. There is also signals that they have an impact on the prices of the related merchandise.
- Twitter trend is more peaky than eBay trend.

To summarize, we observe that e-commerce activity correlates well with social media yet lags it especially in certain domains like 'Sports' and 'DVDs & Movies'. This is more prominent when




(a) Giants' touchdowns drives high peaks on Twitter, followed by increases in eBay search volumes with lags. When the game is over, Giants as the winning team have peaks in both streams and the eBay peak lags Twitter.




(b) Patriots' touchdowns drives high peaks on Twitter, followed by increases in eBay search volumes with lags. The game over peaks are much smaller comparing with Giants' peaks.

Figure 7.13: Twitter and eBay trends for the keyword phrases 'giants' and 'patriots' from 23pm Feb 5 to 6am Feb 6 2012 GMT during the Super Bowl game.

users react to events and happenings. A possible reason which explains the correlation is that the domain of sports and entertainment is eventful and attracts eyeball, making it more likely to generate news of direct commercial values. A reason for the lags might lie in the nature of the two platforms, one is a social media platform which generates and delivers latest news fast while the other one serves the need for purchasing in response to the latest news. We believe that access to online social streams can enable near real-time merchandising of relevant products.

In future work, we plan to predict the sales on e-commerce platforms, according to signals in social media. Ultimately, the system will be able to detect events in social media, classify them into sales-driving and non-sales-driving events and estimate the sales of corresponding products on e-commerce platforms. With such a system, we can recommend relevant items for sellers and buyers

well in advance to increase the transactions on e-commerce platforms. This will involve techniques such as event detection, text mining and machine learning. A first step in this direction would be focusing on the sub-domains (e.g. 'Sports' and 'DVDs & Movies') of queries which demonstrate strong correlations and obvious lags, in order to build up a predictive model that can be generalized.

# CHAPTER 8

# SUMMARY AND CONCLUSIONS

In this thesis, we have studied the 'user-sensed data' generated by the users or their smart devices, aiming to distill credible real world insights out of it. In order to achieve that and make the process generalizable, we described a lightweight framework that assembles essential techniques that are commonly used across many data mining applications that deal with the user-sensed datasets, with a focus on their geo-temporal attributes. We then approach our meta research question: "to what extent can we mine the virtual world to study the real world?". As this is a relatively large topic, we broke it down into quadrants with regard to the subjects of study and the methodologies of study. In each quadrant, we picked one representative problem to address and the topics cover collective knowledge, natural phenomena, large-scale human behaviors and individual behaviors respectively. We created four novel data mining applications, with the framework as a reference.

In Chapter 2, we first surveyed representative work that guided or inspired our research. We then went through work specifically related to the four questions that we were trying to answer.

In Chapter 3, we organized the assembly of relevant techniques into the aforementioned framework to help fast prototyping of data mining projects on geo-temporal user-sensed data. The framework includes the following components: geo-temporal pattern extraction, comparison, clustering, visualization and external phenomenon detection. The techniques are later contextualized when being applied in the four data mining applications described in following chapters.

In Chapter 4, we discovered Flickr photo tag relationships by comparing their usage distributions over time and space, showing that these reveal connections between real world concepts in the physical world. By doing this, we are able to tell why the tags are connected, from a geo-temporal perspective. We clustered the tags based on their geo-temporal usage distributions and visualized the clusters. Evaluations showed that high quality geo-temporal semantics are extracted and the visualizations helped human users better recognize subtle semantic relationships between tags. This approach goes beyond the existing work that explores tag co-occurrences as a tag similarity measure. Potentially, it is useful for building tag suggestion systems as well as exploring any data having geographic and temporal annotations.

In Chapter 5, as opposed to extracting collective knowledge and comparing with subjective human judgments, we estimated natural phenomena and compared our results with large-scale fine grained ground truth data. We measured and quantified the occurrence of ecological phenomena including ground snow cover, snow fall and vegetation density by analyzing the textual and visual features of geo-tagged, time-stamped photos. We evaluated several techniques for dealing with the large degree of noise in the dataset, and improved the results by using machine learning to reduce errors caused by misleading tags and ambiguous visual content. This study can be a possible reference for other work that tries to accurately distill credible information from large, noisy social sharing datasets.

In Chapter 6, we explored individual mobile user behaviors in the human world. Previous research often utilized fine-grained and direct information such as time, location, web browsing and app usage information to infer about the users. We demonstrated that indirect and simple statistics such as entropy and frequency calculated from WiFi, Bluetooth, app usage, cell tower readings collected from the user's phone, can be useful in uniquely identifying the user. With this descriptive power, features can be further associated with user attributes to build personalized services.

In Chapter 7, we moved our focus from individual mobile users onto large-scale human behaviors and made one of the first efforts to bridge the research on social media user behaviors and e-commerce user behaviors. We showed that there are signals from social media for predicting aggregate consumer behavior. We demonstrated that about 5% of general eBay query streams strongly correlate with the corresponding Twitter mention streams, while the percentage jumps higher for trending eBay query streams and some specific eBay query categories. We also discovered that eBay trend lags Twitter. We showed that celebrities' popularities on Twitter correlate well with their relevant search and sales on eBay. The correlations and lags provide predictive insights for future applications on instant merchandising for both sellers and e-commerce platforms. For example, we can potentially predict what merchandise is going to be trending on e-commerce platforms using social media signals and inform relevant sellers or buyers about what to sell or buy ahead of time. In this way, the e-commerce platforms could increase their transactions.

## 8.1 POTENTIAL PITFALLS

The applications that mine the user-sensed datasets, including the applications that we described in this thesis, can fall into potential pitfalls due to design flaws and changes in the data.

We take Google Flu Trends (GFT) as an example. As described in Section 2.1, it finds the search engine queries whose geo-temporal usage distributions correlate well with the estimates of influenza-like illness activity released by the United States Centers for Disease Control and Prevention (CDC) and aggregate their usage data to generate the estimates two weeks ahead of the release of the official data. Though it is hard to define the 'ground truth for flu activity', the CDC estimates can be considered systematic and reliable – CDC collects and compiles nationwide physician visit data to produce the percentage of physician visits in which a patient presents with influenza-like symptoms. Google Flu Trends does not always produce accurate estimates and sometimes they can be off by large margins. For example, in 2009 summer, it underestimated the outbreak of flu where

GFT's estimate is half the official estimate [21]. In January 2013, it overestimated the outbreak where GFT's estimate doubles the official estimate [13, 67].

There has been a lot of discussions on the possible causes for these errors. As we mentioned earlier in Section 2.1, there could be some initial design flaws including the possible overfitting problem. The original GFT algorithm selected only 45 search queries out of 50 million candidates as best matches to fit 1152 CDC data points and the researchers also reported that noisy queries related to 'high school basketball' appeared in the top 100 matches [38, 67]. Another reason is the user search behavior change. The underestimation in 2009 happened during summer which was an unusual season for flu outbreaks and the users appeared to use different terms when they were seeking flu-related information. The original GTF algorithm was trained on data with regular seasonal flu outbreak patterns and failed to adapt to the this user behavior change [21]. While it is hard to predict the changes user behavior, an agile system should be updated often and could alert the researchers about the errors and changes. One other possible source of noise is the service provider itself. The companies changes their services as parts of their business plans. However, when they do so, the non-profitable research projects that rely on their data might be greatly affected, with or without notifications. For example, in 2011, Google introduced related search terms to their search result pages and similarly, in 2012, it started to return diagnoses for search terms containing flu symptoms [67]. When the user clicks on the suggested terms, noise may brought into the GFT analytics as these were not the user's initial intentions.

The design flaws and potential changes in these user-sensed datasets may all result in pitfalls for the data mining applications including ours introduced in this thesis. For example, two of our projects rely on the user-sensed data from Flickr. There could also be user behavior changes and service changes and its user distribution and general popularity are not under our control.

## 8.2 VISION FOR FUTURE WORK

To reduce the risks of failure and even avoid these pitfalls, the applications need to be designed agile and robust, apart from recalibrating the applications once in a while. As we mentioned in Chapter 5, incorporating data from multiple sources may help us build such applications. First of all, the application is designed based on data from multiple sources instead of one single source. If there are changes in one source, the application may still rely on other sources. Secondly, among these sources, we can include the sources that are more scientifically reliable with greater transparency and public availability. This potentially helps us better calibrate the applications. Here we explicitly list the sources that we believe are beneficial to include:

1. Data from multiple social media platforms
2. Conventional data
3. Volunteer-based crowd-sourced data

We will discuss them in the following subsections.

### 8.2.1 DATA FROM MULTIPLE SOCIAL MEDIA PLATFORMS

On one hand, as mentioned above, multiple sources may increase the data density and make the application less sensitive to the changes in a single source. On the other hand, as we discussed in Chapter 5, multiple social media platforms provide user-sensed data with similar attributes and it may ease the difficulty in integrating the data. For example, timestamped text and images are usually available from platforms such as Flickr, Twitter, Tumblr and Instagram and photo tags and tweet hashtags are both short keyword phrases that could be comparable. Besides, most of these services can record geolocations. They can all potentially contribute to geo-temporal user-sensed datasets.

However, we shall still be careful in integrating data from multiple sources to avoid introducing new bias and noises. For instance, a same person may be users on many social media platforms and

we may need to de-anonymize users across different social media platforms [62] to prevent a user from spamming the datasets. More systematic approaches to integrating the heterogeneous web data (data fusion) may be applied for this purpose [11].

### 8.2.2 CONVENTIONAL DATA

The conventional data is compiled and released by domain experts at research institutes or governments. Some examples would be the CDC influenza-like illness activity data as we mentioned in this chapter, the NASA satellite data and NOAA ground weather station data as we introduced in Chapter 5 and the New York City Open Data released by the government[1]. In spite of the limitations such as expenses and area of coverage mentioned in Chapter 1, it is usually consistent and less likely to have bias and noises, comparing with the user-sensed data. We consider it more scientifically reliable the user-sensed data.

Incorporating this data may help improve the performance of the applications and make them more tolerant to changes in these user-sensed datasets. It has been suggested that the GFT could incorporate lagged CDC data into the regression model and the results are better than that from the original GFT or the model based on the CDC data alone [67]. With regard to our projects discussed in this thesis, the conventional data could also help. For example, the regional year-round snow data may help us balance out some of the bias in the user behavior. In a place where it seldom snows, traces of snow may get the users excited and taking snow photos.

### 8.2.3 VOLUNTEER-BASED CROWD-SOURCED DATA

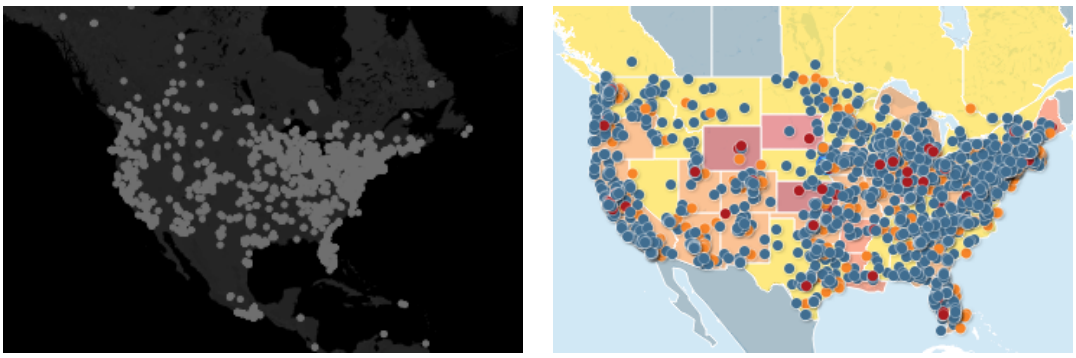
Another possible source is the volunteer-based crowd-sourced data with open access from platforms such as eBird[2] and Flu Near You[3]. eBird, initiated by the Cornell Lab of Ornithology and National

[1]The New York City government releases various public datasets generated by government agencies and City organizations to improve the government and facilitate research. The data covers meta topics such as city government, education, transportation, business, recreation, environment and health. It is accessible to the public at `https://nycopendata.socrata.com`.

[2]`http://ebird.org`

[3]`https://flunearyou.org`

Audubon Society in 2002, is a web platform allowing recreational and professional bird watchers to report and access the information about the presence of birds. It thus provides geo-temporal bird distributions at a continental scale. It is reported that its participants contributed over 3.1 million observations in North America in March 2012[4]. Figure 8.1(a) is a snapshot for the eBird website that shows the distribution of the bird presence in North America on September 2, 2014. This not only only benefits the bird watching community, but also helps research in biology and ecology. Flu Near You is initiated by Boston Children's Hospital in partnership with the American Public Health Association and the Skoll Global Threats Fund to allow the public in North America to report information related with flu weekly. Figure 8.1(b) is a snapshot from the Flu Near You website that demonstrates the distribution of flu reports in the US during the week ending August 31, 2014. It helps the public to track and prevent the flu and also provides data to relevant research communities.



(a) (b)

Figure 8.1: Sample geo-spatial data distributions from eBird and Flu Near You. (a) the distribution of the bird presence in North America on September 2, 2014 from the eBird website, (b) the distribution of flu reports in the US during the week ending August 31, 2014 from the Flu Near You website. Red dots: influenza-like illness; orange dots: symptoms; blue dots: no symptoms.

This data is contributed by citizen volunteers and its characteristics are somewhat in between of the social media data and the conventional data. It is similar to the social media data in a sense that it is crowd-sourced. However, as opposed to the passively collected data from social media, this data is generated by volunteers who are more aware that they are contributing the data to the

[4] `http://ebird.org/content/ebird/about/`, About eBird.

corresponding communities. They may thus take more responsibility and some of them have more expertise in the relevant topics than the common social media users. This makes the data quality better compared with the social media data though not necessarily as good as the conventional data. Besides, its crowd-sourced nature may make up for the shortcomings of the conventional data. For example, the volunteer-based crowd-sourced data is usually updated real time compared with the conventional data that often takes longer time to compile and publish. Meanwhile, this data has far less limitations on the labor resource thus may have much larger coverage.

## 8.3 EXPECTATIONS FOR THE INDUSTRY

As we have been mentioning throughout this thesis, a relatively large portion of the user-sensed data comes from the industry. If the industry can work closely with the research communities, many tasks would be made much easier and less prone to errors. The companies such as Twitter and Flickr are already taking up their social responsibility – they share some of their public data at no cost via their APIs to help application developers and researchers. Our research projects have benefited from the data shared by the industry. But we think that there is still space for improvement in the data sharing process such that various data mining projects can be robust and better avoid the pitfalls that we discussed above. Here we want to advocate the expectations for the industry with regard to:

1. Data quantity
2. Data standardization
3. Data transparency

We will discuss them in the following paragraphs.

***Data quantity.*** We hope the industry can put less limitations on the amount of public data being shared. Currently, there are often restrictions on the portion of public data from the Internet companies that can be accessed at no cost. For example, though Twitter's sample streaming API can be

accessed for free, it provides only a 1% sample of the public tweets being generated on Twitter. 1% of the all the tweets being generated per day is about 2 million. This amount may not always be enough for the research projects. In the temporal user behavior study project described in Chapter 7, to ensure the density of the time series vectors that we extracted, we had to put a threshold on the daily Twitter mentions of our keywords. This greatly reduced the number of candidate keywords that we wanted to track originally. If there were more data, these experiments would have been performed at a larger scale. We encourage the companies to make more public data accessible – though more does not always mean better, it does help in many cases.

***Data standardization.*** We hope that different companies could have standardized APIs for accessing and parsing their data. Currently, different companies have different APIs though most of their functionalities are similar. For example, Twitter has the Twitter Search API for retrieving tweets relevant to the query and Flickr also has an API for finding a list of photos that match the search criteria. But they have different authentication methods and API request formats. Besides this variety, these APIs change from time to time. This creates overhead for developers and researchers – time and effort is spent on looking up various documents as well as adjusting the applications to the API changes. Though unifying all the APIs with a single standard may be hard or unrealistic, we will definitely be glad to see some parts of the API call processes be unified, such as authentication method, request format and return data format.

***Data transparency.*** We hope that the companies can provide more details about the data which would help researchers better understand the data. For instance, it is not clearly documented by Twitter how the 1% data from its streaming API is sampled from all the data being generated. There have been studies on the characteristics of the sampled data trying to figure out whether this sample can be representative for the overall user activity on Twitter [55, 77]. As the data is the ground for many applications and studies, we believe it is important for the companies to give more details on how the data is generated.

Besides, we expect they could notify the researchers about the service changes that may potentially affect the data being provided. Service changes may change the user behavior thus bring in changes to the user-sensed data being generated. One example would be the Google service changes as we mentioned in Section 8.1. They may lead a user to use search terms that was not the user's original intention. This may have affected the GFT's estimates based on users' search behavior. Therefore, it will be helpful if developers and researchers can be notified about the changes in a timely manner.

## 8.4 CONCLUSIONS

With abundance of the observational data generated by the users via the booming social webs and mobile devices, we became interested in this meta research question: "to what extent can we mine the virtual world to study the real world?". To approach it, we divided all the possible problems in this domain into quadrants with our criteria. For each quadrant, we picked one representative problem and addressed them with novel solutions. With these efforts, we showcased the vast possibilities of mining the user-sensed data for real world insights. In summary, we (1) showed that geo-temporal relationships between photo tags reveal real world concept connections; (2) quantified ecological phenomena and compared our estimates with ground truth data at a continental scale; (3) demonstrated that very simple behavioral statistics can characterize and identify mobile users; (4) discovered that social media signals are useful for predicting aggregate consumer behavior. Besides these, we assembled techniques for dealing with geo-temporal user-sensed data and they were contextualized in the aforementioned data mining applications on multiple representative datasets – we hope this thesis can serve as a guiding reference for future data mining applications on large, noisy, biased, geo-temporal user-sensed datasets in order to distill credible real world values. For future work, we hope to incorporate data from various sources in order to create robust and agile applications. We are confident that with the continuous and joint efforts from research communities,

industry and end users, there will be more and more exciting and even game-changing applications that enhance quality of life and push the boundaries of our knowledge.

# BIBLIOGRAPHY


[1] http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/.

[2] The Funf Open Sensing Framework. http://www.funf.org/.

[3] Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Hui-I Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10. ACM, 2007.

[4] Sajid Anwar, Robert McMillan, and Mingli Zheng. Bidding behavior in competing auctions: Evidence from eBay. *European Economic Review*, 50(2):307–322, 2006.

[5] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.

[6] Louise Barkhuus and Anind K Dey. Location-based services for mobile telephony: a study of users' privacy concerns. In *INTERACT*, volume 3, pages 702–712. IOS Press, 2003.

[7] Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop*, pages 15–33, 2006.

[8] Donald J Berndt and James Clifford. Using Dynamic Time Warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370, 1994.

[9] Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, and Danny Weitzner. Creating a science of the web. *Science*, 313(5788):769–771, 2006.

[10] Claudio Bettini, X Sean Wang, and Sushil Jajodia. Protecting privacy against location-based personal identification. In *Secure Data Management*, pages 185–199. Springer, 2005.

[11] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1, 2008.

[12] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[13] Declan Butler. When Google got flu wrong. *Nature*, 494(7436):155, 2013.

[14] John W Byers, Michael Mitzenmacher, and Georgios Zervas. Daily deals: Prediction, social diffusion, and reputational ramifications. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 543–552. ACM, 2012.

[15] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, Ronald A Peterson, Hong Lu, Xiao Zheng, Mirco Musolesi, Kristóf Fodor, and Gahng-Seop Ahn. The rise of people-centric sensing. *IEEE Internet Computing*, 12(4):12–21, 2008.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[17] Ling Chen and Abhishek Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[18] Steve Chien and Nicole Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th international conference on World Wide Web*, pages 2–11. ACM, 2005.

[19] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Proceedings of the 15th Annual International Symposium on Wearable Computers*, pages 29–36, 2011.

[20] Hyunyoung Choi and Hal Varian. Predicting the present with Google trends. *Economic Record*, 88(s1):2–9, 2012.

[21] Samantha Cook, Corrie Conrad, Ashley L Fowlkes, and Matthew H Mohebbi. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PloS one*, 6(8):e23610, 2011.

[22] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[23] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World Wide Web*, pages 761–770. ACM, 2009.

[24] Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "OMG, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the International Workshop on Location Based Social Networks*, pages 73–80, 2009.

[25] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. Predicting personality using novel mobile phone-based metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 48–55. Springer, 2013.

[26] Michel-Marie Deza and Elena Deza. *Dictionary of distances*. Elsevier, 2006.

[27] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R1-PCA: Rotational invariant L1-norm Principal Component Analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288. ACM, 2006.

[28] Trinh Minh Tri Do and Daniel Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 163–172. ACM, 2012.

[29] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. *ACM Transactions on the Web*, 1(2):7, 2007.

[30] Maeve Duggan and Joanna Brenner. The demographics of social media users – 2012. http://pewinternet.org/Reports/2013/Social-media-users.aspx. Accessed: 2013-11-05.

[31] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[32] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[33] Vladimir Estivill-Castro and Ickjai Lee. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *Proceedings of the 6th International Conference on Geocomputation*, pages 24–26, 2001.

[34] Vladimir Estivill-Castrol and Alan T Murray. Discovering associations in spatial data—an efficient medoid based approach. In *Research and Development in Knowledge Discovery and Data Mining*, pages 110–121. Springer, 1998.

[35] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[36] Flickr. Flickr: Camera Finder. http://www.flickr.com/cameras/. Accessed: 2013-11-05.

[37] Nikhil Garg and Ingmar Weber. Personalized tag suggestion for Flickr. In *Proceedings of the 17th international conference on World Wide Web*, pages 1063–1064. ACM, 2008.

[38] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[39] Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

[40] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[41] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[42] Krzysztof M Gorski, Eric Hivon, AJ Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005.

[43] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[44] Stephen Guo, Mengqiu Wang, and Jure Leskovec. The role of social networks in online shopping: information passing, price of trust, and consumer choice. In *Proceedings of the 12th ACM conference on Electronic commerce*, 2011.

[45] Jonna Häkkilä and Craig Chatfield. 'it's like if you opened someone else's letter': user perceived privacy and social practices with SMS communication. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pages 219–222. ACM, 2005.

[46] Dorothy K Hall, George A Riggs, and Vincent V Salomonson. MODIS/Terra Snow Cover Daily L3 Global 0.05Deg CMG V004. Boulder, CO, USA: National Snow and Ice Data Center, 2011, updated daily.

[47] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[48] James Hays and Alexei A Efros. IM2GPS: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[49] Daniel Houser and John Wooders. Reputation in auctions: Theory, and evidence from eBay. *Journal of Economics & Management Strategy*, 15(2):353–369, 2006.

[50] Yan Huang and Pusheng Zhang. On the relationships between clustering and spatial co-location pattern mining. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, pages 513–522. IEEE, 2006.

[51] R.W.G. Hunt. *Measuring Colour*. London: Fountain Press, 1998.

[52] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007.

[53] Xin Jin, Andrew Gallagher, Liangliang Cao, Jiebo Luo, and Jiawei Han. The wisdom of social multimedia: using flickr for prediction and forecast. In *Proceedings of the international conference on Multimedia*, pages 1235–1244. ACM, 2010.

[54] George H John and Pat Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[55] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1% s don't make a whole: Comparing simultaneous samples from twitter's streaming api. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 75–83. Springer, 2014.

[56] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *Proceedings of the 18th international conference on World Wide Web*, pages 801–810. ACM, 2009.

[57] Amy K Karlson, AJ Brush, and Stuart Schechter. Can i borrow your phone?: understanding concerns when sharing mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1647–1650. ACM, 2009.

[58] George Karypis and Vipin Kumar. Parallel multilevel k-way partitioning scheme for irregular graphs. In *Proceedings of the 1996 ACM/IEEE Conference on Supercomputing*. IEEE Computer Society, 1996.

[59] Owen Kaser and Daniel Lemire. Tag-cloud drawing: Algorithms for cloud visualization. *arXiv preprint cs/0703109*, 2007.

[60] Maurice G Kendall and Alan Stuart. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Hafner Publishing Company, 1961.

[61] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How Flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th international conference on Multimedia*, pages 631–640. ACM, 2007.

[62] Mohammed Korayem and David J Crandall. De-anonymizing users across heterogeneous social computing platforms. In *Proceedings of 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[63] Kay Kremerskothen. 6,000,000,000. http://blog.flickr.net/en/2011/08/04/6000000000/. Accessed: 2014-08-03.

[64] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, pages 591–600. ACM, 2010.

[65] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011.

[66] Land Processes Distributed Active Archive Center. MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V005. Sioux Falls, SD: U.S. Geological Survey., 2011.

[67] David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[68] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. MoodScope: building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402. ACM, 2013.

[69] Bo Liu, Quan Yuan, Gao Cong, and Dong Xu. Where your photo is taken: Geolocation prediction for social images. *Journal of the Association for Information Science and Technology*, 65(6):1232–1243, 2014.

[70] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. In *Proceedings of the 18th international conference on World Wide Web*, pages 351–360. ACM, 2009.

[71] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 165–178. ACM, 2009.

[72] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

[73] Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.

[74] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World Wide Web*, pages 641–650. ACM, 2009.

[75] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Pulse of the nation: U.S. mood throughout the day inferred from Twitter. http://www.ccs.neu.edu/home/amislove/twittermood/.

[76] Matt Mohebbi, Dan Vanderkam, Julia Kodysh, Rob Schonberger, Hyunyoung Choi, and Sanjiv Kumar. Google correlate whitepaper. http://www.google.com/trends/correlate/whitepaper.pdf, 2011. Accessed: 2013-11-05.

[77] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[78] Emily Moxley, Jim Kleban, and B. S. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from Flickr. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 24–30. ACM, 2008.

[79] Emily Moxley, Jim Kleban, Jiejun Xu, and BS Manjunath. Not all tags are created equal: Learning flickr tag semantics for global annotation. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, pages 1452–1455. IEEE, 2009.

[80] Andrew Y Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[81] Raymond T. Ng and Jiawei Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 144–155, 1994.

[82] Axel Ockenfels and Alvin E Roth. Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the internet. *American Economic Review*, 92(4), 2002.

[83] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of 4th International AAAI Conference on Weblogs and Social Media*, volume 11, pages 122–129, 2010.

[84] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *International Conference on Language Resources and Evaluation*, 2010.

[85] Nish Parikh and Neel Sundaresan. Scalable and near real-time burst detection from eCommerce queries. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[86] Martin L Parry, Osvaldo F Canziani, Jean P Palutikof, Paul J van der Linden, and Clair E Hanson. *IPCC, 2007: Climate Change 2007: Impacts, Adaptation, and Vulnerability*. Cambridge University Press, 2007.

[87] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010.

[88] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World Wide Web*, pages 337–346. ACM, 2011.

[89] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM, 2012.

[90] Kira Radinsky and Eric Horvitz. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM, 2013.

[91] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM, 2007.

[92] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860. ACM, 2010.

[93] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998.

[94] Mehmet Sayal. Detecting time correlations in time-series data streams. Hewlett-Packard Company, 2004.

[95] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing Flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

[96] Harshit S Shah, Neeraj R Joshi, Ashish Sureka, and Peter R Wurman. Mining eBay: Bidding strategies and shill detection. In *WEBKDD 2002-Mining Web Data for Discovering Usage Patterns and Profiles*, pages 17–34. Springer, 2003.

[97] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266. ACM, 2008.

[98] Choonsung Shin, Jin-Hyuk Hong, and Anind K Dey. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 173–182. ACM, 2012.

[99] Börkur Sigurbjörnsson and Roelof Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336. ACM, 2008.

[100] Gyanit Singh, Nish Parikh, and Neel Sundaresan. Rewriting null e-commerce queries to recommend products. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 73–82. ACM, 2012.

[101] Gyanit Singh, Nish Parikh, and Neel Sundaresn. User behavior in zero-recall ecommerce queries. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 75–84. ACM, 2011.

[102] Vivek K Singh, Mingyan Gao, and Ramesh Jain. Social pixels: genesis and evaluation. In *Proceedings of the international conference on Multimedia*, pages 481–490. ACM, 2010.

[103] Michael F Squires and Jay H Lawrimore. Development of an operational northeast snowfall impact scale. NOAA National Climatic Data Center, 2006.

[104] Jiang Su, Harry Zhang, Charles X Ling, and Stan Matwin. Discriminative parameter learning for Bayesian networks. In *Proceedings of the 25th international conference on Machine learning*, pages 1016–1023. ACM, 2008.

[105] Catherine A Sugar and Gareth M James. Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463), 2003.

[106] James Surowiecki. *The wisdom of crowds*. Random House LLC, 2005.

[107] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[108] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[109] Twitter. Audiences on Twitter. https://business.twitter.com/audiences-twitter/. Accessed: 2013-11-05.

[110] Srinivas Vadrevu, Ya Zhang, Belle Tseng, Gordon Sun, and Xin Li. Identifying regional sensitive queries in web search. In *Proceedings of the 17th international conference on World Wide Web*, pages 1185–1186. ACM, 2008.

[111] Ismael A Vergara, Tomás Norambuena, Evandro Ferrada, Alex W Slater, and Francisco Melo. StAR: a simple tool for the statistical comparison of ROC curves. *BMC bioinformatics*, 9(1):265, 2008.

[112] Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 131–142. ACM, 2004.

[113] Jens Weppner and Paul Lukowicz. Collaborative crowd density estimation with mobile phones. In *Proceedings of ACM PhoneSense*, 2011.

[114] Ryen W White, Susan T Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 132–141. ACM, 2009.

[115] Karen Wickre. Celebrating #twitter7. https://blog.twitter.com/2013/celebrating-twitter7/. Accessed: 2014-08-03.

[116] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 31–40. ACM, 2008.

[117] Xiangye Xiao, Xing Xie, Qiong Luo, and Wei-Ying Ma. Density based co-location pattern discovery. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 29. ACM, 2008.

[118] Jun-Ming Xu, Aniruddha Bhargava, Robert Nowak, and Xiaojin Zhu. Socioscope: Spatio-temporal signal recovery from social media. In *Machine Learning and Knowledge Discovery in Databases*, pages 644–659. Springer, 2012.

[119] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th international conference on extending database technology*, pages 259–270. ACM, 2011.

[120] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. Smartphone bluetooth based social sensing. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 95–98. ACM, 2013.

[121] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[122] Haipeng Zhang, Mohammed Korayem, David J Crandall, and Gretchen LeBuhn. Mining photo-sharing websites to study ecological phenomena. In *Proceedings of the 21st international conference on World Wide Web*, pages 749–758. ACM, 2012.

[123] Haipeng Zhang, Mohammed Korayem, Erkang You, and David J Crandall. Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 33–42. ACM, 2012.

[124] Haipeng Zhang, Nish Parikh, Gyanit Singh, and Neel Sundaresan. Chelsea won, and you bought a t-shirt: Characterizing the interplay between Twitter and e-commerce. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 829–836. ACM, 2013.

[125] Haipeng Zhang, Zhixian Yan, Jun Yang, Emmanuel Munguia Tapia, and David J Crandall. mFingerprint: Privacy-preserving user modeling with multimodal mobile device footprints. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 195–203. Springer, 2014.

# Haipeng Zhang

## Education

Ph.D. Computer Science, Indiana University, Bloomington, US, 2014. (Advisor: Prof. David J. Crandall)

M.S. Computer Science, Indiana University, Bloomington, US, 2012.

B.E. Software Engineering, Nanjing University, China, 2009.

Undergraduate Exchange Program, Hong Kong University of Science and Technology, China, 2007 Fall.

## Research Interests

Data Mining. I am particularly interested in discovering the connections between the user generated contents and the real world.

## Employments

Research Intern, Samsung Research America, San Jose, US, 2013 Summer.

Research Intern, Microsoft Research Cambridge, UK, 2013 Spring.

Research Intern, eBay Research Labs, San Jose, US, 2012 Summer.

Research Assistant, Indiana University, Bloomington, US, Jan 2011 to Jul 2014.

Associate Instructor, Indiana University, Bloomington, US, Aug 2009 to Dec 2010.

Research Intern, National Institute of Informatics, Tokyo, Japan, 2010 Summer.

Intern, eBay China Development Center, Shanghai, China, 2008 Summer.

## Publications

**Haipeng Zhang**, Mohammed Korayem, Erkang You and David J. Crandall, **Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities**, in WSDM2012.

**Haipeng Zhang**, Mohammed Korayem, David J. Crandall and Gretchen Lebuhn, **Mining Photo-sharing Websites to Study Ecological Phenomena**, in WWW2012.

**Haipeng Zhang**, Nish Parikh, Gyanit Singh and Neel Sundaresan, **Chelsea Won, and You Bought a T-shirt: Characterizing the Interplay Between Twitter and E-Commerce**, in ASONAM2013.

**Haipeng Zhang**, Zhixian Yan, Jun Yang, Emmanuel Munguia Tapia and David J. Crandall, **mFingerprint: Privacy-Preserving User Modeling with Multimodal Mobile Device Footprints**, in SBP2014.

## Patents

Jun Yang, Zhixian Yan, Lu Luo, Xuan Bao, **Haipeng Zhang** and Emmanuel Munguia Tapia, **Method and Systems for On-device Social Grouping**, US patent pending.