Title

Student Name, Major; Mentor Name, Department

Background and Research Questions

It is common to see holiday-oriented events planned weeks or months in advance of the holiday (Fotis et. al, 2012). Furthermore, product advertising often begins far before a holiday. For example, a Halloween sale might take place in mid-August. Many organizations plan events carefully around the holidays. However, while it may be clear when the holiday itself occurs, it is not obvious when the general celebration of that holiday begins. For some holidays, the celebration is just that: a day. Other holidays, on the contrary, are seasons that take place over many days or weeks. While ample research exists on some of the questions and observations surrounding holidays, other questions and understandings still lie unexplored (Barber á et. al, 2015; Kouloumpis et. al, 2011). Of particular interest is how holiday celebrations are affected by demographics and geographic regions (Agarwal et. al, 2011). Traditional approaches tend to use polling data to investigate when people express interest about holidays the most. This is unreliable because there is no way to ask people whether they are thinking about a given holiday, as the question itself makes the answer invalid by reminding the person of that holiday. Social media data such as twitter usage has emerged as a new data source for studying social phenomena. There is little research into twitter data concerning holidays. Since ever more people are using social media to share their opinions and activities, the analysis of social media data can reveal interesting patterns to help better understand human behavior related to holidays (Kumar, 2013). This project will process and analyze billions of geo-tagged tweets (spatial Big Data) posted by millions of Twitter users across the US to answer the following **research questions**:

1. How much before and after holidays do Twitter users tweet about them?
2. Where do these tweets originate?
3. From what type of person (demographics, social economics) do these tweets originate?
4. What is the relationship between the above questions? Namely, is there interaction among "when", "who", and "where"?

Present State of Knowledge and Preliminary Study

Though holidays are a well-researched phenomenon, to the best of our knowledge, no studies investigate the trends in holiday celebration using twitter data. The closest study which examines the subject of this project was carried out by Fotis et al. (2012). The authors concluded that for those holidays involving the largest amounts of travel (Christmas, Thanksgiving, July 4th, and New Year) it is unwise to plan events 1-2 days before the date in question since many people are engaging in holiday-related travel. Another related previous finding is that twitter users who tweet about the same things tend to tweet about the same other topics (Barber á2013). This finding indicates that research could be done on specific demographics within twitter (i.e., furries, sports fans) to learn more about those distinct communities. Toward that end, we will investigate how demographics affect holiday celebration patterns. In relation to this project, I have conducted some preliminary twitter data processing and analysis during my internship (supervised by [mentor]). Figure 1(left) shows the spatial distribution of the geo-tagged tweets mentioning Christmas in December, 2016. Figure 1(right) shows the temporal trend of the Christmas-related tweets.

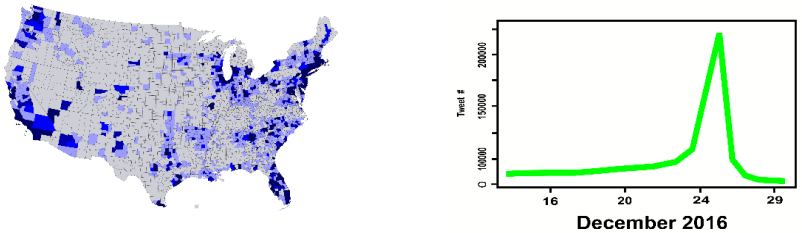

Figure 1. Demonstration of the preliminary results. Left: Spatial distribution of Christmas-related tweets (tweets clustered along coastal regions); Right: temporal trend of Christmas-related tweets (tweets volume peaked on Christmas day).

Project Goals and Objectives

This study is part of a broader research trend of gaining information about society through the analysis of twitter data. The goal of this project is to identify potential spatiotemporal patterns on the geography, demographics, and nature of the timeline of how Americans celebrate holidays. It explores *when* people tweet about specific holidays, *where* they are tweeting, and *who* is tweeting. First, this project aims to model the temporal trend of tweeter activities for fourteen US holidays. For example, based on our preliminary research, interest in Christmas does not occur substantially until December 1st. Nonetheless, many Christmas events take place in November. This kind of planning error could easily be avoided with proper research. Second, I will investigate where these tweets are coming from in order to gain a better understanding of the regional differences in holidays celebration and the degree to which they are celebrated. For instance, it may be reasonable to start a Christmas celebration in late November in some parts of the country, but not in others. Third, I will investigate the demographics of who is posting these tweets by comparing the data from the spatial analysis to census data of the demographics of those regions. This could allow us to get an understanding of which demographics celebrate certain holidays

at which point in time. For instance, we may discover that older people celebrate Christmas earlier, which would affect how marketing companies target specific generations.

Project Significance

While previous studies have explored the relationship between social media and holidays, few have had such a large scope in data volume (over 1 billion geo-tagged tweets), population sample (over 3 million twitter users), time period (one year on a daily basis), and geographic specificity (longitude and latitude of each tweet and twitter user). This study will process billions of tweets and analyse them temporally, spatially, and statistically. The patterns identified in this project can help us better understand how holidays are celebrated in the U.S. regarding the questions when, where, and who. The findings may also be of interest to the entertainment industry as often the economic activity surrounding holidays is a major component of an entertainment company's balance sheet. As such, it is vital that the marketing efforts of these companies utilize this type of research to best enhance holiday sales. Finally, the proposed methodology would contribute to literature in computational social science where social media data is being increasingly used for societal studies.

Project Design

The project consists of four steps, including data query and extraction, temporal trend analysis, spatial pattern analysis, and correlation analysis.

Step 1: Data Query and Extraction: Over one billion geo-tagged tweets have been retrieved through the Twitter Stream Application Programming Interface (API) in 2016. These streamed tweets were handled, stored, and managed in a high-performance computing (HPC) cluster (operated by [mentor name]). The first step is to query against billions of tweets with SQL using spatiotemporal criteria (latitude, longitude, time) and keywords in the tweet message and hashtags to extract the holiday-related tweets. I will search for multiple possible forms of tweets for each holiday, for instance, using "October 31st", "Halloween", or "Sp00k day" to extract Halloween-related tweets.

Step 2: Temporal Trend Modeling: With the extracted data, I will gather the number of tweets relating to a given holiday on each day before and after that holiday and create a statistical model for the tweet frequency. This is performed by using a function which creates a regression equation for the data. The models will involve stepwise functions oriented around the day itself, as the tweet trends follow different models after and before the main day of the holiday. After creating these equations, I will plot them to create a graph which can predict the level of interest in a given holiday based on the date.

Step 3: Spatial Pattern Analysis: Since each tweet contains location information (latitude, longitude), I will aggregate the tweets at the county level for each selected holiday. For example, I will count the number of Christmas-related tweets in each county of the US. The aggregation result for each holiday will be first mapped using ArcMap and relevant scripting language (R) for automation (e.g. Figure 1.left). Next, spatial analysis techniques such as hot sport analysis and cluster analysis will be applied to further identify the potential spatial patterns from the density of tweets.

Step 4: Correlation analysis: In this step, I will link the aggregated tweets (county level tweets density) with the county-level US census data, and analyze the relationship between tweets density and the demographics. Ordinary Least Squares (OLS) regression and Geographically Weighted Regression (GWR) will be used in this analysis. This allows us to map twitter trends for demographics. For example, we may find that county which are predominately Hispanic are more likely to have twitter users who tweet about "*The Day of the Dead*" (a traditional Mexican holiday).

Anticipated Results and Dissemination

We anticipate that a variety of spatiotemporal patterns will be identified in this study, which can help us gain a greater understanding of the complicated timeline and nature of American holiday celebration. For example, the holiday season for a given holiday may begin approximately one month before a given holiday and slowly rise until it peaks, and then quickly dissipate thereafter. The religious holidays will peak more in the South and rural areas, than the North. This will most likely hold more true for Christmas than Easter.

I will present the project results at Discover USC and publish an article in Caravel journal. In collaboration with my mentor, I will also submit a manuscript to a peer-reviewed journal (*ISPRS International Journal of Geo-Information*).

Project Timeline

Months (8 Total)	May	June	July	August	September	October	November	December
Data Query/Extraction								
Temporal Analysis								
Spatial Analysis								
Correlation Analysis								
Report/Presentation/ Paper								

Personal Statements: For most of my life I have been an avid user of social media, whether through my Youtube channel or my twitter account. I have always been attracted to this "world." Some of my more popular videos are on statistics and big data analysis of Youtube channels. As an avid user of social media, this is not just a research opportunity to further my career. I see this is an opportunity for me to learn about my community and aspects of my culture.

Reference

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.

2. Barbera, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23, 76-91.

3. Fotis, J., Buhalis, D., & Rossides, N. (2012). Social media use and impact during the holiday travel planning process (pp. 13-24). Springer-Verlag.

4. Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, 11(538-541), 164.

5. Kumar, S., Morstatter, F., & Liu, H. (2013). *Twitter data analytics*. Springer Science & Business Media. (pp. 13-24). Springer-Verlag.

Magellan Scholar BUDGET FORM

Student's Name:

Student salary	Hours	Rate	Subtotal
	Estimated number of hours student will work	Enter the hourly wage	
Research hours during semesters when enrolled in classes	280	9.5	$2,660.00
Research hours during semesters when NOT enrolled in classes	40	9.5	$380.00
Fringe: Student salary * student fringe rate[1] (what is fringe? See budget instructions or guidebook)			
Enrolled in classes	$2,660.00	0.55%	$14.63
Not enrolled in classes	$380.00	8.29%	$31.50
Materials/Supplies		Enter sub-total from below:	$0.00
Travel		Enter sub-total from below:	$0.00
		TOTAL:	**$3,086.13**
		Amount requested for Scholar award:	**$3,000.00**

Student Salary:
The student will carry out research during the Spring, Summer, and Fall 2017 semesters working 10 hours per week for 32 weeks at a rate for $9.00 per hour.

Materials/Supplies:
All of the materials and supplies needed for this project will be paid for by the mentor's research funds. The money awarded by the Magellan grant will go towards student salary to perform the research.

Travel:
Not requested.