**[Tentative Title][A Study of Tourist Mobility Patterns]**

**Project Proposal Outline**

**Background and Motivation**

Throughout the course we have reiterated time and again the increasing amount of data being generated about the world around us. Insights about such data can lead to further understand the dynamics of various industries and markets.

Increasingly, geo-tagged social media data allow us to perform quantitative studies of such industries. For this project, we are interested in studying the tourism industry leveraging different kinds of social media data available. We are interested in using social media data obtained from twitter, travel sites like Travel Blog and Trip Advisor, and potentially others to analyze the mobility patterns of tourists.

In several regions like Greece, Japan, and Hawaii, tourism is a critical component of the economy. Insights obtained from tourist mobility patterns and the resulting geo-spatial network of destinations can provide stakeholders with useful knowledge regarding the overall touristic industry dynamics on the region. Increasingly, countries like Japan are looking to big data techniques to discover behavioral patterns as potential means of optimizing outreach, increasing the overall volume of incoming tourists, and enhancing the tourist experience.

**Project Objectives**

A critical component of the project is to develop a methodology for mapping and understanding tourist flows within a region using network models. The project’s core goal is to construct a geo-spatial network corresponding to tourist flows within a region. Statistical analysis of this network will be performed to provide insights into relevant questions like:

Given a network of city-by-city tourist flows in a given country, centrality measures and a clustering of such network would reveal important tourist hubs as well as potential transportation issues that might arise given a certain volume of tourists (add).

**Objectives**

As a first objective, we will develop a city-by-city network of tourist flows in a given region. We will further compute different centrality metrics for the network, as well as cluster analysis, as a way of understanding potential similarities across cities.

Another component of the project is studying the mobility traces of the individual tourists. The mobility trace of a tourist describes the set of locations and corresponding times associated with a tourist’s movement in a particular region. Through data about such traces in a particular region, we can provide personalized descriptors as well as infer probabilities of the path which that particular tourist will follow. We can also understand and potentially classify users based on their mobility profiles. Such a classification is important from a region’s point of view because as it enables such regions to provide contextualized services that cater to the different profiles.

**Must-Have Features**

1. A city-by-city network of tourist flows with statistical metrics, such as centrality measures and clustering that identify important hubs as well as provide some measure of similarity and relatedness across tourist destinations in a given region.
2. A comprehensive analysis of aggregated mobility traces of the users.

**Optional Features**

1. Statistical inference analysis on the trace that individual travelers follow.
2. Inference on the detailed traveler trace within a city or destination, based on text analysis of blogs and twitts, e.g., attractions visited in New York City

**What Data**

There two types of data that we will explore for building the geo-spatial network. Through blog sites like TravelBlog, we can obtain data on traveler traces across cities and destinations in a particular region. In addition to movement data, we can obtain text descriptions of the traveler’s experience in the locations visited.

A second class of data that we plan to explore is Twitter. There have been several studies performed using geo-located tweets as a medium to study tourists and their mobility traces in their destinations. Twitter provides a short descriptive text (tweet) as well as location data for some individuals. An aggregation of such data for tourists would provide data that can be used for building our geo-spatial network.

**Design Overview**

Overview of Statistical and Computational methods

Various computation of centrality scores for each node in a network, i.e., eigenvector centrality, betweenness centrality, and degree centrality will be explored.

Clustering and community analysis will also be performed on the network to identify similarity and relatedness of cities and destinations. Here different clustering techniques such as k-means, spectral clustering, and others will be considered.

The mobility traces for different users can be classified using supervised learning techniques so as to get a sense for the different kinds of travelers observed. In addition, supervised learning techniques will be employed in order to infer the traces that individual travelers follow.

**Verification**

[To be discussed]

**Visualization & Presentation**

Here different network tools would be explored for presenting the insights learned from the study of the city-by-city network. Tools like Gephi, D3js, as well as networkx would be explored for presenting the network and the different clusters obtained.

In presenting our findings based on the network observed, we hope to emphasize the macrostructure and as well as the substructure of the network observed. These two levels of structure would present different insights about the similarity between cities as well as their relative importance for an overall region. Gephi provides a suitable interface for clear visual presentation of such information.

**Schedule**

We have about four weeks and some days to work on the project. We should shoot for weekly (twice a week as the deadline approaches) updates/evaluations for the entire project so that we can access the status of the project. We should also keep personal ipython notebooks for analysis done. As we get more results we can update a central notebook on github with results for the overall project.

Rough schedule

Week 1

Setup

Define specific areas and methodologies that each team member would like to own.

Begin data collection and preliminary explorative analysis of the data.

Further research into specific techniques to be used in the analysis phase especially for the mobility traces.

Week 2

Continue data collection

Construct a preliminary city-by-city network for a given region

Network analysis, ie centrality, clustering and others

Aggregate mobility traces

[Review meeting with the TF]

Week 3

Refine analysis with additional data

Recompute critical network metrics.

Begin validation analysis

Classification of mobility traces and potential prediction of next user location.

Evaluate the status of the results so far.

Week 4

Perform additional analysis as time permits for optional features

Assembly the overall set of results

Put together presentation material [screencast and presentation]