COACTION PUBLISHING

Inferring travel from social media

Alessio Signorini*, Philip Polgreen and Alberto Segre

Computer Science, University of Iowa, Iowa City, IA, USA

Objective

To use sequential, geocoded social media data as a proxy for human movement to support both disease surveillance and modeling.

Introduction

The spread of infectious diseases is facilitated by human travel. Disease is often introduced by travelers and then spread among susceptible individuals. Likewise, uninfected susceptible travelers can move into populations sustaining the spread of an infectious disease.

Several disease-modeling efforts have incorporated travel and census data in an effort to better understand the spread of disease. Unfortunately, most travel data are not fine grained enough to capture individual movements over long periods and large spaces. Alternative methods (e.g., tracking currency movements or cell phone signals) have been suggested to measure how people move with higher resolution but these are often sparse, expensive and not readily available to researchers.

FourSquare is a social media application that permits users to 'check-in' (i.e., record their currentlocation at stores, restaurants, etc.) via their mobile telephones in exchange for incentives (e.g., location-specific coupons). FourSquare and similar applications (Gowalla, Yelp, etc.) generally broadcast each check-in via Twitter or Facebook; in addition, some GPS-enabled mobile Twitter clients add explicit geocodes to individual tweets.

Here, we propose the use of geocoded social media data as a real-time fine-grained proxy for human travel.

Methods

Sixty-eight million geocoded entries (tweets and check-ins) from 3.2 million users were collected from the Twitter streaming API for the period from September 11, 2010 through January 28, 2011. The Twitter API provides a random sample of tweets; non-geocoded tweets or tweets originating from outside the United States were discarded. In addition, users with fewer than 6 records, or those who check in too frequently (more than once in 5 seconds) or travel too quickly (faster than 1800 km/hr) were removed to exclude automated bots or other location spam.

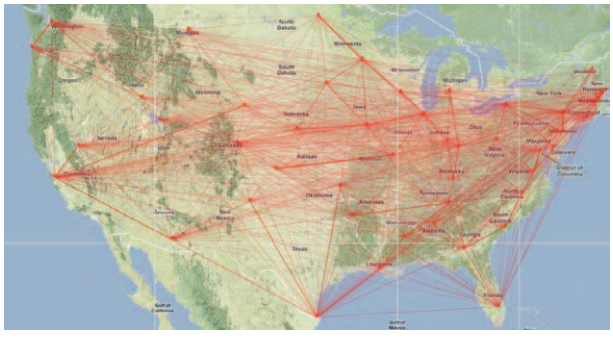

Fig. 1. User Transitions between U.S. States.

Results

We analyzed a 5-week subset of the data (September 11, 2010 through October 26, 2010) consisting of 3 million record intervals from 165,000 users.

We display intrastate travel by aggregating each user's consecutive records within each state and plotting only transitions between states (Fig. 1). The denser edges represent more frequent transitions, illustrating the pattern of travel on a national scale. We also constructed a heat map representation of Manhattan (Fig. 2) by aggregating users' check-ins with 500 m resolution. A larger bubble represents a denser set of records in that geographic area.

By linking each individual users' consecutive location records together, we computed the statistical distribution of time interval and distance traveled between records. About half of the checkins are less than 6 hr and no more than 1km apart from each other.

Conclusions

We show that social media location data can be used as multiscale proxy for travel at the national, state and urban level. These data are inexpensive and easily obtained. Furthermore, they can be used not only to understand historical travel but also to monitor in real-time changes in travel behavior to help inform disease surveillance.

Future work, currently underway, will validate this source of information against other sources of travel data and will investigate its value to better understand the spread of infectious diseases for disease monitoring and surveillance purposes.

Keywords

Social media; travel; modeling; surveillance; twitter

***Alessio Signorini**

E-mail: alessio-signorini@uiowa.edu

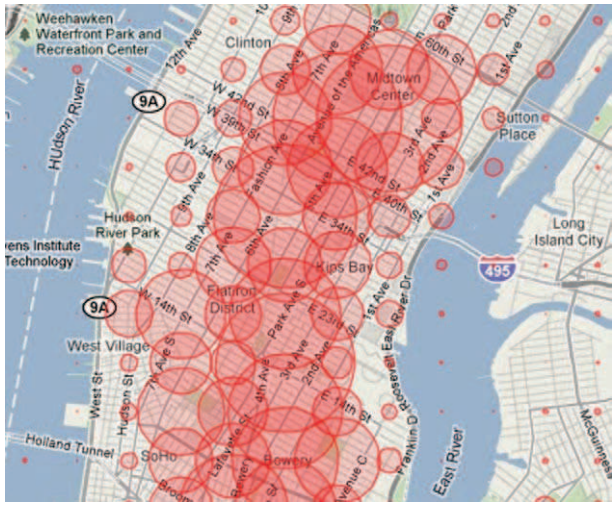

Fig. 2. Density of Check-ins over Manhattan, New York City, NY.

Citation: Emerging Health Threats Journal 2011, **4**: 11126 - DOI: 10.3402/ehtj.v4i0.11126