# Tracking Global Spread of Disease through Air Travel

Sean Brennan

Dec. 18, 2012

## 1 Introduction

### 1.1 Problem Statement

In this project I hope to leverage data mining and machine learning techniques to track the global spread of disease via airports and air travel. This project will require a combination of past techniques devised by Adam Sadilek to accurately identify indicators of illness in geotagged tweets, as well as some degree of geospatial inference to decide whether or not someone is tweeting from an airport. Additionally, this project will necessitate a rich model of temporal health inference, as well as a rich probabilistic model for user encounters.

### 1.2 Problem Importance

The rise of affordable mass transit in the previous century has completely changed the scope of epidemiology, both spatially and temporally - people are now able to travel much farther and faster than they could have possibly imagined traveling even 100 years ago. Air travel in particular has introduced complications, as our highly connected system of airports has created conceptual "global communities". The SARS epidemics of 2002 and 2003 are salient examples of how disease can propagate through this new global transit network.

## 2 Research

The increased risk of disease transmission on commercial planes has been known since at least 1979. In a seminal case study, passengers that were stuck aboard a plane delayed for three hours were found to have an incredibly high incidence of influenza within 72 hours of landing, simply due to one passenger aboard being ill with influenza initially. [4] Similar results were found in 2009, at the height of the H1N1 influenza scare. In a case study of two flights to Australia in this time frame, even just 2% of passengers having some influenza-like illness increased the post-flight incidence rate of influenza to 5%; additionally, it was found that sitting within 2 seats of a passenger with influenza would increase the likelihood of you contracting influenza to almost 8%. [3]

A big challenge in this line of research is modeling the new ways in which humans travel in the 21st century. In past eras, the spread of disease could be modeled as a fairly intuitive and smooth spatial diffusion phenomenon. The bubonic plague's course over Europe is a classic example of this, in that it would travel somewhere around 200 miles a year in every direction. [2] However, with the convenience of air travel, this smooth diffusion is now intermingled with short periods of very fast travel over long distances, yet many epidemiological models do not account for this so-called "super-diffusive" movement. [1] Later on in the proposal I will discuss how our model will handle this.

Of course I would be remiss to not mention the work that we have done related to inferring sickness from Twitter data. Our approach is a semi-supervised one and involves training two support vector machines over a huge corpus of tweets: one is heavily penalized for false positives and thus becomes very good at positively identifying sick tweets; the other is heavily penalized for false negatives and thus becomes very good at identifying tweets that are not about sickness. Training data for these SVMs was gathered by having users on Amazon's Mechanical Turk label the initial set as "sick" or "other". [5]

# 3 Methodology

## 3.1 Data Acquisition

Data will predominantly be acquired through harvesting geotagged tweets via Twitter's search API over a period of several weeks. The search API is powerful in that it enables users to pull in a huge volume of tweets, and these tweets are typically no more than a week or so old, so it's all very up-to-date. Of course, I will only concern myself with tweets in the English language. I will also only concern myself with geotagged tweets within cities that are served by the airports which I'm tracking.

In addition to the Twitter data, some initial collection related to airports will need to be performed. I will only track the (as of 2010) 50 most heavily trafficked airports in the world, since the majority of interesting cases are most likely to be in these locations.

## 3.2 Algorithms

The fundamental approach that defined our earlier work will still be at the heart of this new branch of research: Adam's support vector machine algorithm for classifying sick and healthy individuals. We will use this classification to obtain the initial health rating of an individual tweet.

One enhancement to our model that we will add in this project is the temporal relation between a user's tweets; that is, if a certain user's tweet is marked as having some health risk, how does that affect our estimate of their past and future health risks? We decided to treat a tweet as a "median point" in an individual's sick period and apply a Gaussian distribution around it, which allows us to project the health risk of an individual at some arbitrary point in time. Risk of infecting others is highly dependent on several confounding factors, such as what disease you have, what its period of contagiousness is, and what its incubation period is. However, the Gaussian model will likely give us good results on average, and trying to account for such factors given the sparsity of detailed information about what disease a user has is difficult and likely not a very rewarding approach.

The next and final step of the algorithm is its "special sauce". For every single day, we can represent the encounters as a graph where each vertex represents a unique user and every edge represents an interaction between two users. Each vertex is initialized to have some probability $D$, which represents the likelihood of that user being sick at that time. Each edge is then given some weight $M$ that represents the probability of that encounter occurring. In a smaller scale local case study, this probability would mostly be a function of time and location, e.g. two people tweeting from the same area within five minutes of each other were much more likely to have encountered each other than if they had tweeted from the same place two hours apart. However, since we are now dealing with international flights, travel time and distance must also be taken into account. For example, an encounter between someone who tweeted at Heathrow and someone who tweeted at JFK six hours later is now far more likely. For the sake of computational tractability, we can eliminate encounters that have an $M$ which falls below a certain threshold. Finally, the last bit

in our probabilistic model is the probability $S$, here representing the likelihood of a disease being spread via some interaction.

With our probability model fully initialized, we can commence an iterative algorithm that will fully refine the health risk of every individual. For every vertex $V$ in the graph, we examine all of its encounters $(V, W)$ and "boost" the health risk of $W$ by some function of $D_V$, $S$, and $M_{V,W}$. We will iteratively refine the health risks until convergence, or rather until the average change in probabilities between iterations falls below some threshold of insignificance.

# 4 Evaluation

The most important thing to validate is the accuracy of our approach: how does our new model correlate with results from previous research in this area? A metric that we typically correlate with is Google Flu Trends, since this has been shown time and time again to be accurate and representative of what's occurring in the real world. Since we have already validated the accuracy of our health risk rating system, this measure of correlation will primarily validate our model of user interaction: that is, is it appropriate to extrapolate local interaction patterns to a global scale?

## References


[1] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462–465, January 2006.

[2] V. Colizza, A. Vespignani, and E.F. Hardy. Impact of air travel on global spread of infectious diseases, 2007.

[3] A.R. Foxwell, L. Roberts, K. Lokuge, and P.M. Kelly. Transmission of influenza-like illness on international flights. *Emerging Infectious Diseases*, May 2009.

[4] M.R. Moser, T.R. Bender, H.S. Margolis, G.R. Noble, A.P. Kendal, and D.G. Ritter. An outbreak of influenza aboard a commercial airliner. *American Journal of Epidemiology*, 110:1–6, July 1979.

[5] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.