The Pennsylvania State University
The Graduate School
Eberly College of Sciences

DATA SCIENCE WITH SOCIAL MEDIA FOR EPIDEMIOLOGY AND PUBLIC HEALTH

A Dissertation in
Biology
by
Todd Bodnar

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2015

The dissertation of Todd Bodnar was reviewed and approved* by the following:

Marcel Salathé
Professor of Biology and Computer Science
Dissertation Advisor

Andrew F. Read
Professor of Biology
Chair of Committee

Conrad S. Tucker
Professor of Industrial Engineering

John Yen
Professor of Information Sciences and Technology

Richard J Cyr
Interim Head of Biology

*Signatures are on file in the Graduate School.

Abstract

The emergence of large scale web data has opened the door to novel research questions to be answered in many fields. However, most work has been done based on either improving a user's experience with a service or increasing the likelihood of click-able advertisements with other fields both showing and receiving less interest in these big data web analytics skills. In this dissertation we apply this type of methodology to epidemiology and public health, a field which, aside from genomics and disease surveillance, tends to focus on traditional methodology such as field research, animal experiments, and mathematical modeling.

We begin with a review of previous work in this area and its shortcomings. Specifically, we consider modern web-based disease surveillance systems such as Google Flu Trends and related academic work. We find inconsistencies in the literature around how to measure the accuracies of the models, making cross-publication comparisons difficult. To address this, we reproduce several methods on our own datasets. We find inconsistencies in the expected performance, along with unusual results, and determine to develop a better, more validated approach.

We note that these papers do *not* actually focus on validated individual diagnoses, but instead with simply fitting the population's web behavior with the population's disease incidence. We address this by collaborating with a local health provider to obtain medical records related to patients that had been previously, professionally diagnosed with influenza. We then obtained approximately all Twitter data related to these patients. We then developed classifiers based on this data that could accurately determine if the patient had influenza or not at a given time.

We then applied this classifier to a multi-million user dataset consisting of approximately 10 terabytes of Tweets covering a four year span of users located in the United States. To scale to this size, we developed a map-reduce version of the classifier implemented with Apache Hadoop and stored the results in Apache Hive. We used this information to (1) build a novel disease surveillance system that works at any arbitrary geographic scale and (2) to analyze geographic spread of influenza.

With this information about disease transmission, we then considered a separate public health question of *how* to efficiently spread accurate information about a disease. To do this, we studied the spread of information on Twitter related to three public health events: the detection of a novel strain of influenza (H7N9), a measles outbreak related to vaccine refusal, and Autism Awareness Month. As with actual diseases, we start with a network analysis of the spread of information. We then extend these traditional spreading models to include information about the content of the Tweet and the type of person spreading it to develop a more accurate model of information spread.

Finally, we conclude with a discussion of future paths that this research could follow such as employing deep learning artificial neural networks to increase the accuracy of our disease diagnosis system or performing experimental manipulation of Twitter users to encourage healthy activities.

Table of Contents

List of Figures

List of Tables

Acknowledgments

A project of this magnitude can not solely be the work of one man, and there are many colleagues I would like to thank and acknowledge for their contributions to this dissertation. First, I am grateful to my advisor, Marcel Salathé for his mentoring over the past five years. Of note, I'd like to acknowledge the significant professional opportunities, intelligent conversations, financial support and swiss-made pizza provided. I'd like to thank Charles Fischer, Steven Schaeffer and Kathryn McClintock for their administrative support. I acknowledge the University Health Services for allowing us access to their records, without, chapters 3 and 4 wouldn't exist. I want to thank Tim Leso and Eric Charles for guiding me to do more research in the first two years of my undergraduate degree, starting me out on this crazy path of research. I thank my committee members for the many discussions about this dissertation and their helpful advice. Additionally, I'd like to thank Conrad Tucker for including me into some of his other projects.

I want to thank our group's programmers, Brian Lambert, Issac Bromley, and Shashank Khandelwal, for their work on developing many applications our group has used. I'd also like to thank members of the Open Source Software community that have ever contributed to LaTeX, Weka, R, Open Office, Python, scikit-learn, GIMP, or were part of the Apache Software Foundation or any of the *nix distributions that I've used during my time as a grad student. I wish that your contributions were more widely acknowledged in academia.

I'd like to thank the Santa Fe Institute for hosting me during the summer of 2013. I'd like to specifically thank Juniper Lovato, John Paul Gonzales, Tom Carter and Sander Bais for organizing the many lectures, networking opportunities, and other events. I'd like to thank Owen Densmore, Stephen Guerin, John Driscoll, Alfred Hubler, David Masad, Mengsen Zhang, Alastair Jamieson-Lane, Nix Barnett, Yan Xu, Carissa Flocken and everyone else that I've met there for the amazing conversations during my time at Santa Fe and after.

I'd like to thank Susie Lim, Zhuojie Huang, Cosme Adrover, Timo Smieszek, Orhan Kislal, Eun-Kyeong Kim, Spencer Carran, Owen Shartle, Kezia Manlove,

Yafei Wang, and Kurt Vandegrift for their various types of support. Finally, I'd like to thank my mother and sister for their continuing support.

Chapter 1
Introduction

The emergence of social media coupled with big data methodology has opened the door for the next generation of surveillance systems. We develop methods to create such systems on the popular social network, Twitter. We have found [1] issues with previous approaches to use Internet data for disease surveillance (for example, [2–5]). First, these systems have been shown to perform poorly [1, 6–9]. Second, these systems tend to not use or generate in depth knowledge about the disease. Third, these systems do not provide information on how to influence activity related to the disease. Here, we provide a further discussion of these issues and present ways in which they can be addressed.

1.1 Issues with Current Approaches towards Digital Epidemiology

1.1.1 Lack of Validation

Google Flu Trends (GFT)–a real time influenza surveillance system based on Google searches–is generally considered the baseline for web-based disease surveillance [2]. However, GFT has been shown to perform poorly at times, its most 'famous' gaft being the January 2013 overestimate of more than twice the true level of influenza prevalence [6, 7, 9]. One explanation for this overestimate provided by the Google Flu Trends team [7] and others [6, 9] is that a particularly strong public interest in the disease confused the system. This interest was due in part to a feedback loop caused by GFT's high estimates. Google has since modified their algorithm to

dampen these types of events [7], but Lazer et al. [9] have argued that this both doesn't fix the issue and further obfuscates the model.

Another issue of concern is the difficulty in model comparison. For example, the original GFT paper was published in 2008 [2] and finds a mean correlation of 0.9 while Signorini et al. [5] report their model as having an average absolute error of 0.28% on the 2009-2010 flu season[1]. However, when we attempt to standardize these two paper's methods on Twitter data from the 2011-2012 flu season [1], we find the Google's model has a mean correlation of 0.88 and Signorini's model has a mean correlation of 0.76 (mean absolute error = 5.7%). Additionally, one of Culotta's [3] models ("simple-freq-corr") was found to have a correlation of 0.65 in our dataset compared to his finding of a correlation of -0.034, providing a rare counter example of publication bias. This general lack of reproducibility of results, besides GFT's results, is not due to fraud or incorrect implementation, but due to some of these models showing a high sensitivity to the initial data—resulting in a questionable efficacy of comparing different models.

1.1.2 Lack of Semantic Knowledge

As mentioned above, GFT does not specifically differentiate between searches about *having* influenza and *news* about influenza. A similar issue occurs with Twitter analysis: rate limiting makes it impossible for a third party to gain *all* tweets during a period of time. This is commonly addressed by searching for Tweets which contain keywords such as "flu" or symptoms of influenza. Note that this does not filter out Tweets about news about influenza, although work has been done in filtering out these messages [3, 10]. However, the bag-of-words approach for Twitter disease modeling has one major issue: we've been able to replicate the results from datasets based on flu keywords by using tweets that were collected using completely irrelevant search terms, specifically keywords related to zombies (see Figure 1.1).

Additionally, we've compared an unfiltered, random sample of tweets and generated time varying frequencies simply by

$$x_{i,t} = (sin(t * \alpha_{i,1} + \alpha_{i,0}) + \mathcal{N}(0, .1))/1000 + .001 \tag{1.1}$$

[1]The reader should also note that the average flu rate during this period was approximately 3.9%, resulting in a relative error of about $0.28/3.9 = 7.2\%$.

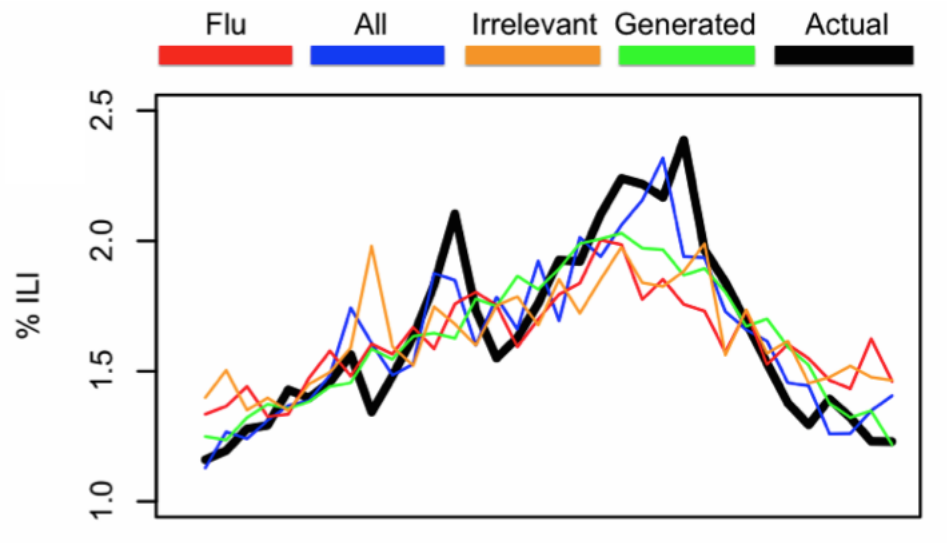


Figure 1.1: SVM-regression for the 2011-2012 Influenza season using multiple datasets, from [1].

where $\alpha_{i,0}$ and $\alpha_{i,1}$ are randomly generated weights, $\mathcal{N}(0, .1)$ is a normally distributed random variable with standard deviation of 0.1 and $x_{i,t}$ is the value of curve i at time t. While this is essentially building a model based simply off of the initial, disease frequency's time series, the fact that its performance is equivalent to more complex, data mining approaches brings the value of the later approaches into question. Further discussion and methods are available in chapter 2.

Besides human determined keywords, all papers surveyed do not encode any semantic knowledge into their models which may be one explanation for issues related to model generalization outside of the initial dataset.

1.1.3 Obsession with Disease Tracking over Knowledge Generation

Infectious disease studies using big data have strongly focused on surveillance systems [11] to either predict the future or current state of a disease's spread. [2, 3, 5, 12–14] These models tend to strongly correlate with the actual disease's prevalence (for example, Yaari et al. [15] report $r > .97$) limiting the value of further optimizing these methods or the likelihood that improvements will be statistically significant. Other topics–such as disease transmission [16–18], the

use of vaccines [19–24], disease movement patterns [25–27], behavioral effects of diseases [28–30] or the economic impact of diseases [31–33]–are widely studied topics in traditional epidemiology, but have been neglected when it comes to big data analytics. This is particularly odd considering that the basis for these other topics, measurements of a population's illness, are also the basis for disease surveillance systems.

1.1.4 Inability to Influence Disease

Using messages to influence behavior is a well studied technique, [34–39] indeed it is the basis of the advertisement industry. However, work on public health messages has lagged more commercial topics [40]. Work has generally focused on an individual's knowledge about a disease, as discerned through surveys [29] or a simple analysis of events [41]. While an understanding of the population's opinion on a topic can be the basis for crafting an influential message, it is a relatively primitive approach compared to large scale A/B testing [42] or modeling a user's behavior using hundreds of thousands of variables. [43]

1.2 Proposed Solutions

1.2.1 On the ground validation through professional diagnoses

Others have worked on a top-down approach by using region-wide disease prevalence data from which Twitter data can be fit. [2–5] However, they do not aim to say whether a specific Twitter user is ill, limiting the usefulness of such methods. In chapter 3, we consider a bottom-up approach to disease surveillance by diagnosing users based off of their Twitter information [44]. To do this, we survey 104 people that were *professionally* diagnosed with influenza through Penn State's University Health Services. As a control group, we survey an additional 122 individuals that were *not* diagnosed with influenza. From this data, we were able to collect Twitter data from 104 users.

We employed machine learning algorithms such as naive Bayes and support vector machines to use the aggregated set of messages a user posted during the time that she was ill, or an equivalent aggregation when she was healthy, to attempt a

diagnosis. Next we looked at non-message information. We performed anomaly detection on a user's rate of Tweeting to see if their usage changes when sick. We find a significant difference in activity, however the results are too noisy to be considered a viable method to diagnose influenza. We consider the aggregated messages from a user's friends *followers* and perform text analysis on them, as was done on the user's messages. We find that a user's follows are a much stronger signal toward a user's health than the accounts that a user follows. This hints at the Twitter's network structure: I follow both celebrities and my friends, but only my friends follow me. Finally, we employ meta-classifiers which use the output of each of the previous classifiers discussed as an input for a final classifier to get better classifier accuracies.

1.3 Quantifying Disease Dynamics

In chapter 4, we combine this classifier with a dataset of 2.7 billion Tweets to track 16 million users' disease states over a period of four years. Additionally, we employ geographical analysis of the users' supplied location information to approximate his real-world social network, [45–47] specifically with respect to disease, which has previously been approached through measuring real world contacts. [16, 48, 49] We do this by seeing if other users near a given user can inform us about the given user's future health status. That is, are they likely to get sick when people near them have recently been sick?

This question can be answered based on probabilities, but instead we decide to model it based on the base reproduction rate, R_0. This rate is the basis of most disease spread models, however, it is generally calculated by fitting curves to measured disease prevalence rates. [50] Attempts to study peer to peer transmission are limited in size (generally by the costs of tracking a group's disease states) to either 10's or 100's of individuals [16, 17, 51, 52]. Instead, we exploit an already developed social network platform, Twitter, as the basis of our collection, allowing us to preform these studies at web-scale.

1.4 Targeting Messages towards Disease Related Individuals

In chapter 5, we consider the effects of various aspects of a message on retweeting rates, a signal for reader interest. [53–56] We consider network structure, the type of user that posted a message and the textual content of the message. Additionally, sentiment analysis is often used to measure how positive or negative a message is. [54, 55, 57–59] Again, this has been shown to be a signal for retweeting rates. However, in the case of disease information, messages tend to be predominately negative. Instead, we develop a model based on four dimensions of emotional content [60–62] and apply it towards the retweet prediction problem.

Chapter 2
Validating Models for Disease Detection Using Twitter[1]

2.1 Introduction

The rapid adoption of social media and the internet in general has opened the door for novel developments in epidemiology [19, 64–70]. Much of this research has been aimed at data mining social media services such as Twitter or Facebook. Due to its openness, Twitter has been of particular interest [19, 71–73]. The site's microblogging and mobile communication features make it particularly useful for determining current levels of disease.

Given the rapid rise of social media usage, assessing disease prevalence using social media will become increasingly important. It is therefore prudent to continuously validate the underlying models [66, 67, 69]. Methods for validation assume that the training and testing data are independent of each other. While this assumption is never completely true, it is often sufficient. However, due to the strong spatial and temporal nature of infectious disease dynamics – along with a lack of multiyear social media datasets – this assumption may result in an inaccurate model.

In this paper, we take previously published models [68, 69, 74] and perform a battery of tests to check for potential issues. We do this by comparing the results of a traditional influenza related tweet dataset to a dataset of tweets that has not been filtered for a specific topic, a dataset of tweets related to a topic that is irrelevant to influenza, and a set of frequencies generated from random sine waves.

[1] A version of this chapter [63] was previously presented at WWW2013, ©2013 International World Wide Web Conferences Steering Committee.

In addition, we compare 10 fold and leave-one-out validation where the testing data are either from a different region or time than the training data. We find that (i) seemingly irrelevant tweets are moderately successful in assessing influenza prevalence, (ii) generated frequencies are often as good as measured frequencies from social media, and (iii) the choice of the validation method greatly affects the model's reported performance.

2.2 Data Sets

2.2.1 Influenza Prevalence

The CDC defines ILI (influenza like illness) as an illness with a fever and a cough or sore throat without a known cause other than influenza. Because ILI is indistinguishable from influenza, except through expensive tests, most data is reported as ILI prevalence instead of influenza prevalence. We used the percentage of doctor's visits that were for ILI between October 2, 2011 and May 26, 2012, as reported by the CDC,[2] to serve as the ground truth. The CDC provides this data both on a national level and as a set of 10 HHS regions.

2.2.2 Tweets

We collected 238,506,796 tweets from the continental United States between October 2, 2011 and May 26, 2012 – a 34 week span – through Twitter's API. The tweets were acquired by requesting all tweets with high-resolution geospatial information within a bounding box that covers the continental United States. By limiting our requests to tweets with high-resolution geospatial data, we potentially introduced a bias in the data. However, this allowed us to avoid being rate limited by Twitter, guaranteeing that the dataset contained every tweet from Twitter, subject to the above parameters.

Each tweet consists of geospatial information, the time that the tweet was sent, and the contents of the message tweeted, along with other information such as the user's profile picture and sign up date. The quality of the tweet's geospatial information varies greatly based on how the user sent it. For example, a tweet sent

[2]http://www.cdc.gov/flu/weekly/pastreports.htm

from a laptop may only have information from which city or state it originated from. In our case, we limited our search to tweets with longitude and latitude coordinates indicating that the tweets most likely came from gps equipped devices such as cell phones. A tweet contains at most 140 characters of text. Note that we did not limit our collection to tweets with a specific set of keywords.

We trained our models on 6 subsets generated from this data. The first subset was simply the entire dataset grouped by each week. The second subset was limited to tweets that contained at least one of the following ILI related keywords: 'flu', 'cough', 'fever', 'headache' or 'head ache'. We defined a third subset of the data using the keywords 'zombie', 'zed', 'undead' and 'living dead'. Since these keywords are presumably unrelated to ILI, this subset serves as a test for odd model behavior. The other three subsets are the same as the first three, but also divided based on which region the tweet came from. We used the 1000 most common words in each of the subsets as the list of keywords for the models. In the ILI dataset, this includes all of the words that were tweeted an average of at least one time per week. The other subsets were also of the top 1000 words to avoid biasing caused by a difference in the amount of data being fed into the models. We did not filter out stop words because, as mentioned by Culotta [69], stop words such as 'I' or 'have' provide valuable information if the tweets also have an ILI keyword. Because of daily fluctuations in Twitter use, all keyword trends are measured by their frequency.[3]

In addition to these 6 datasets, we 'simulate' keyword frequencies by generating another two datasets. We generate one-thousand sine curves with random wavelengths and add noise generated by a normal distribution with a standard deviation of 0.1 to each point. They are then divided by 1000 to be of the same scale as the actual frequencies. We add .001 to each point to avoid negative frequencies. We also generate one for regional data where the wavelengths are fixed across regions but the noise is not. As with the irrelevant tweet dataset, these serve as control groups.

[3]The datasets and associated code are available at http://github.com/salathegroup/w3cRio

Y	1	2	3	4	5	6	7	8	9	10
R	.87	.88	.63	.91	.98	.95	.95	.98	.89	.90

Table 2.1: Predicting region Y's ILI prevalence simply based on the other 9 regions' current prevalences with a multivariable regression illustrates the strong relationship between the regions' disease levels.

2.3 Models

2.3.1 Regression on Tweet Count

Following previous work [69, 74], we first consider using a linear regression of the raw count of tweets that contain at least one of the keywords, as defined above, to predict the CDC's ILI prevalence:

$$logit(CDCRate) = \beta_0 + \beta_1 logit(x) + \epsilon \tag{2.1}$$

Where β_0 and β_1 are coefficients, ϵ is the error function,x is the number of Tweets containing at least one of the keywords and $logit(x) = log(x/(1-x))$.

2.3.2 Multivariable Regression

To gain more information from the tweets, we consider multivariable regression [69, 75].

$$logit(CDCRate) = \beta_0 + \sum_{i=1}^{n} \beta_i logit(x_i) + \epsilon \tag{2.2}$$

Where x_i is the frequency of the i^{th} keyword.

2.3.3 Select Best Keyword

It has been argued that multivariable regression is prone to overfitting [69, 74]. An alternative solution to multivariable regression is to perform regression on the keyword that correlates the best with the training data, and use it for the regression model.

2.3.4 SVM Regression

We consider a form of regression that utilizes a SVM (support vector machine) which has been shown to predict ILI prevalence well [68]. A SVM defines a multi-dimensional hyperplane that divides the training data. While this hyperplane is generally used in classification problems to divide two classes, it also allows for regression based on a sample's distance from the plane [76].

2.4 Model Validation

Models are evaluated by dividing the dataset into training and validation sets. The way that the sets are divided has the potential to greatly affect the measured performance of the models [75]. Examples of these issues are

1. Too much data used in training results in few points to compare to the model's results.
2. Too little data used in training results in a poorly fitted model.
3. If the testing data is too similar to the training data, overfitting may not be detected.
4. If the testing data is too different from the training data, the model will perform badly regardless of it's sophistication.

As a concrete example of issue 3, consider the commonly used method of reserving one region's data for validation. As information about other regions gives a fair bit of information about a region (see table 2.1), there is a risk that a model will present good results in the testing data even if the model has not learned the system's underlying dynamics.

Aside from a simple percentage split, we allow for 10-fold-cross validation. In cross validation, the dataset is divided into k equally sized splits. Each split is used to test a model that was trained on the remaining $k-1$ parts. In addition, we allow for leave-one-out cross validation where each datapoint is used to test a model that was trained by the remaining data.

Model	Flu	All	Irrelevant	Generated
Count	.4184	.4344	.0089	.3529
Multi	.7681	.8774	.6300	.8367
Best	.6946	.6583	.7991	.7313
SVMr	.7557	.8580	.7382	.8766

Table 2.2: Average correlation of the models' predictions and the CDC's national ILI prevalence.

Model	Flu	All	Irrelevant	Generated
Multi	.3493	.6468	.2860	.2713
Best	.1575	.2381	.3158	.6653
SVMr	.4538	.7378	.4270	.7113

Table 2.3: Mean correlation of the results of a model trained on 9 regions and evaluated on the last.

2.5 Results

We first evaluate the models with the national level data through leave-one-out validation (figure 2.1) and 10-fold-cross validation (table 2.2). In the case of 10 fold validation, we repeated the evaluation 100 times with different, randomly generated splits. With both validation methods, multivariable and SVM regression performed similarly. We corroborate Culotta's finding that multivariable regression performs better than regression on just the count of relevant tweets contrary to Ginsberg et al.'s findings in Google search queries. For a discussion on why this may be, see [69]. Because of its much lower performance, we ignore it for the rest of the analysis.

When we repeat this procedure on a regional level with each region being a 'fold', we observe similar behavior (table 2.3, fig 2). However the accuracies are lower in every case. This suggests that a model with what appears to be better performance may not necessarily be better than one with a lower level of performance if the first model's testing set was temporally separated while the second model's testing set was spatially separated.

It may appear that both multiple regression and SVM regression have similar accuracies in the regional data, however their results from the generated dataset are noticeably different. The intuitive conclusion would be that SVM regression performs better than multiple regression. This is not necessarily so. In the case of

SVM regression, real Twitter data is barely a better predictor of ILI than generated sine curves. This calls into question the benefits of using social media with SVM regression, if randomly generated data performs nearly as well.

Interestingly, the dataset that was not filtered resulted in a higher correlation in 3 of the 4 models. This may be due to the filtering process removing potentially insightful tweets that do not contain any of the keywords. Another possibility is that reducing the number of tweets makes the data's trends more susceptible to random fluctuations and thus noisier.

2.6 Conclusions

In this paper we evaluated several well known regression models on their ability to accurately assess disease prevalence from tweets. We found that even irrelevant tweets and randomly generated datasets were able to assess disease levels comparatively well. This could serve as a ground level for evaluating other models: if a model can do only slightly better with seemingly relevant data than with seemingly irrelevant or random data, then it is probably not learning much from the tweets and its ability to fit the data can be attributed to other factors.

The ability for even randomly generated curves to fit the data may be explained by either spatial or temporal autocorrelation. For example in 5 fold cross validation, a model may simply interpolate between points in the dataset instead of gaining information from the tweets. Future work could look at other diseases that have less predictable long term dynamics, such as gastroenteritis or asthma. Another possibility is that – especially in the full dataset – tweets about other events that happened around the same time that ILI peaked could be chosen by the model as a predictor, but clearly this would not be expected to replicate across multiple years.

Finally, we found that the way that the training and testing data were divided had a strong effect on the reported performance of a model. Future work could build a mathematical model to explore these effects and develop a method to evaluate models in a way that best measures their true performance.

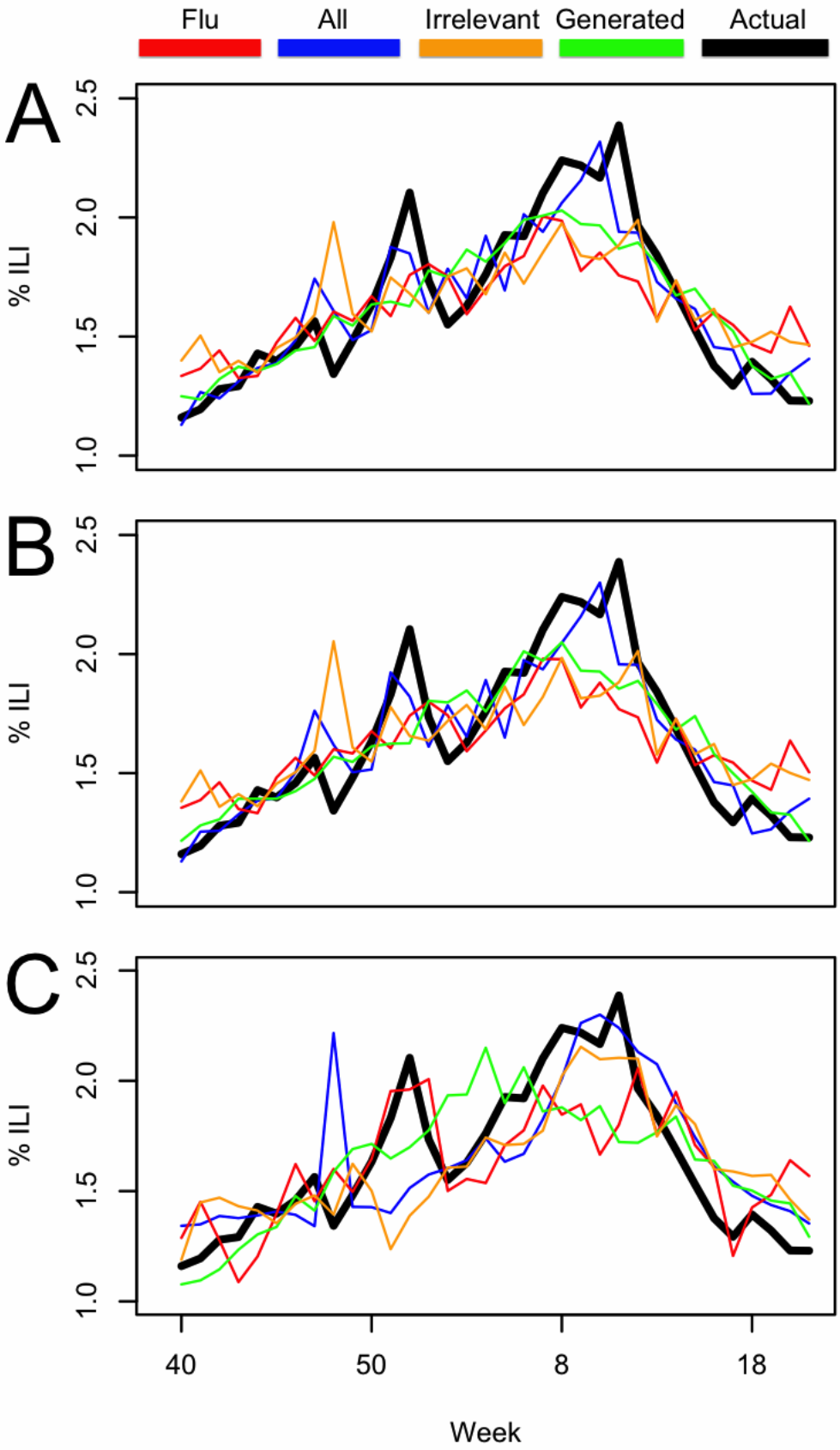


Figure 2.1: Results from (a) SVM regression, (b) multivariable regression, and (c) single regression for each dataset compared to the CDC's national reported ILI levels during the 2011-2012 influenza season. Each data point is the result of a

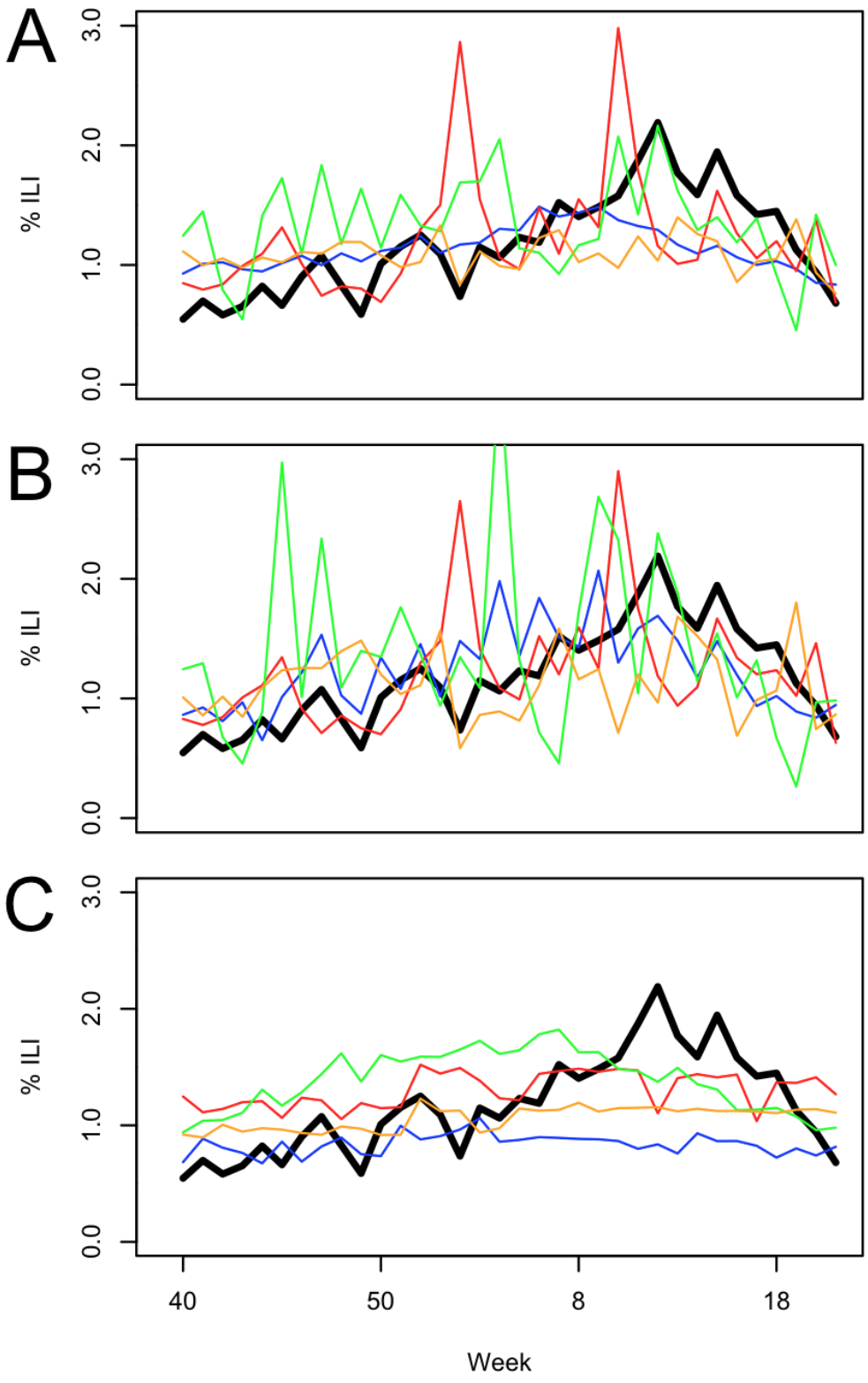


Figure 2.2: As with figure 2.1, but results for region 10 from models trained on regions 1-9.

Chapter 3
On the Ground Validation of Online Diagnosis with Twitter and Medical Records[1]

3.1 Introduction

Disease surveillance systems – which traditionally rely on reports from medical practitioners – are an important part of disease control. However, these traditional surveillance systems are often costly and slow to respond [64,78,79]. The widespread adoption of the Internet by the general public has provided opportunities for the development of novel disease surveillance methods. Compared to traditional systems, where data is provided by medical diagnosis, these new systems provide either semi-automatic – through long term self reporting systems [80, 81] – or fully automatic – through data mining search queries or social media [63, 65, 66, 69, 82] – disease surveillance. While these methods are cheaper, faster and cover a larger number of individuals than traditional systems, one can be less confident about their results than the results from a system based on professional diagnosis. In this paper, we develop a system that performs long term surveillance on Twitter users with classifiers trained on professionally diagnosed data that combines the advantages of all three of these systems.

Previous work with data mining social media has focused on methods to replicate

[1] A version of this chapter [77] was previously presented at WWW2014, ©2014 International World Wide Web Conferences Steering Committee.

the patterns found in traditional surveillance networks [63, 65, 69]. However, these methods have several limitations. First, they generally do not differentiate between an individual with an illness and an individual that is worried about an illness; which may have resulted in a predicted influenza rate that was much higher then the actual 2013 influenza rate [63,66,82,83]. Second, these methods cannot be extended to areas without a previous surveillance network to train the model. Finally, these methods are fundamentally incapable of detecting diseases that do not show strong spatial-temporal patterns such as mental illness, obesity or Parkinson's disease. Instead of top-down methods to measure levels of disease in a population, we approach this problem from the bottom-up. This addresses all three of these issues: we only diagnose individuals that are likely to have the disease, and not just interested in the disease; we do not require previous data when applying these methods to new problems or locations; and these methods can easily generalize to diseases that do not show strong spatial or temporal patterns because we focus on an individual level.

Participatory systems, such as InfluenzaNet or Flu Near You, use self-reported symptoms to diagnose an individual and also work from a bottom-up approach [80, 81]. These systems have the potential to be better than traditional surveillance systems because they update in near-real-time and can detect cases even when the user has not gone to their doctor. These systems require the user to sign up which allows for long term studies which are not normally able to be done with Tweets or search queries. However this reduces the number of users studied compared to data mining approaches. For example, Flu Near You had a total of 9,456 users report during the week ending on 29 December 2013. Marquet et al. [80] have shown a large drop out rate with only 53% of users participating for five or more weeks. While this amount of data is sufficient for many purposes, a system based on Twitter's millions of active users would open the door to more applications.

We develop such a system as follows. In section 2 we describe the collection of an individual's professional diagnoses of influenza and the collection of their Twitter information. In section 3 we consider extracting textual information from Tweets as a method for diagnosing influenza. Previous work has focused on this area. Additionally, we consider other methods for detection. In section 4 we consider anomalies in a user's Tweeting behavior as a signal for diagnosing influenza. In section 5 we extend these methods to other users on a person's social network to

diagnose the original person. In section 6 we aggregate the results of the previous classifiers to develop a more accurate meta-classifier.

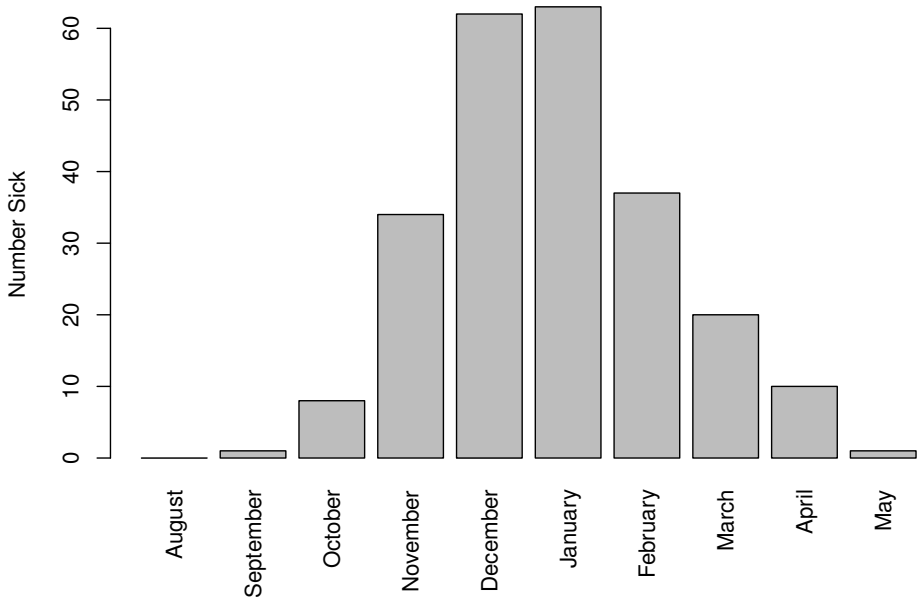


Figure 3.1: The professionally diagnosed Influenza cases during the 2012-2013 season in our sample.

3.2 Data Collection

3.2.1 Medical Records

We received information from the Pennsylvania State University's Health Services about 104 individuals that were diagnosed with influenza by a medical professional during the 2012-2013 Influenza season. Due to privacy concerns, we were limited to knowing which month an individual was diagnosed (see figure 3.1). For comparison, we also obtained information from 122 individuals that were *not* diagnosed with influenza during this time. The participants were mostly students (72% were between the ages of 18 and 22) and slightly more female than expected ($133/226 \approx 58.8\%$.) Data collection was approved through the Pennsylvania State University's IRB (approval #41345.) Twitter handles were available for 119 of these individuals.

3.2.2 Twitter Records

While we received a total of 119 Twitter accounts, 15 were discarded because the associated accounts were either non-existent, banned or private. For each of the remaining 104 accounts, we pulled their profile information, their friends and followers information, their most recent 3000 tweets, and their friends' and followers' profiles and tweets. Some users did not tweet during the month that they were sick; we kept those accounts as part of the control group. We were limited to the most recent 3000 tweets by Twitter's time line query, but this only effected two accounts – both of which posted multiple times per hour and were thrown out because we could only look back a few days.

We collected data through the Twitter API. Tweets, profile and follower information queries have separate rate limits and were collected in parallel. Since users continued to Tweet during data collection, each account was queried no more than once every three days for new Tweets. When all accounts could not be queried due to rate limiting, the accounts that had been queried the least recently were updated. Additionally, the 104 seed accounts collected above were given higher priority over their friends and followers. In total, we collected 37,599 tweets from the seed accounts and 30,950,958 tweets from 913,082 accounts that they either followed or were followed by.

3.3 Text Based Signals

In this section, we consider diagnosis based on the content of a user's tweets. Such analysis can be approached by keyword analysis, where the presence of absence of a keyword predicts disease, or through text classification, where the tweets are classified as being about disease or not about disease. We begin by dividing the tweets into two sets: tweets that were posted the same month that a user was sick and tweets that were posted other times. We find a total of 1609 tweets from 35 users in the first category.

First, we use the occurrence or absence of keywords as features for classification. A set of keywords are defined that are possibly signals of influenza. We chose {flu, influenza, sick, cough, cold, medicine, fever} as our set of keywords. These keywords include the names and symptoms of the illness in addition to "medicine"

Word	Total	Odds Ratio	Significance
flu	25	40.14	<0.0001
influenza	1	0.00	0.8325
sick	128	5.22	<0.0001
cough	18	4.48	0.0094
cold	82	1.45	0.4154
medicin	9	11.20	<0.0001
fever	13	26.20	<0.0001

Table 3.1: Probability of keywords being Tweeted by a user during the month that he or she was diagnosed with influenza.

and serve as a set of keywords that may have been chosen by a domain expert. We use Fisher's exact test to compare keyword occurrence in months when the user is sick or not sick and find a significant effect for six of the seven keywords (See table 3.1). Additionally, we try algorithmically selecting keywords by first finding the 12,393 most common keywords in the data set (words that occur atleast twice). We then rank them based off of information gain on predicting influenza and choose the top 10, 100 or 1000 keywords from the list. A list of the most statistically significant, positive signals is available in appendix A. In all of these cases, we pre-process the data by tokenizing the text on spaces, tabs and line breaks and the characters ".,;':"()?!/\", remove stop words[2], perform Porter stemming [84] and convert the text to lower case. We use Naive Bayes, random forest, J48 (a Java implementation of C4.5), logistic regression and support vector machines to classify a user as being sick in a given month or not (see figure 3.2).

Second, we consider analysing the content of a tweet's text for messages giving hints about being sick such as "another doctor's appointment Wednesday ... have to #treatmyflu" or "I didn't realize how bad it feels to have the flu, should have gotten a flu shot[3]" that would not be detected through simple bag-of-words techniques. Computational approaches for natural language processing are available. However, because our dataset is relatively small, we use a 'human' classifier by hand rating all 1609 tweets that were posted by individuals during the time of their illness. We also sample a randomly selected set of 1609 tweets from times when the users did not have influenza as a control. We find 58 tweets from 17 ($17/35 = 48.57\%$)

[2]Stop words were taken from Weka's stop list version 3.7.10.

[3]These examples are based off of real tweets, but changed to keep our participants anonymous.

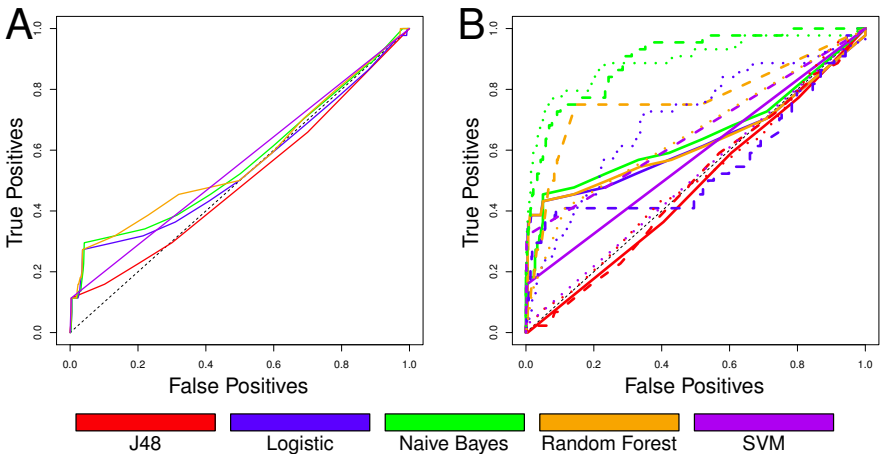


Figure 3.2: The ROC of classifiers that use hand chosen keywords (a) and algorithmically chosen keywords (b) to determine if an individual is ill. The top 10 (solid line), 100 (dashed line) and 1000 (dotted line) were selected as the features.

Sick	Not Sick	
17	18	Sick
0	66	Not Sick

Table 3.2: Confusion matrix of a Tweet-Classification based diagnosis system. Rows are of true values, columns are of predicted values.

individuals in our study that are about the user being sick. We also find zero tweets about the user having influenza during times when they did *not* have influenza. Because humans are very good at extracting information from text, hand rating tweets allows for an approximately 100.0% accurate classification, although it clearly does not scale well. Extracting information from text using machine learning is a complex problem where finding solutions that perform as well as humans is rare. Thus, the human classifier gives us an upper limit to the accuracy of a health monitoring system based off of tweet classification (see table 3.2.)

3.4 Frequency Based Signals

In addition to illness affecting the content of individuals' tweets, it is likely that illness also affects the rate at which individuals tweet. To detect this, we perform one-dimensional anomaly detection on each user's monthly tweeting rate as follows.

First, we calculate the number of tweets in each month in the study period and discard any months where the user tweets less than ten times. This avoids issues caused by the user starting or stopping their use of Twitter. We then calculate the z-score of the tweeting rate of the month that the user is ill by

$$z = \frac{|x - \bar{x}|}{\hat{s}} \tag{3.1}$$

Where $\bar{x}$ and $\hat{s}$ are the estimated mean and standard deviation of the user's tweeting rate for each month during the study [85]. We repeat this process for months when the user is not sick. We then classify the user as sick if $z > 1.411$ where 1.411 was chosen through leave one out cross validation. We find a significant difference between the z-scores for months when a user had influenza and months when the user did not ($p = 0.01303$, two-sample Kolmogorov-Smirnov test). Most of the time individuals are not sick (219 / 258 = 84.88% of the months), resulting in a highly biased sample. Thus we optimize based on the F_1 score instead of accuracy. The optimal z-score cutoff results in an area under the ROC curve of .6218 and $F_1 = 35.0\%$. (See table 3.3.)

Sick	Not Sick	
14	25	Sick
27	192	Not Sick

Table 3.3: Confusion matrix of the classifier based on anomalous tweeting rates. Rows are of true values, columns are of predicted values.

3.5 Network Based Signals

Even if a user is not currently active on Twitter, users on her social network may give clues to her health status. Twitter's social network is one directional, allowing for users to follow other users without the other users having to follow them back. Accounts that follow a user are referred to as her 'followers,' and accounts that a user follow are referred to as her 'friends.' We consider all text that a user's friends or followers tweeted and perform keyword analysis. The analysis was performed the same way as we analyzed the user's tweets in section 3.3, except we normalize the counts here by the total number of characters her followers or friends tweeted. This

controls for the number and activity of a users friends or followers, which should not have an effect on her health status. We find that most of the tested classifiers are able to detect a signal in both the user's followers' and friends' streams (see figure 3.3.)

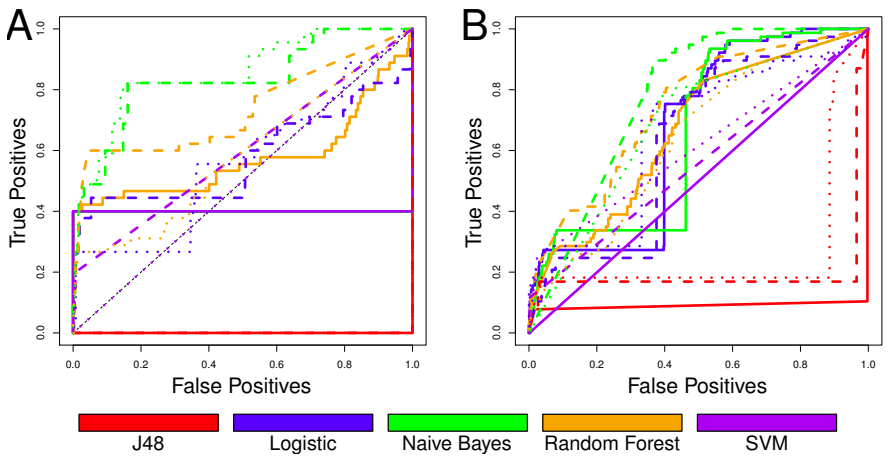


Figure 3.3: The ROC of classifiers based off Tweets from (a) accounts that follow a user and (b) accounts that a user follows. Line coloring and style are equivalent to figure 3.2.

We further analyse the strength of these classifiers by building each classifier using 10 fold cross validation and calculate their performance by measuring area under the ROC curve. We repeat this 100 times to generate a distribution of each classifier's performance. We then perform an analysis of variance test to examine the differences between the sources of data (followers or friends), the number of keywords used and the classifier's algorithm (see table 3.4.) We find that the choice in classifier and the length of the feature vector have a significant effect on performance. We find that classifiers that use tweets from accounts that follow the user are significantly better at diagnosing the user than classifiers that use tweets from accounts that the user follows. This may be because Twitter users often follow celebrities and news organizations – and celebrities and news organizations rarely follow personal Twitter accounts – which could introduce excess noise.

	Df	Sum Sq	F value	Pr(>F)
Source	1	107.16	1290.82	$<2^{-16}$
Keyword Size	1	72.19	869.66	$<2^{-16}$
Classifier	3	752.55	3021.61	$<2^{-16}$
Residuals	109194	9602		

Table 3.4: Results from an analysis of variance of the area under the ROC curve for classifiers based on tweets from an individual's social network. Factors are whether the data is from the user's friends or followers, the number of keywords chosen and the classifier.

Classifier	Area under ROC	Accuracy
AdaBoost	.9961	99.53
Bayesian	.9078	92.08
Decision Tree	.9877	99.22
Logit Boost	.9986	99.22
Weighted Voting	.9783	93.17
Baseline	.8544	89.72

Table 3.5: Performance of the meta classifiers. The presented baseline is the classifier based on datamined keywords – the highest preforming individual classifier.

3.6 Meta Classifier

So far we have considered five separate methods for detecting illness based off of a user's Twitter activity: hand-chosen keyword analysis, datamined keyword analysis, hand classified tweets, anomaly detection and network analysis. However, there is no reason that we cannot combine these methods to get a stronger signal. For example, while mining the user's text is the best of the five methods, she may stop tweeting while sick, which would be detected by the frequency-based anomaly classifier. Aggregating multiple classifiers by a 'meta-classifier' has been shown to be an effective method for increasing classification accuracy [86, 87].

We start by selecting the classifier from each of the previous five approaches that has the largest area under the ROC curve (see figure 3.4.a.) We then use the predicted distributions from these classifiers as the feature vector for the meta classifier. We use Ada Boost, Bayesian classification, J48 decision trees, logit boost, and weighted voting to evaluate the meta-dataset. We then evaluate these methods

with leave-one-out cross validation and see an increase in area under ROC and accuracy compared to the best individual classifier (see figure 3.4.b.) We find that AdaBoost has the highest accuracy (99.53%) and logit boost has the highest area under it's ROC curve with .9986 (see table 3.5.)

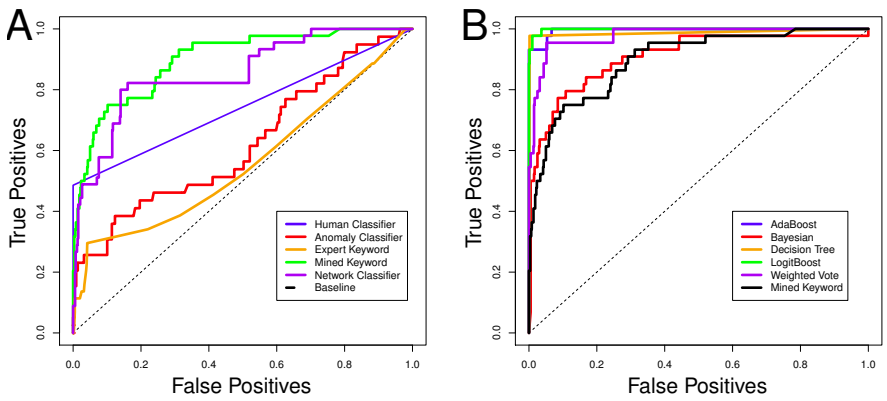


Figure 3.4: The accuracy of the previous classifiers (a) and the accuracy of various classifiers that use the previous classifier's results as features (b).

3.7 Conclusions

In this paper, we have shown that it is possible to diagnose an individual from her social media data with high accuracy. Computational approaches to aid in disease diagnosis has been approached before, however they have been developed with a medical setting in mind. That is, the question addressed was "can we diagnose an individual based off data gathered from medical tests run on her?" instead of "can we diagnose an individual solely based off of publicly available social media data?" While we focus on the relatively benign case of remotely reconstructing a confidential diagnosis of influenza, these methods could also be applied to stigmatized diseases, such as HIV [88], where being able to determine if an individual is HIV positive without her knowledge and with only her Twitter handle could result in serious social or economic effects. Half of the users explicitly stated that they were sick, and we were able to confidently determine illness in the other half of the cases through their data. It would seem that simply avoiding discussing an illness is not enough to hide one's health in the age of big data.

Chapter 4
A longitudinal study of 15 million people to measure disease dynamics

4.1 Introduction

An accurate model of disease transmission is necessary for efficient methods of disease control. [11, 16, 50, 89–94] However, disease transmission is often either estimated at a large scale, through population trends, [50, 89, 95, 96] or in small subsamples of estimated peer to peer transmission. [16, 51, 52, 91, 97, 98] Both of these approaches have flaws, however. Large scale, population trends require the development of costly disease surveillance systems [11, 50] and are only studied on a regional level. Additionally, these systems, which are often built on top of regional health providers, assumes that their samples are representative of the population as a whole. This may not be the case for relatively mild diseases, such as influenza, where only a small amount of the population may actually visit their doctor for treatment. For example, the CDC only recommends people in high risk groups to visit a health provider. [99] Small, peer-to-peer studies are also flawed due to issues related to determining who has actually come in contact with an infectious individual, the effort required to perform a study and the tendency for these studies to be retroactive. [16, 91, 94]

Here, we develop a novel method of influenza transmission detection through a longitudinal analysis of 15 million Twitter users over a four year period. This

approach addresses five of these six issues. Our approach can exploit data collected through third party social media platforms (in this chapter, Twitter) which is readily accessible instead of needing to partner with, or develop, a disease surveillance system. We use information provided by each user's GPS equipment to gain hyper-local information about the disease, compared to city, state or national regions provided by traditional surveillance systems. Twitter access is essentially free, compared to going to a health care provider which potentially reduces biases caused by socio-economic factors, although the Twitter population sample will introduce its own biases. We do not address the issue of determining actual contacts, although our results appear stable over a range of potential contact distances. As our system is automated, it scales well compared to traditional disease spread surveys. Finally, our data is collected in real time, eliminating issues related to subjects having a faulty recollection of past events.

In this chapter, we focus on the base reproductive number R_0. Traditionally, research has calculated either R_0 which is defined as the number of individuals a sick individual will infect over her disease's time span with the assumption that all other individuals are susceptible to infection or the effective reproductive rate R_E. [89] The effective reproductive rate is easier to observe "in the wild" as it controls for the effects of unobserved resistance to disease or non-uniform mixing of the population. [95, 96] Hence, we can track each individual in our subpopulation to estimate the amount of resistance and network effects, allowing us to calculate the base reproductive number to be 2.6945.

The remainder of this chapter is organized as follows. In section 4.2.1, describe the modifications to the classifier built in chapter 3 to scale to our larger dataset. In section 4.2.2 we describe the collection, storage and processing of our large Tweet dataset. In section 4.2.3.1, we describe a classical method of determining R_0 and our system can replicate this method. In section 4.2.3.2, we develop a novel form of disease modeling to determine R_0 that employees additional data our Twitter-based-surveillance-system provides. In section 4.3, we compare the results of our methods to previous studies' results. Finally, we conclude by describing potential biases–due to our inferred contact networks (section 4.4.1) and spam messages (section 4.4.2)–along with our approaches to controlling for them.

4.2 Methods

4.2.1 Building a Validated Diagnosis System

We begin by developing a system that is capable of accurately assessing influenza cases in individuals based on their Twitter feeds. To do this, we collected data on 104 Twitter users that were *professionally* diagnosed with influenza, and 122 Twitter users that self reported as *not* having any influenza-like-illnesses during a one year period of data collection as described above, in chapter 3. Next we trained a machine-learning classifier to differentiate between these two user groups. To allow the classifier to scale, we do not include the long term tweeting rates, followers network, or hand rated Tweets into this version of the classifier. The simplified version of the classifier performed with an accuracy of 89.72% and an area under the ROC curve of 0.8544. (See the previous chapter for further details.) Since an influenza infection is acute, we grouped each user's time line into monthly slices, which is defined as being a time of illness if the user was professionally diagnosed during that non-overlapping window.

We then apply the classifier to each user's time-line on the 4-week sliding window, with each step of the sliding window being one day. The classifier assigns a score to the day where the sliding window begins based on the Tweets the user has posted within the window. For example, when the sliding window first encounters a user's Tweet that says "I am getting sick with the flu," the classifier will heavily lean toward her being sick. Later, the user may Tweet "I am no longer sick" which will give a strong signal that the user is no longer sick which will tend to outweigh the user's previous "sick" Tweet even if they both occur in the same window. Of course, it is rare that such strong signals are in the data, so the classifier is built on an amalgamation of many weaker signals—mentioning going to a party as a not-sick signal, for example—which, while weaker, are more prevalent. We chose a step size of one day in order to increase the temporal granularity of the classifier. Users that are inactive for more than 30 days are not included for *any* analysis during that time window.

4.2.2 Twitter Data Collection

We collected almost all Tweets from the continental United States with high-resolution geo-spatial information over a four year period from March 3^{rd} 2011 to March 4^{th} 2015. Twitter allows users with GPS equipped devices such as mobile phones to opt in to sharing high-resolution geospatial data with each of their Tweets. This data is mainly ($> 99\%$ of the time) as a point defined by longitude and latitude, with the remaining portion consisting of bounding boxes. For compatibility, we converted these bounding boxes to the midpoint of the box. To collect this data we queried Twitter's streaming API with a request for a bounding box that covers the area in interest, thus insuring the collection of geo-taged Tweets with high resolution geo-spatial information. While this limits our data collection to a subset of all Twitter users, it has two substantial advantages. First, the high geospatial resolution allows us to study patterns that occur over short distances, such as disease transmission. Second, this geo-filter allows us to obtain almost all Tweets that match our filtering criteria without invoking Twitter's rate limits. [100] A total of 2,732,174,105 Tweets from 15,560,328 users were collected during this four year time period.

The collected Tweets were then processed through a combination of custom Hadoop programs and Apache Hive scripts on Amazon's Web Services. Each Tweet, when collected through the Twitter API, is associated with a variety of metadata. We parsed each Tweet and stored a simplified version containing user id, the Tweet's text, the time the Tweet was posted and the latitude and longitude describing the location where the Tweet was posted and stored it on a Hadoop Distributed File System. Each user–identified by user id–is then tested for disease based on the method described above in section 4.2.1 for each day that he or she was active. We discard users from our dataset who Tweeted less than 10 times over the entire four year period, as it is unlikely that they will have provided enough data to be useful. Finally, we determine the location of each user based on her longitude and latitude and store her state and HHS (Health and Human Services) region. This was then done by finding the mean location of each user by averaging all of the locations of her Tweets. Then, the user's location is then mapped to a state (or District of Columbia) using a hive plug-in [1]. Users that were *not* in any state tended to be

[1] https://github.com/ToddBodnar/GeoUDF

in northern Mexico or southern Canada and were discarded from analysis. Each HHS region is simply an aggregation of several states, making it relatively easy to convert from state to region.

4.2.3 Estimating Transmission Parameters

4.2.3.1 Parameter estimation through curve fitting

While the above system has been shown to perform well on the case study data [77], the data may be biased, limiting the classifier's ability to generalize to an arbitrary set of Twitter users. To show that this isn't the case, we consider applying our classifier to our full, high-resolution geospatial Twitter dataset. If trends discerned from this application match previously measured real-world results from the CDC, then we will gain some confidence on our system's ability to generalize. Here, we consider the traditional approach of determining the basic reproduction number, R_0, at the population level. To do this, we consider a SIR (Susceptible-Infectious-Recovered) model of disease. First, we build these models using the standard transition equations for the SIR model:

$$\frac{dS}{dt} = -SI\beta, \qquad \frac{dI}{dt} = SI\beta - I\gamma, \qquad \frac{dR}{rt} = I\gamma \tag{4.1}$$

Where S,I and R are the frequencies of susceptible, exposed, infectious and recovered individuals, respectively. [96] Note that the parameters β and γ are the transition probabilities from being susceptible to having the disease and recovering from the disease, respectively. Additionally, we could consider SEIR models for parameter fitting. However, recent work has shown that inferences made from SIR models out perform inferences made by more complex models such as SEIR [50, 95] and that solutions for SEIR models are not unique [89]. Additionally, "[a] single-age class SIRS model inference system is able to reliably infer the leading eigenvalue of the effective reproductive number." [50] As the purpose of this section is to just give a base-line for user-based analysis, we do not further consider SEIR models as SIR models are simpler and more accurate.

We can now find the parameters of β, γ and starting values for S, I and R that cause the model to best fit our data using a multi-grid search. Specifically, we search through three variables, the two transmission parameters γ and β and $S(0)$,

the initial susceptibility rate which may be less than 1 due to innate immunity or previous vaccination. Next, we generate a logarithmically spaced 25 by 25 by 25 grid of potential values over this range. We then set $I(0)$ to be the same as the first infection value in the data and $R(0)$. We then solve an SIR model, with each of the parameter combinations. Finally, we compare the results of the model to the data based on

$$error = \sum_t (I_{\gamma,\beta}(t) - I_{data}(t))^2 \tag{4.2}$$

We then find the value of γ and β which have the smallest error. We then recenter the variable's ranges on these new values and reduce the range to search by an order of magnitude. This process is repeated 25 times, at which point the minimum and maximum range tested differed less than the machine's precision ($\approx 2 \times 10^{-16}$).

We can generalize this method to either arbitrary subset of any (from the CDC or generated from Twitter) disease incidence curve by comparing the SIR model for a given γ, β pair to just a subset of the time-steps or to multiple incidence curves. Note that one could employ more efficient methods of parameter fitting such as Euler's method, [95] likelihood estimation [50] or genetic algorithms, [101] however, the small number of parameters and the reasonably small number of time steps in each incidence curves allow for a naive grid-search to be executed quickly.

Because the influenza virus constantly mutates between flu seasons, previous infection does not necessarily confer a resistance to future infections. For example Cowling et al. [102] do not find a significant difference in seasonal influenza rates based on the previous year's vaccination. To simplify our analysis, we model each year separately, not assuming any cross-protection. We consider both sharing the same values of parameters for every year and refitting the model to each year. As we have geographical information, we additionally build these models for each of the 10 HHS (U.S. Health and Human Services) regions. As a proof of concept, we also perform these methods on the county level. Specifically we use Tarrant County in the state Texas and King county in the state of Washington as two case studies. These locations were chosen due the public availably of surveillance data from the local public health institutions. Note that both of these counties have large population sizes, with Seattle, Washington and Fort Worth, Texas being two large cities in those two counties, respectively. From each of these models, we calculate R_0 based on the estimated parameters by using the relationship

$$R_0 = \frac{\beta}{\gamma}. \tag{4.3}$$

4.2.3.2 Parameter estimation through individual user analysis

Next, we consider employing both the high-resolution spatial data and the individual Twitter-based diagnoses to develop a network-based model of disease transmission. We begin by constructing a pseudo-contact network between each of the Twitter users based on their geo-spatial information in our dataset. We consider two users to be in contact if they are less than a distance d from each other. We assign the health state (sick or not sick) to each user to the network at a daily temporal resolution. From this, we can determine whether or not being located near a sick user increases one's likelihood of getting sick shortly after the potential exposure. While the Twitter users in our dataset are only a small subsample of the general population, we would still expect to see this effect in our data due to the close (droplet-borne or airborne) proximity requirement of influenza transmission. [98, 103, 104]

Additionally, we estimate R_0 for the disease from the individual infections as follows. As with the disease transmission likelihood estimation above, we begin by considering neighboring users that were ill previous to the time that an user became ill as potential candidates as being the exposing individual. Since we cannot determine which of the i ill neighbors infected the individual, we assign each ill neighbor $\frac{1}{i}^{th}$ of the responsibility for infecting the individual. Each user's cumulative responsibility for infecting *all* other cases of illness is then the expected number of people the user has infected. We then calculate the average R_0 of the disease by averaging the expected number of infections for all of the users that were sick.

When determining the responsibility of infection, we assume that an individual became infected the first day that she is classified as being ill, minus a lag l_t, that we vary in the model. This lag is due to both the incubation period and the fact that individuals do not necessarily report symptoms to Twitter on their first day of illness. Additionally, we assume that an individual gains an immunity to the specific strain of influenza after becoming ill and will thus not become infected again until the next flu season.

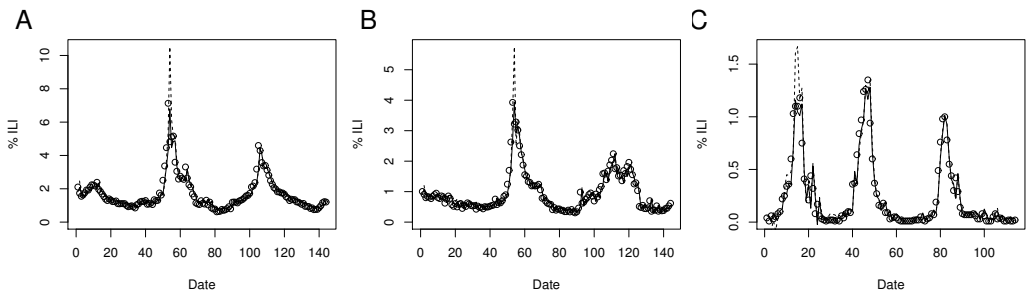


Figure 4.1: Comparison of Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) to the CDC's reported Influenza rates (circles) for national (A), HHS Region 1 (B), and Seattle area (C).

4.3 Results

Flu Season	CDC Data	Twitter Data
2011-2012	2.053	1.997
2012-2013	2.044	2.200
2013-2014	2.044	2.178
Combined	1.854	2.087

Table 4.1: Estimated R_0 based on CDC and Twitter data.

Year	γ	β	Sum Square Error
2011-2012	0.1732	0.1749	0.0001047
	0.1176	0.1195	0.0001323
2012-2013	0.7715	0.9626	0.0009402
	0.7317	0.9020	0.0009492
2013-2014	0.6054	0.7288	0.0003114
	0.6046	0.7264	0.0003026
Combined	0.6998	0.8225	0.003719
	0.6765	0.7935	0.003252

Table 4.2: National best-fit parameters for each year from the CDC's data (white) and Twitter data (gray).

County	Flu Season	CDC Estimates	Twitter Estimates
Fort Worth	2011-2012	1.14	1.14
	2012-2013	1.542	1.576
	2013-2014	1.454	1.443
	Combined	1.375	1.38
Seattle	2011-2012	6.073	4.934
	2012-2013	11.7	11.7
	2013-2014	1.000	1.117
	Combined	4.053	3.415

Table 4.3: Estimated R_0 based on CDC and Twitter data with 5% and 95% percentiles in parentheses for the two proof-of-concept county datasets. Additionally, T-Tests and Kolmogorov-Smirnov Tests are performed to compare results.

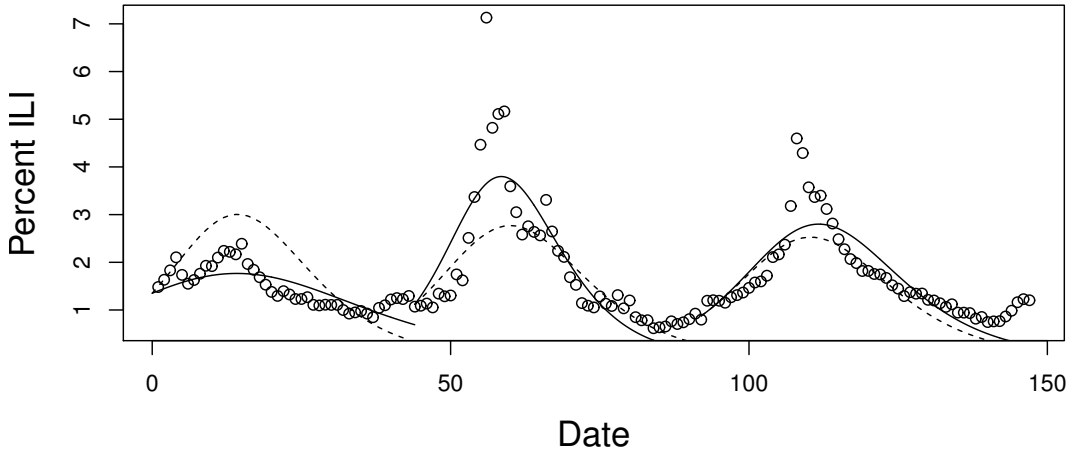


Figure 4.2: The CDC's estimates (circles) of influenza rates for a three year period compared to the best fit SIR models from the Twitter data using combined (dashed line) or yearly (solid line) parameters.

4.3.1 User Activity Summaries

Before discarding low-activity users, our dataset consisted of 2,732,174,105 Tweets from 15,560,328 users, resulting in a mean of 175.59 Tweets posted per user. The tweet count distribution was highly dispersed, with a maximum of 1,119,384 Tweets (approximately 767 Tweets per day), a median of 10 Tweets, and a minimum of one Tweet over the study period. On the lower end of the spectrum, it is unlikely that

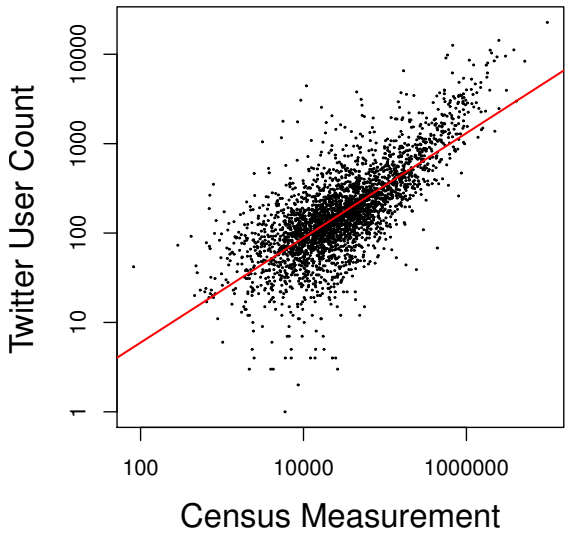


Figure 4.3: The relationship between the US Census's population count and number of Twitter users in our dataset.

users who posted less than 10 Tweets over the 4 year period will have provided us much information. Therefore, as described above, we discarded them from our dataset for our analyses. On the other hand of the spectrum, the group of most prolific Twitter users may contain automated spam bots. As we describe in section 4.4.2, we used Twitter's API to identify spam bots and found 3% of the users to be spam-bots, but removing them did not significantly effect our main results.

The geo-spatial distribution of the Twitter users appears consistent with the U.S. population distribution. To validate this observation, we determined in which county the user is located by comparing, for each user, the mean longitude and latitude of their Tweets to the US Census Bureau's county shape files.[2] We then compared the count of users in each county with the 2010 Census population count and find a strong relationship (Spearman's $\rho = 0.708$, see figure 4.3). As both the Twitter user count and Census population count follow a long tail distribution, they are first log-transformed before compared by Pearson's coorelation coefficient.

[2]https://www.census.gov/geo/maps-data/data/tiger-line.html

Distance (km)	$\bar{K}$	Variance(K)	$1 + Variance(K)/(\bar{K}^2 - \bar{K})$
0.1	42110.48	4.1911×10^{10}	3.3635
0.5	42144.56	4.1973×10^{10}	3.3632
1.0	42155.25	4.1989×10^{10}	3.3629
5.0	42383.88	4.2267×10^{10}	3.3529
10.0	43089.55	4.3101×10^{10}	3.3214
50.0	63776.20	6.9418×10^{10}	2.7067
100.0	112829.72	1.6611×10^{11}	2.3049

Table 4.4: The average and variance of the number of other users that are within a given user of a distance and the R_0 modification factor.

4.3.2 Diagnostic Validation

Before we use our Twitter diagnoses as a proxy for actual disease measurements to determine disease transmission parameters, we must first validate that our diagnoses match expected trends. We have previously shown that our diagnoses classify professional diagnoses of a small sample of Twitter users relatively well, but it is possible that the small sample size of that study biases our diagnostic method in a way that limits its ability to generalize to the full population of Twitter users. To investigate this possible bias, we aggregated all of our users in an area, use our classifier to assess disease status and then calculate the disease incidence based on our classification. If there were a strong bias to our classifier, these Twitter incidence curves would *not* be expected to match with the CDC's incidences curves.

Here, we present the comparison of those incidence curves, i.e. from our Twitter data and the CDC data. The incidence of disease is not normally distributed, hence we use Spearman's rank correlation coefficient instead of Pearson's correlation. We note that our measurements perform well, with the fit at the national scale ($\rho^2 = 0.90761$) out performing Google Flu Trends ($\rho^2 = 0.8207$). We find that the regional ($0.8113 \leq \rho^2 \leq 0.9467$ in each HHS region) and local measurements for Tarrant county ($\rho^2 = 0.8944$) and for the Seattle area ($\rho^2 = 0.7295$) also fit well. Additionally, we look at the predictive power of the model by training the model on data from weeks $1...t-1$ and assessing its ability to predict the incidence at time t. While these models are not as effective as models that fit to the entire dataset, they do perform well on the national ($\rho^2 = 0.90761$), regional ($\rho^2 = 0.9385$), and local ($\rho^2 = 0.6976$) scale. For brevity, only one representative regional and local fit

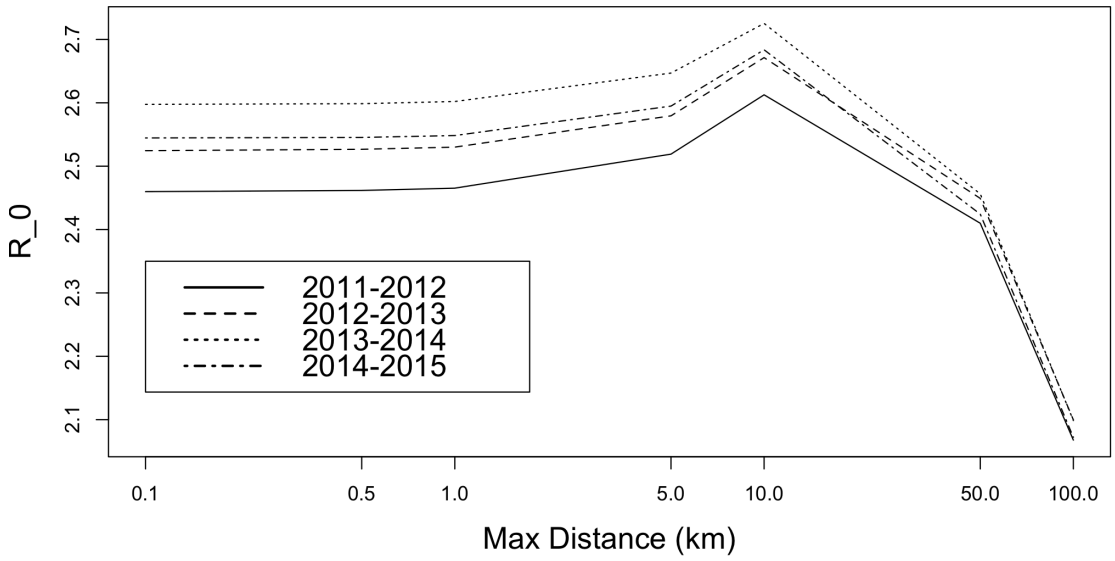


Figure 4.4: Estimated peer-to-peer transmission rates based on maximum distances between users.

are presented in the main paper (see figure 4.1). Full model fits are available in appendix figures C.1 and C.2.

As described in section 4.2.3.1, we estimate national-level influenza disease parameters,β and γ, based on the CDC and Twitter incidence curves. We find that the resulting estimate R_0, based on the CDC data and our Twitter data do not differ substantially on the national scale (min difference = 0.002, max difference = 0.015, see table 4.1) or for the regional scale data (min difference = 2.725×10^{-4} , max difference = 0.106, mean difference = 0.0254, see table B.1), except for one region and one year (HHS Region 3, 2011-2012) where the Twitter estimate for R_0 was 501% higher than the R_0 estimate from the CDC, possibly due to relatively small data size in the first year or the mild flu season. Additionally, the R_0 for the two sample counties is estimated well by our Twitter data (min difference = 2.754×10^{-4} , max difference = 1.139, mean difference = 0.2168, see table 4.3), but not as closely as the regional or national datasets, possibly due to the small scale of county level data. Note that we also estimated values of R_0 based on the assumption that the transmission parameters do not vary between seasons, but the model does not fit the data as well (see figure 4.2) and thus is not included in the analysis.

4.3.3 Individual Parameter Fitting

The base derivation of R_0, as in equation 4.3, assumes that each individual has the same number of contacts. However, it's been long know that this assumption is often invalid and that contact networks can exhibit substantial variation in the distribution of contacts. This increased variance, when not taken into account will lead to an underestimation of R_0 [90, 105, 106] and is thus typically adjusted using the formula

$$\hat{R_0} = R_0 \left(1 + \frac{variance(k)}{\bar{k}^2 - \bar{k}}\right) \tag{4.4}$$

where $\hat{R_0}$ is the adjusted version of R_0 and k is the number of contacts. [90, 106] Note that in the case of a uniform distribution of contacts, the variance of k is zero and the adjusted version $\hat{R_0}$ is identical to R_0.

Here, we begin to calculate this adjustment by determining $k_u(d)$, the number of neighbors of user u given a search distance of d. This is done by iterating through all users and counting the number of additional users that are within d of each user. Note that since Twitter's geolocation is based on longitude and latitude, the distances must be converted into kilometers before compared. We can now calculate the mean and variance of k between each user for a given distance and calculate the necessary adjustment to R_0 (see table 4.4).

We then apply this long tail adjustment to the mean cumulative infection responsibility calculated above (see table 4.4) to determine the value of R_0 for a given maximum distance and incubation period. Note that these effects are similar for a range of distances and lags (see figure 4.5). The incubation is not significantly different for values less than or equal to 12 days ($p > .05$ after adjustment, pair wise, logs scaled, t-test). As it would be unwieldy to report all combinations of these values, we choose to only report based on a lag of 7 days and a maximum distance of one km.

With these parameters set, we can now study the flow of disease between users. We find a total of 182,801 infectious users. Note that we cannot detect all infections, as there are individuals that a user may have infected that were not in our dataset. We find that the adjusted number of expected infections an individual causes follows a long tail distribution (min = 0, max = 379.6000, mean = 2.6940, median = 0.8407, std = 6.8069, see figure 4.6).

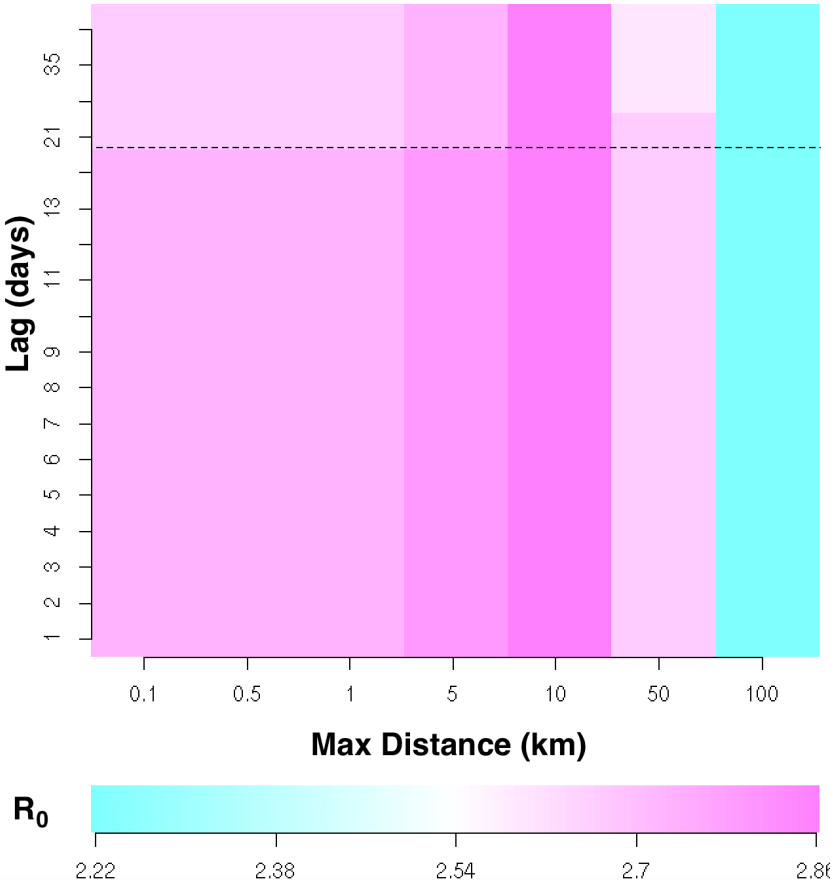


Figure 4.5: Effects of differing time and temporal windows on predicted R_0. Note the increase in time after 14 days.

4.4 Discussion and Future Work

4.4.1 Describing Pseudo-contacts

Note that our definition of contact for disease transmission does not necessarily represent true contacts. Indeed, as we are only working with a subset of the entire population—people in the United States that are active Twitter users with geotagging activated—we can make no claims about *all* contacts and transmissions being recorded. Hence, we likely incorrectly observe transmissions from user A to user B which actually involve a transmission from user A to user C to user B. This may also account for our model being consistent when the incubation period is set to two weeks, which is longer than the infection's length. [107]

Additionally, we are limited to the accuracy of the user's GPS when it comes to

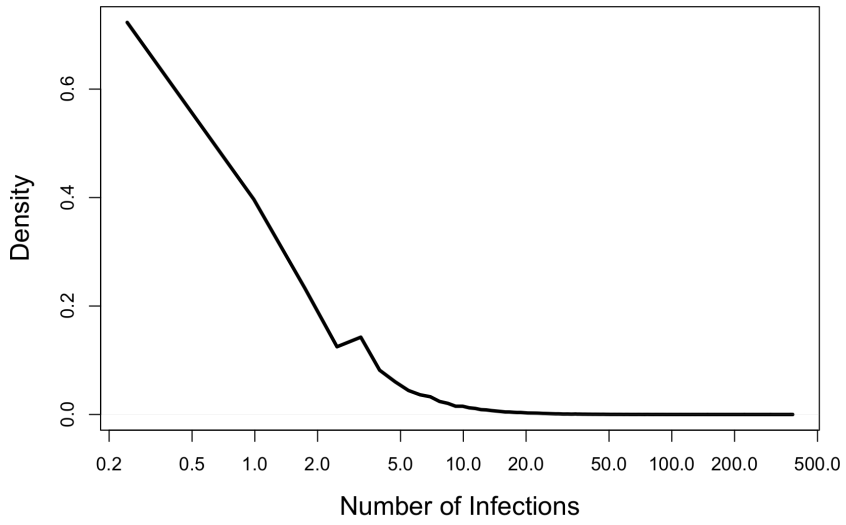


Figure 4.6: The adjusted number of individuals that a person is likely to infect during her disease. Note that the log-transformed x-axis does not include cases where zero transmission occurs.

determining the distance between two users. Previous work by Glidden et al. [108] find that, unlike previous claims [109–111], personal GPS devices do *not* have sub-meter accuracy. For example, they find that more than 80% of iPhone users are within 0.5 miles of where their GPS reports them to be and 4.5% of iPhone users were more than 2 miles away from their reported location. Indeed, this appears that Twitter controls for this by limiting the reported locations. We find that 99% of users with two tweets Tweet in the *exact* location, with a mean of 8.79074e-5 degrees ($\approx$ 9.8 meters). This may also be an attempt by Twitter to limit user's from releasing private information about themselves (for example, see [112]). While this limits our ability to accurately work on sub-kilometer distances, it also allows us to average each user's location across time to simplify the contact network to a static set of connections (although users/nodes vary over time due to user activity).

4.4.2 Spam Removal

So far, we have assumed that all of the collected Twitter accounts may provide accurate information about a person's disease state. However, this is not the case. Specifically, spam "social" bots [113] may attempt to emulate real users, but their disease state would be meaningless. Indeed, much work [114–116] has been done on

the topic of automated spam-account detection on Twitter. However, reproducing these papers' work appears difficult as they tend to rely on large, unaccessible hand coded data. Instead, we consider using Twitter's own API as a method to detect spam bots.

While Twitter does not directly supply a spam-API, attempting to look up a user that was banned results in a specific error returned, instead of the user's data. This method will overestimate the incidence of false accounts, as Twitter users can be banned for non-spam actions such as hate speech or copyright infringement. However, for our case this is acceptable since any effects from removing banned accounts will overestimate the effects from removing the subset of banned accounts that were bots. Additionally, while no spam detection system is perfect, it is likely that Twitter, itself, has a stronger motivation, more resources, and access to additional meta data to develop accurate spam detection than third party researches.

Here, we re-run the individual based R_0 calculation from the 2011-2012 influenza season with banned accounts removed. This choice in year allows for a smaller number of queries to Twitter's API to test if accounts are banned and sidesteps issues with new accounts that will be, but have not yet been, banned. Of the 45,086 Twitter users active in this first year, 1331 (2.95%) were banned as of April, 2015. When we repeat the analysis in 4.2.3.2 on the remaining 43755 users, we find that the mean of R_0 was reduced from 2.465 to 2.417, but this amount is not statistically significant (log-adjusted two-sample t-test, $p = 0.4497$). Hence we conclude that additional spam-removal for the remaining influenza seasons is unnecessary.

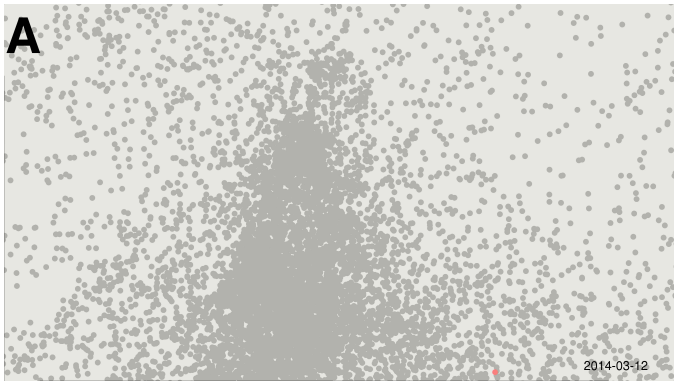
4.4.3 Visualizing Disease Spread

Our methodology allows us to visualize disease spread in a novel way. Here we present a potential method of visualization. Here, we present three time slices of disease dynamics over three different geographic scales. First, we consider an outbreak in Seattle, Washington in March, 2014 (see figure 4.7). Note the clustering in the south center region of new disease cases in the later two weeks. Second, we consider a region consisting predominately of Pennsylvania with two major metropolitan areas (see figure 4.8): Pittsburgh is in the southern left quadrant and Philadelphia is in the southern right quadrant. Note the variation in disease

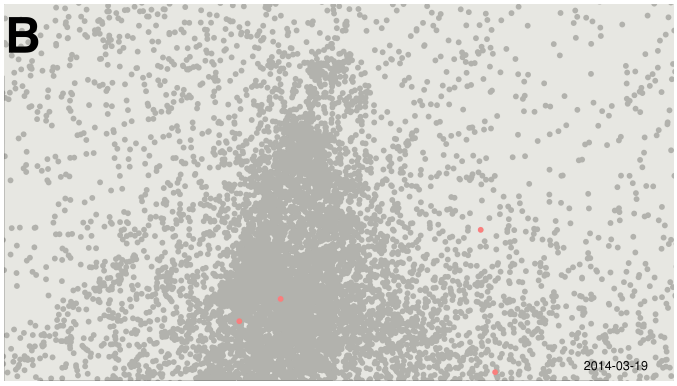
incidence between the the two cities. This implies a difference in disease rate between the two cities, although it is difficult to validate due to Pennsylvania not supplying regional data. Finally, we visualize the entire United States (our entire dataset) over a six month period (see figure 4.9). Note that disease rates appear lowest in September and higher during the winter months, as one would expect.

4.5 Conclusions

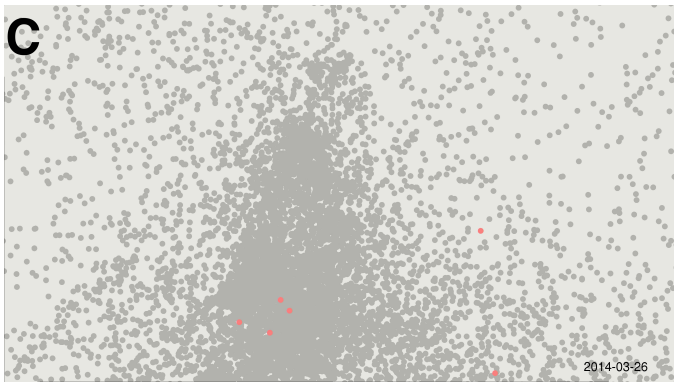
In this chapter, we presented a novel method of capturing peer to peer disease transmission using a combination of a large Twitter dataset and a tweet-based diagnosis system trained on medical records. We find that this is able to to both (1) replicate traditional disease surveillance systems on an arbitrary geographic level and (2) inform us about the dynamics of influenza infection. We provide analysis of R_0 on various levels of distance and incubation period (see figure 4.5) which both serves as a sensitivity analysis and provides novel insights about human mobility. We find the value of R_0 to be 2.6945 when we assume a one kilometer distance and a one week incubation period. Finally, we find evidence for super spreaders of influenza due to the long tail distribution of measured infections.



2014-03-12



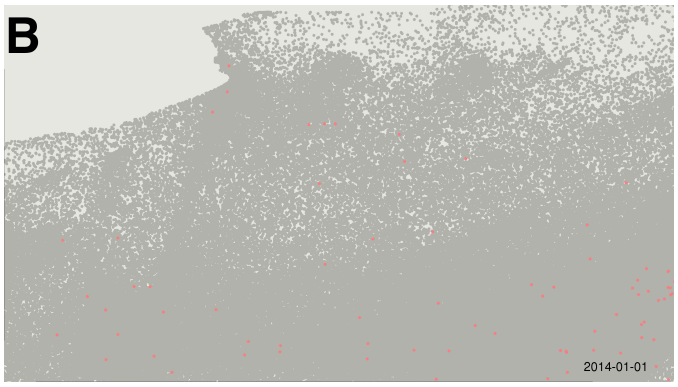
2014-03-19



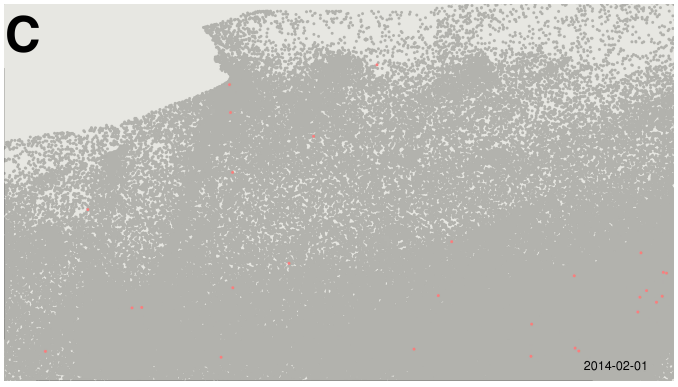
2014-03-26

Figure 4.7: Example of a cluster of influenza in Seattle over a two week period.

2013-12-01



2014-01-01



2014-02-01

Figure 4.8: Example of differing disease rates in nearby cities in Pennsylvania. Note the differences in Pittsburgh (south west) compared to Philadelphia (south east).

2013-9-01

2013-12-01

2014-03-01

Figure 4.9: Disease rates for three days over a 6 month period of our dataset. Best viewed in full screen.

Chapter 5
Beyond Network and Sentiment Analysis for Retweet Prediction

5.1 Introduction

Up to this point, we have only considered measuring behaviors and disease using social media. However, this knowledge isn't useful if it cannot be used to influence future disease prevalence or behaviors. Here, we consider Twitter as a source of information from which users may gain information about a disease and potentially modify their behaviors that relate to the disease. [40, 117] For example, previous work has shown that influenza vaccination rates can be inferred from Twitter [118] and that users modify their sentiment of vaccinations in a way that is statistically related to their exposure to other Tweets about vaccination. [117] Thus, a public health official would be interested in exposing as many Twitter users to as many messages about getting vaccinated as a way to *influence* their behavior. [40] A traditional approach to this would be purchasing ad-space, but this can be expensive. Instead, one could post a message on Twitter with the hope that others would read it and find it important enough to share with their friends. However, Twitter users will not blindly retweet a Tweet. Hence, we set out to find factors about a message that would encourage others to further spread the message which a Twitter user could use while crafting her Tweets.

We first collect Tweets from three health related events: the initial announcement of the novel influenza strain H7N9, a measles outbreak that was caused by a subset

of the population's refusal to receive the MMR (mumps,measles,rubella) vaccine,[1] and Autism Awareness Month. In addition to Tweets, we collected retweets and *inferred* additional message replication through text similarity metrics. These three events were chosen because they all occurred in April 2013—limiting effects of changing Twitter usage— and because they cover a range of health interests. For example, H7N9 incidence was limited to south east Asia at the time, so the majority of English speaking Twitter users was not directly affected by it. On the other hand, users discussing Autism often were tweeting about someone they know that has the disorder.

A Twitter user, that is interested in a large impact in one of these public health areas, would want to reach as many people in each of these three target groups. A trivial way to do this would be to send as many messages as possible, however, this is considered bad practice and may annoy others. Instead, several aspects of a message can be crafted to be more appealing to other readers. In this chapter, we consider both aspects that are hard to change (number of followers and what kind of account the message seems to come from) and aspects that are easy to change (message content and sentiment). We find that we are able to accuratly predict how many times a tweet will be retweeted ($0.4667 \leq r \leq 0.6665$, depending on the dataset).

5.2 Background

The task of predicting a Tweet's propogation is a well studied problem [53, 54, 56, 116, 119–123], even predating Twitter's implementation of retweeting (reposting a message) with the study of, for example, URL's included in a message [123] or a message's text's simularity to other Tweets [88, 124]. While the use of alternative methods of finding message propogation has fallen into disuse in the literature since Twitter released an API to detect retweets, we find that the retweet API fails to find about 10% copies of the message (see table 5.7). Indeed, Twitter appears to have recently begun removing these near-copies under copyright grounds, as it appears to be a bot strategy.[2] In this chapter, we consider tracking both retweets

[1] The vaccine that is incorrectly believed to cause autism.

[2] http://www.theverge.com/2015/7/25/9039127/twitter-deletes-stolen-joke-dmca-takedown Archived at http://www.webcitation.org/6ajikbuI5

along with "hidden" message reproductions. Preliminary analysis on the H7N9 dataset finds that models perform comparably on both datasets, (see Appendix D) so here we consider *only* the combined number of times a message is reproduced through either retweets or other methods.

To determine which attributes are related to higher message reproduction, we first consider four methods others have used to predict retweet count. First, the topology of the follower network is considered [121, 123]. That is, if a user has more followers, there are more users that may see his messages and may retweet them. This is generally modeled in the form of

$$log(\text{retweets}) \propto log(\text{out degree}) \tag{5.1}$$

due to the long tail distributions of both retweets and out degree [123]. In this chapter, we only consider models to predict the log-transformed retweet count. Second, we model the influence of different accounts by considering the account's purpose. We label each account as either being a news agency (News), a health organization (Health), a personal account (Personal) or an other account (Other). This was done with a combination of Amazon Mechanical Turk and machine learning. In the H7N9 dataset, we find a strong difference between these different accounts. For example, we find that an average of 3.98 Personal accounts retweet each Tweet a health organization posts, but only an average of 0.0017 Health accounts retweet each message a Personal account posts. Also of note is that health accounts tend to get more retweets per tweet than news organizations despite having less followers. Third, we consider keyword analysis to predict retweet rates. [54, 56] For example, Kim et al. [54] found that tweets that contained vulgarities were less likely to be retweeted. In the H7N9 dataset, we find that more formal phrases such as "human-to-human" and "transmission" have a positive relation to retweet count while more shocking words like "killed" and "lethal" actually had a slight negative relation to retweet count, although this could be confounded with the type of account. Fourth, we considered the sentiment expressed in a message as a predictor. [54, 125] However, in the H7N9 dataset we were not able to find an effect, possibly due to an overwhelmingly negative sentiment about the topic.

Sentiment analysis has been valuable for many applications such as measuring vaccination rates [118] or determining political affiliation. [55] However, in the case

of infectious disease, sentiment tends to mainly be negative, limiting the power of sentiment analysis. Here, we develop a more general version of sentiment analysis which we base on emotion research and affective science. [61, 62] Specifically, we consider Plutchik's model of emotion [61] which is based on four dimensions: valence, arousal, dominance, and aptitude. Valence corresponds to the attractiveness of a message, which is roughly equivalent to sentiment. Arousal corresponds to the energy behind the message. For example, we may find that messages with multiple exclamation marks show higher arousal than messages with out exclamation marks. Dominance corresponds to how argumentative the message appears. [60] Finally, aptitude corresponds to the inferred maturity of a message. [126] Note that we are interested in the emotion displayed by the Tweet, which does not necessarily correspond to the emotion of the person that made the tweet. For an example of both extremes of each of these four dimensions see Table 4.

5.3 Tweet Collection

We analyzed 947,967 tweets from March 31, 2013 to April 27, 2013. This period of time included the WHO's announcement of H7N9, autism awareness month, and an outbreak of measles. We did not hit any rate limits, so it is likely that this dataset contains all relevant, public tweets from the timespan. Tweets were selected if they contained one of a set of keywords. We defined the set of H7N9 related tweets as any tweet that contained the word "H7N9." Similarly, we defined the set of autism related tweets as any tweet that contained the word "autism", "autistic", or "Aspergers" and defined the set of measles related tweets as ones that contained the word "mmr", "mumps", "measles", or "rubella." We did not require that the text matched case with the keyword. Each tweet consists of a message of up to 140 characters long, the time it was tweeted, information about the origin tweet, if the tweet is a retweet, and information on the user that posted it, such as the number of her followers and her language.

5.4 Metric Measurement

5.4.1 Retweet Measurement

Retweets are the reposting of another users' Tweet on one's own Twitter account, generally signaling that the retweeter finds the message interesting enough to pass on to his or her own readers. The definition is somewhat vague, however, as there are multiple ways that a user could choose to repost a message. First, a user can use the build in "Retweet" interface that Twitter provides. This is generally considered a retweet, by definition. Second, a user can copy and paste another user's Tweet, often prepending "RT: @username" ("Retweet: User *username* said..."). This has the advantage of being able to add one's own commentary (for example: "This is scary. RT: @cnn_brk 'New cases of H7N9 announced."'), but has the disadvantage of potentially going over Twitter's 140 character limit. This manual retweeting predates Twitter addition of built in retweeting functionality. Third, a user could copy another Tweet without attribution to, for example, steal a joke or simulate that the account is used by a human.

Initially, we consider if a message has been reproduced by using Twitter's built in retweet interface. When a user retweets a message, the original, unedited message - along with the original poster's information - appears on her time line. However a user may copy a tweet without using the retweet tools or only partially copy a message. These additional types of copying are hidden from the Twitter API queries. For example, tweets that quote each other ("WHO announces new cases in China" and "Oh no, this looks bad! @cnnbrk 'WHO announces new ca...'") are clearly related but not retweets. However, tweets with different content are unlikely from the same original message ("19 new cases announced on April 9th" and "9 new cases announced on April 10th"). We develop a classifier to determine if two messages are of this type of reproduction as follows.

We create a training set of pairs of tweets and hand rate them as being related or not. We define 'being related' as having text that is similar enough that they are likely from the same origin. Because it is unlikely that a random pair of tweets are related, we sample from the dataset as follows. For each of the metrics above, we determine the theoretical minimum and maximum values. Two tweets can have a longest common substring of a length between 0 and 140 characters, for example.

We then divide this range into ten evenly sized bins. When we select a random pair of tweets, we first evaluate them based on the six metrics and place them in the appropriate bin or bins. However, we do not include, or hand rate, a pair if all of the bins they fill already contain 100 or more rated pairs. We repeat this process until all sixty bins contain at least 100 pairs. 4,897 pairs of tweets were hand rated in total.

We consider three metrics to determine relatedness: Levenshtein distance, longest common substring, and the number of shared keywords between two tweets. In addition, we calculate the normalized version of each of these three metrics by dividing by the number of characters in the longer string. We consider simple classification schemes based off of a cutoff of a single metric. In addition, we consider logistic regression, Ada Boost, C4.5, Naive Bayes, random forests and neural network classifiers that use the six metrics as the feature vector. We evaluate the classifiers with leave-one-out cross validation and find C4.5 to perform the best with an accuracy of 93.62% (See Table 5.1).

We then use this classifier to detect hidden network flow as follows. We begin by considering each tweet i that is not tagged as a retweet by Twitter. We then select all tweets T in our dataset that are both from a user that the user that posted i follows are were posted before tweet i was posted. It follows that tweets in the set T are the only candidates for hidden information flow on Twitter. For each tweet in the set, $j \in T$ we apply the classification algorithm – defined above – to i and j. Each candidate tweet is then stored in a set P_i of potential parent tweets from tweet i.

If the set P_i of potential tweets for tweet i is empty, then we have failed to find a tweet where tweet i originated from and assume that it is an original message. If P_i contains exactly one candidate j, then we determine tweet j to be the parent of tweet i. If P_i contains more than one tweet, then we recursively determine the origin tweet by transversing both retweet and hidden edges for all potential parents $j \in P_i$. If and only if all potential parents j have a common ancestor k, then we determine k to be the origin of tweet i. Since retweets point back to the original tweet, we convert combination networks to a star topology with the oldest common ancestor as the root when comparing them to retweet networks, making the determination of the exact parent unnecessary. If we cannot find a common ancestor, then we flag the message as unable to be processed and ignore it during analysis.

Classifier	Accuracy
Longest Substring	92.52
Longest Substring Normalized	91.52
Levenshtein Distance Normalized	88.95
Levenshtein Distance	85.37
Matching Word Count Normalized	75.42
Matching Word Count	75.13
Logistic Regression	92.07
Ada Boost M1	92.60
C4.5	93.62
Naive Bayes	91.01
Random Forest	92.89
Neural Network	93.22

Table 5.1: Accuracy of different classification metrics from leave one out cross validation. In parenthesis are the optimal cutoff for a decision rule based solely on that metric.

Category	Selection Criteria
Personal	Personal Twitter Accounts, Bloggers
News Organization	Newspapers, News channels
Health Organization	Doctors, Hospitals, Health Experts, Drug Companies, Government run health groups (such as the CDC)
Other Category	Charities, Businesses, Any accounts that can't be put in the other categories

Table 5.2: Selection criteria shown to the Amazon Turk workers when coding the Twitter accounts.

5.4.2 User Type Classification

User accounts were assigned one of four categories: Personal Account, News Organization's Account, Health Organization's Account, or Other (for a list of criteria for each category, see table 5.2). A training set was developed by randomly selecting 3,378 users. Each account was sent to Amazon Mechanical Turk, an online micro-task marketplace, to be hand rated by two people. If there was a disagreement, the account was ranked by a third Amazon Turk worker.

Users were classified by Weka. Common machine learning algorithms [127–129] such as naive Bayes classifier, random forest classifiers, C4.5 decision trees, and support vector machines with a polynomial kernel and Gaussian kernels were considered. Each classifier was trained on a feature vector that included:

- log(1 + number of followers)
- log(1 + number of favourite tweets)
- log(1 + number following)
- log(1 + number of tweets)
- log(1 + number of lists the user has)
- top level domain (".com", "other", or "no website given")
- the 100 most significant principle components from the user's description field.

As well as the user's classification. The principle components were generated by a principle component analysis of the 1000 most common words from the training set. The accuracy of each classifier was determined by selecting 10% of the data for testing while training the model on the remaining 90%. This was repeated ten times, in a process called 10-fold cross validation. [75] The support vector machine with a Gaussian kernel was found to have the highest accuracy (83%) and was chosen to classify the users. (Table 5.3)

5.4.3 Keyword Analysis

While the number of followers and the type of account a user has may be strong predictors about how often a message may be retweeted, they cannot be easily modified by the user that wants to propagate her Tweets, limiting their usefulness. Instead, one can consider the textual content of a message for its ability to effect the chance that a message will be retweeted.

Here, we model retweets based on unigrams of the original Tweet's text. $n > 1$-grams were considered, but were not included because a preliminary analysis found no significant increase in predictive power in the H7N9 dataset. Since many n-grams

Classifier	Accuracy (st. dev.)
Guess Mode	72.99 (0.14)
C4.5	75.36 (1.92)
Naive Bayes	29.79 (2.87)
SVM Polynomial Kernel	82.52 (1.35)
SVM Gaussian Kernel	73.70 (0.46)
Random Forest	79.46 (1.48)

Table 5.3: Accuracy and standard deviation of the models considered for classifying user type determined by ten repetitions of 10-fold cross validation. "Guess Mode" is a simple rule that always classifies a user as the most common user type ("personal") and provides a base line to compare the classifiers against. "SVM" = "Support Vector Machine"

will be too rare to be useful for training, we define a minimum number of times the n-gram must occur to be included in the feature vector. Additionally, we do not want to impose any biases by arbitrarily selecting the cutoff. Thus we consider a model parameter, MIN-N-GRAM $\in \{100, 1000\}$ which we will iterate through during model selection, corresponding to a roughly $\frac{1}{10}\%$ or 1% occurrence in each Tweet, respectively. Similarly, we cannot choose an arbitrary regression model, so we iterate through the following: support vector regression with a polynomial (x, x^2 and x^3) kernel, multi-variable linear regression, regression trees (with a max depth of 2,5, or 10) and gradient boosting. Stop words from the scikit-learn stop list [131] are removed, the text is tokenized on non-alphanumeric characters, and the keywords are not case sensitive.

These 6 combinations of models and MIN-N-GRAM are then compared for performance. To avoid biases from model selection, a 10% hold out set is randomly sampled from each of the three datasets. The remaining 90% of each dataset are then used to evaluate the models using 3-fold cross validation. Since the content of each of the three datasets may be different, we consider model building and selection for each dataset independent of the other ones. For computational reasons, the H7N9 and autism training sets are limited to a random subsample of 100,000 instances. Additionally, we consider a generalized model based on all three datasets. The models are then evaluated based on the Pearson's correlation coefficient and mean absolute error. The model that performs the best is then evaluated against the hold out set.

Axes	Level	Example Tweet
Valence	High	Autistic child shows off inspirational art skills at local gallery
	Low	WHO reports 12 new fatalities due to H7N9.
Arousal	High	Excited for today! Let's go out and run #forthecure!!
	Low	5 new cases of measles? I guess that's bad or whatever :/
Dominance	High	We must go out and fight big pharma! Do everything to stop vaccines now.
	Low	I'm not sure if vaccines are good or bad, it might be worth thinking about, but I don't know...
Aptitude	High	Scientific analysis of new data shows no statistical relation between vaccinations and autism.
	Low	U wot m8? Why u no get yo kid's a flu shot?

Table 5.4: Representative messages for high and low states of activation for each of the four axes.

5.4.4 Emotion Tagging

We consider Plutchik's model of emotion [61] which is based on four dimensions: valence, arousal, dominance, and aptitude. Valence corresponds to the attractiveness of a message, which is roughly equivalent to sentiment. Arousal corresponds to the energy behind the message. For example, we may find that messages with multiple exclamation marks show higher arousal than messages without exclamation marks. Dominance corresponds to how argumentative the message appears. Finally, aptitude corresponds to the inferred maturity of a message. Note that we are interested in the emotion displayed by the Tweet, which does not necessarily correspond to the emotion of the person that made the tweet. For an example of both extremes of each of these four dimensions see Table 4.

We can then develop classifiers for each of these four dimensions based off of each message's content. We tag each message as either low, neutral, or high for each of the four dimensions. Additionally, a message may be tagged as irrelevant. Initially, we considered a 5 level Likert scale for rating each dimension, however, a

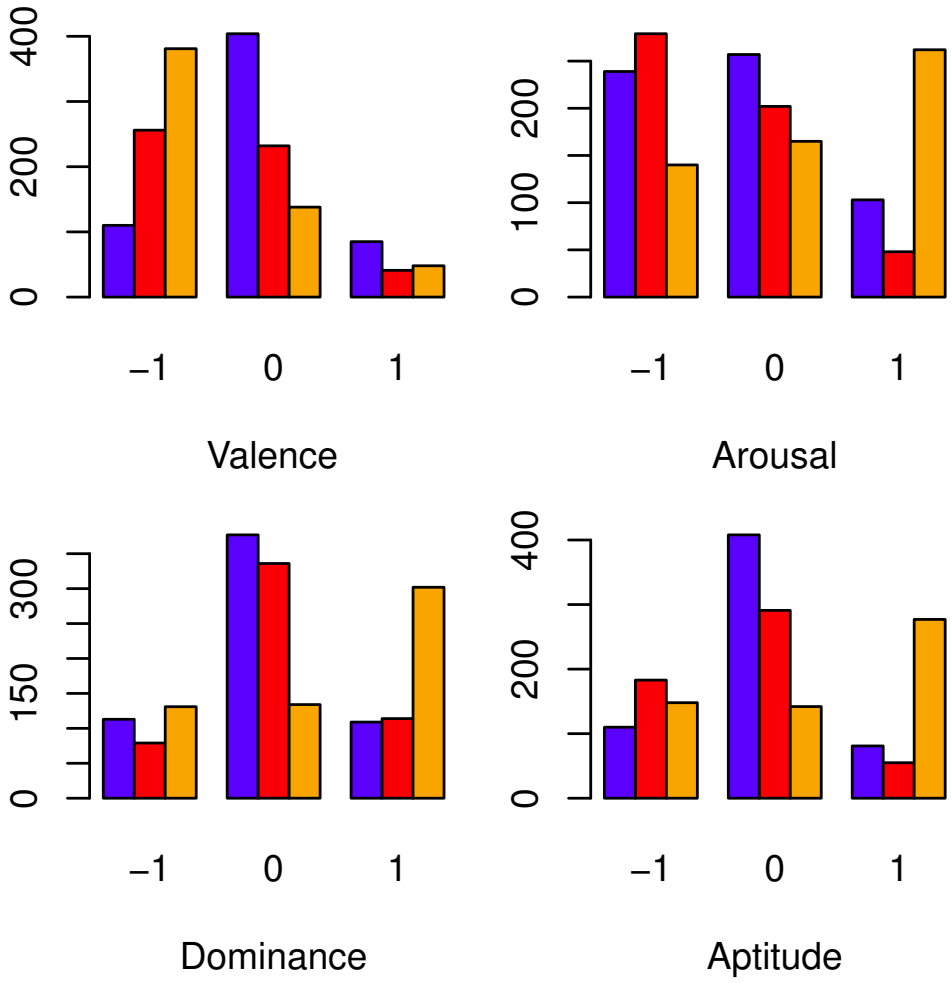


Figure 5.1: The distribution of each of the four parameters of emotion from the hand rated training sets for the Autism (blue), H7N9 (red) and MMR (orange) datasets.

proof of concept study was performed where one of the authors rated 100 MMR messages on this five point scale for each of the dimensions. When they were re-rated by the author, accuracy was low but much higher when the low and very low (along with high and very high) ratings were combined. Thus, it was concluded that a 3-point scale would be a better approach.

A random sample of 600 Tweets from each of the three topics was then hand rated. For rating counts, see figure 5.1. Rating through a majority vote of 3 Amazon Mechanical Turk workers was considered, however agreement with a hand-coded gold standard, along with agreement between the raters, was low, with rates similar

to previous work's analysis [130].

These ratings were then used as a basis for training machine learning classifiers. Each of the four dimensions were evaluated independently. The attribute of a message was evaluated in two steps. First, a binary classifier determines if the text either has a relevant, emotional signal or not. That is, messages that are either irrelevant, or do not show a positive or negative polarity on the given dimension. Messages not filtered are then passed through a second binary classifier to determine if they contain a positive or negative form of the emotional dimension in question.

To transform the text into a machine readable form, the text is converted to a binary keyword vector using scikit-learn's [131] CountVectorizer class. English stop words were removed. The tokens were converted to n-grams of size 1 to max_n where we varied max_n from 1 to 5. Similarly, we varied the number of times an n-gram needed to occur to be included in the final feature vector by min_df, which we varied between 1 and 20. Additionally, we considered including or not including extra features based on the length of the tweet and the counts of numbers, exclamation marks ('!'), question marks ('?'), periods ('.'), hashtags ('#') and at signs ('@') in the text. This results in a total of 200 possible text vectorizers considered. This keyword vector was then applied to random forest (with 5 or 10 decision trees), logistic regression, nearest neighbor, naive bayes, and support vector (with a radial basis function kernel) classifiers to classify the text as described above.

These 1800 possible combinations were trained and evaluated against 500 hand rated messages using 5-fold cross validation. The combination that had the best accuracy, for a given step and emotional dimension, was then chosen as the "best" classifier. The models were then trained on the whole 500 element training/testing set and evaluated against a 100-Tweet hold out set. Additionally, we built a combined dataset that is based on the three datasets.

We find that the models trained on the combined dataset outperform classifiers that were specialized for a specific dataset (see tables 5.5 and 5.6). This is probably due to the three times increase in training data. Since the distributions tend to be non-uniform (see figure 5.1), we evaluate the classifiers on the mean of the true positive and true negative rate, or

$$mean\_accuracy = (\frac{TP}{TP + FN} + \frac{TN}{TN + FP})/2 \tag{5.2}$$

Emotion	Dataset	Best Classifier	Max n-Gram	Min Occurrences	Include Extras?	Mean Accuracy
Aptitude	Autism	Naive Bayes	3	8	False	0.5385
	H7N9	Naive Bayes	1	5	True	0.5280
	MMR	Naive Bayes	1	1	False	0.5690
	Combined	Naive Bayes	1	8	False	0.6831
Arousal	Autism	KNN	1	2	False	0.4727
	H7N9	Naive Bayes	1	1	False	0.5476
	MMR	KNN	1	2	True	0.5073
	Combined	KNN	1	15	True	0.5119
Dominance	Autism	Naive Bayes	3	10	True	0.7696
	H7N9	Naive Bayes	3	7	False	0.4996
	MMR	Naive Bayes	3	1	False	0.5471
	Combined	SVM	3	10	False	0.6993
Valence	Autism	Naive Bayes	3	8	False	0.6023
	H7N9	Naive Bayes	1	3	True	0.5888
	MMR	KNN	3	1	False	0.4940
	Combined	Naive Bayes	3	14	True	0.6141

Table 5.5: The performance for each of the neutral/non-neutral classifiers selected from the test/train set on the validation set.

where TP and TN are the number of true positives and negatives and FP and FN are the number of false positives and negatives. Note that a completely biased estimator that either always estimates positive or negative will have a mean accuracy of 0.5. Since the classifiers to find either the existence of arousal or its polarity are essentially 0.5, we are unable to develop an accurate classifier and ignore that dimension in further analysis. Additionally, we discard dominance because of the difficulty in determining polarity. Thus, we work with a two-dimensional model of emotions based on valence and aptitude.

5.5 Effects of metrics on Retweets

5.5.1 Network Effects

Of the total 947,967 tweets collected, 339,292 tweets were rated as retweets by the twitter API. An additional 39,563 hidden retweets were detected. We observe that

Emotion	Dataset	Best Classifier	Max n-Gram	Min Occur-rences	Include Extras?	Mean Accu-racy
Aptitude	Autism	KNN	1	1	False	0.5979
	H7N9	SVM	1	3	False	0.5000
	MMR	SVM	3	2	False	0.7903
	Combined	LR	1	1	True	0.8292
Arousal	Autism	NB	1	5	False	0.6197
	H7N9	NB	1	14	False	0.5068
	MMR	KNN	1	1	False	0.6455
	Combined	LR	3	1	True	0.7595
Dominance	Autism	NB	1	2	True	0.5136
	H7N9	KNN	1	2	False	0.4298
	MMR	NB	3	1	False	0.4737
	Combined	KNN	3	1	False	0.5625
Valence	Autism	LR	1	1	False	0.5611
	H7N9	LR	1	3	True	0.7384
	MMR	NB	1	16	True	0.6094
	Combined	LR	3	1	False	0.6561

Table 5.6: The performance for each of the emotion polarity classifiers selected from the test/train set on the validation set. (LR = Logistic Regression, NB = Naive Bayes)

the log-transformed lag between Tweet and Retweet is unimodal for traditional retweets, but bi-modal when we include hidden retweets. While the later peak is similar to the one detected in API retweets, the earlier peak is located at less than two seconds. It is unlikely that a human would be able to read and copy a message in this short of a time period. Instead, it would seem that this subset of hidden retweets is an attempt to make an automated account look human by copying topical conversations. Indeed, upon inspection of these accounts, we find an unusually large number of total tweets (mean = 34352.7092 Tweets per user) compared to accounts that do not show this behavior (mean = 12319.2674 Tweets per user). Future work will determine the purpose of these fake accounts. Here, we determine that these spam accounts do not actually consume or produce information – in the traditional sense – and we thus discard any retweets that happen within 10 seconds of the initial post (see figure 5.2). All analysis in the paper outside of this section is on datasets with spam accounts removed.

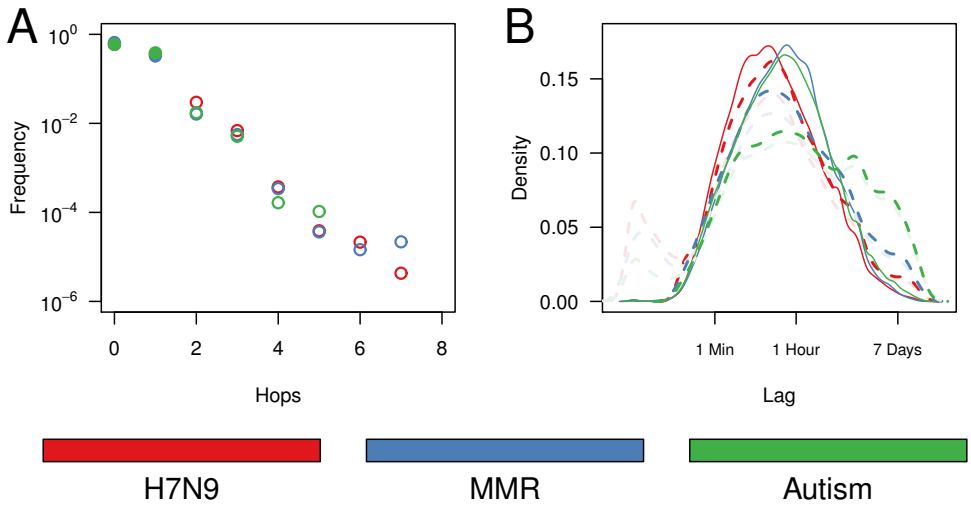


Figure 5.2: A comparison between regular retweets (solid) and hidden retweets (dashed). Hidden retweets show an exponential distribution in the distance between the tweet and it's message's origin, and regular retweets show a star topology, never being more than one hop from the origin (A). The time between when a message is posted and when a message is reposted is similar between the two types of reposts (B). However, hidden retweets may be multimodal with an additional high point in the seconds range, possibly indicating false positives.

Dataset	Retweets	Hidden	Retweets + Hidden	% Hidden
H7N9	74444	18190	92634	19.646
MMR	41961	5627	47588	11.824
Autism	215927	15747	231674	6.797
Total	332332	39564	371896	10.639

Table 5.7: The total number of retweets, hidden reproductions and total reproductions of a message.

As mentioned above, we model the number of times a message is propagated based on the number of users that follow the initial poster's account. The number of followers also approximates the number of people that will see a Tweet. We find that this model explains approximately 10% of the variation in retweets ($0.3142 \leq r \leq 0.4169$, $p < .0001$ in each case, see table 5.13).

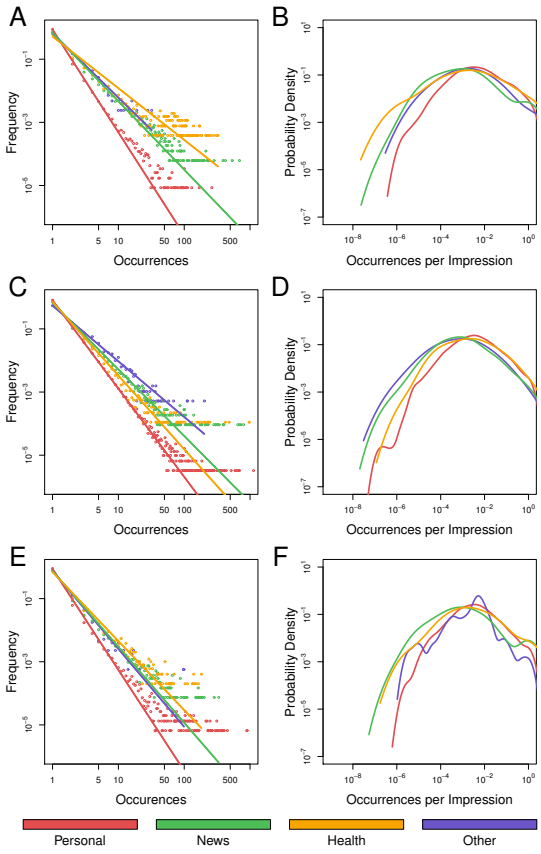


Figure 5.3: The frequency of reposts of a message by each of the four types of users (A) and the estimated probability density function of a message's posts normalize by the number of individuals that the original poster is followed by (B). Note that the non-normalized post counts have a power-law-like distribution with the dashed line representing a log-log function fit to the data. All distributions are significantly different. Occurrences are defined by the number of retweets + 1. Where we add one to account for the initial posting.

5.5.2 User Type Effects

While News and Health accounts are a minority (3.33%) of users, their messages result in 31.47% of the retweets (see figure 5.3). One may expect this as News and hHealth accounts tend to have more followers (see table 5.9). These networks effects can be controlled for by considering the frequency of retweets per follower. However, even when this control is applied, we find that Health and News accounts still tend to have more retweets. Indeed, we find that messages from health accounts tend to generate more retweets even than News accounts, despite less followers. This may

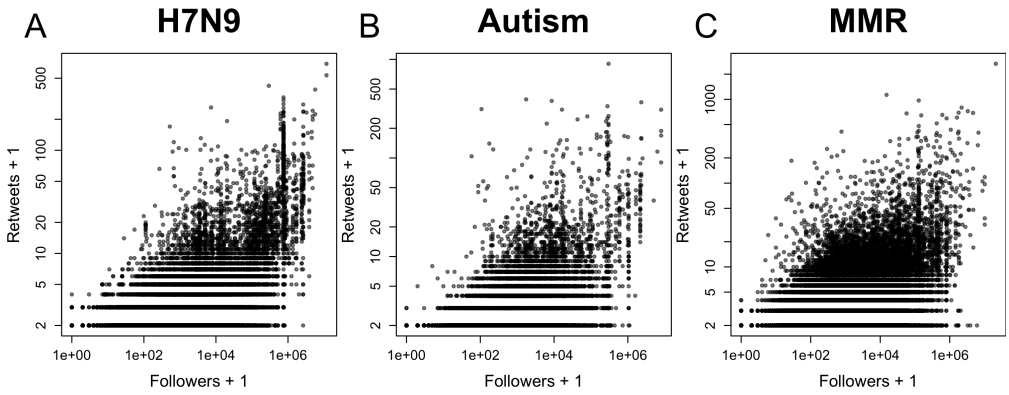


Figure 5.4: The number of times a tweet is posted based on the number of followers the original poster has.

Dataset	Personal	Other	News	Health
H7N9	87602 (94.99%)	141 (0.15%)	3453 (3.74%)	1028 (1.11%)
MMR	57705 (95.36%)	195 (0.32%)	1651 (2.73%)	964 (1.59%)
Autism	269575 (96.51%)	940 (0.34%)	5264 (1.88%)	3535 (1.27%)
Total	397838 (96.38%)	1202 (0.29%)	8851 (2.14%)	4896 (1.19%)

Table 5.8: The number of each type of users in each of the three datasets. Note that the total counts are lower due to duplicates between the three datasets.

be due to several factors such as end users assigning Health organizations a higher authority than News organizations or Health organizations having faster access to important news.

As there are a fixed number of cases for the feature ACCOUNT-TYPE, we can model the effect that the type of account has on the number of retweets by using a multivariable regression based on ACCOUNT-TYPE. Since the predictive variable, ACCOUNT-TYPE is a categorical variable, this regression model can be simplified

User Type	Median Followers	Mean Followers	Max Followers
Personal	246.0	1467.89	6424948
Other	1094.5	15271.77	2503732
News	2573.0	58323.01	21125687
Health	302.0	3762.29	1064952

Table 5.9: The number of each type of users in each of the three datasets. Note that the total counts are lower due to duplicates between the three datasets.

Model	Min-N	MMR	H7N9	Autism	Combined
SVMR	100	0.0401	0.0015	0.1224	0.0987
	1000	0.0232	-0.0157	0.1066	0.0642
Regression Tree (2)	100	0.0619	0.0649	0.1021	0.0907
	1000	0.0914	0.0590	0.1036	0.0969
Regression Tree (5)	100	0.0816	0.1123	0.1160	0.1099
	1000	0.0998	0.0958	0.1222	0.1025
Regression Tree (10)	100	0.0826	0.0712	0.1076	0.0896
	1000	0.0891	0.0850	0.0975	0.0990
Gradient Boosting	100	**0.1800**	**0.1674**	**0.1783**	**0.1513**
	1000	0.1547	0.1399	0.1652	0.1416

Table 5.10: Performance of various regression models to based on correlation coefficients of predicted and actual log(retweet) rates given the Tweet's textual content. Min N = Minimum number of times a word must appear to be included. SVMR = Support Vector Regression,.

to a look up table defined as:

$$Retweets = \begin{cases} \text{mean(retweets of tweets from } News) & \text{If posted by } News \\ \text{mean(retweets of tweets from } Personal) & \text{If posted by } Personal \\ \text{mean(retweets of tweets from } Health) & \text{If posted by } Health \\ \text{mean(retweets of tweets from } Other) & \text{If posted by } Other \end{cases} \tag{5.3}$$

We find that this model correlates to the true retweet count with a coorelation coefficient of between 10.73 and 25.22 for each of the datasets, corresponding to between 1.15% and 6.36% of the retweet variation being explained by the type of user that sent the message (See table 5.13).

5.5.3 Keyword Effects

Textual content appears to have a mild effect on over all retweeting rates (see table 5.10). While this may seem surprising at first, there is a likely explanation: The words within each message are likely similar, as they are all about the same topic. For example, there are many Tweets in the autism dataset of the form "Celebrate Autism Awareness Month." Since there is a wide variation in who posts these similar messages, there is also a large variation in the number of reposts each Tweet

gets that is *not* explained just by textual content. Later, in section 5.6, we will combine the textual content with the other features to build a better preforming model.

5.5.4 Emotional Effects

Alone, emotional content does not seem to be much of a predictor of retweet rates (see table 5.13), although the predictions are statistically significant. Here, instead, we discuss various relations between emotional content and our Twitter data. First, we find a small ($r = -0.1185$) but significant ($p \leq 10^{-16}$) relationship between valence and aptitude in our data. This may be in line with previous psychology work which assumes that these two variables are independent. [61] This small relationship may be due to either biases in our classifiers or due to our large dataset providing more statistical power (a sample size of $n \geq 274$ is required to resolve an effect of this size). Note that with an approximate r^2 of 1.4%, it is unlikely that one could infer much about one of the variables from the other, justifying the, generally employed, simplifying assumption of independence.

We find differences in the average valence and aptitude from users of different types in each of the datasets (see figure 5.5). This shows predictable trends–such as health practitioner's appearing more knowledgeable than others–along with less expected trends. For example, we find that Health accounts tend to post more messages with negative valence while News organizations tend to stay more neutral. Additionally, note that Other accounts seem to have similar aptitude levels to that of Health accounts, but much different valence levels. We find no statistical difference (p=0.185) between aptitude in H7N9 tweets from Health and News accounts, possibly due to both groups focusing on the "breaking" news. Additionally, we find no statistical difference between Health, News and Personal accounts for valence about Autism ($p \geq 0.1586$ in each case) while Other accounts have a statistically significant difference in valence from Personal ($p < 0.0001$) and Health ($p = 0.0006$) accounts but not News accounts ($p = 0.77035$). All other comparisons are significant in the $p < 0.001$ level. All comparisons were performed using pairwise-t-tests with the Holm-Bonferroni correction for multiple tests. Additionally, we find a negative correlation between follower count and valence (Spearman's rank test $\rho = -0.05594$, $p < 2^{-16}$) and a positive relation

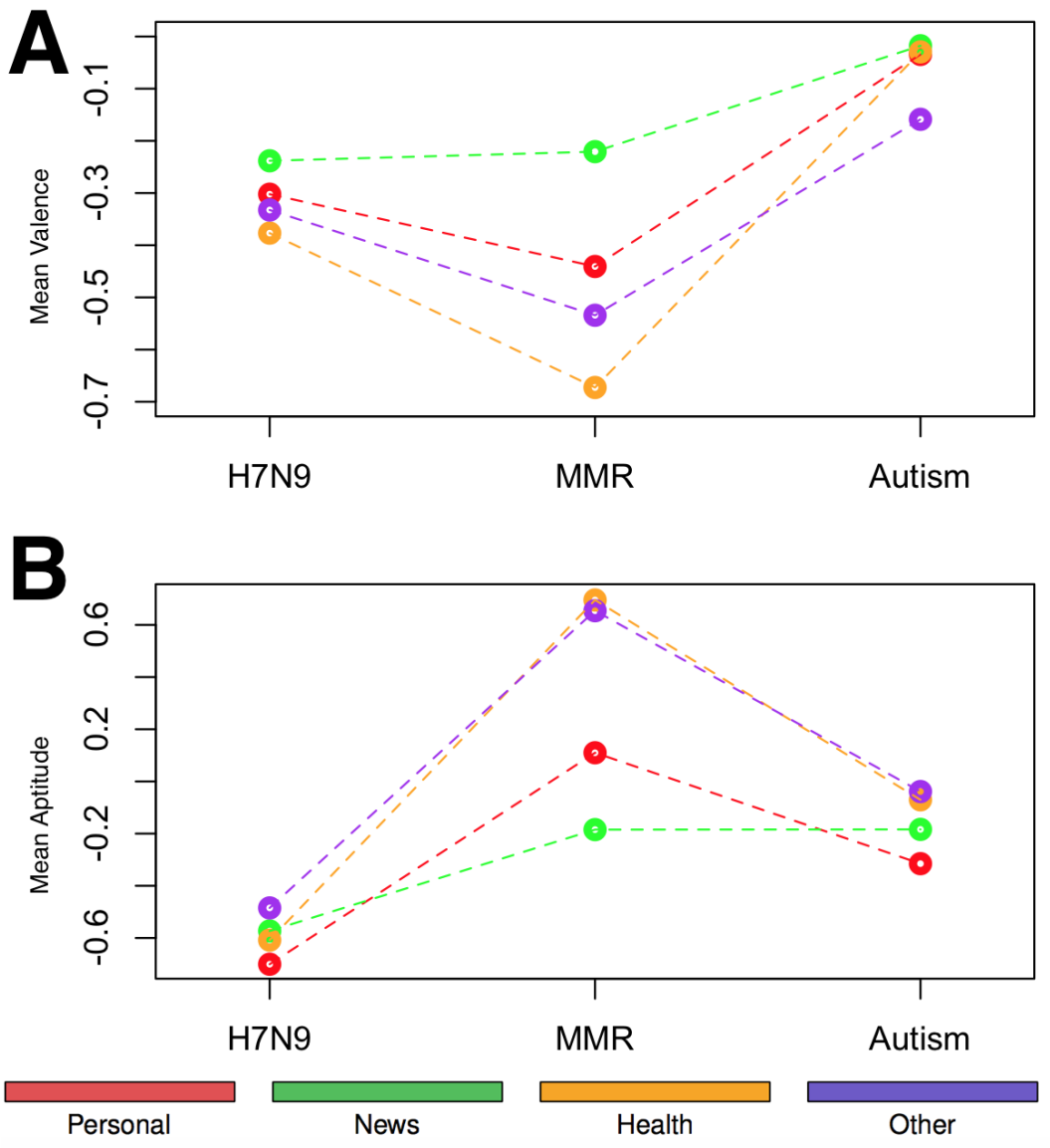


Figure 5.5: The mean Valence (A) and Aptitude (B) of tweets from each of the four user types in each of our three datasets.

between follower count and the aptitude of the messages (Spearman's rank test $\rho = 0.01391$, $p < 2^{-7}$).

5.6 Prediction of Tweet Effects

We now combine the features described in section 5.5 to build a final predictive model. To deal with issues related to overfitting, in this chapter we worked with three subsets of each of the three datasets. We divided the datasets into an 85%, 10%, and 5% train, test, and validate sub sets. All results up to this point were based on the 85% training dataset. Since we perform model selection regarding keywords, user types, and emotions, we cannot accurately discern the models' performances. That is, since we are choosing models that have the best correlation in the training set, any final estimates of correlation would be biased.

Data Set	Followers	User Type	Text	Emotion	Correlation
				X	0.0692
			X		0.1845
			X	X	0.1902
		X			0.1737
		X		X	0.1866
		X	X		0.2559
		X	X	X	0.2564
MMR	X				0.4996
	X			X	0.5018
	X		X		0.5095
	X		X	X	0.5114
	X	X			0.5164
	X	X		X	0.5172
	X	X	X		**0.5222**
	X	X	X	X	0.5201
				X	0.0169
			X		0.1712
			X	X	0.1723
		X			0.2689
		X		X	0.2817
		X	X		0.4022
		X	X	X	0.4019
H7N9	X				0.6819
	X			X	0.6837
	X		X		0.6836
	X		X	X	0.6820
	X	X			0.6827
	X	X		X	0.6846
	X	X	X		0.6874
	X	X	X	X	**0.6878**

				X	0.0775
			X		0.1932
				X	0.0769
			X		0.1744
			X	X	0.1756
		X			0.0984
		X		X	0.1243
		X	X		0.2103
Autism		X	X	X	0.2091
	X				0.4067
	X			X	0.4144
	X		X		0.4508
	X		X	X	0.4523
	X	X			0.4097
	X	X		X	0.4175
	X	X	X		0.4533
	X	X	X	X	**0.4541**
				X	0.0737
			X		0.1476
			X	X	0.1474
		X			0.1412
		X		X	0.1663
		X	X		0.2415
		X	X	X	0.2406
Combined	X				0.4844
	X			X	0.4914
	X		X		0.5111
	X		X	X	0.5109
	X	X			0.4882
	X	X		X	0.4940
	X	X	X		0.5128
	X	X	X	X	**0.5130**

Table 5.11: Combined model performance on the test sets.

Included Variable	Δr
Followers	0.2866
User Type	0.01985
Text	0.01525
Emotion	-0.000175

Table 5.12: Decrease in mean correlation when a feature is removed from the full model.

Model	H7N9	Autism	MMR	Combined
Followers	0.4169	0.3142	0.3411	0.3466
User Type	0.2522	0.1410	0.1073	0.1363
Keyword	0.1674	0.1738	0.1800	0.1513
Emotion	0.0221	0.0753	0.0810	0.0761

Table 5.13: The correlation coefficient for each of the models described. Models are compared with either each dataset individually or the combined dataset

Since we've found that gradient boosting performs the best with the keyword studies, we base our final predictive model on gradient boosting. For each of the datasets, we build a regression model to predict log-transformed retweets based on the initial poster's log-transformed follower count, user type, textual content and emotional rating. Additionally, we consider feature selection using a factorial design to include or exclude each of these four features (see table 5.11). We can also use this to estimate the predictive effects of each of the four features by comparing the decrease in model performance when a feature is removed to the full model. We find that follower count has the strongest effect with a mean decrease of 0.2866 between our datasets (see table 5.12).

Finally, we can select the model that predicts retweeting rates the best for each of the four models. As before, we must now score these regression models against an independent dataset. Thus we evaluate the final models against the remaining 5% validation set and find a mean correlation of 0.5521 (see table 5.15), compared to a baseline of 0.3547.

5.7 Conclusions

In this chapter we considered the problem of retweet prediction applied to health related messages. In addition to normal retweeting behavior, we also described an

Model	H7N9	Autism	MMR	Combined
Followers	0.1875	0.2128	0.1656	0.1993
User Type	0.1718	0.1607	0.2222	0.2027
Keyword	0.5134	0.5768	0.4862	0.5503
Emotion	0.2497	0.3254	0.2240	0.2922

Table 5.14: The mean absolute error for each of the models described. Models are compared with either each dataset individually or the combined dataset

Data Set	r
MMR	0.5565
H7N9	0.6665
Autism	0.4667
Combined	0.5185

Table 5.15: Performance of the final regression models on the validation datasets.

additional form of message propagation based on text similarity. We then used standard features (follower count and keyword information) and novel features (account type and emotional information) to build a model that is able to predict retweet rates with up to a correlation of 0.6665 to the true retweet rates. This predictive model, along with information discussed about user type and emotional effects, could be employed by a public health practitioner to increase the reach of public health messages at minimal additional costs.

Chapter 6
Future Work

6.1 Future Directions

6.1.1 Extended Diagnostic Methods

While our diagnostic system, described in chapter 3, provides a proof of concept for validated diagnosis of a disease using a combination of Twitter and medical records, there are many things that could be done to improve it. First, our system would be improved with a larger dataset. While we have shown that we can generate a good fit on a small user set, there is still the question of generalizable. Extending the dataset beyond one season of college student disease diagnoses at one university could help to address the issue of generalizable.

This larger dataset could allow us to measure other factors besides simply the disease, such as vaccination rates or hygiene practices, which may effect the chance of someone becoming ill. [132] Previous work [133–135] has looked into this using surveys and other traditional data gathering methods. Due to their limited sample sizes, they tend to not be able to generate interesting insights because they can only detect "strong" correlations between risk factors and disease. We attempted this approach on our small dataset, but were unable to find statistical significant results. Additionally, some of the results we found seemed counterintuitive. For example, users who discuss smoking tended to be less likely to be diagnosed with influenza than ones that discussed exercise or yoga. However, this may be explained by a bias introduced in our sampling methods: users who are more health conscious may be more likely to visit a medical practitioner to get an official diagnosis. If we were to apply these risk factor detectors and influenza diagnosis systems to the

general Twitter dataset–for example, the one used in chapter 4– we may be able to get around these biases that would be inherent in *any* medical-based survey study. Additionally, we could apply these risk factors to improve our diagnosis' accuracy. For example, a user that previously said that she was going to get an influenza vaccine is probably less likely to be later be diagnosed with influenza.

We could automate this by employing Deep Learning [136] neural networks to the dataset. An artificial neural network could be designed with L.T.M. (Long-Term-Memory) neurons [136–139] to keep a persistent model of each user over a long period of time. The neural network would then employ a user's current tweets along with information stored in the L.T.M. neurons to provide a diagnosis *and* update the same L.T.M. neurons. One drawback of this approach, however, is that other text processing systems [140, 141] tend to require hundreds of thousands or millions of data points before such a system can accurately learn these subtle, temporal relations.

Finally, we could extend this system to use other data sources than Twitter or to diagnose other diseases.

6.1.2 Experimental Validation of Message Propagation

While modeling retweet behavior is a well studied problem, there does not seem to be much academic research on an experimental validation of these behavioral models. One potential approach to such an experiment would be to coordinate with various Tweet creators–for example, CNN or the CDC in our health study–and see if they would be able to modify their messages in ways that the models predict would increase (or decrease) the expected number of retweets. These messages would then be "released into the wild" to see how many Twitter users decide to retweet the message. Clearly, many of these experiments are being done in the industrial side of research, but differing end-goals result in a lack of publication of these results.

Appendices

Appendix A
Keyword Recommendations

While our system should be trusted more than one based simply off of aggregated tweets, it is more computationally intensive than simply pulling data from a keyword stream. These systems require the user to select a specific set of keywords before data collection can begin. Keywords representing symptoms such as “flu”, “cough”, “sore throat”, and “headache” are often chosen. We suggest the thirty keywords with the highest positive predictive value (see table A.1) be chosen as the parameters for a keyword stream. In addition to keywords related to symptoms (e.g. “flu” or “sick”) we also find keywords related to treatments (e.g. “health,” “prayer” or “recovery”) and keywords related to negative mood (e.g. vulgarities) to be more commonly tweeted when a user is ill.

Keyword	Ratio
flu	34.424
health	11.360
sick	5.019
track	10.952
stud	3.508
asshol	9.090
ton	9.090
particip	20.667
salt	20.667
recov	40.118
fuck	2.963
sham	13.64
row	10.180
win	2.947
rt	3.077
walk	3.077
childr	6.820
incred	6.820
meal	6.820
longer	6.820
succes	26.765
accis	26.765
holida	26.765
luv	26.765
oblig	26.765
path	26.764
pract	26.764
prayer	26.765
reserv	26.765
riot	26.765

Table A.1: The thirty keyword stems with the highest positive predictive power ranked by significance. The Twitter API limits searches to at most thirty keywords. Ratio is calculated as the rate of occurrence when a user is sick over the rate when a user is not sick.

Appendix B
Regional R0 Levels

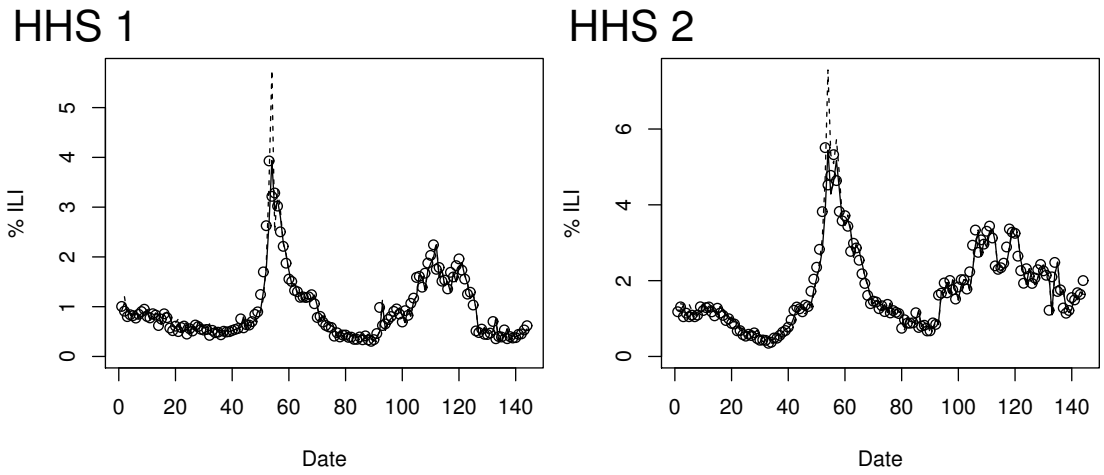
HHS Region	Flu Season	CDC R_0	Twitter R_0	CDC R_E	Twitter R_E
1	2011-2012	1.105	2.015	1.054	1.085
	2012-2013	2.177	2.034	1.325	1.295
	2013-2014	1.933	2.071	1.257	1.264
	Combined	1.74	2.009	1.209	1.222
2	2011-2012	2.2	1.665	0.9395	0.987
	2012-2013	2.034	1.987	1.382	1.381
	2013-2014	2.155	2.2	1.354	1.347
	Combined	1.849	1.862	1.259	1.259
3	2011-2012	2.2	2.141	1.128	2.141
	2012-2013	2.044	2.184	1.398	1.376
	2013-2014	1.986	2.2	1.227	1.272
	Combined	2.2	2.2	1.248	1.248
4	2011-2012	1.772	2.004	1.056	1.149
	2012-2013	2.2	2.2	1.399	1.39
	2013-2014	1.767	1.982	1.296	1.293
	Combined	2.18	2.192	1.292	1.292
5	2011-2012	2.013	1.358	1.151	1.115
	2012-2013	2.2	2.2	1.376	1.335
	2013-2014	2.2	2.176	1.271	1.265
	Combined	2.171	1.842	1.269	1.23

HHS Region	Flu Season	CDC R_0	Twitter R_0	CDC R_E	Twitter R_E
6	2011-2012	2.2	2.181	1.173	1.129
	2012-2013	2.2	2.2	1.492	1.488
	2013-2014	2.2	2.136	1.429	1.371
	Combined	2.2	2.151	1.365	1.327
7	2011-2012	1.847	1.797	1.296	1.248
	2012-2013	1.356	1.324	1.356	1.324
	2013-2014	1.782	1.277	1.248	1.199
	Combined	1.277	1.361	1.254	1.244
8	2011-2012	2.123	2.18	1.166	1.163
	2012-2013	2.137	2.006	1.368	1.337
	2013-2014	2.2	1.55	1.304	1.206
	Combined	2.183	1.369	1.292	1.203
9	2011-2012	2.2	2.199	1.125	1.231
	2012-2013	2.073	2.074	1.317	1.317
	2013-2014	2.196	2.2	1.297	1.279
	Combined	2.128	2.2	1.26	1.277
10	2011-2012	2.102	1.296	1.231	1.146
	2012-2013	1.275	1.251	1.234	1.226
	2013-2014	1.405	1.383	1.258	1.242
	Combined	2.153	2.199	1.292	1.275
Combined	2011-2012	2.2	2.151	1.065	1.086
	2012-2013	2.2	2.07	1.38	1.351
	2013-2014	2.2	2.2	1.27	1.25
	Combined	2.2	2.2	1.266	1.254

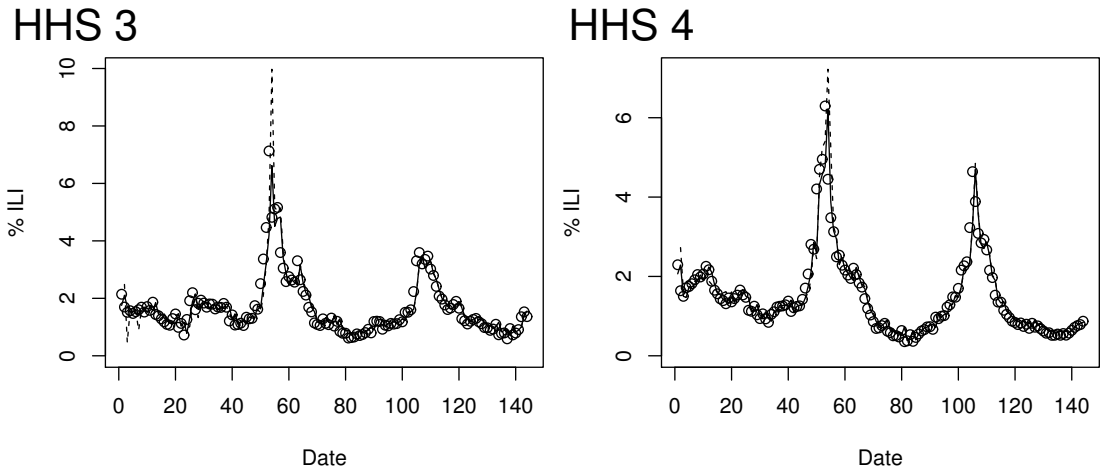
Table B.1: Estimated R_0 based on CDC and Twitter data with 5% and 95% percentiles in parentheses for each of the ten HHS regions. Note that the "Combined" region is *not* equivalent to the national estimates as it is calculated based on each of the ten region's incidents instead of a single, aggregated incident curve.

Appendix C
All Regional Fits

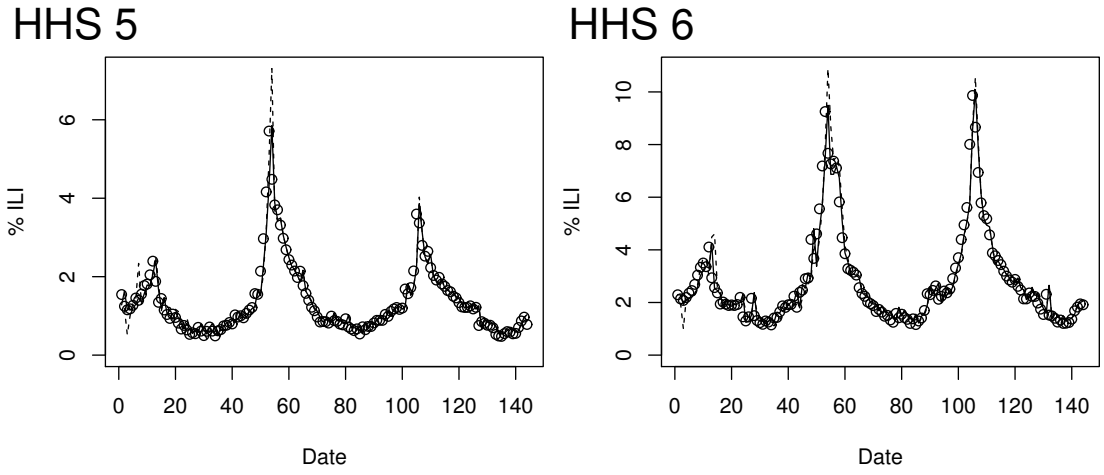
For space reasons, we did not include model fits for all of the potential regions in figure 4.1. Here, we provide full figures for all of the HHS regions (see fig C.1) and both of the counties that provided disease data to compare against (see fig C.2).


HHS 1
% ILI
0 1 2 3 4 5
0 20 40 60 80 100 120 140
Date

HHS 2
% ILI
0 2 4 6
0 20 40 60 80 100 120 140
Date


HHS 3
% ILI
0 2 4 6 8 10
0 20 40 60 80 100 120 140
Date

HHS 4
% ILI
0 2 4 6
0 20 40 60 80 100 120 140
Date


HHS 5
% ILI
0 2 4 6
0 20 40 60 80 100 120 140
Date

HHS 6
% ILI
0 2 4 6 8 10
0 20 40 60 80 100 120 140
Date

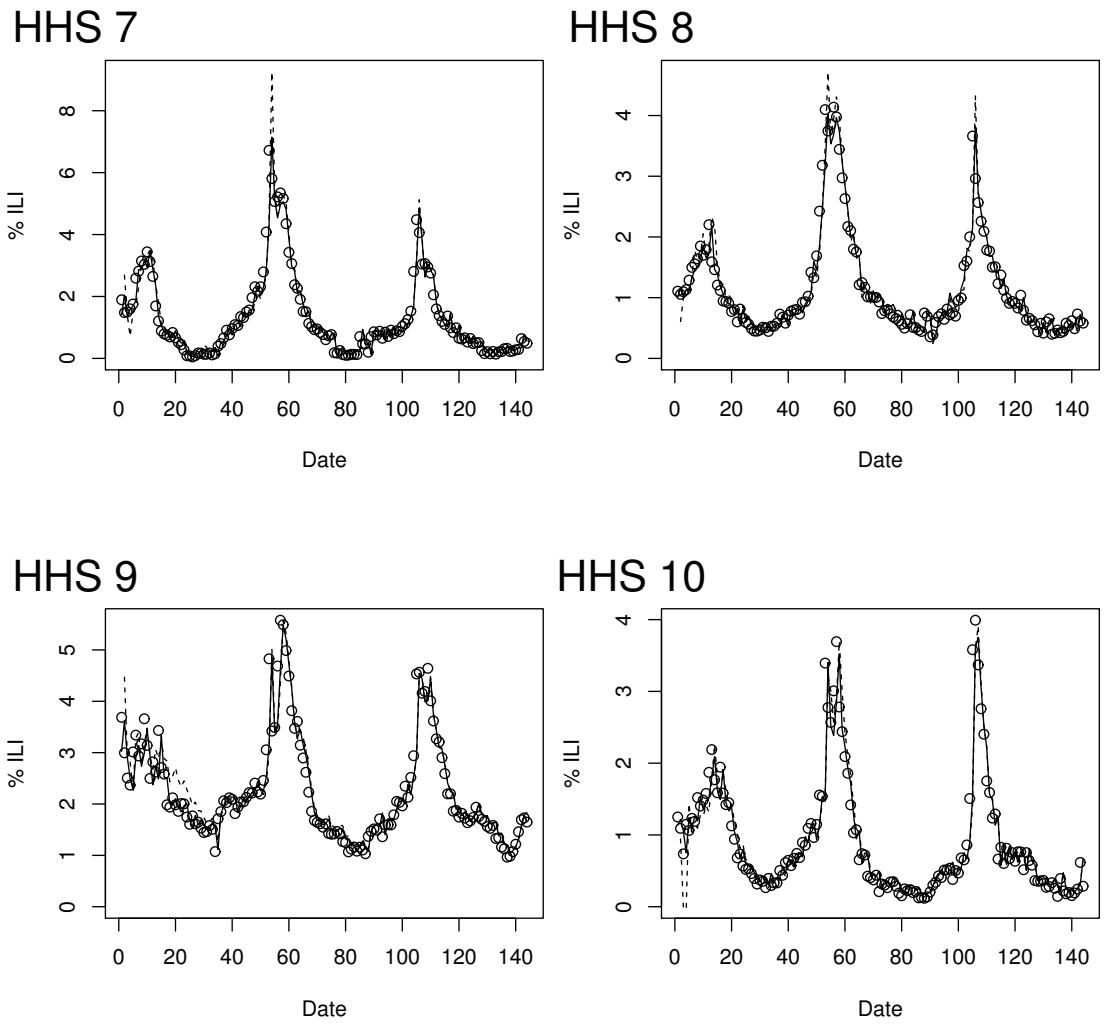


Figure C.1: Comparison of Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) to the CDC's reported Influenza rates (circles) for each of the 10 HHS regions.

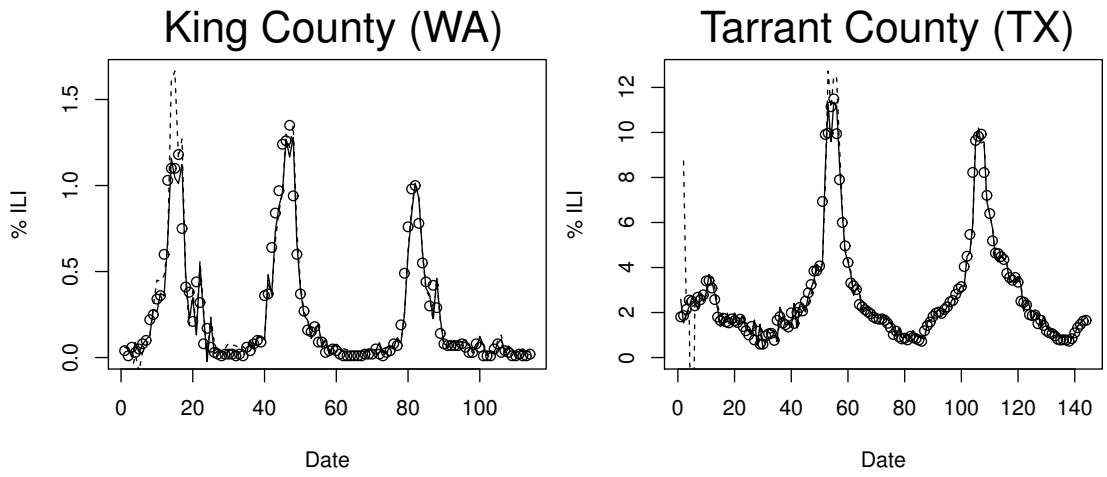


Figure C.2: Comparison of Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) to the CDC's reported Influenza rates (circles) for King County and Tarrant County.

Appendix D
A Comparison of Three Types of Message Propagation Models

Here we repeat the analysis in chapter 5 on the H7N9 dataset using three different types of propagation models: Retweets (Retweets), Retweets+Hidden flow (Combined), Retweets+Hidden Flow - Spam (No Spam). There are 158364, 139959, and 140174 independent, original messages detected in the models, respectively.

Due to the large data sizes, we decide to *not* present the statistical significance between the different models as it may be deceptive to the reader. For example, we use the No Spam model in the main paper, which can be predicted using follower count with a correlation of 0.4169. Given the size of the data, the Retweet dataset is different in a statistically significant sense if the correlation is outsize of the range $.4109 \leq r \leq .4229$ (Fisher z-transformation, two-tailed z-test), approximately a 1.4% change, which is unlikely to be operationally different.

Here, we repeat the steps described in chapter 5 by replicating the modeling of followers, user type and emotions on retweet rates (see table D.2) and perform keyword model selection (see table D.1). Instead of doing analysis on an 85%, 10%, 5% split as described in section 5.11, we simply work with the full datasets. Thus we do not include the aggregated models, as it would be deceptive to present the fits without a hold out dataset. Note that in 3 out of 4 models, the Combined dataset (the one without spam Tweet copies removed) is the most poorly fit. This loss of preventiveness may be due to a random selection of what Tweets to repost by a simple spam bot compared to the other two datasets, where a human provides

Model	Min-N	Retweets	Combined	No Spam
SVMR	100	0.0132	0.0059	-0.0029
	1000	-0.0007	0.0013	-0.0217
Regression Tree (2)	100	0.0688	0.0497	0.0747
	1000	0.0653	0.0475	0.0637
Regression Tree (5)	100	0.0829	0.0719	0.0962
	1000	0.1033	0.0742	0.0927
Regression Tree (10)	100	0.0707	0.0742	0.0642
	1000	0.0891	0.0960	0.1066
Gradient Boosting	100	0.1552	0.1473	0.1719
	1000	0.1605	0.1408	0.1600

Table D.1: Correlation of the output of various regression models used to predict log(retweet) rates given the Tweet's textual content on the three types of tweet propagation: Base API reposts (Retweets), Base reposts plus similar messages (Combined) and Combined with spam removed (No Spam).

Model	Retweets	Combined	No Spam
Followers	0.3963	0.4151	0.4255
User Type	0.2466	0.2420	0.2503
Keyword	0.1605	0.1473	0.1719
Emotion	0.02339	0.02119	0.02279

Table D.2: The correlation coefficient of models to predict the propagation count from messages in the H7N9 dataset.

some selective pressure in what he or she chooses to retweet.

Bibliography

[1] Bodnar, T. and M. Salathé (2013) "Validating models for disease detection using twitter," in *Proceedings of the 22nd international conference on World Wide Web companion*, International World Wide Web Conferences Steering Committee, pp. 699–702.

[2] Carneiro, H. A. and E. Mylonakis (2009) "Google trends: a web-based tool for real-time surveillance of disease outbreaks," *Clinical infectious diseases*, **49**(10), pp. 1557–1564.

[3] Culotta, A. (2010) "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proceedings of the first workshop on social media analytics*, ACM, pp. 115–122.

[4] ——— (2013) "Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages," *Language resources and evaluation*, **47**(1), pp. 217–238.

[5] Signorini, A., A. M. Segre, and P. M. Polgreen (2011) "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PloS one*, **6**(5), p. e19467.

[6] Butler, D. (2013) "When Google got flu wrong." *Nature*, **494**(7436), p. 155.

[7] Copeland, P., R. Romano, T. Zhang, G. Hecht, D. Zigmond, and C. Stefansen (2013) "Google Disease Trends: An update," *Nature*, **457**, pp. 1012–1014.

[8] Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014) "Twitter: Big data opportunities Response," *Science*, **345**(6193), pp. 148–149.

[9] Lazer, D. M., R. Kennedy, G. King, and A. Vespignani (2014) "Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season," *Available at SSRN*.

[10] Lamb, A., M. J. Paul, and M. Dredze (2013) "Separating Fact from Fear: Tracking Flu Infections on Twitter." in *HLT-NAACL*, pp. 789–795.

[11] World Health Organization et al. (2006) "Communicable disease surveillance and response systems: guide to monitoring and evaluating," . URL http://www.who.int/csr/resources/publications/surveillance/WHO_CDS_EPR_LYO_2006_2.pdf

[12] Chunara, R., C. C. Freifeld, and J. S. Brownstein (2012) "New technologies for reporting real-time emergent infections," *Parasitology*, **139**(14), pp. 1843–1851.

[13] Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts (2010) "Predicting consumer behavior with Web search," *Proceedings of the National Academy of Sciences*, **107**(41), pp. 17486–17490.

[14] Kim, E. K., J. H. Seok, J. S. Oh, H. W. Lee, and K. H. Kim (2013) "Use of hangeul twitter to track and predict human influenza infection," *PloS one*, **8**(7), p. e69305.

[15] Yaari, R., G. Katriel, A. Huppert, J. Axelsen, and L. Stone (2013) "Modelling seasonal influenza: the role of weather and punctuated antigenic drift," *Journal of The Royal Society Interface*, **10**(84), p. 20130298.

[16] Salathé, M., M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones (2010) "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences*, **107**(51), pp. 22020–22025.

[17] Cauchemez, S., A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, D. Swerdlow, and Pennsylvania H1N1 working group (2011) "Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza." *Proceedings of the National Academy of Sciences of the United States of America*, **108**(7), pp. 2825–2830.

[18] Ferrari, M. J., S. E. Perkins, L. W. Pomeroy, and O. N. Bjørnstad (2011) "Pathogens, social networks, and the paradox of transmission scaling." *Interdisciplinary perspectives on infectious diseases*, **2011**, p. 267049.

[19] Salathé, M. and S. Khandelwal (2011) "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control." *PLoS computational biology*, **7**(10), p. e1002199.

[20] Wells, C. R., E. Y. Klein, and C. T. Bauch (2013) "Policy resistance undermines superspreader vaccination strategies for influenza," *PLoS computational biology*.

[21] Seale, H., A. E. Heywood, M.-L. McLaws, K. F. Ward, C. P. Lowbridge, D. Van, and C. R. MacIntyre (2010) "Why do I need it? I am not at risk! Public perceptions towards the pandemic (H1N1) 2009 vaccine," *BMC infectious diseases*, **10**(1), p. 99.

[22] Larson, H. J., D. M. Smith, P. Paterson, M. Cumming, E. Eckersberger, C. C. Freifeld, I. Ghinai, C. Jarrett, L. Paushter, J. S. Brownstein, and L. C. Madoff (2013) "Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines," *The Lancet Infectious Diseases*, **13**(7), pp. 606–613.

[23] Bansal, S., B. Pourbohloul, and L. A. Meyers (2006) "A comparative analysis of influenza vaccination programs." *PLoS medicine*, **3**(10), p. e387.

[24] Salathé, M. and S. Bonhoeffer (2008) "The effect of opinion clustering on disease outbreaks," *Journal of The Royal Society Interface*, **5**(29), pp. 1505–1508.

[25] Huang, Z., U. Kumar, and T. Bodnar (2013) "Understanding population displacements on location-based call records using road data," in *MobiGIS '13: Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, ACM Request Permissions.

[26] Balcan, D., V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco, and A. Vespignani (2009) "Multiscale mobility networks and the spatial spreading of infectious diseases," .

[27] Afzal, S., R. Maciejewski, and D. S. Ebert (2011) "Visual analytics decision support environment for epidemic modeling and response evaluation," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 191–200.

[28] Funk, S., M. Salathé, and V. A. Jansen (2010) "Modelling the influence of human behaviour on the spread of infectious diseases: a review," *Journal of the Royal Society Interface*, **7**(50), pp. 1247–1256.

[29] Jones, J. H., M. Salathe, et al. (2009) "Early assessment of anxiety and behavioral response to novel swine-origin influenza A (H1N1)," *PLoS one*, **4**(12), p. e8032.

[30] Gilad, E. and C. Watkins (2009) "The spread of awareness and its impact on epidemic outbreaks." *Proceedings of the National Academy of Sciences*, **106**(16), pp. 6872–6877.

[31] Tudor, A. (2003) "A (macro) sociology of fear?" *The Sociological Review*, **51**(2), pp. 238–256.

[32] Grüne-Yanoff, T. (2011) "Agent-Based Models as Policy Decision Tools: The Case of Smallpox Vaccination," *Simulation & Gaming*, **42**(2), pp. 225–242.

[33] Glasser, J., M. Meltzer, and B. Levin (2004) "Mathematical modeling and public policy: responding to health crises." *Emerging infectious diseases*, **10**(11), pp. 2050–2051.

[34] Eyssartier, C., A. H. Ladio, and M. Lozada (2008) "Cultural transmission of traditional knowledge in two populations of North-western Patagonia." *Journal of ethnobiology and ethnomedicine*, **4**, p. 25.

[35] Cavalli-Sforza, L. L., K. H. Chen, and S. M. Dornbusch (1982) "Theory and observation in cultural transmission." *Science (New York, N.Y.)*, **218**(4567), pp. 19–27.

[36] Rendell, L., R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland (2010) "Why copy others? Insights from the social learning strategies tournament," *Science*, **328**(5975), pp. 208–213.

[37] Helbing, D. and W. Yu (2010) "The future of social experimenting," *Proceedings of the National Academy of Sciences*, **107**(12), pp. 5265–5266.

[38] Centola, D. (2010) "The spread of behavior in an online social network experiment." *Science (New York, N.Y.)*, **329**(5996), pp. 1194–1197.

[39] Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012) "A 61-million-person experiment in social influence and political mobilization," *Nature*, **489**(7415), pp. 295–298.

[40] Timimi, F. K. (2013) "The Shape of Digital Engagement: Health Care and Social Media," *The Journal of ambulatory care management*, **36**(3), pp. 187–192.

[41] Kinsman, J. (2012) ""A time of fear": local, national, and international responses to a large Ebola outbreak in Uganda," *Global Health*.

[42] Bakshy, E., D. Eckles, and M. S. Bernstein (2014) "Designing and Deploying Online Field Experiments," in *Proceedings of the 23rd ACM conference on the World Wide Web*, ACM.

[43] McMahan, H. B., G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al. (2013) "Ad click prediction: a view from the trenches," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1222–1230.

[44] Bodnar, T., V. C. Barclay, N. Ram, C. S. Tucker, and M. Salathé (2014) "On the ground validation of online diagnosis with Twitter and medical records," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, International World Wide Web Conferences Steering Committee, pp. 651–656.

[45] Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti (2014) "Geo-located Twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science*, **41**(3), pp. 260–271.

[46] Leetaru, K., S. Wang, G. Cao, A. Padmanabhan, and E. Shook (2013) "Mapping the global Twitter heartbeat: The geography of Twitter," *First Monday*, **18**(5).

[47] Tatem, A. J. (2014) "Mapping population and pathogen movements," *International health*, **6**(1), pp. 5–11.

[48] Isella, L., M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, et al. (2011) "Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors," *PloS one*, **6**(2), p. e17144.

[49] Smieszek, T., V. C. Barclay, I. Seeni, J. J. Rainey, H. Gao, A. Uzicanin, and M. Salathé (2014) "How should social mixing be measured: comparing web-based survey and sensor-based methods," *BMC infectious diseases*, **14**(1), p. 136.

[50] Yang, W., M. Lipsitch, and J. Shaman (2015) "Inference of seasonal and pandemic influenza transmission dynamics," *Proceedings of the National Academy of Sciences*, **112**(9), pp. 2723–2728.

[51] Moser, M. R., T. R. Bender, H. S. Margolis, G. R. Noble, A. P. Kendal, and D. G. Ritter (1979) "An outbreak of influenza aboard a commercial airliner," *American journal of epidemiology*, **110**(1), pp. 1–6.

[52] Klontz, K. C., N. A. Hynes, R. A. Gunn, M. H. Wilder, M. W. Harmon, and A. P. Kendal (1989) "An outbreak of influenza A/Taiwan/1/86 (H1N1) infections at a naval base and its association with airplane travel," *American journal of epidemiology*, **129**(2), pp. 341–348.

[53] Suh, B., L. Hong, P. Pirolli, and E. H. Chi (2010) "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," , pp. 177–184.

[54] Kim, J. and J. Yoo (2012) "Role of Sentiment in Message Propagation: Reply vs. Retweet Behavior in Political Communication," in *Proceedings of the 2012 International Conference on Social Informatics*, SOCIALINFORMATICS '12, IEEE Computer Society, Washington, DC, USA, pp. 131–136.
URL http://dx.doi.org/10.1109/SocialInformatics.2012.33

[55] Stieglitz, S. and L. Dang-Xuan (2012) "Political Communication and Influence Through Microblogging–An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior," in *Proceedings of the 2012 45th Hawaii International Conference on System Sciences*, HICSS '12, IEEE Computer Society, Washington, DC, USA, pp. 3500–3509.
URL http://dx.doi.org/10.1109/HICSS.2012.476

[56] Gransee, S., R. McAfee, and A. Wilson, "Twitter Retweet Prediction," http://www.webcitation.org/6ajmw1B7o, accessed: 2014-10-08.

[57] Zhao, K., J. Yen, G. Greer, B. Qiu, P. Mitra, and K. Portier (2014) "Finding influential users of online health communities: a new metric based on sentiment influence," *Journal of the American Medical Informatics Association*, **21**(e2), pp. e212–e218.

[58] Bollen, J., H. Mao, and A. Pepe (2011) "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." *ICWSM*.

[59] Ofek, N., C. Caragea, L. Rokach, P. Biyani, P. Mitra, J. Yen, K. Portier, and G. Greer (2013) "Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon," , pp. 109–113.

[60] Russell, J. A. and A. Mehrabian (1977) "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, **11**(3), pp. 273–294.

[61] Plutchik, R. (2001) "The nature of emotions," *American Scientist*, **89**(4), pp. 344–350.

[62] Cambria, E., A. Livingstone, and A. Hussain (2012) "The Hourglass of Emotions," in *Proceedings of the 2011 International Conference on Cognitive Behavioural Systems*, COST'11, Springer-Verlag, Berlin, Heidelberg, pp. 144–157.
URL http://dx.doi.org/10.1007/978-3-642-34584-5_11

[63] Bodnar, T. and M. Salathé (2013) "Validating Models for Disease Detection Using Twitter," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 699–702.
URL http://dl.acm.org/citation.cfm?id=2487788.2488027

[64] Salathé, M., L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, and A. Vespignani (2012) "Digital epidemiology." *PLoS computational biology*, **8**(7), p. e1002616.

[65] Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts (2010) "Predicting consumer behavior with Web search." *Proceedings of the National Academy of Sciences of the United States of America*, **107**(41), pp. 17486–17490.

[66] Butler, D. (2013) "When Google got flu wrong," *Nature*, **494**(7436), pp. 155–156.

[67] Carneiro, H. A. and E. Mylonakis (2009) "Google Trends: A WebâĂŘBased Tool for RealâĂŘTime Surveillance of Disease Outbreaks," *Clinical Infectious Diseases*, **49**(10), pp. 1557–1564.

[68] Signorini, A., A. M. Segre, and P. M. Polgreen (2011) "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic." *PloS one*, **6**(5), p. e19467.

[69] Culotta, A. (2010) "Towards detecting influenza epidemics by analyzing Twitter messages ," in *the First Workshop*, ACM Press, New York, New York, USA, pp. 115–122.

[70] Dugas, A. F., Y. H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman (2012) "Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics," *Clinical Infectious Diseases*, **54**(4), pp. 463–469.

[71] Hawn, C. (2009) "Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care," *Health Affairs*, **28**(2), pp. 361–368.

[72] St Louis, C. and G. Zorlu (2012) "Can Twitter predict disease outbreaks?" *BMJ (Clinical research ed.)*, **344**, p. e2353.

[73] De la Torre-Díez, I., F. J. Díaz-Pernas, and M. Antón-Rodríguez (2012) "A content analysis of chronic diseases social groups on facebook and twitter." *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, **18**(6), pp. 404–408.

[74] Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2009) "Detecting influenza epidemics using search engine query data," *Nature*, **457**(7232), pp. 1012–1014.

[75] Marsland, S. (2014) *Machine learning: an algorithmic perspective*, CRC press.

[76] Vapnik, V. (1997) "Support vector regression machines," *Advances in neural information processing systems*, **9**, pp. 155–161.

[77] Bodnar, T., V. C. Barclay, N. Ram, C. S. Tucker, and M. Salathé (2014) "On the Ground Validation of Online Diagnosis with Twitter and Medical Records," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 651–656.

[78] Chan, E. H., T. F. Brewer, L. C. Madoff, M. P. Pollack, A. L. Sonricker, M. Keller, C. C. Freifeld, M. Blench, A. Mawudeku, and J. S. Brownstein (2010) "Global capacity for emerging infectious disease detection," *Proceedings of the National Academy of Sciences*, **107**(50), pp. 21701–21706, `http://www.pnas.org/content/107/50/21701.full.pdf+html`. URL `http://www.pnas.org/content/107/50/21701.abstract`

[79] Heymann, D. L. and G. R. Rodier (2001) "Hot spots in a wired world: {WHO} surveillance of emerging and re-emerging infectious diseases," *The Lancet Infectious Diseases*, **1**(5), pp. 345 – 353.

[80] Marquet, R. L., A. I. Bartelds, S. P. van Noort, C. E. Koppeschaar, J. Paget, F. G. Schellevis, and J. van der Zee (2006) "Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003-2004 influenza season," *BMC public health*, **6**(1), p. 242.

[81] van Noort, S. P., M. Muehlen, H. Rebelo de Andrade, C. Koppeschaar, J. M. Lima Lourenco, and M. G. Gomes (2007)

"Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe," *Euro Surveill.*, **12**(7), pp. 5–6.

[82] Olson, D. R., K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen (2013) "Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales," *PLoS computational biology*, **9**(10), p. e1003256.

[83] Lamb, A., M. J. Paul, and M. Dredze (2013) "Separating Fact from Fear: Tracking Flu Infections on Twitter," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Atlanta, Georgia, pp. 789–795.
URL `http://www.aclweb.org/anthology/N13-1097`

[84] Porter, M. F. (1980) "An algorithm for suffix stripping," *Program: electronic library and information systems*, **14**(3), pp. 130–137.

[85] Grubbs, F. E. (1969) "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, **11**(1), pp. 1–21.

[86] Todorovski, L. and S. Džeroski (2003) "Combining Classifiers with Meta Decision Trees," *Mach. Learn.*, **50**(3), pp. 223–249.
URL `http://dx.doi.org/10.1023/A:1021709817809`

[87] Tsirogiannis, G., D. Frossyniotis, K. Nikita, and A. Stafylopatis (2004) "A Meta-classifier Approach for Medical Diagnosis," in *Methods and Applications of Artificial Intelligence* (G. Vouros and T. Panayiotopoulos, eds.), vol. 3025 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 154–163.

[88] Adrover, C., T. Bodnar, Z. Huang, A. Telenti, and M. Salathé (2015) "Identifying Adverse Effects of HIV Drug Treatment and Associated Sentiments Using Twitter," *JMIR Public Health Surveillillance*, **1**(2), p. e7.
URL `http://publichealth.jmir.org/2015/2/e7/`

[89] Bilge, A. H., F. Samanlioglu, and O. Ergonul (2014) "On the uniqueness of epidemic models fitting a normalized curve of removed individuals," *Journal of mathematical biology*, pp. 1–28.

[90] Newman, M. E. and J. Park (2003) "Why social networks are different from other types of networks," *Physical Review E*, **68**(3), p. 036122.

[91] Smieszek, T. and M. Salathé (2013) "A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks," *BMC medicine*, **11**(1), p. 35.

[92] Fraser, C., S. Riley, R. M. Anderson, and N. M. Ferguson (2004) "Factors that make an infectious disease outbreak controllable," *Proceedings of the National Academy of Sciences of the United States of America*, **101**(16), pp. 6146–6151.

[93] Read, M., P. S. Andrews, J. Timmis, R. A. Williams, R. B. Greaves, H. Sheng, M. Coles, and V. Kumar (2013) "Determining disease intervention strategies using spatially resolved simulations," .

[94] Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. (2008) "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med*, **5**(3), p. e74.

[95] Diekmann, O., H. Heesterbeek, and T. Britton (2012) *Mathematical tools for understanding infectious disease dynamics*, Princeton University Press.

[96] Heesterbeek, J. (2002) "A brief history of R 0 and a recipe for its calculation," *Acta biotheoretica*, **50**(3), pp. 189–204.

[97] Cattuto, C., W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani (2010) "Dynamics of person-to-person interactions from distributed RFID sensor networks." *PloS one*, **5**(7), p. e11596.

[98] Liem, N. T., I. A. I. I. Team, and W. L. Vietnam (2005) "Lack of H5N1 avian influenza transmission to hospital employees, Hanoi, 2004," *Emerging infectious diseases*, **11**(2), p. 210.

[99] Center for Disease Control (2014).
URL `http://www.cdc.gov/flu/takingcare.htm` `(http://www.webcitation.org/6axXG9nMV)`

[100] Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley (2013) "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," *arXiv preprint arXiv:1306.5204*.

[101] Banzhaf, W., P. Nordin, R. E. Keller, and F. D. Francone (1998) *Genetic programming: an introduction*, vol. 1, Morgan Kaufmann San Francisco.

[102] Cowling, B. J., S. Ng, E. S. Ma, C. K. Cheng, W. Wai, V. J. Fang, K.-H. Chan, D. K. Ip, S. S. Chiu, J. M. Peiris, et al. (2010) "Protective efficacy of seasonal influenza vaccination against seasonal and

pandemic influenza virus infection during 2009 in Hong Kong," *Clinical infectious diseases*, **51**(12), pp. 1370–1379.

[103] Weinstein, R. A., C. B. Bridges, M. J. Kuehnert, and C. B. Hall (2003) "Transmission of influenza: implications for control in health care settings," *Clinical Infectious Diseases*, **37**(8), pp. 1094–1101.

[104] Cowling, B. J., K.-H. Chan, V. J. Fang, C. K. Cheng, R. O. Fung, W. Wai, J. Sin, W. H. Seto, R. Yung, D. W. Chu, et al. (2009) "Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial," *Annals of Internal Medicine*, **151**(7), pp. 437–446.

[105] Olinky, R. and L. Stone (2004) "Unexpected epidemic thresholds in heterogeneous networks: The role of disease transmission," *Physical Review E*, **70**(3), p. 030902.

[106] Molina, C. and L. Stone (2012) "Modelling the spread of diseases in clustered networks," *Journal of theoretical biology*, **315**, pp. 110–118.

[107] Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, C. B. Bridges, N. J. Cox, and K. Fukuda (2004) "Influenza-associated hospitalizations in the United States," *Jama*, **292**(11), pp. 1333–1340.

[108] Glidden, M. and J. Blomo (2012), "GPS vs WiFi: The Battle for Location Accuracy Using Yelp Check-Ins," `http://engineeringblog.yelp.com/2012/08/gps-vs-wifi-the-battle-for-location-accuracy-using-yelp-check-ins.html`.

[109] Zandbergen, P. A. and S. J. Barbeau (2011) "Positional accuracy of assisted gps data from high-sensitivity gps-enabled mobile phones," *Journal of Navigation*, **64**(03), pp. 381–399.

[110] Djuknic, G. M. and R. E. Richton (2001) "Geolocation and assisted GPS," *Computer*, **34**(2), pp. 123–125.

[111] Modsching, M., R. Kramer, and K. ten Hagen (2006) "Field trial on GPS Accuracy in a medium size city: The influence of built-up," in *3rd Workshop on Positioning, Navigation and Communication*, pp. 209–218.

[112] Street, J. E. (2010) "Deceiving the heavens to cross the sea," DEFCON 18.
URL `https://www.youtube.com/watch?v=EzGwO5L9oq4`

[113] Boshmaf, Y., I. Muslukhov, K. Beznosov, and M. Ripeanu (2011) "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th Annual Computer Security Applications Conference*, ACM, pp. 93–102.

[114] Krause, B., C. Schmitz, A. Hotho, and G. Stumme (2008) "The anti-social tagger: detecting spam in social bookmarking systems," in *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, ACM.

[115] Bíró, I., J. Szabó, and A. A. Benczúr (2008) "Latent dirichlet allocation in web spam filtering," , pp. 29–32.

[116] Castillo, C., M. Mendoza, and B. Poblete (2011) "Information credibility on twitter," , pp. 675–684.

[117] Salathé, M., D. Q. Vu, S. Khandelwal, and D. R. Hunter (2013) "The dynamics of health behavior sentiments on a large online social network," *EPJ Data Science*, **2**(1), pp. 1–12.

[118] Salathé, M. and S. Khandelwal (2011) "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," *PLoS computational biology*, **7**(10), p. e1002199.

[119] Ediger, D., K. Jiang, J. Riedy, D. Bader, C. Corley, R. Farber, W. N. Reynolds, et al. (2010) "Massive social network analysis: Mining twitter for social good," in *Parallel Processing (ICPP), 2010 39th International Conference on*, IEEE, pp. 583–593.

[120] Weng, L., A. Flammini, A. Vespignani, and F. Menczer (2012) "Competition among memes in a world with limited attention." *Scientific reports*, **2**, p. 335.

[121] Osborne, M. and V. Lavrenko (2011) "Rt to win! predicting message propagation in twitter," in *ICWSM*.

[122] Lerman, K. and R. Ghosh (2010) "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks." *ICWSM*, **10**, pp. 90–97.

[123] Kwak, H., C. Lee, H. Park, and S. Moon (2010) "What is Twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*, ACM, pp. 591–600.

[124] Dou, W., X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou (2012) "Leadline: Interactive visual analysis of text data through event identification and exploration," , pp. 93–102.

[125] Fan, R., J. Zhao, Y. Chen, and K. Xu (2013) "Anger is more influential than joy: sentiment correlation in Weibo," *arXiv preprint arXiv:1309.2402*.

[126] Thoits, P. A. (1989) "The sociology of emotions," *Annual review of sociology*, pp. 317–342.

[127] Burger, J. D., J. Henderson, G. Kim, and G. Zarrella (2011) "Discriminating gender on Twitter," in *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.

[128] Go, A., R. Bhayani, and L. Huang (2009) "Twitter sentiment classification using distant supervision," *CS224N Project Report*.

[129] Krause, B., C. Schmitz, A. Hotho, and G. Stumme (2008) "The Anti-Social Tagger - Detecting Spam in Social Bookmarking Systems," in *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*.

[130] Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng (2008) "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, pp. 254–263.

[131] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011) "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, pp. 2825–2830.

[132] Viboud, C., P.-Y. Boëlle, S. Cauchemez, A. Lavenu, A.-J. Valleron, A. Flahault, and F. Carrat (2004) "Risk factors of influenza transmission in households," *British Journal of General Practice*, **54**(506), pp. 684–689.

[133] Mounts, A. W., H. Kwong, H. S. Izurieta, Y.-y. Ho, T.-k. Au, M. Lee, C. B. Bridges, S. W. Williams, K. H. Mak, J. M. Katz, et al. (1999) "Case-control study of risk factors for avian influenza A (H5N1) disease, Hong Kong, 1997," *Journal of Infectious Diseases*, **180**(2), pp. 505–508.

[134] Dinh, P. N., H. T. Long, N. T. K. Tien, N. T. Hien, L. T. Q. Mai, L. H. Phong, L. Van Tuan, H. Van Tan, N. B. Nguyen, P. Van Tu, et al. (2006) "Risk factors for human infection with avian influenza A H5N1, Vietnam, 2004," *Emerging infectious diseases*, **12**(12), p. 1841.

[135] OâĂŹRiordan, S., M. Barton, Y. Yau, S. E. Read, U. Allen, and D. Tran (2010) "Risk factors and outcomes among children admitted to

hospital with pandemic H1N1 influenza," *Canadian Medical Association Journal*, **182**(1), pp. 39–44.

[136] LeCun, Y., Y. Bengio, and G. Hinton (2015) "Deep learning." *Nature*, **521**(7553), pp. 436–444.

[137] El Hihi, S. and Y. Bengio (1995) "Hierarchical Recurrent Neural Networks for Long-Term Dependencies." in *NIPS*, pp. 493–499.

[138] Ferrari, S. and M. Jensenius (2008) "A constrained optimization approach to preserving prior knowledge during incremental training," *Neural Networks, IEEE Transactions on*, **19**(6), pp. 996–1009.

[139] Pascanu, R., T. Mikolov, and Y. Bengio (2012) "On the difficulty of training Recurrent Neural Networks," *arXiv.org*, `1211.5063v2`.

[140] Sutskever, I., J. Martens, and G. E. Hinton (2011) "Generating text with recurrent neural networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1017–1024.

[141] Graves, A., A.-r. Mohamed, and G. Hinton (2013) "Speech Recognition with Deep Recurrent Neural Networks," *arXiv.org*, `1303.5778v1`.

Vita

Todd Bodnar

For more information, visit ToddBodnar.com.

Education

- PhD, Biology, Fall 2015, Pennsylvania State University, For details: see above
- Complex Systems Summer School, June 2013, Santa Fe Institute
- B.S., Computer Science, May, 2012, Pennsylvania State University, Mathematics Minor

Peer Reviewed Publications

- Cosme Adrover, **Todd Bodnar**, Zhuojie Huang, Amalio Asensio Telenti, Marcel Salathé. *Identifying Adverse Effects of HIV Drug Treatment and Associated Sentiments Using Twitter* Journal of Medical Internet Research 2015
- **Todd Bodnar**, Conrad Tucker, Kenneth Hopkinson, and Sven G. Bilén. *Increasing the Veracity of Event Detection on Social Media Networks Through User Trust Modeling* IEEE BigData 2014
- **Todd Bodnar**, Victoria Barclay, Nilam Ram, Conrad Tucker and Marcel Salathé. *On the Ground Validation of Online Diagnosis with Twitter and Medical Records* WWW 2014
- Zhuojie Huang, Udayan Kumar, **Todd Bodnar** and Marcel Salathé. *Understanding Population Displacements on Location-Based Call Records Using Road Data* SIGSPATIAL 2013,
- **Todd Bodnar** & Marcel Salathé. *Validating Models for Disease Detection Using Twitter* WWW 2013 Companion, May 13-17, 2013, Rio de Janeiro, Brazil.
- Marcel Salathé, Linus Bengtsson, **Todd J. Bodnar**, Devon D. Brewer, John S. Brownstein, Caroline Buckee, Ellsworth M. Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L. Mabry, Alessandro Vespignani. *Digital Epidemiology* Plos Computational Biology, 2012
- **Todd Bodnar** & Marcel Salathé. *Governing the Global Commons with Local Institutions.* PloS One, 2012