Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Ilya Shats Date

Incorporating Social Relationships from Call Detail Records into Infectious Disease Spread Simulators

By

Ilya Shats
Master of Science

Mathematics/Computer Science

_____

Li Xiong
Advisor

_____ _____

Michele Benzi
Committee Member

Vicki Hertzberg
Committee Member

_____ _____

Samuel Jenness
Committee Member

Ymir Vigfusson
Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

Incorporating Social Relationships from Call Detail Records into Infectious Disease Spread Simulators

By

Ilya Shats

Advisor: Li Xiong, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Mathematics/Computer Science
2016

Abstract

Incorporating Social Relationships from Call Detail Records into Infectious Disease Spread Simulators

By Ilya Shats

Traditionally, mathematical models and surveying have been central to studying the spread of infectious diseases. In this work, we used an anonymized call-detail-record (CDR) dataset, which contains metadata about phone calls, text messages, and data transmissions, as the foundation for predicting spread of influenza-like-illness (ILI) during the 2009 Flu Pandemic in Iceland. The CDR provides population's mobility patterns and in addition to a basic contact tracing, this data can be used to infer people's social networks. Here we show that social strength has an impact on disease spread, supporting the perhaps intuitive idea that an infected individual is likely to transmit the disease to people socially closest to him or her. To simulate ILI spread throughout populations, we built several discrete event simulators (written in the Python programming language) that are described in the second part of the thesis. Though there is still work to be done in improving the models' accuracy in predicting the spread, it is a step forward in the novel area of using cell phone metadata to model infectious disease dynamics.

Incorporating Social Relationships from Call Detail Records into Infectious Disease Spread Simulators

By

Ilya Shats

Advisor: Li Xiong, Ph.D.

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Mathematics/Computer Science
2016

Acknowledgements

I would like to acknowledge all the people who helped me along the way.

First and foremost, my advisor Dr. Li Xiong who provided encouragement and guidance throughout the research process. In addition, I would like to thank Dr. Ymir Vigfusson for his continued support and for providing me with the data and foundation for the project, and Dr. Becky Mitchell for her support and her insight into epidemiology.

Table of Contents

Tables and Figures

Section 1. Introduction

In just the United States, over 200,000 people are hospitalized every year due to complications brought about by influenza [1]. With the recent 2009 flu pandemic, reliable methods of surveillance are becoming more and more important to intercepting influenza early on.

Conventional methods of tracking and predicting the spread of infectious diseases throughout populations rely on surveying the infected individuals and their families to track their inter- and intra-community movement as well as on mathematical and computational models utilizing a variety of data, such as demographics, immigration and emigration patterns, vaccination data, etc. [2] Surveying, however, is a tedious process, and its efficacy is heavily limited by the quality and quantity of data on user interactions.

1.1 Motivation

With the rapid expansion of the Internet and mobile technology, new possibilities of building models that utilize new types of data appear.

For example, Google Flu Trends, introduced by Google in 2008 provided a novel way of tracking flu by studying search queries. While the idea was promising, since around 1.2 billion people worldwide use Google [3], it ultimately failed when it drastically missed the peak of 2013 flu [4]. One of the possible reasons is that the intent behind Google queries is not known. People may query “flu” for a variety of reasons: curiosity, research project, etc.

In contrast to Google data, metadata of mobile phone calls and text messages does not contain subjective data. Every time a call is made or a text message is sent, a service provider saves the call or text message's metadata into a Call-Detail-Record (CDR). CDR contains the origin and destination addresses, the time the call was initiated, the duration of the call, the unique ID of a cell tower used, and many more attributes. Though currently it is challenging to acquire, CDR has been used in various studies to predict carbon footprint, caller's gender [5], and personality [6] and study urban dynamics [7]. Mobile data, CDR included, has been used to study human mobility patterns [8-13]. CDR also has utility in epidemiological studies relying on mobility patterns [14-18] and has been used to study the effect of government intervention on inhibiting the spread of H1N1 in Mexico [2,19].

There are about 6.8 billion mobile phone subscriptions worldwide [20]. Simply put, more people are calling and texting than are submitting Google queries. With the ability to capture the individuality of human mobility, leveraging mobile phone data can prove to be invaluable in tracking the spread of infectious diseases [21,22].

In this work, we used an anonymized CDR dataset as the foundation for simulating the spread of influenza-like-illness (ILI) during the 2009 Flu Pandemic in Iceland.

Influenza A virus (H1N1) infected millions of people worldwide during the 2009 flu pandemic. In Iceland there were 8,650 confirmed cases, which roughly translates to 28 confirmed cases per 1,000 inhabitants [23].

With the massive amounts of data that CDR provides about user interactions and mobility patterns, we can build a novel surveillance method that will allow for premature threat analysis and intervention.

1.2 Contribution

In this project, we investigated the impact of social connections on disease propagation and built a framework that attempts to model disease spread using a contact tracing approach integrated with disease and social network information (e.g. social strength).

The first part of the thesis provides descriptive work that investigates the impact of social networks on disease onset dates. Do closely connected people infect each other more often than those they are not connected with, and more specifically, can we use the data to extract variables that act as a reliable metric for measure social strength?

The second part of the thesis describes several Discrete Event Simulators that were built to model the spread of H1N1 in Iceland during the 2009 Flu Pandemic.

Our hypothesis was that social strength has an impact on disease propagation, and the social network information extracted from the CDR dataset can be used to improve the accuracy of disease simulators that lack that information.

Section 2. Understanding the impact of social strength on disease spread

Before building discrete event simulators that use social network information to predict infection, we first investigated whether social relationships have a connection with disease offset. Specifically, does a pair with a strong social connection increase the likelihood that the pair of users get infected close to each other? Of course, we do not know when the users were infected, but we use the diagnosis information as a proxy for infection.

2.1 Description of data

This work makes use of two datasets.

The first dataset, the CDR dataset, contains metadata about user's cellphone records. Specifically, it contains metadata about calls and texts for roughly half of the population between February 2009 and June 2012 - about 41 months. This includes data such as the unique user ids (UID) of parties participated in a phone call, the duration of the phone call, and the cellphone tower the user connected to. There are a little over 970 million records in the CDR dataset.

Table 1. A fragment of the CDR dataset. The direction IN means that UID 1 called UID2. OUT means that UID2 called UID1.

UID 1	UID 2	Timestamp	Direction	Tower ID	Duration (sec)
10884794	8106006	2009-02-01 00:00:00	IN	274010019BCD	10
10557308	4992194	2009-03-05 05:11:31	OUT	27401100CD6E	512
52090496	11334959	2009-03-04 11:02:59	OUT	27401012D0BD	31

The cell phone tower coordinates serve as an approximation of the location of the user who initiated the call. Depending on the density of the cell phone towers in the area, this may be accurate within a couple hundred meters or a couple kilometers.

The second dataset, which we will call the ILI-onset dataset, lists individuals and dates they were diagnosed with ILI during the 2009 flu pandemic. It contains 4,346 records and covers diagnoses from March 2009 to November 2010. It is linked by UID with the CDR-dataset, meaning that we can match a user diagnosed with ILI with records in the CDR data.

Table 2. A fragment of the ILI-onset dataset.

UID	Disease Onset Date
4205788	2009-10-13
4227181	2009-08-17
4227464	2009-10-26

Table 3. CDR dataset statistics.

CDR Dataset	
Total Records	972,800,043
Number of unique UIDs	2,170,454
Spanning Period	41 months

Table 4. ILI-Onset dataset statistics.

ILI-Onset Dataset	
Total Records	4,294
Maximum number of people diagnosed on one day	194
Date of highest number of diagnoses	10/20/2009
Mean/Variance of number of diagnoses per day	16 / 909.857
Spanning Period	20 months

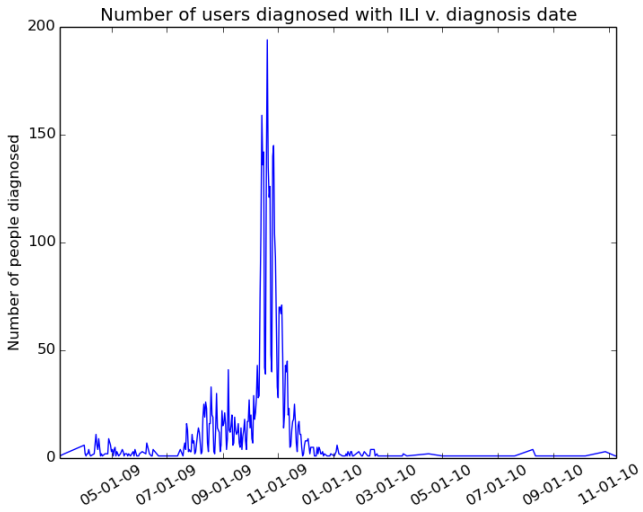


Figure 1. Daily disease onset curve

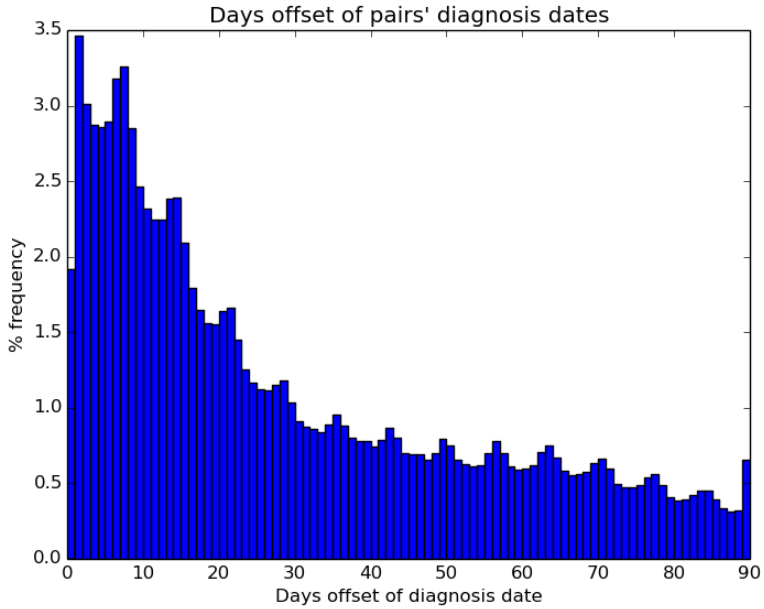


Figure 2. Distribution of disease onset offsets among all infected pairs

2.2. Data Limitations

In Iceland, there were over 8,650 confirmed cases of ILI during the 2009 flu pandemic [23]. The ILI-onset dataset contains 4,346 records, which is roughly half of confirmed ILI cases. In addition, we can assume that many people who contracted H1N1 did not go to the doctor, and thus were not recorded in the ILI-onset dataset. In fact, the estimated proportion of the population that was infected ranges from 10% to 22% [24], though it appears to be 3% if we look at just the confirmed cases. These infected people provided a vector of transmission that the models cannot capture. This relates to the second limitation that the CDR-dataset does not have data on every person in Iceland, so those missing data points could act as indirect vectors of transmission. There is a similar problem with the ILI-onset dataset, namely, not all infected individuals are represented. Infected individuals may neglect to go to the doctor and thus are never accounted for.

That being said, infectious individuals who did not see the doctor, because they are asymptomatic are not as important to disease spread as are symptomatic individuals [25].

The coordinates in the CDR dataset represent the location of the cell-phone towers, so we do not know the exact location where the users were making calls or sending text messages from. The accuracy of the approximation depends on the density of cellphone towers in the region.

The CDR dataset has no information on who a user texted. Considering the popularity of texting especially among teenagers, we lose out on the ability to use text messaging frequency and patterns to infer social strength among users.

Cell phone tower switches are not captured in the data. If a user initiates a call and then during the call connects to another cell phone tower, the data does not show those towers. This may be more prevalent when people are driving, but since it is unlikely that people transmit the disease to other people on the road, the missing data points are not hindrances.

Considering that the population of Iceland in 2010 was around 320,000 [26], and there are over one million unique UIDs in the CDR dataset, many of these UIDs may represent not individuals but businesses. Also, phone ownership in Iceland in 2010 was at 107% [27] and some users own multiple phones. Currently, we have no reliable way of cleaning the CDR dataset.

The data we have only represents the adult population (at least 18 years of age). Knowing the role schools played in H1N1 spread [24], access to data on schoolers would significantly improve our model.

The last limitation worth briefly discussing is sparsity. Given that coming into contact with infected individuals is a requirement for disease transmission in H1N1 and influenza-like illnesses, we need to know when users cross each other's paths. In the CDR data, intersections are not as common. This is because people are not using their phones all the time.

2.3. Data preprocessing

We wrote several Python scripts to clean and filter the data. This involved removing records where the tower_id or coordinates were missing, which brought down 1,702 towers to 1,502, condensing the data to remove redundant fields, and filtering by requiring a minimum number of contacts, minimum number of records, minimum number of active days, etc. In addition, data files were transformed into hashmaps to turn the linear-time operation of scanning files line by line, into a constant-time lookup operation. For instance, the ILI-onset dataset was transformed into a hashmap where the key was the UID and the value was the date of disease onset. The towers file was turned into a hashmap where the key was the towerID and the value was a tuple of the coordinates.

The three variables that were extracted from the CDR-dataset are *call duration*, *call frequency*, and *co-occurrence*. Two users co-occur if they have a record at the same tower within 30 minutes of each other. These variables are used as a proxy for measuring social strength between two users. Logically, we would expect that people who call more frequently and call for a longer period of time are more closely connected than those who do not. The logic behind using co-occurrence as an indication of social strength is not as straight-forward. With phone calls, one reciprocated call suggests that there exists a connection between two users. Co-occurrences on the other hand do not necessarily suggest that. Certainly, people who make phone calls on the way to work would co-occur with everyone making phone calls near them. However, since physical proximity is a requirement for ILI transmission, it is interesting to investigate this variable. Perhaps, after a certain number of co-occurrences, the likelihood of strong connections increases.

2.3.1. Building hashmaps from datasets

The process for extracting the variables and populating the hashmap was as follows.

Call duration:

For each CDR record, we extract the two UIDs of users, who interacted with each other, along with the duration in seconds, and added an entry to the hashmap. The first UID is the key. The second UID is a key to the nested hashmap and the duration is the value. The final object looks like this.

{UID1: {UID2: 300, UID3: 500}}

In this example, UID1 called UID2 for 300 seconds and called UID3 for 500 seconds.

The reciprocated entry was added into the hashmap: if UID1 talked to UID2, then UID2 talked to UID1 for the same amount of time.

Call frequency:

The method for extracting call frequency was very similar. If an interaction occurred between UID1 and UID2, we incremented their call frequency by 1. Only calls are considered since the data includes only the user who sent the text, not the user who received it.

Co-occurrence:

How does one know if two users co-occur? Is meeting up at the exact same place at the exact same time a co-occurrence? What if records are five minutes apart, should that still count as a co-occurrence?

We define *co-occurrence* as follows: Two users co-occur if they both have records at the same tower within 30 minutes of each other. Two users who call each other multiple times at the same place within the 30 minute range, count as having co-occurred once.

The structure of the hashmap is the same as the structure above. A UID has several contacts each of which have a number of co-occurrences.

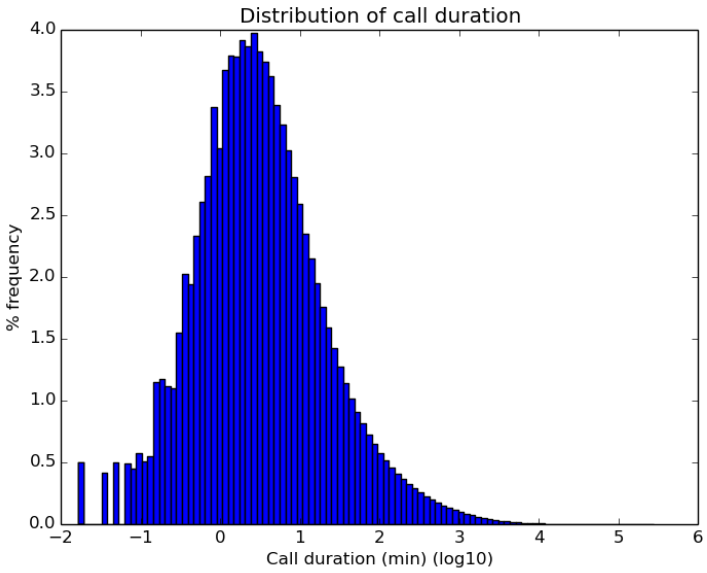


Figure 3: Distribution of call duration

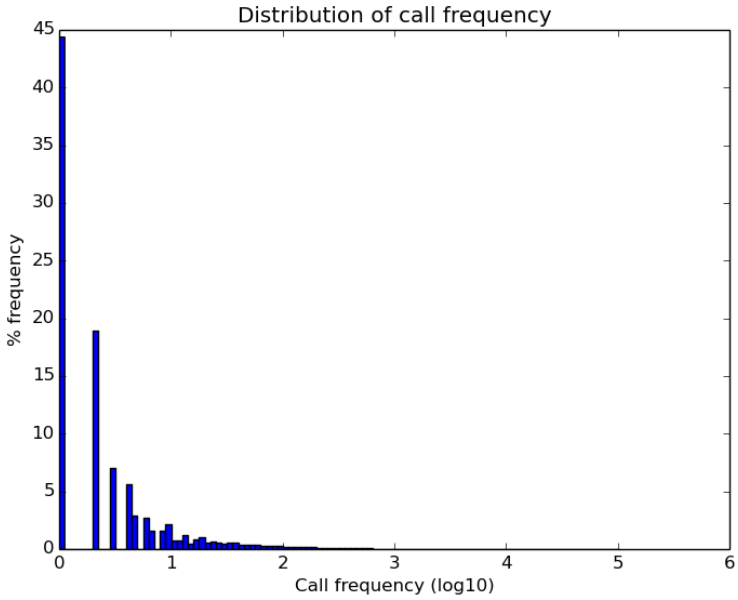


Figure 4: Distribution of call frequency

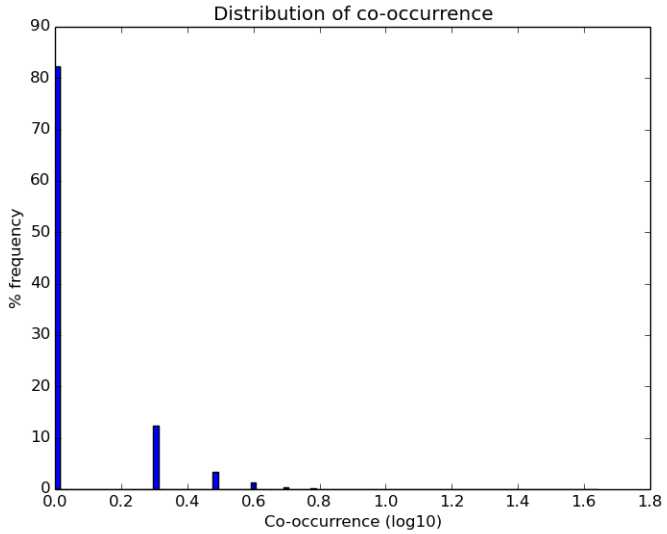


Figure 5: Distribution of co-occurrence

The peak of Figure 3 amounts to about 2.5 minutes. Most pairs who exchanged calls, called for a very short period of time. At the right tail of the distribution, the maximum time was about 277,000 minutes which amounts to 3.75 hours per day. It is likely that the pairs at the right tail are companies.

The peak of Figure 4 tells us that most people who called a contact, called him or her only once. The six outliers at the right-most side exchanged a little over 100,000 calls over the 41 month period.

From Figure 5, we see that most pairs co-occurred only once, and the maximum number of co-occurrences was 44.

We built three weighted undirected graphs, a call duration graph, a call frequency graph, and a co-occurrence graph, each weighted by call duration, frequency, and co-occurrence, respectively. We display statistics about the data as well as the graphs

below. Since the only difference between the call duration and call frequency graphs are the weights (a pair connected in one of those graphs will be connected in the other), we display statistics about the call duration and call frequency graphs in one table.

Table 5. Call Duration and Call Frequency statistics.

Call Duration and Frequency	
Number of users who called someone	2,153,738
Percentage of total users who called	99.230%
Average number of people a user called	40.513
Maximum number of people a user called	128,122
Minimum number of people a user called	1
Average/Variance number of calls per user	15.045
Maximum number of calls	102,315
Minimum number of calls	1
Average call duration per user (seconds)	2071.091
Maximum call duration (seconds)	1,6642,897
Minimum call duration (seconds)	1
Number of connected components	8
Average clustering coefficient	0.053
Average degree centrality	1.553e-05

Table 6. Co-occurrence statistics.

Co-occurrence	
Number of users who co-occurred with someone	36,587
Percentage of total users who co-occurred	1.686%
Average number of people a user co-occurred with	181
Maximum number of people a user co-occurred with	2720
Minimum number of people a user co-occurred with	1
Average number of co-occurrences per user	1.265
Maximum number of co-occurrences per user	44
Minimum number of co-occurrences per user	1
Number of connected components	287
Average clustering coefficient	0.374
Average degree centrality	0.008

2.3.2. Dependency between variables

Before investigating the impact of these variables on disease spread, we first study the correlation or dependency between these variables

In the Figures below, when we say *average*, what we mean is that we bucket the data, average the y-values in the bucket, and plot that.

From Figure 6, we see a clear correlation between call frequency and call duration. This makes sense, because every call will have a positive call duration, so a higher call frequency will increase the number of total minutes spent calling. The call duration is

thresholded at about 33,000 minutes, which focuses in on the pairs who talked on the phone for a realistic duration. 33,000 minutes corresponds to about an hour of talking to one person every day for a period of 18 months.

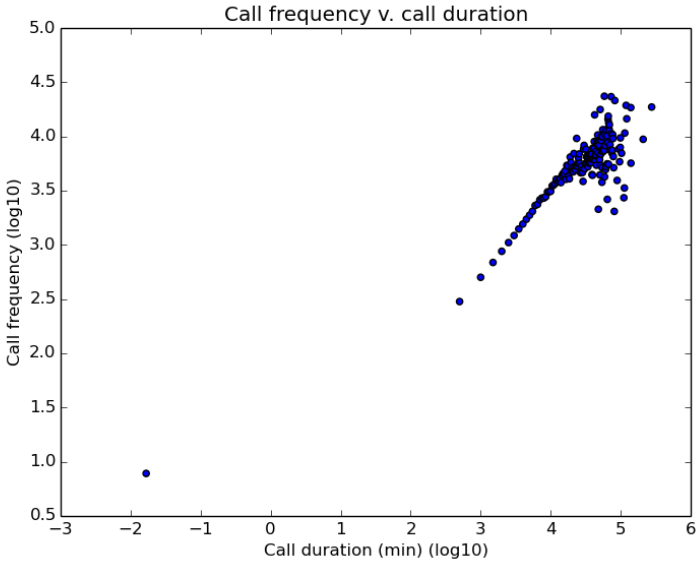


Figure 6. Call Frequency vs Call Duration
Spearman Correlation Coefficient: 0.725
p-value: < 0.001

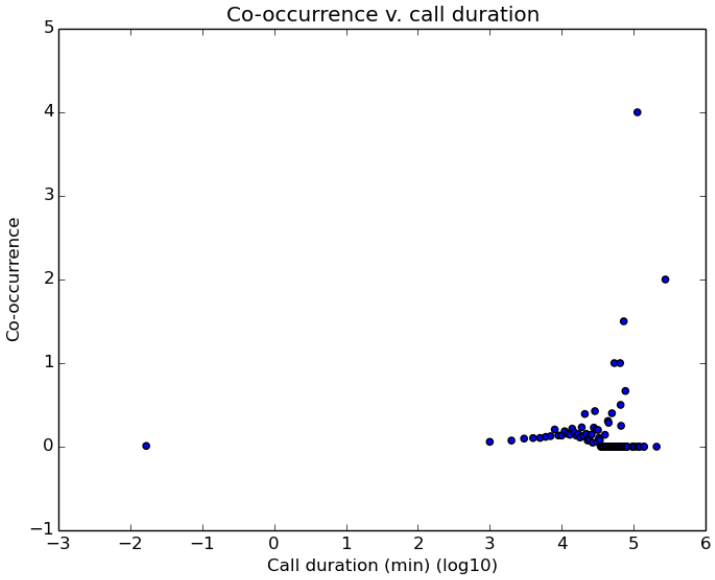


Figure 7. Co-occurrence vs Call Duration
Spearman Correlation Coefficient: 0.032
p-value: < 0.001

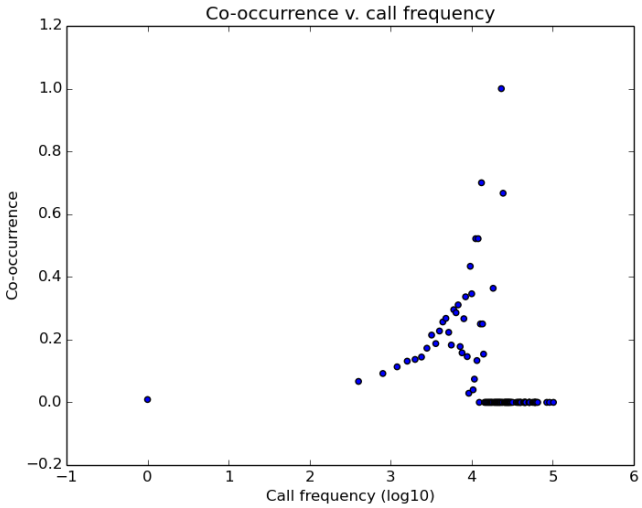


Figure 8. Co-occurrence vs Call Frequency
Spearman Correlation Coefficient: 0.060
p-value: < 0.001

Figure 7 shows virtually no correlation between co-occurrence and call duration. A high

co-occurrence does not imply a high call frequency. After all, two people making calls on the way to work are co-occurring, and this does not suggest anything about the strength of their relationship. We would expect two people who call each other frequently to co-occur more often than those who don't, but it does not appear that they do. The low co-occurrence could be a result of the sparsity of the data. The data captures a small percentage of the co-occurrences that actually happened. It is important to note that in Figure 7 and Figure 8, we filter out pairs that had a positive call duration, but did not co-occur.

2.4. The impact of social strength

Here, we examine how call duration, call frequency, and co-occurrence correlate with the time difference (in number of days) when two users got infected. To generate the following graphs, we iterated through every record in the ILI-onset dataset and retrieve their contacts, which are the users they called. We filtered out the users that never got diagnosed with ILI and then got the diagnosis dates of his/her contacts. Subtracting the two dates and taking their absolute value gives us the offset. We then constructed a hashmap, where the key is the offset and the value is a list of every pairs' call duration. We then binned the data into 350 buckets and plotted the average call duration of each bucket. The same technique is used for the other variables.

There is a negative correlation between call duration and the offset meaning that the people, who call each other for a longer time, tend to get infected temporally near each

other. Though we do not have the ground truth to verify that people with high call duration are more closely connected, it is fair to assume that the premise is true. In the figure, we see that if we use call duration as a measure of social strength, Social strength is negatively correlated with the closeness of diagnosis.

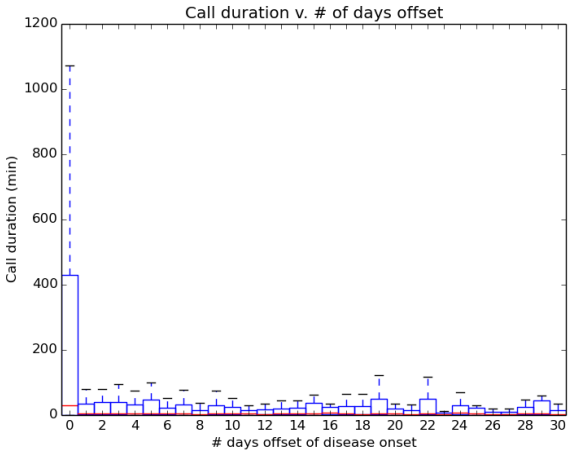


Figure 9. Call Duration vs Days Offset of Disease Onset
Spearman Correlation Coefficient: -0.531
p-value: 0.00213

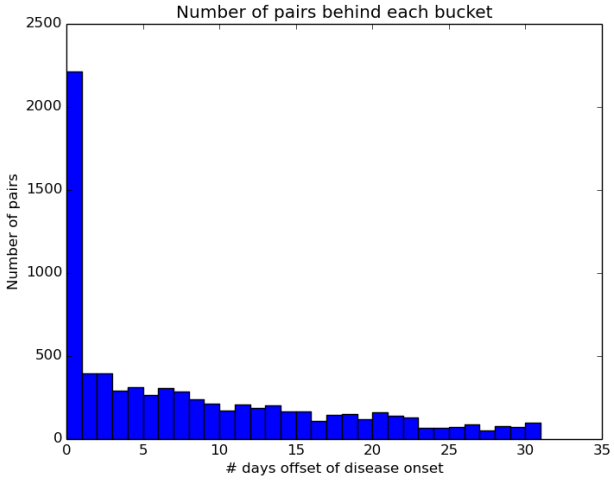


Figure 10. Number of pairs in each bucket in Figure 9

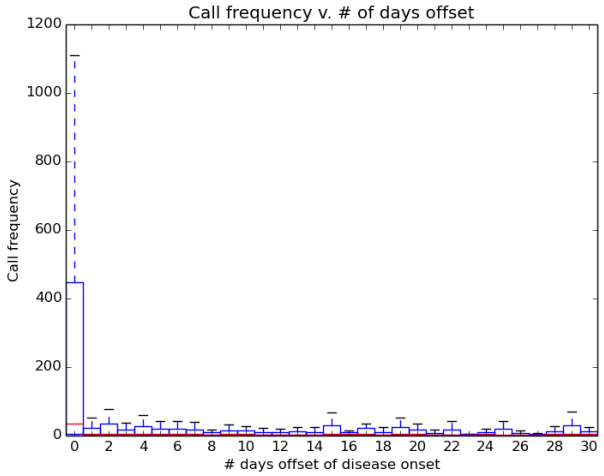


Figure 11. Call Frequency vs Days Offset of Disease Onset
Spearman Correlation Coefficient: -0.538
p-value: 0.00180

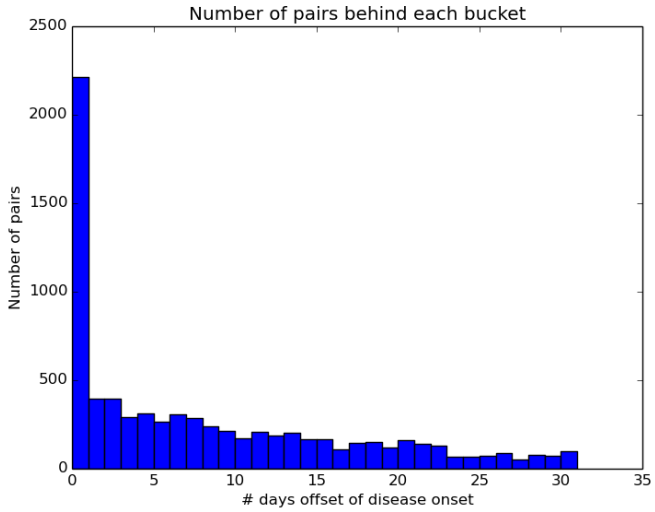


Figure 12. Number of pairs behind each bucket in Figure 11

We see a similar trend if we use call frequency as a measure of social strength. The infected users who called their contacts more frequently had a closer date of diagnosis. This is especially evident at offset 0.

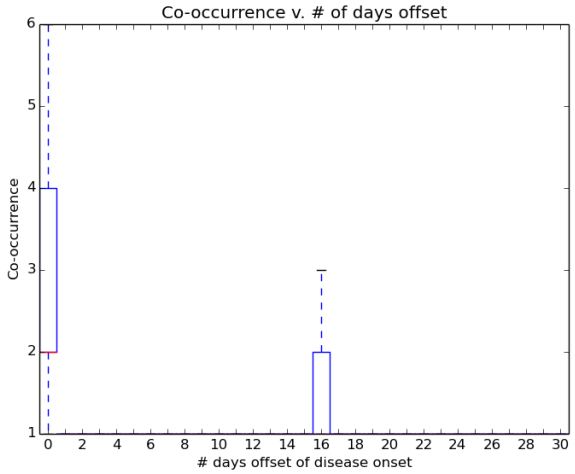


Figure 13. Co-occurrence vs Days Offset of Disease Onset
Spearman Correlation Coefficient: -0.198
p-value: 0.287

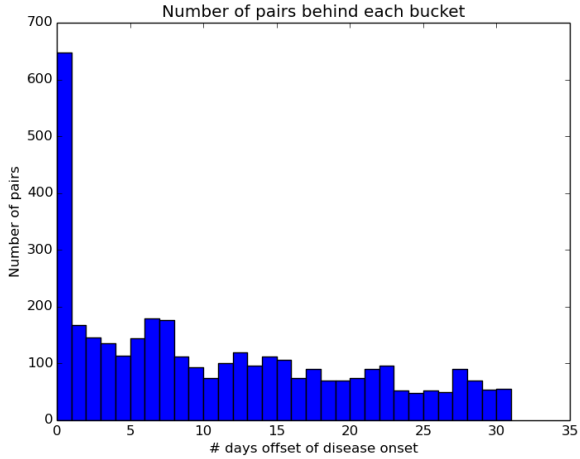


Figure 14. Number of pairs behind each bucket in Figure 13

In Figure 13, while we do see an unusual spike at offset 0, the Spearman coefficient is fairly low. There does exist a negative correlation, but it is not nearly as strong as the correlation of call duration and call frequency vs. disease offset.

While co-occurrence does not appear to be a strong indicator of social strength, call duration and call frequency do appear to be correlated somewhat strongly.

Next, we study what percentage of users' top contacts got infected within a range. The baseline figures below are calculated by shuffling users' relationships. So if previously a user had five contacts that user would still have five contacts, but they would be selected randomly. First, we attempt to answer the following question: *Of the pairs that had a higher call volume, did a higher proportion of them get infected near each other?* The idea is that high call volume suggests closer social connection, which implies more frequent contact and closer physical proximity, which is necessary for the disease to propagate.

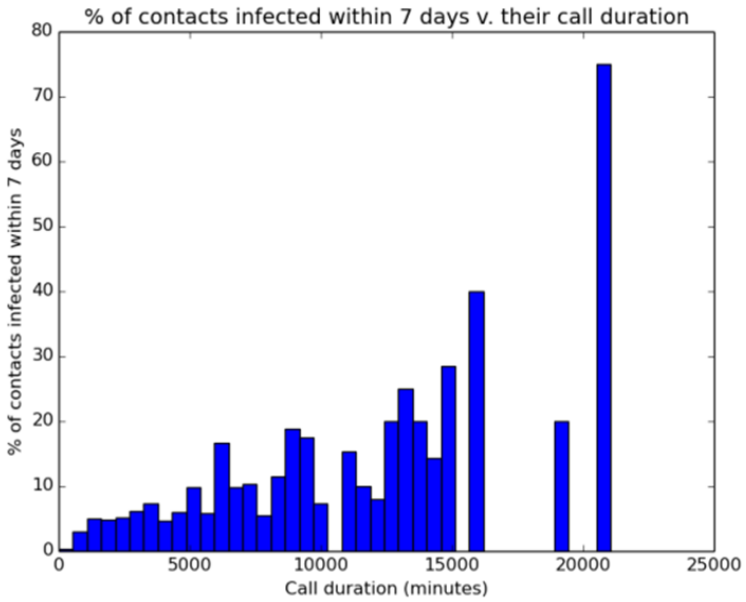


Figure 15. Percentage of Contacts Infected within 7 days vs Call Duration

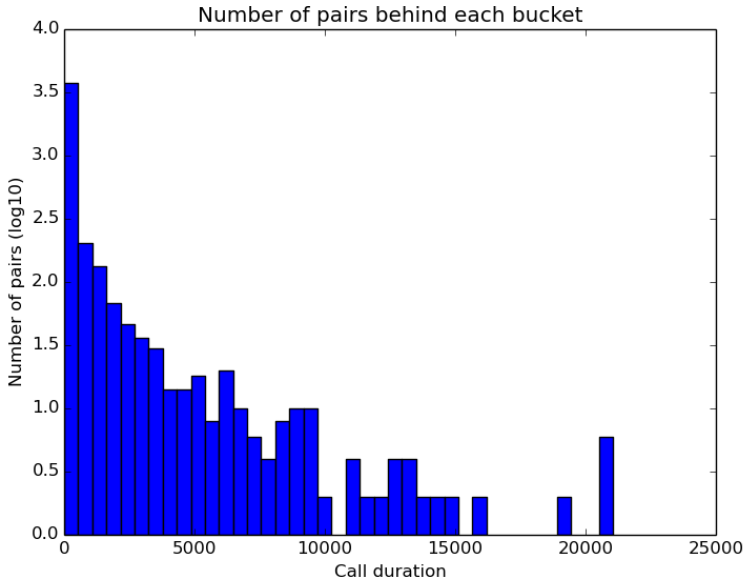


Figure 16. Number of pairs in each bucket in Figure 15

We can see that there appears to be a positive correlation between the number of minutes users spent talking to someone on the phone and the relative number of users' contacts that got infected within a week.

What if we look at call frequency rather than duration?

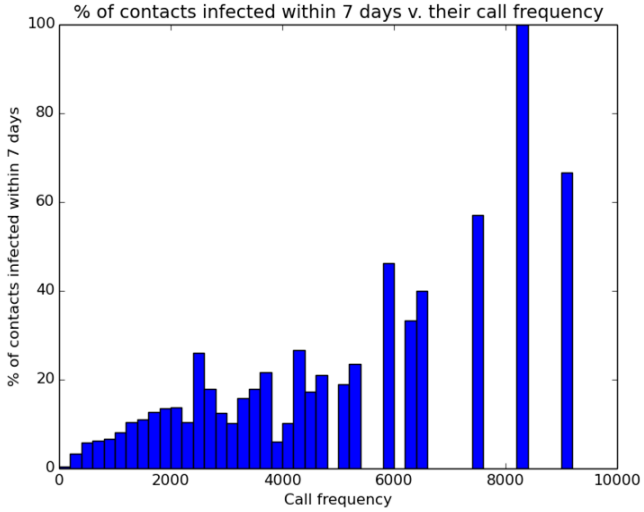


Figure 17. Percentage of Contacts Infected within 7 days vs Call Frequency

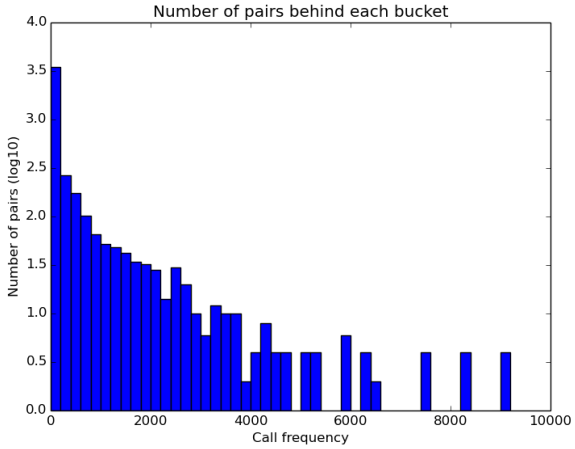


Figure 18. Number of pairs in each bucket in Figure 17

Again, we see a similar trend in Figure 14. A higher percentage of the pairs that talked more frequently on the phone, were diagnosed within a week.

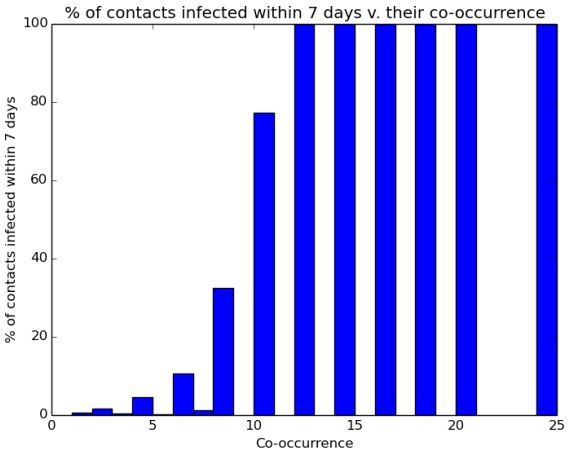


Figure 19. Percentage of Contacts Infected within 7 days vs Co-occurrence

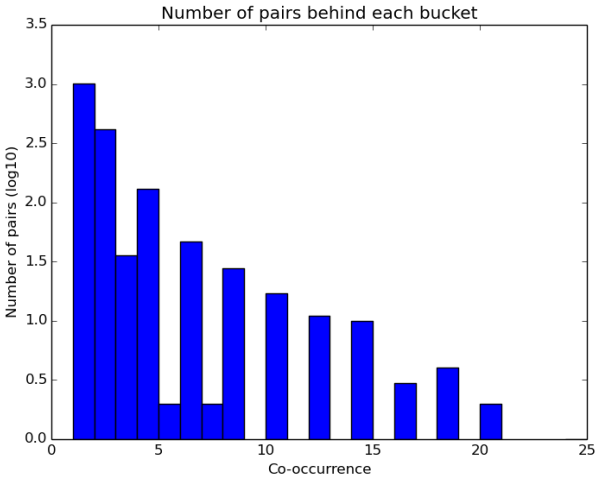


Figure 20. Number of pairs in each bucket of Figure 19

The implication is that, close friends are more likely to be infected closer to one's onset date and call frequency and call duration are both adequate metrics for measuring social strength.

To continue studying social strength's connection to closeness of diagnoses dates, we answer the question, *what percentage of a user's contacts got diagnosed close to the user's onset date?*

To answer this question, we run the following algorithm. We shuffle all infected users and iterate through them. For each infected user *u*, we retrieve *u*'s contacts *F* and select the top five infected contacts *F'* (ranked by call duration). We then calculate a corresponding set of offsets *O'* by taking the absolute value of the difference between the contact's and *u*'s onset dates. Keeping track of how many contacts were infected at each offset and how many were not, after iterating through the set of infected users, we can calculate the percentage of contacts infected at each offset. We rerun the algorithm

to get a control. This is done by replacing each user's top five contacts with five random users. We threshold the maximum offset to 90.

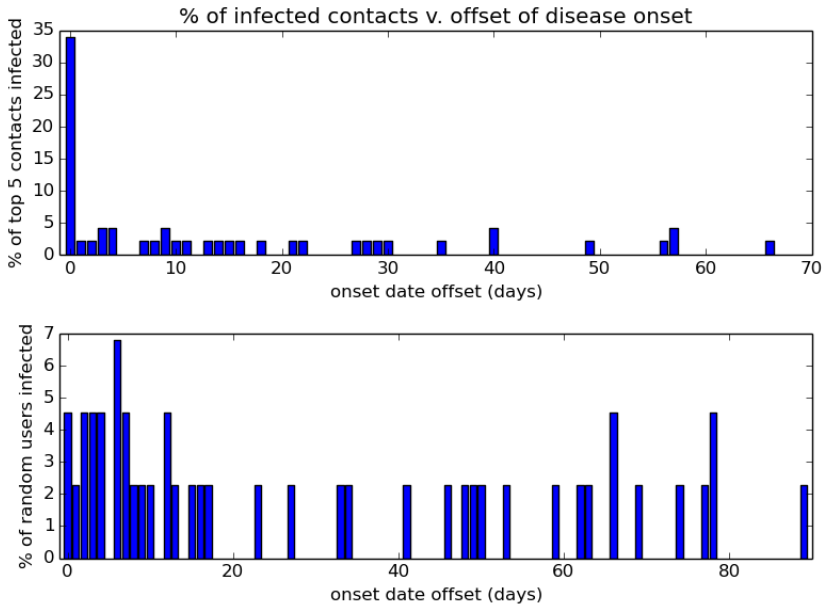


Figure 21 (25 users). *<u>Top subplot:</u> Percentage of Top 5 Contacts Infected vs Offset of Disease Onset;*
<u>Bottom subplot (Control):</u> Percentage of Random Users Infected vs Offset of Disease Onset

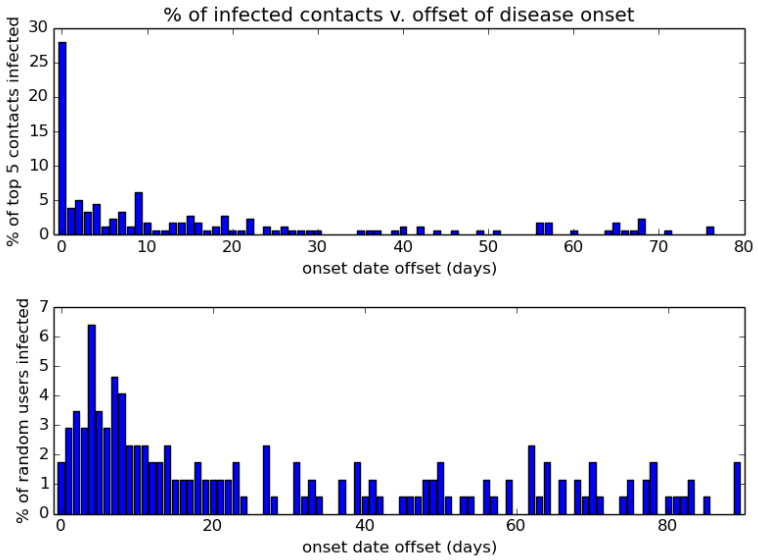
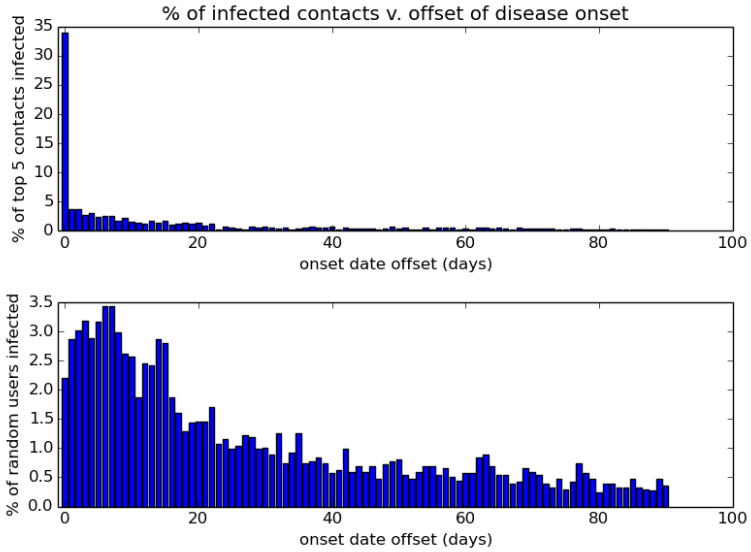


Figure 22 (100 users). <u>Top subplot:</u> Percentage of Top 5 Contacts Infected vs Offset of Disease Onset;
<u>Bottom subplot (Control):</u> Percentage of Random Users Infected vs Offset of Disease Onset



*F*igure 23 (2,125 users). Top subplot: Percentage of Top 5 Contacts Infected vs Offset of Disease Onset*;*
Bottom subplot (Control): Percentage of Random Users Infected vs Offset of Disease Onset

We see a very peculiar trend. In the top subplot, we have a sharp peak at offset 0. In Figure 23, we see that 34% of the users' infected top contacts were diagnosed on the exact same day. While we would expect close friends to be infected near each other, the result shows that they went to the doctor on the exact same day. Notice, the percentage drops severely at offset 1. The offset 0 bucket contains 1,202 users while the offset 1 bucket contains 129. What could explain such a steep decline? Why are there so many users going to the doctor on the same day?

First, we investigate how this peak changes when we consider a different number of top contacts. If we look at the top 20 contacts, will we see the same trend?

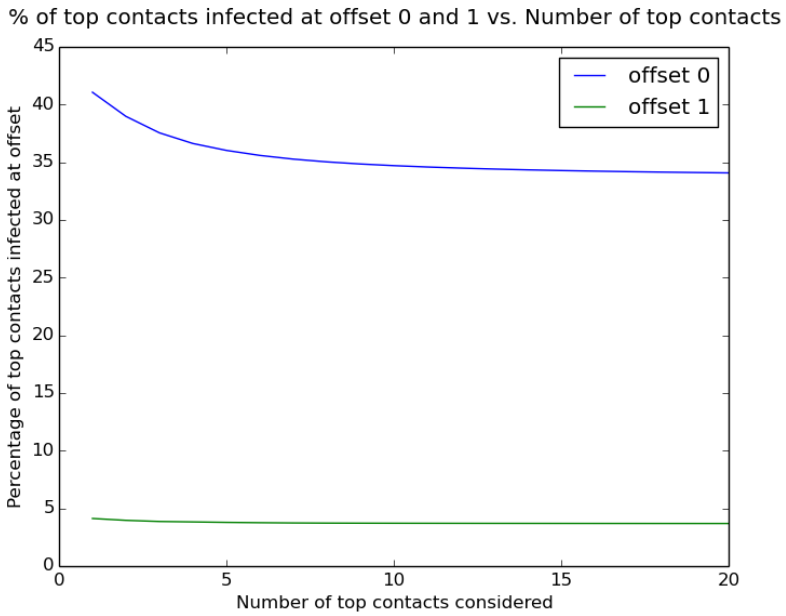


Figure 24. % top contacts infected at offset 0 and 1 vs. Number of top contacts

Figure 24 does show a slight decrease in the peak at offset 0, but ultimately there is a very slight change.

Fortunately, we have access to a smaller dataset that can help us answer this question. This dataset maps UIDs of infected users to family numbers. Specifically it tells us which users belong to the same family. Since family members would be expected to be a user's top contacts, it is possible that there are so many people going to the doctor on the same day, simply because they are a family and it is more convenient for them to visit the doctor together.

Table 7. UID to Family ID relation.

UID 1	Family ID
5030493	15
4990153	15
10816576	246
11340924	246
10553121	246

Data in Table 7 shows that the top two users are part of the same family as are the bottom three.

Table 8. Data describing the Family Dataset.

Family Dataset	
Number of families	754
Average/Variance of number of family members per family	2.366 / 0.590
Maximum number of family members per family	10

First, we do a quick analysis of the family data to see if family members are getting diagnosed on the same day. To do this, we simply loop through every family and check if each family member of a given family has the same disease onset date.

We see that for 90% of all infected families, all family members went to the doctor on the same day. We now rerun the algorithm used for generating Figures 21-23 and check if at offset 0, many of the users belonged to the same family.

We do this by slightly modifying the algorithm to output additional information along the way. For a user *u* and his or her infected contacts *C*, we find how many users in *C* are family members, *F*, of *u*: $M = C \cap F$. The size of the resulting set *M* tells us how many of *u*'s top contacts are part of *u*'s family. We keep track of how many are in the family and how many are not, and calculate the proportion. We find that only 0.187% are part of the same family, contrary to the hypothesis that many of the pairs at offset 0 are part of the same family.

Section 3. Predicting disease spread with social strength

Models are invaluable in studying spread of disease and can be helpful in prevention and control. In this section, we introduce several spatio-temporal discrete event simulators written in Python to attempt to model A(H1N1) spread in Iceland.

Initially the simulators were run on the majority of the CDR traces, but the magnitude of the data proved to be a major issue, as the simulators would not finish the simulations and in fact, took an inordinate amount of time to perform one step. To resolve this, we took a small subset of the data that spanned one year and binned the data into 30 minutes intervals. The modal location during each 30 minute interval was the location that was used in the record. The size of this subset was around 6 million records.

3.1. Seedset

The seedset is the initial set of infected patients that all models use.

Given a starting seed date, the ILI-onset dataset is queried for all people infected on that date, and that result set becomes the seed set that is used by the models. In addition to the UIDs, the algorithm finds and saves each infected user's initial data point. The first model iteration starts at the minimum date of those initial data points. A graphic is included below to better illustrate this.

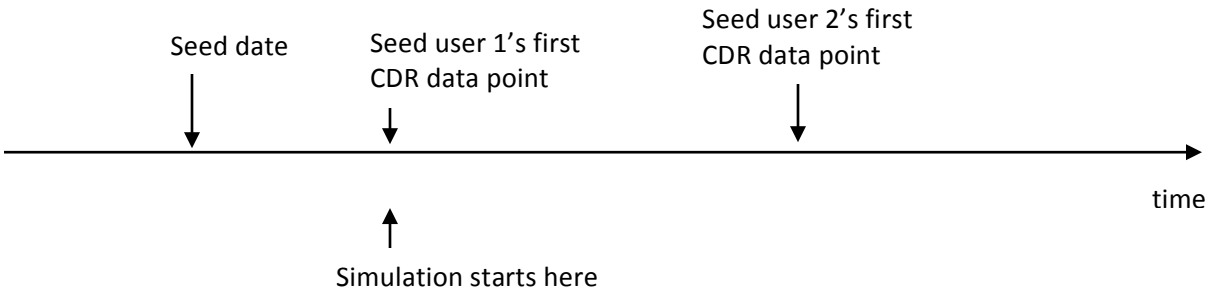


Figure 25. Simulation Start

Initially, the seed date chosen the first onset date in the ILI-onset dataset. However, because the size of the seed set was very small, we chose random seed dates to see if larger seed sets positively impacted the models' predictions.

3.2. Proposed Frameworks

3.2.1. Baseline

The baseline model is the fundamental model that is inherited by the rest. It uses a contact tracing approach to simulate the spread of A(H1N1). An infected individual

whose trajectory intersects with a susceptible individual will deterministically transmit the disease.

This approach requires a clear definition of what it means for two users to intersect. Two users intersect if they are spatio-temporally close to one another. Specifically, we allow a time window (for example, 30 minutes) and a distance window (for example 0.25 kilometers). These are the two major parameters in this model along with the start seed date. Considering cell phone tower density, a window of 0.25 km will generally require two users to be in range of the same cell phone tower.

Indirect transmission includes infection through soil contamination and through touching a contaminated surface. There are many challenges in integrating that information into models [28], and so we do not explicitly define parameters to capture that. The models do however indirectly capture this transmission in part due to the time window. Depending on how long the virus can persist on surfaces, allowing users to become infected even when appearing several minutes after an infected user, allows the model to capture indirect transmission.

In the baseline model, there is only an infected set and an implied susceptible set (the complement of the infected set). One key disadvantage is that individuals can never transition from an infected state to a cured state. Once infected, an individual stays infected and, in theory, the illness spreads until it has been transmitted to the entire population.

This model is not expected to be good at predicting spread, since it ignores all disease information. Though this is not realistic, it should allow us to see the infected set grow until it contains the entire population. It would be interesting if the baseline model failed to simulate this growth, as it may imply a sparsity of intersections in the data.

Overview of Model:

$\theta = \{\varsigma, \Delta t, \Delta d\}$ where

Δt is time a user can be from another to be considered intersecting, Δd is the distance a user can be from another to be considered intersecting, and ς is the seed date.

$$\begin{cases} if\ |u_i.t - u_j.t| \leq \Delta t\ \wedge\ |u_i.c - u_j.c| \leq \Delta d & infect \\ otherwise & do\ not\ infect \end{cases}$$

Algorithm 1 "Move" Contact Tracing Model forward

```
procedure NEXT(infectedSet)
    for each tuple userid, time, coord in infectedSet do
        spatioCloseUsers = Grid[coord]
        for each userid, timestamp in spatioCloseUsers do
            if timestamp + delta_time > time then
                break
            end if
            if inRange(timestamp, delta_time) then
                infectedSet.add((userid, timestamp, coord))
            end if
        end for
        nextDataPoint = getNextDataPoint(userid, time)
        updateUserInfo(userid, nextDataPoint, infectedSet)
    end for
    Return infectedSet
end procedure
```

Algorithm 2 Get a user's next data point

```
procedure GETNEXTDATAPOINT(userid, time)
    userFile = openUserFile(userid)
    for each line in userFile do
        currentTime, currentCoord = extractTimeAndCoord(line)
        if currentTime > time then          ▷ data is sorted by timestamp
            Return currentTime, currentCoord
        end if
    end for
    Return ()
end procedure
```

Figure 26. Baseline Pseudocode

3.2.2. Disease Base Model

Mathematical models are imperative in studying the epidemics of infectious diseases. A classical deterministic model that is widely used in the public health community is the Susceptible-Infected-Recovered (SIR) model. In the SIR model, the population is divided into three classes: Susceptible, Infected, and Removed. The rate of moving between the classes is dependent on the transition rate and the recovery rate.

This model makes improvements on the baseline contact tracing approach by incorporating information about the disease. This model as well as future models incorporate ideas from the popular SIR approach.

In addition to the infected and susceptible sets that were present in the baseline model, we add a recovered set.

Figure 27. Susceptible, Infected, and Recovered set. Components of the SIR model.

Specifically, it uses a parameter β, which represents the duration of the infectious period, meaning how long does it take for an infected individual to transmission between the infected and recovered sets. According to literature, the estimated β of H1N1 is 60 hours [29].

When an individual moves from susceptible to infected, they are given 60 hours until they move to the recovered set. Once an individual has been cured of the disease, he or she cannot contract it again.

In addition to infectious period, the model uses a parameter, α, which represents the number of hours a user was infected for before being diagnosed. Since, the individual

was likely infected several days before being diagnosed by the doctor [30], the model subtracts this parameter from the infectious period to calculate how much time the individual has left before moving to the recovered set.

As with any model, we need to make assumptions to reduce the models' complexity. This model makes several assumptions, which mirror assumptions made in the SIR model. All people are born into susceptible class – no one is inherently immune.

In addition, while individual immune responses certainly vary and are important in disease [31] dynamics, we are not taking into account individual immune response, since it adds unnecessary complexity. Instead, we assume that everyone responds the same to disease and everyone has the same 60 hours infectious period.

Contact with the disease moves users to the infected class, meaning that there is no latency window or no exposed class like in the SEIR (Susceptible-Exposed-Infected-Recovered) model. The individuals in the recovered class cannot be infected again – they are immune for life. The host population is closed. No one is flying in or out of Iceland.

The size of the recovered set is initially zero. The infected set is non-empty – this is the seed set. Everyone else in the population is in the susceptible set.

3.2.3. Disease Model 2

The second disease model incorporates the transmission probability of the disease. Rather than naively passing on a disease whenever users intersect, there is a 26%

chance of transmitting the disease [32]. This accounts for the fact that the definition of intersecting users may be too lenient and a user may have a higher resistance to the disease.

Overview of Disease Models:

$\theta = \{\varsigma, \Delta t, \Delta d, \alpha, \beta, \gamma\}$ where γ is the transmission probability.

$$\begin{cases} if\ |u_i.t - u_j.t| \leq \Delta t \wedge |u_i.c - u_j.c| \leq \Delta d \quad infect\ with\ probability\ \beta \\ otherwise \quad do\ not\ infect \end{cases}$$

3.2.4. Social Network Base Model

The base social network model uses two additional parameters. One, *p1*, is the probability of passing the disease from user "a" to user "b" if a spatio-temporally close user "b" is in the social network of user "a". The other parameter, *p2*, is the probability of transmitting the disease if the two users are close, but user "b" is *not* in the social network of user "a". *p1* and *p2* are parameters in the model and can vary, but from literature [21], we decide to set *p1* to 0.9 and *p2* to 0.1.

Overview of Social Network Model:

$\theta = \{\varsigma, \Delta t, \Delta d, \alpha, \gamma, p1, p2, \omega\}$ where ω is the minimum call duration for two users to count as being in each other's social networks.

$$\begin{cases} if\ |u_i.t - u_j.t| \leq \Delta t\ \wedge\ |u_i.c - u_j.c| \leq \Delta d\ \wedge\ u_i \in SN_{u_j} & infect\ with\ probability\ p1 \\ if\ |u_i.t - u_j.t| \leq \Delta t\ \wedge\ |u_i.c - u_j.c| \leq \Delta d\ \wedge\ u_i \notin SN_{u_j} & infect\ with\ probability\ p2 \\ otherwise & do\ not\ infect \end{cases}$$

3.2.5. Augmented Social Network Model.

This model augments the previous social network models in an attempt to counter the infected set's lack of growth over the course of the simulation. The baseline contact tracing model's inability to infect the entire population (defined by the CDR), means that the future models have no chance of capturing H1N1's growth more accurately, since the models incorporate more and more constraints. This motivates an augmented social network model.

Rather than relying solely on the intersections given by the CDR data and hoping that the data's sparseness does not negatively influence the result, we can augment the model in an effort to spike the size of the infected set and more closely match the epidemic curve.

One of the tricks that is used to spike the growth involves randomly infecting users. One of the consequences of the sparseness of the CDR data is that users' intersections may not be seen in the data. It is fair to assume that even though the data does not show this, people are coming into contact with each other. At each step of the simulation, we retrieve the user whose timestamp comes first in the infected set and look at their top

five contacts (ranked by call duration). Iterating through those contacts, with some small probability, we infect at most one of those contacts.

In addition to this, we construct social strength profiles for every pair of individuals. This involves retrieving the number of co-occurrences, the number of calls, and the duration of the calls between a pair *P*. Then we normalize the values so that they lie between 0 and 1 and weight them by the importance of the parameters, which is determined by the correlations of the parameter vs. offset. By default, the weights are 1/3. In that case that the pair called each other, but never co-occurred the default weights become 1/2. Taking the sum gets us probability *p* that the infected user *u in P* will transmit to the contact *c* in *P*.

Overview of Augmented Social Network Model:

$\theta = \{\varsigma, \Delta t, \Delta d, \alpha, \gamma, p1, p2, \omega, r, P\}$ where r is the probability of infecting a random person at every step of the model, and $P$is a matrix of probabilities, where P_{ij}gives us the probability that $u_i$infects u_j given that they intersect.

Algorithm 1 "Move" Augmented Social Model forward

```
1: procedure NEXT(infectedSet)
2:     userid, timestamp, coord = getFirstDataPointFromInfectedSet()
3:     randInfectedUsers = chooseXRandomUsersFromInfectedSet(X)
4:     for each infectedUser in randInfectedUsers do
5:         for each susceptibleContact of infectedUser do
6:             prob = getIndividualProbabilityOfInfection(infectedUser, susceptibleContact)
7:             infectWithProbability(susceptibleContact, prob)
8:         end for
9:     end for
10:    for each line in dataFile do
11:        u, t, c = extractData(line)
12:        if t < timestamp then
13:            continue
14:        end if
15:        if userid == u or u ∈ infectedSet or u ∈ removedSet then
16:            continue
17:        end if
18:        if intersecting(userid, u, timestamp, t, coords, c, deltaTime, deltaDist)
   then
19:            if inNetwork(userid, u) then
20:                addToInfectedSetWithProbability((u, t, c), p1)
21:            end if
22:            if not inNetwork(userid, u) then
23:                addToInfectedSetWithProbability((u, t, c), p2)
24:            end if
25:        end if
26:    end for
27:    nextDataPoint = getNextDataPoint(userid, time)
28:    updateUserInfo(userid, nextDataPoint, infectedSet)
29:    updateSets()
30: end procedure
```

Figure 28. Augmented Social Model Pseudocode

Algorithm 2 Calculate probability of infection from individual profile

```
1: procedure GETINDIVIDUALPROBABILITYOFINFECTION(userid, contact)
2:     cooccurrence = getCooc(userid, contact)
3:     callDuration = getCallDuration(userid, contact)
4:     callFrequency = getCallFrequency(userid, contact)
5:     prob = normalizeWeightAndSum(cooccurrence, callDuration, callFrequency)
   return prob
6: end procedure
```

Figure 29. GetIndividualProbabilityOfInfection Pseudocode

3.3. Evaluation

To evaluate the models, we generate several graphs to visualize the models' accuracy

in predicting H1N1 spread.

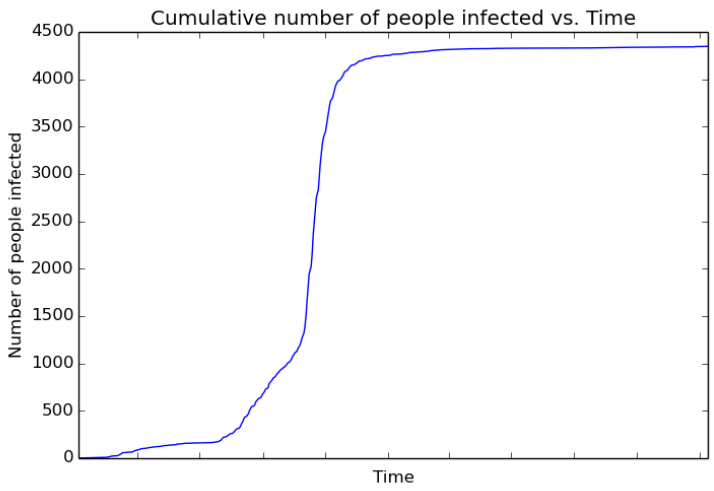


Figure 30. Cumulative number of people infected

The graph shows the amount of people infected throughout the disease period. Note that due to the nature of disease dynamics, 4,500 people will not be infected by the end of the period, some will have transitioned to the recovered set. It is a good starting point for evaluating the efficacy of the disease. At the very least, the baseline contact tracing model, if it works well, should fit nicely on the curve.

More sophisticated evaluation techniques will be required to better evaluate the models.

Precision/recall

We run the simulation for *d* days. After *d* days, we compare the resulting infected set with the ground truth from the ILI-onset-dataset by computing the precision and recall.

Precision: of those in the infected set, how many were actually infected by this time

Recall: of those infected by this time, how many were predicted to be in the infected set

Plot of predicted onset date vs user

Next, we look at every user that was infected and the date on which they were infected. Taking the difference of the predicted onset date and the actual onset date, we can plot how far off the simulators were at predicting the date of onset. A point at $y = 0$ means that the onset date was predicted exactly. A value of $y > 0$ means that the user was predicted to have contracted the disease y days later than he or she actually did. A negative y value means the model predicted disease onset too early.

While this is a better strategy than the previous ones, this says nothing about those individuals that were supposed to be infected but were not. In a sense, it provides an analogue to precision, but fails to represent recall.

3.4. Results and Future Work

The baseline contact tracing model performed poorly. The size of the infected set increased for a few days and then plateaued. The main reason for this is that the sparsity of the data limited the number of intersections the simulators encountered. Though the baseline simulator was not expected to accurately capture the A(H1N1) epidemic curve, the results of its performance are still surprising, because the size of the infected set failed to grow. The precision and recall scores were very high for the baseline, but that is simply because most of the users in the infected set were part of the seed set, which was generated from the ground truth. Because most of the infected

users were part of the seed set, the predicted onset dates were very accurate, though this does not say much.

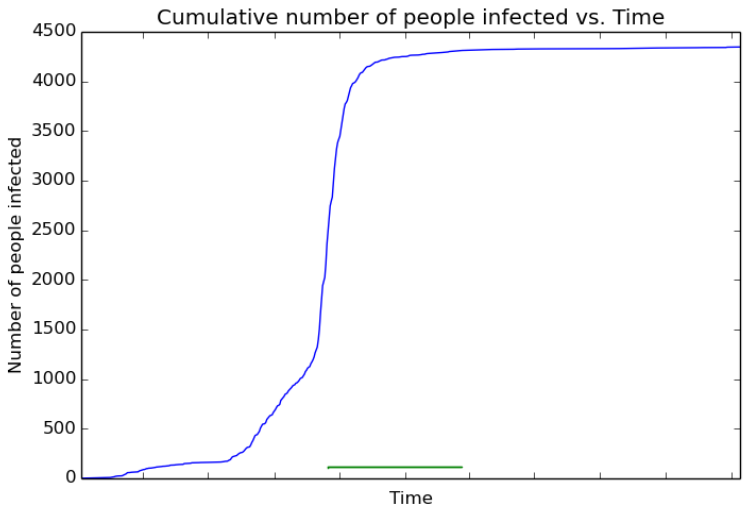


Figure 31. Performance of baseline model with $\theta = \{Oct\ 14\ 2009, 30\ min, 0.25\ km\}$

The disease and social network base models fared poorly as well. With the addition of the *infectious period* parameter, the infected set size dropped down to zero rather than plateauing. Again, precision and recall were high for the same reasons.

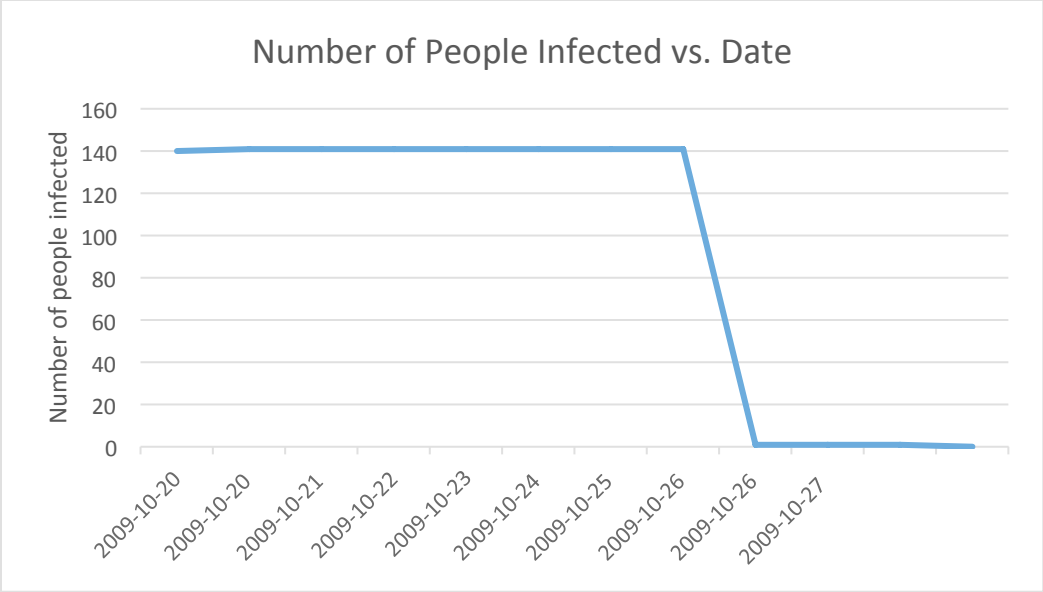


Figure 32. Performance of disease model with $\theta = \{Oct\ 20\ 2009, 30\ min, 0.25\ km, 40\ hours, 26\%, 168\ hours\}$

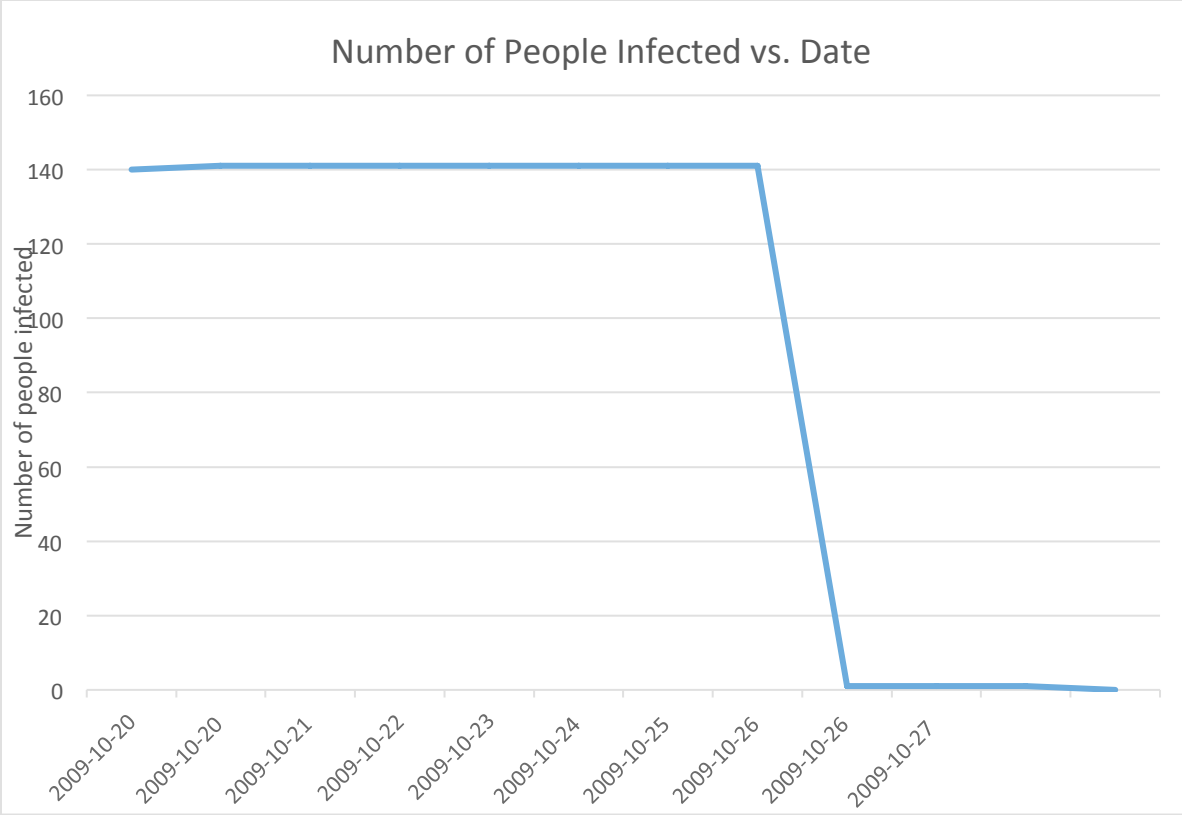


Figure 33. Performance of social network base model with $\theta = \{Oct\ 20\ 2009, 30\ min, 0.25\ km, 40\ hours, 168\ hours, 90\%, 10\%, 1000\ minutes\}$

The augmented social network model failed to capture the A(H1N1) epidemic curve. The size of the infected set did not fall to zero immediately, like in the previous models. Rather it increased briefly showing that the augmentation had a noticeable effect, and then eventually fell to zero.

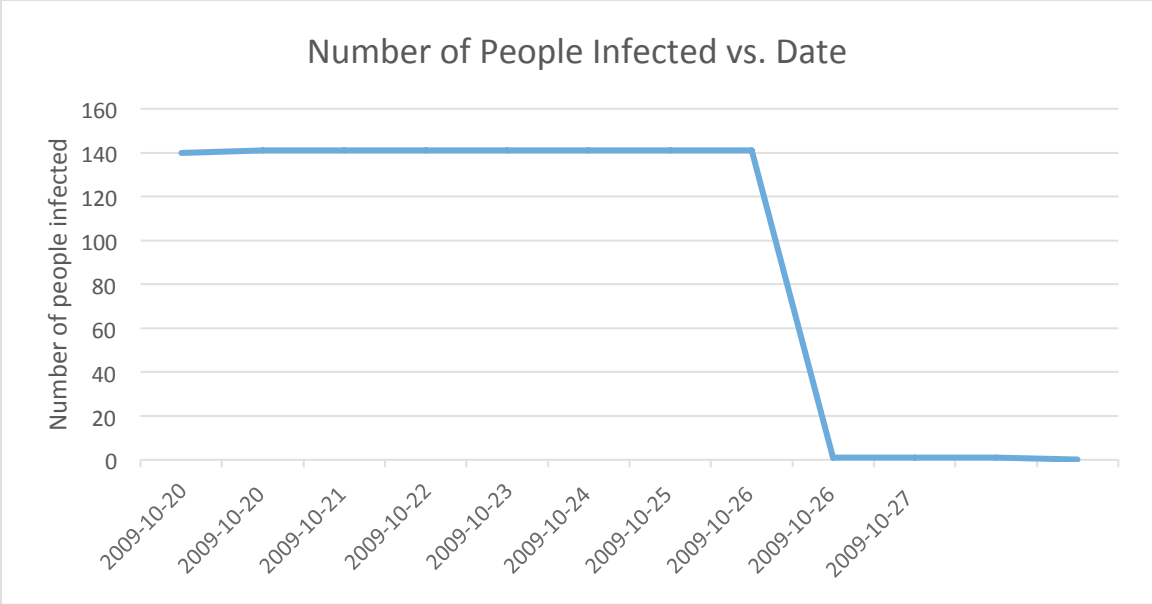


Figure 34. Performance of augmented social network model with $\theta =$

$\{Oct\ 20\ 2009, 30\ min, 0.25\ km, 40\ hours, 168\ hours, 90\%, 10\%, 1000\ minutes, 5\%, P\}$

The major problem the discrete event simulators ran into was the lack of intersections they encountered. Since intersections were the foundation of the simulators, the sparsity of the data limited the robustness of the simulators. Since the simulators ran on only a small subset of the data, running with the complete CDR dataset would likely improve results as more intersections would be captured by the data.

It is probable that a better algorithm would need to be devised for efficiently finding intersecting users. An offline approach involving building an intersection graph should be explored. This should greatly reduce runtime since we would not need to scan through the file for every step of the simulation.

Though we are not using vaccination rate in the models currently, it may be interesting to incorporate average nationwide vaccination rates for influenza-like-illnesses. We could make use of census data, infer home locations of users, which has been done in previous work, map coordinates to zip codes and say X% of these people are vaccinated. Alternatively, we could base it on population density – more people in densely-populated regions will get vaccinated than people in rural regions.

Section 4. Conclusion

The first part of the thesis aimed to show a connection between social strength and temporal closeness of H1N1 diagnosis. We defined three different measures of social strength: call duration, call frequency, and co-occurrence. Call duration and Call frequency are correlated with disease offset. Pairs who make more calls or talk for longer periods of time tend to get sick closer to each other than those that do not. Co-occurrence however is not a reliable measurement of social strength.

Using this information, we built a framework for predicting the spread of infectious diseases. We built several Discrete Event Simulators in Python that leverage CDR coupled with information about the disease and people's social networks in an attempt to track the spread of H1N1 throughout Iceland. The simulators were not successful at accurately tracking the spread of the disease, but after further data cleaning and research effort, we are hopeful that the epidemic curves of A(H1N1) can be modeled accurately. The simulators are a stepping stone. The framework is a step forward in the novel area of using cell phone metadata to model infectious disease dynamics.

References

1. Centers for Disease Control and Prevention. Seasonal Influenza Q&A. Available at http://www.cdc.gov/flu/about/qa/disease.htm. Last Accessed on 3/31/2016.
2. Cruz-Pacheco G, Duran L, Esteva L, Minzoni A, Lopez-Cervantes M, Panayotaros P, Ahued Ortega A, Villasenor Ruiz I. (2009). Modelling of the influenza A(H1N1)v outbreak in Mexico City, April-May 2009, with control sanitary measures. Euro Surveill. 2009 Jul 2;14(26). pii: 19254. PubMed PMID: 19573510.
3. Richter, Felix. "Infographic: 1.17 Billion People Use Google Search." Statista Infographics. N.p., 12 Feb. 2013. Web. 28 Mar. 2016.
4. Lazer, David, and Ryan Kennedy. "What We Can Learn From the Epic Failure of Google Flu Trends." Wired.com. Conde Nast Digital, 1 Oct. 2015. Web. 28 Mar. 2016.
5. Jahani E, Sundsay PR, Bjelland J, Iqbal A, Pentland A and de Montjoye YA. Predicting Gender from Mobile Phone Metadata. Netmob'15 Conference, Massachusetts Institute of Technology, US. April 2015. Book of Abstracts:Oral, page 110. Available online at http://netmob.org/assets/img/netmob15_book_of_abstracts_oral.pdf. Last accessed on March 29, 2016.
6. De Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture*

Notes in Bioinformatics), *7812 LNCS*, 48–55. http://doi.org/10.1007/978-3-642-37210-0_6

7. Calabrese F., Ferrari L., Blondel VD. Urban sensing using mobile phone network data: A survey of research. ACM Computing Surveys (CSUR) 47(2), 25 (2014).

8. Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW. Mobile phones and malaria: modeling human and parasite travel. Travel Med Infect Dis. 2013 Jan-Feb;11(1):15-22. doi: 10.1016/j.tmaid.2012.12.003. Epub 2013 Mar 9. Review. PubMed PMID: 23478045; PubMed Central PMCID: PMC3697114.

9. Cho E, Myers SA, Leskovec J. Friendship and mobility: user movement in location-based social networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, August 21-24, 2011, San Diego, California, USA. Doi 10.1145/2020408.2020579.

10. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature. 2008 Jun 5;453(7196):779-82. doi: 10.1038/nature06958. PubMed PMID: 18528393.

11. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L. *Approaching the limit of predictability in human mobility*. Sci. Rep. 3, 2923 (2013)

12. Belik V, Geisel T, Brockmann D. Natural human mobility patterns and spatial spread of infectious diseases. Physical Review X. 2011; 1(1):011001.

13. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. Quantifying the impact of human mobility on malaria. Science. 2012 Oct

12;338(6104):267-70. doi: 10.1126/science.1223467. PubMed PMID: 23066082; PubMed Central PMCID: PMC3675794.

14. Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ. Commentary: Containing the Ebola Outbreak – the Potential and Challenge of Mobile Network Data. PLOS Currents Outbreaks. 2014 Sep 29. Edition 1. doi: 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.

15. Wesolowski A, Stresman G, Eagle N, Stevenson J, Owaga C, Marube E, Bousema T, Drakeley C, Cox J, Buckee CO. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. Sci Rep. 2014 Jul 14;4:5678. doi: 10.1038/srep05678. PubMed PMID: 25022440.

16. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. (2012) Digital Epidemiology. PLoS Comput Biol 8(7): e1002616. doi:10.1371/journal.pcbi.1002616.

17. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, Rebaudet S, Piarroux R. Using mobile phone data to predict the spatial spread of cholera. Sci Rep. 2015 Mar 9;5:8923. doi: 10.1038/srep08923. PubMed PMID: 25747871; PubMed Central PMCID: PMC4352843.

18. Wesolowski A, Metcalf CJ, Eagle N, Kombich J, Grenfell BT, Bjørnstad ON, Lessler J, Tatem AJ, Buckee CO. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proc Natl Acad Sci U S A. 2015

Sep 1;112(35):11114-9. doi: 10.1073/pnas.1423542112. Epub 2015 Aug 17. PubMed PMID: 26283349; PubMed Central PMCID: PMC4568255.

19. Enrique Frias-Martinez, Graham Williamson and Vanessa Frias-Martinez, "Simulation of Epidemic Spread Using Cell Phone Call Data: H1N1 Case Study", NetMob 2013
20. "Mobile Cellular Subscriptions (per 100 People)." WorldBank. N.p., n.d. Web. 28 Mar. 2016. <http://data.worldbank.org/indicator/IT.CEL.SETS.P2/countries?order=wbapi_data_value_2014%20wbapi_data_value%20wbapi_data_value-last&sort=desc&display=graph>.
21. Frias-Martinez E, Williamson G and Frias-Martinez V. An agent-based model of epidemic spread using human mobility and social network information. *Proc. SocialCom*, pp. 57-64, 2011.
22. Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW. Mobile phones and malaria: modeling human and parasite travel. Travel Med Infect Dis. 2013 Jan-Feb;11(1):15-22. doi: 10.1016/j.tmaid.2012.12.003. Epub 2013 Mar 9. Review. PubMed PMID: 23478045; PubMed Central PMCID: PMC3697114.
23. 2009 flu pandemic by country. Wikipedia. Available at https://en.wikipedia.org/wiki/2009_flu_pandemic_by_country. Last accesses on 3/31/2016.
24. Sigmundsdottir G, Gudnason T, Ólafsson Ö, Baldvinsdóttir GE, Atladottir A, Löve A, Danon L, Briem H. Surveillance of influenza in Iceland during the 2009

pandemic. Euro Surveill. 2010;15(49):pii=19742. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19742.

25. Lau, Lincoln L. H., Benjamin J. Cowling, Vicky J. Fang, Kwok-Hung Chan, Eric H. Y. Lau, Marc Lipsitch, Calvin K. Y. Cheng, Peter M. Houck, Timothy M. Uyeki, J. S. Malik Peiris, and Gabriel M. Leung. "Viral Shedding and Clinical Illness in Naturally Acquired Influenza Virus Infections." The Journal of Infectious Diseases J INFECT DIS 201.10 (2010): 1509-516. Web.

26. "Population, Total." WorldBank. N.p., n.d. Web. <http://data.worldbank.org/indicator/SP.POP.TOTL/countries/IS?display=graph>

27. "Mobile Cellular Subscriptions (per 100 People)." WorldBank. N.p., n.d. Web. 31 Mar. 2016. <http://data.worldbank.org/indicator/IT.CEL.SETS.P2/countries/1W-IS?page=1&display=default>

28. Hollingsworth TD, Pulliam JR, Funk S, Truscott JE, Isham V, Lloyd AL. Seven challenges for modelling indirect transmission: vector-borne diseases, macroparasites and neglected tropical diseases. Epidemics. 2015 Mar;10:16-20. doi: 10.1016/j.epidem.2014.08.007. Epub 2014 Aug 30. PubMed PMID: 25843376; PubMed Central PMCID: PMC4383804.

29. Balcan D, Hu H, Goncalves B, Bajardi P, Poletto C, Ramasco JJ, Paolotti D, Perra N, Tizzoni M, Van den Broeck W, Colizza V, Vespignani A. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. BMC Med. 2009 Sep 10;7:45. doi:

10.1186/1741-7015-7-45. PubMed PMID: 19744314; PubMed Central PMCID: PMC2755471.

30. Ip DK, Lau LL, Chan KH, Fang VJ, Leung GM, Peiris MJ, Cowling BJ. The Dynamic Relationship Between Clinical Symptomatology and Viral Shedding in Naturally Acquired Seasonal and Pandemic Influenza Virus Infections. Clin Infect Dis. 2016 Feb 15;62(4):431-7. doi: 10.1093/cid/civ909. Epub 2015 Oct 30. PubMed PMID: 26518469; PubMed Central PMCID: PMC4725380.

31. Grassly NC, Fraser C. Mathematical models of infectious disease transmission. Nat Rev Microbiol. 2008 Jun;6(6):477-87. doi: 10.1038/nrmicro1845. Review. PubMed PMID: 18533288.

32. Centers for Disease Control and Prevention (CDC). Introduction and transmission of 2009 pandemic influenza A (H1N1) Virus--Kenya, June-July 2009. MMWR Morb Mortal Wkly Rep. 2009 Oct 23;58(41):1143-6. PubMed PMID: 19847148.