Cross-Correlation Networks to Identify and Visualize
Disease Transmission Patterns

Tsung-Chien Lu

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2011

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington
Graduate School

This is to certify that I have examined this copy of a master's thesis by

Tsung-Chien Lu

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final examining
committee have been made.

Committee Members:

____
Neil Abernethy

____
Anne Turner

Date:________________________________

Signature ________________________

Date ____________________________

University of Washington

Abstract

Cross-Correlation Networks to Identify and Visualize Disease Transmission Patterns

Tsung-Chien Lu

Chair of the Supervisory Committee:
Assistant Professor Neil Abernethy
Department of Medical Education and Biomedical Informatics

Influenza-like illness (ILI) has been a major threat to the public health around the world. To inform influenza response by enhancing and supporting disease surveillance, a syndromic surveillance system collects case counts that are aggregated from multiple sources and jurisdictions. Although each jurisdiction has their own planned uses of the data, most systems focus on early detection of the outbreak in regional level response and the algorithms they are using often do not point to a route of transmission. In this work, we seek to develop approaches to aid comparison of data among jurisdictions to improve detection of geographic patterns in disease spread. Using cross-correlation to assess the pairwise similarity between regional case counts, we introduce a cross-correlation network based on ILI activity to reveal potential spatio-temporal patterns in disease transmission. The resulting

networks were plotted and visualized in the map with the R statistical package. To evaluate the feasibility and utility of this approach, we validate these networks against population-level variables influencing the spread of infectious disease, including flight passenger volume, census worker flow, and geographic distance. In our analysis, the spatio-temporal transmission of ILI correlated more closely with state-to-state census worker flows and distance between states than with flight passenger flows. We demonstrate how this visualization motif might enhance existing tools used for the purpose of syndromic surveillance. Finally, limitations of the approach, broader implications for disease surveillance and informatics, and future directions for this research will be discussed.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgement

This thesis will not be possible without the help of the following people. First of all, I would like to thank my advisor Prof. Neil Abernethy who guided me throughout the two-year period of my learning process in the institute. With his trust and support, I can finally concentrate my idea and work-out the thesis finally. I would also like to thank Dr. Anne Turner for her support and precious comments. And with cordiality, I dedicate this thesis to my wife Jia-Yin and my daughter Joyce, for their continuous encouragement, understanding and patience.

CHAPTER 1 INTRODUCTION

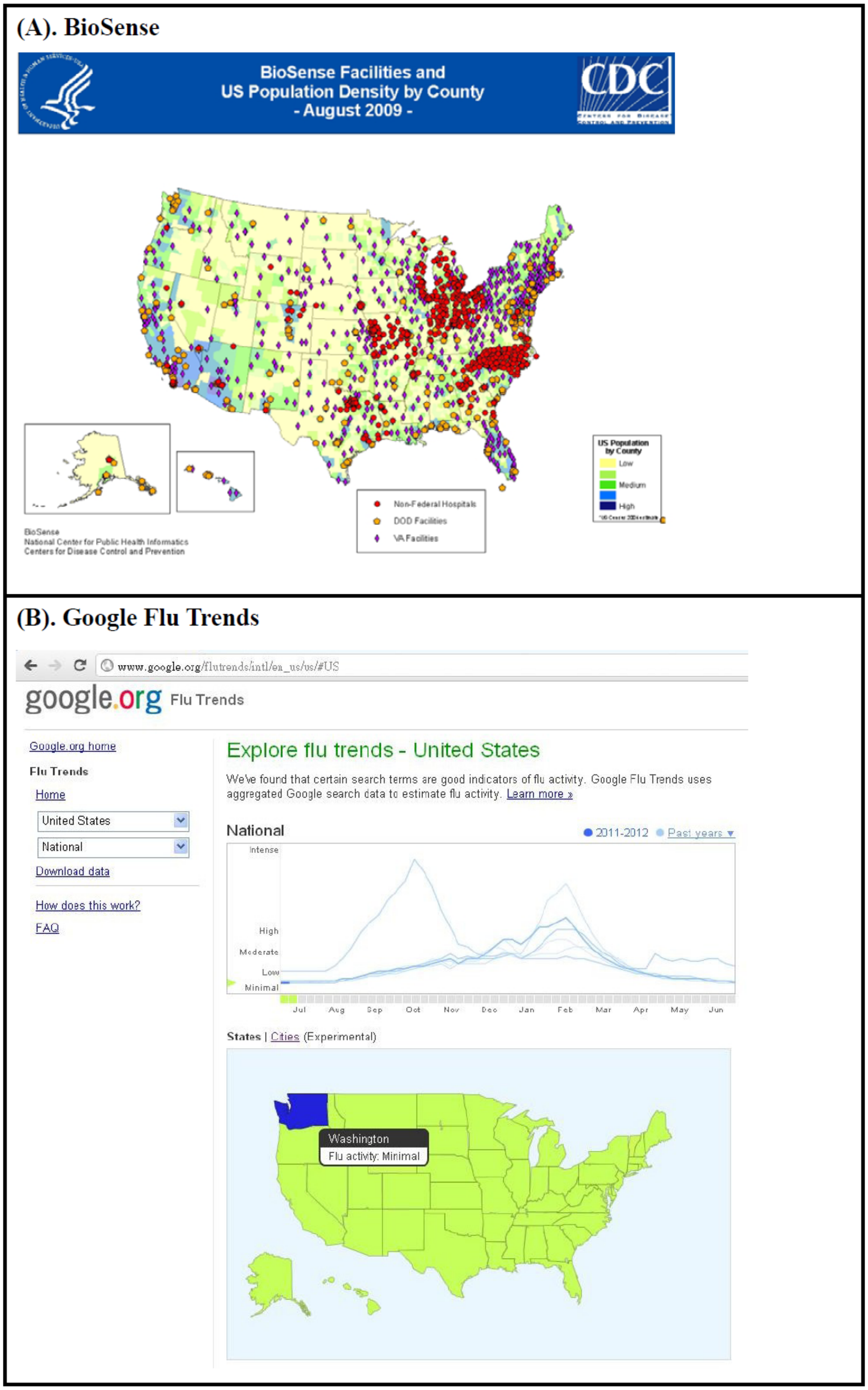
Syndromic surveillance is a novel epidemiologic method which relies on detection of clinical case features that are discernable before diagnoses are confirmed (Mandl, 2004). Unlike traditional surveillance systems such as notifiable disease reporting system that are based on clinical or laboratory diagnosis, syndromic surveillance is an aggregate surveillance system of a disease or health event by collecting summary data on group of cases (Porta, 2008). Given that syndromic surveillance systems are characterized by flexibility and simplicity since they collect summarized counts from existing health data, such surveillance system has emerged as a prospective method to inform public health response in different disease syndromes. Potential data sources include surrogate data sources (e.g., over-the-counter prescription sales or school absenteeism), nurse's hotlines, outpatient visit data, Emergency Department (ED) data (Henning, 2004), and web search queries (Eysenbach, 2006).

With the increasing convenience of transportation and traveling around the world, an outbreak of a novel infectious disease can impact people living around the globe. One of the major tasks of the Public Health Informaticians and epidemiologists is discovering what diseases will threat the public as early as possible and understanding their spread through the population, no matter whether a disease is introduced naturally, accidentally, or intentionally (Franz, 2009). In an era of pandemic-causing strains of influenza or Severe Acute Respiratory Syndrome (SARS), infectious disease surveillance is critical to early detection and response. Since many different bioterrorist agents present with flu-like symptoms, most syndromic surveillance systems incorporate influenza-like illness (ILI) as one of the major syndromes to be monitored (Mandl, 2004).

Currently the major systems that use syndromic surveillance to monitor ILI activity as part or a whole of a comprehensive surveillance program in the U.S.A. include Influenza-like Illness Surveillance Program (ILINet, 2011), Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) (Lombardo, 2003), the CDC BioSense Program (Bradley, 2005) (Figure 1A), and Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking and Evaluation Project (DiSTRIBuTE, 2011). In 2008, Google launched Google Flu Trends as a syndromic surveillance system based on flu-related searches, to supplement the deficiency of traditional surveillance system that generally focus only on local trends (Google Flu Trends, 2011) (Figure 1B). By counting the frequency of search queries, Google claimed that they can estimate how much flu is circulating in different countries and regions around the world. Although over the years there are so many novel syndromic surveillance systems have been put into practice, with varying degrees of success in terms of timely detection and early response to the outbreak (Doornik, 2010; Buehler, 2009; Overhage, 2008), there have been no research reports describing ILI spread across regions using these particular surveillance data sets.

During the 2009 H1N1 flu pandemic, the public was facing the threat of contracting the so called “Swine Flu” and serious complications which might occur among previously healthy individuals during the outbreak. To enhance and support disease surveillance, influenza surveillance data are aggregated from multiple sources and jurisdictions. Although each jurisdiction has their own planned uses of the data collected for their system, especially focusing on early detection of the outbreak in regional level response, there is a lack of knowledge of how diseases spread across different region or jurisdiction. In addition, most surveillance systems focus on early detection of the outbreak in regional level response. The algorithms they use rarely point to a route of transmission. Because understanding the route of transmission is the key to better adopt optimal intervention and control strategies before disease spread, we are interested in making effective use of disease aggregated data and developing a visualization tool that can point to a route of disease spreading across regions. By using this kind of tool, public health practitioners can identify and visualize spatio-temporal patterns in those aggregate data and also help them to communicate with each other. In addition this might assist in early intervention and control measures before disease spreads from one region to another.

In this study, we investigated the feasibility of network visualization to highlight geographic patterns by cross-correlation analysis on data sets with ILI counts collected for influenza surveillance during the period of 2009 H1N1 pandemic. The primary aim is to construct a cross-correlation network of ILI activity and explore how disease is transmitted in a spatio-temporal manner. The secondary aim is to validate this method against population-level variables influencing the spread of infectious disease, including flight passenger volume, census worker flow, and geographic distance.



(A). BioSense

(B). Google Flu Trends

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Figure 1. Screen snapshots showing (A). BioSense and (B).Google Flu Trends.

CHAPTER 2 LITERATURE REVIEW AND BACKGROUND

2.1. Syndromic Surveillance and ILI

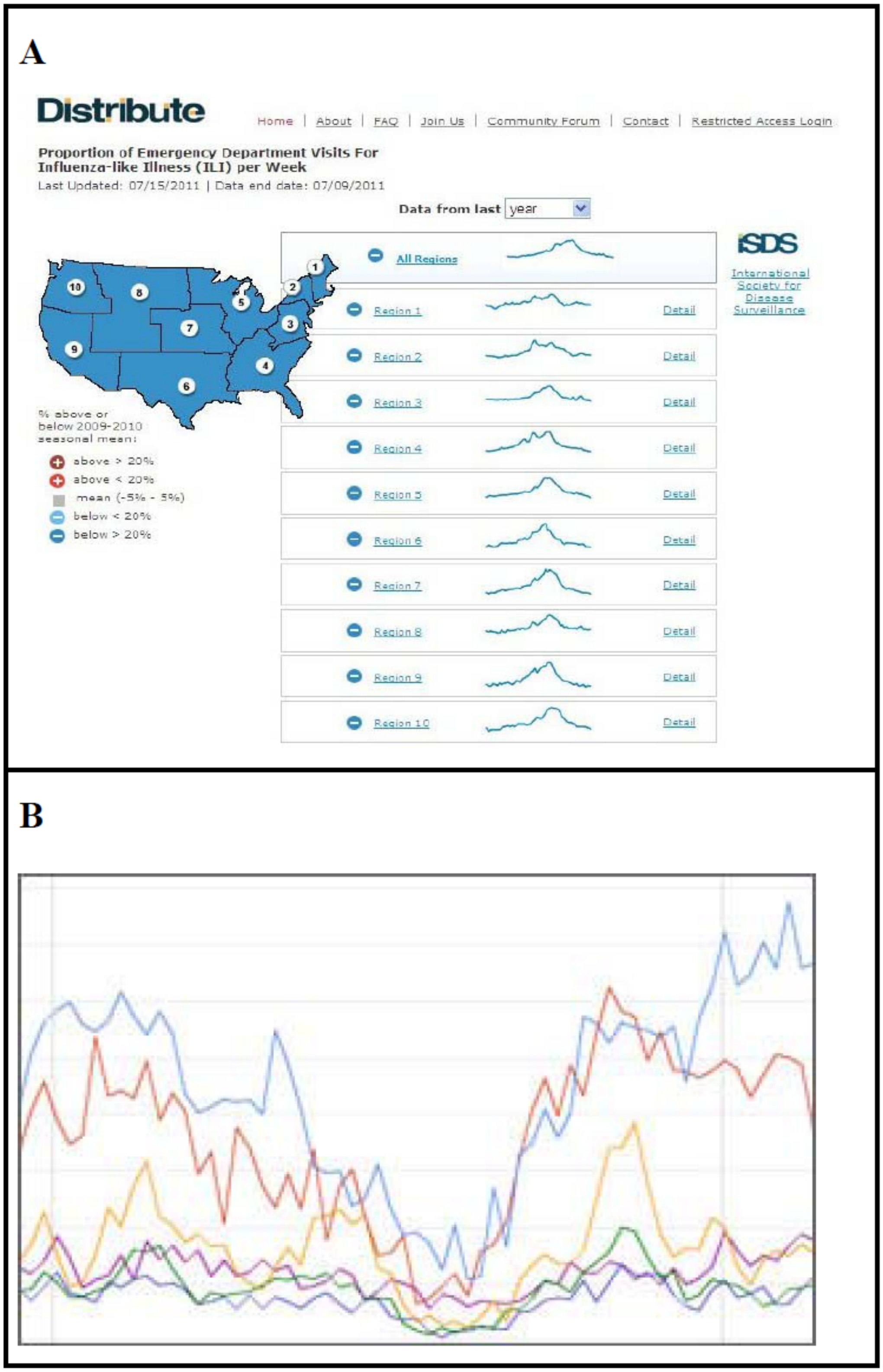
Syndromic surveillance was first developed for bioterrorism preparedness and outbreak detection with the goal of expanding and improving upon traditional public health surveillance (Henning, 2004). Although the choice of data sources could affect the ability of a syndromic surveillance system to detect outbreaks earlier than conventional surveillance methods, the most valuable data sources will be those that are electronically stored, allow robust syndromic grouping, and are available in a timely fashion (Mandl, 2004).

As the key component of the syndromic surveillance system, data from EDs have shown promising research results with adequate balance between sensitivity and specificity for outbreak detection. Specific data elements related to the ED that have been used in syndromic surveillance systems include patient's chief complaint (in either free text or structured format), ED discharge diagnostic code (International Classification of Diseases, ICD), ambulance dispatch notes, and telephone triage service (Beitel, 2004; Reis, 2004; Fleischauer, 2004; Brownstein, 2005; Lu, 2008; Lemay, 2008; South, 2008; May, 2010; Greenko, 2003; Yih, 2009). For ILI syndrome, researchers have demonstrated that ED data can be used to detect disease outbreaks 1-2 weeks earlier than through conventional disease reporting methods (Tsui, 2001; Teich, 2002). Furthermore, a study conducted by Bellazzini *et al* showed that data collection from ED electronic medical records (EMR) by using chief complaint and ICD-9 diagnostic code could detect unexpected ILI before laboratory confirmation, and hence can serve as an adjunct to traditional laboratory-guided public health alerts (Bellazzini, 2011).

In addition to the above mentioned systems that used data sources mainly from clinical domain, researchers sought to develop a more robust and real time monitoring system that could capture user behavior in web search for health-related information. Since 2008, Google launched Google Flu Trends aimed at monitoring health-seeking behavior in the form of queries to online search engine - Google. They developed a system of analyzing large numbers of Google search queries to tract ILI activity in different geographic populations. By counting how often people search for flu-related topics reflecting on aggregate search queries, the system estimates flu circulation in different countries and regions around the world. The results, which appeared in Nature, showed that Google web search queries can be used to estimate ILI percentages accurately in different regions of the United States. The resulting ILI estimates were consistently 1–2 weeks ahead of CDC ILI surveillance reports (Ginsberg, 2009). With the popularity of social networking and microblogging service, researchers are now seeking to tract influenza activities by analyzing Twitter messages and the results are promising (Signorini, 2011).

2.2. The DiSTRIBuTE system as an Example of How Syndromic Surveillance Works

With the goal of improving surveillance and informing influenza response for a more timely response and investigation, the International Society for Disease Surveillance (ISDS) is currently working in partnership with the Centers for Disease Control (CDC) and other public health organizations to launch the DiSTRIBuTE project, which collects ED-related ILI counts from participating local and state health departments. The DiSTRIBuTE system offers publicly accessible visualization graphs of the flu trends in the prior 4 weeks in terms of different HHS regions over their official website (DiSTRIBuTE, 2011). The system currently covers 22 states and 11 cities. The system also provides the public with additional graphs, including the daily time series graphs, weekly time series graphs, and age-group surface plotting that depict relative increases/decreases in ED ILI syndrome visits as observation/baseline by jurisdiction and age (Figure 2A). In addition to the publicly accessible website, the restricted site is available to data contributors of the Distribute system. It contains descriptions of some of the characteristics of each data contributor and also some interactive visualization features that allow comparison queries of data and data timeliness (Figure 2B). Although such a web-based visualization system could serve as a valuable information resource for the surveillance community to explore further research and practice topics, users may be confronted with comparisons of data across different jurisdictions and the existing tools may not help users quantify and assess the relevance of similarities/differences between case count signals. The information flow of the DiSTRIBuTE system is depicted as Figure 3.



A
Distribute
Home | About | FAQ | Join Us | Community Forum | Contact | Restricted Access Login

Proportion of Emergency Department Visits For Influenza-like Illness (ILI) per Week
Last Updated: 07/15/2011 | Data end date: 07/09/2011

Data from last year

% above or below 2009-2010 seasonal mean:

- ● above > 20%
- ● above < 20%
- ■ mean (-5% - 5%)
- ● below < 20%
- ● below > 20%

Figure 2. Snapshots of Distribute project showing visualization features of (A). Public Distribute site and (B). Example of a Distribute Restricted Chart.

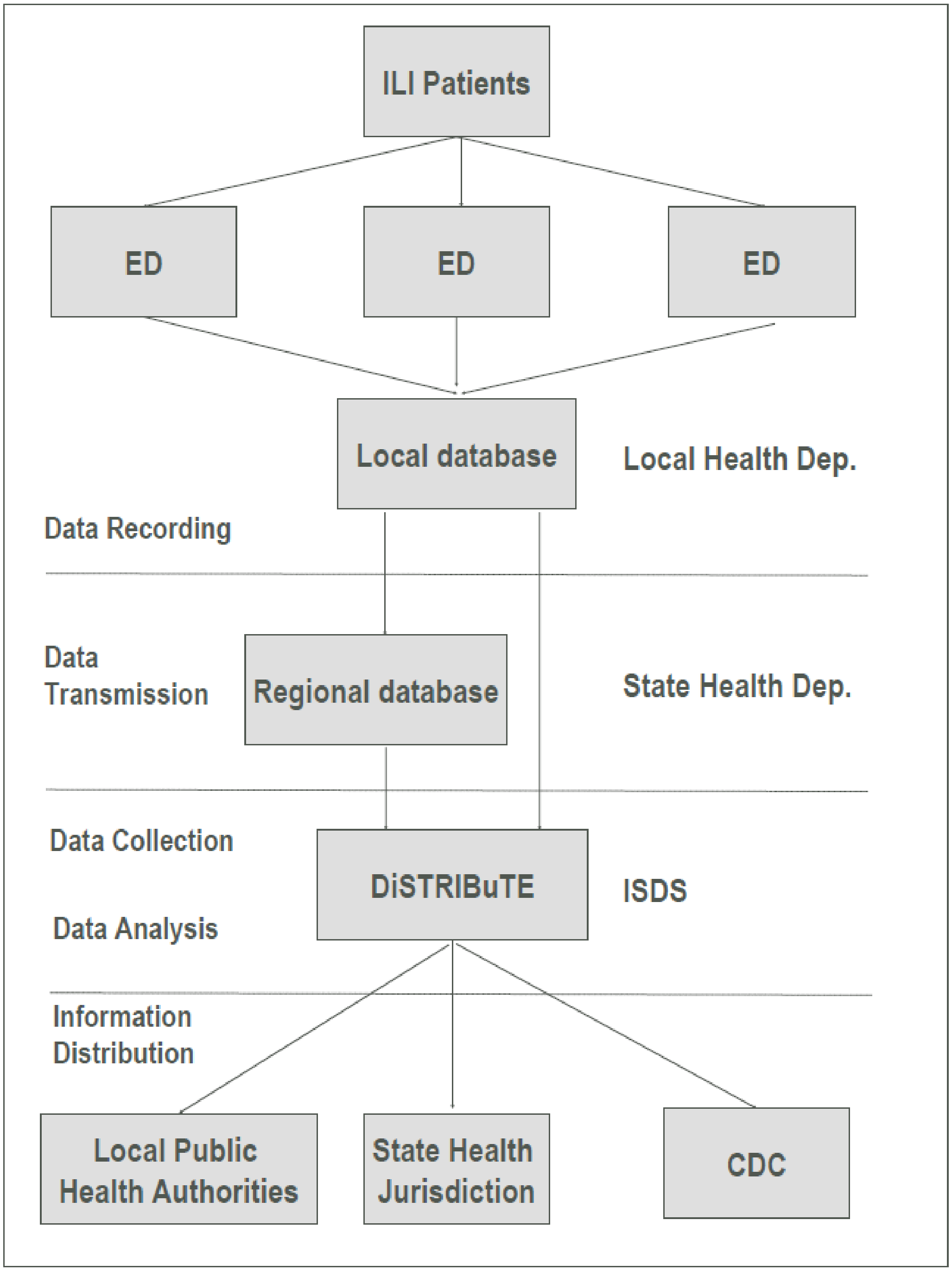


Figure 3. Information flow of the DiSTRIBuTE system.

2.3. Cross-Correlation Analysis as Timely Evaluation for Outbreak Detection

Cross-correlation is a statistics analysis that measures the correlation between two time series and is a commonly used similarity measurement of two waveforms in signal processing (Shumway, 2000) The analysis helps identify one series which is leading indicator of other series or how much one series is predicted to change in relation the other series. The cross-correlation analysis of two time series datasets involves repeated measurements of the Pearson correlation coefficient *r* by time-shifting the one dataset relative to the other dataset. Each shift is called a "lag", and the lag time is simply the time unit of the sampling period in collecting the two time-series datasets (SCRC web site, 2011). The typical cross-correlation graph, which is called "correlogram", shows enough lags in both negative and positive directions to represent the cyclical relationship of the two sets of data (Figure 4). A negative time lag implies the first series in the pair occurred first in R Statistics Package.

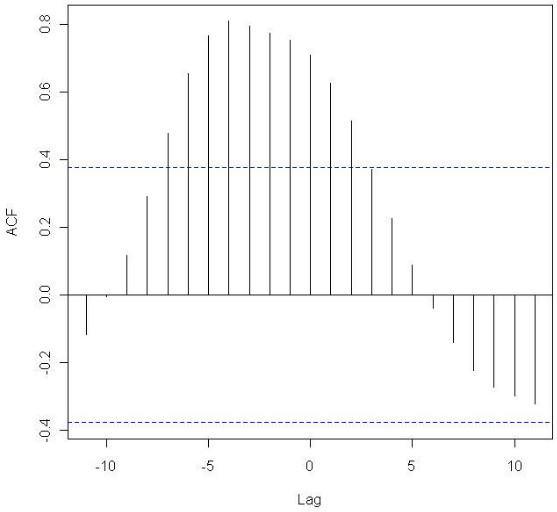


Figure 4. Example correlogram showing the correlation as a function of time lags.

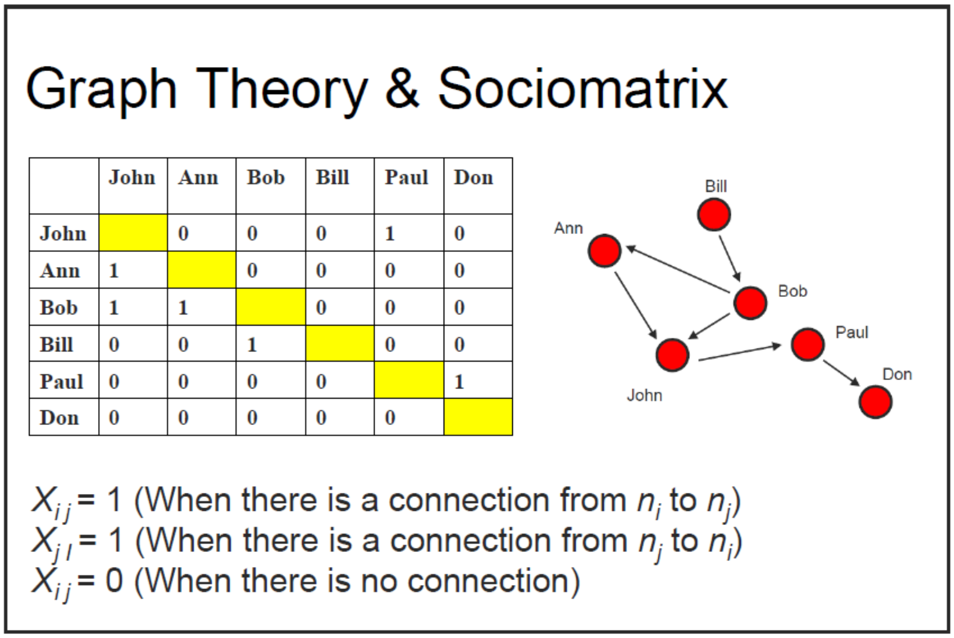
Timeliness is generally defined as the difference between the time an event occurs and the time the reference standard for that event occurs (Dailey, 2007). Time series cross-correlation analysis of syndromic surveillance signals has been applied to assess timeliness as one of the performance measures in the literature. Tsui et al performed cross-correlation analysis to evaluate the timeliness for early detection of epidemics and found that using ICD-9-coded chief complaint was one week earlier than using data from Pneumonia and Influenza deaths (Tsui, 2002). Espino *et al* compared the timeliness of ED telephone triage (TT) data with influenza data from the CDC using cross-correlation analysis, and the results showed that ED TT calls were one to five weeks ahead of CDC surveillance data (Espino, 2003). Lemay and colleagues compared four age groups and six ILI symptoms captured by CC by cross-correlation analysis using reference signal from laboratory-confirmed influenza cases, and found that children younger than 5 years consulting ED mainly for fever and for respiratory symptoms peaked 1 to 4 weeks before the isolation of influenza virus in the community (Lemay, 2008).

A study conducted by Doroshenko *et al* evaluated a syndromic surveillance that captured data from National Health Service (NHS) Direct, a national telephone health advice service in the UK, for surveillance of 10 syndromes commonly occurring in the community (Doroshenko, 2005). Using cross-correlation analysis, they found that an increase in consultations for ILIs recorded by the Royal College of General Practitioners Weekly Returns Service (WRS) is preceded by the increase in calls to NHS Direct for ILI by 1--3 weeks, where WRS is a well-established national clinical surveillance system. In Australia, Zheng and colleagues found that monitoring time series of ED visits clinically diagnosed with influenza could potentially provide three days earlier warning compared with surveillance of laboratory-confirmed influenza by cross-correlation analysis (Zheng, 2007).

To construct a cross-correlation network of ILI activity spreading across regions, this thesis will rely on publicly accessible and well-validated datasets from U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) and Google Flu Trends. Before going further, I would like to review graph theory and sociomatrices of relevant to the construction of cross-correlation network.

2.4. Graph theory and social networks analysis

Graphical models are a marriage between probability theory and graphic theory for solving problems of uncertainty and complexity that occur throughout applied mathematics and engineering (Murphy, 2001). The purpose of graphical modeling is to exploit the statistical relationships of the entities being modeled for representational and computational efficiency (Gimpel, 2006). A graph consists of points called nodes (or vertices) and lines called edges (or arcs) connecting two vertices. An edge is directed if it runs in only one direction and undirected if it runs in both directions. Directed edges can be thought of as sporting arrows indicating their orientation. A graph is directed if all of its edges are directed. An undirected graph can be represented by a directed one having two edges between each pair of connected vertices, one in each direction (Newman, 2003). In sociology, a sociometry is a quantitative method for measuring social relationships and a sociogram is a systematic method for graphically representing individuals as points/nodes and the relationships between them as lines/arcs. In public health, the graph theory has been widely used in analysis of social networks and contact investigation (Abernethy, 2005; Cauchemez, 2011). An example social networks analysis using graph theory with the resulting sociogram is depicted as Figure 5.



Graph Theory & Sociomatrix

	John	Ann	Bob	Bill	Paul	Don
John	🟨	0	0	0	1	0
Ann	1	🟨	0	0	0	0
Bob	1	1	🟨	0	0	0
Bill	0	0	1	🟨	0	0
Paul	0	0	0	0	🟨	1
Don	0	0	0	0	0	🟨

X_{ij} = 1 (When there is a connection from n_i to n_j)
X_{jl} = 1 (When there is a connection from n_j to n_i)
X_{ij} = 0 (When there is no connection)

Figure 5. Sociomatrix (left) and the corresponding directed graph showing relationship with connections between individuals (right).

CHAPTER 3 DATA SETS

To provide task-specific context for this thesis, this chapter describes the data sets used for the analysis and evaluation. Given the reasons that ED sources might be more sensitive and also reliable data sources for surveillance purposes, data sets such as the DiSTRIBuTE system may provide a more reliable source for routine analysis. However, due to issues related to data accessibility and jurisdiction, I decided to use the publicly accessible and well-validated data sets, the ILINet and Google Flu Trends, for analysis. To evaluate the feasibility and utility of this approach, we will validate these networks against population-level variables influencing the spread of infectious disease, including flight passenger volume, census worker flow, and geographic distance.

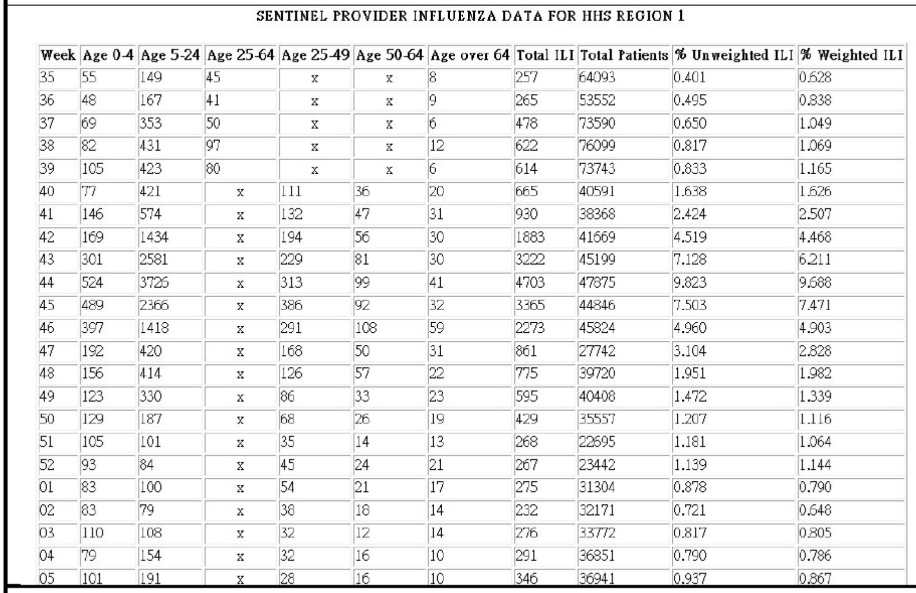
3.1. Data Sets for Analysis

We analyzed weekly patient visits to health care providers for ILI collected through the US Outpatient Influenza-like Illness Surveillance Network (ILINet) and also weekly query counts from flu-related searches collected by Google flu trends. Both data sets are publicly available through the CDC or Google website.

3.1.1. ILINet

Previously known as a sentinel influenza surveillance provider, ILInet is the U.S. Outpatient Influenza-like Illness Surveillance Network conducted by CDC in collaboration with health care providers around the United States (ILINet, 2011). It consists of more than 3,000 healthcare providers and approximately 1,800 outpatient care sites in all 50 states and the District of Columbia (DC), ILINet provides a nationwide picture of influenza virus and ILI activity. Each week, ILINet providers report the total number of patient visits and the total number of patient visits for ILI by age group (0-4 years, 5-24 years, 25-49 years, 50-64 years, and ≥ 65 years). For this system, ILI is defined as fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat in the absence of a *known* cause other than influenza. Sites with electronic records use an equivalent definition as determined by state public health authorities, meaning that slightly different definitions could sometimes result in different syndrome detection rates. The ILINet data sets can be accessed online via CDC seasonal influenza website (ILINet, 2011), allowing routine use of this data for future analyses. A sample table and flu trends of influenza weekly reports from CDC for HHS region-1 is shown as Figure 6.

(A). Snapshot of ILINet chart view for HHS Region 1



SENTINEL PROVIDER INFLUENZA DATA FOR HHS REGION 1

Week	Age 0-4	Age 5-24	Age 25-64	Age 25-49	Age 50-64	Age over 64	Total ILI	Total Patients	% Unweighted ILI	% Weighted ILI
35	55	149	45	x	x	8	257	64093	0.401	0.628
36	48	167	41	x	x	9	265	53552	0.495	0.838
37	69	353	50	x	x	6	478	73590	0.650	1.049
38	82	431	97	x	x	12	622	76099	0.817	1.069
39	105	423	80	x	x	6	614	73743	0.833	1.165
40	77	421	x	111	36	20	665	40591	1.638	1.626
41	146	574	x	132	47	31	930	38368	2.424	2.507
42	169	1434	x	194	56	30	1883	41669	4.519	4.468
43	301	2581	x	229	81	30	3222	45199	7.128	6.211
44	524	3726	x	313	99	41	4703	47875	9.823	9.688
45	489	2366	x	386	92	32	3365	44846	7.503	7.471
46	397	1418	x	291	108	59	2273	45824	4.960	4.903
47	192	420	x	168	50	31	861	27742	3.104	2.828
48	156	414	x	126	57	22	775	39720	1.951	1.982
49	123	330	x	86	33	23	595	40408	1.472	1.339
50	129	187	x	68	26	19	429	35557	1.207	1.116
51	105	101	x	35	14	13	268	22695	1.181	1.064
52	93	84	x	45	24	21	267	23442	1.139	1.144
01	83	100	x	54	21	17	275	31304	0.878	0.790
02	83	79	x	38	18	14	232	32171	0.721	0.648
03	110	108	x	32	12	14	276	33772	0.817	0.805
04	79	154	x	32	16	10	291	36851	0.790	0.786
05	101	191	x	28	16	10	346	36941	0.937	0.867

(B). Snapshot of ILINet flu trends for HHS Region 1

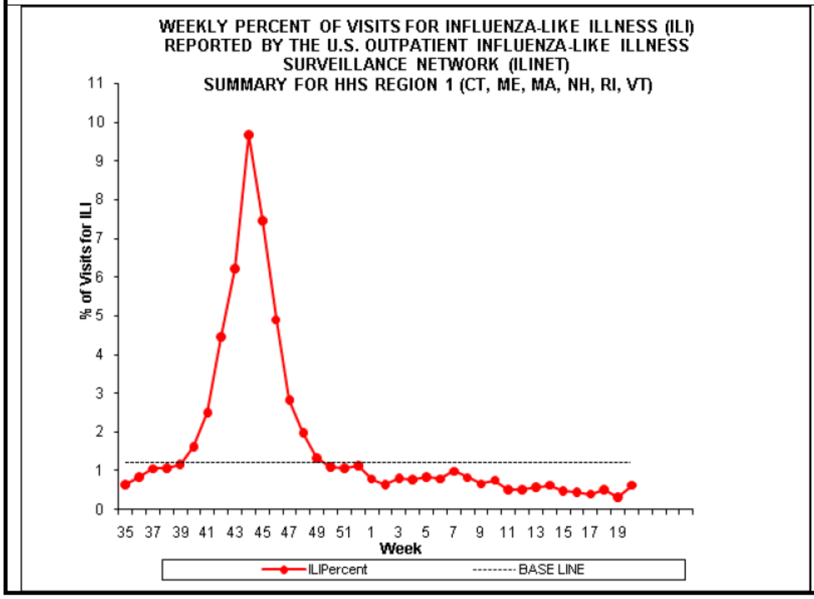


Figure 6. Snapshots of (A).sample table and (B).flu trends of influenza weekly reports from CDC for HHS region-1.

3.1.2. Google Flu Trends

Google Flu Trends is a system conducted by Google Inc., an American multinational public corporation invested in Internet search, cloud computing, and advertising technologies. Google Flu Trends tracks flu-related search queries that are thought to be correlated with the spreading of influenza virus. This information is collected and used to estimate flu activity in the United States and around the world. By counting how often people use the search queries and reporting query counts on a weekly basis, Google Flu Trends has been found to able to estimate flu activity as compared to that of the traditional flu surveillance systems (Ginsberg, 2009). The Google Flu Trends data sets can be accessed via their official website (Google Flu Trends, 2011).

3.2. Data Sets for Evaluation

For the evaluation purposes, I utilized those publicly accessible data sets collected by the US Census Bureau or Bureau of Transportation Statistics (BTA), such as flight passenger volume, census worker flow, and distance between centroids. Since influenza (and other communicable diseases) can spread by social contact, likely including infected surfaces and droplets, we take flight passenger volume and census worker flow as indicatives of actual contact, and distance as a proxy for likelihood of contact.

3.2.1. Airline Origin and Destination Survey (DB1B)

The Airline Origin and Destination Survey (DB1B) is a 10% sample of airline tickets from reporting carriers collected by the Office of Airline Information of the Bureau of Transportation Statistics. Data includes origin, destination and other itinerary details such as number of passengers transported. This annually updated database can be accessed via the website of Research and Innovative technology Administration (RITA), BTA website (DB1B, 2010).

3.2.2. Census 2000 Worker Flow Files

The Census 2000 worker flow data sets consisting of the data that are reported as total number of workers commuting between counties of residence and counties of work for residents of the 50 states and the DC. It is updated every ten years and the current available data sets are year 2000. For the purpose of this study, state-to-state file format can be selected and accessed via the Census Bureau Home Page (Census Worker Flow, 2000).

3.2.3. Census 2010 Centers of Population by State

A population centroid is the center point of the region's population that describes the mean center or the center of gravity of population in a given geographic area (Thaper, 1999). The Census Centers of Population by State files are the data sets updated by the US Census Bureau every ten years and can be used to calculate the centroid distance between states. The 2010 Census Centers of Population by State is the most updated file and can be accessed via the website of the US Census Bureau (Census Centers of Population by State, 2010).

CHAPTER 4 ANALYSIS METHODS

4.1. Cross-Correlation Analysis

Cross-correlation is a test that measures the similarity between two signals or waveforms as a function of a time lag. For discrete time series signals, cross-correlation between two signals *x*[*n*] and *y*[*n*] is calculated as

$$\boxed{r_{xy}[l] \overset{\triangle}{=} \sum_{n=-\infty}^{\infty} x[n]\, y^*[n-l] = \sum_{n=-\infty}^{\infty} x[n+l]\, y^*[n], \qquad l = 0, \pm 1, \pm 2, \ldots,}$$

where *l* is called time lag.

I used the cross-correlation function (CCF) that provided by **R**, an open source Language and Environment for Statistical Computing Software, to measure the correlation between pairs of time series data in different region, each of which derived their curve with ILI counts as the trends to represent the influenza activity along a specified time period. In addition to measuring the degree of correlation, cross-correlation analysis also finds the time lag between two time series that maximize the correlation. A plot of the sample correlations versus the time lags is called a correlogram (Figure 4).

4.2. Data Sources and Study Duration

4.2.1. ILINet

Time series data that consist of the weighted percentage of weekly patient visits to healthcare providers for ILI relative to different population size and the total regional patient visits in each of the ten Human and Health Services (HHS) regions were obtained from the CDC ILINet website. The study duration consisted of a total of 27 weeks duration from the 35th week of 2009 to the 9th week of 2010, which covered the initial wave of traditional flu season (Influenza Season, 2009-2010). I used the cross-correlation analysis (in R, autocorrelation, the ACF, is the built-in function used for CCF) to measure the maximal correlation and the corresponding time lag between pairs of the ten HHS regions, giving a 10 x 10 correlation matrix. The symmetric correlation matrix will be further processed and transformed into an asymmetric one for the construction of cross-correlation networks, as described in the section 4.3. A schematic diagram representing how cross-correlation analysis works on the pairwise ILINet data sets is shown in Figure 7.

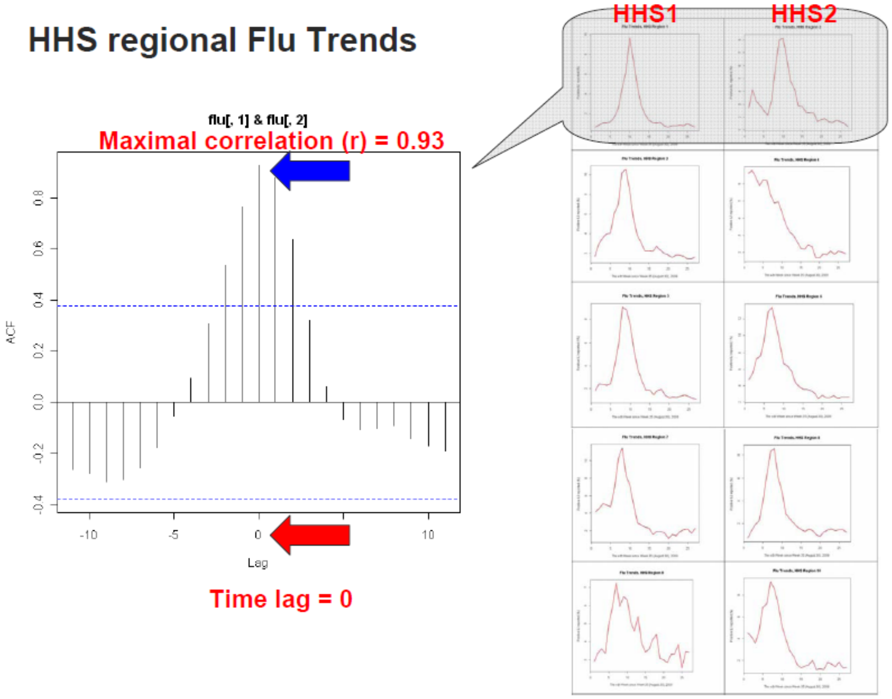


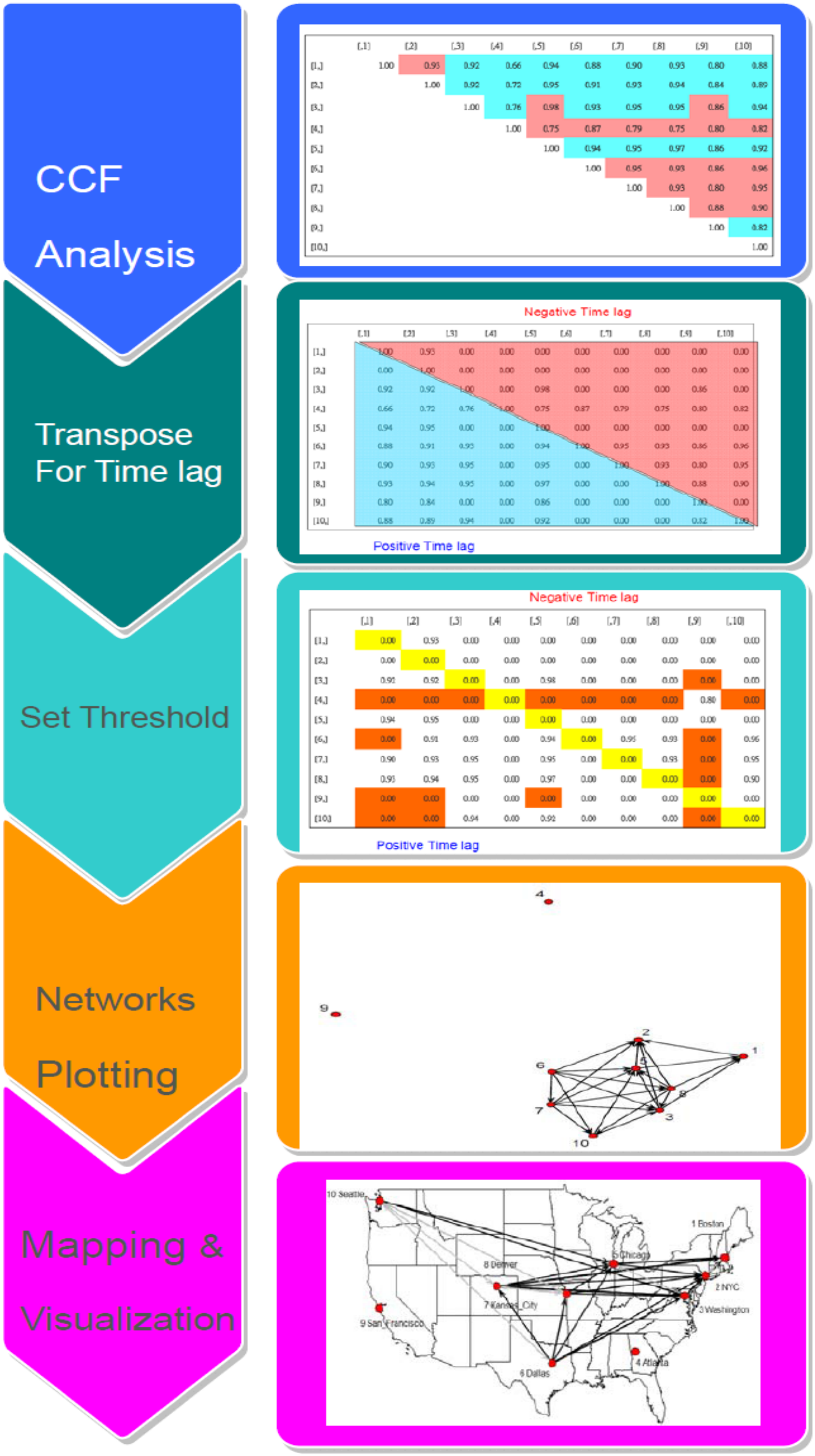
Figure 7. HHS Regional Flu Trends and the Cross-Correlation analysis between pairs. Example cross-correlation analysis applied in HHS1 and HHS2 shows the maximal correlation 0.93 (blue arrow) with the corresponding time latency 0 (red arrow).

4.2.2. Google Flu Trends

The weekly query counts from flu-related searches collected by Google flu trends were obtained from Google Flu Trends website. Data used for cross-correlation analysis and further networks modeling involves 51 state-level data sets (including 50 states and DC) in the United States. The study duration consisted of a total of 52 weeks duration from April 26th 2009 to April 18th 2010. Using cross-correlation analysis to measure maximal correlation and the corresponding latency between each pair of states, the resulting 51 x 51 matrix was further processed using the method described in section 4.3.

4.3. Construction and Visualization of Cross-Correlation Networks

The first step of networks construction involved cross-correlation analysis. Maximal cross-correlation (CCMAX) values with the corresponding hidden time lag were extracted and the symmetric correlation matrix was produced, as described in section 4.2.1. The diagonal of a correlation matrix always consists of values one (The diagonal values were later set as zeros). In the second step, those correlation values with negative time lag remained in the upper triangle (the ones above and to the right of the diagonal) and those with positive time lag were switched to the lower triangle (the ones below and to the left of the diagonal). By setting the threshold value (for instance, 0.9), the third step involved setting those values as zeros if the cross-correlation values were below that of the threshold. In the fourth step, the resulting matrix was then processed and visualized by the built-in networks function provided by **R**, giving cross-correlation networks as desired. Finally networks were plotted and mapped in the geographic map. The process of construction and visualization of cross-correlation networks using ILINet 10 HHS Regional Flu Trends data is depicted in Figure 8. The corresponding R code was illustrated in Figure 9.



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	1.00	0.93	0.92	0.66	0.94	0.88	0.90	0.93	0.80	0.88
[2,]		1.00	0.92	0.72	0.95	0.91	0.93	0.94	0.84	0.89
[3,]			1.00	0.76	0.98	0.93	0.95	0.95	0.86	0.94
[4,]				1.00	0.75	0.87	0.79	0.75	0.80	0.82
[5,]					1.00	0.94	0.95	0.97	0.86	0.92
[6,]						1.00	0.95	0.93	0.86	0.96
[7,]							1.00	0.93	0.80	0.95
[8,]								1.00	0.88	0.90
[9,]									1.00	0.82
[10,]										1.00

Negative Time lag

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	1.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[2,]	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[3,]	0.92	0.92	1.00	0.00	0.98	0.00	0.00	0.00	0.86	0.00
[4,]	0.66	0.72	0.76	1.00	0.75	0.87	0.79	0.75	0.80	0.82
[5,]	0.94	0.95	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
[6,]	0.88	0.91	0.93	0.00	0.94	1.00	0.95	0.93	0.86	0.96
[7,]	0.90	0.93	0.95	0.00	0.95	0.00	1.00	0.93	0.80	0.95
[8,]	0.93	0.94	0.95	0.00	0.97	0.00	0.00	1.00	0.88	0.90
[9,]	0.80	0.84	0.00	0.00	0.86	0.00	0.00	0.00	1.00	0.00
[10,]	0.88	0.89	0.94	0.00	0.92	0.00	0.00	0.00	0.82	1.00

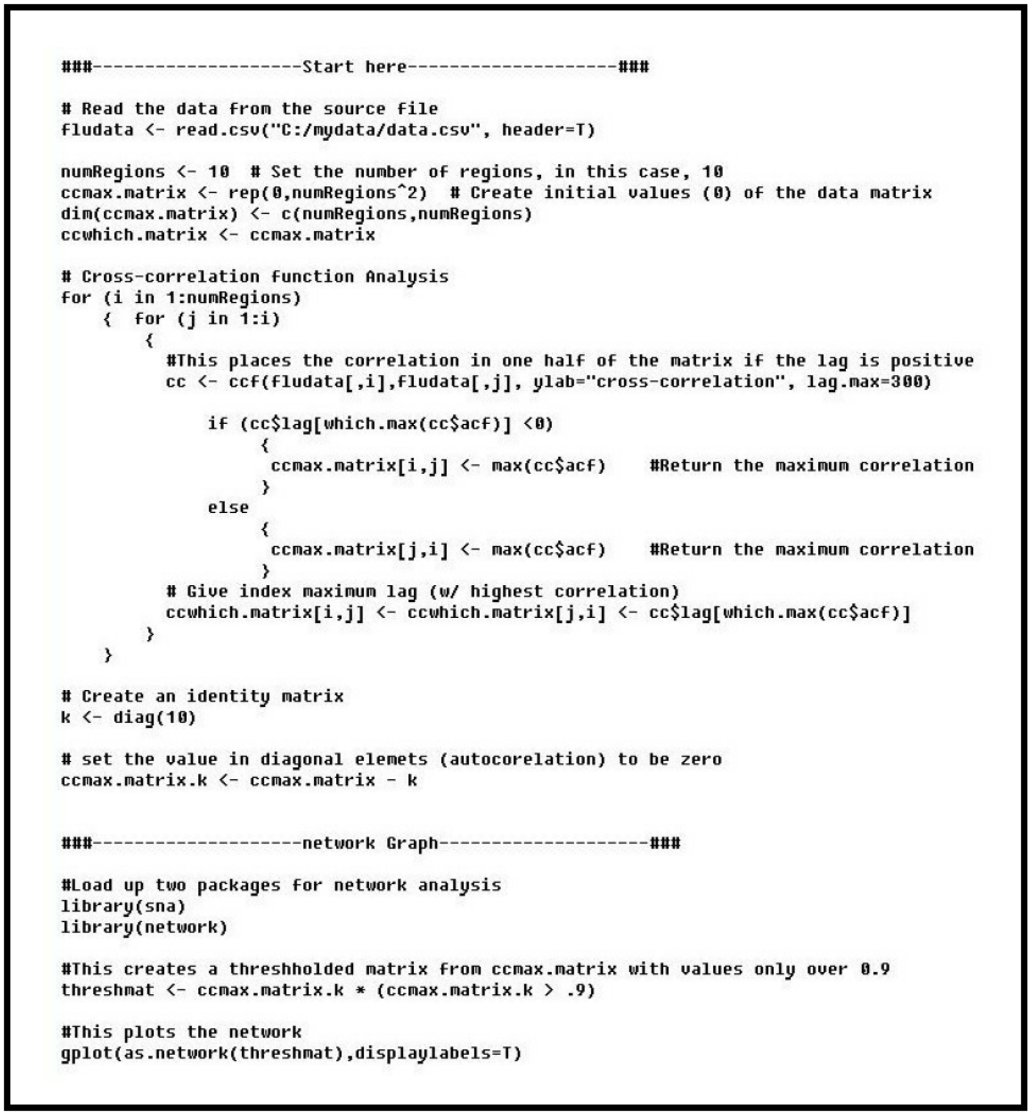
Positive Time lag

Negative Time lag

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[2,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[3,]	0.91	0.92	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00
[4,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00
[5,]	0.94	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[6,]	0.00	0.91	0.93	0.00	0.94	0.00	0.95	0.93	0.00	0.96
[7,]	0.90	0.93	0.95	0.00	0.95	0.00	0.00	0.93	0.00	0.95
[8,]	0.93	0.94	0.95	0.00	0.97	0.00	0.00	0.00	0.00	0.90
[9,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[10,]	0.00	0.00	0.94	0.00	0.92	0.00	0.00	0.00	0.00	0.00

Positive Time lag

Figure 8. The process of construction and visualization of cross-correlation networks using ILINet 10 HHS regional flu trends data as an example.



```
###--------------------Start here--------------------###

# Read the data from the source file
fludata <- read.csv("C:/mydata/data.csv", header=T)

numRegions <- 10  # Set the number of regions, in this case, 10
ccmax.matrix <- rep(0,numRegions^2)  # Create initial values (0) of the data matrix
dim(ccmax.matrix) <- c(numRegions,numRegions)
ccwhich.matrix <- ccmax.matrix

# Cross-correlation function Analysis
for (i in 1:numRegions)
    {  for (j in 1:i)
        {
          #This places the correlation in one half of the matrix if the lag is positive
          cc <- ccf(fludata[,i],fludata[,j], ylab="cross-correlation", lag.max=300)

              if (cc$lag[which.max(cc$acf)] <0)
                   {
                    ccmax.matrix[i,j] <- max(cc$acf)    #Return the maximum correlation
                   }
              else
                   {
                    ccmax.matrix[j,i] <- max(cc$acf)    #Return the maximum correlation
                   }
          # Give index maximum lag (w/ highest correlation)
          ccwhich.matrix[i,j] <- ccwhich.matrix[j,i] <- cc$lag[which.max(cc$acf)]
        }
    }

# Create an identity matrix
k <- diag(10)

# set the value in diagonal elemets (autocorelation) to be zero
ccmax.matrix.k <- ccmax.matrix - k


###--------------------network Graph--------------------###

#Load up two packages for network analysis
library(sna)
library(network)

#This creates a threshholded matrix from ccmax.matrix with values only over 0.9
threshmat <- ccmax.matrix.k * (ccmax.matrix.k > .9)

#This plots the network
gplot(as.network(threshmat),displaylabels=T)
```

Figure 9. The R code for the construction and visualization of cross-correlation networks using ILINet 10 HHS regional flu trends data as an example.

4.4. Data analysis

Time series analysis with cross-correlation measurement was performed using the **R** statistical package version 2.13.0 (R Development Core Team).

CHAPTER 5 ANALYSIS RESULTS

5.1. Cross-Correlation Networks Using ILINet Data Set

By setting a minimum threshold value (0.9) of CCMAX to highlight regions having the most aggregate curve in ILI, an example network showing ILI correlations between 10 HHS regions was plotted and then mapped as a way of visualizing ILI disease transmission pattern during the study period (Figure 10). Regions 4 and 9 are disconnected from the core network due to weak correlation with other regions. These outliers reflect an early peak (region 4) and weak transmission (region 9) of ILI cases in hospital visit data. The abstract describing part of this study had been presented in Annual Conference 2010 of International Society for Disease Surveillance (ISDS) at Park City, Utah. (Appendix A.).

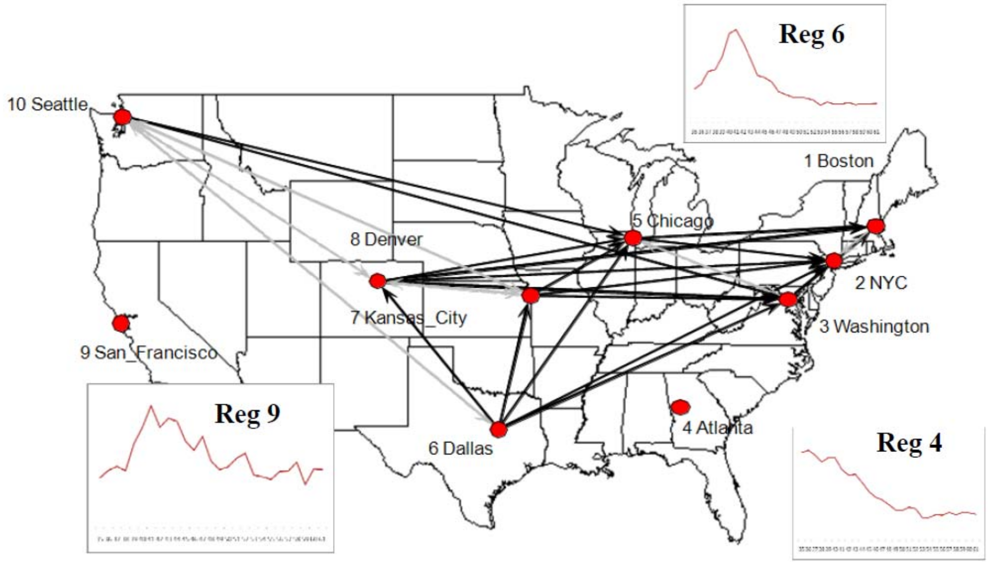


Figure 10. Cross-correlation network representing fraction of ILI cases in the 10 HHS regions during the 2009-2010 influenza season. The corresponding latency between time series is represented by directed ties in the graph. Grey (bidirectional) edges represent regions correlated with no time lag.

5.2. Cross-Correlation Networks Using Google Flu Trends Data Set

Use of aggregate case counts collected by Google Flu Trends and further cross-correlation analysis, cross-correlation networks for modeling ILI transmission pattern during the 2009-2010 influenza season was constructed and shown in Figure 11. The selection of threshold is 0.97 given the highly correlated signals between states. The states of Alaska, Georgia, Hawaii, Maine, Nevada, and Tennessee are disconnected from the core network due to weak correlation with other regions. Cross-correlation networks visualizing at different thresholds using state-level Google Flu Trends data is shown in Appendix B.

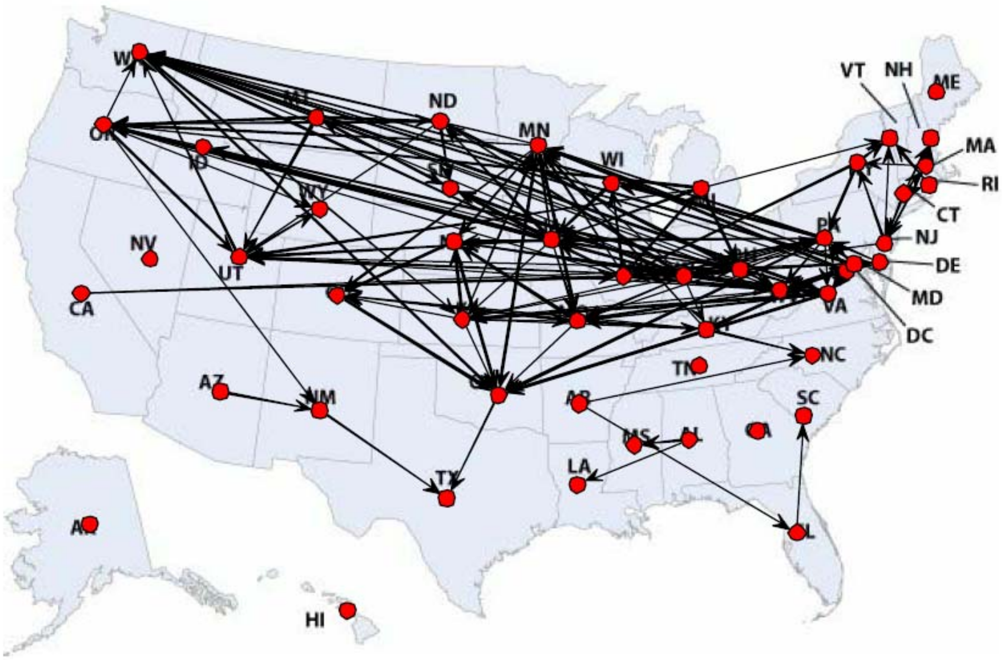


Figure 11. Cross-correlation network representing fraction of ILI cases in the 51 state level regions during the 2009-2010 influenza season.

CHAPTER 6 EVALUATION METHODS

6.1. Data sets to be used for evaluation

To validate that the cross-correlation based method be a feasible way of visualizing how ILI was transmitted in a spatio-temporal pattern, datasets from flight passenger volume, census worker flow, and distance between centroids will be evaluated as the indicative or proxy for influenza transmission. The state-level cross-correlation network constructed using the Google Flu Trends dataset was selected for evaluation.

6.2 Data Collection and Processing

6.2.1 Airline Origin and Destination Survey (DB1B)

DB1B data sets in state-to-state passenger's volume that covered the duration of one year from the third quarter 2009 to the second quarter 2010 were collected. Data were initially stored as a 51 x 51 asymmetric matrix that consisting of pairwise origin-to-destination state level data elements, and then further processed to yield a symmetric matrix that stands for relative flight passenger's volume between states.

$$X_{i,j} = X_{j,i} = \frac{X_{i,j}(\text{original}) + X_{j,i}(\text{original})}{(\text{Population}_{(i)} + \text{Population}_{(j)})}$$

where i and j represent different state and $\text{Population}_{(i)}$ represents the population size of state i.

6.2.2. Census 2000 Worker Flow Files

The Census 2000 worker flow data sets consisting of worker's volume commuting between states were collected from the website as the link stated before. Data were initially stored as a 51 x 51asymmetric matrix that consisting of pairwise origin-to-destination state level data elements, and then underwent further processing to a symmetric matrix that stands for relative worker flow volume between states. The method of data processing is as the following:

$$X_{i,j} = X_{j,i} = \frac{X_{i,j}(\text{original}) + X_{j,i}(\text{original})}{(\text{Population}_{(i)} + \text{Population}_{(j)})}$$

where i and j represent different state and $\text{Population}_{(i)}$ represents the population size of state i.

6.2.3. Census 2010 Centers of Population by State

The Census 2010 Centers of Population by State data sets were retrieved from the link stated before, giving a 51 x51 symmetric matrix in a form of pairwise state-to-state data elements that stands for centroids distance between states.

6.3. Data Transformation

Maximal cross-correlation values (CCMAX) were retrieved from the first step of network construction and then processed by logit transformation, and will be used as the outputs for linear regression.

$$\operatorname{logit}(Y_{i,j}) = \log(\frac{Y_{i,j}}{1 - Y_{i,j}})$$

where $Y_{i,j}$ is the maximal cross-correlation value between state i and state j.

The set of inputs *X* were data elements retrieved from symmetric matrices calculated using three different data sets, e.g., flight passenger flow, census worker flow, and distance between centroids. Logarithmic transformation, represented as $\log(X_{i,j})$, was performed in individual data elements of input set for further statistics analysis.

6.4. Statistics Analysis

A total of three data sets using input variables from flight passenger flow, census worker flow, and distance between centroids will be evaluated for significance. For each evaluation set, simple linear regression will be applied using $\log(X_{i,j})$ as the input and $\text{logit}(Y_{i,j})$ as the output. The best-fit line of regression was determined using the method of least squares and the coefficient of determination R^2 was used to measure the proportion of variability in a data set that is accounted for by the regression model. All samples with missing data will be eliminated before the analysis. Those input variables with value of zero before logarithmic transformation will also be eliminated since log 0 has no definition.

6.5. Data analysis

Data were entered, processed and analyzed with SPSS for Windows (Release 16.0, SPSS Inc., Chicago, IL, USA). Simple linear regression analysis was performed to determine independent predictors of CCMAX. A significant difference was accepted as a two sided P-value of less than 0.05.

CHAPTER 7 EVALUATION RESULTS

7.1. Sample Size

The evaluation analysis focused on 51 state-level cross-correlation networks constructed by using Google Flu Trends dataset. The data set downloaded from DB1B flight data had no record related to Delaware state. There were a total of 5 flight volume records and 18 worker flow records yielded a zero value and were hence eliminated before data transformation and analysis. The total sample size left to be evaluated was 1202 (e.g., (50 x 50 -50)/2 – 5- 18 = 1202).

7.2. Linear Regression

7.2.1. Census Worker Flow

Using linear regression analysis, the calculated best-fit line of regression (logit-ccmax V.S. log-worker plot) had a slope of 0.148, with a R^2 of 0.088 (Figure.12).

7.2.2. Distance between Centroids

Linear regression analysis showed that the calculated best-fit line of regression (logit-ccmax V.S. log-distance plot) had a slope of -0.425 and a R^2 of 0.157 (Figure.12).

7.2.3. Flight Passenger Flow

The calculated best-fit line of regression (logit-ccmax V.S. log-flight plot) had a slope of 0.007 and a R^2 <0.001 (Figure.12).

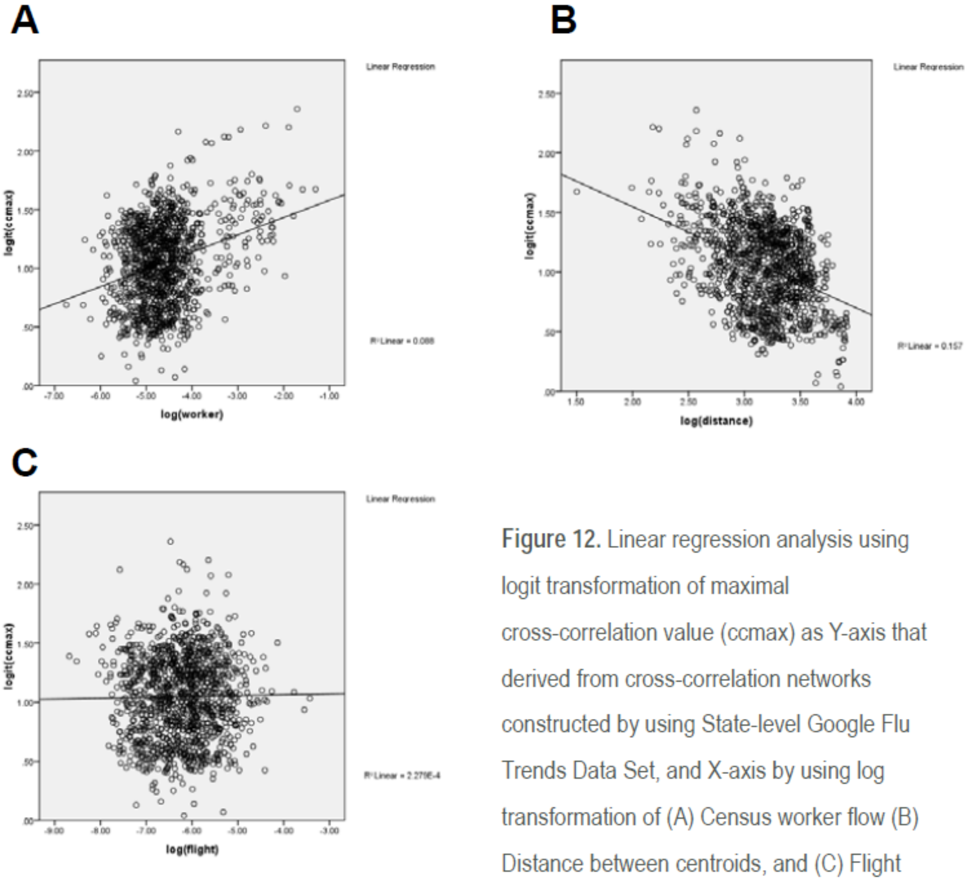


Figure 12. Linear regression analysis using logit transformation of maximal cross-correlation value (ccmax) as Y-axis that derived from cross-correlation networks constructed by using State-level Google Flu Trends Data Set, and X-axis by using log transformation of (A) Census worker flow (B) Distance between centroids, and (C) Flight passengers' volume.

7.2.4. Summary of the Linear Regression

Summary report on linear regression analysis is shown in Table 1. Census worker flow and distance between centroids, but not flight passenger flow, were associated maximal cross-correlation values.

Table 1. Summary report on the results of linear regression

	Simple Linear Regression				
	r	R^2	**Slope**	**P**	**95% CI**
Census Worker Flow					
	0.296	0.088	0.148	<0.001*	[0.121, 0.174]
Distance between Centroids					
	0.396	0.157	-0.425	<0.001*	[-0.481, -0.369]
Flight Passenger Flow					
	0.015	<0.001	0.007	0.599	[-0.020, 0.034]

Abbreviations: r stands for correlation coefficient; R^2 stands for R square; CI stands for confidence interval. * P value <0.05

CHAPTER 8 DISCUSSIONS

This study proposed a new method to interpret the correlation between case rates in geographic regions. By using cross-correlation analysis on those public accessible data sets, we developed and modeled disease transmission patterns in ILI that help epidemiologists quickly identify and visualize the similarities in case rates across different regions. As the indicative or proxy for influenza transmission, variables from other data sets that stand for flight passenger flow, census worker flow, and distance between centroids were used to compare the pairwise correlation values derived from our cross-correlation analysis. We found that census worker flow and distance between centroids, but not flight passenger flow, were significantly associated with the maximal cross-correlation values we were relying on for networks construction, given the evaluation analysis using state-level Google Flu Trends data.

Our results show that census worker flow and distance between centroids as the independent predictors are significantly associated with observed pairwise correlation values derived from our cross-correlation analysis, however, the R^2 coefficient of determination for both of the models being constructed are not high (0.088 and 0.157, respectively). In statistics, R^2 is a measure of how well the regression line approximates the observed data points and can also be used to measure the unexplained variance. An R^2 of 1.0 indicates that the regression line perfectly fits the data (Everitt, 2002). In these instances, our evaluation results can support the method we are using for the construction of cross-correlation networks, however, there are likely to be some unexplained factors governing the nature of ILI disease transmission in the season we chose for analysis, given the resulting R^2 values. For the data sets that have been chosen, there is collinearity present in the data on the explanatory variables. A more complex (multivariate) model would likely yield greater insight.

Network analysis has a role in several aspects of infectious disease modeling, including simulation, contact investigation, and sampling. Our analysis focused on using real world aggregate data for the construction of cross-correlation networks as a tool to visualize the spatio-temporal patterns of ILI transmission in the population level. By case investigation and phylogenetic analysis, it is possible to reconstruct the transmission networks in individual level or small scale outbreak (Bon, 2010; Chalmet, 2010), however, such approach can be time and resource consuming in large scale disease transmission such ILI. Although a study exists that utilized the phylogenetic analysis for modeling disease transmission patterns in the population level (Flavia 2011), our study utilized those data sets that could be easily accessible from the web either for analysis or evaluation. With data sets comprised of geocoded strains circulating during an ILI outbreak across different regions, it may also be possible to validate cross-correlation networks based on phylogenetic analysis (Appendix C).

The reasons for using census worker flow, distance between centroids, and flight passenger flow are based on the reasons that ILI is mainly transmitted by contact (Valleron, 2010). In this study, we took flight passenger volume and census worker flow as indicatives of actual contact, and distance as a proxy for likelihood of contact. Until now, there is no studies related to census worker flow or flight passenger flow being investigated for their role in infectious disease transmission, however, studies evaluating the impact of social networks on spreading of infectious diseases have been elaborated in the literature (Cauchemez, 2010; Abernethy, 2005). Our study found that census worker flow, but not flight passenger flow, is associated with the cross-correlation networks being constructed, which could serve as new research themes for investigators.

Distance as a factor in infectious disease transmission has been well elaborated in either experimental setting or real world disease transmission patterns (Spekreijse, 2011; Tuite, 2011), suggesting that infectious disease transmission over a long distance is a less likely route of spread. In our study, we used distance between state centroids as one of the variables to be evaluated the influence on cross-correlation networks. Result shows that distance between states is inverse proportional to disease spreading that modeled by cross-correlation networks, and that further support the approach we are using can be a feasible way of modeling infectious disease transmission in a spatio-temporal manner.

One of the major drawbacks of using DB1B as the flight passenger flow is the lack of specificity. Those airports (and their associated states) with high volume passenger flow may also serve as the international transportation stations for passengers' transit. Sick passengers flying to other nations may exert little influence on the impact of disease spreading in the Destination State, although the case counts will be reflected in the DB1B data sets and hence be used for our evaluation. Our study found no strong association between cross-correlation networks and flight passenger flow. This doesn't preclude the possibility of their association given we didn't have detailed data related to the "actual" flight passenger flow. Future research trends may include simulation data that can actually model the ILI case counts in each geographic region and also passengers' behavior with sufficient details suitable for subgroup analysis.

For simulation, agent-based models (ABM) can be explored as an alternative approach. ABM is a computational method for simulating the actions and interactions of autonomous decision-making agents (either individual or groups) in an attempt to assess the behavior of the system being modeled as a whole (Bonabeau, 2002). Our evaluation analysis utilized the aggregate data sets of ILINets and Google Flu Trends. These two systems that need to be analyzed are complex without finer levels of granularity once they are organized into accessible data sets. As an alternative approach in the experiment, we could use agent-based simulation to track individual episodes of disease transmission between regions. The aggregate surveillance data can also be generated by this approach. For the next step, we would compare the region-to-region transmission matrix to the observed cross-correlation network to determine to what extent the observed correlations resemble the actual transmission events from the simulation. Since we lack a gold standard to measure transmission events in aggregate data collected mainly for syndromic surveillance, it is possible to re-create and predict the appearance of complex phenomena like ILI activity by simulation approach. Such a simulation experiment exhibits the potential to have far-reaching effects on the way that public health practitioners use informatics to support decision-making during infectious disease outbreak with the knowledge of how disease spreading.

Here we used cross-correlation analysis and cross-correlation networks as the way to identify and visualize disease transmission patterns in ILI. During networks construction, we plotted the connections with correlation values above a pre-selected threshold in order to highlight the most important similarities between case rates. The selection of the threshold was tentative for balancing the sensitivity and specificity, in an attempt to identify "true" transmission links based on standard not presently available. The higher the threshold value selected, the more specific (and less sensitive) the links are modeled. Such a threshold selection strategy can also be employed using the ratio of connections between contiguous states versus total connections. This value did increase as a function of the threshold being selected (Strategies in selecting the thresholds are shown in Appendix D).

One of the major problems using simple linear regression is the requirement to fulfill the assumption of statistical independence of observations. Our data sets used to evaluate the validity of cross-correlation networks (census worker flow, flight passenger's volume, or distance between centroids) all contain mutually dependent observations. Here we take census worker flow as an example. Suppose the number of workers commuting from California to Oregon is A, the number commuting from California to Washington is B, and the number commuting from Oregon to Washington is C. Although A and B are independent, jointly C is not independent since it is constrained by the other two (A and B). In our evaluation analysis, we used univariate analysis by simple linear regression and did not consider the statistical dependence issue. Although this might be a problem, yet this can be solved by performing the regression on a very limited sample of 25 state pairs with no repeated states, or by doing 50 single state analyses.

There are limitations in this study. First, assumptions were simplified on the selection of threshold, although it was necessary for modeling based on probability of "true" network connection. Secondly, aggregate case counts used for networks modeling were ILI visits (ILINet) and web-queries (Google Flu Trends) rather than based on the gold standard confirmed cases. Besides, there was no state-level aggregate data collected by ILINet. A data set like DISTRIBUTE might be utilized in further study if accessibility is approved. Thirdly, Data sets used for networks evaluation are those proxies of influenza transmission, which don't necessarily represent the "true" disease transmission route. Studies relying on phylogenetic or simulated data can be further explored in the future. Fourthly, simple linear regression might not be the suitable method of evaluating the association between variables, especially when our data sets can not satisfied the assumption of independency. A method using non-linear method, or sub-group analysis using independent pairs could be explored as an alternative approach.

CHAPTER 9 CONCLUSIONS

We introduce a new approach to interpret the correlation and time lag between time series data that collect aggregate case counts for syndromic surveillance. This method can be a tool to identify and visualize patterns in disease spreading across different geographic areas. The evaluation results further explain spatio-temporal spread of disease using cross-correlation networks being constructed. The networks may then serve as a basis to evaluate intervention options during outbreak of pandemic influenza or other emerging infectious diseases. Using cross-correlation analysis, we would like to extend this approach to studying patterns in different diseases or matching disease trends to other syndrome signals in the future.

Bibliography

Abernethy N. Automating social network models for tuberculosis contact investigation. Stanford University, 2005.

Beitel AJ, Olson KL, Reis BY, et al. Use of emergency department CC and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatr Emerg Care. 2004;20:355-60.

Bellazzini MA, Minor KD. ED syndromic surveillance for novel H1N1 spring 2009. Am J Emerg Med. 2011;29:70-4.

Bon I, Ciccozzi M, Zehender G, et al. HIV-1 subtype C transmission network: the phylogenetic reconstruction strongly supports the epidemiological data. J Clin Virol. 2010;48:212-4.

Bonabeau E. Agent-based modeling: methods and techniques for simulating human systems. Proc Natl Acad Sci U S A. 2002;99 Suppl 3:7280-7.

Bradley CA, Rolka H, Walker D, et al. BioSense: implementation of a National Early Event Detection and Situational Awareness System. MMWR Morb Mortal Wkly Rep. 2005;54 Suppl:11-9.

Brownstein JS, Kleinman KP, Mandl KD. Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. Am J Epidemiol. 2005;162:686-93.

Buehler JW, Whitney EA, Smith D, et al. Situational uses of syndromic surveillance. Biosecur Bioterror. 2009;7:165-77.

Cauchemez S, Bhattarai A, Marchbanks TL, et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proc Natl Acad Sci U S A. 2011;108:2825-30.

Census Centers of Population by State. United States Census Bureau. http://www.census.gov/geo/www/2010census/centerpop2010/statecenters.html. Accessed on June 21, 2011.

Census Worker Flow. United States Census Bureau. http://www.census.gov/population/www/cen2000/commuting/index.html. Accessed on June 21, 2011.

Chalmet K, Staelens D, Blot S, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis. 2010;10:262.

Dailey L, Watkins RE, Plant AJ. Timeliness of data sources used for influenza surveillance. Am Med Inform Assoc. 2007;14:626-31.

DB1B. Airline Origin and Destination Survey. Research and Innovative technology Administration. Bureau of Transportation Statistics. http://www.transtats.bts.gov/Tables.asp?DB_ID=125&DB_Name=Airline Origin and Destination Survey (DB1B). Accessed on June 21, 2011.

DiSTRIBuTE (Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking and Evaluation) project. http://distribute.syndromic.org/. Accessed on July 17, 2011.

Doornik JA. Improving the timeliness of data on influenza-like illnesses using Google search data. 8th Oxmetrics User Conference, March 18th, 2010. http://www.gwu.edu/~forcpgm/JurgenDoornik-final-Doornik2009Flu-Jan31.pdf. Accessed on June 21, 2011.

Doroshenko A, Cooper D, Smith G, et al. Evaluation of syndromic surveillance based on National Health Service Direct derived data--England and Wales. MMWR Morb Mortal Wkly Rep. 2005;54 Suppl:117-22.

Espino JU, Hogan WR, Wagner MM. Telephone triage: a timely data source for surveillance of influenza-like diseases. AMIA Annu Symp Proc. 2003:215-9.

Everitt, B.S., 2002. The Cambridge Dictionary of Statistics. 2nd edition. Cambridge University Press, Cambridge, UK.

Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc. 2006:244-8.

Flavia GB, Natarajaseenivasan K. Phylogenetic analysis of H1N1 sequences from pandemic infections during 2009 in India. Bioinformation. 2011;5:416-21.

Fleischauer AT, Silk BJ, Schumacher M, et al. The validity of chief complaint and discharge diagnosis in emergency department-based syndromic surveillance. Acad Emerg Med. 2004;11:1262-7.

Franz DR, Midwest Research Institute. Disease Surveillance and International Biosecurity. In: Countering Terrorism: Biological Agents, Transportation Networks, and Energy Systems. Summary of a U.S.-Russian Workshop. Washington DC, National Academy of Sciences 2009:73-78.

Gimpel K. Statistical Inference in Graphical Models. 2006. http://www.cs.cmu.edu/~kgimpel/papers/da-02-draft.pdf. Accessed on July 10, 2011.

Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012-4.

Google Flu Trends. How does this work? http://www.google.org/flutrends/about/how.html. Accessed on June 21, 2011.

Greenko J, Mostashari F, Fine A, et al. Clinical evaluation of the Emergency Medical Services (EMS) ambulance dispatch-based syndromic surveillance system, New York City. Urban Health. 2003 ;80:i50-6.

Henning KJ. Overview of Syndromic Surveillance. What is syndromic surveillance? MMWR Morb Mortal Wkly Rep. 2004;53 Suppl:5-11.

ILINet. U.S. Outpatient Influenza-like Illness Surveillance Network. http://www.cdc.gov/flu/weekly/overview.htm. Accessed on June 21, 2011.

Influenza Season, 2009-2010. Center for Disease Control and Prevention. http://www.cdc.gov/flu/pastseasons/0910season.htm. Accessed on June 21, 2011.

Lemay R, Mawudeku A, Shi Y, et al. Syndromic surveillance for influenzalike illness. Biosecur Bioterror. 2008 Jun;6(2):161-70.

Lemay R, Mawudeku A, Shi Y, et al. Syndromic surveillance for influenzalike illness. Biosecur Bioterror. 2008;6:161-70.

Lombardo J, Burkom H, Elbert E, et al. A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). J Urban Health. 2003;80:i32-42.

Lu HM, Zeng D, Trujillo L, et al. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. J Biomed Inform. 2008;41:340-56.

Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11:141-50.

May LS, Griffin BA, Bauers NM, et al. Emergency department chief complaint and diagnosis data to detect influenza-like illness with an electronic medical record. West J Emerg Med. 2010;11:1-9.

Murphy KP. An introduction to graphical models. 2001. http://www.cs.ubc.ca/~murphyk/Papers/intro_gm.pdf. Accessed on July 10, 2011.

Newman MEJ. The structure and function of complex networks. SIAM Review 2003:45;167–256.

Overhage JM, Grannis S, McDonald CJ. A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. Am J Public Health. 2008;98:344-50.

Porta M. A Dictionary of Epidemiology. 5th ed. USA: Oxford University Press; 2008.

R Development Core Team: R: A language and environment for statistical computing [http://www.r-project.org].

Reis BY, Mandl KD. Syndromic surveillance: the effects of syndrome grouping on model accuracy and outbreak detection. Ann Emerg Med. 2004;44:235-41.

Shumway RH, Stoffer DS. Time series analysis and its applications. New York: Springer; 2000.

Signorini A, Segre AM, Polgreen PM. 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. 2011;6:e19467.

South BR, Chapman WW, Delisle S, et al. Optimizing A syndromic surveillance text classifier for influenza-like illness: Does document source matter? AMIA Annu

Symp Proc. 2008:692-6.

Spekreijse D, Bouma A, Koch G, et al. Airborne transmission of a highly pathogenic avian influenza virus strain H5N1 between groups of chickens quantified in an experimental setting. Vet Microbiol. 2011 Apr 22. [Epub ahead of print]

Spinal Cord Research center (SCRC) - Cross-correlation Analysis of Filtered and Rectified Waveforms. http://www.scrc.umanitoba.ca/doc/tutorial/tutorial_14.html. Accessed on June 15th, 2011.

Teich JM, Wagner MM, Mackenzie CF, et al. The informatics response in disaster, terrorism, and war. J Am Med Inform Assoc 2002;9:97-104.

Thaper N, Wong D, Lee J. The changing geography of the population. centroids in the United States. Geographical Bulletin 1999;41: 45-56.

Tsui FC, Wagner MM, Dato V, et al. Value of ICD-9-Coded Chief Complaints for Detection of Epidemics. Proc AMIA Symp. 2001; 711-5.

Tuite AR, Tien J, Eisenberg M, et al. Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions. Ann Intern Med. 2011;154:593-601.

Valleron AJ, Cori A, Valtat S, et al. Transmissibility and geographic spread of the 1889 influenza pandemic. Proc Natl Acad Sci U S A. 2010;107:8778-81.

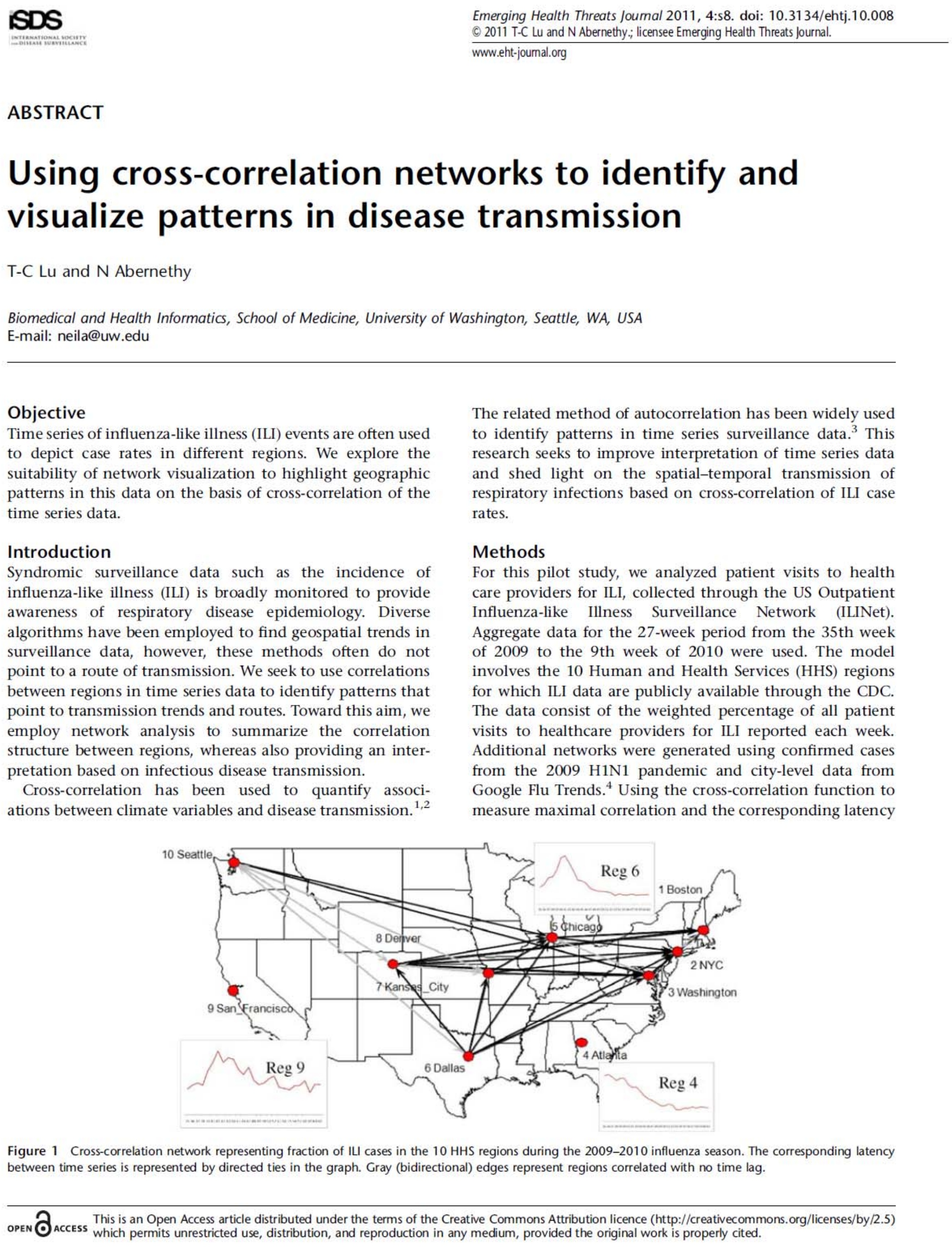
Yih WK, Teates KS, Abrams A, et al. Telephone triage service data for detection of influenza-like illness. PLoS One. 2009;4:e5260.

Zheng W, Aitken R, Muscatello DJ, et al. Potential for early warning of viral influenza activity in the community by monitoring clinical diagnoses of influenza in hospital emergency departments. BMC Public Health. 2007;7:250.

Appendices

Appendix A:

The abstract presented in Annual Conference 2010 of International Society for Disease Surveillance (ISDS).



ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Emerging Health Threats Journal 2011, 4:s8. doi: 10.3134/ehtj.10.008
© 2011 T-C Lu and N Abernethy.; licensee Emerging Health Threats Journal.
www.eht-journal.org

ABSTRACT

Using cross-correlation networks to identify and visualize patterns in disease transmission

T-C Lu and N Abernethy

Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, WA, USA
E-mail: neila@uw.edu

Objective

Time series of influenza-like illness (ILI) events are often used to depict case rates in different regions. We explore the suitability of network visualization to highlight geographic patterns in this data on the basis of cross-correlation of the time series data.

Introduction

Syndromic surveillance data such as the incidence of influenza-like illness (ILI) is broadly monitored to provide awareness of respiratory disease epidemiology. Diverse algorithms have been employed to find geospatial trends in surveillance data, however, these methods often do not point to a route of transmission. We seek to use correlations between regions in time series data to identify patterns that point to transmission trends and routes. Toward this aim, we employ network analysis to summarize the correlation structure between regions, whereas also providing an interpretation based on infectious disease transmission.

Cross-correlation has been used to quantify associations between climate variables and disease transmission.[1,2] The related method of autocorrelation has been widely used to identify patterns in time series surveillance data.[3] This research seeks to improve interpretation of time series data and shed light on the spatial–temporal transmission of respiratory infections based on cross-correlation of ILI case rates.

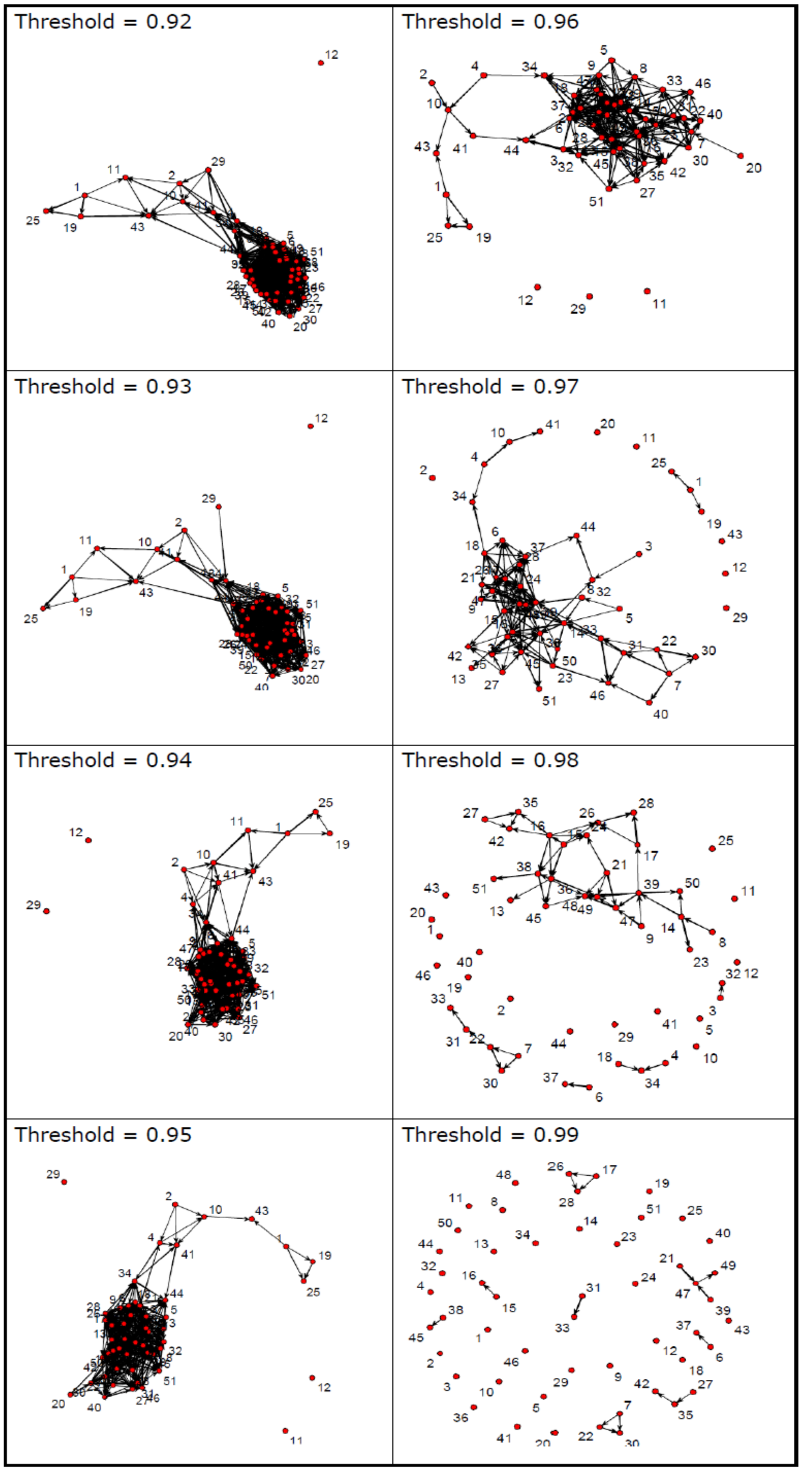
Methods

For this pilot study, we analyzed patient visits to health care providers for ILI, collected through the US Outpatient Influenza-like Illness Surveillance Network (ILINet). Aggregate data for the 27-week period from the 35th week of 2009 to the 9th week of 2010 were used. The model involves the 10 Human and Health Services (HHS) regions for which ILI data are publicly available through the CDC. The data consist of the weighted percentage of all patient visits to healthcare providers for ILI reported each week. Additional networks were generated using confirmed cases from the 2009 H1N1 pandemic and city-level data from Google Flu Trends.[4] Using the cross-correlation function to measure maximal correlation and the corresponding latency

Figure 1 Cross-correlation network representing fraction of ILI cases in the 10 HHS regions during the 2009–2010 influenza season. The corresponding latency between time series is represented by directed ties in the graph. Gray (bidirectional) edges represent regions correlated with no time lag.

Appendix B

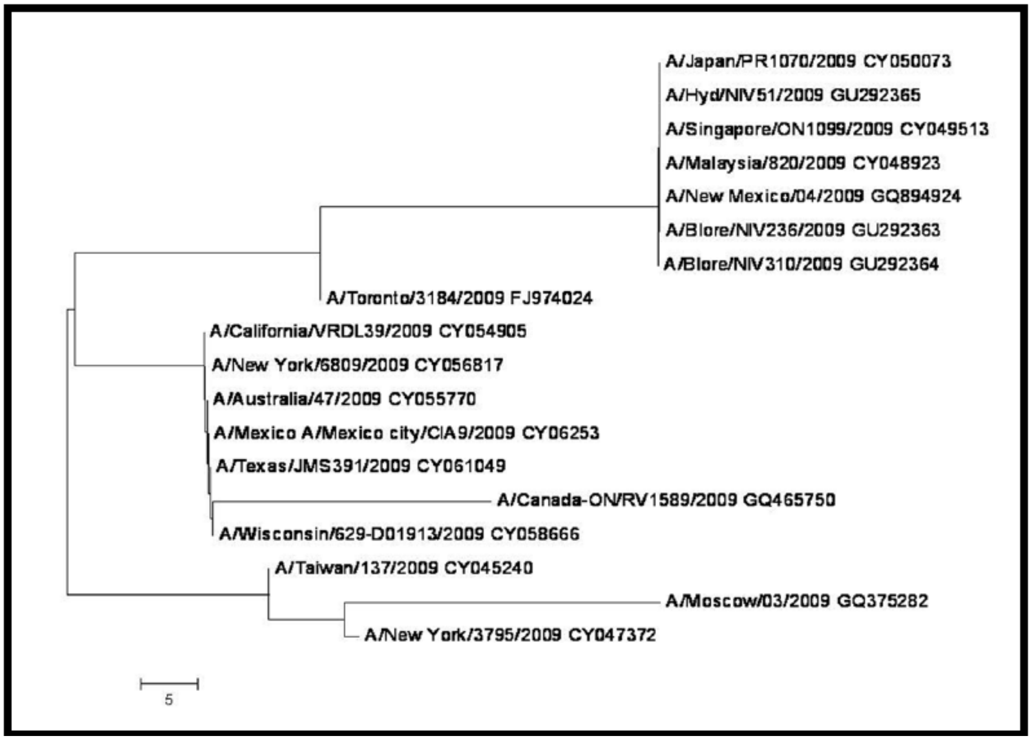
Cross-correlation networks visualization on selecting different thresholds using state-level Google Flu Trends data.



Appendix C

ILI transmission networks construction based on geocoded database using phylogenetic tree analysis.

A phylogenetic approach can be carried out by using data like the figure shown below (Flavia 2011), with the sample of strains that was geocoded. Since the strains are tagged with the geographic source of the strain, it is possible to validate the cross-correlation networks of ILI based on phylogenetic tree analysis.


A/Japan/PR1070/2009 CY050073
A/Hyd/NIV51/2009 GU292365
A/Singapore/ON1099/2009 CY049513
A/Malaysia/820/2009 CY048923
A/New Mexico/04/2009 GQ894924
A/Blore/NIV236/2009 GU292363
A/Blore/NIV310/2009 GU292364
A/Toronto/3184/2009 FJ974024
A/California/VRDL39/2009 CY054905
A/New York/6809/2009 CY056817
A/Australia/47/2009 CY055770
A/Mexico A/Mexico city/CIA9/2009 CY06253
A/Texas/JMS391/2009 CY061049
A/Canada-ON/RV1589/2009 GQ465750
A/Wisconsin/629-D01913/2009 CY058666
A/Taiwan/137/2009 CY045240
A/Moscow/03/2009 GQ375282
A/New York/3795/2009 CY047372
5

Appendix D: Threshold Selection Strategies

1. How and why to Select the Threshold

In order to hide the connections between weakly correlated regions, the following threshold selection strategies were adopted in an attempt to identify "true" transmission links based on standard not presently available. We are using contiguity as a proxy for a gold standard under the assumption that ILI transmission locally is more likely than over a long distance (Spekreijse, 2011; Tuite, 2011).

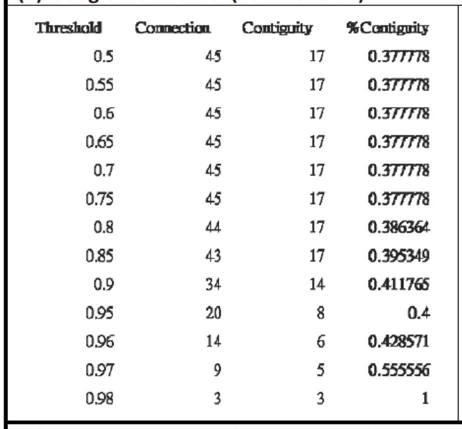
2. Percentage of Contiguity: the ratio of connections between contiguous states versus total connections.

$$\%\,\text{Contiguity} = \frac{\text{Connections between two Contiguous Regions/States}}{\text{Total Connections}}$$

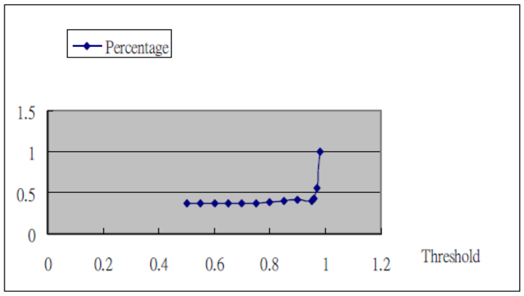
3. Selection Threshold for Google Flu Trends

Using Google Flu Trends data sets (aggregates counts data collected from April 26th, 2009 to April 18th, 2010) as an example for the construction of cross-correlation networks, the results by analyzing the percentage of contiguity are shown as (a).10 HHS-level and (b).51 state-level. With the increase in selecting the value of the threshold, the number of total connections and contiguous connections decrease. However, the percentage of contiguity increases.

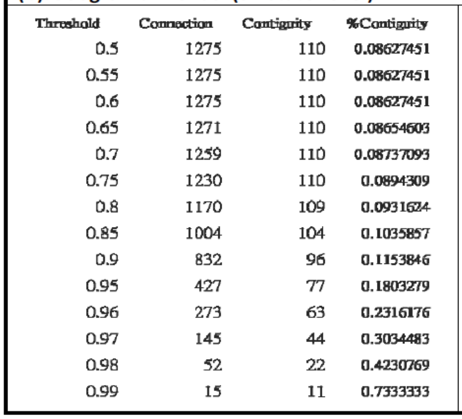
(a).Google Flu Trends (10 HHS-level)



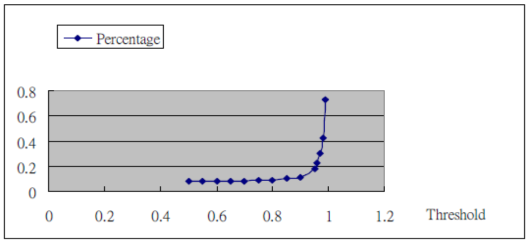
Threshold	Connection	Contiguity	%Contiguity
0.5	45	17	0.377778
0.55	45	17	0.377778
0.6	45	17	0.377778
0.65	45	17	0.377778
0.7	45	17	0.377778
0.75	45	17	0.377778
0.8	44	17	0.386364
0.85	43	17	0.395349
0.9	34	14	0.411765
0.95	20	8	0.4
0.96	14	6	0.428571
0.97	9	5	0.555556
0.98	3	3	1



(b).Google Flu Trends (51 state-level)



Threshold	Connection	Contiguity	%Contiguity
0.5	1275	110	0.08627451
0.55	1275	110	0.08627451
0.6	1275	110	0.08627451
0.65	1271	110	0.08654603
0.7	1259	110	0.08737093
0.75	1230	110	0.0894309
0.8	1170	109	0.0931624
0.85	1004	104	0.1035857
0.9	832	96	0.1153846
0.95	427	77	0.1803279
0.96	273	63	0.2316176
0.97	145	44	0.3034483
0.98	52	22	0.4230769
0.99	15	11	0.7333333


Percentage
Threshold