11-1-2012

# Graph Matching Based Decision Support Tools For Mitigating Spread Of Infectious Diseases Like H1N1

Jomon Aliyas Paul
*Kennesaw State University*, jpaul17@kennesaw.edu

Kedar Sambhoos
*CUBRC*

*Journal of Homeland Security and Emergency Management*

*Volume* 9, *Issue* 2 2012 *Article* 9

# Graph Matching Based Decision Support Tools For Mitigating Spread Of Infectious Diseases Like H1N1

**Jomon Aliyas Paul,** *Kennesaw State University*
**Kedar Sambhoos,** *CUBRC*

**Recommended Citation:**
Paul, Jomon Aliyas and Sambhoos, Kedar (2012) "Graph Matching Based Decision Support Tools For Mitigating Spread Of Infectious Diseases Like H1N1," *Journal of Homeland Security and Emergency Management*: Vol. 9: Iss. 2, Article 9.
DOI: 10.1515/1547-7355.1978

# Graph Matching Based Decision Support Tools For Mitigating Spread Of Infectious Diseases Like H1N1

Jomon Aliyas Paul and Kedar Sambhoos

## Abstract

Diseases like H1N1 can be prevented from becoming a wide spread epidemic through timely detection and containment measures. Similarity of H1N1 symptoms to any common flu and its alarming rate of spread through animals and humans complicate the deployment of such strategies. We use dynamic implementation of graph matching methods to overcome these challenges. Specifically, we formulate a mixed integer programming model (MIP) that analyzes patient symptom data available at hospitals to generate patient graph match scores. Successful matches are then used to update counters that generate alerts to the Public Health Department when the counters surpass the threshold values. Since multiple factors like age, health status, etc., influence vulnerability of exposed population and severity of those already infected, a heuristic that dynamically updates patient graph match scores based on the values of these factors is developed. To better understand the gravity of the situation at hand and achieve timely containment, the rate of infection and size of infected population in a specific region needs to be estimated. To this effect, we propose an algorithm that clusters the hospitals in a region based on the population they serve. Hospitals grouped together affect counters that are local to the population they serve. Analysis of graph match scores and counter values specific to the cluster helps identify the region that needs containment attention and determine the size and severity of infection in that region. We demonstrate the application of our models via a case study on emergency department patients arriving at hospitals in Buffalo, NY.

**KEYWORDS:** graph matching, H1N1, disaster planning, hospital demand estimation, public health, optimization

# 1. Introduction

Outbreaks of the H1N1 virus in 2009 have so far led to a total of 18,449 confirmed deaths around the world (World Health Organization (2009)). Timely detection and containment of such outbreaks can prove very helpful in averting the economic damage and human deaths and prevent it from becoming a widespread pandemic. However, limited success has been achieved in this regard because of the following challenges (Medicine Net (2009)): a) The symptoms of H1N1 flu are similar to common flu; b) Certain groups of individuals like children, pregnant women, those suffering from chronic conditions, elderly, etc., are more vulnerable to the virus; c) It can spread from animals to humans as well as from humans to humans at an alarming rate; d) Medication available to treat this virus cannot be dispensed randomly to people because a non-infected person might end up becoming more susceptible to developing H1N1; e) Sufficient quantities of vaccination for all of the susceptible population might not be available at all the times. Therefore, prioritization and careful determination of the population that receives the vaccine might become absolutely critical for efficient preparedness and response.

Research on H1N1 has taken two main directions. One: preparedness by simulating hypothetical scenarios to estimate expected cases and necessary resources required (Hagenaars et al. (2004), Longini et al. (2005), Feighner et al. (2008), Pan-InfORM (2009)). And two: timely detection of flu virus by studying the effect of size of infected population on the rate of infection (Mohtashemi et al. (2006), Reis et al. (2007)). Though the existing research provides some assistance to public health departments in decision making, it has the following shortcomings. Firstly, this body of literature bases its results on analysis of data from a single hospital and/or analysis of all hospitals in a region collectively. Analysis of a single hospital is not a good strategy when it comes to determining the size of infected or susceptible population in a region because all the infected patients might not receive treatment from the same hospital. Similarly, analysis of all hospitals together without a well-defined grouping logic might lead to errors and in many instances to a lot of false alarms. Secondly, number of infected is not a reliable measure of rate of infection because it does not precisely model the effect of severity of infection. Therefore, use of such a metric might result in inaccurate estimates of the rate of infection. In addition, two groups of infected individuals or those showing up with flu like symptoms, can have different rates of infection depending on whether they are part of high risk groups (chronic conditions, children, pregnant women, etc.) or not.

We overcome the above challenges through dynamic graph matching. We propose models that would perform continuous scanning of hospital patient data with the objective of finding patterns by comparing it with templates containing

Published by De Gruyter, 2012

the basic symptoms seen in infected patients. In this study, the incoming patient symptom data acts as sensor data. The data gathered from the hospital is assumed to be accurate and transcends the need to be tracked. The hypothesis of this research, infection template, is created using the help of Subject Matter Experts (SME) in this case doctors. Graph match scores developed using models proposed in this paper take into account not only the number of matching symptoms but also the severity of infection and therefore is a more reliable metric. Successful graph matches update values of counters. When values of the counter surpass the threshold values, public health officials are notified of a possible outbreak. This process acts as a situation and threat assessment tool for the analyst in this case public health officials.

As discussed earlier, if data of all the hospitals are aggregated and analyzed together, it can not only lead to lot of false alarms but also add to challenges in accurately identifying the origin and size of outbreak in a region. We overcome this challenge with the help of a grouping algorithm that estimates the degree of belonging of hospitals to specific clusters they serve. This algorithm enables the data of hospitals in the same cluster to be analyzed together and therefore, increases the probability of detection and helps reliably identify the location of outbreak. This would help government officials take decisions like closing of schools or colleges in a region to avoid wide scale spread. As additional decision support, we propose continuous monitoring and comparison of the total graph match scores of patients coming to hospitals with flu like symptoms with normal week data to detect any anomalies and also get a reliable estimate of the rate of infection. To summarize, we formulate a comprehensive disaster methodology that can be used for improved detection and containment of infectious diseases like H1N1. Paul et al. (2009) have developed similar models but those are specific to bioterrorist attacks. In addition, they achieve matching through the use of a heuristic as graphs involved in their study are very large. For infectious diseases like H1N1 that involve smaller graphs, optimal solutions can be obtained using exact models proposed in this paper. Accuracy is of utmost importance especially when dealing with diseases like H1N1 because the symptoms are very similar to commonly occurring flu.

The rest of the paper is organized as follows: Section 2 provides the problem background and gives the motivation behind the proposed methodology which is described in Section 3. The operational framework for the modeling logic as well as the details on our nominated algorithms of graph matching and hospital grouping is discussed in section 3. Section 4 demonstrates the methodology and results from its application via a realistically simulated case study. Finally, we provide additional decision making strategies in Section 5 followed by our conclusions and recommendations in Section 6.

## 2. Background

Diseases like H1N1 run the risk of becoming an epidemic if not detected in the earlier stages. Its similarity to common flu, ability to spread via different transmission agents and modes, differences in susceptibility of individuals due to age, health status, pregnancy condition, etc., make its timely containment a considerable challenge. Prior studies focusing on these critical issues include a metapopulation stochastic epidemic model developed by Colizza et. al. (2007). This research focused on the temporal and spatial evolution of a pandemic subject to different levels of infectiousness and initial outbreak conditions (both geographical and seasonal). The objective of their work was to study the worldwide spread of a pandemic and its possible containment at a global level using air travel data. In a similar work, Das et al. (2007) developed a stochastic simulation model on the propagation of an epidemic and then proposed a Markov Decision Process model with a reinforcement learning framework to study possible mitigation strategies. Mathews et al. (2007) used Monte Carlo Markov Chain methods to estimate the clinical attack rate of influenza subject to variation in immunity and asymptomatic nature of infection in the population considered. Ekici et al. (2008) developed disease spread and location models to estimate the food needs and the food distribution network setup during an influenza outbreak. Medlock and Galvani (2009) used a parametric model to determine the optimal allocation of vaccines during an influenza outbreak. The model used survey based contact data and mortality data from influenza pandemics to determine the allocation for five outcome measures: deaths, infections, years of life lost, contingent valuation and economic costs.

Mohtashemi et al. (2006) modeled the short-term dynamic interaction between different subpopulations with respect to an infectious disease using a nonlinear system of difference equations. They used the model to detect anomalous deviations from historically observed events to estimate the rate of infection. Reis et al. (2007) developed a class of epidemiological network models to monitor the relationships among different health-care data streams instead of monitoring the data streams themselves. The extra information present in the relationships between the data streams was used to enhance the detection capabilities of the system and at the same time increase the system's robustness to unpredictable baseline shifts therefore increasing the likelihood of detection of a pandemic. Most of the extant literature in pandemic planning has focused on simulating the disease propagation and estimation of resources required to efficiently handle the outbreak. Though some work has been done to enable early and efficient detection there are still a number of questions associated with detection and containment that are unanswered and under researched (discussed in the introduction section).

Published by De Gruyter, 2012

While there are a lot of lingering questions, most of them can be addressed through a mechanism that is able to detect these infections at an early stage. We propose a graph matching approach to this effect. We use graphs for representing data since it maintains relationships and entities and gives a sense of overall view at a glance to the person. In this research area, we use attributed graphs to represent important data. The nodes represent important information while the properties are represented as attributes. The relationships within nodes are represented as edges, where the extra information can be represented as edge attributes. There has been a vast amount of literature focusing on representation and analysis of graphical data representations (Si et al. (1991), Li et al. (2005), Tong et al. (2007)).

Graph matching can be classified into two categories: exact graph matching and inexact graph matching. Various inexact graph matching algorithms are available in extant literature for solving these problems (Conte et al. (2004), Hlaoui and Shengrui (2002), Romanowski and Nagi (2005), Cesar et al. (2005), Salcedo et al. (2006), Cross et al. (1997), Gold and Rangarajan (1996), Wilson and Hancock (1997), Finch et al. (1998)). These algorithms although are mainly designed to solve large sized graph matching problems. The exact sub-graph matching problems provide exact matches but they have been proven to be NP complete (take exponential time for large sized problems). The patient symptom data however is smaller in size and hence we propose an exact approach towards inexact graph matching. This algorithm finds inexact matches with greater accuracy. If there exists a match between the patient symptom graph and the H1N1 symptom graph, it will be definitely found by this approach. In the past such an approach was considered unthinkable, but with modern day fast processors such problems can now be easily executed.

As a result of the similarity of H1N1 symptoms to any other flu, analysis of patient data of all hospitals in a region collectively can result in false alarms. This challenge could be overcome if there exists a mechanism that could accurately estimate demand distribution for medical care in all the sub regions of a region. For example, if a significantly high number of patients show up with H1N1 symptoms from the same sub region, this would suggest a higher probability of outbreak in that specific sub-region compared to same number of patients presenting to the hospital from the entire region. There are several methods available in literature to study the demand distribution problem. For instance, Cohen and Lee (1985) developed a multinomial logit model to help predict hospital utilization. They considered patient travel time to hospitals, physician specialty, patient characteristics, features attributing to hospital attractiveness to patients etc. in their model. Hunt-Mc cool et al. (1994) used four functional forms with each model incorporating demand for physician outpatient services as a response variable with socioeconomic variables such as income, age,

etc., urbanization and self-reported health indicators as predictor variables. In another study, real population or the demographic structure and population denominators such as tourist load or floating population has been shown to have a profound impact on the healthcare service utilization (Perea-Milla et al. (2007)). Another approach available in literature involves use of gravity models for representing patient flows as a function of patient demand, available resources, indices of accessibility and proximity with an aim to reconfigure emergency services (Congdon (2001)). These models have been used to study the effect of closure of existing sites, addition of new sites, addition of beds at existing sites etc. on the patient flows. Paul et al. (2009) in their study focusing on planning for bioterrorist attacks have proposed a grouping algorithm that uses a gravity model for demand estimation. In this study, we use a variant of their approach to calculate the degree of belonging of a patient cluster or a sub-region to a hospital. Specifically, we propose an MIP model that assigns demand clusters to hospitals based on their proximity and available Emergency Department (ED) capacity. This grouping enables the data of hospitals in the same cluster to be analyzed together. In addition to the hospital grouping approach, our graph matching algorithms are designed such that they assign greater weights to patients with larger number of symptoms of H1N1, age and health status that makes an individual more susceptible to acquiring H1N1, etc., thereby further reducing the chances of false alarms. Next, we present the methodology designed to aid timely and efficient detection and containment of H1N1.

## 3. Methodology

In this section, we first highlight the characteristics of H1N1 that need to be considered for any planning effort especially the ones focusing on detection and containment. Secondly, we present a macro view of the operational framework of our models. Thirdly, we present the graph matching algorithm used to determine patient graph match scores including the dynamic score generating heuristic that is part of this macro model. Finally, we present the hospital grouping algorithm that enables the efficient use of the patient graph match scores.

### 3.1. Symptoms

Symptoms of swine flu are similar to regular flu and include fever, cough, sore throat, runny nose, body aches, headache, chills, and fatigue. Some H1N1 patients have had diarrhea and vomiting. All the infected patients exhibit at least two of these symptoms. Like seasonal flu, pandemic swine flu can cause neurologic symptoms in children. Though rare, such instances can be very severe and often fatal. Symptoms include seizures or changes in mental status (confusion or sudden

cognitive or behavioral changes). Lab tests performed by State Health departments only can definitively show whether a patient has swine flu. For most people, swine flu is a mild illness. Some people get better by staying in bed, drinking plenty of water and taking over-the-counter flu medication. However, some groups of people are more at risk of serious illness if they catch swine flu, and will need to start taking antiviral medication as it is confirmed that they have it. These groups include those that might be suffering from chronic diseases of lung, heart, kidney, liver, neurological disorders (for example, motor neurone disease, multiple sclerosis and Parkinson's disease), immuno suppression (whether caused by disease or treatment), diabetes mellitus etc. Also at risk are patients who have had drug treatment for asthma within the past three years, pregnant women, people aged 65 and older, and young children under five. The other important elements that need to be considered are the duration and severity of illness. For example, if person develops cough and after a couple of days it becomes severe and person has yellow sputum or blood then the person has a greater chance of suffering from H1N1 than a person with a mild cough.

### 3.2. Proposed Modeling Architecture

Continuous processing of hospital patient data by comparing it with symptoms of H1N1 can prove invaluable in detection of an H1N1 outbreak. The basic process that we use to accomplish this objective is as follows. First, incoming patient data at hospital ED is compared with H1N1 symptom template and scores measuring the strength of this match is generated using dynamic graph matching algorithms. These scores are then used to update counters. Since H1N1 symptoms are very similar to commonly occurring diseases like influenza and severity of incoming patient illness could depend on multiple factors (discussed in prior section), only scores that exceed fixed thresholds are considered. Once the counter surpasses a threshold value, alerts are sent to public health officials warning them about an outbreak. Figure 1 presents the basic operational framework of models proposed in this study.

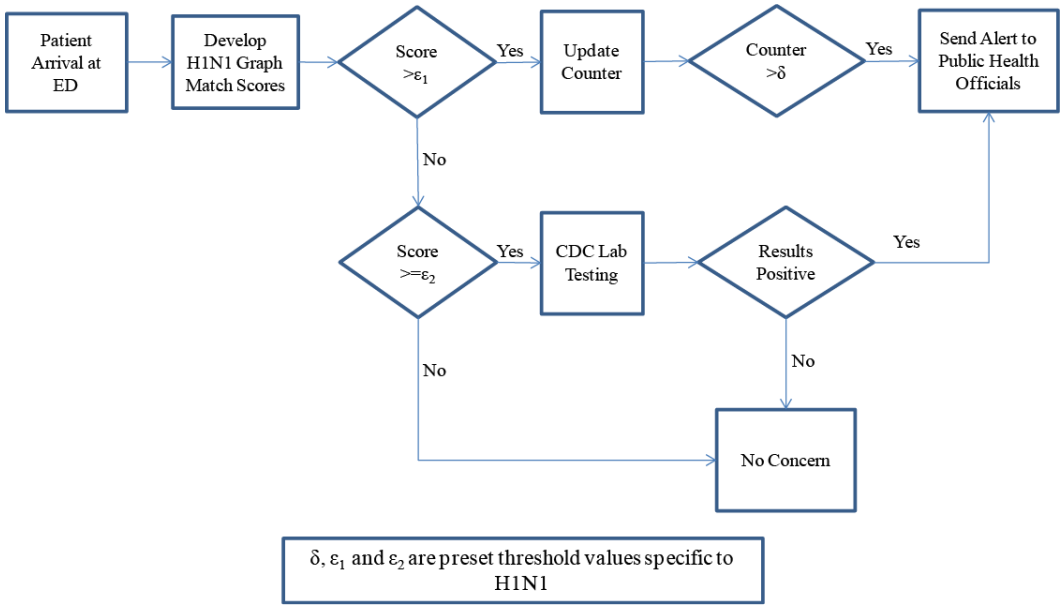



Fig 1. Process Flow adapted from Paul et al. (2009).

### 3.3. Graph matching

Time is one of the most critical components in response to a H1N1 event. The sooner the authorities know about a H1N1 outbreak, the sooner they can curb the spread. To aid in early detection, the most popular technique researchers have developed is syndromic surveillance especially in the field of bioterrorism. The work on syndromic surveillance has a few shortcomings like lack of theoretical evaluation which raises concerns about its validity and performance (Stoto et al. (2004)). Graphical techniques have been around for a while and have been successfully applied to various problem domains. Recent research has demonstrated the ability of graphical techniques in effective handling of streaming dynamic data (Stotz et al. (2009)).

Data representation in graphs is one of the most important aspects of this research. The incoming instantaneous patient data and standard H1N1 patient symptom data is represented as attributed graphs. The attributed graph structure is given as $G = (V, E, A, a_V, a_E)$, where V is a set of nodes, E is a set of arcs, A is a set of node attributes, and $a_V$: $V \rightarrow A$, $a_E$: $E \rightarrow A$. Graph matching is used when the recognition is based on comparison with a model: one graph represents the model or data graph and another one the template or hypothesis where recognition has to be performed. In our research the model is the hospital patient symptom data while the hypotheses are the standard symptoms of H1N1. Figure 2 shows a simple example of a data graph and a template.

Published by De Gruyter, 2012

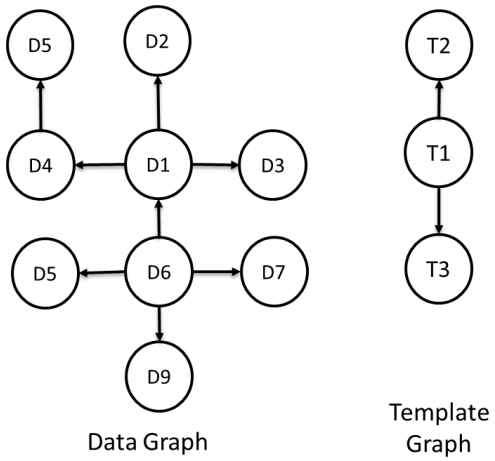



Fig 2. An example of data graph and template.

*3.3.1. Exact Approach towards Inexact Graph Matching*

Many different graph matching techniques can be applied to search for an isomorphism that will exactly match the Target Graph and some portion of the data graph generated from the hospital patient symptom data. However, in asymmetric problem environments with high levels of uncertainty in the observational data, the isomorphism condition can be too strong in many real problems and cannot be expected between both graphs. In these cases, such problems call for inexact graph matching, and the aim of such an approach is to search for the best homomorphism possible. The inexact graph matching problem has been proved to be NP-hard, and therefore heuristic algorithms that provide an approximation to acceptable solutions are required. However, problems involving small graphs can be formulated as an MIP model. This approach yields the best solution that any inexact graph matching approach can aim for. This approach is suitable in our problem scenario since the size of the data graphs is relatively smaller in comparison to those that require heuristics (Paul et al. (2009)). The mathematical formulation is shown as follows:

**Parameters:**
Data graph $G_D = (V_D, E_D)$
Template graph $G_T = (V_T, E_T)$
$V_D$, $V_T$ = Set of nodes in data graph and template graph respectively
$E_D$, $E_T$ = Set of edges in data graph and template graph respectively
$S_{ij}$ = Similarity score between Data Graph node $i$ and template graph node $j$.
$C_{ij,uv}$ = Similarity score between data graph edge $(i, u)$ and template edge $(j, v)$. The value is negative infinity if there is no edge $(i, u)$ or $(j, v)$.

**Variables:**

$x_{ij} = \begin{cases} 1 & \text{if data graph node } i \text{ is associated with template graph node } j \\ 0 & \text{otherwise} \end{cases}$

$y_{ij,uv} =$

$\begin{cases} 1 & \text{if data graph edge } (i,u) \text{ is associated with template graph edge } (j,v) \\ 0 & \text{otherwise} \end{cases}$

where,

$i, u \in V_D$

$j, v \in V_T$

$(i,u) \in E_D$

$(j,v) \in E_T$

**Formulation:**

$Max\left(\sum_i \sum_j S_{ij} x_{ij} + \sum_{(i,u)} \sum_{(j,v)} y_{ij,uv} C_{ij,uv}\right)$

**Subject To:**

$x_{ij} \geq y_{ij,uv} \quad \forall (i, u) \in E_D, (j, v) \in E_T$ **(1)**

$x_{uv} \geq y_{ij,uv} \quad \forall (i, u) \in E_D, (j, v) \in E_T$ **(2)**

$\sum_i x_{ij} = 1 \quad \forall j \in V_T$ **(3)**

$\sum_j x_{ij} \leq 1 \quad \forall i \in V_D$ **(4)**

$x_{ij} = 0 \text{ or } 1 \quad \forall i \in V_D, j \in V_T$ **(5)**

$0 \leq y_{ij,uv} \leq 1 \quad \forall (i, u) \in E_D, (j, v) \in E_T$ **(6)**

The objective function maximizes the similarity score between data graph node *i* and template graph node *j*. Here $S_{ij}$ is the similarity score between data graph node *i* and template graph node *j* and $C_{ij,uv}$ is the similarity score between data graph edge *(i, u)* and template edge *(j, v)*. Constraints 1 and 2 make sure that the edges are assigned if and only if the corresponding nodes are assigned. Constraint 3 ensures that at least one of the data graph nodes is assigned to a given template graph node. Constraint 4 makes sure that no more than one data graph node gets assigned to multiple template graph nodes. Constraints 5 and 6 are bounds on the *x* and *y* variables. In the formulation, *y* is defined as continuous to reduce the complexity of the mathematical model. In an optimal solution *y* will take on 0 or 1 values, by definition of constraints 1 and 2 and positive $C_{ij,uv}$ values make sure that y will never take an intermediate value. The formulation is coded in C++ and executed using CPLEX 9.0, leading mixed integer linear programming software, to solve the problems.

When analyzing patient symptom data that is dependent on a number of dynamic factors, using a static template to match patient data is not meaningful. The template needs to be continuously adjusted to incorporate the effect of factors like difference in severity of symptoms, time since infection, terrorism risk of the region the incoming patients belong to etc. This pre-processing dynamic step

should be completed to make sure that the importance factor for each of the symptoms is taken into consideration. It would also help in differentiating between common symptoms like fever and cough from flu season or influenza outbreak from H1N1 symptoms. We achieve this using a dynamic score updating heuristic discussed in the next subsection.

### *3.3.2. Dynamic Score Updating Heuristic*

Two patients with the same symptoms but showing up at hospitals at different infection stages (severity) cannot be compared to the template in the exact same manner. Similar situation arises if one patient belongs to risk groups mentioned in 3.1 and other does not. To address this, we suggest a heuristic that dynamically updates the scores. One key element of the heuristic is to give variable significance to scores based on the time since the beginning of symptoms, severity of symptoms and whether they belong to a risk group or not. We use average incubation period length and type of the patient i.e. risk group or not, and severity of symptoms patients shows up with as standards to rate the overall severity of the patient. If the length of period since the patient had the symptom is almost close to the average incubation period for H1N1 patients we rate it as more severe than if it was less than the incubation period length. Similarly, belonging to a risk group and/or showing up with severe cough or other flu like symptoms would increase the probability of getting infected compared to those not belonging to the high risk groups and/or those showing up with less severe symptoms. Thus, we follow a multi parameter approach to modify the scores. The heuristic that we use is provided in Figure 3.

Importance calculator (Patient Match Score is modified after patient data arrives, specific to each patient))

SET

$Updated\_Match\_Score = Match\_Score * Severity\_Index*((t_{arrival}-t_{symstart})/t_{inc});$

WHERE:

***Severity_Index*** = weightage based on severity of patient condition and if he/she belongs to the risk group or not, values 1 or 0.

***Risk group*** = {children, pregnant women, elderly, those with chronic conditions, people with chronic condition}, Severity index is set at 1 (if patient belongs to risk group irrespective of patient condition)

$t_{arrival}$ = Time at which patient arrived at the hospital,

$t_{symstart}$ = Time at which symptoms started,

$t_{inc}$ = Average length of incubation period,

Fig 3. Heuristic for graph match score update.

### *3.3.3. Sample Templates*

Figure 4 describes a sample template for H1N1.

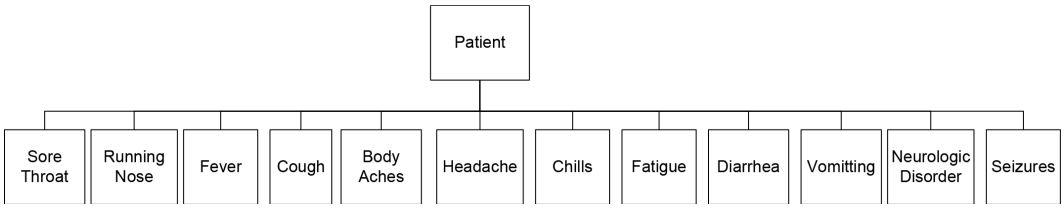



Fig 4. Template for H1N1.

## 3.4. Hospital Grouping

One of the issues with relying merely on graph matching for the base process flow described in Figure 1 is that it does not correlate findings from network of hospitals in the region. This might result in delays in detection and at the same time lead to an inaccurate estimation of spread of infection in a region. However, if we group data of all the hospitals in a region, it might lead to a lot of false alarms. This is partly because hospitals in a region might only serve populations from a certain sub region. In this study, we use a variant of the approach developed by Paul et al. (2009) to group hospitals that serve the same demand clusters. Specifically, we develop an MIP model that determines the allocation of a demand from a population cluster to different hospitals. This grouping enables the data of hospitals in the same cluster to be analyzed together and increases the probability of detection (minimizes false alarms) of an outbreak. In addition, it also gives a much better estimate of total infected people in a region. This piece of information is not only important for containment but also in determining the rate of infection and the necessary resources therefore to control an outbreak.

As patients arrive at a hospital, their address information (cluster that they belong to) is recorded along with other details. If the dynamic graph matching algorithm indicates that the patient has symptoms similar to H1N1, counters specific to the cluster are updated. This is based on the assumption that patient belonging to a certain cluster arrives at a hospital belonging to the grouping identified by the gravity model as serving that particular cluster. When counter values surpass the threshold value, the public health officials are notified of a possible outbreak. This helps officials determine which region needs more attention with regard to containment. The values of the counters represent the approximate size of infected population. The graph match scores at the same time indicate the severity of infection and also the size of infected population. The

Published by De Gruyter, 2012

higher the number of infected and severity of infection, the larger the total graph match scores of patients on any day will be.

The region under study is divided into clusters using k-means clustering algorithm (MacQueen (1967)). The k-means clustering algorithm works as follows:

- The data-set is partitioned into *C* clusters and the data points are randomly assigned to the clusters such that each cluster has roughly the same number of data points. The centroid of a cluster is determined and it can be assumed to be the mean of all the data points in the cluster.
- For each data point:
    - Calculate the distance from the data point to each cluster.
    - If the data point is closest to its own cluster, leave it in the same cluster. Otherwise, move it into the closest cluster.
- Repeat the above step until a complete pass through all the data points result in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

Since there is a possibility that an inappropriate value of k could lead to incorrect clustering, we used silhouette index to determine the best value of k, the total number of clusters. Silhouette refers to a method of interpretation and validation of clusters generated using a clustering algorithm like k-means. The technique provides a succinct graphical representation of how well each object lies within its cluster (Rousseeuw (1987)). It is a based on a comparison of average dissimilarity of a data point with all the other data within the same cluster to that of data belonging to all the remaining clusters. Let $a(i)$ and $b(i)$ be these two measures respectively for a data point $i$. $a(i)$ can be interpreted as how well matched $i$ is to the cluster it is assigned (the smaller the value, the better the matching). $b(i)$ is selected as the minimum value out of those obtained for each remaining cluster that $i$ is not a part of. $b(i)$ can be interpreted as the neighboring cluster of $i$ as it is, aside from the cluster $i$ is assigned, the cluster $i$ fits best in. The silhouette index using $a(i)$ and $b(i)$ can be defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \textbf{(7)}$$

This can be simplified as follows:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & if\ a(i) > b(i) \end{cases} \quad \textbf{(8)}$$

From the above definition it is clear that

$$-1 \leq s(i) \leq 1 \quad \textbf{(9)}$$

As $a(i)$ is a measure of how dissimilar $i$ is to its own cluster, a small value means it is well matched. Furthermore, a large $b(i)$ implies that $i$ is badly matched to its neighboring cluster. A value of $s(i)$ close to one indicates that the datum is appropriately clustered. If $s(i)$ is close to negative one, then it indicates that $i$ would be more appropriate if it was clustered in its neighboring cluster. An $s(i)$ near zero would mean that the datum is on the border of two natural clusters. The average $s(i)$ of a cluster is a measure of how tightly grouped all the data in the cluster are. Thus the average $s(i)$ of the entire dataset is a measure of how appropriately the data has been clustered. Too few or too many clusters highlight a poor choice of k and therefore some of the clusters will display much narrower silhouettes than the rest. Thus the silhouette index is a powerful tool for determining the natural number of clusters within a dataset. Once the clusters are developed, we use the following formulation to group hospitals that serve the same cluster.

**Notations:**
$i$ = demand cluster,
$j$ = hospital facility,
$I$ = set of demand clusters,
$J$ = set of hospital facilities,
$s$ = fictitious site (serves as a reserve location in case the hospitals in the region cannot serve the demand).
**Parameters:**
$c_i$ = demand at cluster $i$,
$w_j$ = ED capacity at hospital $j$,
$w_j = Round((92*B_j+2267)/EDcapacity\_scaler+OR_j+3))$ (Yi et al. (2010)),
$B_j$ = number of beds in hospital $j$,
$EDcapacity\_scaler$ = constant (depends on the size of the hospital),
$OR_j$ = number of operating rooms in hospital $j$,
$d_{ij}$ = distance from cluster $i$ to hospital $j$.
**Decision variables:**
$y_{ij}$= proportion of demand cluster $i$ that is allocated to hospital $j$.
**Formulation:**

$$Min \sum_{i=1}^{I} \sum_{j=1}^{J \cup s} c_i y_{ij} d_{ij}$$

s.t.

$$\sum_{i=1}^{I} c_i y_{ij} \leq w_j, \quad \forall j \epsilon J,$$

**(10)**

$$\sum_{j=1}^{J \cup s} y_{ij} = 1, \forall i \epsilon I,$$
$$0 \leq y_{ij} \leq 1$$

**(11)**

Constraint 10 makes sure that total population from all the clusters allocated to a hospital is less than equal to the capacity of the capacity. Constraint 11 ensures that all clusters are served. Once we obtain the $y_{ij}$ values from the above formulation, we create a counter that groups all the hospitals that serve a particular cluster *i*. Each hospital could update different counters depending on how many clusters they serve. In order to avoid any errors we always keep track of patient information with regard to the cluster they belong to before updating the counter. We recommend continuous scanning of graph match scores with normal week data to determine any major deviations that indicate a possible pandemic. It is also beneficial to compare data from consecutive days and those within a fixed time window in order to detect any abnormal trends. To this effect, comparison of means and skewness of graph math scores would be really beneficial to public health officials in their planning and preparedness. In some of the papers dealing with detection for example, Mohtashemi et al. (2006), number of infected is used to determine the rate of infection. This might not be the best way to estimate infection rate because two days can have the same number of infected but differ with regard to type of patients and severity of infection. This difference could lead to totally different rates of infection in the affected region. Moreover, our hospital grouping approach should help improve the accuracy in estimating infection than prior detection models that analyze a single hospital data or group all the hospital data for their decision making. We demonstrate the relevance of the continuous scanning in our case study section with a small example.

## 4. Case Study

For the purposes of our case study, we simulated 1000 cases of patients that arrived at the emergency department with various symptoms. We can deploy this approach at all levels. We have shown a city level model (for demonstration purposes) and hence this model can be applied at the state level too. We suggest

this model to be applied at state level at maximum, because the area of infection at outbreak will usually not span more than the area of a typical state. Another reason is health jurisdictions would prevent officials from taking any actions beyond their state boundaries. We chose Buffalo, NY as the region of focus for our case study. We use the demographic information available on Buffalo, NY from the IDcide - Local Information Data Server datasource (IDcide (2011)) to generate the case data. One of the key pieces of information we used is included in Table 1.

Tab 1. Demographic details for Buffalo, NY

| | **Men** | **Women** | **Total** | |
|---|---|---|---|---|
| **Under 20** | 15% | 14% | 29% | 290 |
| **20 to 40** | 14% | 16% | 30% | 300 |
| **40 to 60** | 11% | 13% | 24% | 240 |
| **Over 60** | 7% | 10% | 17% | 170 |
| Total | 47% | 53% | 100% | 1000 |

Each patient record included the following information: latitude, longitude or zip code (indicating location), symptoms they come in with, age, pregnancy status (yes or no), chronic conditions (yes or no). Based on the silhouette index and application of k-means clustering algorithm, we were able to assign the 1000 patients to 20 clusters (0.526). We noticed that the silhouette index improved with larger number of clusters but then it lead to too few patients per cluster. In addition, a silhouette index of 0.526 indicated a good grouping. So we chose 20 as the final number of clusters. This information was then applied to the MIP hospital grouping model. The information we collected for hospitals that were part of our case study is provide in Table 2.

Tab 2. Hospitals considered in the case study.

| **Hospital** | **Latitude** | **Longitude** | **No of Beds** | **No of OR** |
|---|---|---|---|---|
| Buffalo General Hospital | 42.9008 | -78.8657 | 1558 | 58 |
| Erie County Medical Center | 42.9271 | -78.8292 | 1137 | 12 |
| Mercy Hospital of Buffalo | 42.8473 | -78.8127 | 349 | 16 |
| Kenmore Mercy hospital | 42.9776 | -78.8824 | 184 | 8 |

The region considered for the problem along with the hospital and cluster locations is shown in Figure 5. It was generated using Arc GIS Explorer.

Published by De Gruyter, 2012

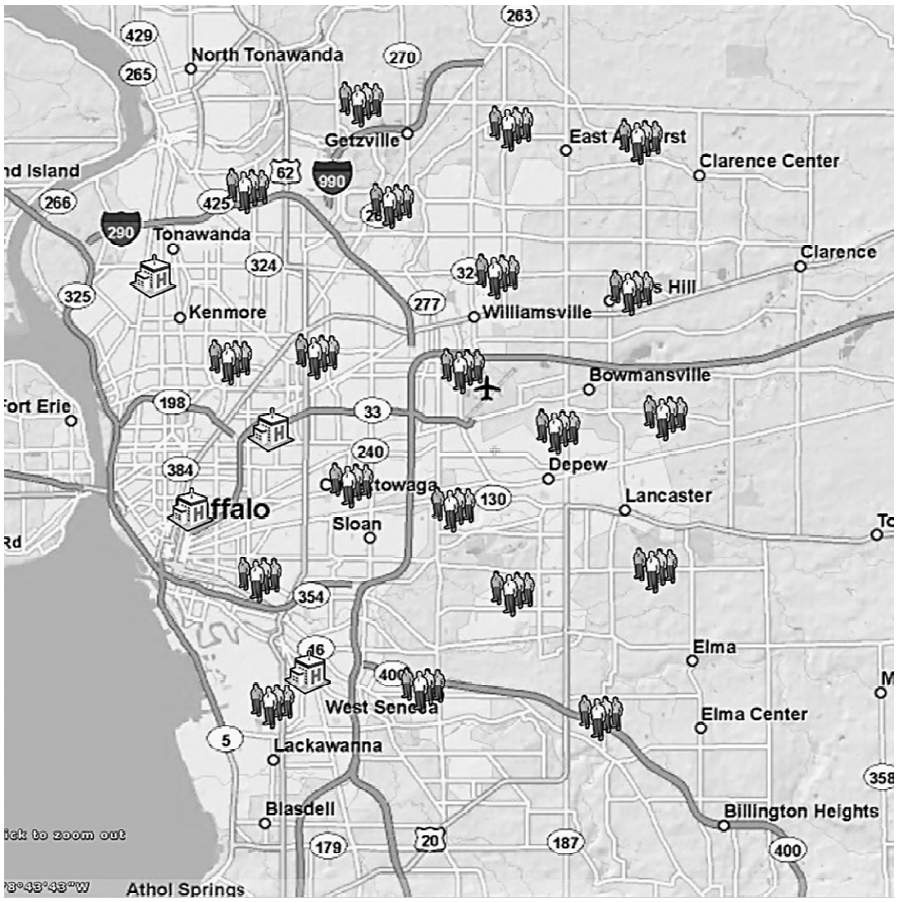



Fig 5. Problem Study Region.

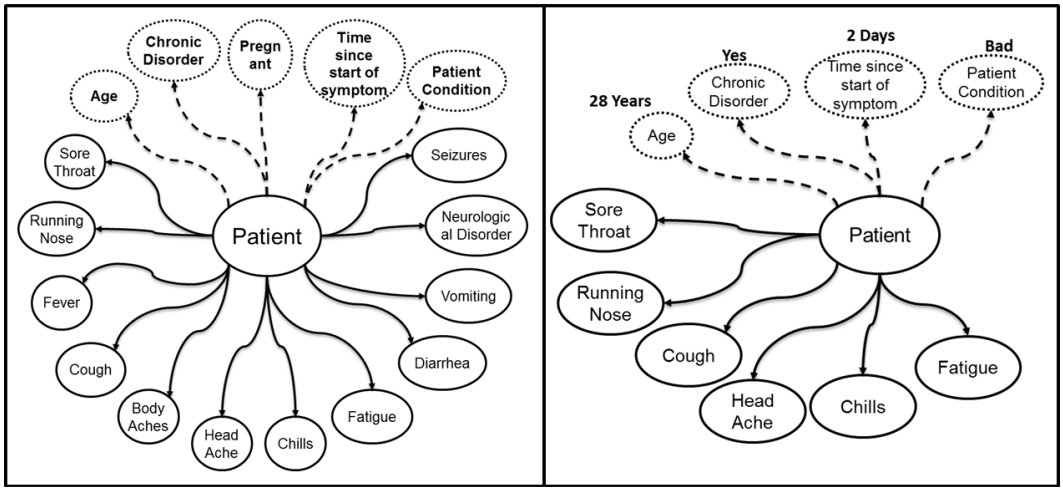
The 1000 simulated patient cases and the hospital information in Table 2 were used as inputs to the hospital grouping MIP model (discussed in the section 3). The counters that were created as outputs from the grouping model are given in Table 3. The first subscript represents the cluster and the second one represents the hospital it is served by. As can be noted, the clusters 4 and 10 are served by multiple counters and therefore counter values for these clusters need to be grouped together for decision analysis.

Tab 3. Counters specific to clusters and hospitals.

| Cluster | Hospital | Counters |
|---|---|---|
| 1 | 1 | C1_1 |
| 2 | 1 | C2_1 |
| 3 | 2 | C3_2 |
| 4 | 1 | C4_1 |
| 4 | 2 | C4_2 |
| 4 | 4 | C4_4 |
| 5 | 2 | C5_2 |
| 6 | 2 | C6_2 |
| 7 | 1 | C7_1 |
| 8 | 1 | C8_1 |
| 9 | 4 | C9_4 |
| 10 | 1 | C10_1 |
| 10 | 3 | C10_3 |
| 11 | 1 | C11_1 |
| 12 | 1 | C12_1 |
| 13 | 2 | C13_2 |
| 14 | 2 | C14_2 |
| 15 | 1 | C15_1 |
| 16 | 1 | C16_1 |
| 17 | 2 | C17_2 |
| 18 | 3 | C18_3 |
| 19 | 1 | C19_1 |
| 20 | 1 | C20_1 |

Similarly, the patient information on symptoms, age, etc., was used as input to the graph matching model to obtain their scores. The patient's symptom and H1N1 template graphs are stored as XML files. The exact approach to graph matching along with the dynamic score updating heuristic was coded in C++ while the optimization was run using CPLEX 9.0 interface. Examples of patient symptom graph and the template graph against which they were compared are provided in Figure 6.

Published by De Gruyter, 2012




a) Template Graph b) Sample Patient Graph

Fig 6. Graphs depicting the template and a randomly selected patient.

The distribution of scores obtained using the graph matching algorithm and dynamic graph generation heuristic can be noted from the histogram shown in Figure 7.

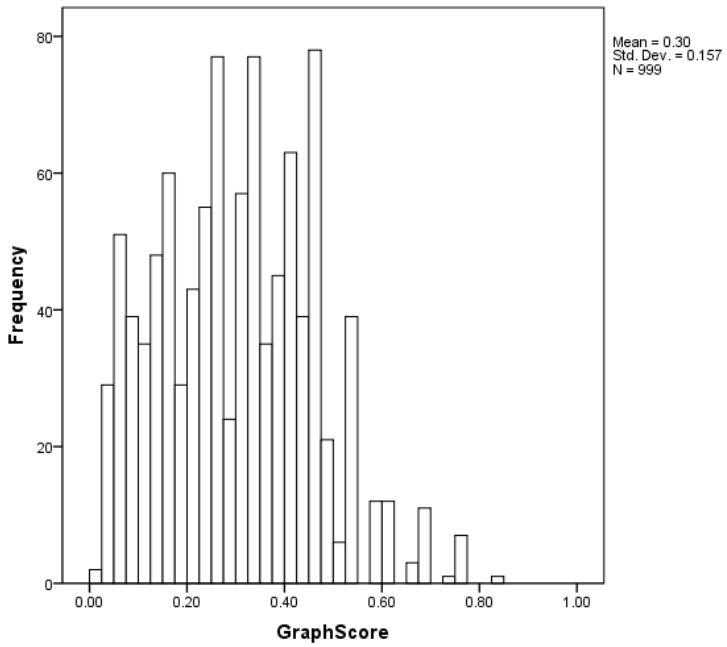



Fig 7. Histogram of graph match scores.

Once the scores were developed, the next step was to determine the threshold value for these scores i.e. $\varepsilon_1$ and $\varepsilon_2$ as per Figure 1. These values are extremely important as inappropriate values could lead to lot of false positives and negatives. The threshold value $\varepsilon_1$ would be a limit such that if patient graph match score is above it, it indicates patient has H1N1. The following factors were considered in deciding value of $\varepsilon_1$ and $\varepsilon_2$. a) An infected patient would at the least have two of the symptoms resulting in a score at least equal to 2/12 or 0.167 b) A perfect match is never possible (score of 1). c) Even a score above 0.6 would be too difficult for an infected patient as it would require a patient to have more than six of the 12 symptoms (based on discussions with physicians). Therefore, a value of 0.4 was selected for $\varepsilon_1$ as it lies approximately in the center of the interval [0.1667, 0.6] and a value of 0.1667 was selected for $\varepsilon_2$. Any score above $\varepsilon_2$ and less than $\varepsilon_1$ would require additional medical testing for validation of patient condition.

In addition to the threshold values, we also had to determine $\delta$, a threshold for the counter values (Figure 1). Based on our discussions with physicians, we selected a value of 5 for $\delta$. Once these values were decided, the scores were then used to update the appropriate counters. We include two sets of counters in our results. The first set of values indicates the number of infected in each cluster. The second set of values show the patient cases that need additional testing for validation. The second piece of information could be used by public health officials to determine resources they need or would need to provide to the hospital. The counter values indicating the size of infection is shown in Figure 8. Similarly, the size of population belonging to each cluster that would need further testing for validation is shown in Figure 9. These pieces of information would prove invaluable for public health officials as well as hospitals in the Buffalo region.

Published by De Gruyter, 2012

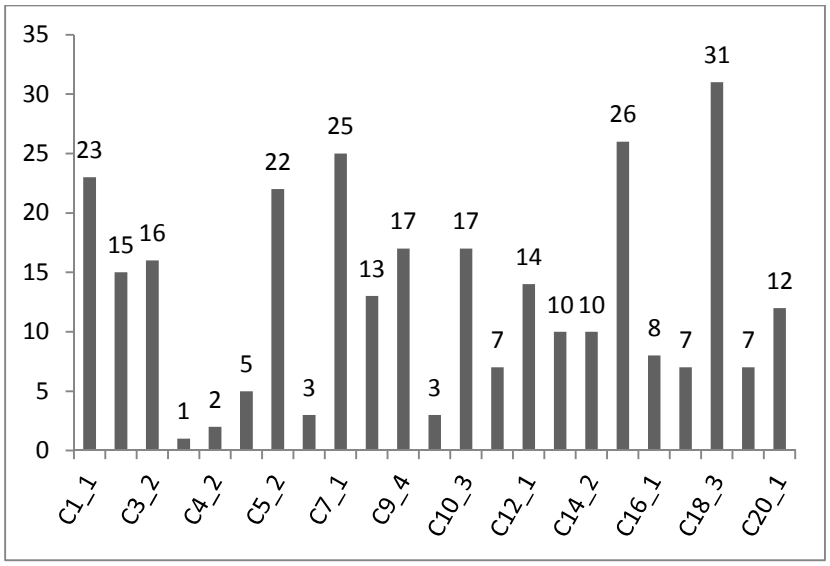



Fig 8. Size of infection (by cluster and hospital they get served)

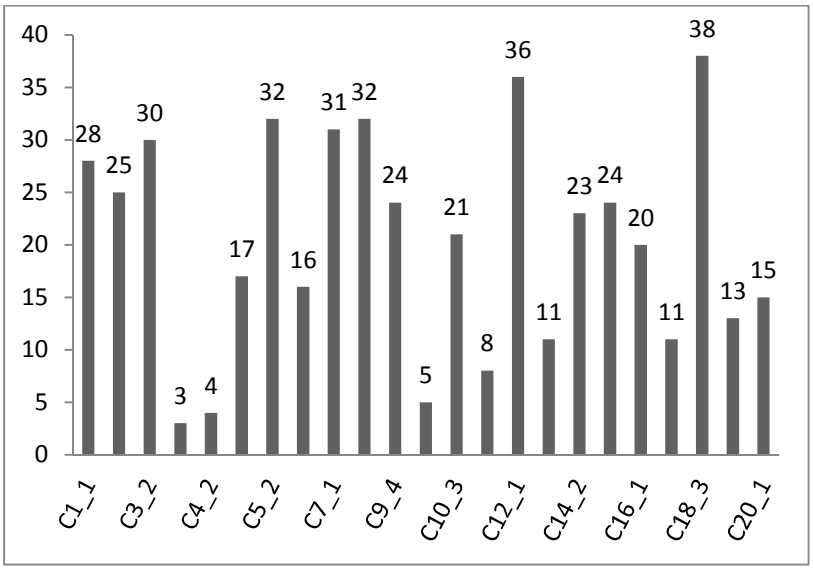


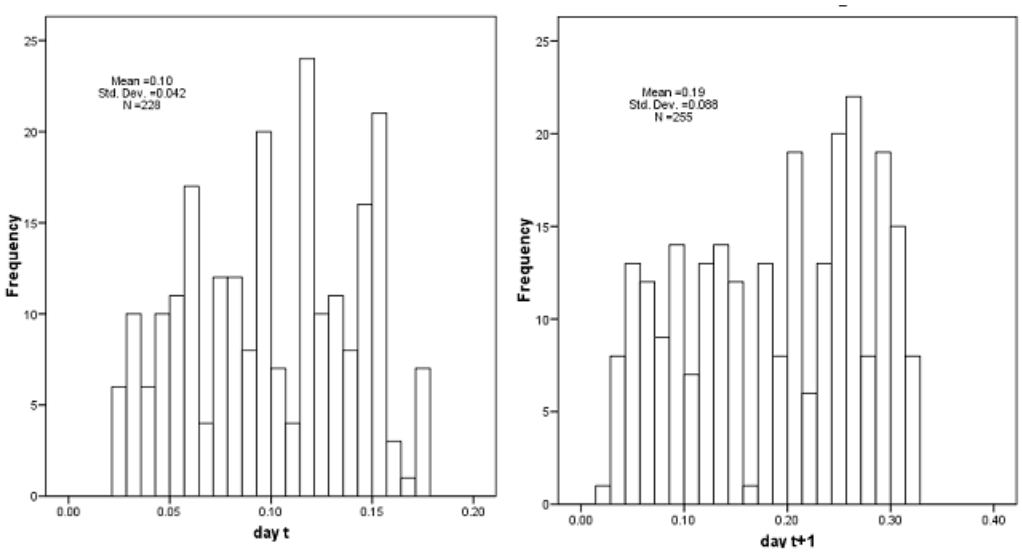

Fig 9. Size of population needing testing (by cluster and hospital they get served)

Alerts were sent out to the public health officials for all the counters that were above $\delta$. Thus based on the score analysis, it was possible to determine the location and size of infection.

## 5. Additional Decision Making Strategies

It is possible that the values of counters do not exceed the threshold in the initial stages of infection due to very small size of infected population or very low severity of infection. However, with passage of time, the symptoms might become severe thereby elevating the graph match scores and leading to a larger infected population. One of the ways to detect infection in the early stages without totally relying on counters is to check for anomalies in graph match scores through continuous comparison of scores from consecutive days and scores of days within a certain moving time window. This is due to the possibility of scores being insignificant on consecutive days but being different from each other when compared with multiple days since diseases like H1N1 generally have an incubation period. This can be achieved through comparison of average scores for consecutive days using t-test and post hoc comparison tests and comparison of skewness of data for multiple days (more than two) within a predefined time window. The time window for infectious diseases could be set equal to its incubation period which for H1N1 is approximately seven days. Any significant change in the mean, skewness of the distribution during such comparisons could be used to alert public health officials. We demonstrate one possible scenario through comparison of average graph match scores from two random days (Figure 10). As can be noted, not only has the average changed significantly from 0.10 to 0.19 (p-value $<0.01$) but so has the skewness. Such major changes in the variability of score data or average could provide useful information to public health officials.

Published by De Gruyter, 2012




a) Distribution for day t b) Distribution for day t+1

Fig 10. Comparison of scores from two consecutive days.

The decision support tools developed in this paper (summarized in Figure 11) can assist the public health officials with the following:

a) Provide timely alerts to public health officials for implementing precautionary actions without singularly relying on lab test outcomes.
b) Identify the location and size of H1N1 outbreak. This will make quarantine and/or containment measures more effective.
c) Help prioritize the population that requires vaccination or increased medical attention in a hospital.
d) Assist in prioritization of patient samples requiring lab testing by identifying patients that have a higher probability of being infected with H1N1. This is important because the existing protocol requires hospitals to send samples for lab testing for anyone that comes with flu like symptom. Random testing might lead to delays in detection of an outbreak.

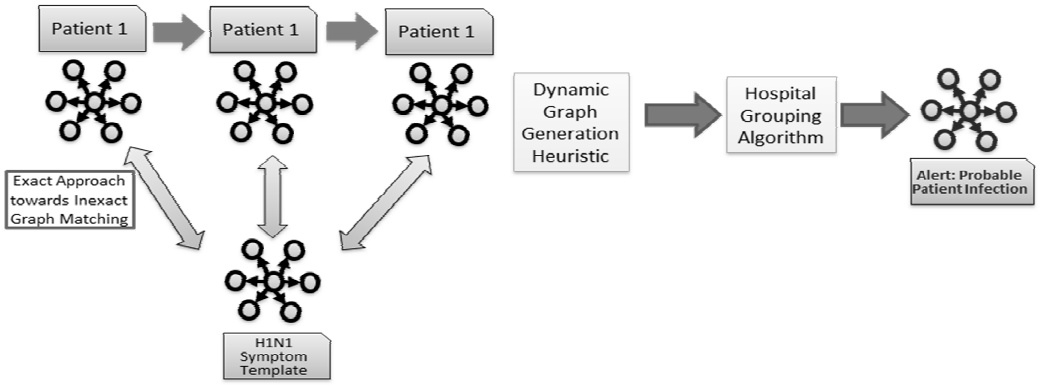



Fig 11. Process flow of the H1N1 detection algorithm.

## 6. Conclusions

This research spans the areas of situation assessment and threat estimation wherein the situation is perceived through the analysis and understanding of the patient symptoms and the area of infection. Specifically, we propose dynamic graph matching algorithms that compare patient data to H1N1 symptom template. Based on the graph matching scores generated, threat alerts are sent to public health officials in case of an outbreak. Hospital grouping models along with graph matching algorithms assist in identifying and isolating the infected areas. To summarize, they are able to improve the precision of syndromic surveillance and mitigate the source of infection.

Applications of the proposed research extend well beyond H1N1 planning. With minor modifications, these models could be used for epidemics such as SARS, bird flu, future variations of H1N1 virus, etc., as well as early detection and isolation of terrorist attacks. It can also aid in the optimal allocation of resources to prevent and mitigate the effects of chronic diseases like Diabetes, HIV/AIDs, etc. Specifically, the models can help estimate the size and characteristics of the affected population in a region. In the current study, due to data availability issues, we had to rely on a simulated data set. In the future extensions of this research effort aimed at providing implementation guidelines for our models to prospective users, we plan to use real hospital data to increase the user confidence. In addition, we plan to partner with medical personnel to develop more sophisticated mechanisms to estimate the preset threshold values currently based on H1N1 symptom templates.

Published by De Gruyter, 2012

## References

1. World Health Organization, “Pandemic H1N1 2009 – update 112”, 2010 [cited 2010 September 18]; Available from: http://www.who.int/csr/don/2010_08_06/en/.
2. Medicine Net. Swine Flu, 2009 [cited 2010 September 14]; Available from: http://www.medicinenet.com/swine_flu/page4.htm#prevention.
3. Hagenaars TJ, van Genugten MLL, Wallinga J., “Pandemic Influenza and Health Care Demand: Dynamic Modeling”, International Congress Series. 2004. p. 235.
4. Longini IM, Jr., Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DAT, “Containing Pandemic Influenza at the Source” Science, 2005; 309(5737):1083-7.
5. Feighner BH, Murphy SP, Skora JF, “The Pandemic Influenza Policy Model, a Planning Tool for Military Public Health Officials”, John’s Hopkins APL Technical Digest, 2008; 27(4 ):374-82.
6. Pan-InfORM, “Modelling an Influenza Pandemic: A Guide for the Perplexed”, Canadian Medical Association. 2009; 181:3-4.
7. Mohtashemi M, Szolovits P, Dunyak J, Mandl KD, “A Susceptible-Infected Model of Early Detection of Respiratory Infection Outbreaks on a Background of Influenza”, Journal of Theoretical Biology, 2006; 241(4):954-63.
8. Reis BY, Kohane IS, Mandl KD, “An Epidemiological Network Model for Disease Outbreak Detection”, PLoS Med, 2007; 4(6):e210.
9. Paul JA, Sambhoos K, Hariharan G., “Using Dynamic Graph Matching and Gravity Models for Early Detection of Bioterrorist Attacks”, Journal of Homeland Security and Emergency Management, 2009; 6(1-17).
10. Colizza V, Barrat A, Barthelemy M, Valleron A-J, Vespignani A., “Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions”, PLoS Medicine, 2007; 4:0095-110.
11. Das TK, Savachkin A, Zhu Y., “A Large Scale Simulation Model of Pandemic Influenza Outbreaks for Development of Dynamic Mitigation Strategies”, IIE Transactions, 2007; 40(9):893-905.
12. Mathews JD, McCaw CT, McVernon J, McBryde ES, McCaw JM., “A Biological Model for Influenza Transmission: Pandemic Planning Implications of Asymptomatic Infection and Immunity”, PLoS ONE. 2007; 2(11):e1220.
13. Ekici A, Keskinocak P, Swann JL, “Pandemic Influenza Response” Winter Simulation Conference; 2008; Miami, Florida.
14. Medlock J, Galvani AP., “Optimizing Influenza Vaccine Distribution” Science, August 20, 2009:1175570.

15. Si Wei L, Ren Y, Suen CY., "Hierarchical Attributed Graph Representation and Recognition of Handwritten Chinese Characters", Pattern Recognition. 1991; 24(7):617-32.
16. Li Y, Blostein D, Abolmaesumi P, "Asymmetric Inexact Matching Of Spatially-Attributed Graphs", 2005; Heidelberg, D-69121, Germany: Springer Verlag.
17. Tong H, Faloutsos C, Gallagher B, Eliassi-Rad T, "Fast Best-Effort Pattern Matching in Large Attributed Graphs", 2007, New York, NY 10036-5701, United States: Association for Computing Machinery.
18. Conte D, Foggia P, Sansone C, Vento. M., "Thirty Years of Graph Matching in Pattern Recognition", International Journal of Pattern Recognition and Artificial Intelligence, 2004; 18(3):265-98.
19. Hlaoui A, Shengrui W, "A New Algorithm for Inexact Graph Matching", Proceedings of 16th International Conference on Pattern Recognition; 2002.
20. Romanowski CJ, Nagi R, "On Comparing Bills Of Materials: A Similarity/Distance Measure for Unordered Trees", IEEE Transactions on Systems, Man and Cybernetics, Part A, 2005; 35(2):249.
21. Cesar JRM, Bengoetxea E, Bloch I, Larranaga P., "Inexact graph matching for model-based recognition: Evaluation and comparison of optimization algorithms", Pattern Recognition, 2005; 38(11):2099.
22. Salcedo-Sanz S, Xu Y, Yao X., "Hybrid Meta-Heuristics Algorithms for Task Assignment in Heterogeneous Computing Systems", Computers & Operations Research. 2006; 33(3):820.
23. Cross ADJ, Wilson RC, Hancock ER., "Inexact Graph Matching Using Genetic Search" Pattern Recognition. 1997; 30(6):953.
24. Gold S, Rangarajan A., "A Graduated Assignment Algorithm for Graph Matching" IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996; 18(4):377.
25. Wilson RC, Hancock ER., "Structural Matching by Discrete Relaxation" Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1997; 19(6):634.
26. Finch AM, Wilson RC, Hancock ER., "Symbolic graph matching with the EM algorithm", Pattern Recognition. 1998; 31(11):1777.
27. Cohen MA, Lee HL, "The determinants of spatial distribution of hospital utilization in a region Medical Care", 1985; 23(1):27-38.
28. Hunt-McCool J, Kiker Bf, Chung Y., "Estimates of the demand for medical care under different functional forms", Journal of Applied Econometrics. 1994; 9:201-18.

29. Perea-Milla E, Pons SM, Rivas-Ruiz1 F, Gallofre A, Jurado EN, Ales MAN, "Estimation of the Real Population and its Impact on the Utilization of Healthcare Services in Mediterranean Resort Regions: An Ecological Study", BMC Health Services Research 2007; 7:13.
30. Congdon P., "The Development of Gravity Models for Hospital Patient Flows Under System Change: A Bayesian Modeling Approach", Health Care Management Science, 2001; 4:289-304.
31. Stoto MA, Schonlau M, Mariano LT., "Syndromic Surveillance: Is it Worth the Effort?" chance. 2004; 17(1):19-24.
32. Stotz A, Nagi R, Sudit M, "Incremental Graph Matching for Situation Awareness" Information Fusion; 2009; Seattle, WA.
33. MacQueen JB, "Some Methods for Classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press. 1967; 1:281-97.
34. Rousseeuw PJ., "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", Computational and Applied Mathematics. 1987; 20(53–65).
35. Yi P, George SK, Paul JA, Lin L., "Hospital Capacity Planning for Disaster Emergency Management", Socio-Economic Planning Sciences. 2010; 44(3): 151-160.
36. IDcide - Local Information Data Server, 2011 http://www.idcide.com/citydata/ny/buffalo.htm.