Journal of Biomedical Informatics 93 (2019) 103161

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Social network analysis for better understanding of influenza

Branimir Ljubic, Djordje Gligorijevic, Jelena Gligorijevic, Martin Pavlovski, Zoran Obradovic*

Temple University, Center for Data Analytics and Biomedical Informatics (DABI), Philadelphia, PA, USA

ARTICLE INFO

Keywords:
Influenza
Social networks
Heatmaps
Infectious disease spreading
Demographics
Biomedical informatics

ABSTRACT

Introduction: The objective of this study is to improve the understanding of spatial spreading of complicated cases of influenza that required hospitalizations, by creating heatmaps and social networks. They will allow to identify critical hubs and routes of spreading of Influenza, in specific geographic locations, in order to contain infections and prevent complications, that require hospitalizations.

Material and methods: Data were downloaded from the Healthcare Cost and Utilization Project (HCUP) – SID, New York State database. Patients hospitalized with flu complications, between 2003 and 2012 were included in the research (30,380 cases). A novel approach was designed, by constructing heatmaps for specific geographic regions in New York state and power law networks, in order to analyze distribution of hospitalized flu cases.

Results: Heatmaps revealed that distributions of patients follow urban areas and big roads, indicating that flu spreads along routes, that people use to travel. A scale-free network, created from correlations among zip codes, discovered that, the highest populated zip codes didn't have the largest number of patients with flu complications. Among the top five most affected zip codes, four were in Bronx. Demographics of top affected zip codes were presented in results. Normalized numbers of cases per population revealed that, none of zip codes from Bronx were in the top 20. All zip codes with the highest node degrees were in New York City area.

Discussion: Heatmaps identified geographic distribution of hospitalized flu patients and network analysis identified hubs of the infection. Our results will enable better estimation of resources for prevention and treatment of hospitalized patients with complications of Influenza.

Conclusion: Analyses of geographic distribution of hospitalized patients with Influenza and demographic characteristics of populations, help us to make better planning and management of resources for Influenza patients, that require hospitalization. Obtained results could potentially help to save many lives and improve the health of the population.

1. Introduction

Infectious diseases, like influenza, can have devastating consequences on populations. Influenza is associated with substantial morbidity and mortality [1]. In addition to clinical impact, Influenza has significant economic impact. Starting from 2010, each year CDC estimates the burden of influenza. The burden of influenza disease in the United States can vary widely and is determined by a number of factors including the characteristics of circulating viruses, the timing of the season, how well the vaccine is working to protect against illness, and how many people got vaccinated. CDC [2] estimates that influenza has resulted in between 9.3 million and 49.0 million illnesses, between 140,000 and 960,000 hospitalizations and between 12,000 and 79,000 deaths annually since 2010. The estimated number of flu illnesses during the 2017–2018 season was 49 million, flu hospitalizations – 960,000, and flu deaths – 79,000.

Standard medical treatments and vaccines are often not sufficient to stop flu infections. Equally important, is to understand how the pathogen spreads in the population [3]. Understanding the nature of human contact patterns is crucial for predicting future pandemics and developing effective control measures [4]. Explorations of spatio-temporal spread are also, very important in order to explain, are Influenza infections more spatially synchronized and widespread in populous highly connected areas, compared to smaller, more isolated ones [5]. Reasons for geographical trends of spreading of Influenza could be explained in terms of population size, connectivity and demographics. Understanding the spatio-temporal spread of infectious disease is important both for design of control strategies and to deepen fundamental

* Corresponding author at: Temple University, SERC Building, Center for Data Analytics and Biomedical Informatics (DABI), 1925 N 12th St, Philadelphia, PA 19122, USA.

E-mail addresses: branimir.ljubic@temple.edu (B. Ljubic), gligorijevic@temple.edu (D. Gligorijevic), jelena.stojanovic@temple.edu (J. Gligorijevic), tuh27103@temple.edu (M. Pavlovski), zoran.obradovic@temple.edu (Z. Obradovic).

https://doi.org/10.1016/j.jbi.2019.103161
Received 5 December 2018; Received in revised form 17 March 2019; Accepted 29 March 2019
Available online 30 March 2019
1532-0464/

knowledge about the interaction between infectious diseases dynamics and spatial mixing of the population [6]. It's important to locate geographic hotspots (so called hubs) for Influenza infections [7]. Analyses of geographic distribution of Influenza and demographic characteristics of geographic hotspots, help us to make better planning of hospital resources for complicated cases of the flu and better management of health-care systems. Environmental factors, population sizes and, demographics, are major determinants in disease spread and potential complications, which can result in admissions to hospitals. Influenza causes many complications, that can worsen the disease, require hospitalizations and complicate outcomes. Respiratory, cardiovascular, digestive system and other complications have been studied [8–10].

The objective of this research is to develop a novel method, that leverages options of heatmaps and network science in explanations of spatial distribution of patients with complications of Influenza. The goal is to visualize complicated flu cases throughout the particular geographic region. Public health experts, doctors, and other medical scientists, by using the results of the visualization tools, like heatmaps, could rapidly recognize, how flu cases are distributed, and how they expand geographically. This knowledge would help them plan resources for earlier detection of flu and reduce the future impact of influenza [11]. Further, we will utilize network science options to analyze correlations among zip codes of the NY state, considering numbers of hospitalized flu cases over the 10 years period. The goal is to show observed and calculated correlations in the network, as nodes (zip codes) linked, based on strength of correlations. That will allow calculation of nodes degrees, with the objective to find the most connected nodes and hubs, based on results of these network centrality measures. Adequate measures should be applied to isolate (treat) zip codes that represent discovered hubs in order to decrease the number of complicated flu cases in the future. The final objective is to present results of our novel method to health professionals and researchers, which will help them to plan adequate resources to contain flu infection and prepare appropriate hospital resources for patients with complications. Results could potentially save many lives and improve the health of the population. The experiments were conducted on the state of New York data, but the proposed method is scalable and could easily be generalized to any other geographic region in the U.S. and all over the world.

1.1. Background and significance

According to Centers for Disease Control (CDC), seasonal influenza infects approximately 5–20% of the U.S. population every year [12]. Connections between Influenza and networks is a well-studied topic, that dates back to the mid-1980s and many papers describe the association between influenza and networks. To assess the influence of network effects, the predictions were compared, from the detailed network model, consisting of fixed contacts of known weights, to several simplified alternatives [4]. Spatial spread of influenza infections and geographic transmission hubs were analyzed and described in recent publications [5–7]. Numerous research studies evaluated the risk and development of complications associated with influenza virus infections [8,9]. Many of those complicated cases require a hospital treatment. Researchers described influenza-associated critical illness hospitalizations [10]. Demographic factors associated with influenza A(H1N1) infection have been studied [13,14]. Many authors assessed the network configuration, network stability, and changes in risk configuration and risk behavior, using social network analysis and visualization techniques. The evolving science of social networks has evident potential to help researches to explain the spread of infectious diseases [15]. Gligorijevic and colleagues studied the importance of the confidence of predictions in longer-term forecasting in health and climate domains [16]. They presented an effective novel iterative method developed for Gaussian structured learning models, for propagating uncertainty in temporal graphs, by modeling noisy inputs (most of inputs in the field of infectious diseases). Good planning of hospital resources will also need a prediction of length of hospital stay, for individual patients, in addition to prediction of numbers of potential hospitalizations. Stojanovic and collaborators described how to learn low-dimensional vector representations of patient conditions and clinical procedures in an unsupervised manner, and generate feature vectors of hospitalized patients, useful for predicting their length of stay, total incurred charges, and mortality rates [17]. Barabasi and Kleinberg published robust analytical and numerical framework to mathematically model the spread of pathogens [18,19]. Meyer developed power law models to better capture dynamics of infectious disease spread [20]. He demonstrated power law model frameworks and spatial distribution heatmaps on meningococcal bacterial meningitis in Germany and on influenza virus infections in the Southern Germany. Many papers described social network applications and geographical distributions in explaining of other diseases besides human types of Influenza. Poolkhet in his study described social network analysis for assessment of avian influenza spread and trading patterns of backyard chickens in Thailand [21]. Song and colleagues used Pearson's correlation to measure the impact of socioeconomic factors on AIDS diagnosis rates in certain geographic areas. Their correlation-based method discovered the complexity of contribution of socio-demographic determinants of health and geographic area based measures to AIDS diagnosis [22].

In our study, we proposed a novel method for better understanding of geographical distribution of hospitalized Influenza cases in a specific geographic region, using the combination of geographically specific heatmaps and social network analysis. A combination of social network analysis and visualization of findings on interactive geographical heatmaps is a novelty, that provides quick and efficient information about hubs and spatial distribution of hospitalized flu patients. We calculated correlations among zip codes in the state of New York. Zip codes represent specific geographic localities in a specific state, with very characteristic demographics in each of them. Our research used detailed demographics from 2010 Census. If we consider that the particular zip code has specific demographic characteristics, then we can assume that the specific geographical location of the zip code and demographic characteristics, affect numbers of cases and hospitalizations of Influenza. We then find correlations of these zip codes, knowing that actually we calculate correlations of numbers of cases affected by geographic locations and demographic characteristics of zip codes.

We took a period of 10 years, since we can draw conclusions about distribution of Influenza cases in so long period. Based on our findings we can expect similar patterns in the next decade.

In order to visualize findings, our method uses contemporary Google maps as the base for heatmaps. Researchers can clearly see names of cities, roads, rivers, mountains. Visualization is more effective than description of regions, because researches who are not familiar with all places in one state, can quickly see where the cities are located, how are they connected (highways, roads) and are any geographic features (mountains, lakes, etc) between nearby cities that can slow down spread of infections between 2 cities. This is especially effective if sites are not connected with direct roads.

Our novel method, proposed in this study, is a different methodology from previously published approaches. We constructed heatmaps, that show distribution of flu patients in NY state, which enable easy and fast visualization of zip codes, that are the most affected by flu infection, as well as visualization of the most likely routes of Influenza spreading. We performed a network analysis of distributions of patients through zip codes, where nodes represent affected zip codes and links represent correlation among zip codes. Designed network allows calculation of centrality measures, aimed to provide discovery of hubs and significance of individual zip codes. These findings could help medical professionals to improve planning of resources, needed to treat flu infections and complications and to better allocate resources. Our approach will provide fast and accurate understanding of Influenza in specific geographic areas, which can be the size of one or more states or

more countries. Detailed heatmaps will locate regions, that need the most resources for medical intervention. Our model provides more geographic details about the distribution of flu infection, than previously published research.

We conducted our research on Healthcare Cost and Utilization Project (HCUP) data for the period of 10 years, and we recommend further study of geographical distribution of Influenza on more different datasets in order to better understand geographical distribution of hospitalized influenza patients and what demographic and socio-economic factors contribute to that distribution.

2. Material and methods

Proposed is a novel method for better understanding of geographical distribution of hospitalized Influenza cases in New York state, using the combination of geographically specific heatmaps and social network analysis. We analyzed data from the HCUP, the State Inpatient Databases (SID). HCUP is a family of health care databases that contain data of State data organizations, hospital associations, private data organizations, and the Federal government. The HCUP includes the largest collection of longitudinal hospital care data in the United States and contains information on inpatient stays, emergency department visits, and ambulatory care. The SID are state-specific files that contain all inpatient care (hospital) records in participating states. The State-specific SID encompass more than 97 percent of all U.S. hospital discharges.

Data for this project were downloaded from HCUP - SID New York State inpatient database. We downloaded and analyzed data regarding the influenza infections for the period of 10 years (2003–2012). There were 30,380 cases of influenza registered in the database. Those were patients who required a hospital admission and stay, due to more severe flu or presence of complications.

Influenza cases were depicted on the bar-plot and visualized on heatmaps. We further analyzed these heatmaps with absolute numbers of hospitalized flu patients, to study activity and spatial spreading of the flu. We, also, normalized absolute numbers of patients over the population in each of zip codes. Normalization of results helped us discover which zip codes were the most prune to flu infections. In order to conduct the study, we used demographic data from the Census Bureau from the last census conducted in 2010. Census demographic data match the period of the processed data from HCUP databases. We analyzed a percentage of the population affected by severe flu complications that required hospitalizations. We normalized the number of patients per number of people who lived in individual zip codes to obtain percentages of affected population per zip code.

Next, we constructed a power-law type network. Statistical analysis was performed to determine if the network follows power-laws. A Kolmogorov–Smirnov test was conducted, at the significance level of 0.05. We created a function to estimate the exponent and to plot the log–log data and the fitted line. Networks whose degree distributions follow a power law are called scale free networks. The probability of observing high-degree nodes, or hubs, is very high in this type of network. Scale-free networks, also, have a large number of small degree nodes that tend to connect among themselves and are virtually absent in a random network.

In order to utilize network analysis of the geographic distribution of flu patients, we constructed a matrix (1471 rows and 12 columns). The rows represent 1471 zip codes in the state of NY, from which, patients were registered in the HCUP-SID database. The columns represent 12 months. Suitable for analysis of the flu distribution, through zip codes, was the weighted signed correlation network. To form this network, we constructed a 1471 × 1471 matrix with the aim of calculating the correlation between zip codes. Pearson's correlation (r) was calculated between pairs (x,y) of zip codes (Formula (1)).

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}} \quad (1)$$

We constructed the network from Pearson's correlations of numbers of complicated flu cases among zip codes. The network is first represented as a weighted correlation adjacency matrix.

For detailed network analysis we chose 100 zip codes, with the largest number of patients (70 and more), due to privacy protection of patients. The correlation was calculated among top 100 zip codes. The adjacency matrix, A[ij], encodes whether and how a pair of nodes is connected. For weighted networks, the adjacency matrix reports the connection strength between node pairs.

There are two types of weighted correlation networks: unsigned and signed. Unsigned networks have an absolute value of correlation $|\text{cor}|$ and Signed networks keep the sign of correlation $\frac{(cor+1)}{2}$.

We selected a signed correlation network. The nodes of such a network correspond to zip codes, and edges between them are determined by the pairwise Pearson's correlations between zip codes. The network was created from the adjacency matrix and performed power transformation (normalization) of calculated Pearson's correlations. For normalization, we used signed network normalization $|(\text{corMatrix} + 1)/2|^{\beta}$ for 100 × 100 adjacency matrix. By raising the absolute value of the correlation to a power $\beta \geq 1$ (soft thresholding), the signed correlation network emphasizes high correlations at the expense of low correlations. We tested β-values between 1 and 10, and chose the value of $\beta = 2$, as the best choice to represent the flu infection by the correlation among zip codes in the network. We used the correlation matrix after the transformation (normalization) as adjacent matrices to plot the network. The cutoff for edges, to be plotted on the network, was set to some reasonable number (smaller correlations were not plotted). The cutoff correlation of 0.9 and higher was selected to be plotted as an edge. In order to determine hubs in the network we calculated degrees for all 100 nodes.

The associated network, based on correlation results between zip codes of patients was constructed in R, with the help of WGCNA, Statnet and gplot packages.

Our method brings together Biomedical informatics, Medicine and Network science, in an attempt to illuminate nature of Influenza, in this specific population. This method can be generalized and applied to other infectious diseases and geographic regions in the U.S. and in the world.

3. Results

Data from the HCUP–SID New York State database for the period 2003–2012 were analyzed. We studied the evolvement of flu infections that required hospitalization throughout the year, with monthly breakdown of cases (January through December), for 10 years with a total number of 30,380 inpatient cases, with influenza diagnosis. The display of monthly breakdown of the number of hospitalized cases for the flu is shown on the bar plot (Fig. 1). The highest number of cases was registered in December (6720). Flu virus infections were also very active in January, February and March. Out of the peak of the flu season, cases were sporadic, even in big cities.

The Heatmap (Fig. 2) of the state of NY, which shows the distribution of hospitalized patients with flu complications throughout different zip codes, was constructed for the same period of 10 years. Dots on heatmaps show numbers of patients who reside in particular zip codes. We show only zip codes with more than 20 cases of flu (due to privacy reasons). The total number of zip codes shown on the Heatmap was 443, with the total number of patients of 29,071.

The highest numbers of cases are registered in the most urban zip codes. Red color (large number of cases) on heatmaps is noticeable in big cities: Albany, Buffalo, New York City, Rochester and Syracuse. Also, we can observe that a lot of blue dots (that correspond to zip codes with 20–50 cases) are highly concentrated in big urban areas. An

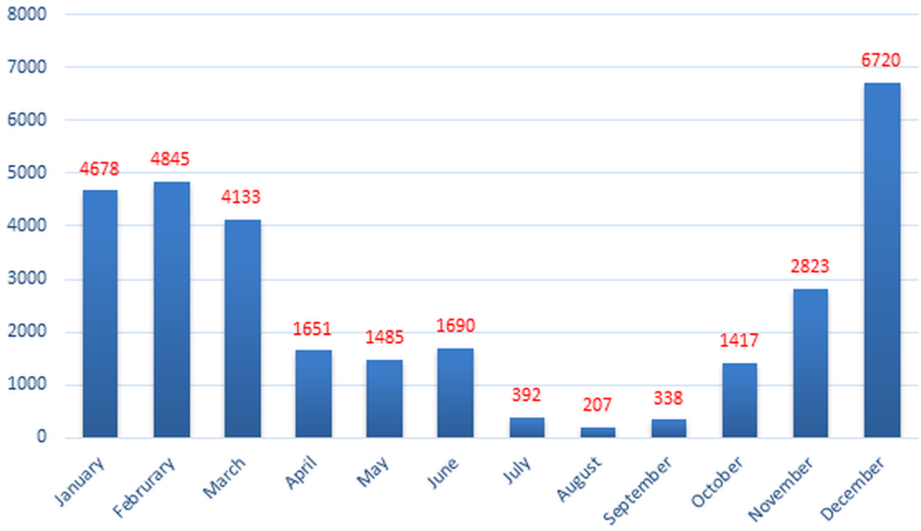

Fig. 1. Bar plot – Number of patients infected by the influenza virus during the 2003–2012 period (monthly distribution), who were hospitalized in the state of New York.

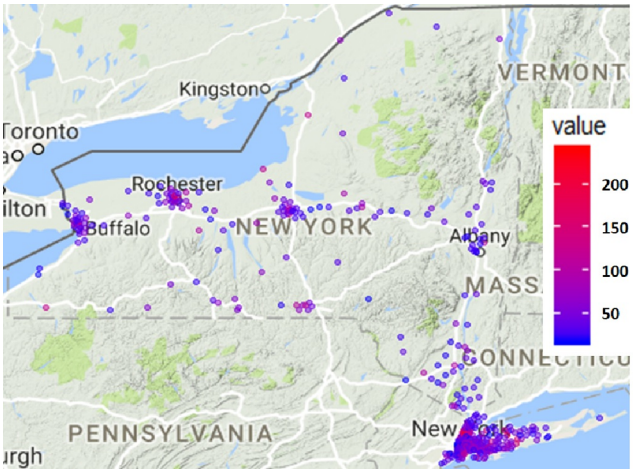

Fig. 2. Heat map of NY state – Distribution of hospitalized flu patients by the zip code (2003–2012), shows that the highest concentration of hospitalized flu patients was in five big urban areas (Albany, Buffalo, New York City, Rochester and Syracuse). The heatmap also shows that routes of spreading follow highways (in particular Highways 81, 86 and 90) as high frequency routes of travelling between places.

important finding on the heatmaps, is that the distribution of hospitalized patients follows highways or other big roads, which indicates that flu spreads along the routes that people use to move from place to place. Rural areas had small numbers of cases. Further, we present heatmaps of big urban areas separately (Fig. 3): (a) Albany area, (b) New York City area, (c) North side of NY State and (d) Buffalo area. These heatmaps can be used for future predictions and healthcare planning for particular zip codes, as the areas with the highest risk for the flu infection outbreaks and spreading. Accordingly, health care providers should plan more resources to deal with hospitalized Influenza patients in these particular areas.

We analyzed the distribution of population per zip codes in the state of NY. Populations of 20 the largest zip codes are shown in Table 1.

We presented 20 zip codes with the largest absolute number of hospitalized patients infected with flu in the period of 2003–2012 in Table 2.

Simple inspection of these tables discovered that the highest populated zip codes didn't have the largest number of patients with flu complications. The most populated zip code at the 2010 census was 11368 (Queens) with 53.6% males and 46.4% females and average age 31.8 (significantly lower than the state average). About 59% were singles. Compared to NY state averages we found that average income was 3 times lower than the average state income. Hispanic population percentage (74%) was significantly above the state average, median age below the state average, foreign-born population percentage significantly above the state average. There are no hospitals in this zip code. Percentage of the population without a health insurance was extremely high 31%, which is much higher than the state average (8.7%). The second highest populated zip code was 11226 (Brooklyn) with 44.9% males and 55.1% females. About 71% of the population were black, and 14% Hispanic. Comparing to the state average: Black race population percentage (71%) significantly above the state average, followed by 14% Hispanics, median age below the state average. Foreign-born population percentage was above the state average. There are few hospitals in this zip code, but the percentage of population without health insurance was 15% (higher than the state average). The third most populated zip code was 11373 (Elmhurst) with 50.4% males and 49.6% females. 22% of population didn't have health insurance, which was significantly higher than surrounding areas. Hospitals are available in this area. Hispanic (43%) and Asian (47%) population percentages were above the state average and foreign-born population percentage significantly above the state average. Average income was significantly below the state average.

Analyses of the top 20 highest populated zip codes in the NY state lead us to discoveries, that all of these zip codes are located in the NY City area, mostly in Brooklyn. Top 5 zip codes have significantly lower average income than the average income in the state. They also have significantly higher percentage of population without health insurance, and a large percentage of foreign born, as well as Hispanic and black population. Average number of household members was larger than the state average. Common conclusion for this population could be that, due to lower income and lower percentage of health insurance, a larger fraction of residents of these zip codes haven't visited hospitals. One of zip codes among the top 20 highest populated zip codes was a zip code 10025 with predominantly white population and larger average income and larger percentage of population with the health insurance. This zip code is not on the list of top 20 zip codes with the largest number of hospitalized patients, which could lead to the conclusion that patients got necessary health care treatment before the flu developed complications, or they had less flu cases due to preventive measures.

The top zip code with the most patients who were hospitalized due to Influenza, was 10457 (Bronx) with 241 patients. Census data for 2010, show that the zip code had the population of 70,496 (65%

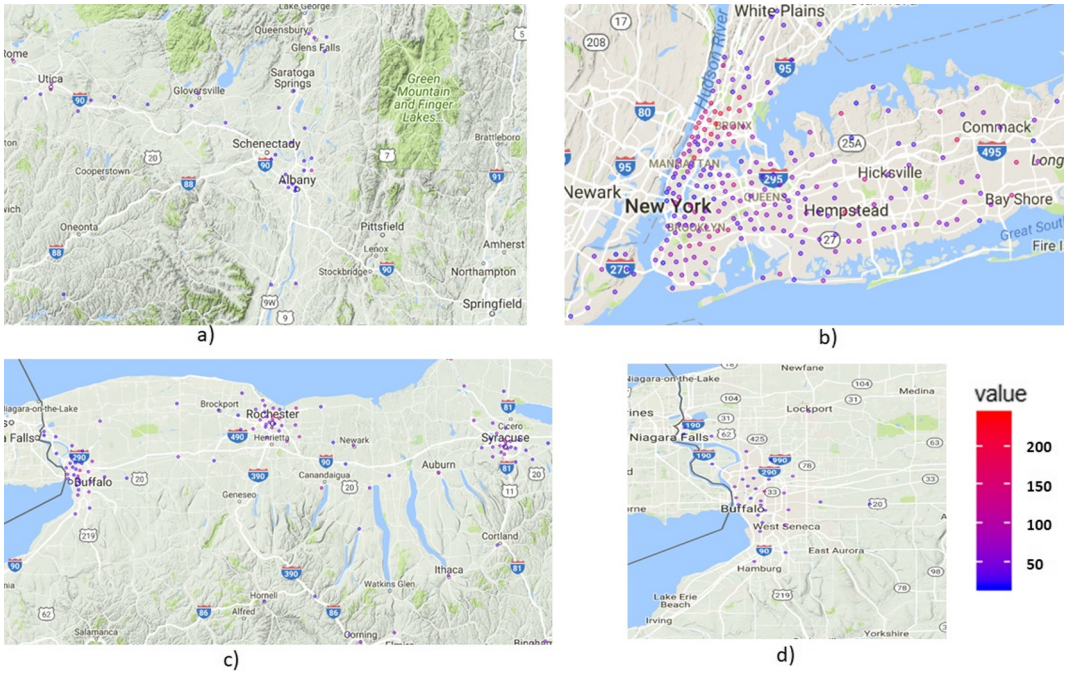

Fig. 3. Heatmaps of (a) Albany area, (b) NY City area, (c) North side of NY state, (d) Buffalo area. Heatmaps show that the distribution of hospitalized flu patients by the zip code in period between 2003 and 2012, was highly concentrated in five big cities (Albany, Buffalo, New York City, Rochester and Syracuse). Heatmaps show that the routes of distribution follow highways (highways 81, 86 and 90, in Albany area highways 87 and 9 and in Buffalo area, highway 190 toward Niagara Falls).

Hispanic, 30% black, 1.5% white and Asian…). Median age was 29.8 with 52.5% women and 47.5% men. Median income was about $24000 (3 times less than the state average income). The second zip code by the absolute number of hospitalized patients from Influenza was 10458 (Bronx) with 221 patients. This zip code had the population of 79,492 (64% Hispanic, 20% black, 10% white, 4% Asian…). Median age was 29.3 with 52% women and 48% men. Median income was 3 times less than the state average income. The third zip code by absolute number of hospitalized patients from Influenza was 10467 (Bronx) with 208 patients. This zip code had the population of 97,060 (48% Hispanic, 33% black, 10% white, 6% Asian…). Median age was 33.6 with 52% women and 48% men. Median income was about 2.5 times less than the state average income. Further analysis revealed that, among top five zip codes, 4 are in Bronx, with 870 patients hospitalized due to flu complications. We can also notice that among top 20 zip codes, most of patients were in Bronx.

Furthermore, we normalized the number of hospitalized patients per population in zip codes and calculated a percentage of population affected by severe flu complications, that required hospitalizations. We show percentages of affected population per zip code in Table 3. This time, the highest percentage of affected population was in the zip code 11509 (Atlantic Beach, NY). In this zip code, we found 29 hospitalized flu patients per 2653 residents. Zip codes with small numbers of residents have a higher percentage of affected people, than more populated zip codes. It's interesting that none of the zip codes from Bronx were in the top 20, despite the fact that they had the highest total numbers of patients.

To further analyze geographical distribution of hospitalized patients with Influenza infections a power-law type network was constructed. In order to create this network, we calculated correlations among zip codes, considering distribution of hospitalized flu cases, over the 10 years period. A common property of power law type networks is that the node degrees span several orders of magnitude. Outliers, or exceptionally high-degree nodes, are not only allowed but expected in these networks [7]. The main reason to construct the power-law type network was to identify highly connected nodes (hubs). If we can locate hubs, that will help us to prepare an adequate public health strategy to eliminate them and decrease the magnitude of hospitalized influenza infections in those spatial regions, which will significantly alleviate the cost that influenza infections impose on populations.

We performed statistical analysis to determine if the network follows power-laws. A Kolmogorov–Smirnov test was calculated at the significance level of 0.05. Obtained results show that at this significance level, the network is of power-laws type ($p = 0.01$). We created the function that helped us to estimate the exponent, plotted the log–log data and the fitted line (Fig. 4). The calculated value of the degree exponent $\gamma = 2.5935$ (t-statistic value = 9.804 (p-value very small), SE = 0.2645, distance distribution = 3.618).

The network is first represented as a weighted correlation adjacency matrix. Initially, we calculated Pearson's correlation among all 1471 zip codes (plot of correlation is shown in Fig. 4).

Detailed network analysis was performed on 100 zip codes with the largest number of patients (70 and more). We used the correlation matrix, after the power transformation (normalization) as adjacent matrices to plot the network. We selected the cutoff correlation of 0.9 and higher to be plotted as an edge. The network is shown on Fig. 5. We picked two different colors to make nodes and labels more visible, with no other meanings.

Table 1
Distribution of population per zip codes (20 the largest) in the state of New York from census data for 2010.

Zip	11368	11226	11373	11220	11385	10467	10025	11208	11236	11207
Population	109,931	101,572	100,820	99,598	98,592	97,060	94,600	94,469	93,877	93,386
Zip	11219	11211	11377	11214	11234	10456	11230	11355	10314	11212
Population	92,221	90,117	89,830	88,630	87,757	86,547	86,408	85,871	85,510	84,500

Table 2
Number of hospitalized flu patients in the state of New York for the period 2003–2012 (HCUP data).

Zip	10457	10458	10467	10029	10456	10032	11550	11746	10466	10463
Population	241	221	208	207	200	190	185	184	178	177
Zip	10468	10453	14621	10452	11542	10033	11717	10031	10469	14609
Population	176	172	168	164	163	160	154	152	152	146

Analyzing the network, we can clearly see nodes with high degrees (hubs) as well as isolated nodes. Hubs correspond to flu cases in big urban areas (cities) in the State of NY. Disconnected nodes correspond to isolated cases in rural areas. By studying the network, it's possible to identify critical clusters, hubs, and routes, that could be subjects of intervention, in order to minimize the spread of Influenza, decrease numbers of complicated cases that require hospitalization and decrease the cost.

In order to determine significance of particular zip codes and which zip codes were hubs in the network, we calculated degrees for all 100 nodes. Zip codes (that represent nodes in the network) with the highest node degrees are: 10465 (Bronx) with the degree of 86, then 11226 (Brooklyn) – degree of 84 and 10027 (NY City) – degree 80. Top 20 zip codes with highest degrees are shown in Table 4.

Census data for 2010, show that the zip code 10465 had the population of 42,230 (51% white, 37% Hispanic, 7% black, 3% Asian…). Median age was 40.7 with 52% women and 48% men. Median income was 1.4 times less than the state average income. The number of people without health insurance was 7.3% which is better than the state average (8.7%). Few hospitals are available in the area. The zip code 11226 had the population of 101,572 (71% black, 17% Hispanic, 6% white, 3% Asian…). Median age was 34.3 with 55% women and 45% men. Median income was about 2.3 times less than the state average income. Uninsured population was 14.7%, which is larger than the state average and there are few hospitals in this area. Zip code 10027 had the population of 59,707 (40% black, 26% Hispanic, 23% white, 8% Asian…). Median age was 30.8 with 53.5% women and 46.5% men. Median income was about $50000 1.5 times less than the state average income. Number of uninsured people was 10%, with few hospitals in the area. Zip code 10457 (Morningside Heights, Uptown, Manhattan, NY City) had population with higher number of uninsured people 13.6% and lower income than the state average. This zip code had the largest number of hospitalized patients (241) in NY state. Zip code 10463 (Bronx) had 10th largest number of hospitalized patients (177). Number of people without health insurance was 9.3% and average income was slightly smaller than the state average. Further inspection of the list of zip codes with highest node degrees in the constructed social network, shows that zip codes were located in the NY City area, with moderate to large population sizes. Most of them had larger than average percentages of population without health insurance. Vast majority of zip codes had lower than average income. Females were majority of population in most of zip codes. And almost all zip codes had significantly larger fraction of Hispanic and black population than the state average.

We constructed heatmaps to visualize geographic locations of top 20 zip codes with the highest node degrees (Fig. 6). We can clearly conclude that all 20 zip codes with the highest node degrees (hubs) are in the NY City area.

4. Discussion

The goal of this analysis of influenza, is to contribute to better understanding of spatial spreading of hospitalized flu cases. If we convey this research into the lower numbers of infected people, the result will be more saved lives, a decrease in medical costs and economic losses. We wanted to show that the ability to more accurately analyze and assess infection levels in geographic regions, that have higher infection risk, in the future, can suggest targeted planning and measures to deal with complications of Influenza. We provided a detailed analysis of hospitalized cases, caused by flu infection in the state of NY. We constructed heatmaps to visualize findings of our research. Heatmaps show that the majority of cases are located in big cities. Other interesting finding, from heatmaps, was that the spreading of Influenza around the state was along highways and big roads. Out of 20 zip codes with the highest absolute number of flu infected patients, who required hospitalizations, most of them were in Bronx. Demographic data from 2010 census show, that the top 3 zip codes (10457, 10458, 10467) with the highest absolute number of patients were located in Bronx. All 3 zip codes had predominantly Hispanic population, followed by black, white and Asian population. Further, all 3 zip codes had significantly lower median age (29.8, 29, 3, 33.6) of population than the median age of NY state, which is 38.4. Average income in all 3 zip codes was significantly lower than the average income in NY state (2.5–3 times). We also found that normalized data per number of residents in zip codes, show that zip codes with the largest percentage of population affected by influenza were different than the zip codes with the largest absolute numbers of patients. The largest percentage of patients was in Atlantic Beach, which has relatively small population. The second largest percentage of patients per population was in Glen Clove. Both of zip codes are in proximity of NY City, but they have small populations.

Constructed power-law network reveals hubs (high degree nodes). Most zip codes among the top 20 by the highest node degree are located in Bronx, NY City and Brooklyn area. This finding confirms that the most hubs are in high populated areas in big cities. Detailed analysis of 3 zip codes with the highest node degrees, show that all 3 of them have significantly lower average income then the state income. Zip code 10465, has predominantly white and Hispanic population, but the average age was higher than the state average (40.7, vs 38.4). Other 2 zip codes (11226, 10027) had predominantly black population, followed by Hispanic and white. Both areas had higher percentage of female population, than the state average percentage of females. Although we presented demographic data for top 3 zip codes (in cases of the highest numbers of hospitalized Influenza cases or highest node degrees) the trend in the top 20 zip codes in both cases was similar. Data revealed that all zip codes with higher number of hospitalized Influenza cases, or higher node degrees in the social network, have significantly lover average income than the rest of population in NY

Table 3
Percentages of affected population for 20 zip codes with the highest percentages of hospitalized flu patients.

Zip	11509	11542	13205	14895	13452	13367	13202	13904	14605	14621
% of patients	0.011	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
Zip	13204	13203	14514	13669	11798	13208	13905	14513	14482	14843
% of patients	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004

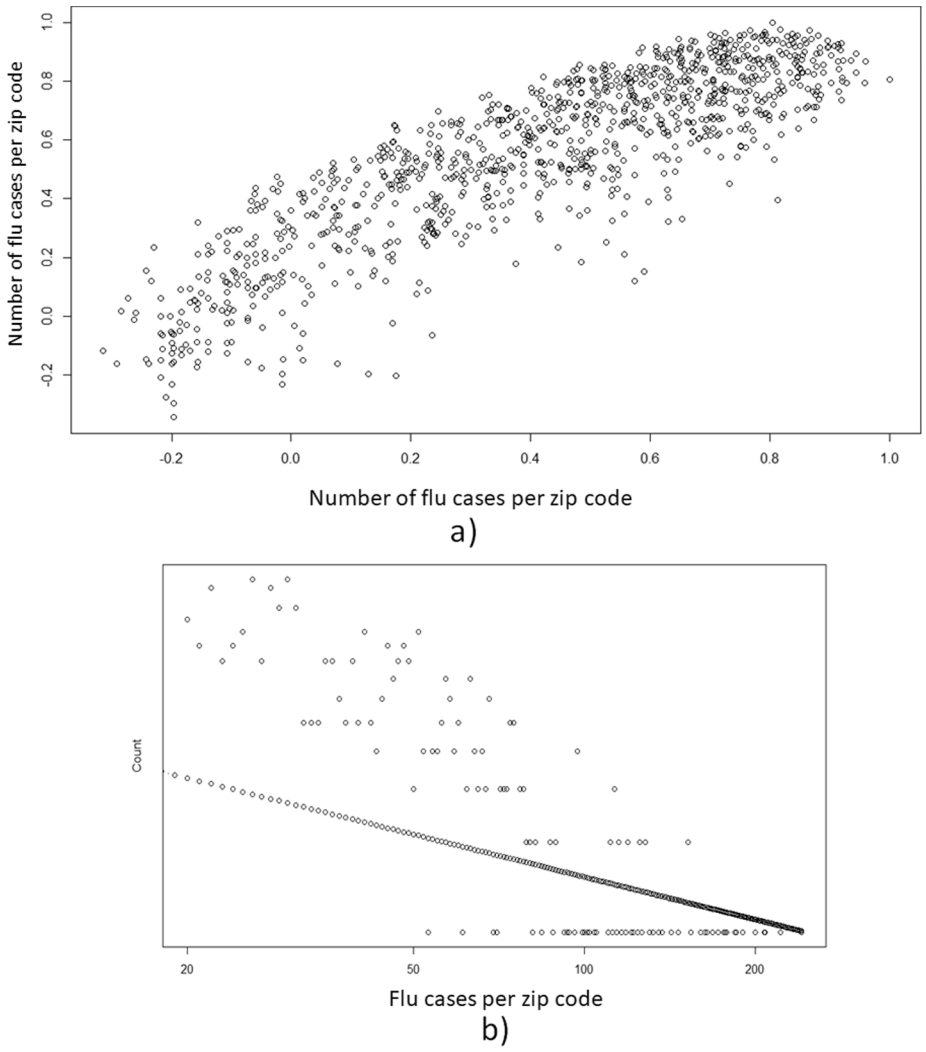

Fig. 4. (a) Plot of correlations between 1471 zip codes with respect to number of hospitalized patients with flu complications in those zip codes between 2003 and 2012, (b) Plotted the estimate of the power law exponent, the log–log data, fitted line at $\gamma = 2.5935$.

state. Also, most of these top zip codes show younger population predominantly black or Hispanic, as well as higher than the state average of foreign born population. Common and important characteristic of zip codes with high node degrees is that almost all of them have significantly higher percentage of population without health insurance than the state average. In case, if white population was majority, those zip codes had higher average age, than the average age in NY state. Findings of the social network analysis were in alignment with findings shown on heatmaps and in 4 tables. All zip codes with the highest node degrees were also among top zip codes with the largest number of hospitalized influenza cases, or the largest number of people living in those zip codes. Discoveries of our social network research suggest some obvious measures that could lead to lower number of complicated flu cases that needed hospitalizations. Discovery that most of zip codes among top 20 with high node degrees had lower income and high percentage of foreign born population suggest that these issues should be addressed. Findings that most of the zip codes had high percentage of uninsured population suggest that health insurance affordability is very important factor and measures that will increase number of insured people should be applied. The next issue that arises in areas with lower incomes and larger number of uninsured people is the availability and affordability of primary care physicians, which also needs to be addressed. More affordable primary care medical offices, that accept patients without insurance or with not so good insurance coverage are needed in these zip codes. Further, a lot of areas have hospitals that are not well ranked, which can be managed by providing more resources to these hospitals. Additional problems are that hospitals in many cases are concentrated in medical centers and there are not enough hospitals in areas where people live. It's known that a lot of people without health insurance do not visit primary care offices and wait till they are very sick to go directly to emergency rooms. During epidemics a lot of medical personal are busy with other patients, so wait periods till patients are seen by doctors could be very long. During that time health conditions can further deteriorate. These problems could be solved by employing more doctors, or trainees like residents, or Physician assistants, Nursing practitioners… Additional problems that people with significantly lower incomes than average and without healthcare insurance face are lower percentages of preventive vaccinations and other preventive measures. Often, zip codes like this don't have enough pharmacy stores in the area, which is the problem during epidemics, because people have to travel longer distances to purchase necessary medications.

Our social network study clearly identified hubs among zip codes, that need some of suggested measures to improve prevention and treatment that will decrease numbers of hospitalized cases.

This study of social networks provides a wealth of information for understanding the influence of population sizes and demographics, in particular zip codes, on the spread of influenza viruses. We used a novel

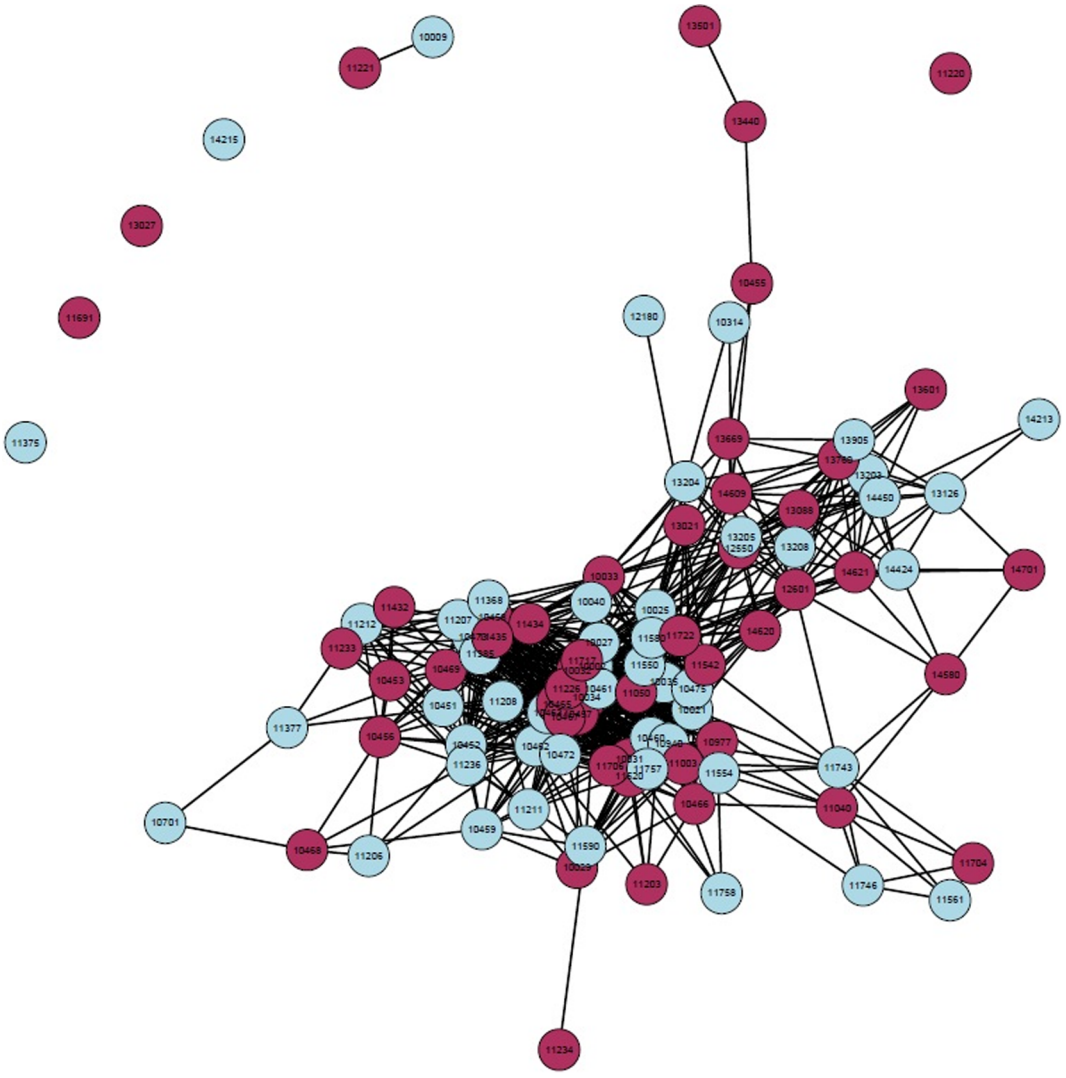

Fig. 5. The power law type network representation of hospitalized flu cases in the State of New York between 2003 and 2012, based on correlation between zip codes. Nodes correspond to zip codes that are linked, based on strength of calculated correlations. Zip codes (nodes in the network) with the highest node degrees are: 10465 (Bronx) with the degree of 86, then 11226 (Brooklyn) with the degree of 84 and 10027 (NY City) – degree 80.

Table 4
Zip code with the highest node degrees.

	Zip code	Township names	Node degree
1	10465	Eastchester Bay, Bronx	86
2	11226	Flatbush, Brooklyn	84
3	10027	NY City	80
4	10457	Morningside Heights, Uptown, Manhattan, NY City	80
5	10463	Riverdale, Bronx	80
6	11580	Valley Stream, NY	80
7	10035	East Harlem, Harlem, NY City	78
8	10467	Van Cortlandt Pk, Bronx	78
9	10032	Washington Heights, Manhattan, NY City	76
10	10460	Bronx Park South, Bronx	74
11	11550	Hempstead, NY	74
12	10034	Inwood, Uptown, Manhattan, NY City	72
13	11722	Central Islip	72
14	10461	Westchester Square, Bronx	70
15	11717	Brentwood	70
16	11003	Elmont	68
17	11520	Freeport	68
18	10025	Upper West Side, West Side, NY City	66
19	10462	Van Nest, Bronx	66
20	11208	City Line, Brooklyn	66

methodology to construct heatmaps and the network representation of hospitalized flu patients in the state of NY (2003–2012), based on correlation among zip codes. Results of this study have important implications for predicting the geographical spread of hospitalized cases of influenza and prioritizing some of suggested public health measures. Results can help adequate planning of resources for infectious disease outbreaks and their efficient control, as well as planning of hospital resources for more severe cases in the future.

5. Conclusion

Our research brings together medicine, biomedical informatics, computer science and social network science, in an attempt to explain the distribution of hospitalized flu cases. The desire to have realistic networks based on spatial distribution of complicated cases, for entire populations, provides important insights how the size of the population and demographics, influence distribution of influenza. The future research framework in this field, would allow for different networks (from different times or different locations) to be compared. It will be important, for further development of the network science and its ability to analyze spread of Influenza, to have effective data collecting protocols and to use the statistical techniques to analyze collected data.

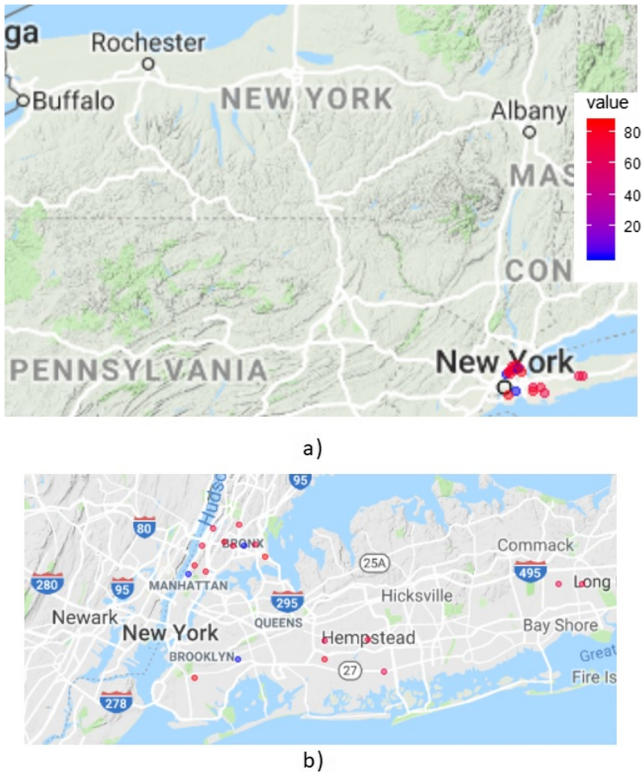

Fig. 6. Geographic location of 20 zip codes with the highest nodes degrees (hubs) in the state of New York. (a) the whole NY State, (b) NY City area – all 20 top hubs are in this area. Zip codes (dots on maps) with the highest node degrees are: 10465 (Bronx) with the degree of 86, then 11226 (Brooklyn) with the degree of 84 and 10027 (NY City) – degree 80.

Our research was conducted on HCUP data, and we recommend further study of geographic distribution of Influenza on more different datasets and on different size of geographic areas in order to improve understanding of distribution of hospitalized influenza patients and spreading of Influenza.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported in part by the NSF grant SES-1659998, SES-1447670 and Pennsylvania Department of Health CURE Health Data Science Research Project. Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality provided data used in this study.

References

[1] Centers for Disease Control and Prevention (CDC), Estimated influenza illnesses and hospitalizations averted by influenza vaccination-United States, 2012–13 influenza season. MMWR, Morbidity and mortality weekly report 2(49) (2013) 997.
[2] Center for Disease Control, Disease burden of Influenza, < https://www.cdc.gov/flu/resource- center/freeresources/graphics/flu-burden.htm > .
[3] S. Cauchemez, A. Bhattarai, T.L. Marchbanks, et al., Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza, Proc. Natl. Acad. Sci. 108 (7) (2011) 2825–2830.
[4] J.M. Read, T.D. Ken, T.D. Eames, et al., Dynamic social networks and the implications for the spread of infectious disease, J. R. Soc. Interface 5 (2008) 1001–1007.
[5] C. Viboud, O. Bjørnstad, D.L. Smith, et al., Synchrony, waves, and spatial hierarchies in the spread of influenza, Science 5772 (2006) 447–451.
[6] J.R. Gog, S. Ballesteros, C. Viboud, et al., Spatial transmission of 2009 pandemic influenza in the US, PLoS Comput. Biol. 10 (6) (2014) e1003635.
[7] Stephen M. Kissler, Julia R. Gog, Cécile Viboud, Vivek Charu, Ottar N. Bjørnstad, Lone Simonsen, Bryan T. Grenfell, Geographic transmission hubs of the 2009 influenza pandemic in the United States, Epidemics (2018), https://doi.org/10.1016/j.epidem.2018.10.002.
[8] R.E. Malosh, E.T. Martin, J.R. Ortiz, et al., The risk of lower respiratory tract infection following influenza virus infection: a systematic and narrative review, Vaccine 36 (1) (2018) 141–147.
[9] D.C. Pearce, J.M. McCaw, J. McVernon, et al., Influenza as a trigger for cardiovascular disease: an investigation of serotype, subtype and geographic location, Environ Res. 156 (2017) 688–696.
[10] J.R. Ortiz, K.M. Neuzil, C.R. Cooke, et al., Influenza pneumonia surveillance among hospitalized adults may underestimate the burden of severe influenza disease, doi:10.1371/journal.pone.0113903.
[11] J. Ginsberg, M.H. Mohebbi, R.S. Patel, et al., Detecting influenza epidemics using search engine query data, Nature 457 (2009) 1012–1015.
[12] M.W. Davidson, D.A. Haim, J.M. Radin, Using networks to combine "big data" and traditional surveillance to improve influenza predictions, Sci. Rep. 5 (2015) 8154.
[13] R.R. Miller, B.A. Markewitz, R.T. Rolfs, et al., Clinical findings and demographic factors associated with ICU admission in Utah due to novel 2009 influenza A(H1N1) infection, Chest 137 (4) (2010) 752–758.
[14] J.J. Jung, R. Pinto, R. Zarychanski, et al., 2009–2010 Influenza A(H1N1)-related critical illness among Aboriginal and non-Aboriginal Canadians, PLoS One 12 (10) (2017) e0184013.
[15] L. Danon, A. Ford, T. House, et al., Networks and the epidemiology of infectious disease, Hindawi. Interdis. Perspect. Infect. Dis. (2011) 1–28. Article ID 284909.
[16] D.j. Gligorijevic, J. Stojanovic, Z. Obradovic, Uncertainty propagation in long-term structured regression on evolving networks, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16) Phoenix, AZ, February 2016, 1603–10.
[17] J. Stojanovic, D. Gligorijevic, V. Radosavljevic, et al., Modeling healthcare quality via compact representations of electronic health records, IEEE/ACM Trans Comput Biol Bioinform. 14 (3) (2017) 545–554, https://doi.org/10.1109/TCBB.2016.2591523.
[18] A. Barabasi, Network Science (book), 2016.
[19] D. Easley, J. Kleinberg, Networks, crowds, and markets: reasoning about a highly connected world, 2010.
[20] S. Meyer, L. Held, Power-law models for infectious disease spread, Ann. Appl. Stat. 8 (3) (2014) 1612–1639.
[21] C. Poolkhet, P. Chairatanayuth, S. Thongratsakulet, et al., Social network analysis for assessment of avian influenza spread and trading patterns of backyard chickens in Nakhon Pathom, Suphan Buri and Ratchaburi, Thailand, Zoonoses Public Health 60 (6) (2013) 448–455.
[22] R. Song, H.I. Hall, K. McDavid-Harrison, et al., Identifying the impact of social determinants of health on disease rates using correlation analysis of area-based summary information, Public Health Rep. 126 (Suppl. 3) (2011) 70–80.