Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A network-based method with privacy-preserving for identifying influential providers in large healthcare service systems

Xiaoyu Qi a, Gang Mei a,*, Salvatore Cuomo b, Lei Xiao a

a School of Engineering and Technology, China University of Geosciences (Beijing), China
b Department of Mathematics and Applications, University of Naples Federico II, Italy

**Abstract**

In data science, networks provide a useful abstraction of the structure of many complex systems, ranging from social systems and computer networks to biological networks and physical systems. Healthcare service systems are one of the main social systems that can also be understood using network-based approaches, for example, to identify and evaluate influential providers. In this paper, we propose a network-based method with privacy-preserving for identifying influential providers in large healthcare service systems. First, the provider-interacting network is constructed by employing publicly available information on locations and types of healthcare services of providers. Second, the ranking of nodes in the generated provider-interacting network is conducted in parallel on the basis of four nodal influence metrics. Third, the impact of the top-ranked influential nodes in the provider-interacting network is evaluated using three indicators. Compared with other research work based on patient-sharing networks, in this paper, the provider-interacting network of healthcare service providers can be roughly created according to the locations and the publicly available types of healthcare services, without the need for personally private electronic medical claims, thus protecting the privacy of patients. The proposed method is demonstrated by employing PhysicianandOtherSupplierDataCY2017, and can be applied to other similar datasets to help make decisions for the optimization of healthcare resources in the response to public health emergencies.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

A network is a collection of nodes/vertices joined together in pairs by links/edges. Networks are used in many fields to represent the patterns of connections between the components of various complex systems [1]. Networks provide a useful abstraction of the structures of many complex systems, ranging from social systems and computer networks to biological networks and physical systems. Much research has been conducted to extract insights from network data especially based on the topology of networks [2].

Healthcare service systems are one of the main social systems that can also be understood using network-based approaches [3]. For example, social network analysis can be used to estimate patient flows, evaluate cooperation between healthcare providers, or identify influential providers in networked healthcare service systems. These extracted insights from networked healthcare systems can be exploited to improve healthcare service utilization and management [3].

Identifying the most influential providers in a healthcare service system is an important step towards optimizing the use of available healthcare resources and ensuring the more efficient delivery of the matched healthcare service. This information can be abstracted to identify the influential nodes in networks, which mainly consists of two stages: (1) creating reasonable networks for the interested healthcare service systems and (2) identifying the influential nodes in the created networks.

Much research has been conducted to abstract healthcare systems as networks, and most of this research has focused on creating “patient-sharing networks” based on electronic medical claims [4]. For example, Pollack et al. [5] first created patient-sharing networks among physicians using medical claims data and then investigated care coordination and costs of care. Ong et al. [6] created a patient-sharing network using commercial healthcare claims spanning the years 2008 through 2011 and elucidated the effect of professional relationships among providers on multiple-provider prescriptions of benzodiazepines.

Moreover, using Medicare Part B claims data, Moen et al. [7] first constructed a network of physicians who care for patients with cardiovascular disease based on patient-sharing relationships and then analyzed a case study of physicians’ adherence to clinical guidelines for the prevention of sudden cardiac death.
Using a dataset consisting of 1306 physicians who practiced within the Chicago hospital referral region in 2016 and who collectively had 12,091 patient-sharing ties, Linde [8] created a patient-sharing network and examined the degree to which hospital affiliation drives physicians’ sharing of Medicare patients.

In addition, Landon et al. [9] identified patient-sharing networks in which physicians share patients, information, and behaviors and presented a social network analysis of Medicare administrative data from 2006 to 2010 in 51 hospital referral regions. Similarly, DuGoff et al. [10] constructed patient-sharing networks using administrative data in which pairs of physicians are considered connected if they both deliver care to the same patient and examined the approaches to conceptualizing, measuring, and analyzing provider patient-sharing networks.

There are typically two procedures in identifying influential nodes in the created networks: (1) ranking nodes according to nodal influence measure metrics, such as degree centrality [11], clustering coefficients [12], H-Index [13,14], and k-shell [15,16]; and (2) evaluating the influence of top-ranked nodes by comparing the structures and functions before and after removing a certain percentage of top-ranked nodes.

Currently, many algorithms have been proposed for identifying influential nodes. Most of the identification algorithms are designed for online social networks [17–19]. In these identifications, some influence measures, such as degree centrality [11], betweenness centrality [20], closeness centrality [21], and Katz centrality [22], are highly dependent on the topological structure of the network. Moreover, Xiao et al. [12] used company behaviors and clustering coefficient to rank nodes. Lu et al. [13] adopted the H-Index [14] to identify vital nodes. Kitsak et al. [23] proposed the theory of k-shell decomposition to identify influential nodes.

In addition, in the development of search engines, there are also important and famous algorithms for ranking nodes, such as PageRank [24], HITs [25], and LeaderRank [26]. These ranking algorithms have been excellently investigated and summarized in several reviews [20,27]. The node ranking and influential nodes identification methods have their own advantages and disadvantages. Users can employ suitable metrics and develop algorithms in their specific applications.

In this paper, we propose a network-based method with privacy-preserving for identifying influential providers in large healthcare service systems. First, the provider-interacting network is constructed by employing publicly available information on the locations and types of healthcare services of providers. Second, the ranking of nodes in the generated provider-interacting network is conducted in parallel on the basis of four nodal influence metrics, including the degree centrality (DC) [11], the companion behaviors (CB) [12], the clustering coefficients (CC) [12], and H-Index [13,14]. Third, the impact of top-ranking influential nodes in the provider-interacting network is evaluated by comparing three network indicators, including the maximum connectivity coefficient [28], network efficiency [29,30] and susceptibility [27]. The proposed method is finally demonstrated by employing Physician and Other Supplier Data CY 2017 [31] and can be applied to other similar datasets to help make healthcare system management decisions, such as the optimization of healthcare resources.

The novelty of the proposed method can be explained as follows. As mentioned above, most of the related research is based on patient-sharing networks, which are created according to electronic medical claims, focusing on the sharing of patient information between providers. The advantage of creating patient-sharing networks is that they are quite precise. However, the data of electronic medical claims are personal and private [32,33], and the data are generally not directly available to the public. Healthcare organizations need to consider regulations and rules of privacy in regard to patient information. Researchers need to deal with personal privacy information carefully before creating the required patient-sharing network.

In contrast, our study is based on a patient-sharing network rather than the provider-interacting network. In the proposed method, the provider-interacting network joined by healthcare providers can be roughly created according to the location and available types of healthcare service, without the need for personal, private electronic medical claims data. We only utilize the relationships between providers to identify influential providers, thus protecting the privacy of patients. Most importantly, the locations and the provided types of healthcare services are publicly available. For example, these data are publicly available from the U.S. Centers for Medicare & Medicaid Services (https://www.cms.gov/). The aforementioned networks connected by healthcare providers are referred to as provider-interacting networks. The “provider-interacting” networks can be created by employing the publicly available information on the locations and the provided healthcare service.

The main contributions of this paper can be summarized as follows.

(1) We construct a provider-interacting network by employing publicly available information on locations and types of healthcare services of providers.

(2) We rank the influential nodes of the created provider-interacting networks using four local metrics.

(3) We evaluate the impact of the top-ranked influential nodes in the provider-interacting network using three indicators.

The rest of this paper is organized as follows. Section 2 describes the details of the proposed network-based method for identifying the influential provider of healthcare service in the provider-interacting network. Section 3 presents the application of the proposed method for the Physician and Other Supplier Data CY 2017 [31]. Section 4 discusses the experimental results and the proposed method. Finally, Section 5 draws several conclusions.

2. Materials and methods

In this section, we will first introduce the data source and then describe the details of the proposed network-based method.

2.1. Data source

In this paper, the data of providers and other suppliers who have effective National Provider Identifiers (NPIs) and submit Part B medical insurance services in the United States from 2012 to 2017 are obtained from the U.S. Centers for Medicare & Medicaid Services. We first extract interesting and important data and summarize them according to the following content:

(1) npi — NPI for the performing provider on the claim. The provider’s NPI is the numeric identifier registered in the NPPES. Each provider has a unique NPI.

(2) nppes_provider_zip — The provider’s ZIP code.

(3) hcpcs_code — Healthcare Common Procedure Coding System (HCPCS) code used to identify the specific medical service provided.

The availability of the obtained datasets is as follows.

Dataset: Medicare Physician and Other Supplier PUF, CY2017, Interactive Dataset [31].

URL: https://www.cms.gov/.
2.2. Overview of the proposed method

In this paper, we propose a network-based method with privacy-preserving for identifying influential providers in large healthcare service systems. The proposed network-based method is composed of three main procedures.

The first procedure is the construction of the network, which uses the ZIP code and HCPCS code of the providers to build the provider-interacting networks. The key step in this procedure is to set a threshold value for the location of the neighboring providers, specify that providers within the threshold range can generate the connection relationship, and add weights to the edges according to the similarity of the types of medical services provided.

The second procedure is to rank influential nodes. The ranking of the nodes in the generated provider-interacting network is conducted in parallel on the basis of four nodal influence metrics, including the DC, CB, CC, and H-Index.

The third procedure is to evaluate the impact of the identified influential nodes. By removing a certain proportion of the top-ranked nodes, the impact on the three indicators of the network is evaluated, including the maximum connectivity coefficient [28], network efficiency [29, 30], and susceptibility [27], to evaluate the effectiveness of the influential nodes ranking algorithm.

The flowchart of the proposed network-based method for identifying influential providers in a healthcare service system is illustrated in Fig. 1.

2.3. Procedures of the proposed method

2.3.1. Construction of the provider-interacting network

First, we cleaned and processed the obtained data. By writing a specific Python script, we extracted the NPI, ZIP code, and HCPCS code of 898,257 providers from Physician and Other Supplier Data CY 2017 [31]. Address matching was used to map geographic entities to addresses (spatial locations). There are three methods for address matching: (1) by street address, (2) by ZIP code, and (3) by boundary information [34–36]. In this paper, we employed the second approach by using the ZIP code of the providers. The centroid coordinates based on ZIP code were converted into longitude and latitude and then represented by nodes on the map (Fig. 2). The size of each node is the weight added according to the number of different medical services provided by the providers. The greater the number is, the greater the weight, and the larger the point (Fig. 3).

We constructed provider-interacting networks $A(\delta)$ for 50 U.S. states ($\delta$), who have an effective NPI and submit Part B medical insurance services. The network can be abstracted as an undirected weighted network. The model of the undirected weighted network is given as a triple $G = (V, E, W)$, where $V = \{v_1, v_2, \ldots, v_n\}$ represents the set of nodes and $v_i \in V$. $E = \{e_1, e_2, \ldots, e_m\}$ represents the set of edges; $W = \{w_{ij}\}$ is the weight matrix of connected edges, where $w_{ii} = 0$ and $w_{ii} \geq 0$, $w_{ij}(i \neq j)$ represents the weight of edges $(v_i, v_j)$ [37]; see Eq. (1).

$$w_{ij} = w_{ji}(i, j = 1, 2, \ldots, n) \quad (1)$$

Each provider is regarded as a node. Node $v_i$ and node $v_j$ are connected by weighted lines. The weighted lines are created by considering (1) ZIP code proximity and (2) HCPCS code similarity.

(1) ZIP code proximity

We converted the centroid coordinates of each provider's ZIP code to latitude and longitude. Node $v_i$'s longitude is $X_i$, the latitude is $Y_i$; node $v_j$'s longitude is $X_j$, and the latitude is $Y_j$. If the absolute value of the difference between the longitudes of two nodes, i.e., $|X_i - X_j|$, and the absolute value of the difference between the latitudes, i.e., $|Y_i - Y_j|$, are both within a certain range, then the two providers are considered potential neighbors; see Eq. (2).

$$|X_i - X_j| \leq \alpha \quad |Y_i - Y_j| \leq \beta \quad (2)$$

According to the division of cities, districts, and streets, people who live in a city are likely to know each other. After statistical calculation, the average value of threshold $\alpha$ is $0.2^\circ$, and the average value of threshold $\beta$ is also $0.2^\circ$. 

**Fig. 1.** Flowchart of the proposed network-based method for identifying influential providers in a healthcare service system.
(2) HCPCS code similarity

The HCPCS is divided into two principal subsystems, referred to as level I and level II. Level I of the HCPCS is comprised of CPT (Current Procedural Terminology), a numeric coding system maintained by the American Medical Association (AMA). Level II codes are also referred to as alpha-numeric codes because they consist of a single letter of the alphabet followed by 4 numeric digits, while CPT codes are identified using 5 numeric digits. For each alpha-numeric HCPCS code, there is descriptive terminology that identifies a category of similar items.

In this paper, we specify that the set of HCPCS codes for medical services provided by node $v_i$ is represented by $S_i$, and the set of HCPCS codes provided by node $v_j$ is represented by $S_j$. If two providers have the same HCPCS code, they are considered potential neighbors, and the weight of the edge is $k$ (Eq. (3)).

$$k = |S_i \cap S_j|$$ (3)

Due to the large number of nodes in the nationwide provider-interacting network, we choose a representative provider-interacting network in California and Nevada for research. The distributions of the providers in California and Nevada are shown in Fig. 4. According to the above two rules for finding neighbors of providers, we generate local provider-interacting networks as shown in Figs. 5 and 6.
2.3.2. Ranking of the influential nodes

In network science, one of the most important research areas is how to rank influential nodes in a complex network [38]. It is interesting to rank nodes when there are a massive number of nodes and find the vital providers that have a strong impact on the function and efficiency of the whole network. For example, it is possible to replace the missing vital nodes in a timely manner to protect the network from paralysis and maintain the stability of the medical system [39].

Moreover, Gao et al. [40] proposed a local structural centrality measure to rank the spreading ability of nodes in the network. Zhao et al. [41] proposed the IM-LPA algorithm to solve the influence maximization problem in social networks with a community structure. Mo et al. [42] proposed a new comprehensive centrality measure based on Dempster–Shafer evidence theory.

The above algorithms can be roughly divided into two categories: one is based on local nodal influence metrics such as the degree centrality, and the other is based on global metrics such as the betweenness centrality, closeness centrality, and k-shell metrics [43].

In this paper, after considering the ranking metrics that are relatively suitable for undirected weighted provider-interacting networks, four nodal influential metrics are used to rank nodes, including (1) the degree centrality [11], (2) the companion behaviors [12], (3) the clustering coefficient [12], and (4) the H-Index [13,14].

1. Degree Centrality (DC)
DC is measured by the degree of a node, which can directly reflect the possibility of the node having direct contact with other nodes in the network (Eq. (4)).

$$DC_i = \frac{K_i}{n - 1}$$

(4)

DC$_i$ indicates the DC of node $v_i$ and $K_i$ indicates the degree of node $v_i$. To reflect the function of weighted edges, we specify how many HCPCS codes are the same between nodes $v_i$ and $v_j$, and how many connecting edges are there between nodes $v_i$ and $v_j$. When calculating the DC of node $v_i$, the number of added connected edges can be equivalent to the number of increased neighbors. For a network where node $v_j$ has $n$ connected edges, the maximum degree of a node is $n - 1$. After the node is divided by $n - 1$ for normalization, there is $0 \leq DC_i \leq 1$. The larger the value is, the stronger the centrality of the nodes is [37].

(2) Companion Behaviors (CB)

The local metric CB was originally proposed by Mei et al. [12] and is based on the calculation of Jaccard Coefficients (JCs) of edges. The JC reflects the difference between the neighbors of two nodes, that is, the relationship strength of two nodes (Eq. (5)). If the JC of two nodes is small, the relationship between the two nodes may be weak. Conversely, if the JC of two nodes is large, the relationship between the two nodes may be strong.

$$JC = \frac{|A \cap B|}{|A \cup B|}$$

(5)

Table 1

| Connection relationship | Number of identical HCPCS codes |
|-------------------------|---------------------------------|
| Node $v_i$ Neighbor 1   | 3                               |
| Node $v_i$ Neighbor 2   | 2                               |
| Node $v_i$ Neighbor 3   | 1                               |
| Node $v_i$ Neighbor $v_j$ | 5                               |
| Node $v_i$ Neighbor $v_j$ | 1                               |
| Node $v_i$ Neighbor 4   | 4                               |
| Node $v_i$ Neighbor 5   | 3                               |
| Node $v_i$ Neighbor $v_j$ | 4                               |

$v_i$ and $v_j$, which means that $v_i$, $v_j$, and $v_i$ are connected with 4 triangles. The edges of node $v_i$ and node $v_j$ with the other nodes are shown in Table 1. In Fig. 7(b), JC = $4/(3 + 2 + 1 + 5 + 1 + 4 + 3 + 4 - 4) = 4/19$. JC represents the weight of the edge whose value is always in the range of 0–1. The weight of the nodes is represented by CB, which means the sum of the JCs of the connected edge of the nodes (Eq. (6)).

$$CB(v_j) = \sum_{j=0}^{n} JC(v_j | v_i)$$

(6)

(3) Clustering Coefficient (CC)

CC reflects the tightness of the connection between neighbors. For a node $v_i$ with a degree of $K_i$, the number of edges between node $v_i$ and $K_i$ nodes is $K_i(K_i - 1)/2$, which is the case with the largest number of edges. If the number of edges between $v_i$ and its neighbors is $E_i$, the CC of a node can be calculated via Eq. (7).

$$C_i = \frac{E_i}{K_i(K_i - 1)/2}$$

(7)

In special cases, if the degree $K_i$ of the node is 0 or 1, the $C_i$ of the node is considered to be 0. Obviously, the nodal CC is also between 0 and 1 [20,44].

(4) H-Index

The definition of the H-Index proposed by J.E. Hirsch [14] is as follows. All $n$ papers published by an author are sorted in
The impact of removing a part of the nodes on the giant connected component (Eq. (9)).

The maximum connectivity coefficient of a network can be calculated by first ranking the nodes according to the nodal influence metrics from the largest to the smallest and then observing the impact of removing a part of the nodes on the giant connected component (Eq. (9)).

\[ G = \frac{R}{N} \] (9)

where \( N \) represents the total number of nodes in the network and \( R \) represents the giant connected component. The smaller the scale of the giant connected component after the removal of nodes, the more obvious the trend is, indicating that the effect of using this method to attack a network is better than that of other methods.

(2) Network efficiency

To investigate the effect of node removal on network efficiency, the network efficiency can be used to evaluate the connectivity of the network (Eq. (10)). To remove nodes and all their corresponding edges in the network, some paths in the network are interrupted, resulting in the shortest path between some nodes becoming larger, and the average path length of the whole network increases, affecting network connectivity [30].

\[ \eta = \frac{1}{N(N-1)} \sum_{i,j \in V} \eta_{ij} \] (10)

where \( \eta_{ij} = 1/d_{ij} \), \( d_{ij} \) is the shortest path between nodes \( v_i \) and \( v_j \), and \( N \) is the number of network nodes.

In this paper, we remove a certain proportion of specific nodes in the network to simulate the effect of a network attack and then calculate the network efficiency decline ratio before and after the attack to quantify the accuracy of each node influence metric.

The proportion of network efficiency decrease is expressed through Eq. (11).

\[ \mu = 1 - \eta / \eta_0 \] (11)

where \( \eta \) represents the network efficiency after node removal, \( \eta_0 \) represents the original network efficiency, and \( 0 \leq \mu \leq 1 \). The higher the value of \( \mu \) is, the worse the network efficiency becomes after removing the node.

(3) Susceptibility

The giant connected component decreases with an increase in the number of removed nodes and vanishes when a critical proportion \( p_c \) of nodes is removed [27]. To find the accurate \( p_c \) value, the susceptibility \( S \) of the network is calculated before and after removing a certain number of nodes (Eq. (12)).

\[ S = \sum_{s \in R} \frac{s^2 n_s}{n} \] (12)

where \( n_s \) is the number of components of size \( s \), \( n \) is the size of the whole network and \( R \) represents the giant connected component.

In general, there is a peak value of \( S \) at the critical proportion \( p_c \) of network collapses. If the network experiences multiple collapses during selective node deletion, there are multiple peaks, and the \( p_c \) value is determined by the highest one. Obviously, according to the objective function on network connectivity, the smaller the value of \( p_c \), the better the sorting algorithm.
2.4. Implementation details of the proposed method

In this subsection, we will introduce more implementation details on the development of the proposed network-based method for identifying influential providers in healthcare service systems. As has been described several times, there are three main procedures in the proposed method: (1) constructing the provider-interacting network, (2) ranking the influential nodes in the generated provider-interacting network, and (3) evaluating the impact of top-ranked influential nodes. This subsection will introduce more details on the above three procedures.

2.4.1. Implementation details of constructing the provider-interacting network

There are two steps in this procedure: the extraction of NPI, ZIP code, and HCPCS data and the construction of the provider-interacting network according to (1) the ZIP code proximity and (2) the HCPCS code similarity.

We wrote a Python script specifically to extract the NPI, ZIP code, and HCPCS of each provider from the raw dataset, and we obtained 898,257 providers from Physician and Other Supplier Data CY 2017 [31]. The entire Python script is presented in Listing 1.

We also specifically wrote C/C++ code to construct the local provider-interacting network in California and Nevada according to (1) the ZIP code proximity and (2) the HCPCS code similarity. After reading in the data of the NPI, ZIP code, and HCPCS of each provider, we first exclude those providers located outside California and Nevada according to the latitude and longitude. Those providers with latitudes in the range of 32.5 and 42 and with longitudes in the range of −124.5 and −114 are kept, while the rest are excluded. All the remaining providers are stored in a node list.

Listing 1: Python code for extracting the NPI, ZIP code, and HCPCS from the raw dataset

```python
import csv
f = open("2017rawdata.txt","r")
data = f.readlines()
print(len(data))
for line in data:
    newline = line.split()
    tempstr = newline[0]
    for i in range(1, len(newline)):
        word = newline[i]
        if word.isdigit():
            if len(tempstr) == 10 and len(word) == 9:
                tempstr += ',
            tempstr += word
            if len(tempstr) == 20 and len(word) == 5:
                tempstr += ',
            tempstr += word
            if len(tempstr) == 27:
                newdata.append(tempstr)
        # Save
        file=open('2017newdata.csv','w')
        file.write(eachline)
        file.close()
```

We then implement a double loop over all the providers stored in the node list to form the links between providers by (1) comparing the HCPCS codes and (2) evaluating the distance between any pair of providers. If two providers have the same HCPCS codes and the providers are also located within a given region (see Eq. (2)), then a link/edge is created and stored in the edge list. If there are duplicate edges, then those edges will be merged into a unique one using sorting and scanning operations, while the weight of the merged edge indicates the number of duplications.

2.4.2. Implementation details of ranking the influential nodes

In the proposed method, the ranking of nodes in the generated provider-interacting network is conducted in parallel on the basis of four nodal influence metrics, including the DC, CB, CC, and H-Index. In this paper, we directly employ the source code presented in our previous work [12] to rank the influential nodes. Much more implementation detail is provided in reference [12].

2.4.3. Implementation details of evaluating the impact of top-ranked influential nodes

In the proposed method, the impact of those top-ranked influential nodes in the provider-interacting network is evaluated by comparing three network indicators: the maximum connectivity coefficient, network efficiency, and susceptibility. This procedure is implemented by invoking the Stanford Network Analysis Platform (SNAP) [48]. Much more implementation detail is provided in reference [49].

3. Results

3.1. Experimental environment

The performance of the proposed method is evaluated on a workstation computer. Detailed specifications of the employed workstation computer are listed in Table 2.

| Specifications | Details |
|----------------|---------|
| CPU            | Intel Xeon Gold 5118 CPU |
| CPU Frequency (GHz) | 2.30 |
| RAM (GB)       | 128 |
| CPU core       | 48 |
| GPU            | Quadro P6000 |
| GPU memory (GB) | 24 |
| CUDA cores     | 3840 |
| OS             | Windows 10 professional |
| Compiler       | VS2015 community |
| CUDA version   | v9.0 |
| Anaconda version | Python 3.7 |

Table 2 Specifications of the workstation computer for testing the proposed method.

3.2. Experimental results

3.2.1. Results of ranking influential nodes

The frequency distribution of the four node-ranking metrics is shown in Fig. 8. Fig. 8(a) shows the frequency distribution of the unweighted node degrees, and Fig. 8(b) shows the frequency distribution of the weighted node degrees.

Fig. 8 shows that the unweighted degree of nodes ranges from 1 to 4050, of which 98.7% are less than 200; the weighted degree of nodes ranges from 1 to 14,247, of which 99.1% is less than 1000; and the CB of nodes ranges from 0 to 40,257, of which 98.6% are less than 2. The distributions of the degree and CB metrics of nodes are extremely irregular from low to high, and most of them are low-value metrics.

The CC of nodes ranges from 0 to 1, of which 20.3% are less than 0.05; the H-Index of nodes ranges from 1 to 99, of which 34.1% are less than 5. The distributions of the CC and H-Index metrics of nodes are relatively regular from low to high, but the proportion of low values is the largest. The general trend is that with increasing of CC and H-Index values, the proportion decreases.


3.2.2. Results of impact evaluation of the top-ranked influential nodes

To evaluate the influence of static attacks (i.e., the node importance remains unchanged) on the maximum connectivity coefficient, network efficiency, and susceptibility of the network, we selected 11 groups of removal ratios (0%, 0.1%, 0.2%, 0.3%, 0.4%, 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1%) for testing; see Tables 3–6 and Fig. 9 for the detailed test results.

Fig. 9 shows the effect of removing the top-ranked nodes with different scales on the maximum connectivity coefficient, network efficiency, and susceptibility. After removing the top-ranked nodes, the trend in the giant connected component becoming smaller is more obvious, the more obvious the decline in the maximum connectivity coefficient is, and the worse the network efficiency is. Fig. 9(a) and (b) show that when removal ratios reach 0.5%–1%, using the metric DC to remove the top-ranked nodes leads to the largest decline in the maximum connectivity coefficient and network efficiency, followed by the metrics CB and H-Index. The metric CC leads to a slight decrease in the maximum connectivity coefficient and network efficiency, which is almost unchanged. In Fig. 9(c), the metrics CC and H-Index reduce the susceptibility by a small margin, which is almost unchanged, while the metric CB reduces the maximum connectivity coefficient and network efficiency by the largest margin, followed by the metric DC. The above results show that, in the provider-interacting networks,

Fig. 8. Frequency distributions of node ranking using four nodal influence metrics.
when selectively removing the top 0.1%-1% of nodes for each metric, the efficiency, connectivity and susceptibility of the network become the worst when using the metrics DC and CB to remove top-ranked nodes.

4. Discussion

4.1. Advantage of the proposed method

The most obvious advantage of the proposed method is that it does not require the information of personally private medical claims.

As mentioned above, patient-sharing networks are created according to electronic medical claims, and the advantage of creating patient-sharing networks is that they are quite precise. However, the data of electronic medical claims are personal and private [32,33]. Personally private information needs to be carefully addressed before creating the required patient-sharing networks [50,51].

In contrast, in the proposed method, the provider-interacting network joined by healthcare providers can be roughly created according to the location and the available types of healthcare services, without the need for personally private electronic medical claims. For example, the data of the location and the provided types of healthcare service are publicly available at the U.S. Centers for Medicare & Medicaid Services (https://www.cms.gov/). This is the main advantage of the proposed method.

4.2. Shortcomings of the proposed method

One of the essential ideas behind the proposed method is the construction of the provider-interacting network, rather than the commonly used patient-sharing networks. Those patient-sharing networks are typically constructed based on medical claims data. Medical claims are usually quite accurate, and the patient-sharing network is also quite accurate in reflecting the relationships between providers.

Compared with those patient-sharing networks constructed based on medical claims data, the provider-interacting network is generated according to the location and available healthcare service of the providers, more specifically, according to the ZIP codes and HCPCS codes of the providers. In the proposed method, several nearby providers located within a certain distance who provide the same HCPCS codes are linked and considered neighbors in the provider-interacting network. This procedure may not be accurate because it is not able to determine the “best” distance threshold for selecting neighboring providers. In fact, the distance threshold is specified. Thus, the generation of the provider-interacting network is not as accurate as that of the patient-sharing network.

4.3. Applicability of the proposed method

The proposed method can be applied to help make decisions for healthcare systems management, such as the optimization of healthcare resources in the response to public health emergencies [52]. For example, since mid-December 2019, coronavirus disease 2019 (COVID-19) has been spreading from Wuhan, China. The epidemic quickly disseminated from Wuhan, and as of 12 February 2020, 45,179 cases have been confirmed in 25 countries, including 1,116 deaths [53]. As of March 1, 2020, it has spread to 56 other countries [54]. Recognizing the Wuhan-focused and nationwide outbreak responses in China, the WHO has encouraged countries with heavy air travel exchanges with Wuhan to take precautionary public health measures and, if there is an imported infection, to undertake activities that could lead to the elimination of the virus in human populations, as occurred during the 2003 SARS outbreak [55]. Rapid and effective collaboration between the clinicians (e.g., the general practitioners attending the cases, emergency hotline clinicians, and infectious diseases specialists), the National Reference Centre and the regional and
national health authorities has played a crucial role in the systemic capacity to quickly detect, isolate and investigate cases to implement adequate control measures [56,57].

In China, during the period of epidemic prevention, the demand for front-line medical staff, disinfection materials, protective equipment and emergency supplies in the respiratory tract infection department increased dramatically. In contrast, many patients with other diseases went to the hospital as little as possible; thus, the investment in diagnosis and treatment in other departments of the hospital decreased, and the required medical resources (e.g., the labor and material resources) decreased accordingly.

It is quite meaningful to employ the proposed method in this paper to build a network of hospital providers, identify the important providers remaining to maintain the normal operation of local medical services, and transfer other redundant providers and available medical materials to support areas heavily affected by the epidemic. In this way, the proposed method can not only effectively ensure that local hospitals maintain their capability to provide healthcare services but can also reasonably distribute medical resources, which plays an important role in epidemic prevention. Moreover, more attention needs to be paid to strengthening the distribution of personnel and facilities in district and community level medical institutions. According to urban planning, the method in this paper ensures that public medical resources can flow between different levels corresponding to different geographical units and can be transferred at the grassroots level to establish a resilient and hierarchical healthcare service system.

4.4. Outlook and future work

Due to the large number of nodes in the provider-interacting network, in this paper, we only choose the providers in California and Nevada for research. Moreover, in the proposed method, several nearby providers located within a certain distance who provide the same HCPCS codes are linked. This procedure may not be accurate because it is not able to determine the "best" distance threshold to select neighboring providers. In fact, the threshold of distance is specified. In the future, we will generate a nationwide provider-interacting network for further research and optimize the selection of a proper distance threshold.

Moreover, with significant advances in communication technologies, an era of “Internet of Things” (IoT) appears [58,59]. A large amount of IoT data can be collected in various ways [60,61] which may be used to generate networks for healthcare service systems. Identifying influential providers in those generated network-based systems can also help optimize and distribute the healthcare service resource. In the future, we will also conduct research work in this field.

5. Conclusions

Identifying the most influential providers in a healthcare service system is an important step towards optimizing the use of available healthcare resources and ensuring the more efficient delivery of matched healthcare services. In this paper, we proposed a network-based method with privacy-preserving for identifying influential providers in large healthcare service systems. First, we constructed a provider-interacting network by employing the publicly available information on the locations and

Fig. 9. The influence of removing the top nodes with different proportions on the (a) maximum connectivity coefficient, (b) network efficiency, and (c) susceptibility.
types of healthcare service of providers. Second, we ranked the nodes in the generated provider-interacting network on the basis of four nodal influence metrics. Third, we evaluated the impact of those top-ranked influential nodes in the provider-interacting network by comparing three network indicators. The proposed method is demonstrated by employing the dataset of Physician and Other Supplier Data CY 2017, and can be applied to other similar datasets to help make decisions for healthcare system management, such as the optimization of healthcare resources in the response to public health emergencies.

Declaration of competing interest

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.

CRediT authorship contribution statement

Xiaoyu Qi: Conceptualization, Data curation, Methodology, Software, Writing - original draft. Gang Mei: Supervision, Conceptualization, Data curation, Methodology, Writing - original draft, Writing - review & editing. Salvatore Cuomo: Investigation, Writing - review & editing. Lei Xiao: Software, Validation.

Acknowledgments

This research was jointly supported by the National Natural Science Foundation of China (Grant Numbers: 11602235), and the Fundamental Research Funds for China Central Universities (Grant Numbers: 2652018091, 2652018107, and 2652018109). The authors would like to thank the editor and the reviewers for their valuable comments.

References

[1] M.E.J. Newman, Communities, modules and large-scale structure in networks, Nat. Phys. 8 (1) (2011) 25–31, http://dx.doi.org/10.1038/nphys2162.
[2] D. Tsioras, Detecting different topologies immanent in scale-free networks with the same degree distribution, Proc. Natl. Acad. Sci. USA 116 (14) (2019) 6701–6706, http://dx.doi.org/10.1073/pnas.1816842116.
[3] D.R. Lo Sardo, S. Thurner, J. Sorger, G. Duftschmid, G. Endel, P. Klimek, Quantification of the resilience of primary care networks by stress testing the health care system, Proc. Natl. Acad. Sci. USA 116 (48) (2019) 23930–23935, http://dx.doi.org/10.1073/pnas.1904826116.
[4] R. Robbins, A. Seixas, A. Schrothaler, The nature and scope of patient-sharing network research: a novel, important area for network science, Transl. Behav. Med. 8 (4) (2018) 626–628, http://dx.doi.org/10.1093/iby/052.
[5] C.E. Pollack, G.E. Weissman, K.W. Lemke, P.S. Hussey, J.P. Weiner, Patient sharing among physicians and costs of care: a network analytic approach to care coordination using claims data, J. Gen. Intern. Med. 28 (3) (2013) 459–462, http://dx.doi.org/10.1007/s11606-012-2104-7.
[6] M.S. Ong, K.L. Olson, A. Cami, C. Liu, F. Tian, N. Selvam, K.D. Mandl, Provider patient-sharing networks and multiple-provider prescribing of benzodiazepines, J. Gen. Intern. Med. 31 (2) (2016) 164–171, http://dx.doi.org/10.1007/s11606-015-3470-8.
[7] E.L. Molen, A.M. Austin, J.P. Bynum, J.S. Skinner, A.J. O’Malley, An analysis of patient-sharing physician networks and implantable cardioverter defibrillator therapy, Health Serv. Outcomes Res. Methodol. 16 (3) (2016) 132–153, http://dx.doi.org/10.1007/s11097-016-9152-x.
[8] S. Linde, The formation of physician patient sharing networks in medicare: exploring the effect of hospital affiliation, Health Econ. 28 (12) (2019) 1435–1448, http://dx.doi.org/10.1002/hec.3936.
[9] B.E. Landon, N.L. Keating, J.P. Onnela, A.M. Zaslavsky, N.A. Christakis, A.J. O’Malley, Patient-sharing networks of physicians and health care utilization and spending among medicare beneficiaries, JAMA Intern. Med. 178 (1) (2018) 66–73, http://dx.doi.org/10.1001/jamaiamert.2017.5034.
[10] E.H. DuGoff, S. Fernandes-Taylor, G.E. Weissman, J.H. Huntley, C.E. Pollack, A scoping review of patient-sharing network studies using administrative data, Transl. Behav. Med. 8 (4) (2018) 598–625, http://dx.doi.org/10.1093/ibm/txy015.

X.Q. Qi, G. Mei, S. Cuomo et al. / Future Generation Computer Systems 109 (2020) 293–305
Xiao Yu Qi is currently a bachelor student and will start her Ph.D study in September 2020 at China University of Geoscience (Beijing). Her main research interests are in the areas of Numerical Simulation and Computational Modeling, Data Mining, and Network Science and Applications.

Gang Mei is an Associate Professor in Numerical Modeling and Simulation at China University of Geosciences (Beijing). He received his Ph.D degree in 2014 from the University of Freiburg in Germany. He obtained both his bachelor and master degrees from China University of Geosciences (Beijing). His main research interests are in the areas of Numerical Simulation and Computational Modeling, GPU Computing, Machine Learning, Data Mining, and Network Science and Applications. He has published more than 50 research articles in journals and academic conferences. He is the IEEE Member, and has served as an Associate Editor for the journals PeerJ Computer Science and IEEE Access since 2018.

Salvatore Cuomo is an Associate Professor in Numerical Analysis at University of Naples Federico II, Italy. His research field interests are in Mathematical models for applied sciences, Numerical Approximation theory and applications, Parallel and Distributed computing. Inverse problems arising from image analysis and Information Technology for automatic data processing in medicine and teaching environments.

Lei Xiao is currently a Ph.D student at China University of Geosciences (Beijing). His research interests are in the areas of Numerical Simulation and Computational Modeling, including Computational Geometry, FEM Analysis, and GPU Computing.