Disease-Based Clustering of Hospital Admission: Disease Network of Hospital Networks Approach

Nouf Albarakati[1,2], Zoran Obradovic[1]
[1]Center for Data Analytics and Biomedical Informatics
Temple University, Philadelphia, USA
{nouf, zoran.obradovic}@temple.edu
[2]Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—To improve the quality of healthcare planning, healthcare systems face challenges in identifying clusters of similar hospitals while considering varying factors. Clustering hospitals based on their admission behavior would be helpful whereas diagnosis of patients is vital in understanding variation in admission. Therefore, grouping hospitals that show similar behavior on their admission distribution while considering similarity among disease symptoms in admission is the objective of our study. This is achieved by a Disease Network of Hospital Networks model which is used to represent hospital admission distribution of multiple diseases as different hospital networks that correspond to disease nodes in a top-layer disease network. This disease network that was extracted from the Human Symptoms Disease Network models the similarity among different disease-specific hospital networks. We assume that disease-specific hospital networks have different underlying clustering structure while share the same underlying clustering structure if corresponding diseases share similar symptoms. Experiments were conducted on more than 14 million electronic health records of monthly admission of 160 diseases over 4 years at 301 hospitals in California. Results of clustering 160 disease-specific hospitals networks that share similar symptoms among corresponding diseases show consistent behavior among these networks when similarity among diseases is considered in clustering process. Patterns of consistent behavior were lacking in results when similarity among diseases is not considered.**

Keywords—hospital admission networks; disease symptom network; graph clustering; network of networks

I. INTRODUCTION

Healthcare planning has been recognized to have a significant impact on community health, and economy [1]. On their mission to improve quality of planning and decision making, healthcare systems face challenges in identifying groups or clusters of similar hospitals considering varying factors and/or different configurations [2]. Analyzing clusters of hospitals that show similar admission behavior plays an important role in helping healthcare facilities to adjust their plans and policies to society and patient needs [3,4].

Clustering hospitals based on admission behavior would help to understand the effect of certain factors that are related to hospitals size, specialty, and/or spatial characteristics [2,5]. On the other hand, diagnosis of patients is vital in understanding variation in admission rate among hospitals [4,6]. Therefore, grouping hospitals that show similar behavior on their admission distribution while considering patient diagnosis in admission is a promising objective. However, in previous work, clustering has been applied for grouping different diseases without considering similarity among disease symptoms [7,8] which is a limitation that is addressed in this study.

This finding invites an obvious question about the benefit of considering the similarity among disease symptoms in analyzing clustering of hospital admission. That is, if there is a way to cluster hospitals that have similar admission behavior for a certain disease, would the similarity among disease symptoms have a beneficial effect on clustering results? Let's assume that there is a different weighted graph of hospitals for every disease. Each graph represents a network of hospitals that admitted patients for a certain disease. Nodes represent hospitals and links represent the similarity between hospitals' monthly admission distributions. To cluster each network into groups of hospitals that show similar behavior on their admission distribution, do these networks share one underlying clustering structure? Are there multiple underlying clustering structures across different networks where some networks may share one underlying clustering structure if they share similar symptoms?

Examining these challenging questions is the main contribution of this work. It analyzes hospitals clustering based on their monthly admission distribution for different types of diseases. For each disease, there is a different clustering result for all hospitals that have admitted patients for that disease. To understand the underlying clustering structure across different networks of disease-specific hospitals, the clustering process was guided by similarity among disease symptoms.

This study relies on two other studies that have significant impact in their fields. First, NoNClus method is proposed as a clustering method that allows capturing multiple underlying clustering structures across different networks [9]. Another fundamental study in conducting this research is the Human Symptoms Disease Network (HSDN) [10]. This symptom-based disease network was used to generate disease network [11] to guide clustering of hospital admission for different diseases.

The main contribution of this study is characterization of the underlying clustering structure of 160 disease-specific hospital networks when diseases they represent share similar symptoms. We show that disease-specific hospital networks share the same underlying clustering structure if corresponding diseases share similar symptoms.

II. Background and related work

For the past 30 years, different types of clustering algorithms were used to define hospital clusters [1,2]. These algorithms were applied in multiple settings using different types of data that range from patient level, to hospital level data. In this paper, filling the gap in studying the effect of similarity among disease symptoms on disease-specific clustering of hospitals led to this combinatorial work between Network of Networks clustering method NoNClus and HSDN, symptom-based disease network.

A. Clustering Algorithm

The revolution of data analytics in the last decade shifted the interest from simple networks towards more advanced heterogeneous information networks where different data forms and sources can be integrated. Many approaches have emerged to explore hidden patterns and valuable information lie within such data. One of these approaches is merging different networks into a multi-layered 'Network of Networks' architecture either into multi-view networks or multi-domain networks [12]. Network of Networks structure is used to illustrate a network at different scales. A largescale network is composed of several sub-networks, and the interconnectivity between these subnetworks are known to be crucial to information distribution [13].

One of the frameworks that was proposed in addressing clustering in multi-networks is NoNClus [9]. It was designed to cluster complex multi-layered networks that has a Network of Networks structure. It allows multiple underlying clustering structures across different networks. On the other hand, most of the previous work that addressed multi-network clustering share the assumption that these networks have a single common clustering structure although different networks usually have different data distributions [14-16].

NoNClus models the clustering structure in the top-layer network, which can be used to regularize the clustering structures in different domain-specific networks that every node in the top-layer network represents. In other words, it partitions the domain-specific networks while respecting the clustering structure obtained initially from the top-layer network.

B. Disease Network

Influenced by the claim that symptoms are critical in clinical diagnosis and treatments, Human Symptoms Disease Network (HSDN) was constructed by Zhou et al. in 2014 [10]. This weighted network was generated using both Medical Subject Headings (MeSH) terminology and a big medical bibliographic literature database, PubMed [10]. MeSH was used to index all articles in PubMed resulting a total of over four thousand disease terms and over three hundred symptoms terms. After identifying the association between diseases and symptoms, every disease was described by a vector of related symptoms. The similarity between two vectors of two diseases was calculated using cosine similarity measure. This measure ranges from 0 with no shared symptoms to 1 which means both diseases shared identical symptoms.

To study the effect of the similarity among disease symptoms in disease-specific clustering of hospital admission, the disease network is used as a top-layer network in Disease Network of Hospital Networks data model. Since the hospital admission networks for each disease was constructed using the California State Inpatient Database, CCS code was used to code diseases. Therefore, disease network that was extracted from HSDN in [11] is used. Glass et al. did the matching between the CCS codes and the MeSH terminology manually [11]. They used the average of similarities in some cases where the matching was not one-to-one.

III. Methodology

This work aims to understand the underlying clustering structure of different disease-specific hospital networks when diseases they represent share similar symptoms. We assume that disease-specific hospital networks have different underlying clustering structure whereas share the same underlying clustering structure if corresponding diseases share similar symptoms. To test this assumption, two clustering settings are proposed: NoN-cluster and Single-cluster.

A. Disease Network of Hospital Networks data model

In the proposed architecture, data is represented in two layered weighted networks: the top-layer disease network layer and a disease-specific hospital networks layer. Every disease node in the top-layer network represents a disease-specific hospital network in the lower level.

The top-layer disease network, was constructed as in [11] using HSDN. Nodes in this network are restricted to only represent diseases that are included in this study. Links among the top-layer disease nodes (sub-networks) quantify the similarity of symptoms between corresponding diseases. They would be beneficial for studying the characterizing interconnectivity and interdependency among sub-networks.

Disease-specific hospital networks have nodes representing hospitals admission for the corresponding diseases considered in this study. Edges between these nodes represent similarities between hospitals' monthly admissions. Gaussian Kernel was used as a similarity measure for pattern analysis to measure similarities as it is shown in equation 1.

$$(h_i, h_j) = \exp\left(\frac{-\|h_i - h_j\|^2}{\sigma^2}\right) \quad (1)$$

It measures the similarity between hospitals i and j. h_i and h_j are input vectors that represent monthly admission distributions for hospitals i and j, respectively. It has a value of 1 if two hospitals have identical monthly admission distributions, and 0 as their admission distribution moves further apart.

B. Clustering Settings

The top-layer disease network and disease-specific hospital networks are used in both proposed settings. This method is explained in detail in the following subsection. The NoNClus method is also applied to both settings. Although the NoNClus method allows specifying different number of clusters among domain specific networks, our experiments were unified and the number of hospital clusters was predefined as t=3 for both settings. This predefined cluster number is a simple number that has been chosen to keep this setting as simple as possible.

NoN-cluster setting is proposed as an application for clustering method, NoNClus. This algorithm allows multiple underlying clustering structures across different disease-specific

hospital networks. The predefined number of disease cluster (DC) for the top-layer network is k=3. This simple number has been tested and given meaningful results. It implies that the underlying clustering structure among different disease-specific hospital networks are different. However, some networks may share the same underlying clustering structure if these networks belong to a bigger group. For example, disease-specific hospital networks that represent diseases that share similar symptoms, i.e. belong to the same top disease cluster, may share the same underlying clustering structure.

Single-cluster setting is a baseline approach which assumes that all disease-specific hospital networks belong to the same underlying disease-level cluster. Single-cluster setting is implemented by pre-specifying the number of disease clusters for the top-layer network to one (k=1). It means that all disease-specific hospital networks share the same underlying clustering structure.

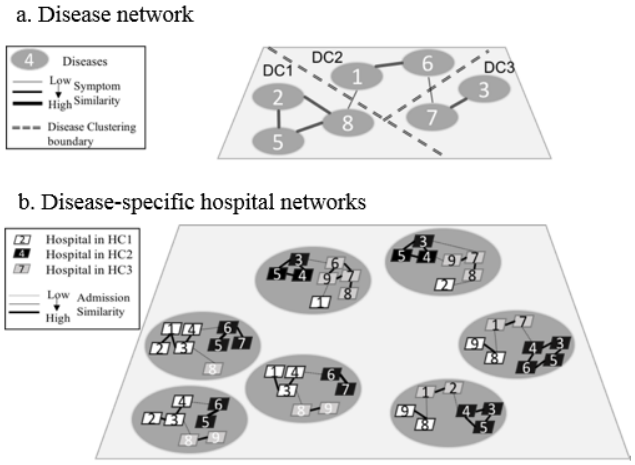

Fig. 1. Disease Network of Hospital Networks model: (a) A main network is disease network extracted from HSDN, nodes represent diseases and links represent symptoms similarity between diseases. (b) Domain-specific networks are disease-specific hospital networks, a network for each disease. Nodes in each network represent hospitals and links represent admission similarity between hospitals for a specific disease. Every network in (b) corresponds to a node in (a).

Figure 1 illustrates the Disease Network of Hospital Networks data model when NoNClus method is applied using NoN-cluster setting. The top-level network in Figure 1 is the top-layer disease network. It shows diseases and symptom similarity among them in the form of nodes and links. In phase I of NoNClus method [9], the disease network is clustered based on the similarity among diseases. Resulted clusters in this level are labeled as Disease Clusters, DC1, DC2, and DC3. This clustering structure is used in phase II to regularize clustering of disease-specific hospital networks shown below disease network. The bottom set of networks are the different disease-specific hospital networks, each network corresponds to a disease node in the disease network at the top-level network. Nodes in these networks represent hospitals that have admitted patients for the specific disease and links represent similarity among hospitals' monthly admission. Resulted clusters in hospital level are labeled as Hospital Clusters, HC1, HC2, and HC3. For better explanation, diseases 2,5, and 8 are grouped in one cluster due to the strong similarity between them as it is shown in Figure 1. This clustering structure affected clustering corresponding disease-specific hospital networks below. Hospitals in the three hospital networks have tended to share similar clustering structure. For instance, hospitals 1,2,3 and 4 are grouped together in the first cluster (white). Hospitals 5,6 and 7 tend to be grouped together (black) while Hospitals 8 and 9 are grouped in the third cluster (light grey). Although hospital 1 does not have admission for disease 5 and hospital 2 does not have admission for disease 8, both hospitals have tendency to be clustered with the same hospitals in similar symptoms diseases when their admission data are available due to flexibility of the NoNClus method.

C. NoNClus Method

NoNClus is a clustering method designed for complex networks that have Network of Networks structure. It improves graph clustering accuracy by integrating multiple graphs or networks while allowing multiple underlying clustering structures across different networks to find non-overlapping clusters [9]. NoNClus models the clustering structure in the main network, which can be used to regularize the clustering structures in domain-specific networks [9]. In our application, it partitions the disease-specific hospital networks while respecting the clustering structure obtained initially from the disease network.

NoNClus works as a two-phase method. In phase I, NoNClus method starts by solving a single network clustering problem to partition the top-layer disease network. It uses a symmetric non-negative matrix factorization by minimizing the following objective function [9]:

$$J_M = \|\mathbf{G} - \mathbf{H}\mathbf{H}'\|_F^2 \tag{2}$$

In equation (2), G is a g×g matrix where g is the number of diseases considered in this study. This matrix represents the similarity of symptoms among all diseases in the top-layer disease network. H is a g×k factor matrix of the disease network, G, where k is the number of clusters defined for the top-layer disease network. This factor matrix, H, defines the probability for each disease node to belong to one of the main clusters. That is, every element, h_{ij}, in H, determines the probability of the disease i^{th} to fit in the j^{th} cluster [9]. For the single-cluster setting, H factor matrix was predefined as a g×1 vector where h_i =1. It sets the probability of the i^{th} disease to belong to the only cluster with 100% probability.

In phase II, the factor matrix of the disease network is incorporated in clustering disease-specific hospital networks. However, NoNClus is developed to handle domain-specific networks that may have different number of nodes and clusters by minimizing the following objective function [9]:

$$\min_{\substack{\mathbf{U}^{(i)}\geq 0\ (i=1,\ldots,g)\\ \mathbf{V}^{(j)}\geq 0\ (j=1,\ldots k)}} J_D = \underbrace{\sum_{i=1}^{g}\|\mathbf{A}^{(i)} - \mathbf{U}^{(i)}(\mathbf{U}^{(i)})'\|_F^2}_{domain-specific\ network\ clustering} + \alpha \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{k} h_{ij}\|\mathbf{U}^{(i)} - \mathbf{V}^{(j)}\|_F^2}_{main\ cluster\ guided\ regularization} \tag{3}$$

This objective function of phase II also applies a symmetric non-negative matrix factorization with a factor matrix, H, as guided regularization to get $U^{(i)}$, an n_i×t factor matrix of $A^{(i)}$, where n_i is the number of hospitals in the i^{th} disease-specific hospital network and t is the number of clusters in disease-specific hospital network. $A^{(i)}$ is the matrix of the i^{th} disease-

specific hospital network that represents the similarity in admission for the i^{th} disease (i=1, …, g) [9].

The first term in the objective function of phase II, equation 3, deals with clustering the disease-specific hospital networks individually based on a similarity matrix of hospitals admission for every i^{th} disease [9]. The second term regularizes $U^{(i)}$, the factor matrix of i^{th} disease-specific hospital networks, using main clustering structure defined in h_{ij} and the underlying clustering structure of domain-specific networks of main cluster j defined in $V^{(j)}$. Since there are k main clusters, $V^{(j)}$ was introduced as a k hidden cluster to represent the underlying structure of disease-specific hospital networks in the main cluster j (j=1, …, k) [9].

IV. Experiments

In this section, the two clustering settings were evaluated on Disease Network of Hospital Networks using NoNClus method.

A. Data

A Disease Network (*one network of 160 nodes*) is the network extracted in [11] from HSDN [10] to represent diseases in CCS code instead of MeSH terminology. As some diseases have incomplete information in their study period and other diseases have no matching MeSH term, the total number of CCS disease codes included in this network is 189 [11]. This number we have reduced to 160 nodes after eliminating diseases that were represented in less than 50% of California hospitals.

Disease-Specific Hospital Networks *(160 networks, each with up to 301 nodes)* vary in structure and in number of hospitals. Hospital data used in this study are extracted from the California State Inpatient Database (SID) as part of the Healthcare Cost and Utilization Project (HCUP) provided by the Agency for Healthcare Research and Quality (AHRQ). For each disease included in this study, hospital monthly admission data between 2008 and 2011 is aggregated for diseases listed as a principle diagnosis disease from 14,534,016 single-patient discharge records included in the California SID. The total number of hospitals used in this study is 301 hospitals out of 500 California hospitals as there some hospitals had no admission records for some diseases over the study period. A hospital was eliminated from the study when it won't be represented in more than 50% of disease-specific hospital networks.

B. NoNClus method settings

All disease-specific hospital networks were clustered into three hospital clusters, t=3. NoNClus method was applied for every year separately of the four years (2011 - 2008).

NoN-cluster setting assumes that the underlying clustering structure across different disease-specific hospital networks is different. Therefore, the top-layer disease network is clustered into three main clusters, k=3. Analysis of these three main clusters showed meaningful groups. One cluster includes diseases related to infections and parasitic, gastrointestinal system, circulatory system and respiratory system. The second cluster includes diseases related to nervous system and sense organs, cerebrovascular accidents, musculoskeletal system and connective tissue, and injuries and poisoning. The last cluster of the disease network includes diseases related to congenital anomalies and complications of pregnancy and childbirth.

Single-cluster setting assumes that different disease-specific hospital networks share the same underlying clustering structure. Therefore, the top-layer disease network is clustered into one and only one cluster, k=1. Also, the H factor matrix was predefined to give the probability of i^{th} disease to belong to the only cluster with 100% probability.

V. Results

The NoNClus method produced a clustering label, HC1, HC2, or HC3, for every hospital in every disease-specific hospital networks to which it belongs. If the hospital does not treat a certain disease or has not admitted a patient to be treated from that disease, that hospital is not included in that specific disease-specific hospital network. For presentation purpose, cluster label is set to 0 for hospitals that have no admission data for a certain disease. Therefore, a table of clustering labels of

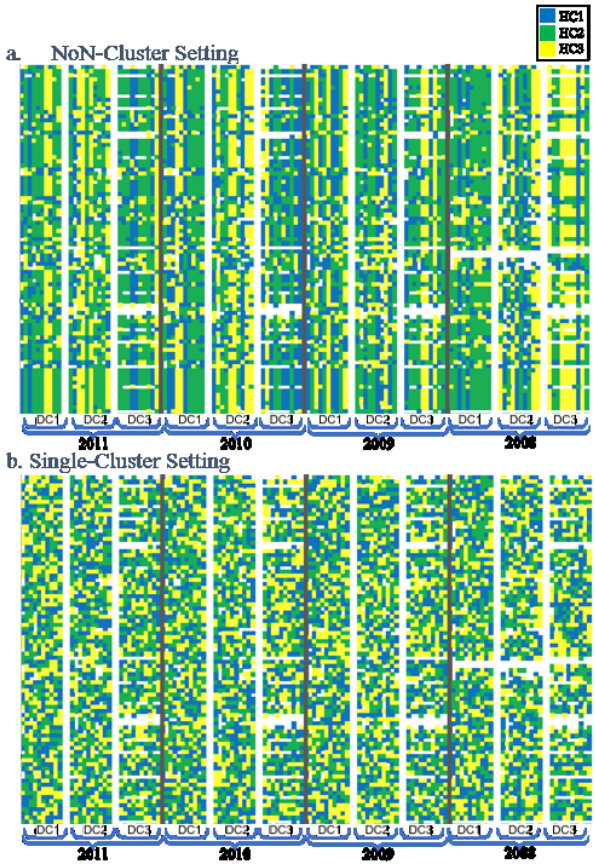

Fig. 2. Clustering results of disease-specific hospital networks for both settings: NoN-cluster and Single-cluster for four years (2011-2008). Results for 10 diseases-specific networks are presented for every disease cluster for both settings. (y-axis: 133 hospitals, x-axis: 10 disease–based networks). Blue: a hospital belongs to HC1, Green: a hospital belongs to HC2, Yellow: a hospital belongs to HC3

301 hospitals (rows) for the different 160 diseases-based hospital networks (columns) was produced for admission data of every four years included in the study.

An analysis of obtained results, shows that NoN-cluster's results were consistent when a group of hospitals that belong to the same cluster in a single disease-specific network tend to stay

together in the same cluster in other disease-specific networks if these networks correspond to diseases that belong to the same top-level cluster of diseases. On the other hand, patterns of consistent behavior were lacking in Single-cluster's results when clustering data produced with no regulation made based on disease network. This finding contradicts the assumption that different domain networks share the same underlying clustering structure.

Figure 2 shows the difference between clustering results produced using both settings for NoNClus method: (a) NoN-cluster setting and (b) Single-cluster setting. Part of both results were displayed for visualizing purposes to give a big picture of results. These results were color coded to give an overview of the clustering behavior over years and among different settings and different disease-specific networks. Blue color denotes hospitals that belong to hospital cluster HC1, green is for hospitals that belong to hospital cluster HC2 and yellow is for hospital cluster HC3. It is important to take into consideration that a cluster in one year might be labeled differently in other years as clustering is done separately for each year. Rows in Figure 2 represent hospitals. It shows consistent behavior in results among the first 133 hospitals shown over years for NoN-cluster setting. As drawn beneath each subfigure in Figure 2, hospital clustering results are shown for four years (2011-2008) admission data. Years were separated by a dark line. In every year, 30 out of 160 disease-specific networks were chosen for presentation. 10 disease-specific networks were chosen for every main cluster: DC1, DC2 and DC3. These main clusters were separated by a white line. These networks correspond to diseases that have high probability to belong to one of these three main clusters. These disease's CCS codes their probability values are shown in Table 1.

To understand the resulted pattern when considering the symptom similarity between diseases corresponding to disease-specific hospital networks, fewer disease-specific hospital networks were selected for further analysis. Hospital clustering of disease-specific hospital networks that represent diseases with high probability ($p>0.80$) to be one of the three disease clusters in the top-layer disease network is analyzed. Four related disease-specific hospital networks that represent four diseases in each of the top-level clusters were chosen for the extensive analysis due to the paper's length restriction. Diseases in the first set (CCS codes 4,99,100,101) are from the top-level DC1 cluster of diseases. They are mainly liver-related diseases. Diseases from the second top-level cluster DC2 are genitourinary-related diseases (CCS codes 42,111,113,145) whereas diseases in the third top-level cluster DC3 are related to pregnancy and childbirth (CCS codes 118,120,122,123). These diseases showed tendency to be grouped together.

TABLE 1. DISEASES' CCS CODES AND THEIR PROBABILITY VALUES TO BELONG TO MAIN DISEASE CLUSTERS 1,2, OR 3 FOR CORRESPONDING CHOSEN DISEASE-SPECIFIC NETWORKS

Main Cluster 1		**Main Cluster 2**		**Main Cluster 3**	
CCS code	**Prob. in DC1**	**CC S code**	**Prob. in DC2**	**CCS code**	**Prob. In DC3**
14	0.834	143	0.924	156	0.967
28	0.852	172	0.938	153	0.967
125	0.857	188	0.952	159	0.967
6	0.857	53	1	155	0.967
17	0.866	187	1	151	0.967
104	0.917	145	1	160	0.967
118	1	81	1	161	0.967
132	1	204	1	148	0.969
133	1	57	1	149	0.969
131	1	56	1	158	0.999

These 12 diseases have high probability to belong to their top-level clusters. Therefore, their regularization effect on clustering related disease-specific hospital networks is high. About 50% of the hospitals tend to be grouped together in some disease-specific hospitals networks that represent diseases belonging to the same top-level cluster of diseases. Therefore, for each top-level cluster of diseases, related disease-specific hospital networks were analyzed in Table 2.

TABLE 2. PERCENTAGE OF HOSPITALS BELONG TO CERTAIN DISEASE-SPECIFIC CLUSTERS FOR THE CHOSEN DISEASE-SPECIFIC HOSPITAL NETWORKS FOR EACH OF THE 3 TOP-LEVEL CLUSTERS OF DISEASES. PERCENTAGE COLOR CODES: RED: >80%, YELLOW: >70%, GREEN: >60%, AND BLUE >50%

		Year 2011				Year 2010				Year 2009				Year 2008			
	CCS	100	101	4	99	100	101	4	99	100	101	4	99	100	101	4	99
	Code	HC3	HC3	HC2	HC2	HC2	HC2	HC2	HC2	HC1	HC1	HC2	HC3	HC2	HC2	HC2	HC2
Diseases of DC1	100	0.814	0.671	0.023	0.003	0.754	0.721	0.701	0.661	0.694	0.641	0.146	0.003	0.738	0.671	0.638	0.585
	101	0.671	0.698	0.023	0.003	0.721	0.784	0.731	0.684	0.641	0.718	0.030	0.013	0.671	0.741	0.654	0.621
	4	0.013	0.013	0.771	0.608	0.701	0.731	0.821	0.654	0.010	0.010	0.817	0.020	0.638	0.654	0.751	0.568
	99	0.133	0.066	0.608	0.664	0.661	0.684	0.654	0.698	0.076	0.060	0.063	0.651	0.585	0.621	0.568	0.628
	CCS	42	145	111	113	42	145	111	113	42	145	111	113	42	145	111	113
	Code	HC3	HC2	HC2	HC3	HC3	HC1	HC1	HC2	HC2	HC2	HC3	HC2	HC3	HC3	HC2	HC3
Disease of DC2	56	0.608	0.056	0.033	0.548	0.701	0.013	0.030	0.130	0.674	0.611	0.017	0.571	0.711	0.645	0.023	0.598
	188	0.027	0.738	0.538	0.020	0.030	0.731	0.658	0.116	0.611	0.731	0.033	0.658	0.645	0.767	0.023	0.668
	143	0.047	0.538	0.608	0.090	0.023	0.658	0.807	0.033	0.043	0.050	0.698	0.060	0.033	0.033	0.605	0.023
	145	0.548	0.020	0.020	0.824	0.030	0.027	0.053	0.807	0.571	0.658	0.053	0.771	0.598	0.668	0.007	0.734
	CCS	120	122	123	118	120	122	123	118	120	122	123	118	120	122	123	118
	Code	HC2	HC2	HC2	HC2	HC1	HC1	HC3	HC2	HC2	HC1	HC2	HC3	HC2	HC3	HC2	HC1
Disease of DC3	120	0.664	0.625	0.585	0.664	0.635	0.605	0.003	0.130	0.638	0.010	0.578	0.020	0.628	0.007	0.565	0.123
	122	0.625	0.648	0.588	0.648	0.605	0.605	0.003	0.143	0.010	0.648	0.050	0.013	0.010	0.638	0.047	0.037
	123	0.585	0.588	0.621	0.621	0.017	0.013	0.608	0.120	0.578	0.056	0.631	0.033	0.565	0.047	0.631	0.066
	118	0.664	0.648	0.621	0.854	0.000	0.000	0.000	0.857	0.000	0.000	0.003	0.794	0.000	0.000	0.000	0.847

Three sets of data for the three top-level clusters are listed. They represent four disease-specific hospital networks related to four diseases of main clusters DC1, DC2 and DC3, respectively. For each of the four similar-symptom-disease-specific hospital networks, a list of hospitals in every domain-specific cluster was compared with lists of hospitals in other clusters at the same network. Also, this list is compared with other lists of hospitals in different clusters of the other three networks. The percentage of same hospitals that belong to both networks is calculated on a yearly basis. A high percentage of hospitals are grouped together at one of the hospital clusters in similar symptom-disease networks. Red color in table 3 shows very high percentage (>80%), yellow, green and blue represents >70%, >60% and >50% respectively.

For example, in the first set of disease-specific hospital networks for similar-symptoms-diseases of the top-level cluster of diseases DC1 in 2011, 81% of the total 301 hospitals included in this study are grouped in hospital-cluster HC3 of hospital admission network for liver-disease-alcohol–related disease (CCS code: 100). Also, 70% of hospitals are grouped in hospital-cluster HC3 of hospital admission network for other liver disease (CCS code: 133). However, the same 67% of the 301 hospitals are grouped in the same hospital-cluster HC3 for both networks of liver disease, alcohol–related disease, and other liver disease (CCS codes 132 and 133). It means that 67% of hospitals share similar admission distribution when considering the symptoms similarity between diseases. This finding confirms that symptoms are critical in clinical diagnosis and treatments [10]. Similarly, the same list of hospitals (60% of all hospitals) are also grouped into the same hospital-cluster HC2 for both networks of Hepatitis and Biliary tract disease (CCS codes 5 and 131). 65% and 55% of hospitals share the same hospital-specific clusters, HC2, in 2010 and 2008 respectively in all four diseases-specific networks in the first set. Networks share the same underlying clustering structure when corresponding to diseases that belong to same top-level cluster of diseases DC1. The same observations can be drawn from the other two sets of the top-level clusters DC2 and DC3.

Analysis of hospital-specific clusters of Single-cluster settings on the same list of disease-specific hospital networks resulted in no significant groupings. In these experiments about 33% of hospitals are shared for the three-different domain-specific clusters. Most hospitals belong to the same hospital-specific clusters during the first year and tend not to stay in the same cluster in the following years which contradicts NoN-clustering results.

Further investigation was proceeded in hospitals that tend to group together most of the time when disease symptom similarity was considered. Factors that seem to affect hospitals for grouping tendency was preliminarily studied and showed a very promising future work.

VI. Conclusion

The aim of this work was to study the tendency of hospitals to be clustered based on admission distributions for different diseases while considering the similarity among disease symptoms. Hospital admission data was extracted from the California State Inpatient Database (SID) for 2008-2011. The Network of Networks data model was used to represent the hospital admission distribution of different diseases that correspond to disease nodes in a disease network that was extracted from the Human Symptoms Disease Network. The NoNClus method was used to test the assumption of existence of underlying groups among different disease-specific hospital networks. This assumption holds in this work and, hence, it opens the road to explore more hidden underlying clustering structures for better planning and utilizing of healthcare resources.

ACKNOWLEDGEMENT

This research was supported in part by DARPA grant FA9550-12-1-0406 negotiated by AFOSR, NSF BIGDATA grant 14476570 and ONR grant N00014-15-1-2729. Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, provided part of the data used in this study.

References

[1] J. W. Thomas, J. R. Griffith, and P. Durance, "Defining hospital clusters and associated service communities in metropolitan areas," *Socioecon. Plann. Sci.*, vol. 15, no. 2, pp. 45–51, 1981.

[2] P. Shay, "More Than Just Hospitals: An Examination of Cluster Components and Configurations," Virginia Commonwealth University, 2014.

[3] P. L. Delamater, A. M. Shortridge, and J. P. Messina, "Regional health care planning: a methodology to cluster facilities using community utilization patterns.," *BMC Health Serv. Res.*, vol. 13, no. 1, p. 333, 2013.

[4] S. Belciug and F. Gorunescu, "Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation," *J. Biomed. Inform.*, vol. 53, pp. 261–269, 2015.

[5] P. J. Phillip, R. Mullner, and S. Andes, "Toward a better understanding of hospital occupancy rates.," *Health Care Financ. Rev.*, vol. 5, no. 4, pp. 53–61, 1984.

[6] L. F. McMahon, R. A. Wolfe, and P. J. Tedeschi, "Variation in hospital admissions among small areas. A comparison of Maine and Michigan.," *Med. Care*, vol. 27, no. 6, pp. 623–631, 1989.

[7] D. S. Morrison and P. McLoone, "Changing patterns of hospital admission for asthma, 1981-97.," *Thorax*, vol. 56, no. 9, pp. 687–90, 2001.

[8] D. Devi, M. and Kumar, "Discovering Disease Pattern in Hospital Data Analysis," *Int. J. Emerg. Res. Manag. &Technology*, vol. 4, no. 5, pp. 235–241, 2015.

[9] J. Ni, H. Tong, W. Fan, and X. Zhang, "Flexible and Robust Multi-Network Clustering," *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15*, pp. 835–844, 2015.

[10] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nat. Commun.*, vol. 5, no. May, p. 4212, Jun. 2014.

[11] J. Glass, M. Ghalwash, M. Vukicevic, and Z. Obradovic, "Extending the Modelling Capacity of Gaussian Conditional Random Fields while Learning Faster," *Proc. Thirtieth AAAI Conf. Artif. Intell. Extending*, pp. 1596–1602, 2016.

[12] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *J. Complex Networks*, 2014.

[13] J. Ni, H. Tong, W. Fan, and X. Zhang, "Inside the Atoms: Ranking on a Network of Networks," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1356–1365.

[14] W. Cheng, X. Zhang, Z. Guo, Y. Wu, P. F. Sullivan, and W. Wang. Flexible and robust co-regularized multi-domain graph clustering. In KDD, 2013.

[15] A. Kumar and H. Daum´ e. A co-training approach for multi-view spectral clustering. In ICML, 2011.

[16] D. Zhou and C. J. Burges. Spectral clustering and transductive learning with multiple views. In ICML, 2007.