Discovering Linkages Among Multiple Disease Networks by Joint Clustering

Nouf Albarakati[1], Hussain Otudi[1], Rafaa Aljurbua[1,2], and Zoran Obradovic[1]
{nouf, hussain.otudi, rafaa.aljurbua, zoran.obradovic}@temple.edu

[1] Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA 19122, USA
[2] College of Computer, Qassim University, Buraydah 52571, Saudi Arabia

Abstract. This study introduces an optimized joint clustering algorithm to identify hospital groupings based on disease-specific monthly admission patterns using data from the California State Inpatient Database (2008–2011). Unlike the previous two-step method that relied on a predefined disease similarity network, the proposed approach dynamically constructs a meta-disease super network while clustering, enabling simultaneous optimization of disease and hospital networks. The method models 145 disease-specific hospital networks across 152 hospitals, forming a multilayer Network of Networks structure. It uses low-rank approximations and regularization to capture both local (hospital-level) and global (disease-level) similarities. Evaluation on synthetic and real-world data shows improved clustering homogeneity (average of $61.2\%, SD \pm 2.1\%$) over the prior method (average $52.4\%, SD \pm 5.6\%$), with statistically significant gains ($p = 0.0038$).These clusters showed temporal stability and meaningful clinical groupings, aiding in referral coordination and resource allocation. The method is robust, interpretable, and extensible to other datasets and healthcare systems.

Keywords: Network of networks, Meta super network, Co-clustering

1 Introduction and Related Work

Healthcare planning significantly impacts both community health and economic outcomes. A critical aspect of improving healthcare efficiency and reducing costs lies in identifying similar hospitals based on their admission patterns. However, accurately capturing the complex and heterogeneous relationships among hospitals remains a major challenge [9]. Clustering analysis has long been applied to healthcare data to group hospitals by shared characteristics and to uncover meaningful variations in hospital behavior [9]. For instance, principal diagnoses at admission have proven to be a strong indicator of hospital-level variation [1, 2, 3, 4]. Yet, hospitals with similar overall admission volumes may diverge significantly in disease-specific patterns. Traditional clustering studies typically rely on single-view data representations. This simplification misses the multi-dimensional nature of hospital behavior. To address this, our study adopts a multi-domain clustering approach that captures hospital similarity across diverse disease domains. In the multi-domain disease-specific hospital networks

setting, a hospital's cluster assignment can vary by disease, and specialized hospitals may appear in only a subset of networks. This richer modeling enables discovery of overlapping and heterogeneous structures across diseases. To jointly model these networks, we adopt a Network of Networks (NoN) framework, where each disease-specific hospital network forms a subnetwork and inter-disease relationships are encoded in a top-layer super network [4, 5]. This enables joint clustering of heterogeneous, multidomain networks by capturing both within-disease and across-disease similarities. In prior work [2], we guided the joint clustering process using a literature-based disease symptom similarity network. We later showed that health record-derived similarity networks outperformed literature-based ones in improving hospital cluster homogeneity [3]. However, both approaches required predefined disease networks, introducing rigidity and potential bias. To overcome this limitation, we proposed a static super network derived from low-rank factor similarity in [1]. Yet, even this two-step method limited joint optimization. Building upon these insights, this study introduces an optimized joint clustering algorithm that eliminates the need for any predefined super network. Instead, it dynamically learns a meta disease similarity network while simultaneously performing joint clustering of hospitals across all disease networks. This integrated, one-step approach enables more accurate, robust, and interpretable clustering by capturing both local hospital-level structure and global disease-level relationships. Prior literature has demonstrated the utility of clustering for healthcare applications, from identifying referral patterns to optimizing resource allocation [6, 8]. Common clustering algorithms, such as k-means, DBSCAN, and hierarchical clustering, have been used to group hospitals based on utilization or mobility patterns, often in a single-view setting [6, 9]. Some studies applied fuzzy or k-means clustering to monitor healthcare systems during the COVID-19 pandemic [8]. However, these methods are not equipped to handle multiple interconnected networks with shared underlying structures. Our proposed method contributes to this line of work by:

1. Developing a novel, optimized joint clustering algorithm that simultaneously clusters hospitals and learns a dynamic meta super network from disease-specific hospital admission patterns.
2. Validating the algorithm using synthetic data to assess its accuracy in recovering known clustering structures.
3. Applying the optimized joint clustering algorithm to a real-world dataset of disease-specific hospital networks to jointly cluster hospitals based on their monthly admission behavior for specific diseases.

2 Methodology

2.1 Problem Definition

Let $A^{(i)} \in \mathbb{R}^{h \times h}$ represent the i-th disease-specific hospital subnetwork, where each matrix captures the similarity in monthly admission behavior among h hospitals for a specific disease $i \in \{1, 2, \ldots, d\}$. The objective is to jointly cluster these networks through the following components:

1. **Disease-specific Hospital Subnetwork Clustering**
   Each $A^{(i)}$ is approximated by a low-rank matrix factorization $U^{(i)}(U^{(i)})^\top$, where $U^{(i)} \in \mathbb{R}^{h\times t}$ and t is the number of hospital clusters. Based on preliminary tests with $t = 3, 5, 7$, $t = 3$ was selected for producing the most interpretable and coherent hospital groupings.
2. **Meta Super Network Learning**
   A disease similarity super network $D \in \mathbb{R}^{d\times d}$ is constructed, where each entry D_{lm} reflects the Euclidean distance $\|U^{(l)} - U^{(m)}\|_F^2$ between diseases l and m. To ensure D^2 is positive semi-definite, it is factorized as $D^2 \approx GG^\top$, where $G \in \mathbb{R}^{d\times k}$. Empirical tests with $k = 3, 5, 7$ showed that $k = 3$ yielded the most meaningful and clinically relevant disease clusters.
3. **Disease Cluster Guided Regularization**
   Each cluster centroid $V^{(j)}$ captures the underlying structure of diseases grouped into cluster g_j, while membership strength g_{ij} quantifies how strongly disease i belongs to cluster j. Regularization is applied by minimizing the distance $\|U^{(i)} - V^{(j)}\|_F^2$, promoting consistency among diseases with similar hospital interaction profiles.

The objective is to solve for the variables $U^{(i)}, V^{(j)}, G$ by minimizing the following function:

$$\min_{\substack{U^{(i)}\geq 0,\ i=1,\ldots,d \\ V^{(j)}\geq 0,\ j=1,\ldots,k \\ G\geq 0}} J_D, \tag{1}$$

where $J_D =$

$$\underbrace{\sum_{i=1}^{d}\left\|A^{(i)} - U^{(i)}\left(U^{(i)}\right)^\top\right\|_F^2}_{\text{(1) hospital subnetwork clustering}} + \underbrace{\left\|D^2 - GG^\top\right\|_F^2}_{\text{(2) meta super network learning}} + \alpha \underbrace{\sum_{i=1}^{d}\sum_{j=1}^{k} g_{ij}\left\|U^{(i)} - V^{(j)}\right\|_F^2}_{\text{(3) disease cluster regularization}} \tag{2}$$

Here, $\alpha > 0$ is a hyperparameter that controls the balance between accurately representing individual disease-specific hospital networks and ensuring consistency with the overall disease clustering structure. All matrices are constrained to be non-negative, ensuring interpretability.

2.2 Optimization Approach

To solve the non-convex optimization problem, a block coordinate descent strategy is used. The algorithm iteratively updates three sets of variables while holding the others fixed, The updates continue until convergence is achieved based on a threshold ε:

- Low-rank representations of the disease-specific hospital networks: $U^{(i)}$
- Low-rank factor of the disease similarity super network: G
- Centroids of the disease clusters: $V^{(j)}$

Initialization All matrices are initialized and projected to be non-negative:

- $U^{(i)}$: Low-rank approximation of $A^{(i)}$
- G: Low-rank approximation of an initial D^2
- $V^{(j)}$: Subset-based initialization from $U^{(i)}$

Iterative Updates continue until convergence is achieved based on ε threshold

- **Update $U^{(i)}$ Given $G, V^{(j)}$**: For each $i \in \{1, \ldots, d\}$, solve:

$$\min_{U^{(i)} \geq 0} \Big[\|A^{(i)} - U^{(i)}(U^{(i)})^\top\|_F^2 + \alpha \sum_{j=1}^{k} g_{ij} \|U^{(i)} - V^{(j)}\|_F^2 + \phi(U^{(i)}) \Big], \quad (3)$$

  where $\phi(U^{(i)})$ accounts for the contribution of $\|D^2 - GG^\top\|_F^2$ via $\|U^{(l)} - U^{(m)}\|_F^2$. Projected gradient descent ensures non-negativity after each step.
- **Update G Given $U^{(i)}, V^{(j)}$**: Define $D \in \mathbb{R}^{d \times d}$ with entries $D_{lm}^2 = \|U^{(l)} - U^{(m)}\|_F^2$. Then D^2 is fixed, and we solve, a symmetric NMF problem on D^2:

$$\min_{G \geq 0} \|D^2 - GG^\top\|_F^2.$$

  using multiplicative updates, The result is projected to non-negativity:

$$G \leftarrow G \odot \frac{D^2 G}{(GG^\top G)},$$

- **Update $V^{(j)}$ Given $U^{(i)}, G$**: With $U^{(i)}$ and G now fixed, solve:

$$\min_{V^{(j)} \geq 0} \Big[\alpha \sum_{i=1}^{d} g_{ij} \|U^{(i)} - V^{(j)}\|_F^2 \Big].$$

  Ignoring non-negativity for a moment, the solution is typically the *weighted average* of the relevant $U^{(i)}$:

$$V^{(j)} = \frac{\sum_{i=1}^{d} g_{ij} U^{(i)}}{\sum_{i=1}^{d} g_{ij}}.$$

  Afterward, project negative entries to zero to preserve non-negativity.

2.3 Model Evaluation: Synthetic Data

Synthetic data is used to validate the model's ability to recover known cluster structures in a controlled setting [7]. The evaluation involves three steps: (1) generating three sets of low-rank factor matrices to represent distinct underlying clustering patterns; (2) creating four adjacency matrices per set by multiplying each factor by its transpose, yielding 12 structured networks; and (3) introducing symmetric additive noise, drawn uniformly from $[0, 0.05]$, to simulate real-world data variability. This controlled noise is symmetrically applied across matrix entries to preserve structural balance. The synthetic setup enables assessment of the algorithm's robustness and capacity to identify latent clustering structure under noisy conditions.

3 Results and Discussion

3.1 Synthetic Data

The optimized joint clustering model was evaluated using synthetic data to assess its ability to recover known cluster structures under varying noise levels. Compared to a previous two-step method that relied on a predefined super network, the new model dynamically learns a unified meta super network that captures both local and global similarities during clustering. In Figure 1, each row contains four synthetic networks that share the same underlying cluster structure. The t-SNE visualizations use color to indicate cluster assignments. It shows that the optimized joint clustering model, which dynamically learns a meta super network during clustering, consistently groups networks with the same structure, evidenced by coherent color blocks across each row. Figure 2 provides a zoomed-in view of the super network clustering results. Panel (a) displays the static super network built from Euclidean distances between low-rank factors, which misclassifies the 12 networks by splitting clusters that should be grouped together. Panel (b), generated by the optimized model, accurately clusters the networks into three distinct groups, recovering the true underlying structure. Together, these figures demonstrate the optimized model's ability to learn and preserve latent cluster structure, even in the presence of noise and structural variation.

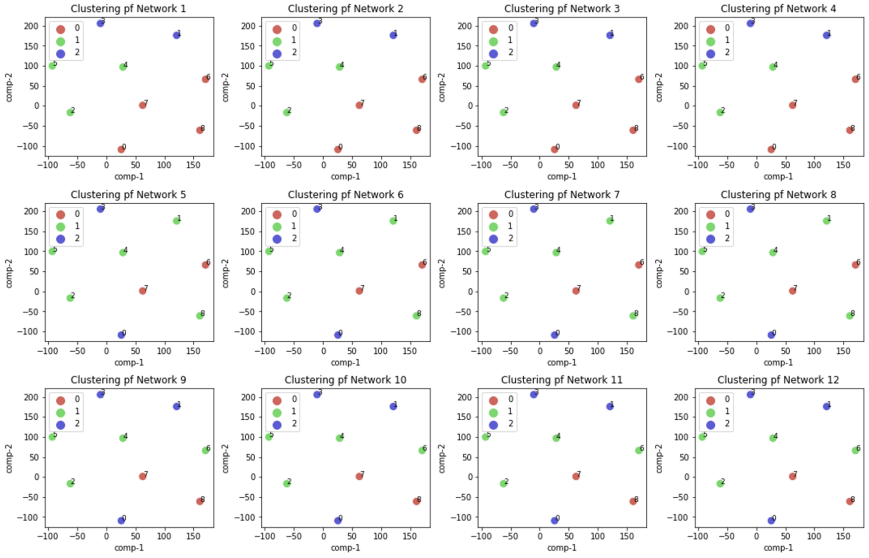
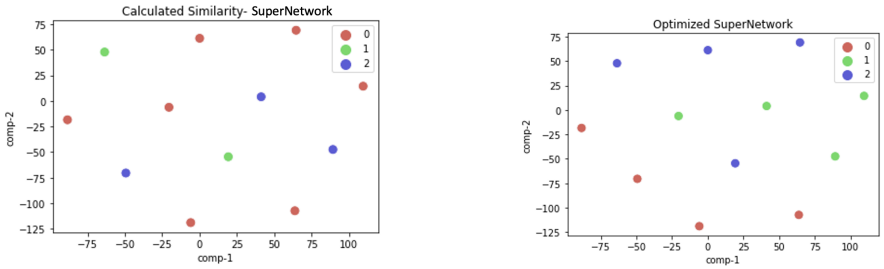


Fig. 1: Clustering of a synthetic network using an optimized joint clustering model.

3.2 Real-World Data: Hospital Clustering

Data The dataset is derived from the California State Inpatient Database (SID), part of the HCUP initiative, containing over 7 million emergency department



(a) Clustering of a predefined super network

(b) Clustering of a dynamically summarized super network

Fig. 2: Comparison of hospital super network clustering approaches.

discharge records from 2008 to 2011. Data includes demographic, medical, and hospital-specific attributes aggregated monthly per disease per hospital. Disease-specific hospital networks were constructed using Kullback-Leibler divergence to measure similarity in admission distributions. All data is fully de-identified and HIPAA-compliant.

Clustering of Disease Network The model identified three stable disease clusters across four years (2008–2011) using dynamically learned meta-super networks. Table 1 highlights the top five diseases within each cluster for 2008, showing strong temporal consistency in their cluster assignments and probabilities across subsequent years. For example, maternal and pregnancy-related diagnoses consistently formed Cluster 1, while vascular and respiratory conditions grouped in Cluster 2. Neurological and injury-related conditions primarily formed Cluster 3. The disease distribution across clusters remains stable over the years, with Cluster 2 containing the highest proportion of diseases (44%–50%). These results demonstrate the algorithm's ability to detect meaningful and temporally stable hospital clustering patterns guided by disease similarities.

Clustering of Disease-Specific Hospital Networks Table 2 presents the results of the optimized joint clustering algorithm of the top five disease-specific hospital networks from each 2008 disease cluster, using meta-super network similarities to guide clustering. The table includes each disease's cluster assignment, probabilities, the number of hospitals treating the condition (out of 152), and their distribution across the three hospital clusters. Results show that Cluster 1 and cluster 3 diseases, related to obstetric care and neurological and trauma, had balanced hospital distributions; Cluster 2, involving respiratory and circulatory conditions, showed broad hospital coverage.

Clustering homogeneity [1, 3] is used to assess the consistency of hospital groupings across disease-specific networks within the same cluster by measuring the proportion of hospitals that remain in the largest consistent group. Higher scores reflect stronger alignment with underlying disease patterns. The optimized joint clustering algorithm demonstrated strong consistency in hospital groupings,

Table 1: Disease meta super network clustering results: Top five diseases based on monthly admission behavior (2008-2011).

Diagnosis (CCS Code)	**2008 Clust**	**2008 Prob**	**2009 Clust**	**2009 Prob**	**2010 Clust**	**2010 Prob**	**2011 Clust**	**2011 Prob**
Prolonged pregnancy(185)	1	0.9692	1	0.9440	1	0.9616	1	0.8708
Diabetes compl preg(186)	1	0.9106	1	0.9156	1	0.9438	1	0.8417
Previous C-section(189)	1	0.8929	1	0.8748	1	0.9096	1	0.8600
Fetal distress(190)	1	0.8917	1	0.8697	1	0.8939	1	0.7952
Polyhydramnios(191)	1	0.8468	1	0.8336	1	0.8243	1	0.7308
Thrombophlebitis(118)	2	0.9785	2	0.8805	2	0.7941	2	0.8900
Veins and lymphatics dis.(121)	2	0.9687	2	0.9560	2	0.8603	2	0.8339
Pleurisy/pneumothorax(130)	2	0.8759	2	0.7848	2	0.7383	2	0.7310
Respiratory failure(131)	2	0.8663	2	0.7932	2	0.7315	2	0.7805
Lymphadenitis(247)	2	0.7842	2	0.7659	2	0.6587	2	0.6656
Multiple sclerosis(80)	3	0.8084	3	0.8275	3	0.8183	3	0.8084
Degenerative nervous sys(81)	3	0.7514	3	0.7816	3	0.7820	3	0.7514
Joint disorders(225)	3	0.7053	3	0.5753	2	0.6136	2	0.5903
Intracranial injury(233)	3	0.6903	3	0.7373	3	0.6635	3	0.6053
Open wounds(235)	3	0.6750	3	0.7551	3	0.6816	3	0.5950

Table 2: Hospital clustering analysis: Top five diseases within each of the three main disease clusters for the 2008 networks.

Disease-specific Hospital Network (CCS Code)	**Clust 2008**	**Clust Prob**	**Num. Hosp. in Net. (%of152)**	**Num. Hosp. Clus. 1 (%)**	**Num. Hosp. Clus. 2 (%)**	**Num. Hosp. Clus. 3 (%)**
Prolonged pregnancy(185)	1	0.97	97(64%)	32(32%)	35(36%)	30(30%)
Diabetes comp pregnancy(186)	1	0.91	65(43%)	18(27%)	25(38%)	22(33%)
Previous C-section(189)	1	0.89	55(36%)	16(29%)	24(43%)	15(27%)
Fetal distress(190)	1	0.89	52(34%)	14(26%)	22(42%)	16(30%)
Polyhydramnios(191)	1	0.85	57(38%)	23(40%)	17(29%)	17(29%)
Thrombophlebitis(118)	2	0.98	141(93%)	52(36%)	41(29%)	48(34%)
Veins & lymphatic dis(121)	2	0.97	149(98%)	52(35%)	44(30%)	53(36%)
Pleurisy/pneumothorax(130)	2	0.88	144(95%)	63(44%)	35(24%)	46(32%)
Respiratory failure(131)	2	0.87	151(99%)	61(40%)	43(28%)	47(31%)
Lymphadenitis(247)	2	0.78	139(91%)	50(36%)	41(30%)	48(35%)
Multiple sclerosis(80)	3	0.81	108(71%)	33(31%)	31(29%)	44(41%)
Degenerative nervous sys(81)	3	0.75	135(89%)	44(33%)	41(30%)	50(37%)
Joint disorders(225)	3	0.71	113(55%)	32(28%)	37(32%)	44(38%)
Intracranial injury(233)	3	0.69	152(100%)	53(35%)	50(33%)	49(32%)
Open wounds(235)	3	0.68	123(81%)	39(32%)	39(32%)	45(36%)

achieving homogeneity scores between 59% and 63% with an average of 61.2% and low variability (standard deviation of $\pm 2.1\%$). Compared to the earlier two-step method, which averaged 52.4% with greater variability (standard deviation of $\pm 5.6\%$), the improvement was statistically significant ($p = 0.0038$), highlight-

ing the optimized model's robustness and value in identifying stable hospital clusters for informed healthcare planning and resource coordination.

Conclusion and Future Work

This study introduced an optimized joint clustering model that groups hospitals based on disease-specific monthly admission patterns while simultaneously learning a unified meta disease super network. The model demonstrated improved clustering consistency, outperforming a previous two-step approach in terms of homogeneity and stability, which underscores its robustness in identifying meaningful hospital groupings. These clusters can inform real-world decisions by supporting regional planning, referral coordination, and resource distribution. Despite its strengths, the study has limitations. It relies on older data (2008–2011) from the California State Inpatient Database, which is geographically specific. To improve generalizability and relevance, future work will apply the model to more recent and international datasets. Additionally, while the study used a brief sensitivity analysis to determine the number of clusters, future research will incorporate systematic methods to refine cluster selection and enhance interpretability.

References

[1] Nouf Albarakati, Avrum Gillespie, and Zoran Obradovic. "Summarizing multiple networks based on their underlying clustering structure to guide joint clustering of hospital admissions". In: *Informatics in Medicine Unlocked* 39 (2023).

[2] Nouf Albarakati and Zoran Obradovic. "Disease-based clustering of hospital admission: disease network of hospital networks approach". In: *The IEEE 30th International Symposium on Computer-Based Medical Systems*. 2017.

[3] Nouf Albarakati and Zoran Obradovic. "Multi-domain and multi-view networks model for clustering hospital admissions from the emergency department". In: *International Journal of Data Science and Analytics* 8.4 (2019).

[4] Florence T Bourgeois et al. "Variation in emergency department admission rates in US children's hospitals". In: *Pediatrics* 134.3 (2014), pp. 539–545.

[5] Wei Cheng et al. "Flexible and robust co-regularized multi-domain graph clustering". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013).

[6] Jan Chrusciel et al. "Making sense of the French public hospital system: a network-based approach to hospital clustering using unsupervised learning methods". In: *BMC Health Services Research* 21 (2021).

[7] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. "Synthetic data in health care: A narrative review". In: *PLOS Digital Health* 2.1 (2023).

[8] Karli Eka Setiawan et al. "Clustering models for hospitals in Jakarta using fuzzy c-means and k-means". In: *Procedia Computer Science* 216 (2023).

[9] Patrick Davis Shay. *More than just hospitals: an examination of cluster components and configurations*. Virginia Commonwealth University, 2014.