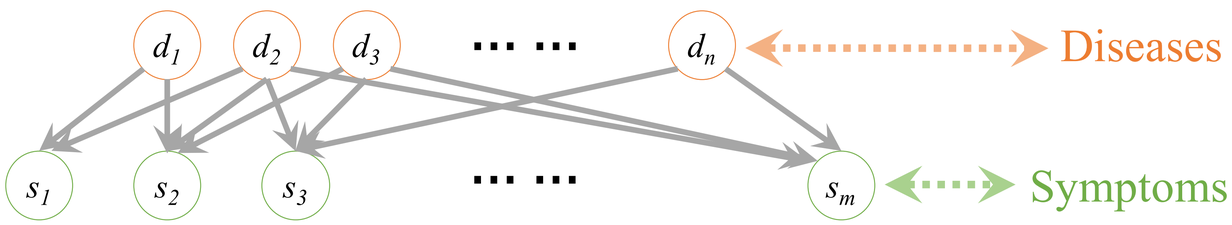
Cross-Network Clustering and Cluster Ranking for Medical Diagnosis

Jingchao Ni[1], Hongliang Fei[2], Wei Fan[2], and Xiang Zhang[1]
[1]College of Information Sciences and Technology, Pennsylvania State University
[2]Baidu Research Big Data Lab

2017 IEEE International Conference on Data Engineering

Motivation

Computational medical diagnosis:



Inferring a probability distribution for the disease nodes given a subset of the symptom nodes.

Limitations:

- Diseases are independent with each other
- Symptoms are conditionally independent with each other
- Sparse disease-symptom associations
- Poor outcome interpretability: mixed ranking of diseases

Introducing Domain Networks

Symptom domain

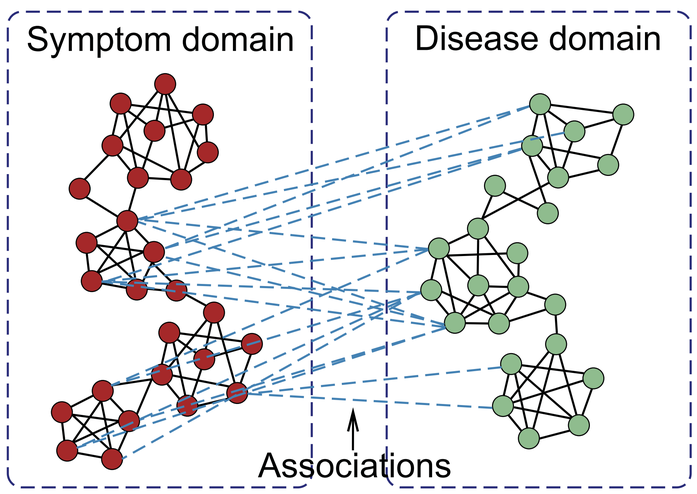
- Each edge represents the similarity between a pair of symptoms

Disease domain

- Each edge represents the similarity between a pair of diseases

Association network

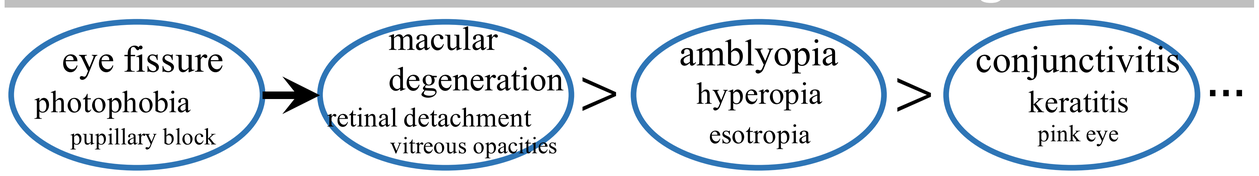
- Each edge is weighed by the correlation between a symptom and a disease



Contribution 1: accounting for the dependencies between diseases (symptoms).

Contribution 2: the rich domain information can alleviate the sparsity problem of the associations.

Cross-Domain Cluster Ranking



Contribution 3: clustered structure in the ranking list gives better interpretability of outcomes and less risk of false identifications.

CCCR Algorithm

Domain network clustering (doubly stochastic matrix decomposition):

$$J_A^{(i)} = -\sum_{(x,y)\in E^{(i)}} \mathbf{A}_{xy}^{(i)} \log \hat{\mathbf{A}}_{xy}^{(i)} - (\alpha-1)\sum_{xu} \log \mathbf{H}_{xu}^{(i)}$$

$$s.t.\ \ \mathbf{H}^{(i)} \geq 0, \quad \mathbf{H}^{(i)}\mathbf{1}_{k_i} = \mathbf{1}_{n_i}$$

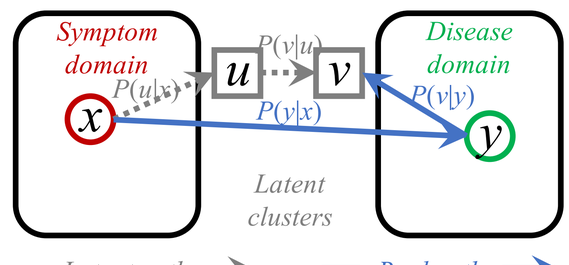
$$\hat{\mathbf{A}}_{xy}^{(i)} = \sum_{u=1}^{k_i} \frac{H_{xu}^{(i)} H_{yu}^{(i)}}{\sum_{z=1}^{n_i} H_{zu}^{(i)}}$$

Cross-network cluster ranking (second-order random walk):

$$J_R^{(ij)} = -\sum_{xv} \underbrace{(\tilde{\mathbf{B}}^{(ij)}\mathbf{H}^{(j)})_{xv}}_{\text{real paths}} \log \underbrace{(\mathbf{H}^{(i)}\mathbf{S}^{(ij)})_{xv}}_{\text{latent paths}}$$

Row normalized adjacency matrix of association network

Cross-domain cluster-cluster conditional probability matrix



A joint optimization problem:

$$\min \quad J(\{\mathbf{H}^{(i)}\},\{\mathbf{S}^{(ij)}\}) = \sum_{i=1}^{g} J_A^{(i)} + \beta \sum_{ij,i\neq j} J_R^{(ij)}$$

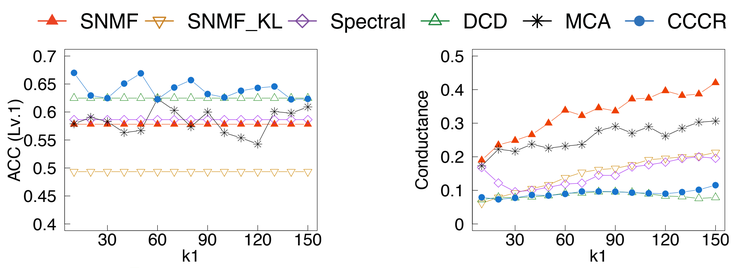
$$s.t. \quad \mathbf{H}^{(i)} \geq 0, \quad \mathbf{H}^{(i)}\mathbf{1}_{k_i} = \mathbf{1}_{n_i}, \quad \mathbf{S}^{(ij)} \geq 0, \quad \mathbf{S}^{(ij)}\mathbf{1}_{k_j} = \mathbf{1}_{k_i}, \quad \forall 1 \leq i,j \leq g, i \neq j$$

Experiments

Real-life symptom-disease network

Disease domain		Symptom domain		Associations
# nodes	# edges	# nodes	# edges	# edges
9,721	29,332	5,093	22,548	5,337

Clustering results



Cluster ranking results

Top ranked disease clusters given by CCCR

	Symptom cluster	1st disease cluster (probability)		2nd disease cluster (probability)		3rd disease cluster (probability)	
(1)	bloating burp stomachache	increased intracranial pressure gastroesophageal reflux gastritis	(0.7877)	duodenal inflammation antral erosion superficial gastritis	(0.1351)	-	(<0.1000)
(2)	eye fissure photophobia pupillary block	macular degeneration retinal detachment vitreous opacities	(0.5002)	amblyopia hyperopia esotropia	(0.1569)	conjunctivitis keratitis pink eye	(0.1330)
(3)	cerebral hemorrhage intracranial hemorrhage increased intracranial pressure	cerebral infarction brainstem infarction stroke	(0.5400)	skull fracture epidural hematoma brain contusion	(0.1449)	diabetes hypertension dyslipidemia	(0.1161)

Results of a QMR-DT algorithm

Symptom #	Top ranked diseases
(1)	***gastritis***, cold, heart disease, fracture, epilepsy
(2)	***cataract, uveitis, ocular trauma, keratitis, pink eye***
(3)	***subarachnoid hemorrhage, aneurysm, hypertension***, cold