Evaluating Community Detection Methods for the Human Symptoms-Disease Network

Sidhant Puntambekar and Behzod Mirpochoev

CSCI 3352

Background

- Many advances have been made to categorize and diagnose human diseases
  - Disease states vary significantly since diseases can occur at various granularities
    - Molecular level (i.e point mutations in DNA/genetic diseases, individual based)
    - Population level (i.e communicable infectious diseases, epidemiological based)
- Important distinguishing features of human diseases are the phenotypes/symptoms patients exhibit
  - Understanding symptom classifications is critical for medical practitioners for correct diagnosis
- Rzhetsky et al. (2007)
  - Inferred the comorbidity links between 161 diseases observed from over 1.5 million hospital patients
  - Proposed models to estimate the genetic overlap between similar disease communities
- Hidalgo et al. (2009)
  - Constructed a disease-phenotype network from over 30 million Medicare patients
  - Highlighted disease progression patterns and related comorbidities
- Example
  - Coughing, sore throat, and fever are shared symptoms of various diseases such as COVID-19, influenza, and the common cold (How do we distinguish one disease from the other?)

Background

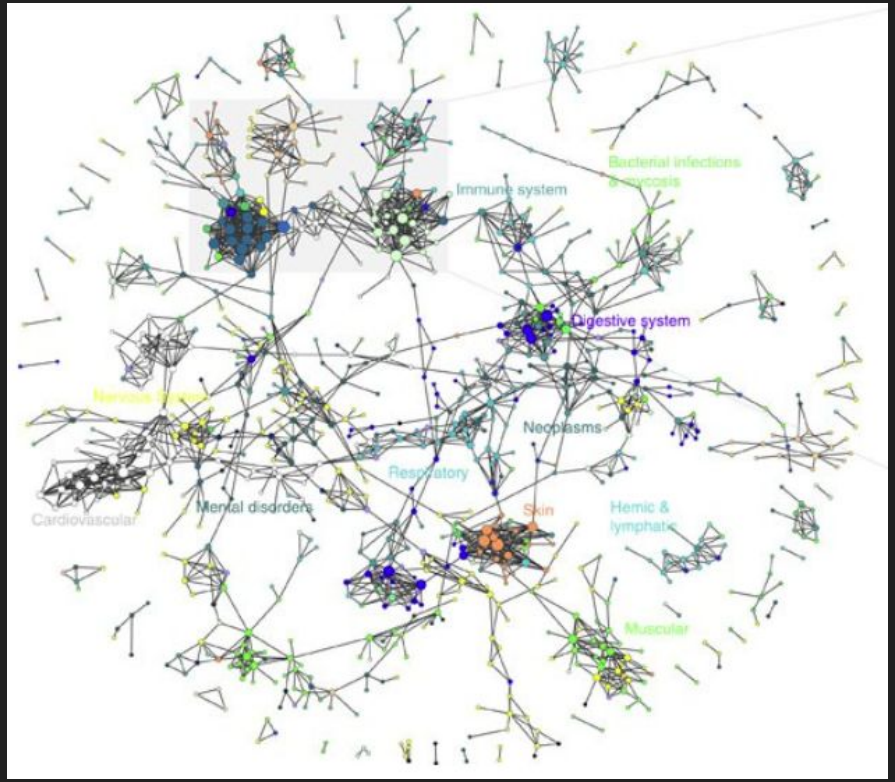
- A network highlighting connections between shared symptoms (phenotypes) and genes/protein–protein interactions of disease states (genotypes) types bridges the gap between biological research and clinical contexts
- Zhou et al. address this discrepancy with a network representation of common human symptom and disease relationships
    - Specifically, the analyze the Medical Subject Headings metadata (MeSH) aggregated from PubMed literature
    - Comprehensive controlled vocabulary for the purpose of indexing journal articles and literature in the life sciences

Zhou et al. (2014) Background

- Generated a bipartite representation of diseases and symptoms (termed Human-Symptom Disease Network)
  - Node sets represent diseases and symptoms
  - Edges represent undirected associated connections linking two disease nodes with a symptom co-occurrence
  - Edges are weighted with their TF-IDF score (term frequency/importance in corpus of PubMed literature)
- Zhou et al. overlayed protein-protein interaction (PPI) data from five external databases on HSDN
  - Investigated the correlations between the symptom similarity of diseases and their degree of shared genes or PPIs

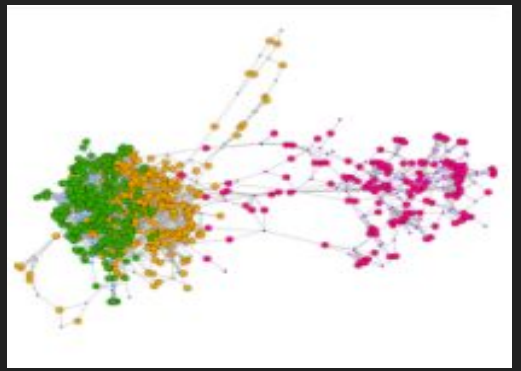
HSDN backbone from Zhou et al.

Key Observation: Highly clustered areas of the network belong to same broad disease categories



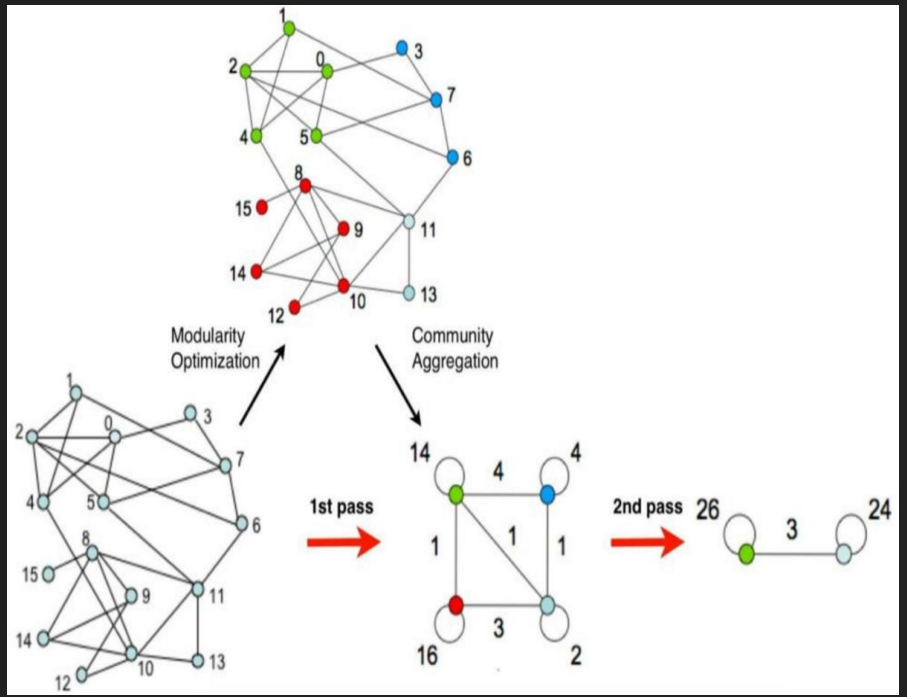
DC-biSBM: Larremore et al. (2014) Background

- Regular DC-SBM models have limitations when applied to bipartite networks
  - These limitations include implicit parameter choices, loss of information through one-mode projections, and general lack of interpretability.
- Larremore et al. highlight a major difference between ordinary SBM networks and biSBM networks
  - DC-biSBMs assume disassortative groups/modules (due to inherent assumption that network is bipartite)
  - Regular SBM models prefer to find either very assortative or very disassortative groups, or some mixture thereof
- Larremore et al. begin by taking the number of nodes in each bipartite set (N_a and N_b) and dividing nodes into distinct K_a and K_b groups.
  - Interrelationships are then represented through a K x K mixing matrix ω (where $K = K_a + K_b$)
  - Entries that would connect nodes of the same type are zeroed out which enforces the bipartite structure
- Construct the DC-biSBM model with:
  - *g* (representing the total number of communities)
  - ω (representing the mixing matrix or stochastic block matrix)
  - Θ (representing the expected degree of every vertex in the real-world network)
  - *T* (matrix representing community partitions).



Community Detection - Louvain Method

- Developed by Blondel et al. (2008) from the University of Louvain in France
- Louvain Method:
  - Each node is given its own module
    - Change in modularity is calculated based on taking node *i* from its own community and placing it in the community of node *j*
    - Modularity calculated as a value in range [-½ , 1] that measures the density of links inside communities compared to links between communities.
  - Community is then aggregated and considered as a collective ‘node’, representing relationships to itself via self-loops and number of nodes connecting to nodes of a different community
  - Whole modularity optimization process is then repeated

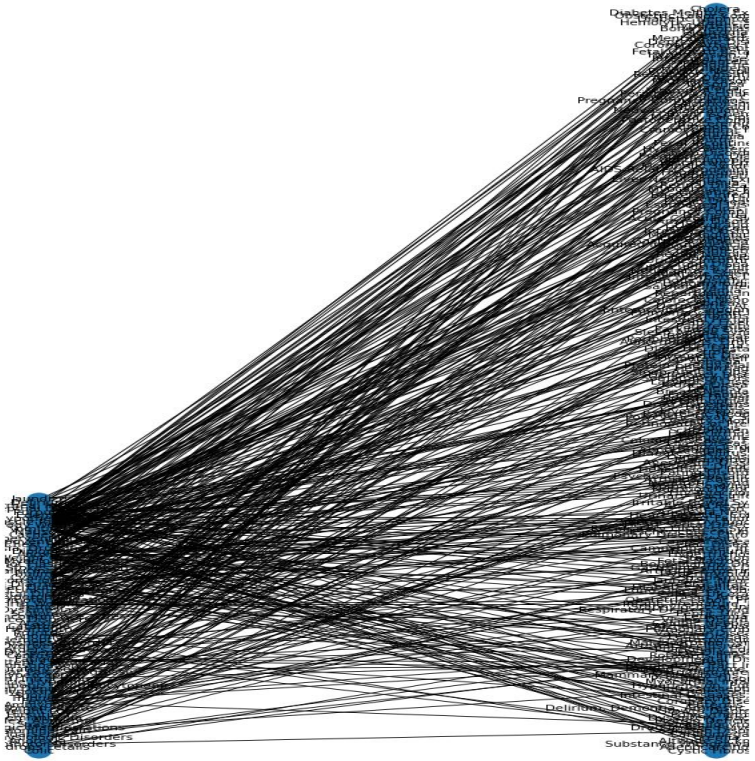


Research Aims & Question

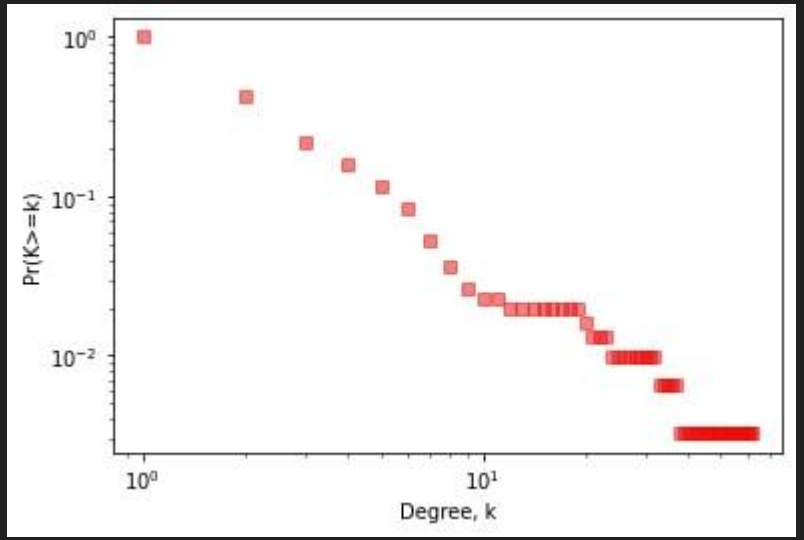
- Research Aims
  - Cluster the network into distinct disease/symptom clusters and identify which ones are the most popular (maximum degree, high clustering coefficients, centrality metrics)
  - Identify correlations and shared states between nodes in clustering communities
  - Using one mode projections of the bipartite network, compute summary statistics to understand projection structure of related diseases and symptoms
- Primary Research Question
  - To what degree does a bipartite degree-corrected stochastic block model (DC-biSBM) generated on disease and common symptom data reveal about the modular nature of interrelated disease states and their potential resultant causes?

Data Quality Control

- Raw HSDN network encompassed 4219 unique disease states, 322 unique symptom states
  - The data, by nature, is a heavy tailed distribution in its PubMed occurrences of a particular symptom or disease
- A popularity filter was introduced (of at least 150 occurrences) along with the isolation of symptoms from diseases in order to focus on relationships between common symptoms and diseases



Full Bipartite Network Summary Statistics




```
number of nodes, n  = 306
number of edges, m  = 394
diameter = 15
mean geodesic distance, <ell> =  4.80
clustering coefficient, C     =  0.00
number of components,   h     =  1
mean degree, k_mean = 2.5751633986928106
max degree, k_max = 62
max degree, k_std =  4.96
```

Figure 1: Summary statistics for full HSDN bipartite network. Note diameter and mean geodesic distance

Bipartite One Mode Projections

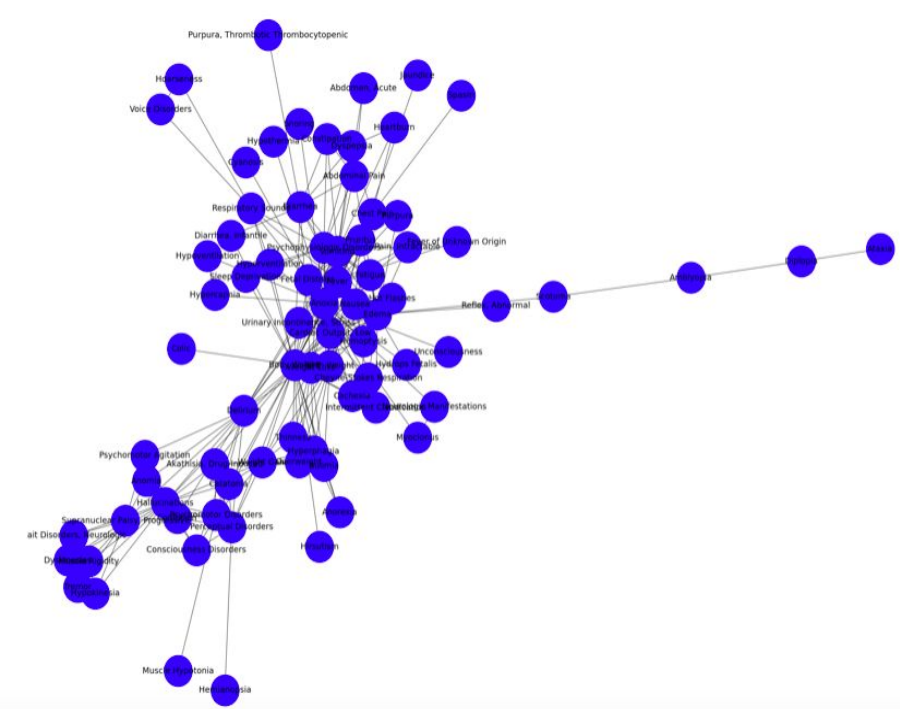


Figure 2: Symptom one mode projection generated from bipartite

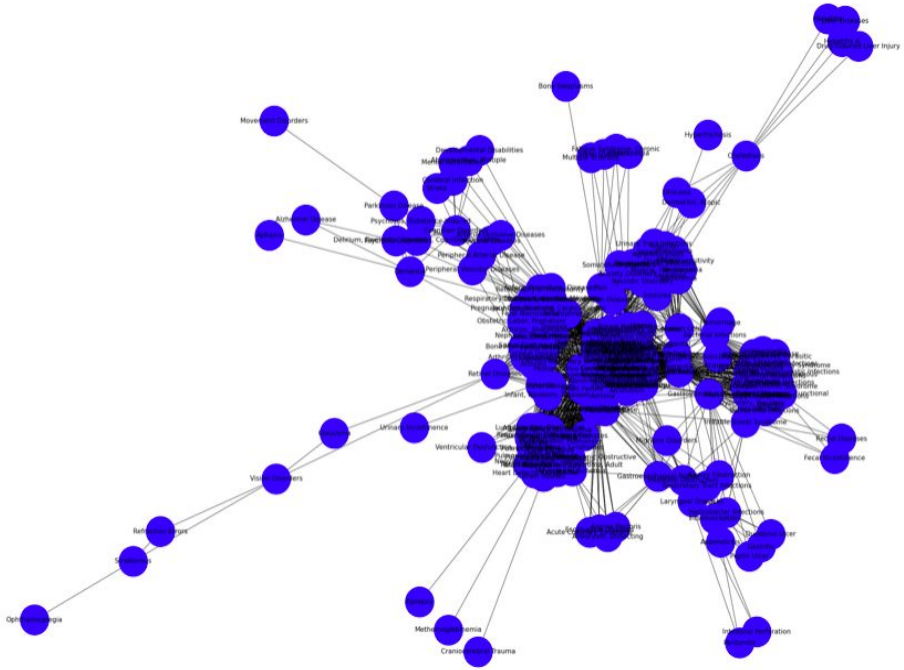
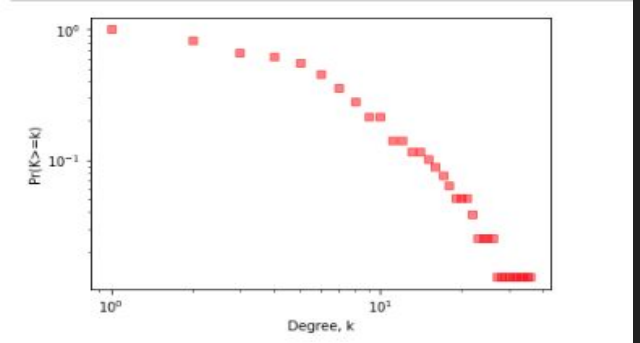


Figure 3: Disease one mode projection generated from bipartite HSDN




```
number of nodes, n  = 78
number of edges, m  = 252
diameter = 7
mean geodesic distance, <ell> =  2.67
clustering coefficient, C     =  0.40
number of components,   h     =  1
mean degree, k_mean = 6.461538461538462
max degree, k_max = 36
max degree, k_std =  6.28
```

Figure 4: Summary statistics for symptom one mode projection

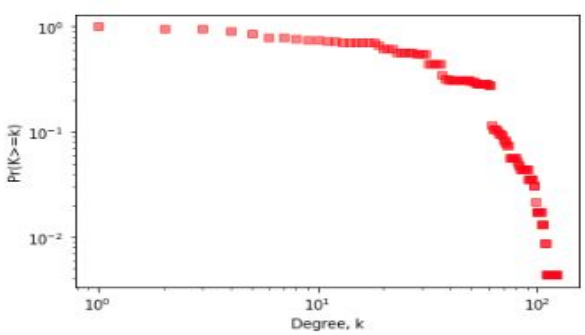


Figure 5: Summary statistics for disease one mode projection

```
number of nodes, n  = 228
number of edges, m  = 3877
diameter = 7
mean geodesic distance, <ell> =  2.29
clustering coefficient, C     =  0.79
number of components,   h     =  1
mean degree, k_mean = 34.00877192982456
max degree, k_max = 123
max degree, k_std = 26.68
```

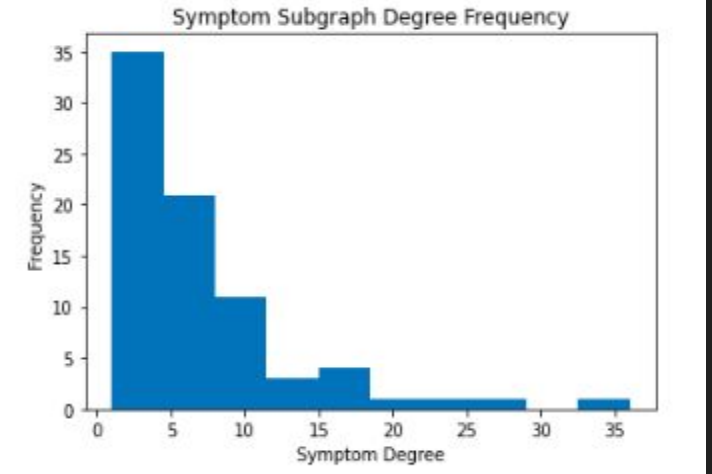


Figure 6: Degree distribution of Symptom one mode projection

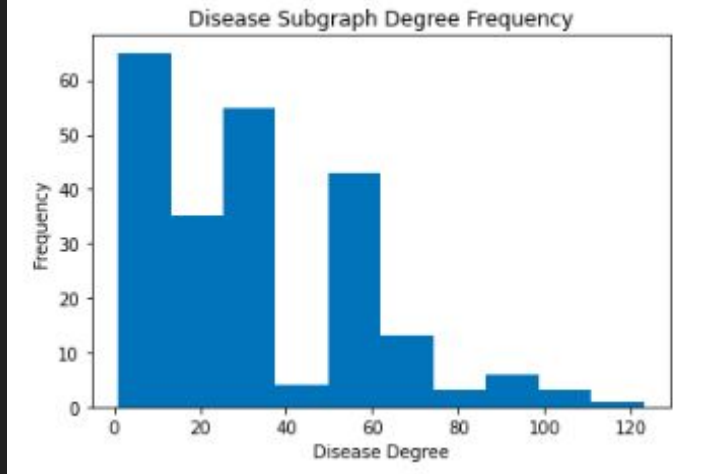


Figure 7: Degree distribution of Disease one mode projection

Degree Analysis

- Sorted the symptoms and diseases by maximum and minimum degrees
- Highest degree symptom and disease nodes are relatively expected as are the symptoms and diseases with the lowest degrees
  - Hypertension, heart disease/failure, HIV infections are all common diseases in the United States
  - Common symptom clusters include fever, fatigue, cough, body aches, vomiting

Highest symptom degrees:

- Body Weight, k = 36
- Anoxia, 26
- Vomiting, 22

Lowest symptom degrees:

- Purpura, k = 1
- Jaundice, 1
- Cyanosis, 1

Highest disease degrees:

- Hypertension, k = 123
- Heart Failure, 109
- HIV Infection, 107

Lowest disease degrees:

- Epilepsy, k = 1
- Craniocerebral Trauma, 1
- Aphasia, 1

Community Detection - Louvain Method

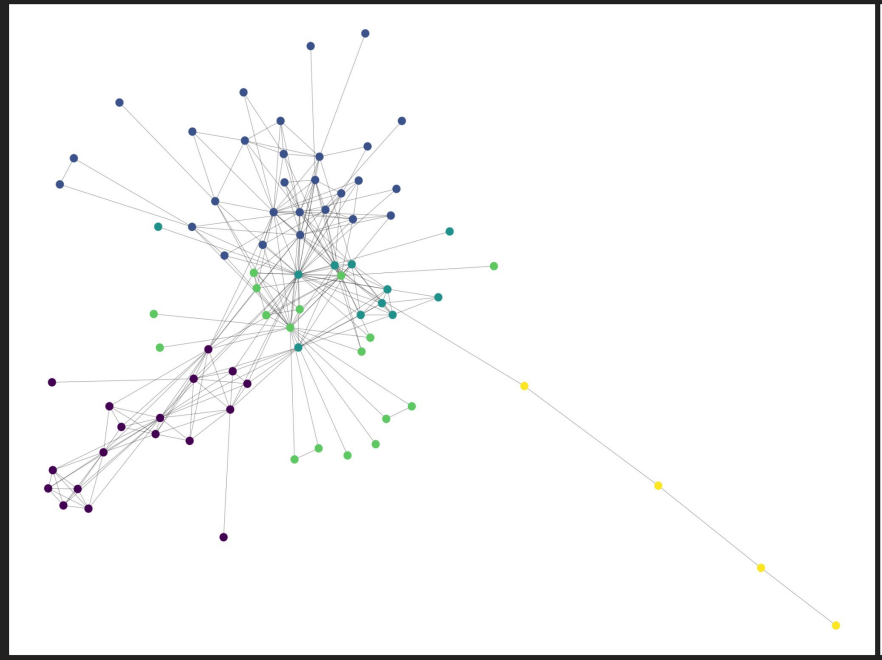


Figure 8: Clustering of symptom network based on Louvain algorithm. Note five symptom communities in total

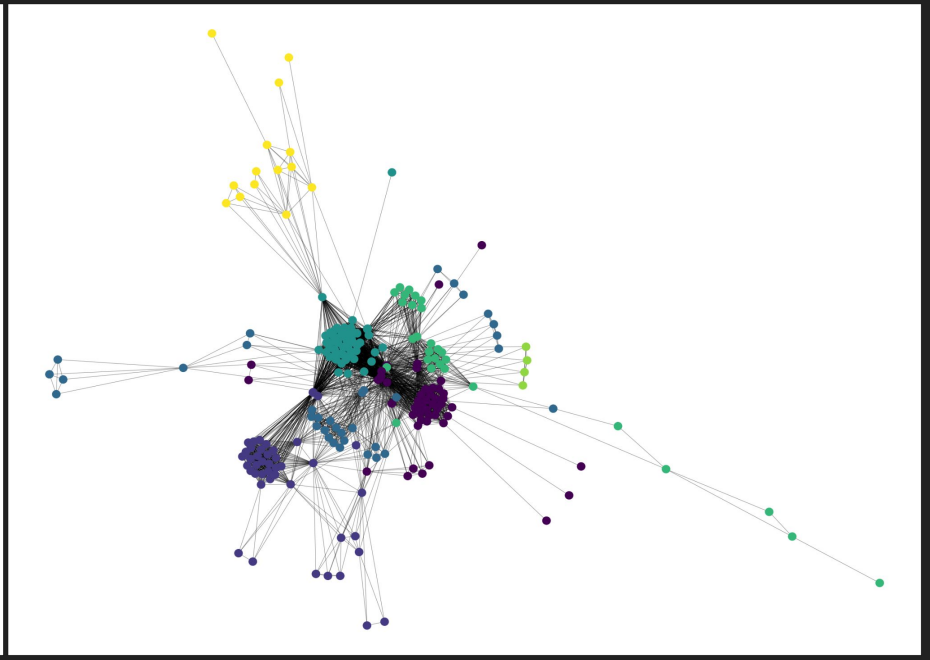


Figure 9: Clustering of disease network based on Louvain algorithm. Note seven disease communities in total

Community Detection - Louvain Method (Symptoms)

- Motor, Muscle Symptoms
  - Community 1: 'Anomia', 'Muscle Hypotonia', 'Hemianopsia', 'Muscle Rigidity', 'Gait Disorders, Neurologic', 'Delirium', 'Psychomotor Disorders', 'Supranuclear Palsy, Progressive', 'Hypokinesia', 'Consciousness Disorders', 'Perceptual Disorders', 'Catatonia', 'Tremor', 'Hallucinations', 'Confusion', 'Dyskinesias', 'Akathisia, Drug-Induced', 'Psychomotor Agitation'

- Bacterial Infections Symptoms
  - Community 2: 'Pruritus', 'Hemoptysis', 'Purpura, Thrombotic Thrombocytopenic', 'Fetal Distress', 'Heartburn', 'Psychophysiologic Disorders', 'Abdomen, Acute', 'Pain, Intractable', 'Respiratory Sounds', 'Constipation', 'Jaundice', 'Voice Disorders', 'Hot Flashes', 'Spasm', 'Diarrhea', 'Chest Pain', 'Hyperventilation', 'Diarrhea, Infantile', 'Dyspepsia', 'Fatigue', 'Nausea', 'Fever of Unknown Origin', 'Hoarseness', 'Fever', 'Vomiting', 'Abdominal Pain', 'Sleep Deprivation'

- Malnutrition, Eating disorder Symptoms
  - Community 3: 'Colic', 'Weight Gain', 'Body Weight', 'Hyperphagia', 'Bulimia', 'Thinness', 'Birth Weight', 'Overweight', 'Hirsutism', 'Weight Loss', 'Anorexia'

- Lung, Heart Symptoms
  - Community 4: 'Hydrops Fetalis', 'Edema', 'Hypoventilation', 'Cyanosis', 'Unconsciousness', 'Snoring', 'Cachexia', 'Hypothermia', 'Reflex, Abnormal', 'Urinary Incontinence, Stress', 'Intermittent Claudication', 'Cardiac Output, Low', 'Myoclonus', 'Anoxia', 'Hypercapnia', 'Neurologic Manifestations', 'Cheyne-Stokes Respiration'

- Vision Symptoms
  - Community 5: 'Amblyopia', 'Diplopia', 'Ataxia', 'Scotoma’

Community Detection - Louvain Method (Diseases)

- Respiratory/Heart Diseases
  - Community 1: Respiratory Distress Syndrome, Adult', 'Heart Failure', 'Infant, Newborn, Diseases', 'Brain Injuries', 'Liver Cirrhosis', 'Altitude Sickness', 'Brain Diseases', 'Pulmonary Disease, Chronic Obstructive', 'Anemia', 'Methemoglobinemia', 'Heart Arrest', 'Heart Septal Defects, Atrial', 'Neovascularization, Pathologic', 'Pulmonary Edema', 'Ventricular Dysfunction, Left', 'Myocardial Ischemia', 'Heart Defects, Congenital', 'Respiratory Insufficiency', 'Aneurysm, Dissecting', 'Apnea', 'Brain Ischemia', 'Pulmonary Heart Disease', 'Ischemia', 'Acidosis', 'Polycythemia', 'Gastroesophageal Reflux', 'Pulmonary Embolism', 'Postoperative Complications', 'Asthma', 'Esophageal Diseases', 'Myocardial Infarction', 'Tuberculosis, Pulmonary', 'Fetal Diseases', 'Lung Diseases', 'Urinary Incontinence', 'Hypertension, Pulmonary', 'Dyspnea', 'Reperfusion Injury', 'Hemorrhage', 'Sleep Apnea, Obstructive', 'Coronary Disease', 'Craniocerebral Trauma', 'Acute Coronary Syndrome', 'Coronary Artery Disease', 'Sleep Apnea Syndromes', 'Angina Pectoris', 'Lung Diseases, Obstructive', 'Epilepsy'

- Gastrointestinal Diseases
  - Community 2: 'Celiac Disease', 'Gastrointestinal Diseases', 'Gastroenteritis', 'Colonic Diseases, Functional', 'Helicobacter Infections', 'Campylobacter Infections', 'Salmonella Infections', 'Fecal Incontinence', 'Cross Infection', 'Intestinal Obstruction', 'Colitis', 'AIDS-Related Opportunistic Infections', 'Duodenal Ulcer', 'Colitis, Ulcerative', 'Intestinal Diseases', 'Peptic Ulcer', 'Enterocolitis, Pseudomembranous', 'Acquired Immunodeficiency Syndrome', 'Crohn Disease', 'Bacterial Infections', 'Hemolytic-Uremic Syndrome', 'Dehydration', 'Migraine Disorders', 'Rotavirus Infections', 'Giardiasis', 'Intussusception', 'Cholera', 'Colorectal Neoplasms', 'Rectal Diseases', 'Dysentery, Bacillary', 'Clostridium Infections', 'Peritonitis', 'Rectal Neoplasms', 'Malabsorption Syndromes', 'Pancreatic Neoplasms', 'Appendicitis', 'Irritable Bowel Syndrome', 'Intestinal Diseases, Parasitic', 'Enteritis', 'Escherichia coli Infections', 'Cryptosporidiosis', 'Gastritis', 'Intestinal Perforation', 'HIV Infections'

- Communicable Diseases
  - Community 3: Sleep Disorders', 'Respiratory Tract Infections', 'Somatoform Disorders', 'Seizures', 'Agranulocytosis', 'Hepatitis A', 'Leukemia', 'Anxiety Disorders', 'Bacteremia', 'Malaria', 'Hepatitis', 'Urticaria', 'Fibromyalgia', 'Neurotic Disorders', 'Headache', 'Sepsis', 'Drug Hypersensitivity', 'Dermatitis, Atopic', 'Common Cold', 'Urinary Tract Infections', 'Multiple Sclerosis', 'Malaria, Falciparum', 'Infection', 'Skin Diseases', 'Cholestasis', 'Pneumonia', 'Fatigue Syndrome, Chronic', 'Drug-Induced Liver Injury', 'Neutropenia', 'Pain', 'Staphylococcal Infections', 'Airway Obstruction', 'Bone Neoplasms', 'Liver Diseases', 'Laryngeal Diseases'

Community Detection - Louvain Method (Diseases)

- Pregnancy Associated Diseases
  - Community 4: 'Hypertrichosis', 'Prostatic Neoplasms', 'Kidney Diseases', 'Abnormalities, Drug-Induced', 'Fetal Growth Retardation', 'Burns', 'Diabetes Mellitus, Type 1', 'Starvation', 'Osteoporosis', 'Kidney Failure, Chronic', 'Depressive Disorder', 'Hypothyroidism', 'Obesity', 'Colonic Neoplasms', 'Nutrition Disorders', 'Breast Neoplasms', 'Cardiomegaly', 'Protein Deficiency', 'Neoplasms, Experimental', 'Schizophrenia', 'Uremia', 'Deficiency Diseases', 'Hypertension, Renal', 'Eating Disorders', 'Vitamin A Deficiency', 'Diabetes Mellitus', 'Pregnancy Complications', 'Lung Neoplasms', 'Cystic Fibrosis', 'Obesity, Morbid', 'Diabetic Nephropathies', 'Hyperlipidemias', 'Hyperglycemia', 'Fatty Liver', 'Hypercholesterolemia', 'Liver Neoplasms, Experimental', 'Alcoholism', 'Metabolic Syndrome X', 'Liver Neoplasms', 'Pregnancy in Diabetics', 'Substance Withdrawal Syndrome', 'Growth Disorders', 'Diabetes Mellitus, Experimental', 'Mammary Neoplasms, Experimental', 'Polycystic Ovary Syndrome', 'Protein-Energy Malnutrition', 'Adenocarcinoma', 'Arteriosclerosis', 'Malnutrition', 'Inflammation', 'Diabetes Mellitus, Type 2', 'Hyperinsulinism', 'Hyperthyroidism', 'Anorexia Nervosa'

- Osteo/Vision Diseases
  - Community 5: 'Obstetric Labor, Premature', 'Vision Disorders', 'Fetal Macrosomia', 'Pre-Eclampsia', 'Arthritis, Experimental', 'Hypertension', 'Pancreatitis', 'Pregnancy Complications, Cardiovascular', 'Jaundice, Neonatal', 'Infant, Premature, Diseases', 'Arthritis, Rheumatoid', 'Arthritis', 'Strabismus', 'Diabetes, Gestational', 'Retinopathy of Prematurity', 'Obstetric Labor Complications', 'Respiratory Distress Syndrome, Newborn', 'Spinal Cord Injuries', 'Retinal Diseases', 'Bone Marrow Diseases', 'Ophthalmoplegia', 'Refractive Errors', 'Nephrotic Syndrome', 'Thrombophlebitis', 'Glaucoma', 'Corneal Diseases', 'Venous Insufficiency', 'Toxemia', 'Eyelid Diseases'

- Vascular Diseases
  - Community 6: Vascular Diseases', 'Peripheral Arterial Disease', 'Peripheral Vascular Diseases', 'Arterial Occlusive Diseases'

- Mental Diseases
  - Community 7: 'Stroke', 'Parkinson Disease', 'Developmental Disabilities', 'Alzheimer Disease', 'Mental Retardation', 'Cerebral Infarction', 'Aphasia', 'Abnormalities, Multiple', 'Psychoses, Substance-Induced', 'Movement Disorders', 'Delirium, Dementia, Amnestic, Cognitive Disorders', 'Psychotic Disorders', 'Cognition Disorders', 'Dementia'

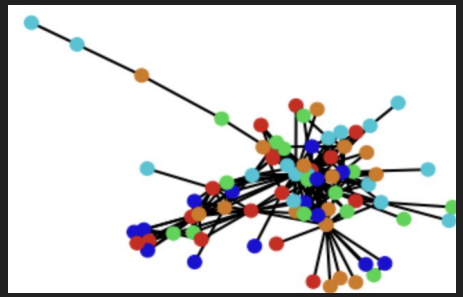
Community Detection - 3352 DC-SBM Methods

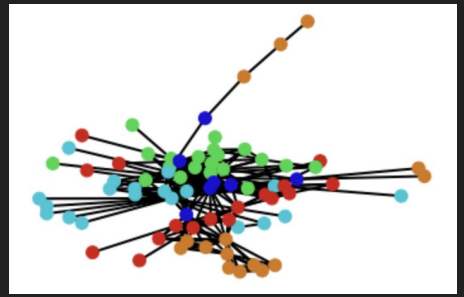
DC-SBM ran on Symptom, Disease One Mode Projections

Random Partition (Baseline)

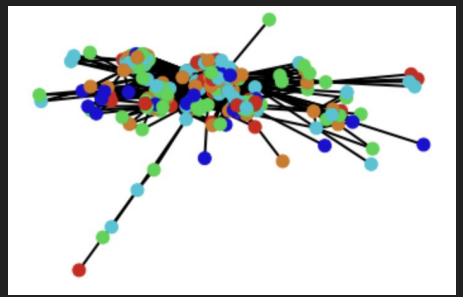
Locally Greedy Partition

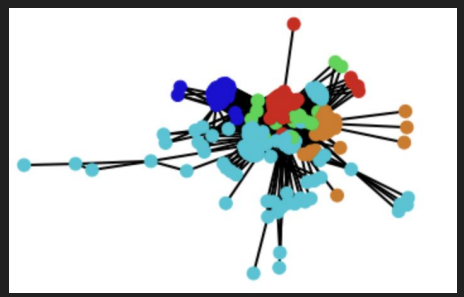
Symptom





Disease





kr Symptom:
[120. 46. 98. 176.
64.]

wrs Symptom:
[[26. 7. 4. 45. 38.]
[7. 24. 11. 0. 4.]
[4. 11. 78. 5. 0.]
[45. 0. 5. 126. 0.]
[38. 4. 0. 0. 22.]]

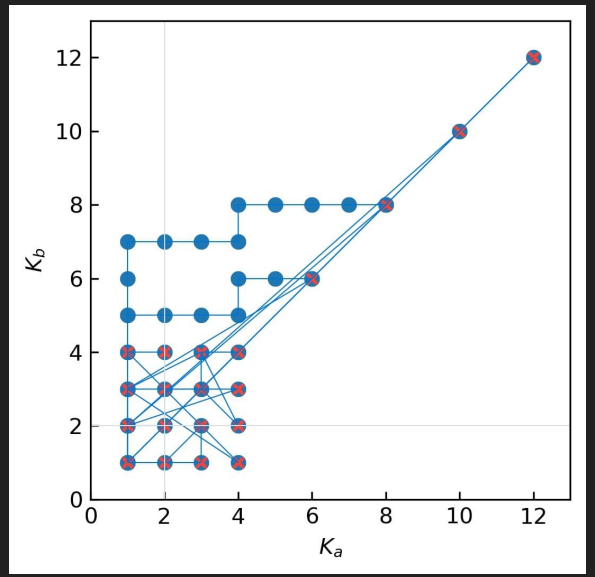
kr Disease:
[434. 1986. 1173.
3262. 899.]

wrs Disease:
[[32. 48. 34. 167. 153.]
[48. 1556. 349. 29. 4.]
[34. 349. 300. 490. 0.]
[167. 29. 490. 2576. 0.]
[153. 4. 0. 0. 742.]]

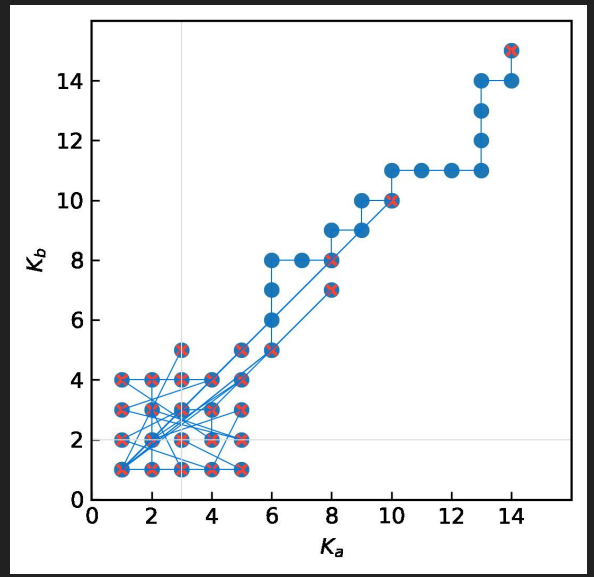
Community Detection - biSBM from Larremore, Yen

- As described in the background, here are the results:

Kernighan-Lin Inference

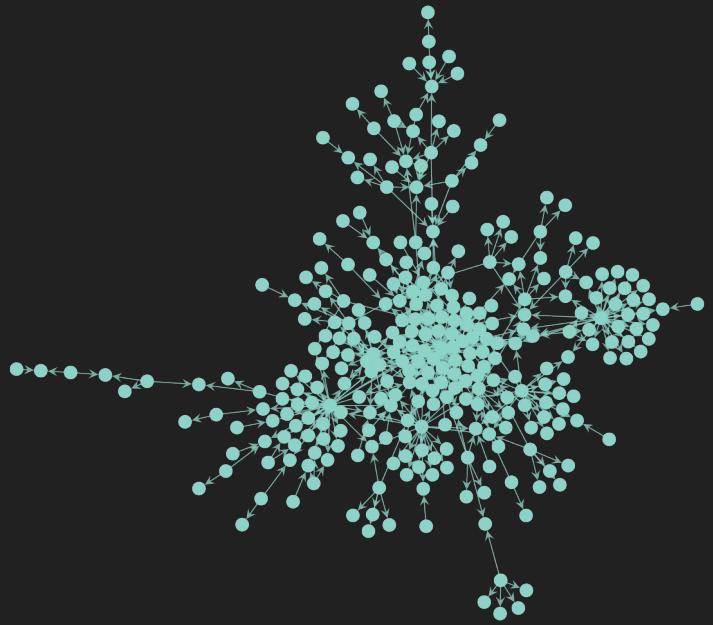


Markov Chain Monte Carlo Inference

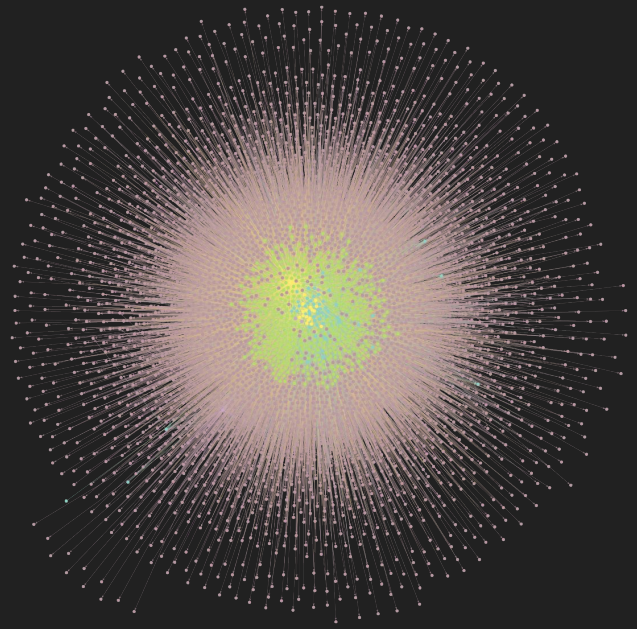


Community Detection - Native graph-tool Method

PubMed occurrences, at least 150:



All PubMed occurrences:



Conclusions

- Louvain clustering indicates more disease communities as compared to traditional DC-SBM methods (7 modules as compared to 5 modules)
  - Louvain is a modularity optimization so it tends toward giving more variable communities whereas DC-SBM makes less assumptions so it is more inline with recreating the network based off the summary statistics only
- Mixing matrices from locally greedy heuristic DC-SBM on one mode projections indicates an assortative community structure
  - This is expected as comorbidities of a particular disease should be related closely together in the same community
- Running the DC-biSBM on the PubMed network w/ the popularity filter gave 2-3 communities per set
- One important caveat to note:
  - Symptoms are subjective!
    - Symptoms represent the surface level representation of a disease that is actually observed by physicians
    - Patients' experience of major classes of symptoms (objectivity of symptom classes) still remains a challenge in the medical field

Next Steps

- Identify further relations between disease and symptom states
  - Is it possible to draw correlations between distant yet related disease states (more than one neighborhood level deep from the original clustering)?
- Overlay protein-protein interaction metadata over disease and symptom one mode projections
  - Effectively would create an interactome of all physical interactions in cells during the disease
    - Potentially identify shared proteins and protein interactions amongst disease states
  - Emulate methods and results similar to Jostins et al. (2012)
    - Studied relations between various inflammatory bowel diseases which identified common proteins between Ulcerative Colitis (UC) and Crohn’s disease

Questions?

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. doi:10.1088/1742-5468/2008/10/p10008

Clauset A., Newman M.E.J., Moore C. Finding community structure in very large networks. Phys. Rev. E, 70 (6) (2004), Article 066111, 10.1103/PhysRevE.70.066111

Hidalgo, C. A., Blumm, N., Barabasi, A. L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS. Comput. Biol.* 5, e1000353 (2009).

Jostins L., et al. “Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease”. Nature 491, 119–124 (2012).

Larremore D.B., Clauset A., and Jacobs A.Z. Efficiently inferring community structure in bipartite networks. Phys. Rev. E 90, 012805 (2014).

Lu, Z., Wahlström, J. & Nehorai, A. Community Detection in Complex Networks via Clique Conductance. *Sci Rep* 8, 5982 (2018).

Peel L., Larremore D. B., and Clauset A., "The ground truth about metadata and community detection in networks." Science Advances 3(5), e1602548 (2017).

References

Rzhetsky, A., Wajngurt, D., Park, N. & Zheng, T. Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA* 104, 11694–11699 (2007).

Yen T., and Larremore D.B. Community Detection in Bipartite Networks with Stochastic Block models,, Physical Review E 102, 032309, (2020).

Zhou X., Menche J., and Barabási AL., “Human symptoms–disease network”. Nat Commun 5, 4212 (2014). https://doi.org/10.1038/ncomms5212