# Evaluating Community Detection Methods for the Human Symptoms-Disease Network

Sidhant Puntambekar, Behzod Mirpochoev

May 2, 2022

## 1 Introduction

Diseases and their associated symptoms in humans are often of great interest for the medical and pathological research communities. Disease association patterns discovered between various academic disciplines has aided in overall human amelioration. As the transition to interdisciplinary study for disease causal analysis becomes more widespread, the need for efficiency and veracity becomes vital. Using a series of symptoms to discern a particular disease classification is the foundational basis of medical diagnosis.

If one examines the literature available in online databases, such as PubMed and the Medical Subject Headings (MeSH), a network model can be formed for associating independent disease states and shared symptoms. Zhou et. al. constructed such a model in their research termed the Human Symptoms-Disease Network (HSDN)[9]. In total, the HSDN comprises 4,219 unique disease nodes, 322 unique symptom nodes, and 147,978 edge connections which represents approximately 98.5% of all symptoms and 95.0% of all diseases contained in the PubMed/MeSH metadata vocabulary[9].

Overall, the primary results from the HSDN study were that the symptom-based similarity of two diseases correlated strongly with shared genetic associations, and their associated first and second order protein interactions[9]. Extending this representation by detecting communities within the HSDN may assist the medical community in finding previously unseen associations and comorbidities. The immediate question then boils down to which community detection method is of most utility. Limitations arise when attempting to utilize traditional methods for this goal, as the HSDN is bipartite by nature[9].

Lu et al. describe such limitations by outlining the issue with implicit parameter choices, loss of information through one-mode projections, and lack of interpretability[8]. Larremore et al. note that a major difference between ordinary community detection methods and those specific to bipartite networks are the underlying assumption of assortativity; a traditional stochastic block model (SBM) prefers to find either very assortative modules, very disassortative modules, or some mixture thereof[3]. Larremore et. al. further describe a method which is particular to the bipartite community detection problem in their bipartite stochastic block model (biSBM), which assumes disassortative modules upfront[3]. Using this model, the paper applies the biSBM to several real world networks including the Southern Women network (interactions of 14 women in Mississippi with various social events), a Malaria network consisting of genes and amino acid substring relationships, and the IMDb network consisting of 53,158 actors and the 39,768 movies they appeared in. The research goal for this study is to leverage the biSBM model framework from Larremore et al.[3] and apply it to the bipartite HSDN network from Zhou et al.[9] in order to perform community detection and explore various modules of interrelated disease groupings.

## 2 Methods

To generate a bipartite network, two sets are created to form connections with nodes of the opposing set. Nodes in the HSDN represented symptom and disease terms, while edges represented a relationship between a disease and a shared symptom. Another disease within the network often has the same symptom thereby forming similarities of disease states[9]. To generate modules within a given network, a stochastic process is used in order to capture partitions through inference.

The techniques employed take in various parameters in order to determine a partitioning that best describes the communities within the network. The stochastic block model (SBM) routinely partitions nodes within a network into some partition $p$ and a mixing matrix of dimension $P \times P$ ($P$ is the number of communities)[2]. Each matrix entry represents a given edge count $e$ such that each $e_{rs}$ is the number of edges between modules $r$ and $s$[2]. Edges, however, are routinely placed with the same probability given a node within one module and that of another. An issue arises when the degree count of each node within a network is considered; the SBM formulation is not reflective of the degree variations of many networks[2]. Thus, a degree-corrected form is proposed to adjust the model, referred to as the degree-corrected stochastic block model (DC-SBM); this form generally takes in an additional parameter $k$ to reflect the degree structure of each node within a network[2].

Another issue arises when SBM partitions are generated from a network which has abnormal underlying structure (for example, the property of bipartiteness involves the separation of node sets). Larremore et al. accounted for this in their DC-biSBM model which employs extended parameter choices to fit a bipartite network[9]. The nodes in each respective set are accounted for in quantities $N_a$ and $N_b$. The nodes are then split into groups of type $a$ in set $K_a$ and $b$ in set $K_b$. The mixing matrix $K \times K$, where $K = K_a + K_b$, is then constructed such that edges which would connect nodes of the same type are zeroed to retain their bipartite property[9]. The total number of communities $g$, the mixing matrix $\omega$, the expected degree of each vertex $\Theta$, and the community partition matrix $T$ are also accounted for in the DC-biSBM[9].

Other community detection algorithms were also accounted for in evaluation of the aforementioned methods. Louvain clustering as described by Blondel et. al. is one such method explored[6]. In the Louvain approach, each node is given its own module. A change in modularity is calculated based on taking node $i$ from its own community and placing it in the community of node $j$[6]. The community which results in the greatest modularity increase (on a scale from $[-\frac{1}{2}, 1]$) for a node is the new community in which the node is assigned[6]. The process is repeated until there is no such modularity increase or a tolerance threshold is met[6]. Communities are then aggregated and considered as a singular ‘node’, representing relationships via self-loops for nodes within the same community that connect to one another[6]. The number of nodes connecting to nodes of a different community are then recorded. After this aggregation, the optimization step is then repeated. Both processes are repeated until there is no apparent increase in modularity[6].

# 3 Results

## 3.1 Data Quality Control

From a general survey of the HSDN, the MeSH metadata occurrences of each symptom and disease term were measured in order to settle on a popularity filter in order to remove the least popular outlier node pairs from the PubMed corpus. The term frequency figure below demonstrates the occurrences per disease term (a similar trend exists for symptom terms as well):

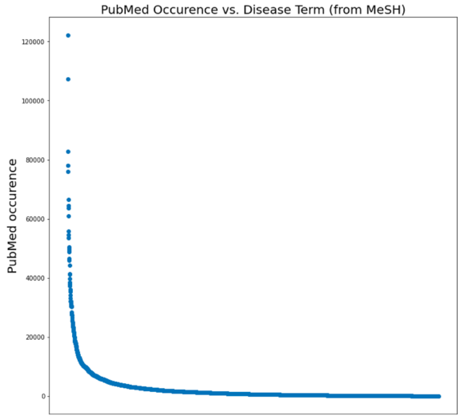
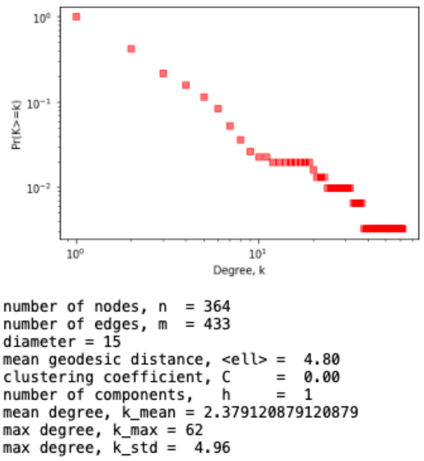


Figure 1: Disease term popularity in MeSH. Many of the diseases have occurrences of less than 150 mentions.

With a large portion of the data spread across a low number of overall PubMed occurrences, filtering the node sets created a more manageable network size while still maintaining a representative mixture of symptom/disease nodes. A popularity filter of greater than 150 PubMed occurrences was chosen since diseases and symptoms that were most frequently diagnosed from patients in the HSDN were of primary interest.

## 3.2 Bipartite Network Representation

A bipartite representation of the filtered HSDN network was constructed and revealed high edge density between the symptom and disease node sets (see Appendix code for full bipartite representation figure). The following summary statistics were calculated that describe the bipartite representation of the HSDN:



```
number of nodes, n  = 364
number of edges, m  = 433
diameter = 15
mean geodesic distance, <ell> =  4.80
clustering coefficient, C     =  0.00
number of components,   h     =  1
mean degree, k_mean = 2.379120879120879
max degree, k_max = 62
max degree, k_std =  4.96
```

Figure 2: Summary statistics of filtered HSDN network.

The bipartite HSDN representation had a diameter of 15 edges and an overall heavy tailed cumulative degree distribution. Additionally, the mean geodesic distance of the network was around 4.8 to 5 edges. The clustering coefficient remained 0 due to the bipartite representation.

## 3.3 Symptom, Disease One-Mode Projections

The bipartite representation of the network was then projected into a symptom one-mode projection and disease one-mode projection respectively. This was done to accommodate the single mode community detection algorithms including the Louvain method and DC-SBM. A visualization of both one-mode projections are displayed below:

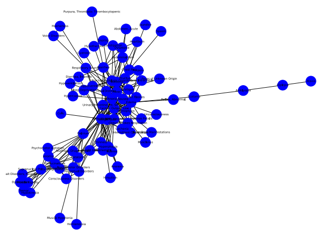


Figure 3: Symptom one-mode projection

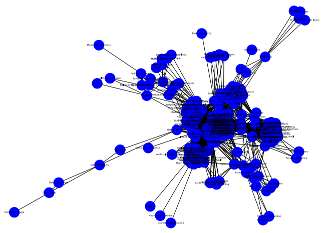
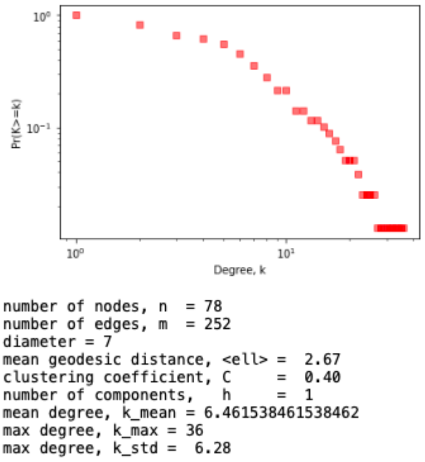


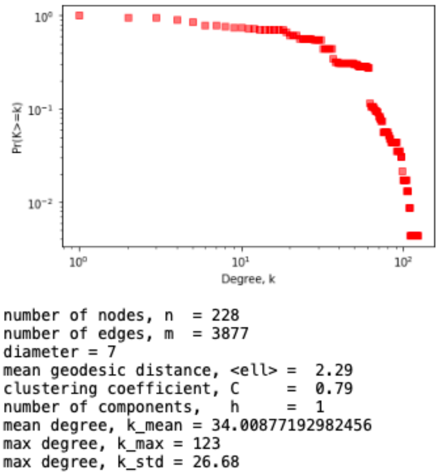
Figure 4: Disease one-mode projection

Overall, both one-mode projections demonstrate areas of high centrality structured around defined symptom and disease node clusters. Both the disease and symptom networks on visual inspection have several clusters that appear assortative in nature meaning that nodes from the same clusters have higher edge connections to each other as compared to the nodes external to that community (community detection from Louvain and DC-SBM will verify this later). The symptom and disease one-mode projections had the following summary statistics:



```
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 5: Symptom one-mode projection summary statistics



```
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: Disease one-mode projection summary statistics

From the one-mode projection summary statistics, it is important to note the high clustering coefficients over the original bipartite network (0.40 for the symptom projection and 0.79 for the disease projection).

### 3.4 Louvain Method

Using the Louvain method on each one-mode projection yielded the following clustering diagrams:

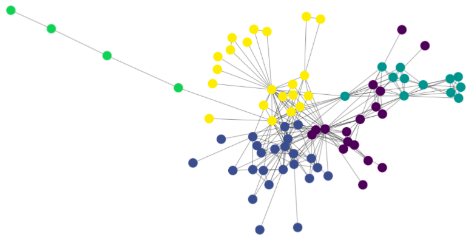


Figure 7: Symptom one-mode projection community detection, Louvain algorithm

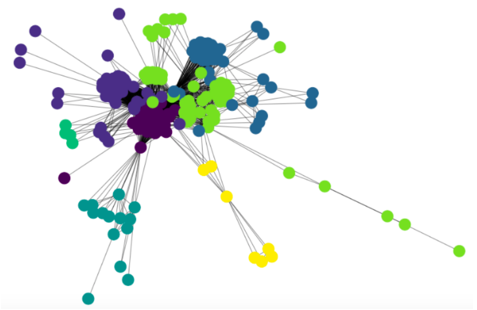


Figure 8: Disease one-mode projection community detection, Louvain algorithm

The Louvain clustering algorithm managed to find a total of five symptom communities and seven disease communities from the projections.

For the symptom clusters, labels were assigned to the five communities detected that included motor/muscular symptoms, bacterial infection symptoms, malnutrition/eating disorder symptoms, respiratory/cardiovascular symptoms, and vision degeneration symptoms.

For the disease clusters, labels were assigned to the seven communities detected that included respiratory/cardiovascular diseases, gastrointestinal diseases, communicable diseases, pregnancy related diseases, osteo/vision disorders, broad cardiovascular disorders, and mental disorders.

### 3.5 DC-SBM Method

Using the DC-SBM method on the symptom and disease one-mode projections, five communities in both were reliably detected using a locally greedy heuristic approach:

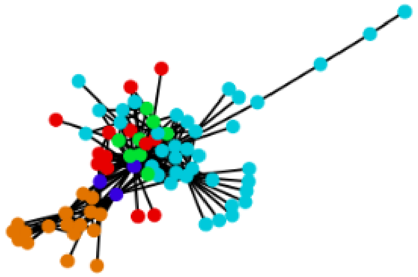


Figure 9: Symptom one-mode projection community detection, DC-SBM algorithm

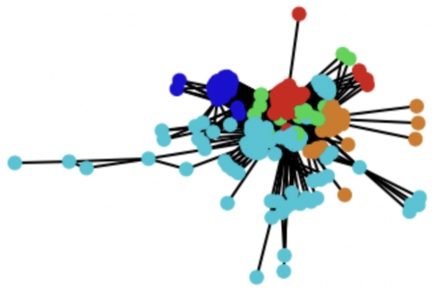


Figure 10: Disease one-mode projection community detection, DC-SBM algorithm

From the DC-SBM results, a mixing matrix was generated as $\omega_{rs} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}\delta_{r,z_i}\delta_{s,z_j}$ (the number of "stubs" of edges whose endpoints are in group $r$ and group $s$) and $\kappa_r = \sum_{s=1}^{c} \omega_{rs}$ (the "expected total degree" of all nodes in group $r$)[2].

For the symptom network, the expected degree structure was computed as $\kappa_r = \begin{pmatrix}120 & 46 & 98 & 176 & 64\end{pmatrix}$ and the mixing matrix was computed as

$$\omega_{rs} = \begin{pmatrix} 26 & 7 & 4 & 45 & 38 \\ 7 & 24 & 11 & 0 & 4 \\ 4 & 11 & 78 & 5 & 0 \\ 45 & 0 & 5 & 126 & 0 \\ 38 & 4 & 0 & 0 & 22 \end{pmatrix}.$$

For the disease network, the expected degree structure was computed as $\kappa_r = \begin{pmatrix}434 & 1986 & 1173 & 3262 & 899\end{pmatrix}$ and the mixing matrix was computed as

$$\omega_{rs} = \begin{pmatrix} 32 & 48 & 34 & 167 & 153 \\ 48. & 1556 & 349 & 29 & 4 \\ 34 & 349 & 300 & 490 & 0 \\ 167 & 29 & 490 & 2576 & 0 \\ 153 & 4 & 0 & 0 & 742 \end{pmatrix}.$$

### 3.6 DC-biSBM with Statistical Inference

In order to examine the network at a higher resolution, the entire bipartite structure should be preserved when the network is under examination. Using Kernighan-Lin inference as a method to aid community detection, the resulting partition and path are generated:

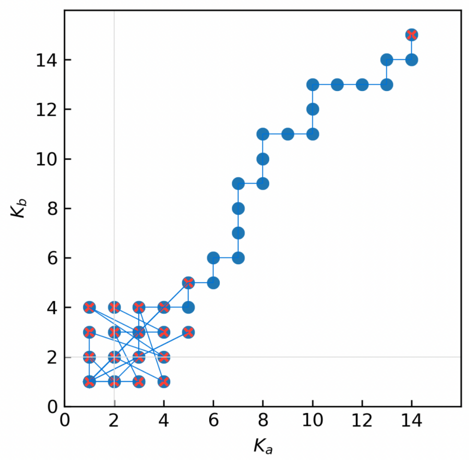


Figure 11: KL partitioning path, biSBM on HSDN

The results arise from an initial number of $K_a, K_b = 14, 15$ communities. To note, a metric for comparison of representations is one which best compresses the model, as measured through the minimum description length (MDL)[8]. Once computed, the smallest description length should be strongly considered for model selection. The inference tool then utilizes a minimization of description length in order to select the best partitioning[8]. The results for the the KL inference model settled on $K_a = 2$ and $K_b = 2$, with an MDL of approximately 1724.66.

Considering another inference tool for model comparison, Markov Chain Monte Carlo inference is used. The resulting partition and path generated are:

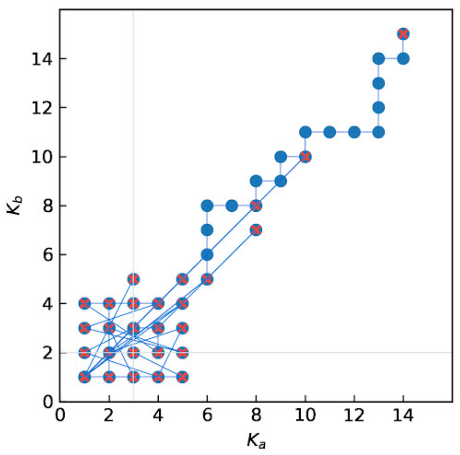


Figure 12: MCMC partitioning path, biSBM on HSDN

The results are for an initial selection of $K_a, K_b = 14, 15$ communities. For the resulting partition, $K_a = 3$ while $K_b = 2$; the MDL was approximately 1703.71.

## 4 Discussion

Some preliminary results to note from the one-mode projection analysis are the number of communities that the Louvain algorithm produced as opposed to the locally greedy DC-SBM model. In terms of the one-mode disease representation, Louvain managed to detect seven communities as opposed to five generated by DC-SBM. This is due to Louvain being a modularity optimization thereby giving more variable numbers of communities whereas DC-SBM makes less assumptions, so it is more inline with recreating the network based off the summary statistics alone[6 10]. In terms of the symptom one-mode projection, both the Louvain and DC-SBM methods produced five distinct communities.

Following a deeper dive into the biological themes that each community contained from the Louvain algorithm, the largest disease cluster was comprised of respiratory and cardiovascular diseases such as pulmonary embolisms, coronary artery diseases, and sleep apnea. The largest symptom cluster was comprised of bacterial infection symptoms such as diarrhea, dyspepsia (discomfort while digesting food), fever, and vomiting. These symptoms had links to prominent gastrointestinal diseases in the HSDN such as gastroenteritis[9].

Overall, communities detected in the disease projection either had very clear associations of interrelated disease states, or were heterogeneous (comprising unrelated disease states). One of the prominent disease communities was mental disorders with nodes such as Parkinson's disease, Alzheimer's, and aphasia (language disorder that affects communication). These diseases had definite links in the HSDN to symptoms such as hallucinations, psychomotor agitation (mood swings), and tremors. On the other hand, one disease community was broadly labeled as ”pregnancy related disorders”, but contained nodes ranging from neoplasms (abnormal growths preceding cancer), schizophrenia, and diabetic precursors such as hyperglycemia (high blood sugar). For convoluted disease communities, further optimizations in community detection can be undertaken to investigate whether these initial communities can be broken down into a finer granularity.

When looking at the results from the community detection algorithms, it is important to note loss of node context when taking one-mode projections. If a one-mode projection is taken, information regarding the projected nodes in the context of the bipartite representation could be lost. This is due to projections representing each node as a group of correlated edges. With these assumptions in mind, projections are still a useful tool for structural analysis similar to substrate-product and reaction networks in metabonets[1].

Both of the projection mixing matrices from DC-SBM had very strong diagonals compared to the anti-diagonal entries leading to a more assortative community structure. This indicates that nodes in certain communities are more interconnected to each other rather than other communities. Overall, this gives credence to Zhou et al.'s argument that symptom and disease groups may co-occur in parallel if a patient is suffering from a particular ailment[9]. Anticipatory actions can be taken in a medical context to protect against comorbidities. However, this should only occur if two diseases are highly connected together in the same community, and if protein-protein interaction data supports a potential common drug target.

As for the results of the DC-biSBM methods, the Kernighan-Lin (KL) inference gives a partitioning of $K_a, K_b = 2$. However, when comparing the minimum description length of Kernighan-Lin (KL) and Markov Chain Monte Carlo (MCMC), the latter appears to be a better descriptor with $K_a, K_b = 3, 2$. This is consistent with the sum $K = K_a + K_b$, as the DC-SBM method partition is $K = 5$. Examining the nature of these modules would be the next logical step by assigning community labels for partitions in each set of disease and symptom nodes and overlaying a protein-protein interaction network.

In terms of next steps, an investigation into breaking down communities with a mix of unrelated diseases should be conducted. This will allow a clearer view into the precise comorbidities of any one disease from the HSDN which has impact when assigning future risk predictions to patients who exhibit symptoms indicative of disease states in a medical context.

Another next step could be to overlay protein-protein interaction data onto the disease/symptom projections. A protein-protein interaction (PPI) approach was investigated by Jostins et al. to identify comorbidities of inflammatory bowel diseases such as ulcerative colitis and Crohn's disease[7]. For this next step, an extrapolation of PPI analysis to the various communities detected through Louvain and DC-SBM clustering could be conducted in order to verify potential drug targets. Overall, the community detection algorithms were able to identify areas of shared interactions in the HSDN. Natural follow up inquiries such as identifying common protein drug targets and fine-graining the communities themselves will provide further insights into disease state relatedness and elucidate potential treatment approaches for a variety of patients in the future.

# 5 Appendix

The Python/Jupyter notebook code was too large to include in the paper. A link to our code repository is available here. Full size versions of each of the respective figures in the Results section can be found in our Python Jupyter Notebook for Louvain and DC-SBM on one-mode projections here and DC-biSBM analysis here.

## References

[1] Clauset A. Lecture notes 11: Metabolic networks, structure. 2022.

[2] Clauset A. Lecture notes 6: Modular networks, inference. 2022.

[3] Larremore D.B. Clauset A. and Jacobs A.Z. Efficiently inferring community structure in bipartite networks. *Phys. Rev. E, 90*, 2014. https://arxiv.org/abs/1403.2933.

[4] Clauset A. Newman M.E.J. Moore C. Finding community structure in very large networks. *Phys. Rev. E, 70*, 2004. https://arxiv.org/abs/cond-mat/0408187.

[5] Peel L. Larremore D.B. and Clauset A. The ground truth about metadata and community detection in networks. *Science Advances*, 2017. https://www.science.org/doi/10.1126/sciadv.1602548.

[6] 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. https://arxiv.org/abs/0803.0476.

[7] Jostins L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 2012. https://www.nature.com/articles/nature11582.

[8] Lu Z. Wahlström J. and Nehorai A. Community detection in complex networks via clique conductance. *Scientific Reports*, 2018. https://www.nature.com/articles/s41598-018-23932-z.

[9] Zhou X. Menche J. and Barabási AL. Human symptoms–disease network. *Nature*, 2014. https://www.nature.com/articles/ncomms5212.

[10] Yen T. and Larremore D.B. Community detection in bipartite networks with stochastic block models. *Phys Review E 102*, 2020. https://arxiv.org/abs/2001.11818.