Names: Behzod Mirpochoev, Sidhant Puntambekar

Background Material:

With the discovery of inherent relationships between the molecular origins of disease to their realized phenotypes in human subjects, the necessity to model these interactions through networks is crucial for identifying unexpected disease associations and their underlying molecular causes as well as initiating the genesis of potential drug design/development. Zhou, et al. in 2014 explored this idea by constructing a large-scale disease-symptom graph with data aggregated through the PubMed Medical Subject Headings (MeSH) metadata catalog (bibliographic data from PubMed papers). After filtering for co-occurrences of at least one disease and symptom metadata term as well as extraction of the specific disease-symptom relationships, the resultant Human-Symptoms Disease Network (HSDN) comprised a set of interactions between disease classifications and their observed clinical symptoms. Each node corresponds to either a particular human disease or associated symptoms. Edges between nodes in the network are weighted by the similarities of their respective symptoms.

In total, the HSDN comprises 4,219 disease nodes, 322 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. Overall, the primary results from the HSDN were that the symptom-based similarity of two diseases correlated strongly with shared genetic associations and their associated first and second order protein interactions. A particular example mentioned by the authors are the associations between diseases in the overarching inflammatory bowel disease (IBD) group such as Crohn's disease (CD) and ulcerative colitis (UC). The HSDN identified close to 78 shared symptoms among both of the diseases and nine associated genetic risk loci. This finding was supported with several genome-wide association studies from Jostins, L. et al. in 2012.

Larremore et al. in 2014 investigated the potential to create degree-corrected stochastic block models with bipartite networks designated as DC-biSBMs in order to solve the bipartite community detection problem. Bipartite networks exhibit community detection patterns similarly to their unipartite counterparts but previous approaches had issues with loss of information through one-mode projections and implicit parameter choices. Larremore et al. point out that a major difference between ordinary SBM networks and biSBM networks with regards to community detection are that biSBMs are assuming disassortative groups/modules while 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 of type *a* into K_a groups and nodes of type *b* into K_b groups. The groups interrelationships are then represented through a K x K matrix ω (where $K = K_a + K_b$) where entries that would connect nodes of the same type are zeroed which enforces the bipartite network structure. Larremore et al. then build a DC-biSBM model with parameters *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), and *T* (matrix representing community partitions). Using this model, the paper applies the

DC-biSBM to several real world networks including the Southern Women network (interactions of 14 women in Mississippi with various social events), the Malaria network consisting of genes and amino acid substring relationships in *Plasmodium falciparum,* and the IMDB network consisting of 53,158 actors and the 39,768 movies they appeared in.

We plan to leverage the DC-biSBM model framework from Larremore et al. and apply it to the bipartite HSDN network from Zhou et al. in order to perform disease module community detection and explore various modules of interrelated disease groupings.

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?

Anticipated Findings:

The HSDN network required a large aggregation of disease and symptoms data from a large aggregation database. In addition, the resultant HSDN presented by Zhou et al. was bipartite in nature (a relation between the diseases themselves and resultant symptoms). We are interested in recreating this network using a degree-corrected stochastic block model network in order to reduce the dimensionality of the disease-symptom relationship data as well as capture the degree structure and modular design of the disease-symptom interactions. We then plan on examining certain disease modules and the connections between them to investigate which ones are tightly coupled in order to hopefully identify common disease states, potential molecular causes of the diseasek and genetic areas of interest.

Data and Algorithms:

We plan to use a bipartite degree-corrected stochastic block model to approximate the structure of the HSDN followed up by running community detection on the network to deduce relationships between groups/modules of common diseases and symptoms. Following the description by Larremore et al. 2014, the construction of an adjacency matrix *A* must be in place, constructed from the nodes and edges resulting from the supplementary data. Further, the vertex types t_i that are found to be *a* are assigned to K_a groups according to a uniformly random manner; the ones found to be *b* are assigned to K_b according to the same uniformly random manner. A vertex may be moved from one type group to another provided that for some T_r and T_s, they are equal. The likelihood function within Larremore et al. 2014 is to be maximized when considering all possible moves. If an improvement cannot be made from one step to the next, the move that least decreases the likelihood function is preferred. The algorithm concludes by running through iterations of attempting to optimize the score and finding a place in which no improvements may be made as described.

In order to create this bipartite degree-corrected stochastic block model network, we plan to use the raw data files provided by Zhou et al. as well as adapting the method from Larremore et al., 2014 to create a bipartite degree-corrected stochastic block model and perform community detection on it to highlight modules of interest. With regards to the testing of accuracy of certain models, we shall use the Area Under Curve method to perform such analysis. If time permits, we may also investigate other algorithms with which to analyze against the DC-biSBM, as outlined in Larremore, et al. The data we plan to analyze is from supplementary sets from the Index of Complex Networks (ICON). From ICON, there is data on symptom instances as compared with disease instances within published medical literature metadata aggregated from PubMed.

Resources:

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

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

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

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

Zhou et al. Supplementary information

- Supplementary Data 1: List of all 4,442 diseases within PubMed MeSH and their occurrence.
- Supplementary Data 2: List of all 322 symptoms within PubMed MeSH and their occurrence.
- Supplementary Data 3**:** Table includes 147,978 records of symptom and disease relationships.