**6/10/2019**

Dataset gene\_disease.tsv

* There are 1,137,270 instances with selected columns where selected columns are the following: geneId, diseaseId, and diseaseClass.

1,137,270 instances

Number of unique value of each columns are as followed

geneId = 16629

diseaseId = 9464

diseaseClass = 815

pmid = 398198

source = 14

**Selected Classes of gene\_disease.tsv is the following, rank by frequency in descending order.**

label = 0050686, count = 157 "organ system cancer"

label = 16, count = 87/115 "integumentary system disease"

label = 3093, count = 75/47 "nervous system cancer"

label = 0080001, count = 67 "bone disease"

label = 863, count = 66 "nervous system disease"

label = 77, count = 65 "gastrointestinal system disease"

label = 0050155, count = 64 "sensory system disease"

label = 0050117, count = 58 "disease by infectious agent"

label = 66, count = 57 "muscle tissue disease"

label = 0080006, count = 50 "bone development disease"

**new\_mega\_classes (highest possible label classes)**

0050117, disease by infectious agent

7, disease of anatomical entity

14566, disease of cellular proliferation

150, disease of mental health

0014667, disease of metabolism

0080015, physical disorder

225, syndrome

**Dataset gene\_disease\_50000\_label\_no\_None.txt**

Dataset of the first 50000 qualified instances of gene\_disease.tsv where all instances belong to 1 of the selected mega classes mention above.

50000 = the first 50000 qualified instances of gene\_disease.tsv.

Qualified instances are instances that have all the selected features: geneId, diseaseId, and other features. Other possible features are diseaseType, diseaseClass, diseaseSemantic Type,and score. Meaning of these features are not clear to me. So,as of now, I decide not to use any features. Many papers do not use additional feature to do node classification.

Label = label column is added at the right of the dataset

It is a bipartite dataset where gene only connects to disease and disease only connects to genes. I am not sure if I should use diseaseClass as a feature when we are trying to do label nodes by classes. DiseaseClass has the following format

no\_None = there is no None class. None class is class that is not in the selected classes.

* Out of the first 50000 stances in the gene\_disease, there are 1995 instances
* Training and testing are separated randomly with no further criteria.
  + In full network where all nodes are included. Number of edges is a lot more than number of uniq\_nodes. However, in the small network, such as data with 50,000 instances. I am not sure if there should have any criteria to choose which nodes should be included in the training and testing.
  + I do not think there should be any criteria other than splitting them randomly and maintain the amount of number instances ratio per label. This is because degree of nodes in 2 bipartite groups hold necessary information for building GCN.

Comments:

+++++++++++++++++

Type of nodes:

1. Disease node
2. Gene ID node
3. Verify that each diease node is mapped to one class lable or there is a possibility to map to multiple disease lables
4. Check whether it is possible to lable Gene nodes, and how?
5. Currently, we have COPD related diease: diseaseId = 9464. This is rather too large, considering human disease network. Please double check the connections are correct.
6. Let’s focus on build the complete network (COPD) using all instances, then we can reduce the network, if it’s too computational expensive to learning the network. (check code to save)
7. Let’s plan to publish COPD network by June 22 through project website (also report to NSF).

+++++++++++++++++++++++++++++++++

**6/14/2019**

disease\_gene.tsv is used to collect disease label as well as disease’s instances.

Disease label/classes are commorbidiy of copd as shown in picture below.

Number of classes is 9.

doid\_label\_dict = {'3393': 'coronary artery disease'**,** '2841': 'Asthma'**,** '9351': 'diabetes mellitus'**,** '0060224': 'atrial fibrillation'**,** '6000': 'congestive heart failure'**,** '10763': 'hypertension'**,** '1168': 'familial hyperlipidemia'**,** '1596': 'mental depression'**,** '6713': 'cerebrovascular disease'}

Step to generate disease Instances with selected classes:

* I manually selected it to only includes the most relevant comorbidities by adding 1 class at time until number of instances is large enough (in this case 11657 instances)

Note: there are 2 ways to increase number of instances

1. Increase number of classes
2. To increase number of instances without increases number of classes, I must substitute old classes with a new classes that contains more classes.

Number of all the instances

{geneId: 11657, geneSymbol: 11657, diseaseId: 11657, diseaseName: 11657, diseaseClass: 11657,pmid: 11657, source: 11657, class: 11657}

Number of uniq value in each classes

{geneId: 2429, geneSymbol: 2429, diseaseId: 61, disesaeName: 61, disesaeClass: 19, pmid: 6973, source: 11, class: 9}

Difficulties

I manually select labels by referring to multiple copd commorbities papers and search for the most relevant copd comorbidities that are gender and age adjusted.

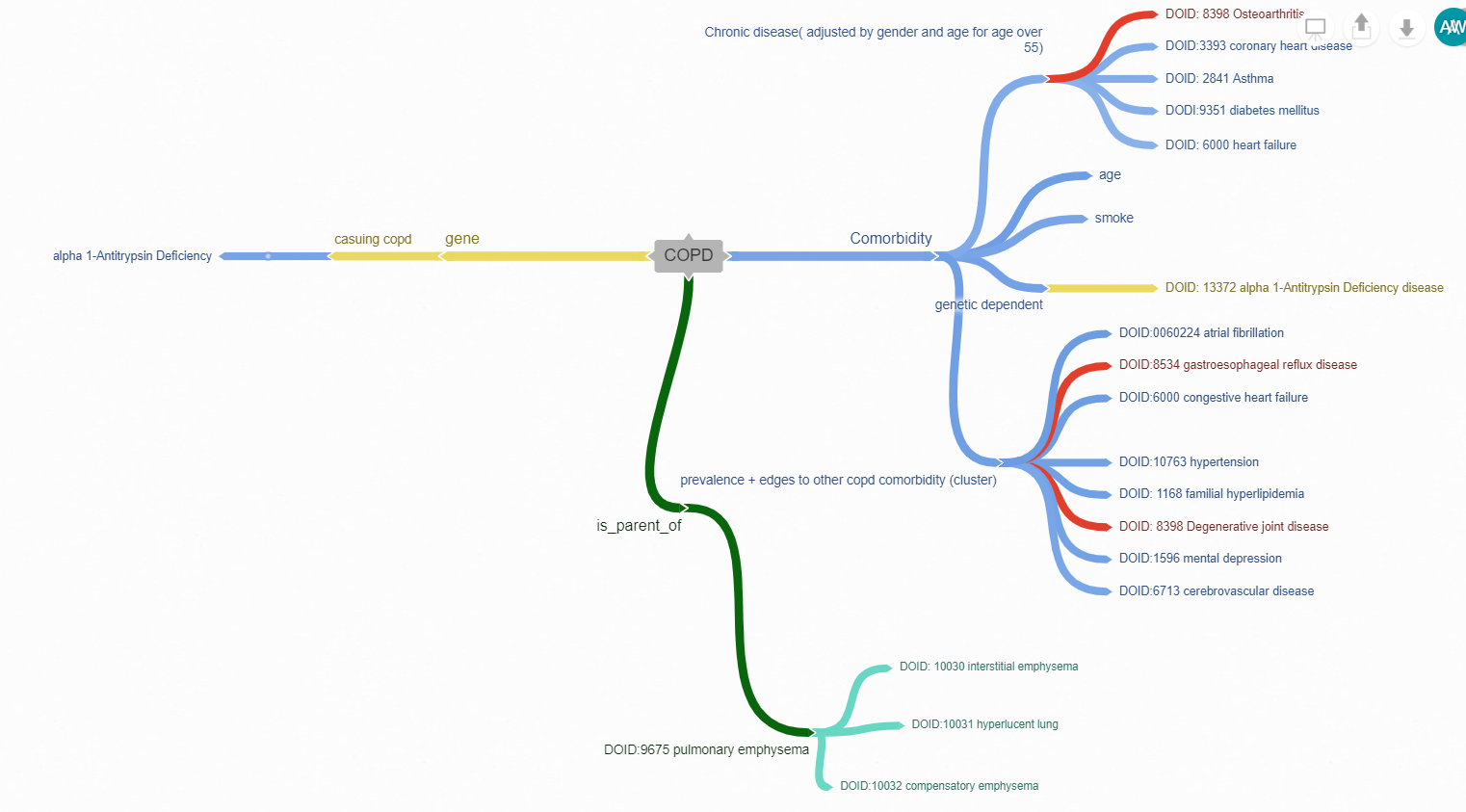
However, copd is commonly found in smoker and older population, so age adjusted that I could find is average of 40 and 55. (on 2 different papers)

I must use copd commorbidites of smoker because 90 + percent of copd patients are older and I cannot find copd comobidity statistic that is specific to younger population who are most likely to be non smoker (due to his/her condition). I do realise that this is more beneficial to our gene-disease analysis case, but no useful infomation can be found.

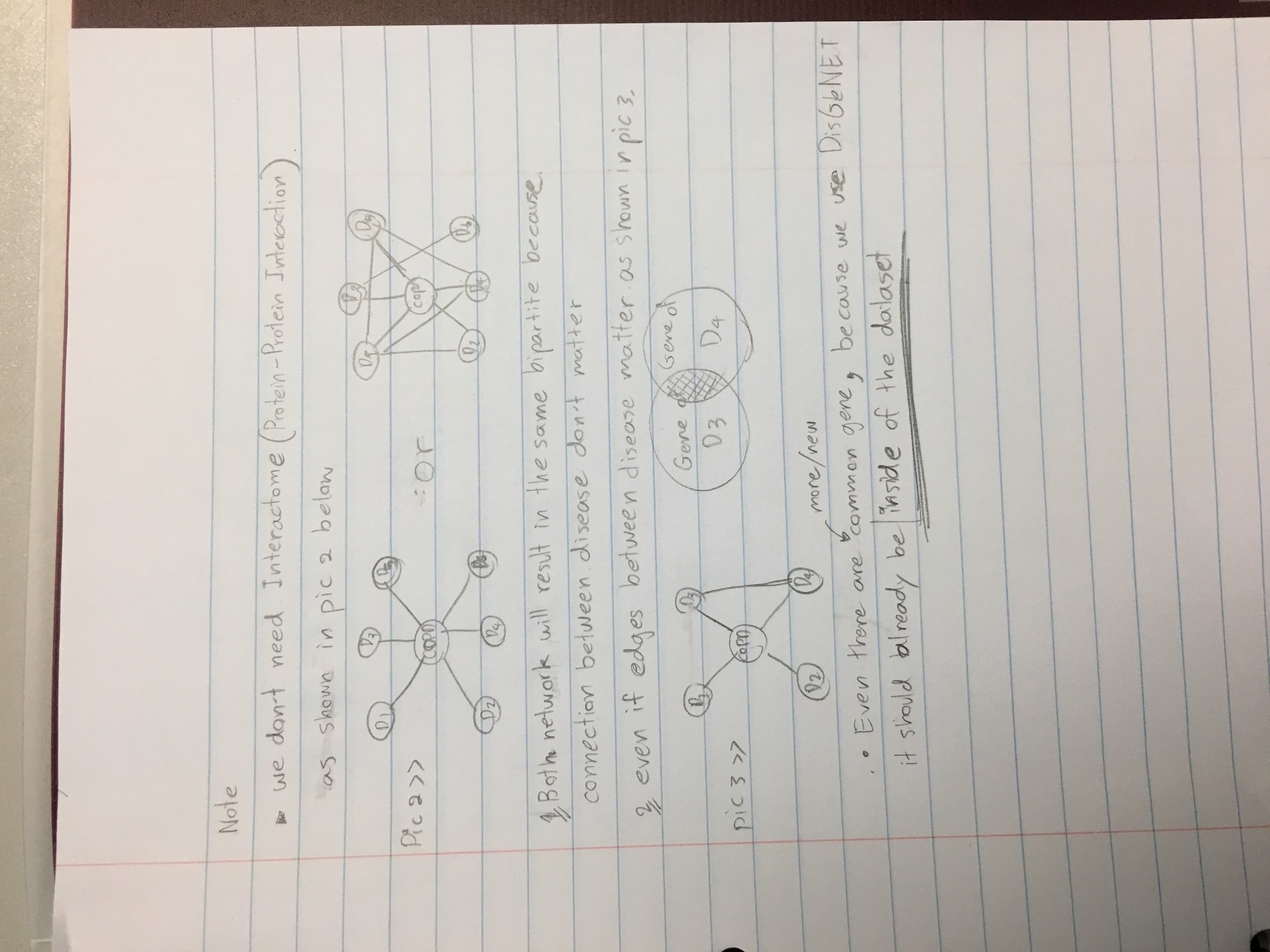
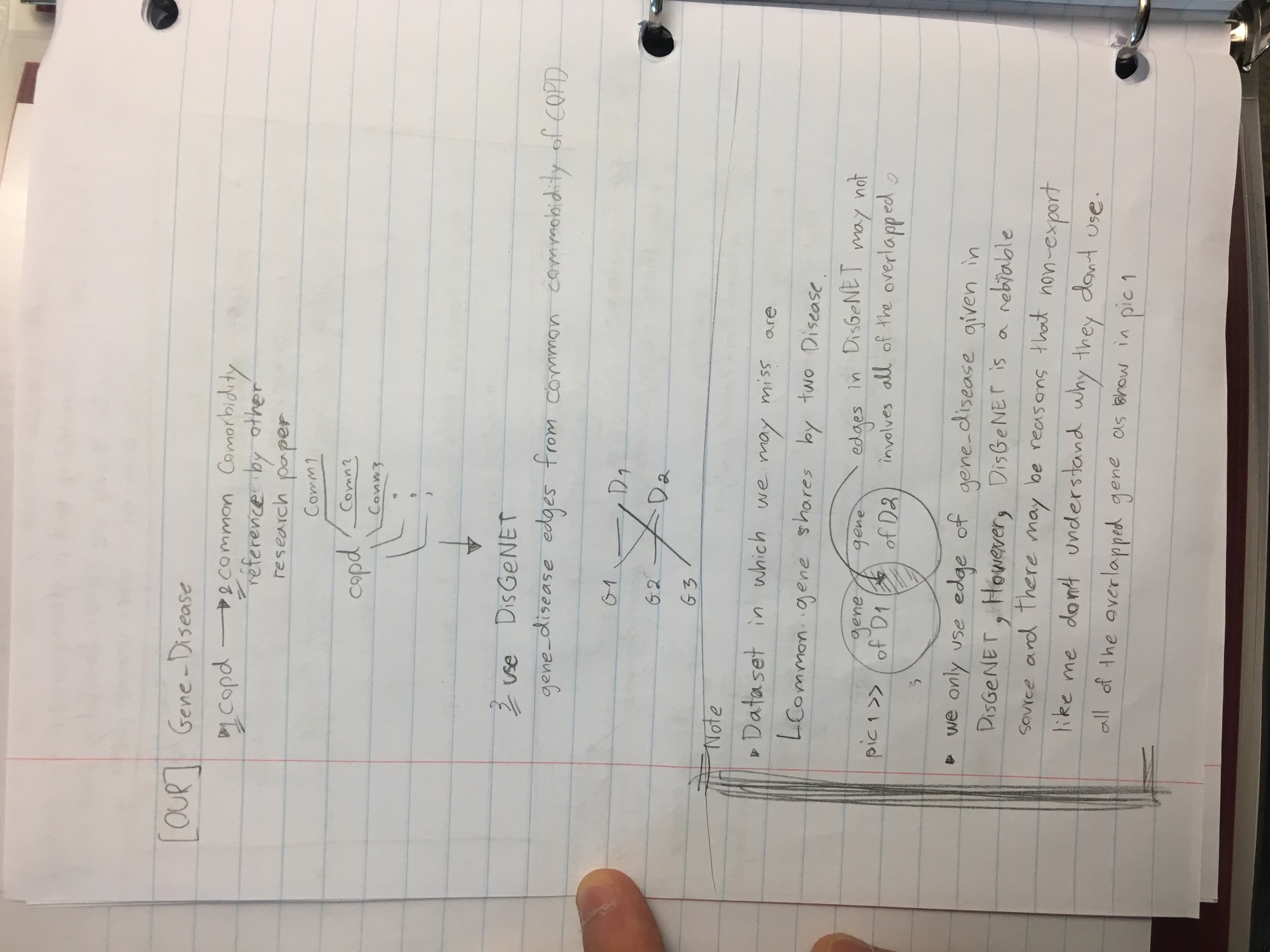
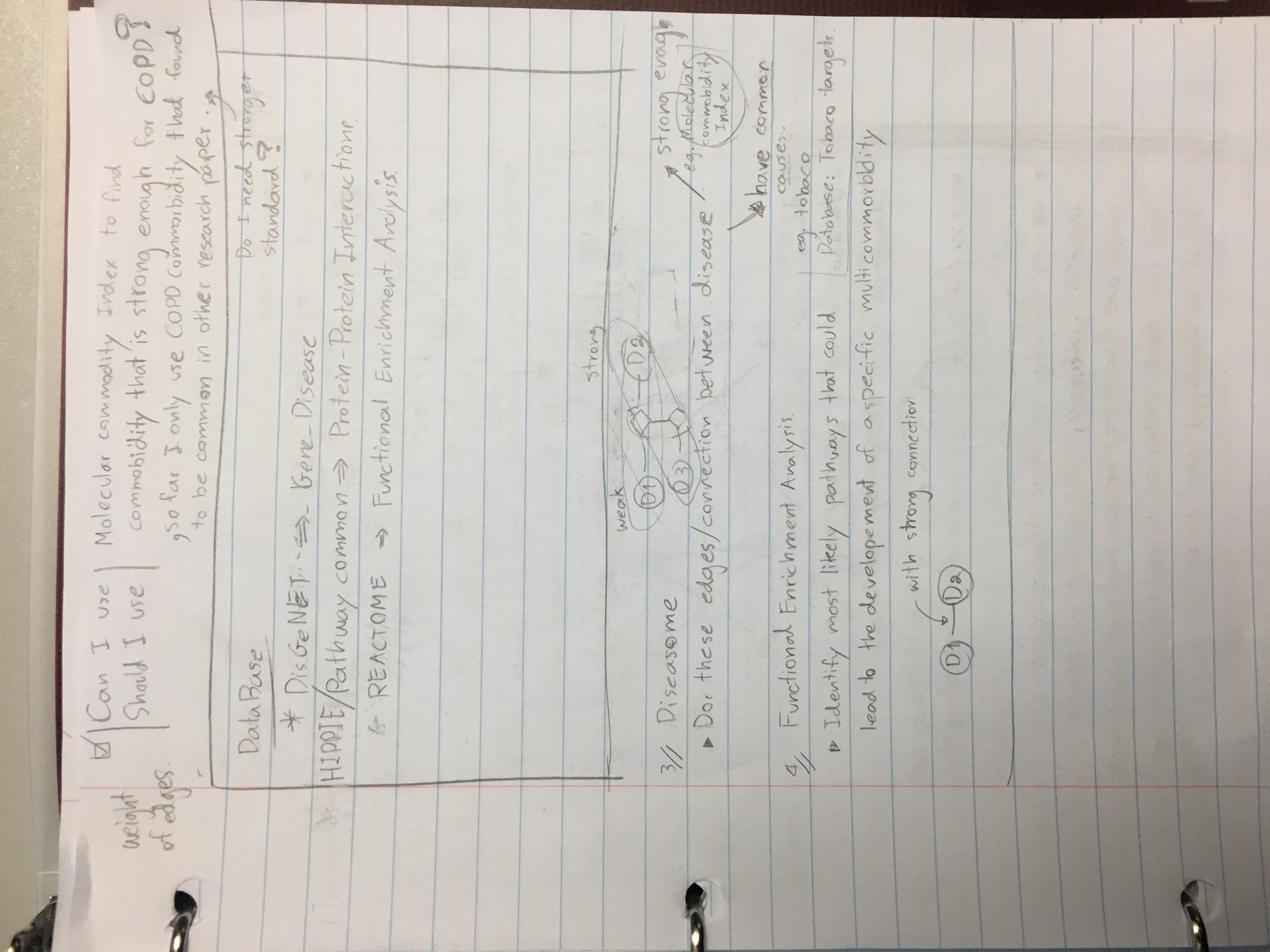
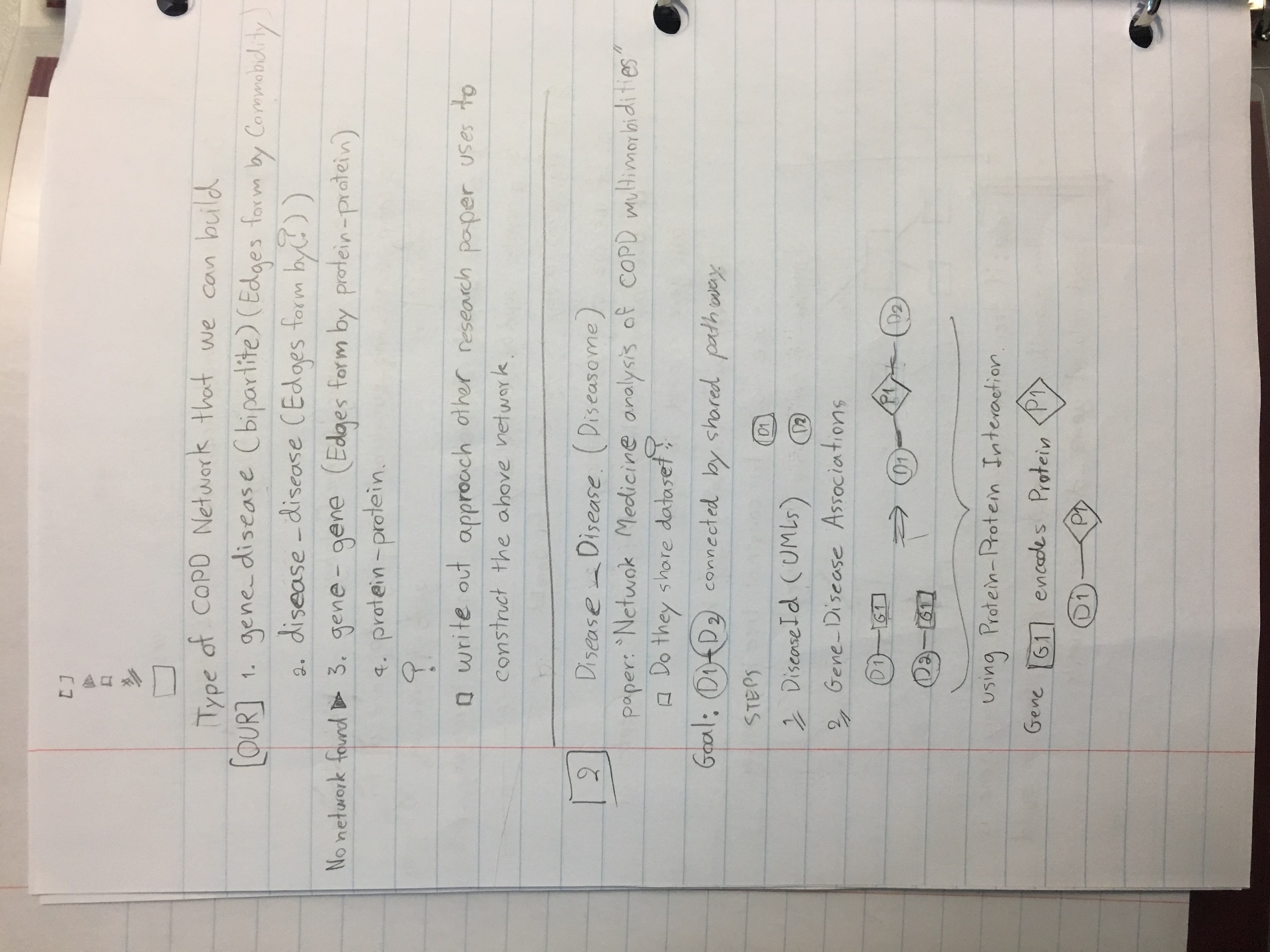
Though, I did find that there is 1 copd gene which is a cause of a cause of "DOID: 13372" or doid of copd, but it is not includes in any of the classes given above, and its doid is a leaf. (no children, cannot be used as label)

for gene, I will link diseaseId to its corresponding gene in disease\_gene.tsv file and then label gene with gene Ontology.

In addition, if possible I will include "alpha 1-Antitrypsin Deficiency" which is a direct cause of copd. However, my concern is that there may be no edges connect to any other nodes in disease because as i mention above doid of copd that is cause my "alpha 1-Antitrypsin Deficiency" is not included in any of the selected classes.



Summarize and comparing our approach to other approaches use in other research papers



**6/17/2019**

* Created unique node and edges in excel
* Created html website to describe our dataset.
* Implemented Graph AttentionWalk with pytorch using chameleon dataset
* Implemented Deep Walk with pytorch using Cora dataset
* Implemented GraphConvolutional Neural Network with pytorch Cora dataset
* Implementing Bipartite Network Embedding (half a day of work)
* Plan to implement GraphSage (half a day of work)
* Experimental structure
  + 80/20 splits on training and test
  + Baseline of embedding algorithm is no embedding
  + Embedding Algorithm that will be used
    - Graph AttentionWalk
    - Node2Vec or DeepWalk
    - BINE
  + Node classification Algorithm
    - GCN
    - GAT (Graph Attention Neural network)
      * How will AttentionWalk and GAT work together?
    - CapsuleGCN (because it claims that capsule can capsulr
      * It cliams that Its “capsule” capsule features of the graph.
      * how will these feature work with feature of embedding Algorithms.?
      * This may require me to manually code the algorithms, because it is not conventional and widely used
    - GraphSage
      * It is a substitution of CapsuleGCN, in case it takes to long to implement.
  + I will use pytorch geometric for algorithms that are available.
  + Visualization
    - Tsne for node classification
    - Loss functions
  + Measurement will include the following
    - Confusion matrix
    - Overall accuracy
  + I am not sure how to split disease and gene nodes.

June 24 2019.

Comments:

1. Change the html file format (simplify):
   1. degree distribution plot (use gene degree distribution plot and disease degree distribution plot).
   2. Draw diagram to show the input dataset (or data sources, such as gene ontology API) so users can understand was the dataset generated
   3. Put nodes in one file (including disease node label).
2. Cross validation (“am not sure how to split disease and gene nodes”)
   1. All genes-diease nodes are kept in the networks. For traingin, split diease nodes as 80% labled, and use them to predict remaining 20% nodes.
3. Change “COPD bipartite graph info” as “COPD Gene-Disease Network”

6/24/2019

1. Make all of the selected model compatible with copd\_label.txt dataset, but 3
   1. CapsGCN
   2. BINE
   3. GraphSage

7/1/2019

* Obtain all embeddings values from BINE, AttentionWalk, and Node2Vec
* Finished building Dataset inherit from pytorch\_geometric.data.Dataset abstract class easily be fed into GCN, GraphSAGE, GAT, and CapsuleGCN.
* (Currently doing) I need to run the files to get result.

7/8/2019

* Add disease and regroup disease under new set of labels to increate “label rate”
  + This helps dealing with class members distribution, and eliminate extreme case where there is only 1 diseases per class in training and testing.
* Visualize embedding to show whether or not dataset can be learnt.
  + It turns out that adding for diseases and regroup disease under new set results in similar embedding space.

**Problem**

* I think labeling gene as non-class is not the best way to label the dataset because clustering visualize from embedded vectors does not follow class label.
  + I am not sure how to relabel it

**Detail description of embedding visualization**

From the labels, there are 0-8 are disease classes, and 61 is unlabel class of genes.

I have a feeling that I should label gene and disease differently. However, I couldn't currently think of other strategy.

There are 60 diseases with 9 classes 0-8.

There are no cluster among disease classes in any embedding technique. Though, there are cluster of genes in node2vec and AttentionWalk.

From the cluster, I noticed that there are 7-9 obvious cluster of genes. I have to confirm further what are the relationship between these cluster and diseases in which there exists edges to the genes within the cluster.

**What I am currently doing**

1. figure out ways to deal with low "label rate" (label:unlabel ratio)

       > so far I have notice that dataset with low label rates have

* higher number of nodes features such as NILL
* more members per class ratio such as PubMed. (**my current approach**)

1. figure out ways to labels.

       > for example, edges in NILL is labels, so the author created hot one vector between nodes and its edge relation to add more features. Though, our each has no labels.

=============================================================================

Type of emb

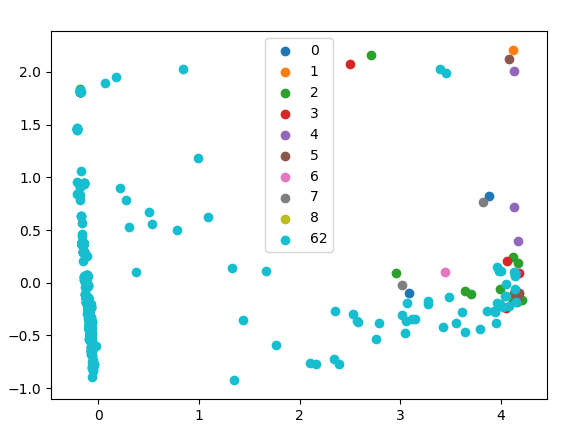
1. Bine
2. Attention\_Walk
3. Node2vec

**Old dataset with class= 0-8 and 61**

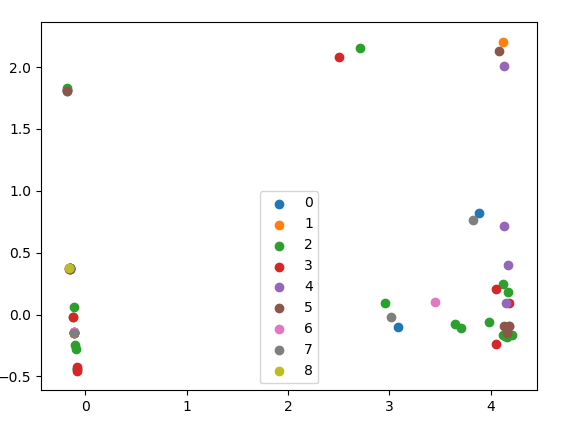
**Attention\_Walk**

1. PCA

* Plot with genes

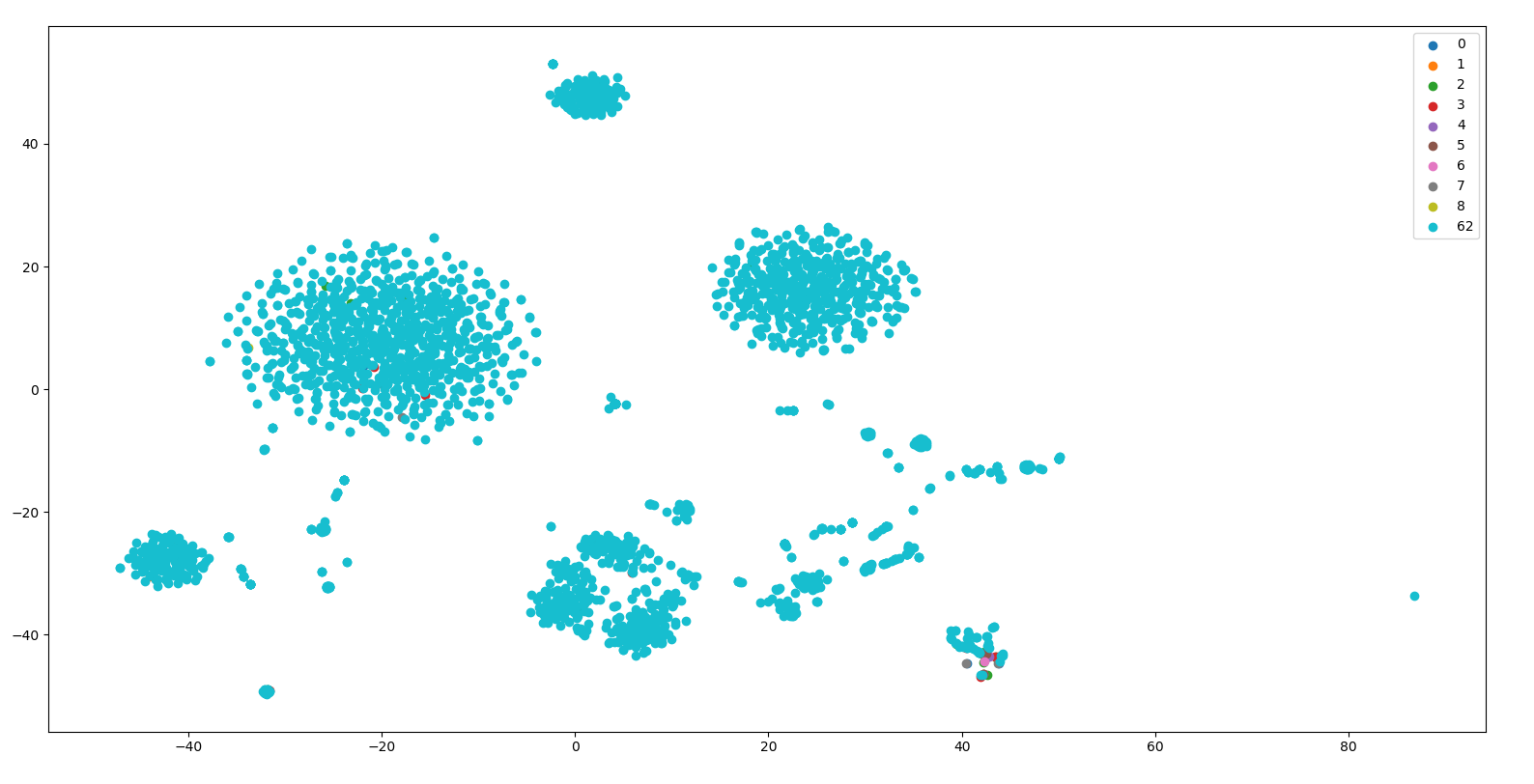


* Plot without genes

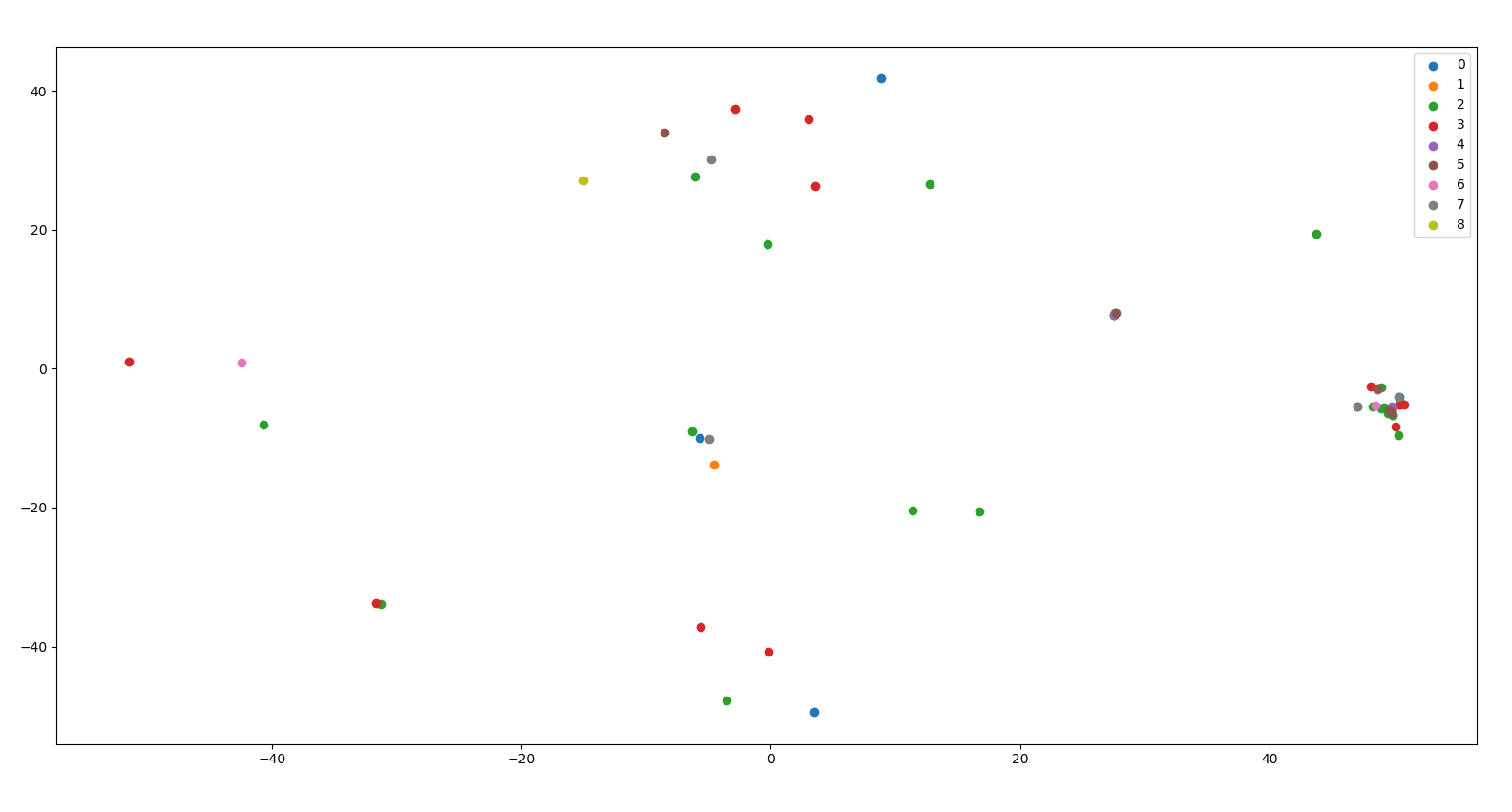


1. TSNE

* Plot with genes
  + There are clustered of gene surrounded disease labels

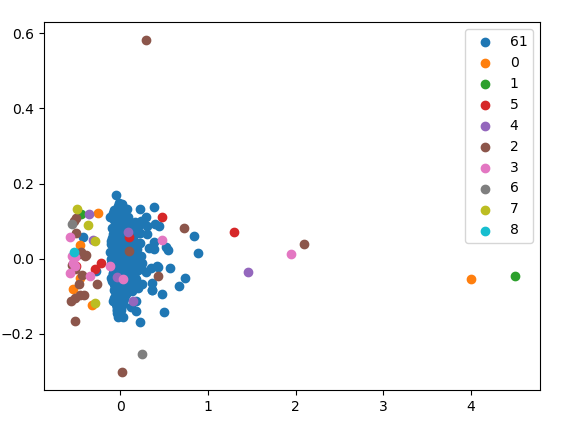


* Plot without genes

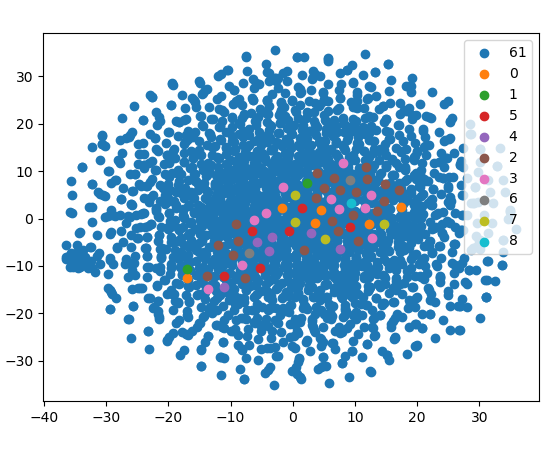


**BINE**

1. PCA
   1. Random no pattern

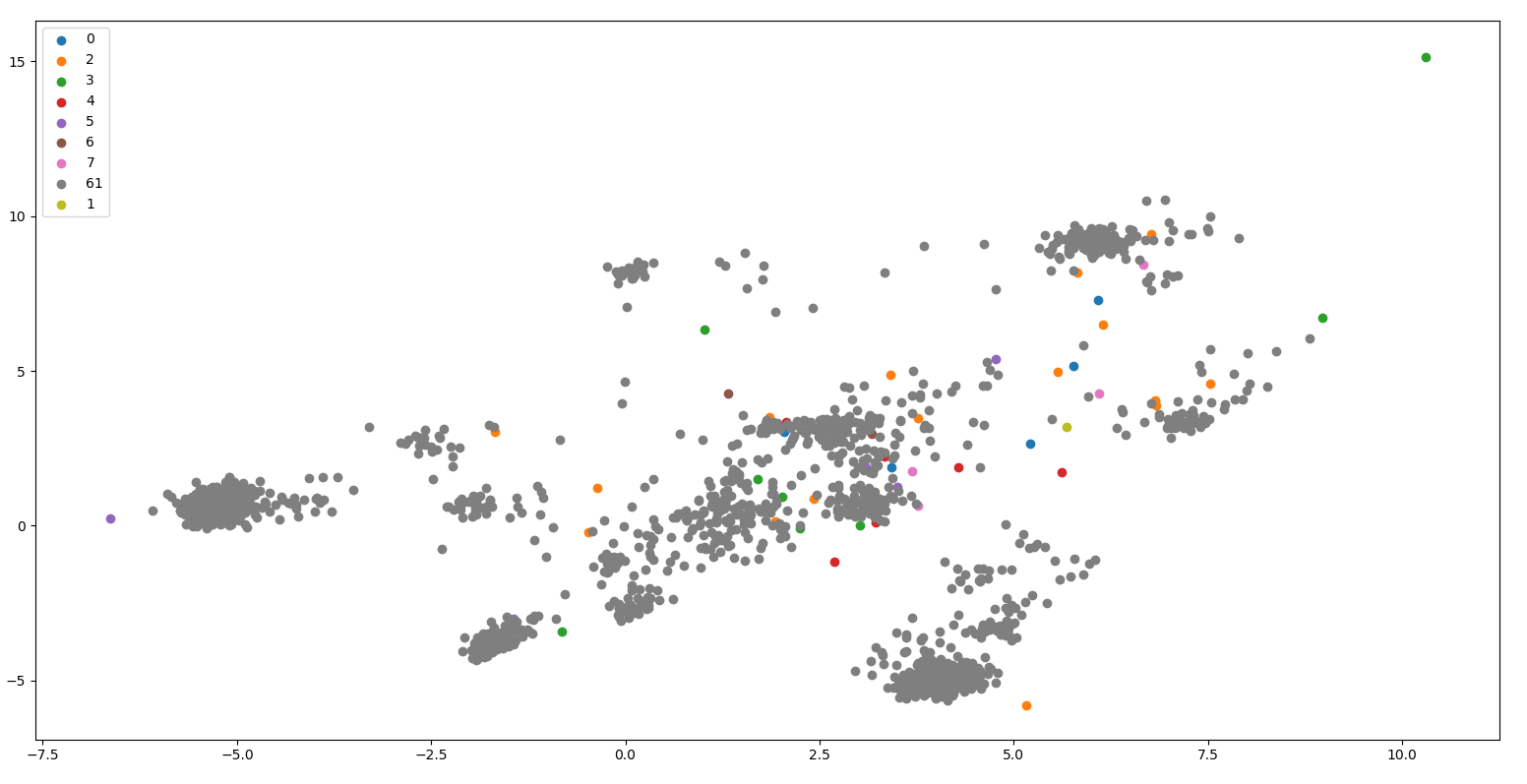


1. TSNE
   1. Random, no pattern.

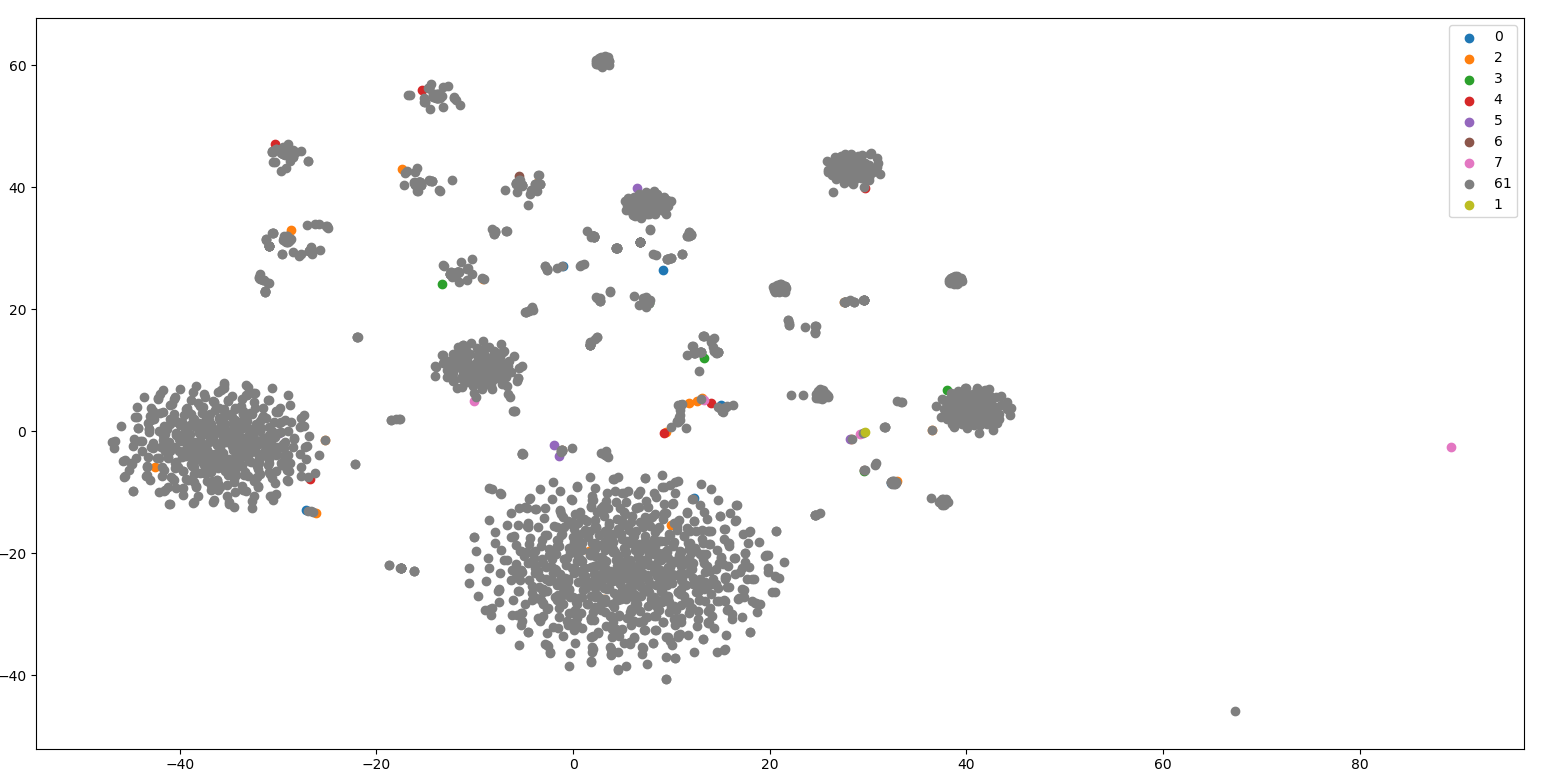


**Node2vec**

1. PCA



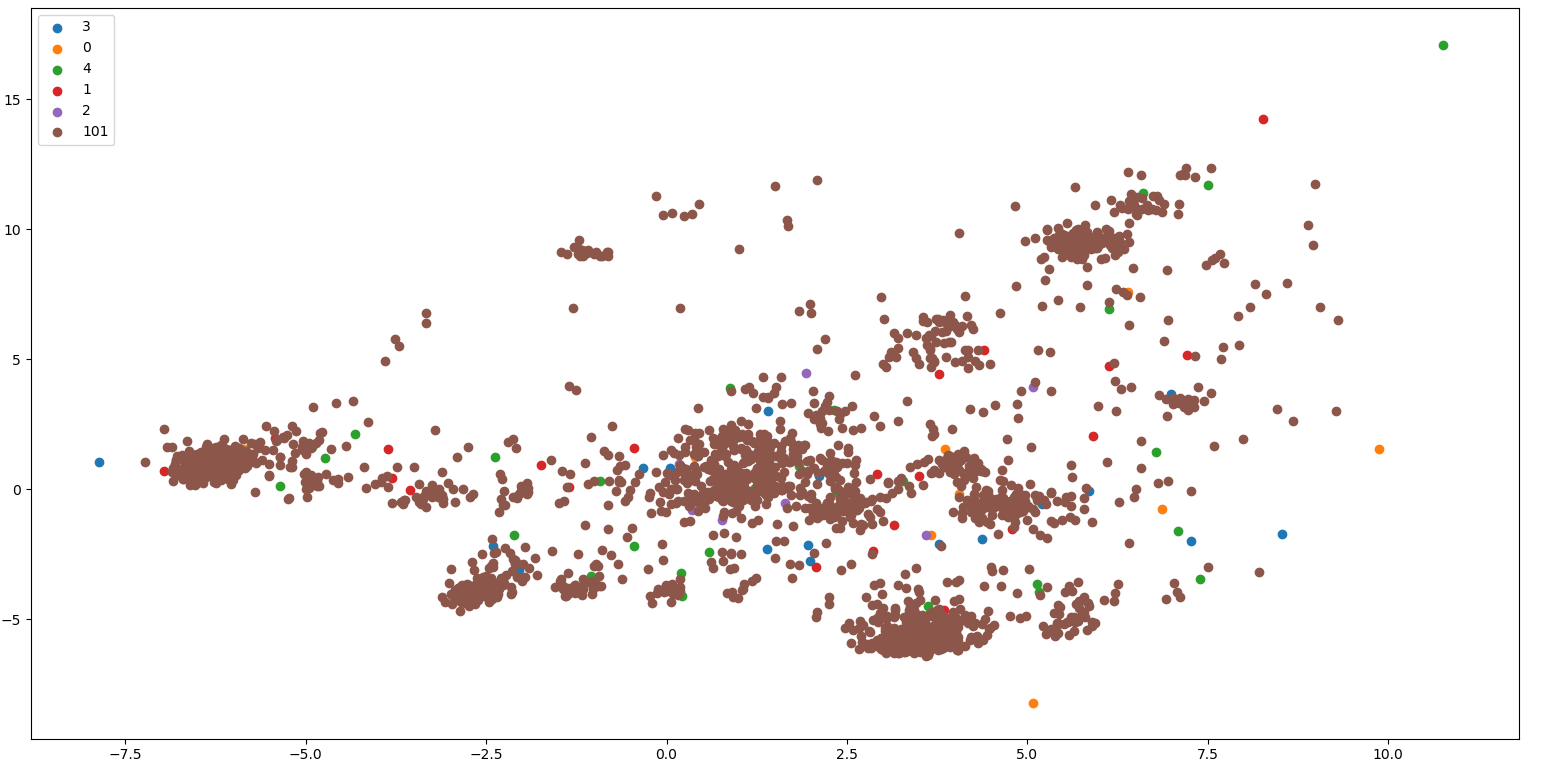
1. TSNE



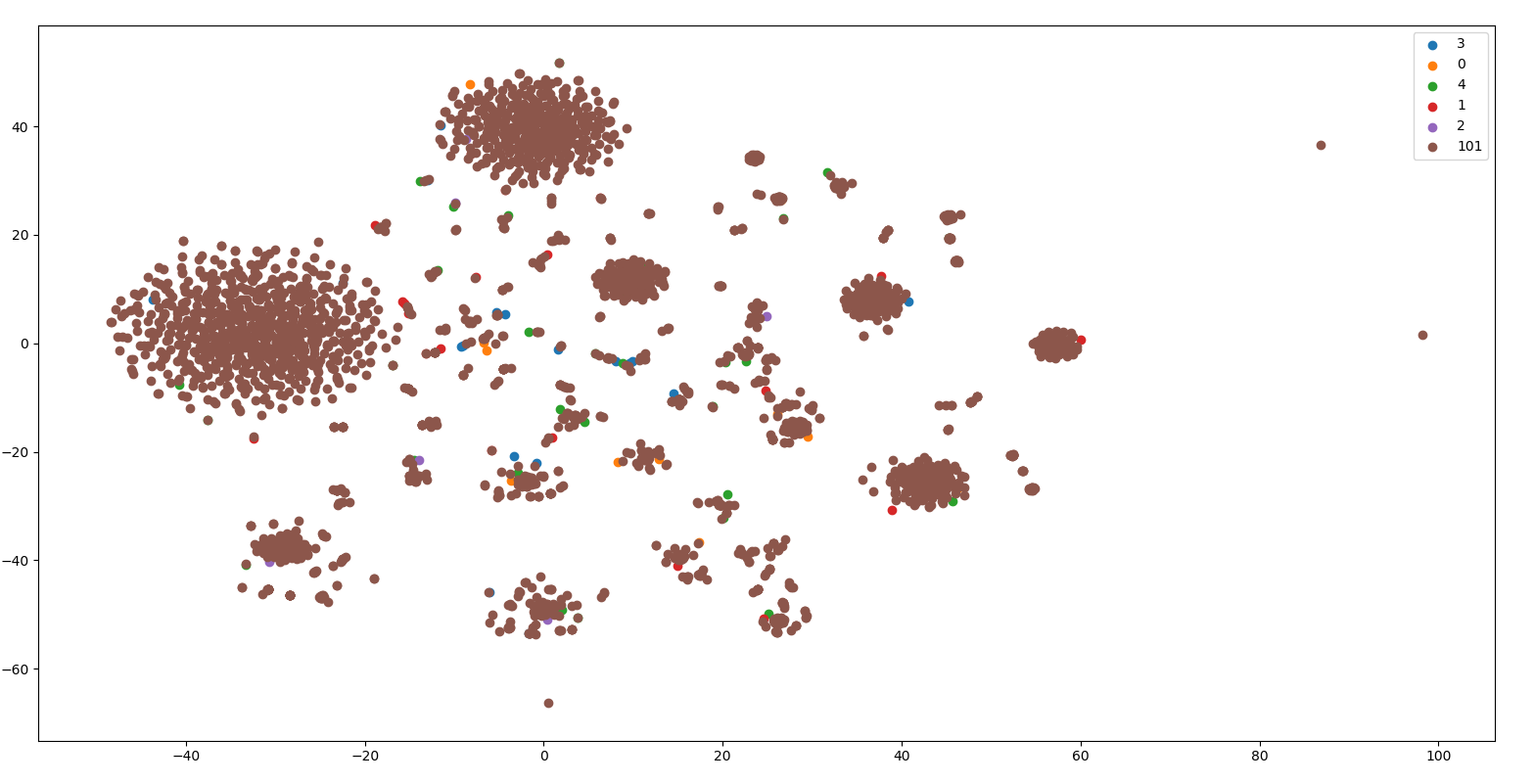
**New dataset with class = 0-4 and 101 where 101 is unlabel class of gene**

**Node2Vec**

1. PCA



1. TSNE



Dr Zhu comments

Hi, Anak,

I felt that 60 diseases are rather too few for cross validation. However, if we cannot expand the number of diease nodes in a short amount of time, I think you can move ahead **to build a classificaiotn model to check**

1.       **The classification of diease nodes using embedding features** (using cross validation), using a network including both diease and gene nodes.

2**.       The above comparisons can be made by using different type of embedding features, node2vec, bine, and Graphical convolutional neural network.**

Once we have the results, we can come back to check whether we should expand diease nodes, or do something else.

You can use your 0-4 (101) dataset.

7/15/2019