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. get distribution of class and its disease members

       > once I understand how relationship of these gene cluster, I may compare distribution of these gene cluster before and after running node embedding.

1. **(HERE)** run the same analysis on the new dataset

    > new dataset is expected to have higher "label rate"

* Run node2vec, attentionWalk and bine. (in order)

1. Create Copd\_dataset() inherite from torch.data.Data.
2. 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 or more members per class ratio such as PubMed.

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

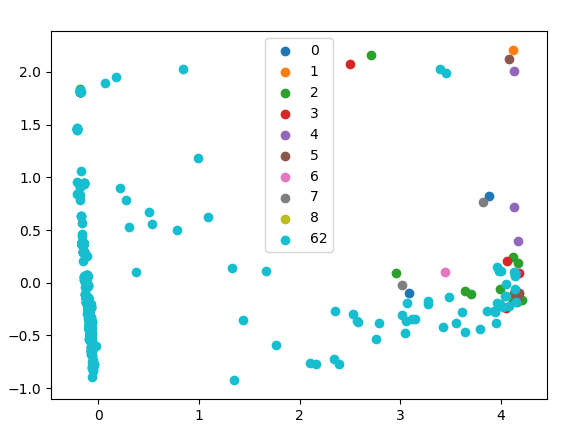
1. Bine
2. Attention\_Walk
3. Node2vec

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

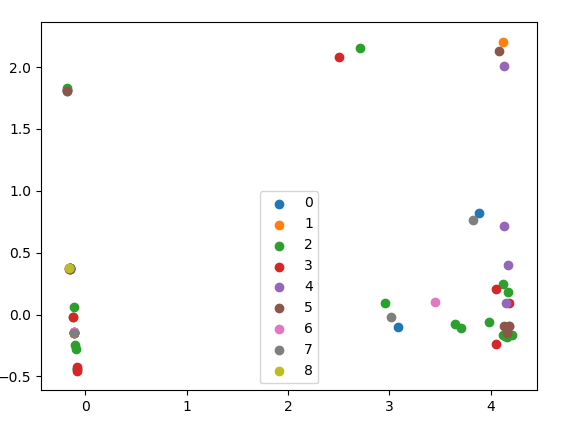
**Attention\_Walk**

1. PCA

* Plot with genes

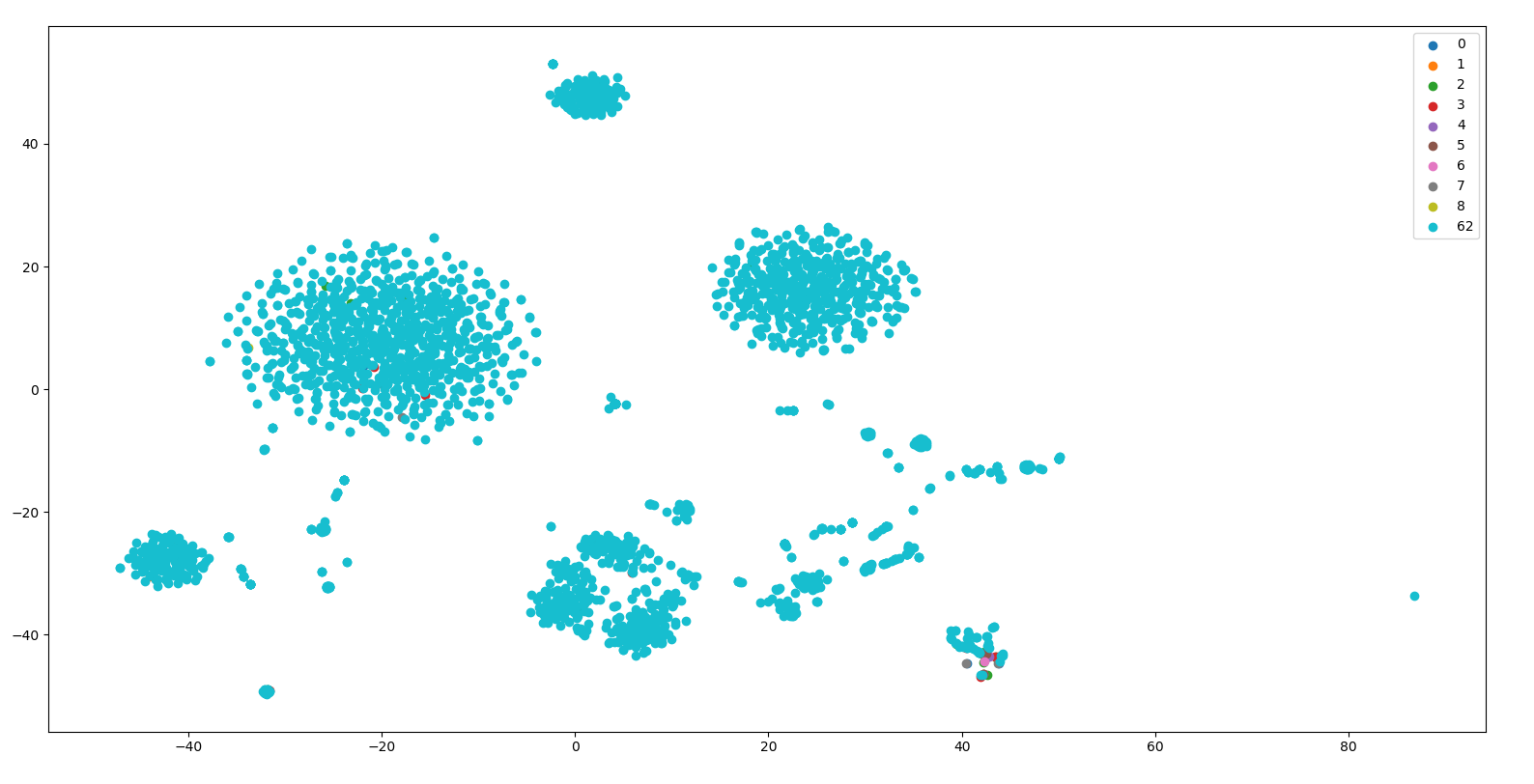


* Plot without genes

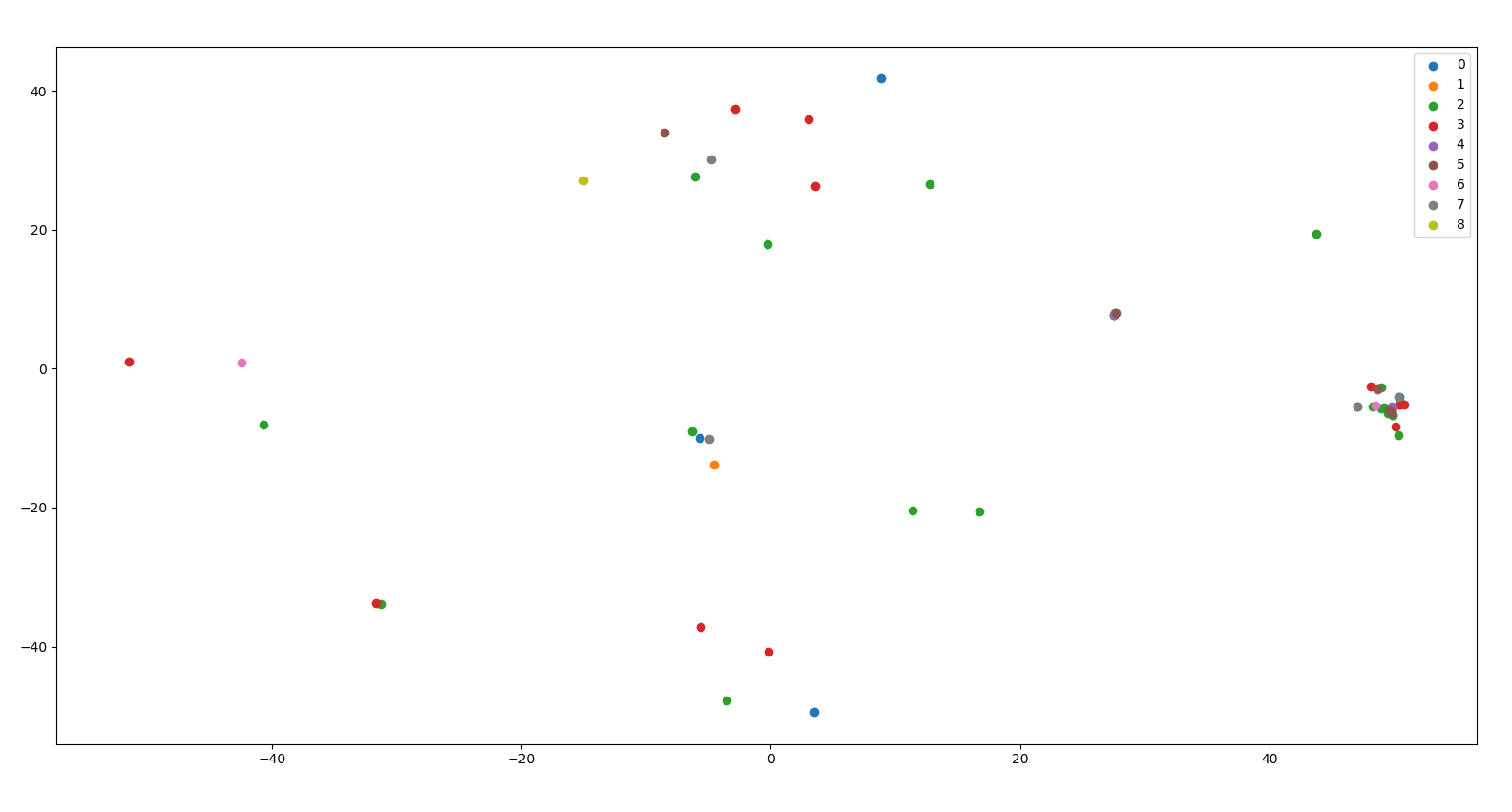


1. TSNE

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

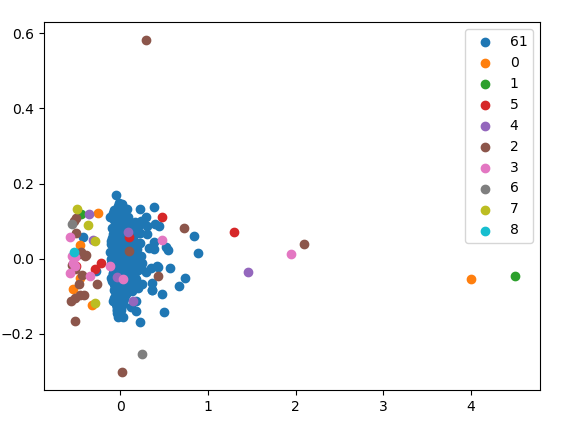


* Plot without genes

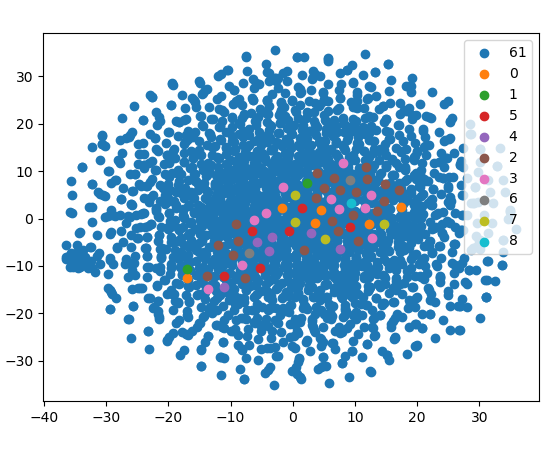


**BINE**

1. PCA
   1. Random no pattern

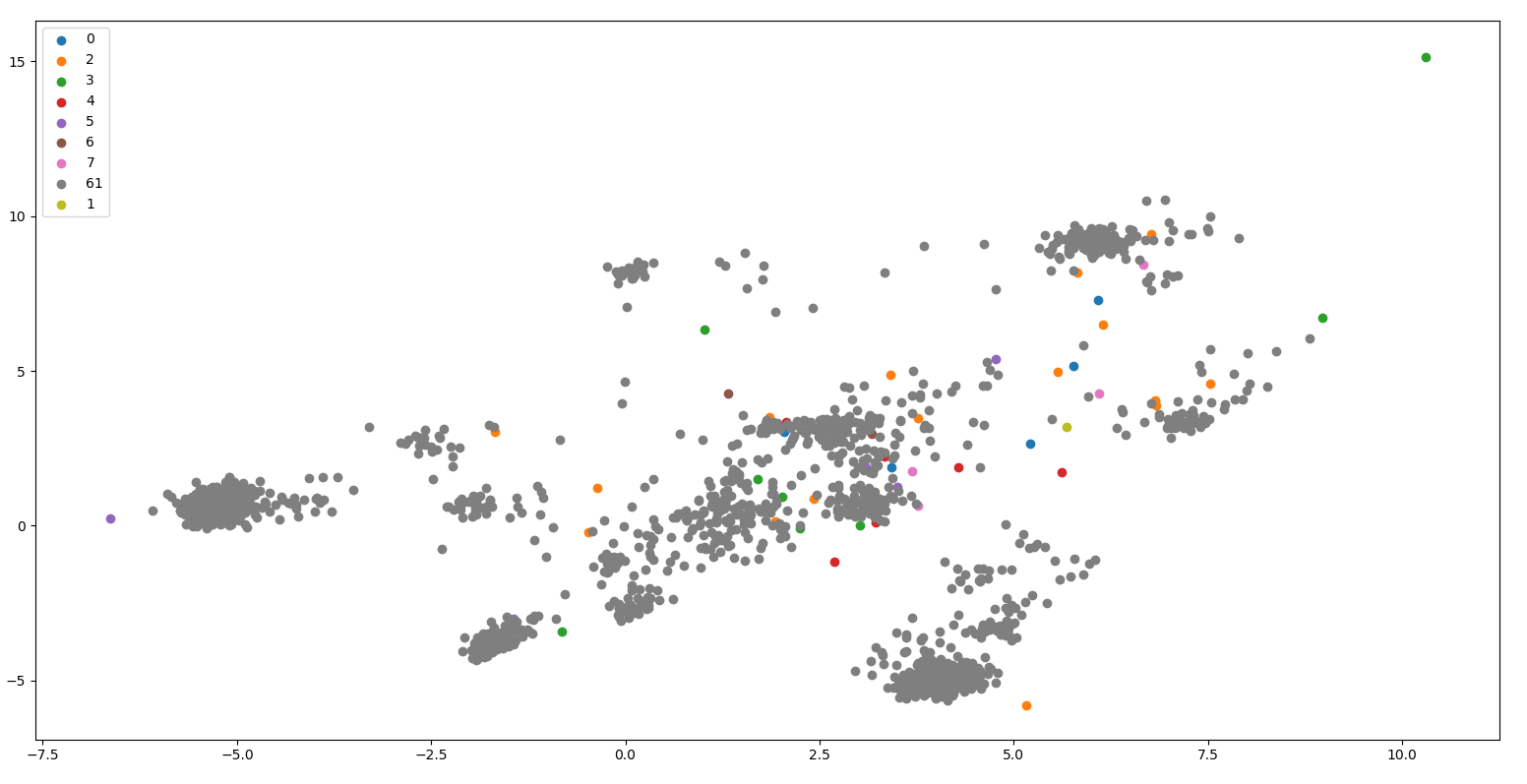


1. TSNE
   1. Random, no pattern.

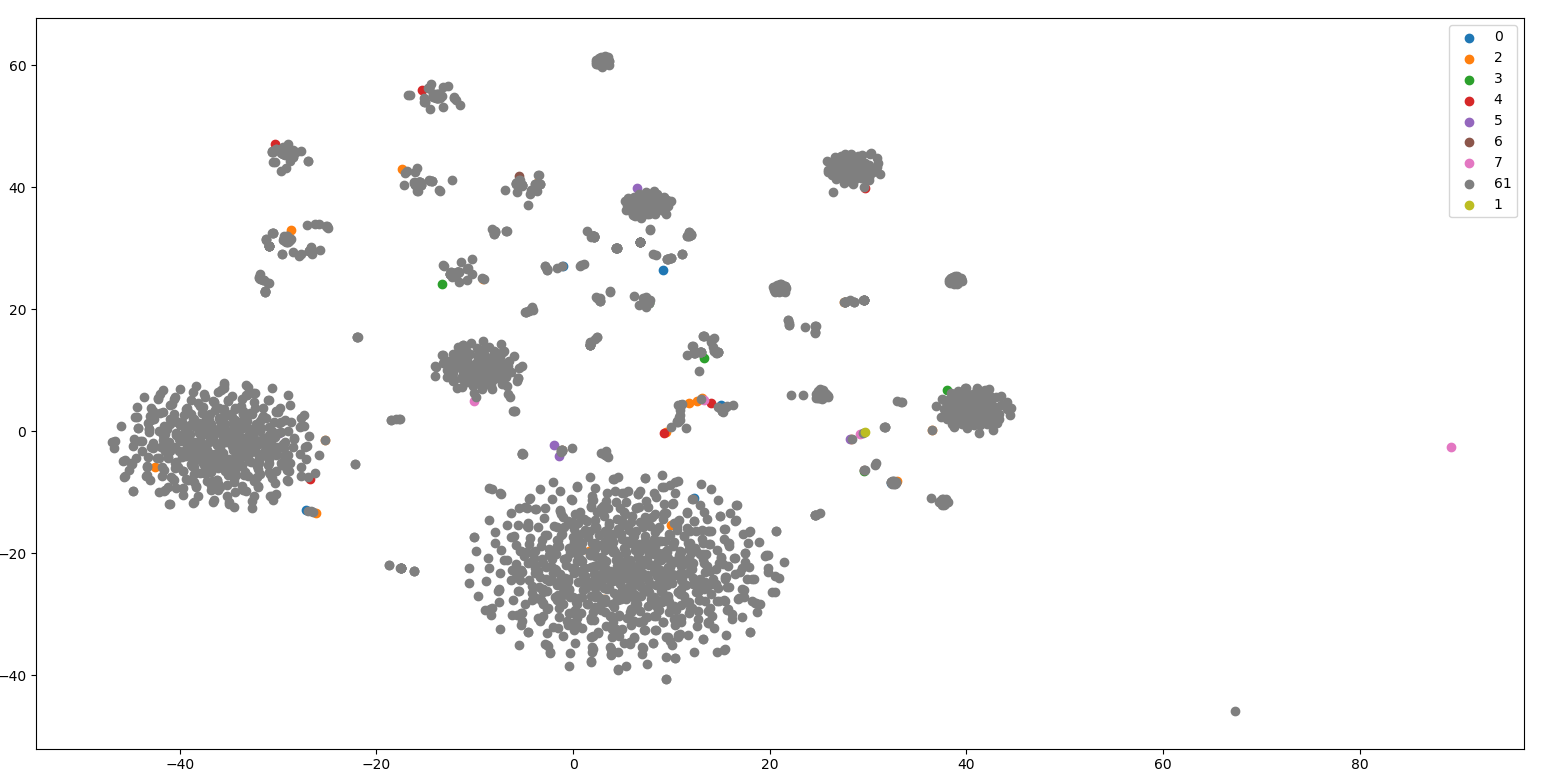


**Node2vec**

1. PCA



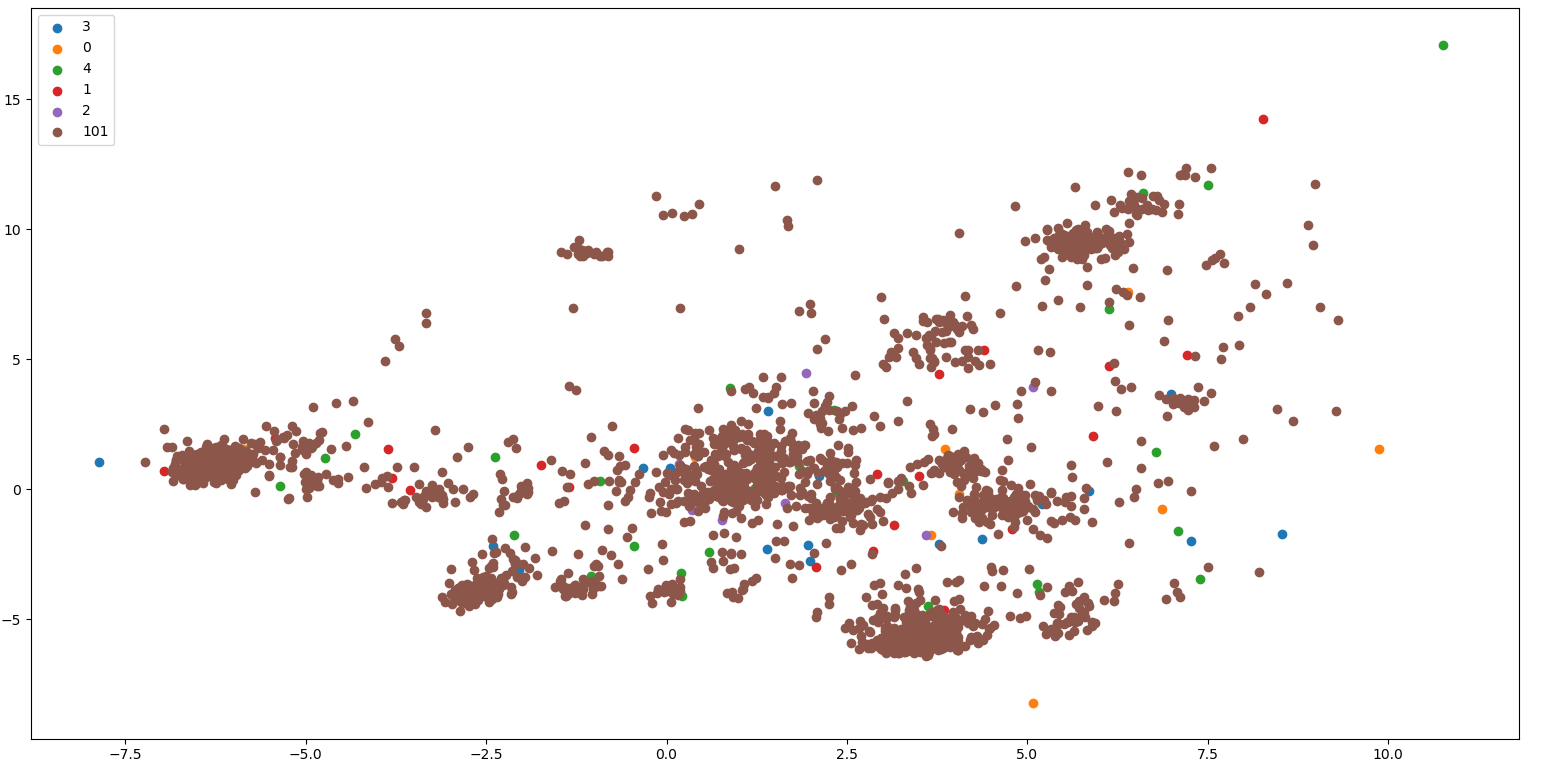
1. TSNE



**New dataset with class = 0-4 and 101**

**Node2Vec**

1. PCA



1. TSNE

