Diseases and Genes

Sky Martin and Kiersten Johnson

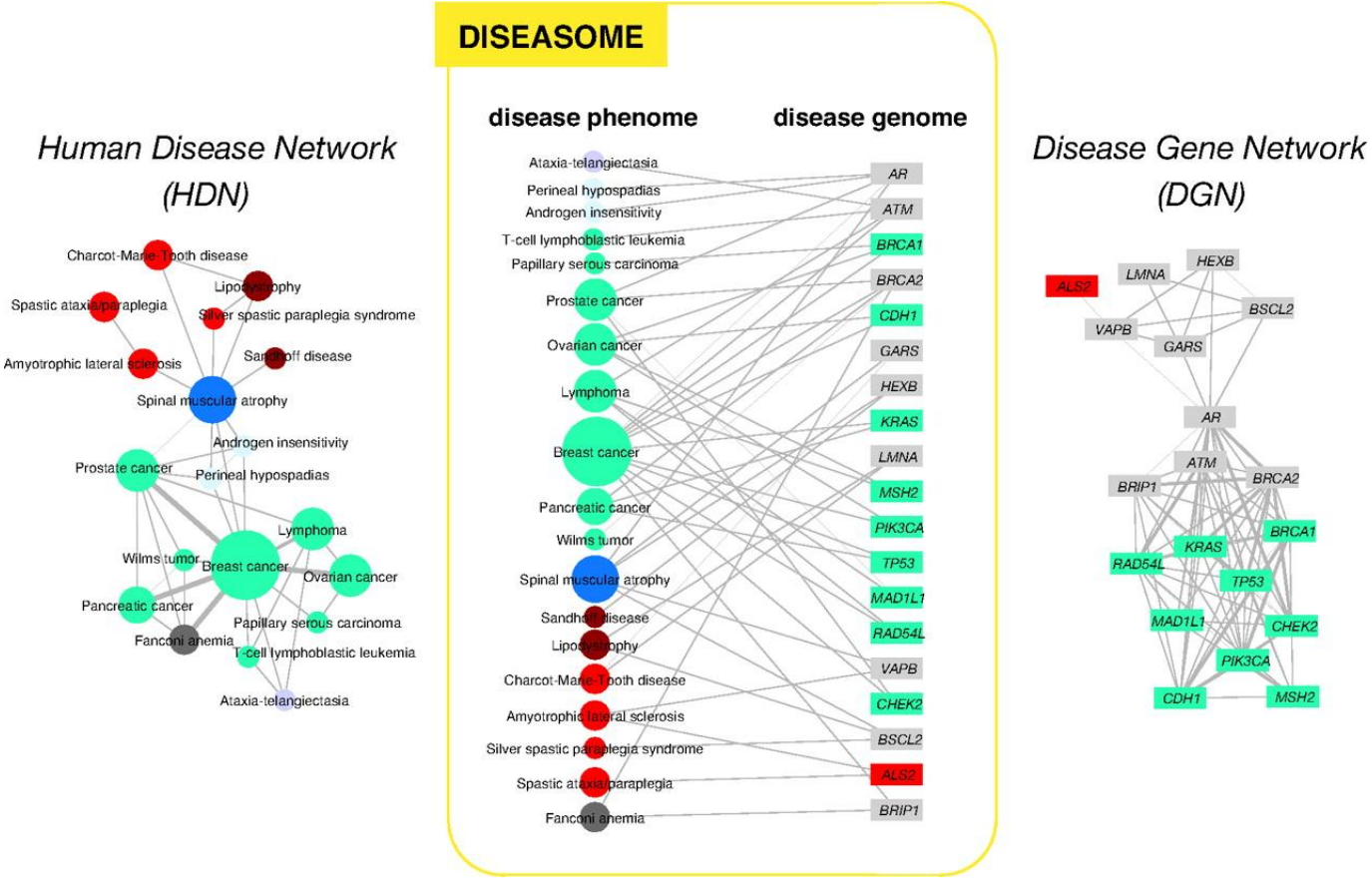
Online Mendelian Inheritance in Man

The Human Gene Disease Network

2007

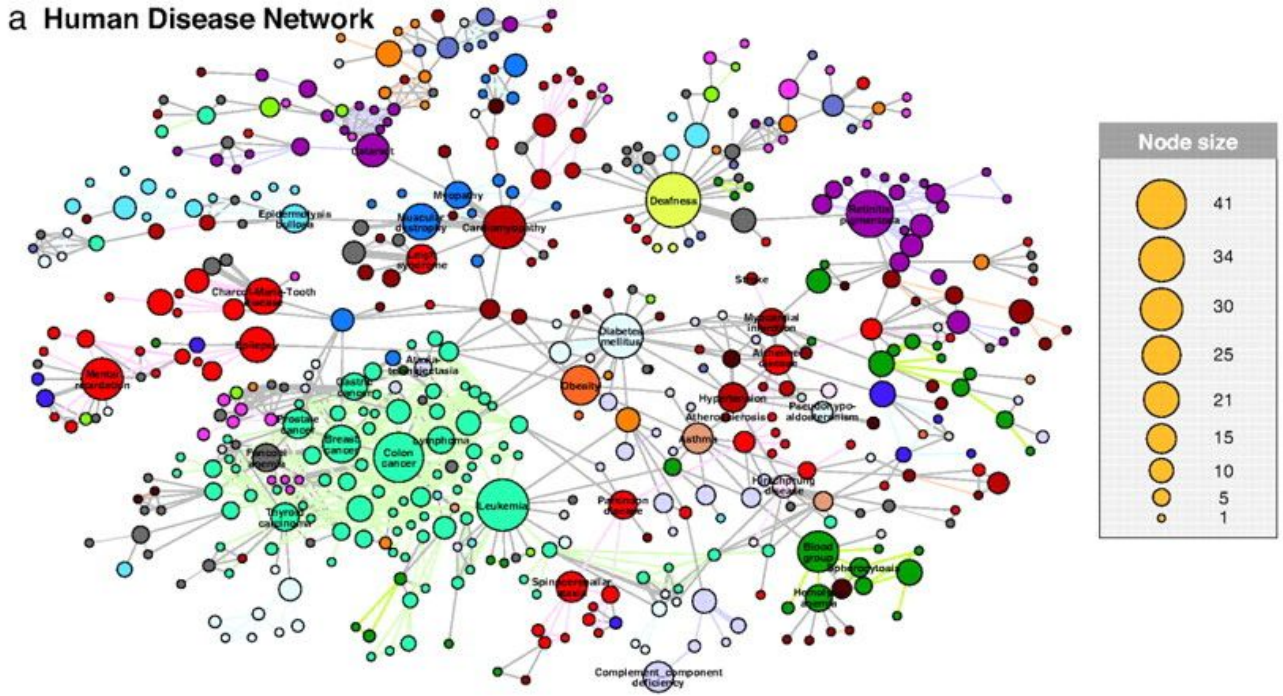
Hand labeled subset of OMIM

Diseases: 1,284
Genes: 1,777



Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási

Human Disease Network (HDN)

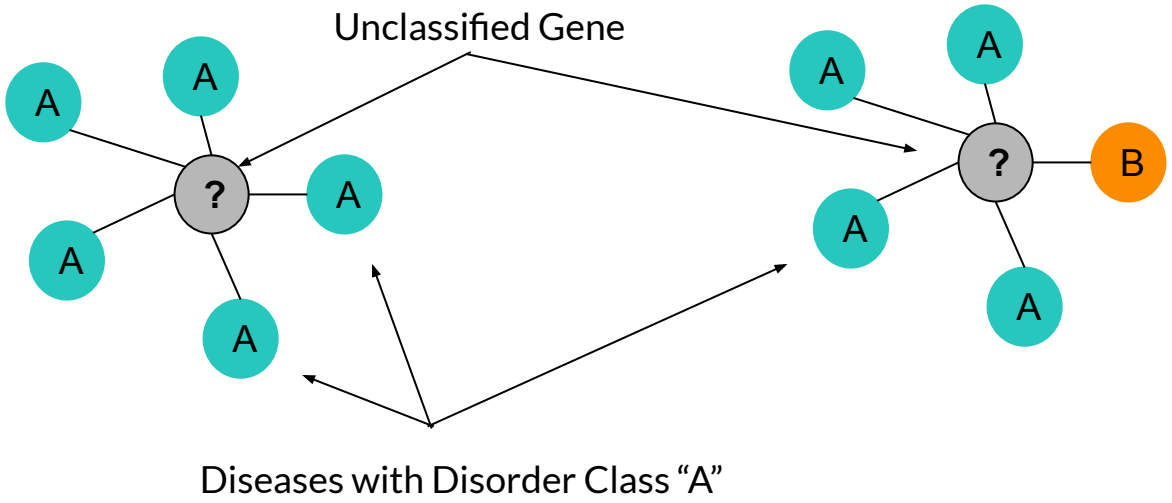


Node: Disease

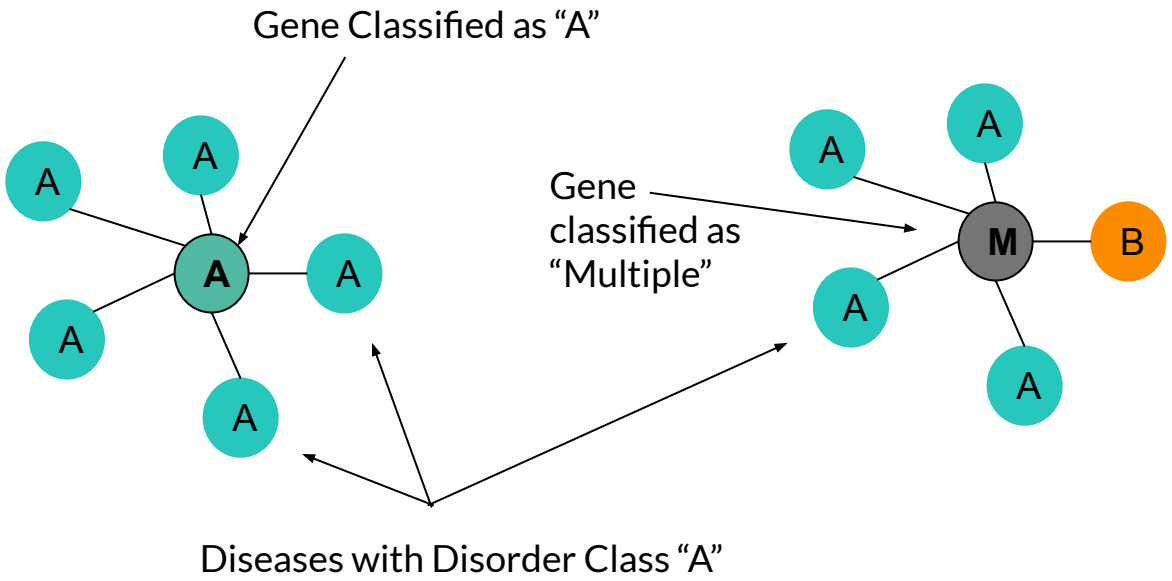
Edge: Shared gene

https://www.pnas.org/content/104/21/8685/tab-figures-data

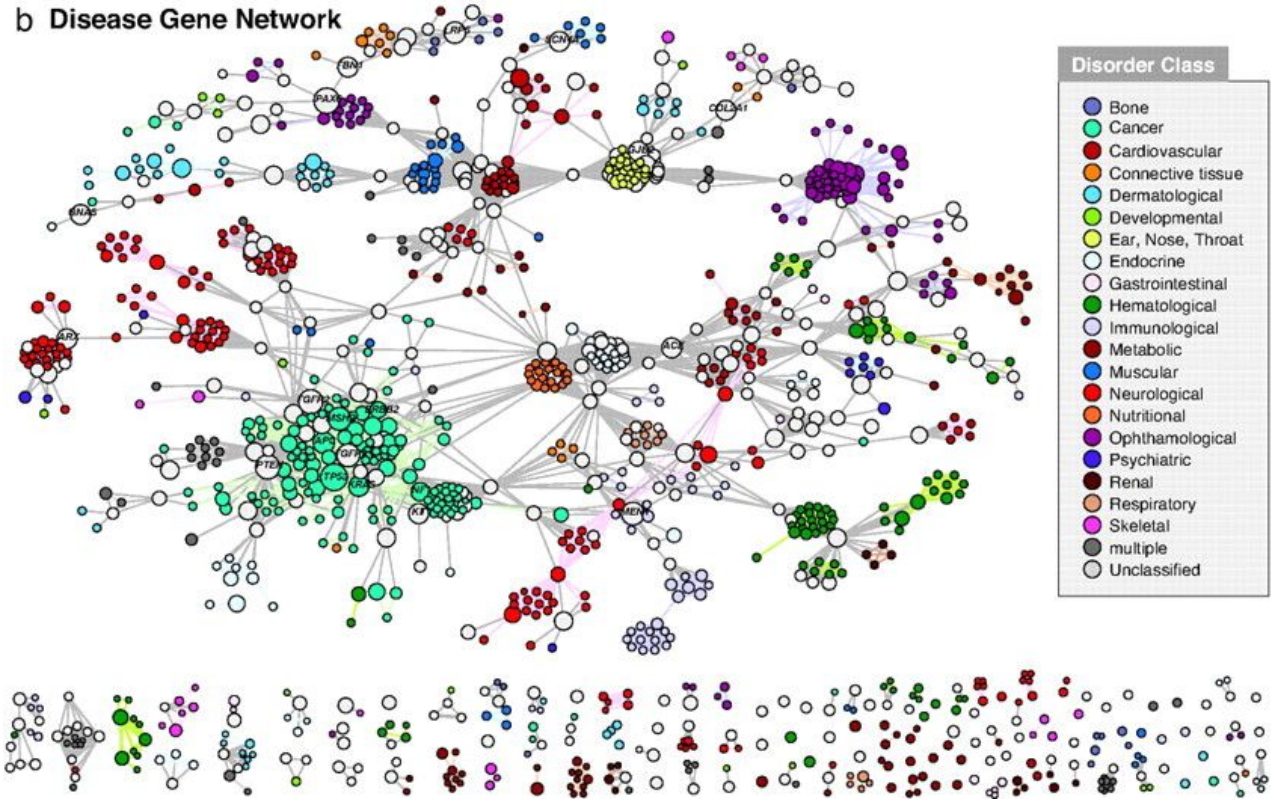
Disorder Class Assignment to Genes



Disorder Class Assignment to Genes



Disease Gene Network (DGN)



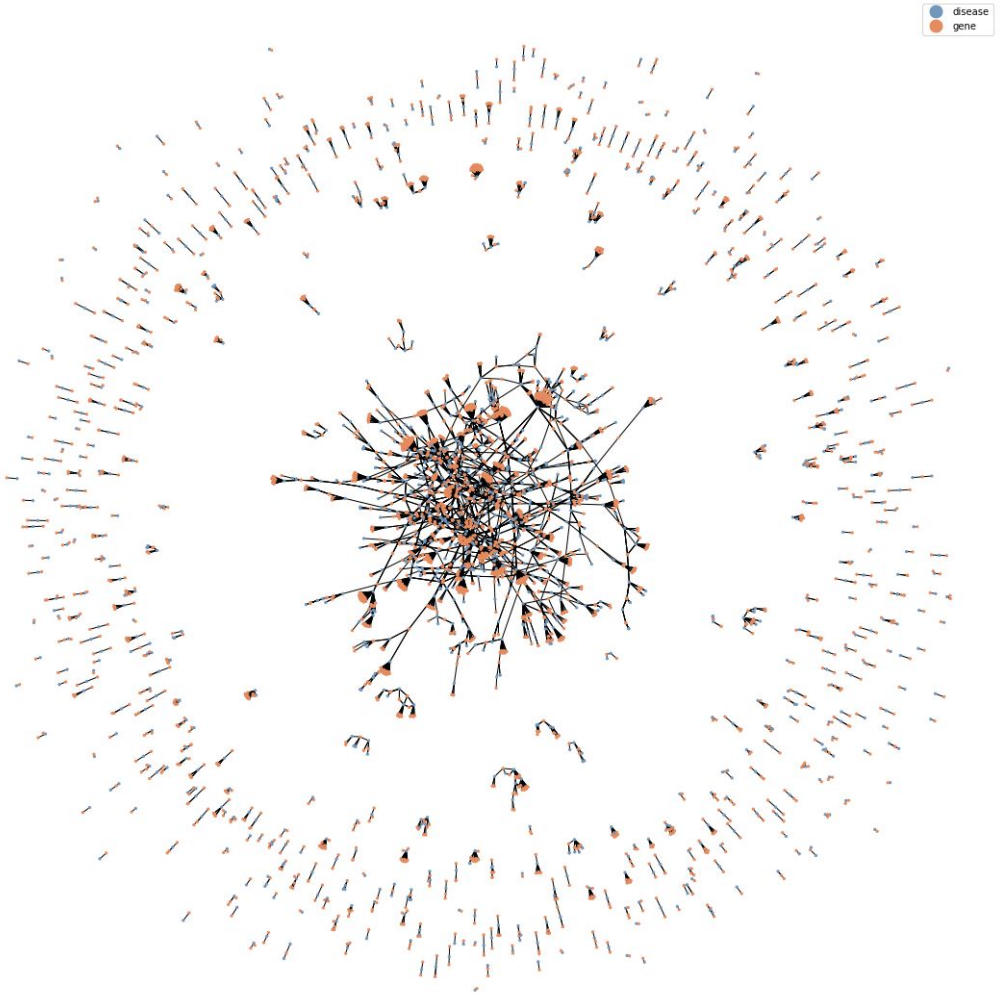
Node: Genes

Edge: Shared disease

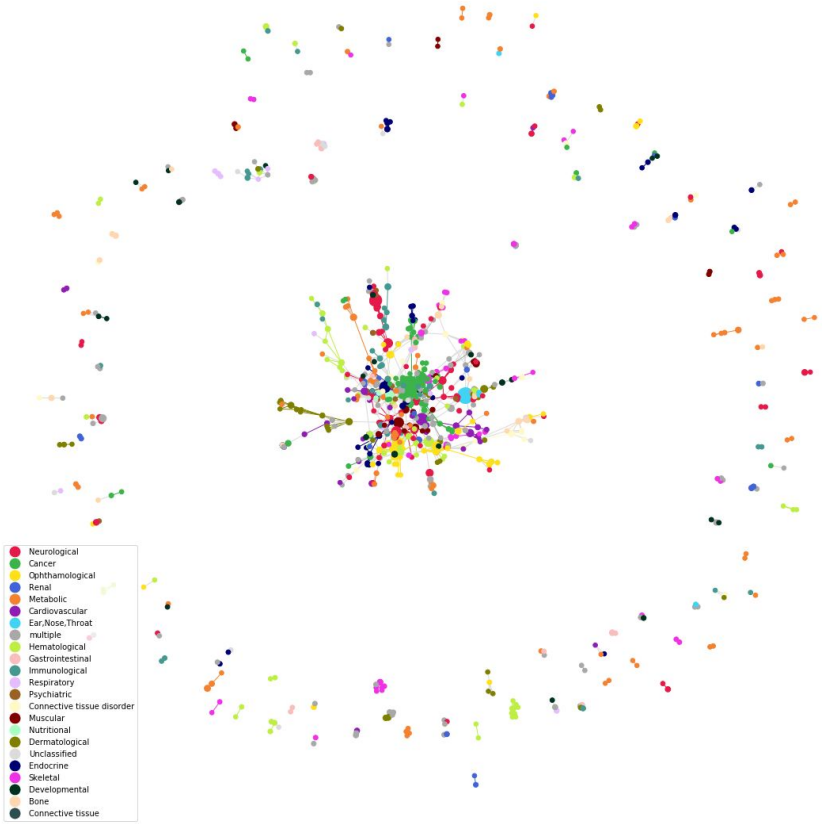
Our Visualizations and Analysis

Human Gene Disease Data

Updated genes from OMIM

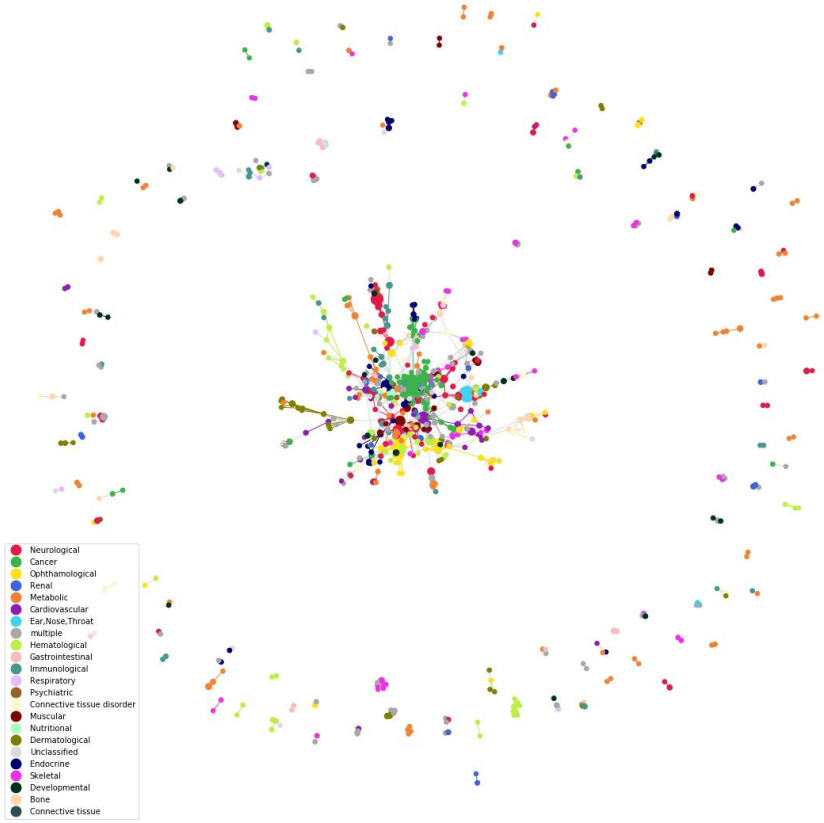


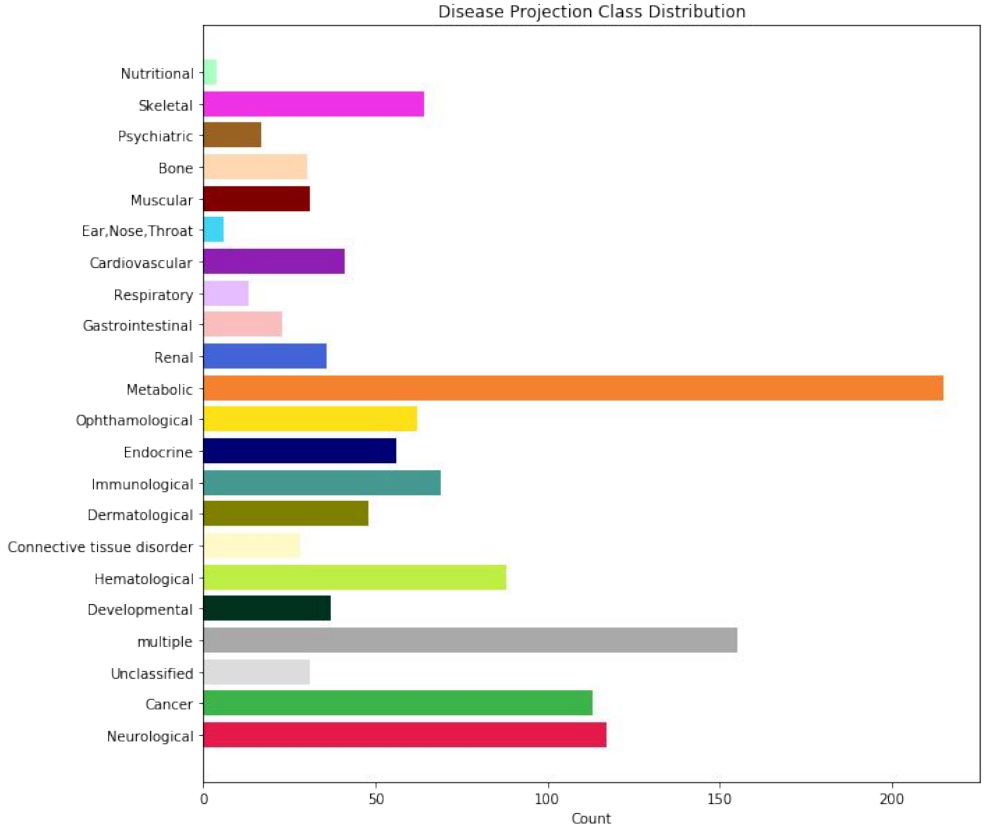
Disease Network



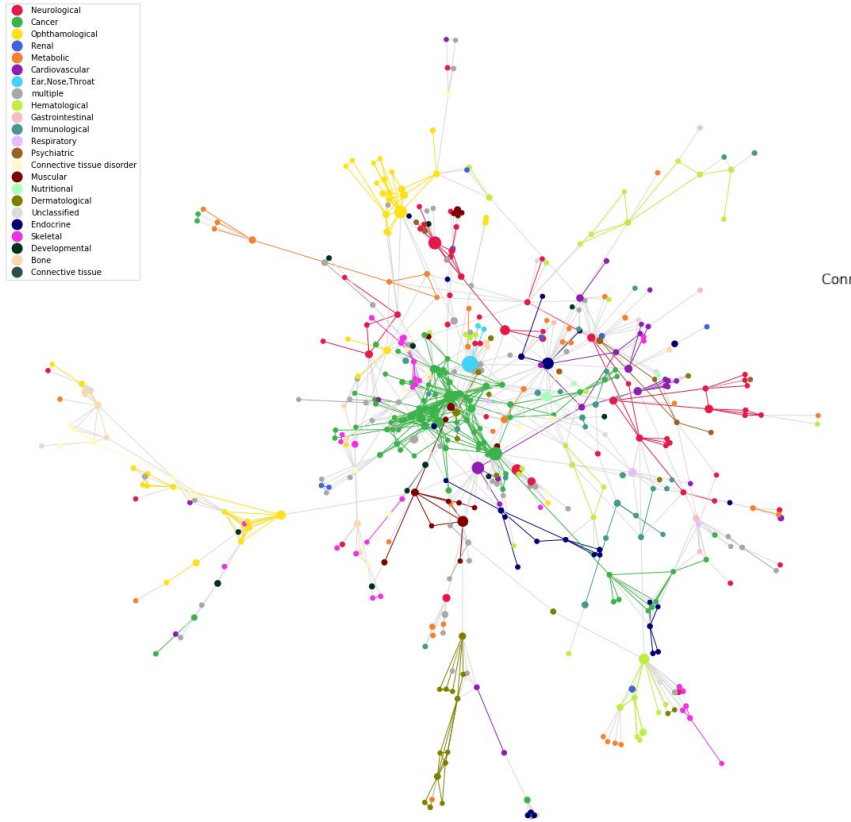
Gene Network

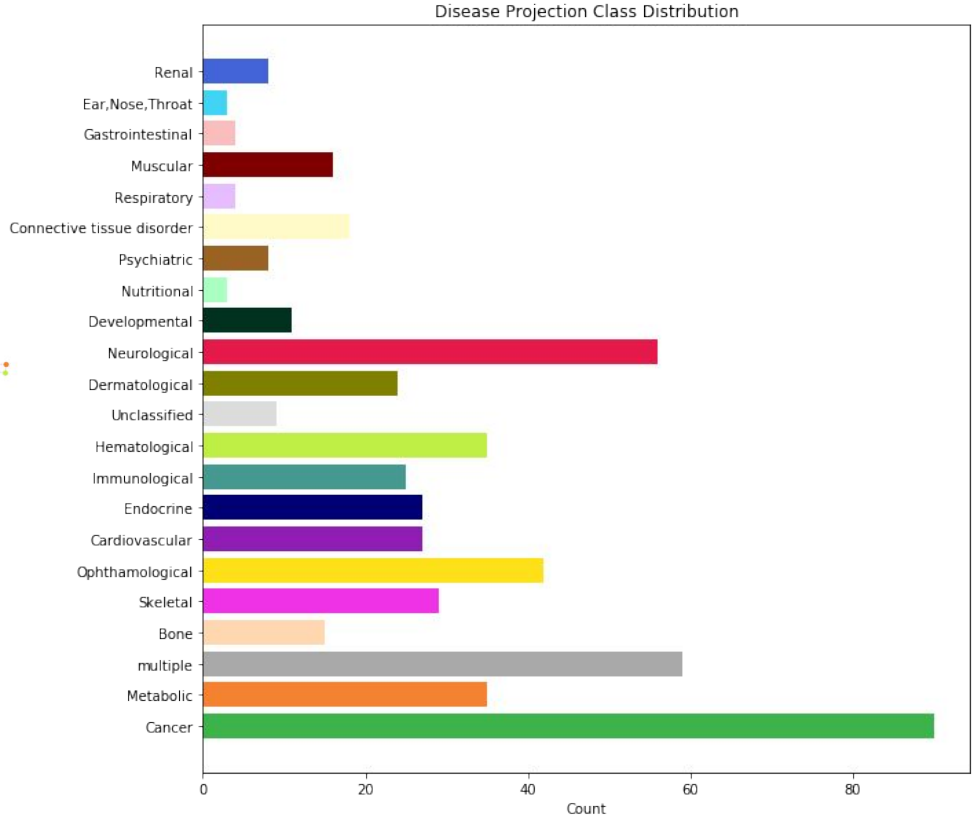
Disease Network



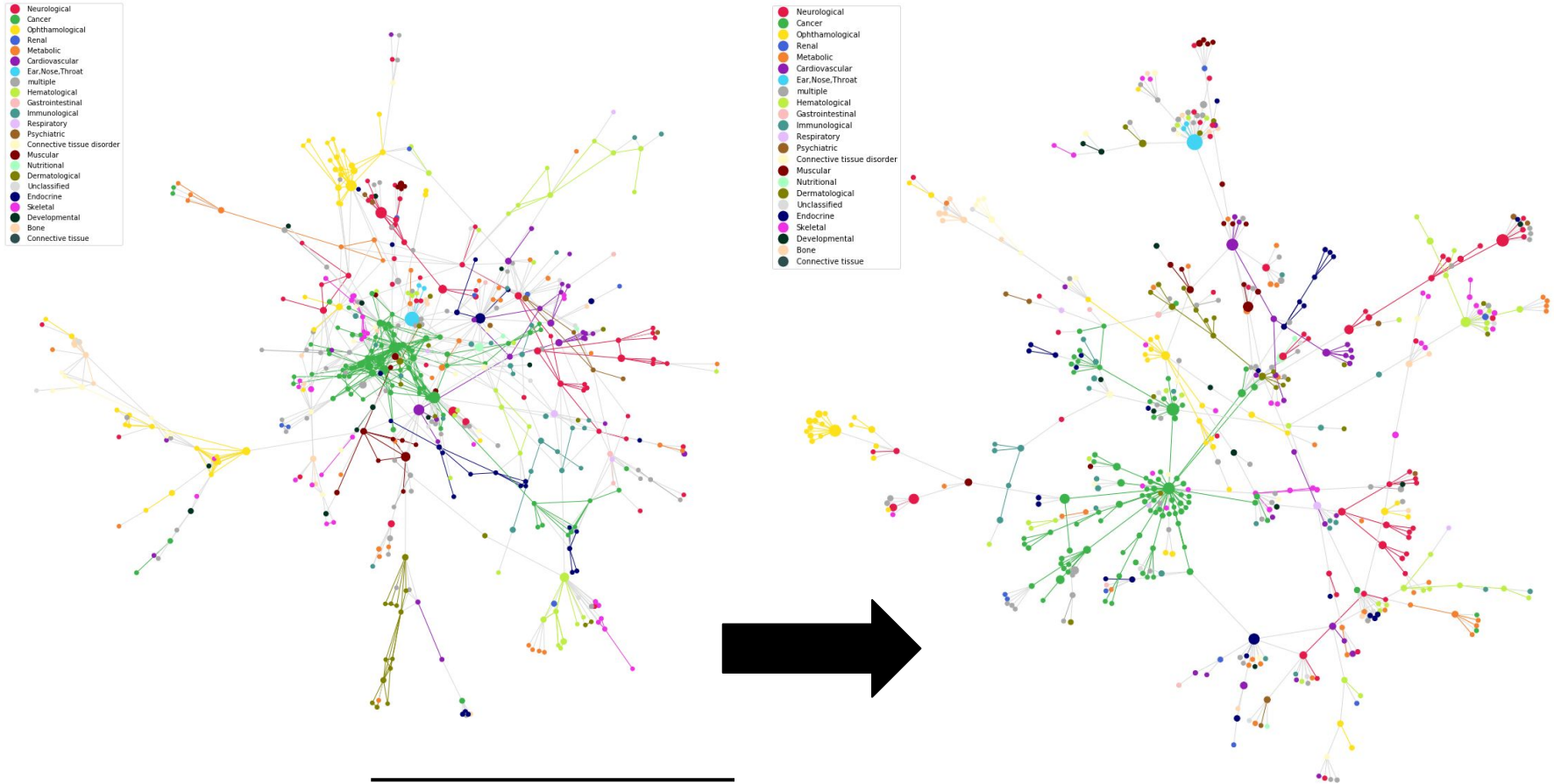


Connected Component: Disease Network

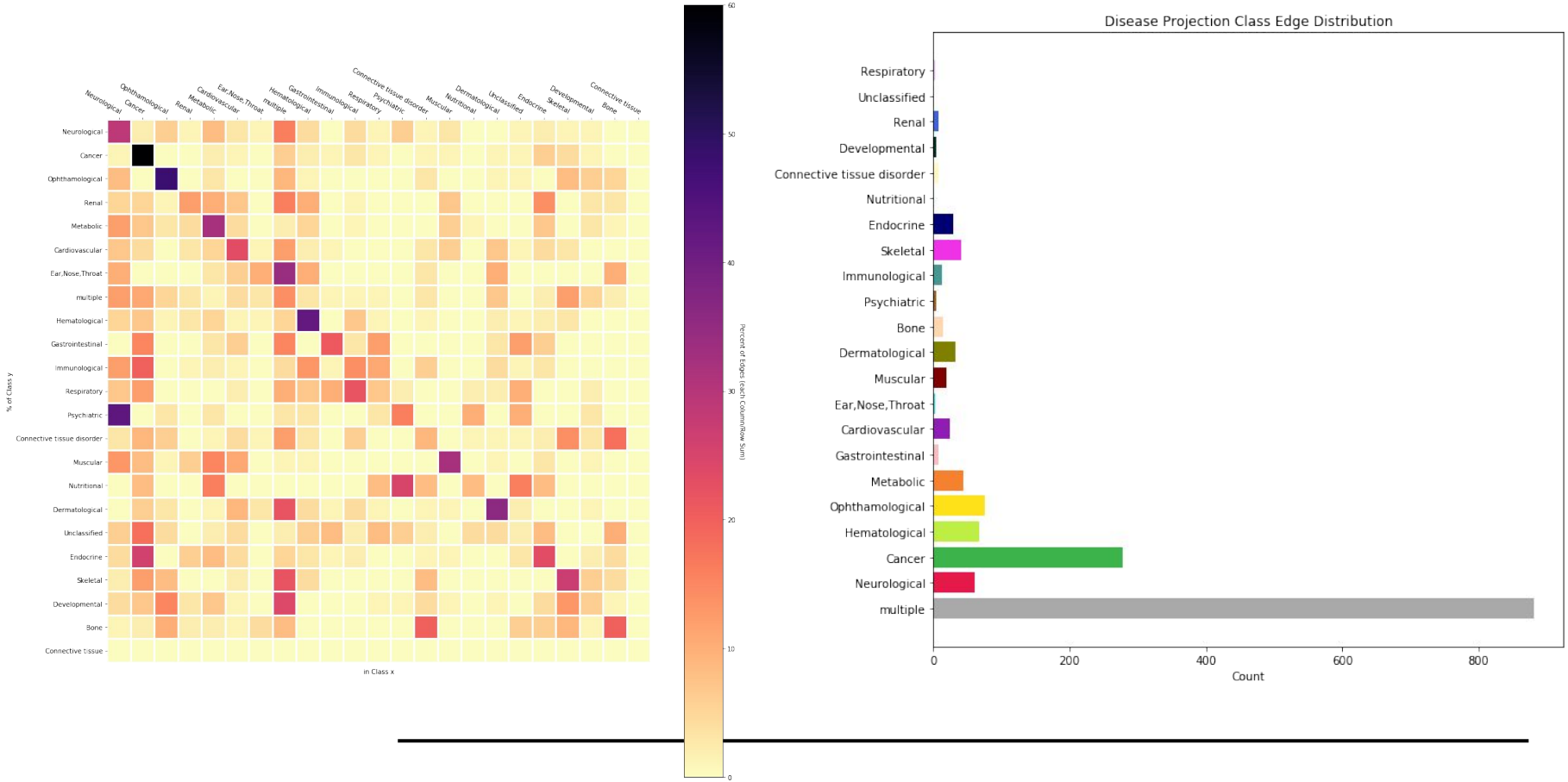




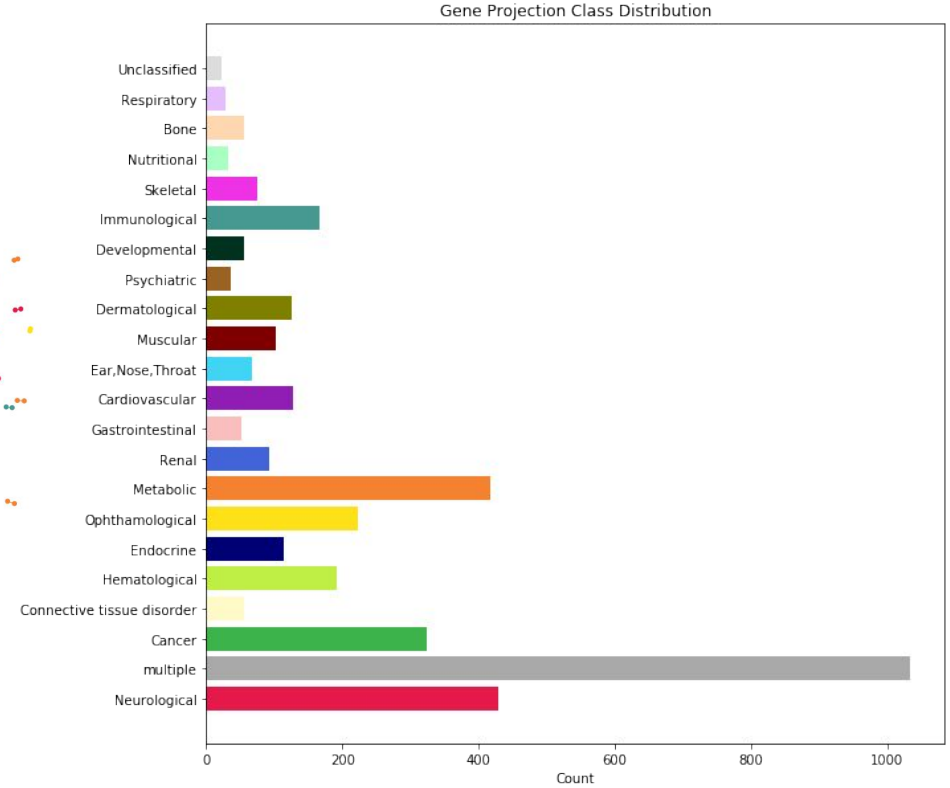
Breadth First Tree Traversal: Disease Network



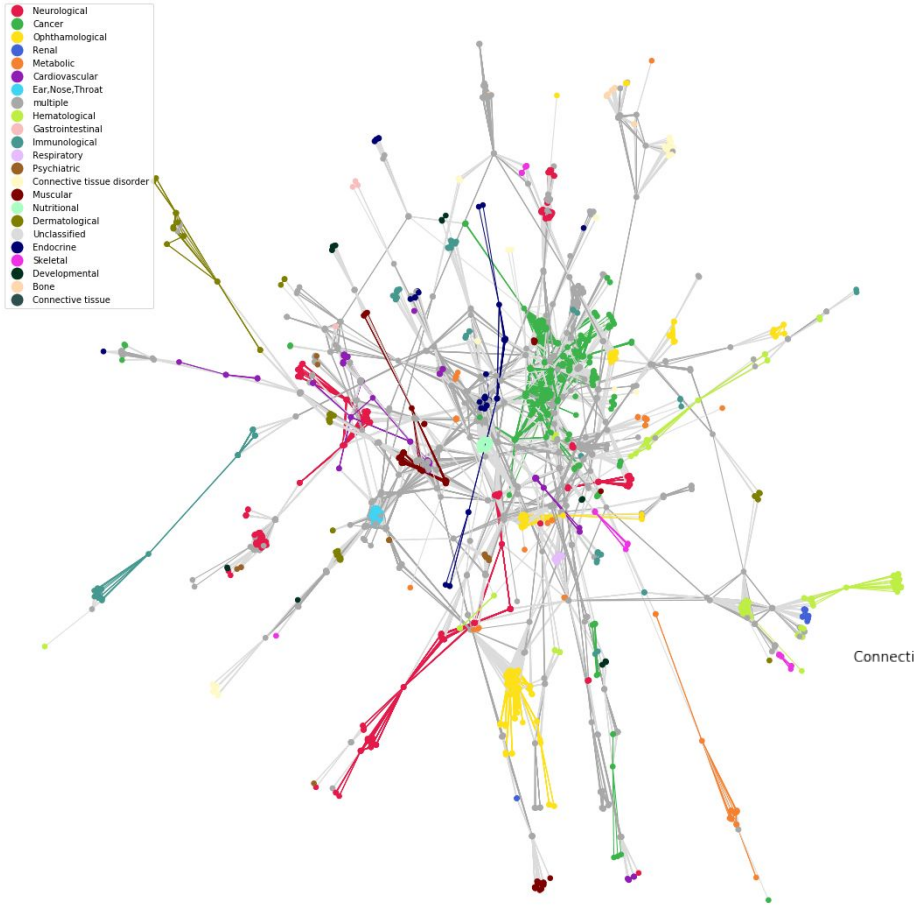
In v. Out Edges: Disease Network

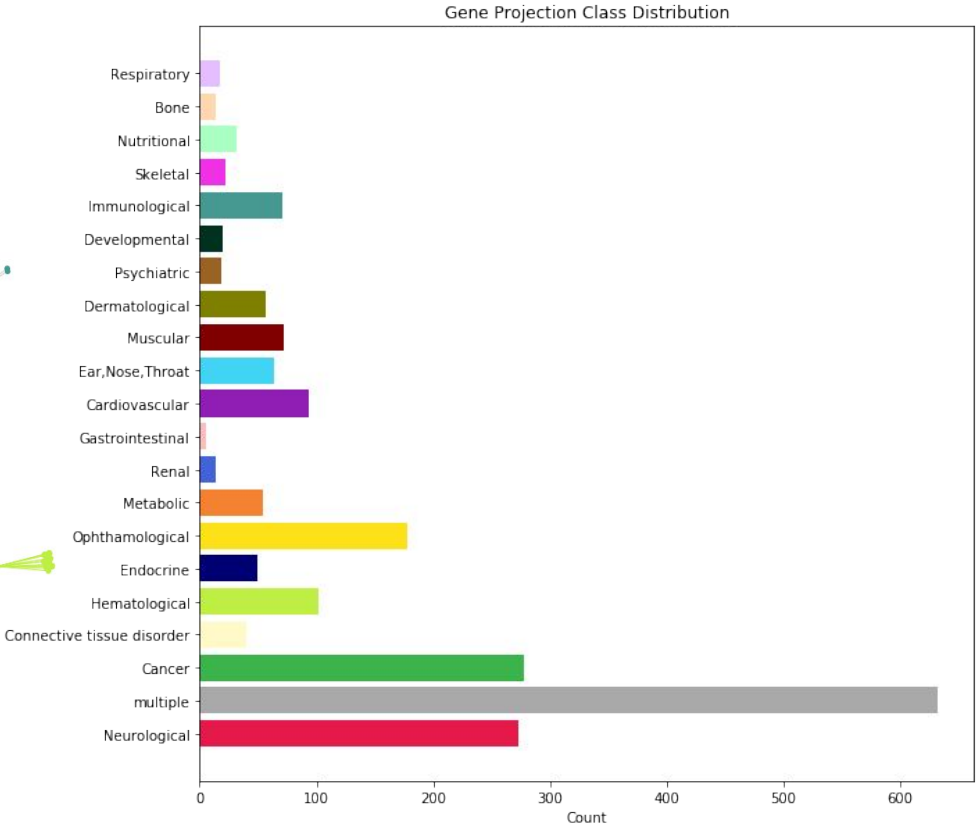


Gene Network

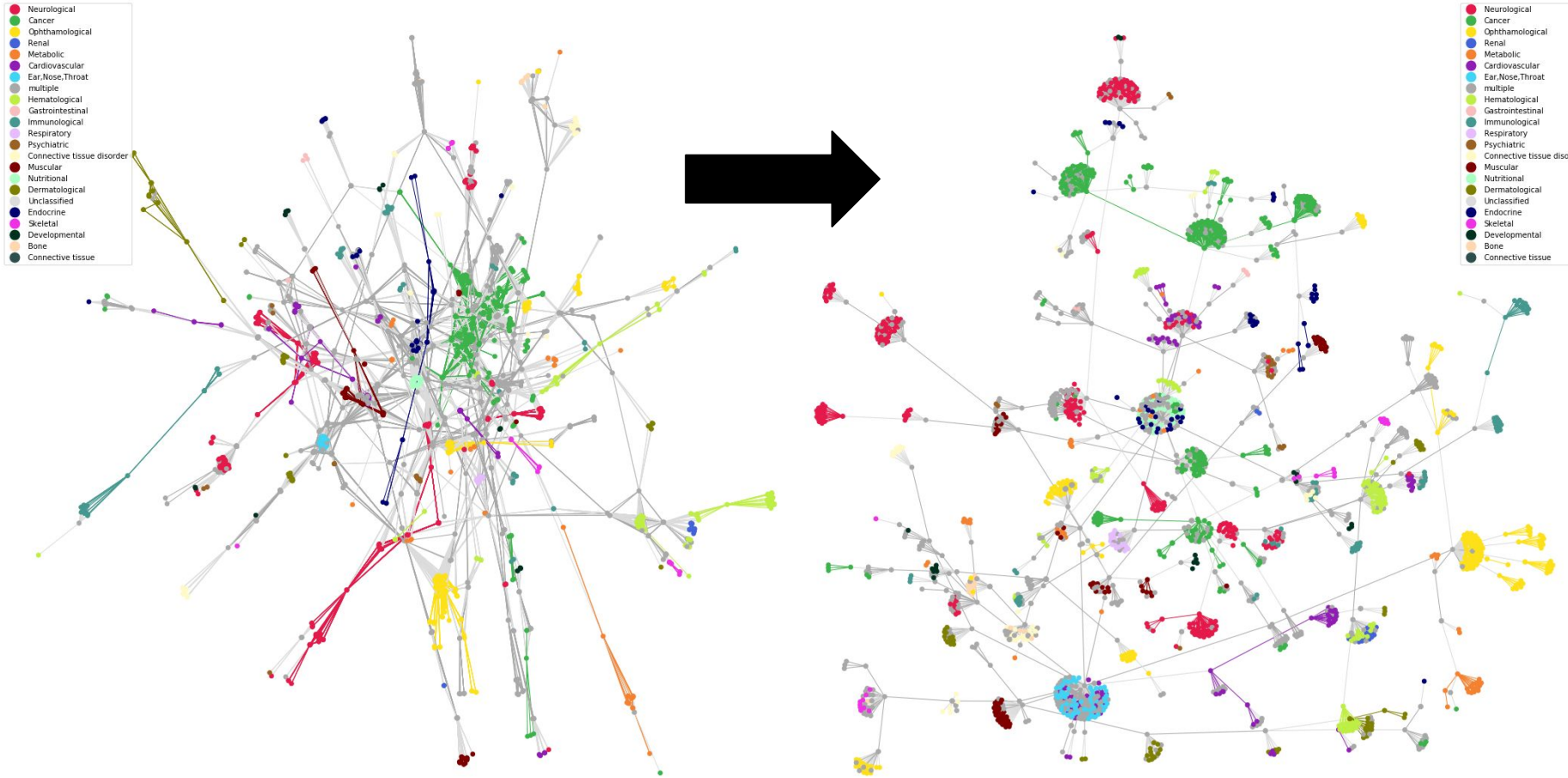


Connected Component: Gene Network

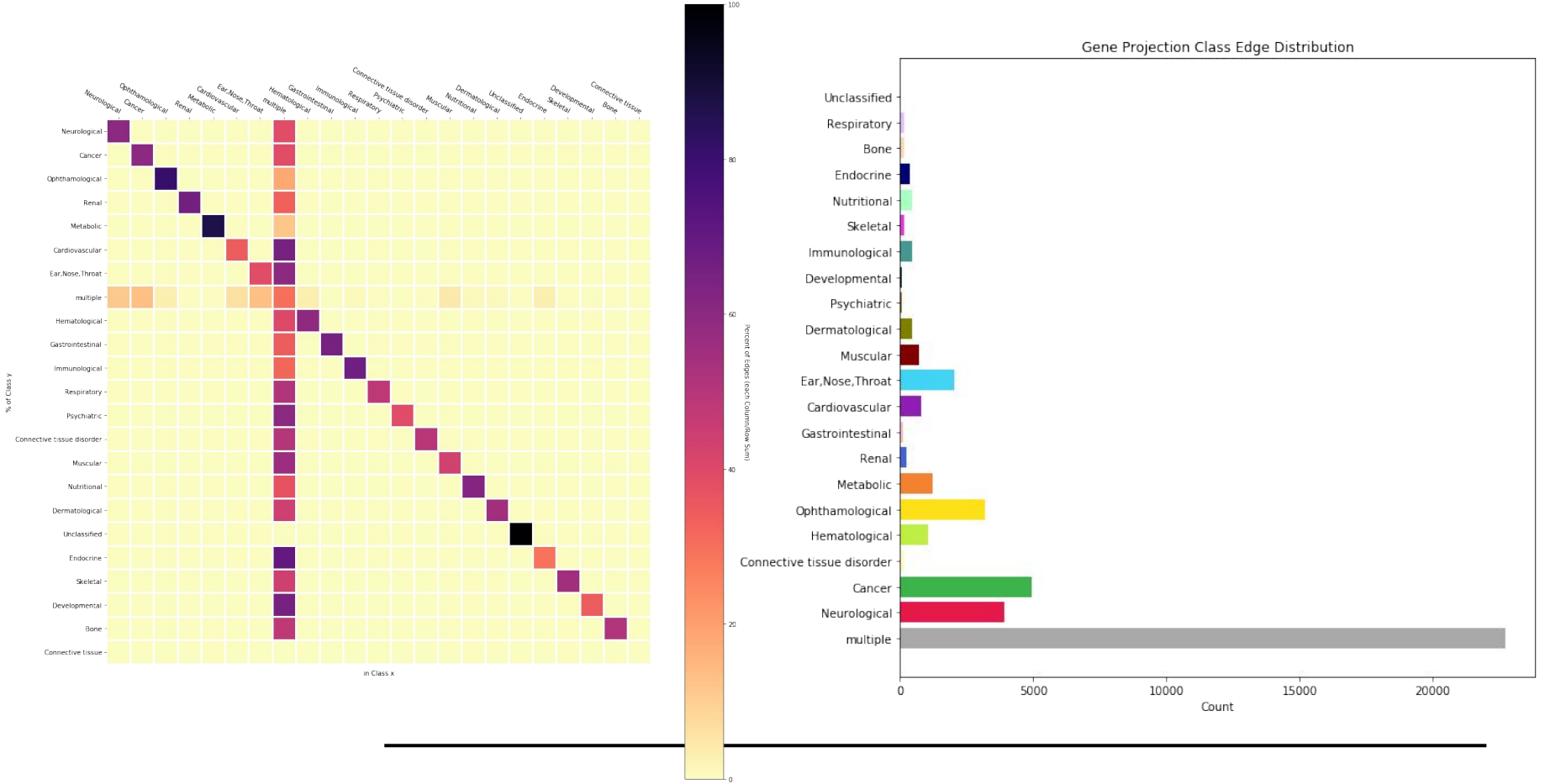




_ Breadth First Tree Traversal: Gene Network



In v. Out Edges: Gene Network



Attribute - Disorder Class Prediction

why

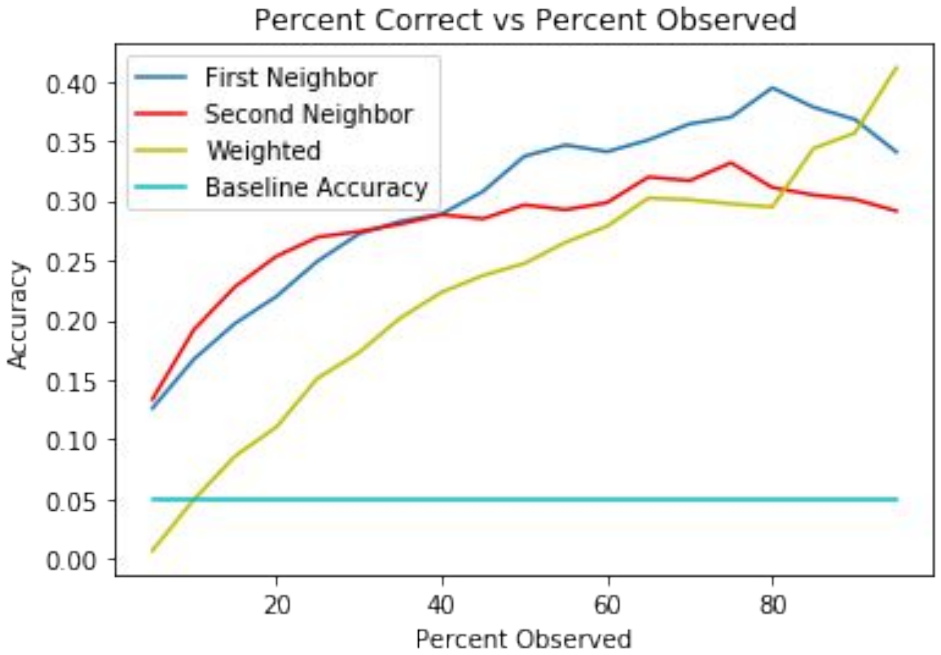
Assign disorder class to full OMIM data set.

Prediction Algorithms

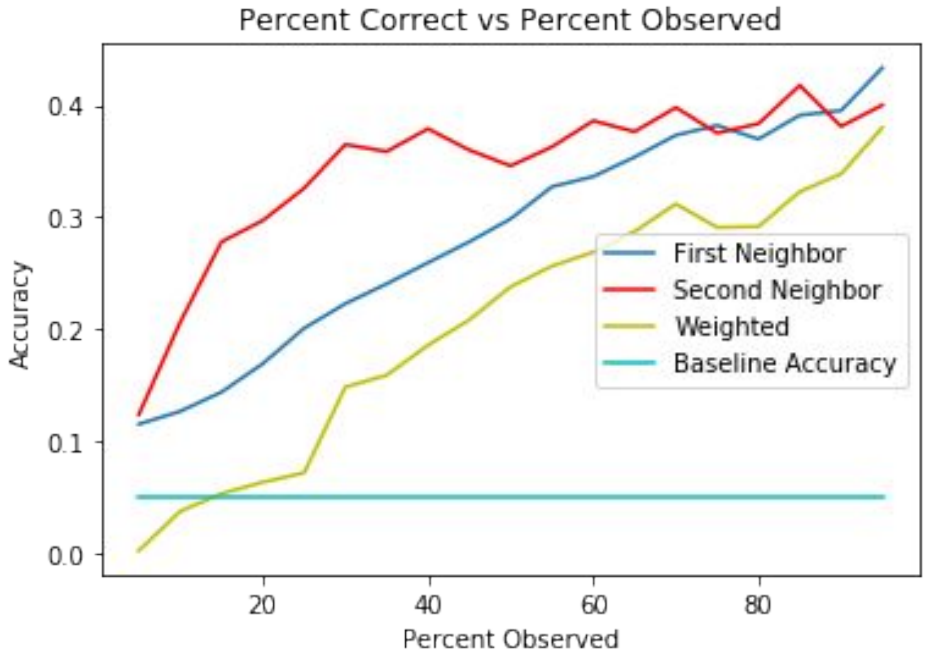
1. First Neighbor
2. Second Neighbor
3. Weighted
4. Baseline (random)

Predicting Classes of Disorders

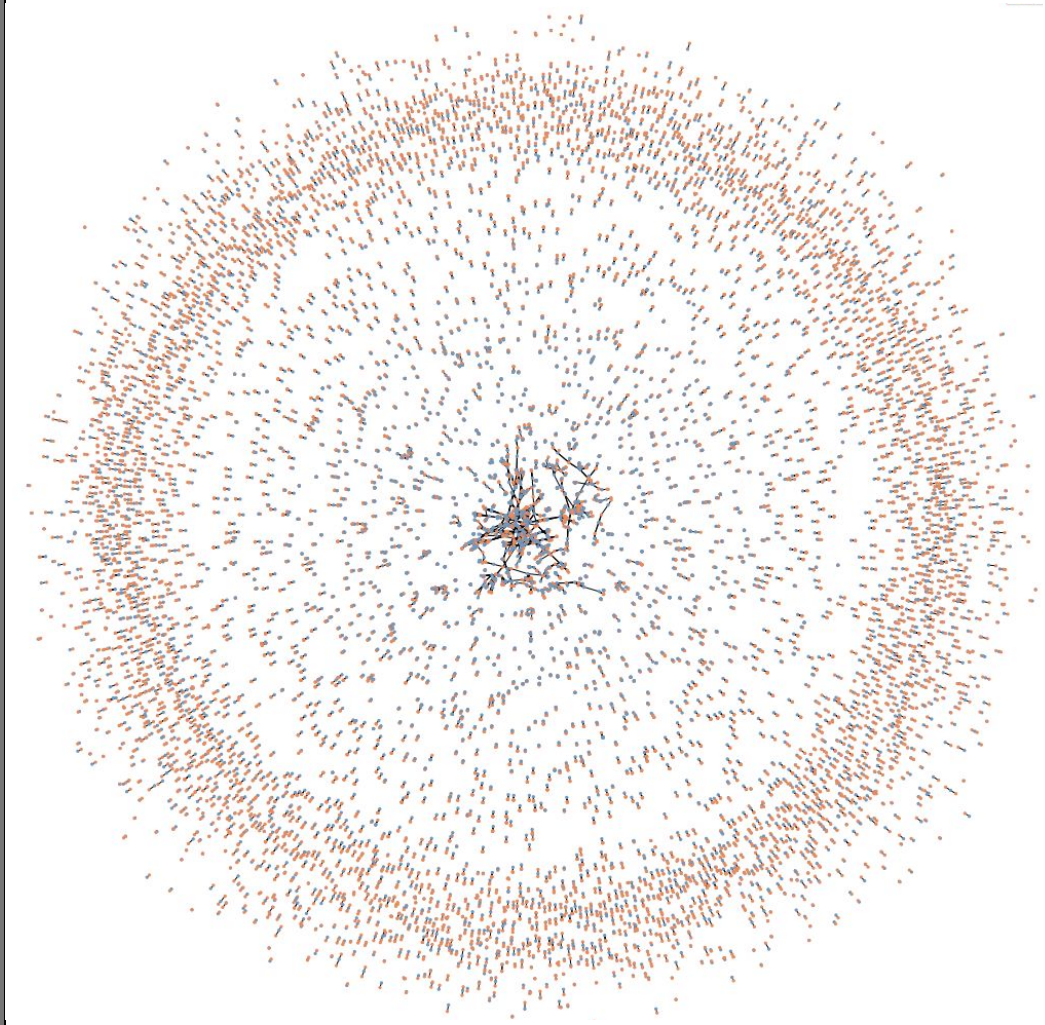
Normal Projection



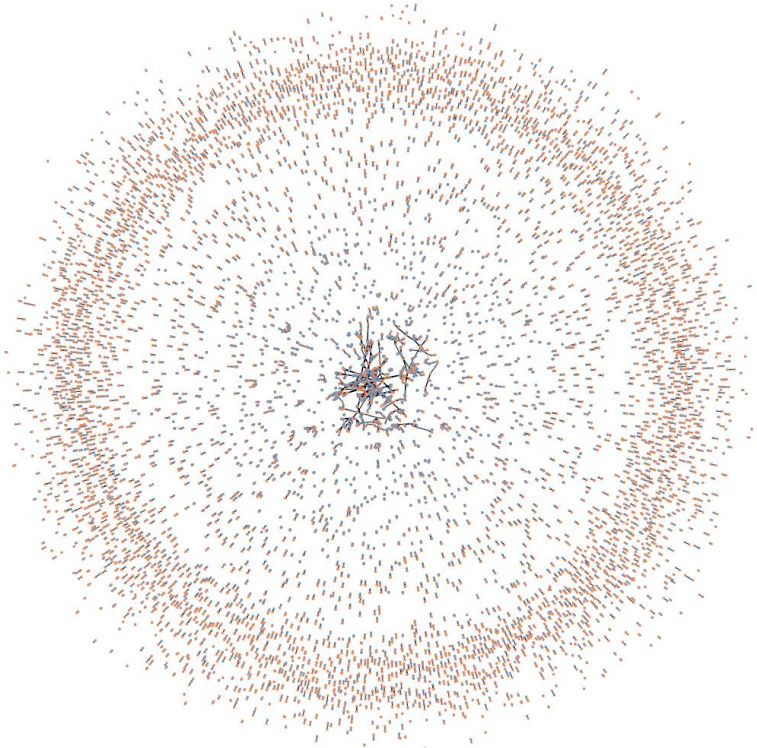
Breadth First Traversal



OMIM Data

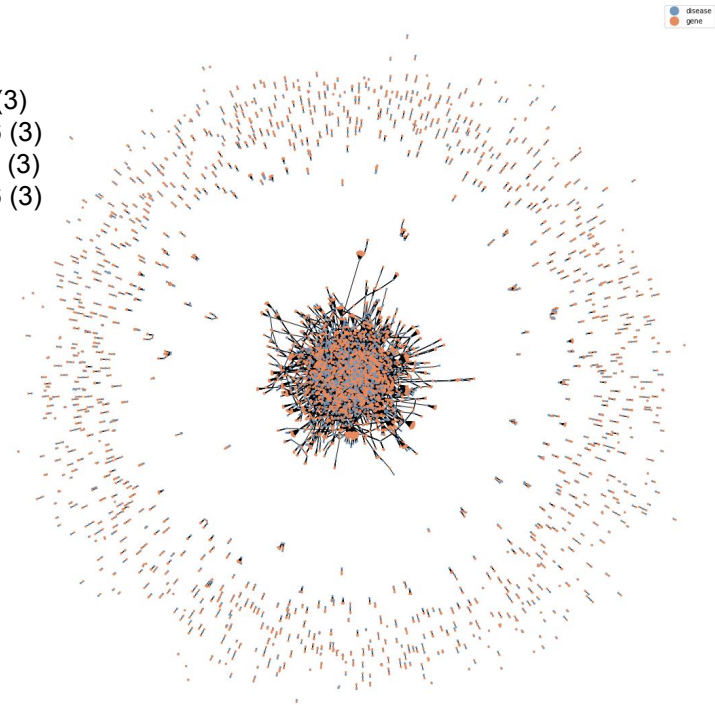


Condensing Diseases



Spinocerebellar ataxia 1, 164400 (3)
Spinocerebellar ataxia 10, 603516 (3)
Spinocerebellar ataxia 11, 604432 (3)
Spinocerebellar ataxia 12, 604326 (3)

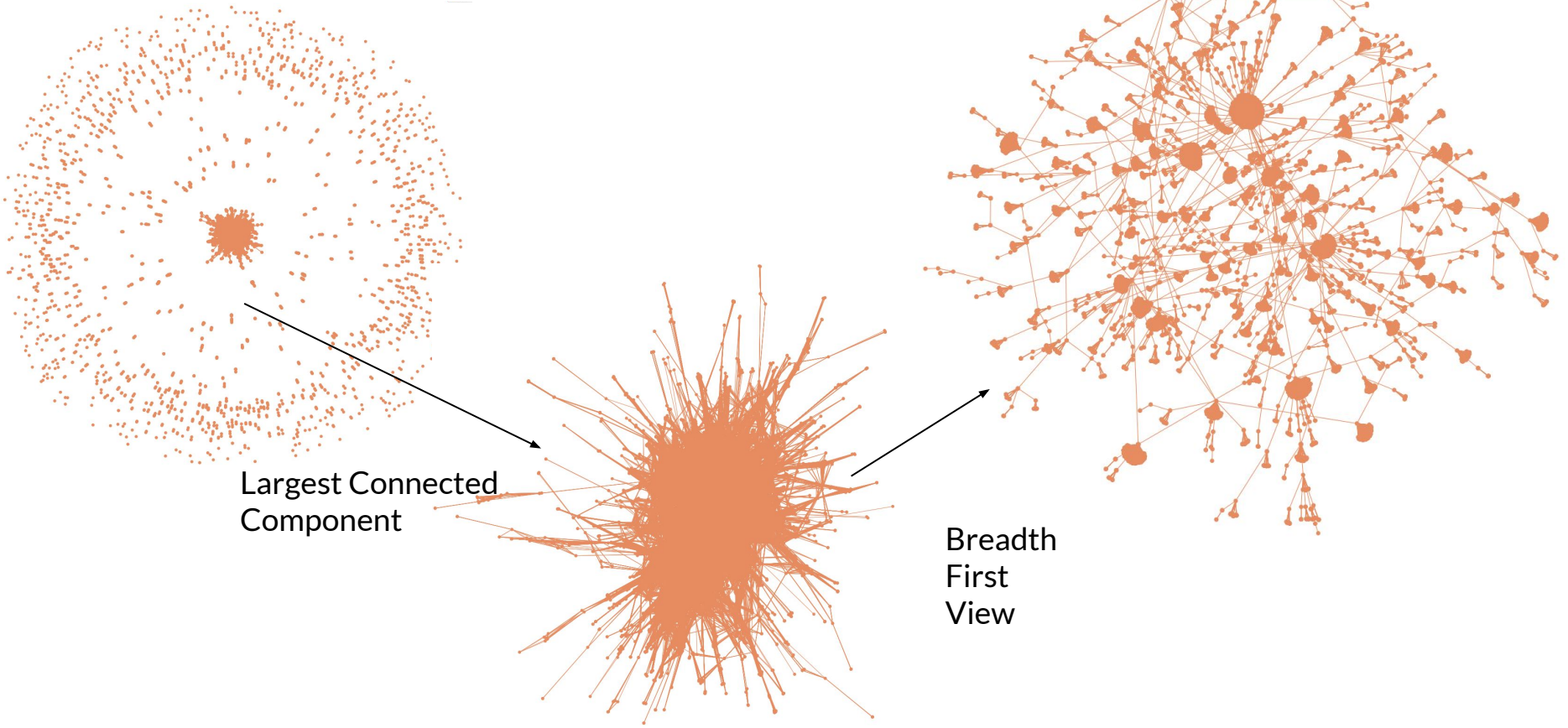
Spinocerebellar ataxia



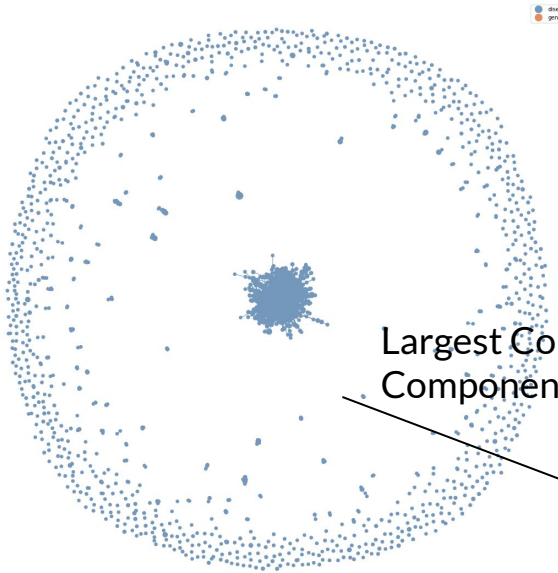

```
number of nodes, n   = 21733
number of edges, m   = 24398
mean degree,     <k> =  2.25
mean geodesic distance, <ell> =  1.67
clustering coefficient, C     =  0.00
```

```
number of nodes, n   = 16844
number of edges, m   = 22804
mean degree,     <k> =  2.71
mean geodesic distance, <ell> =  6.39
clustering coefficient, C     =  0.00
```

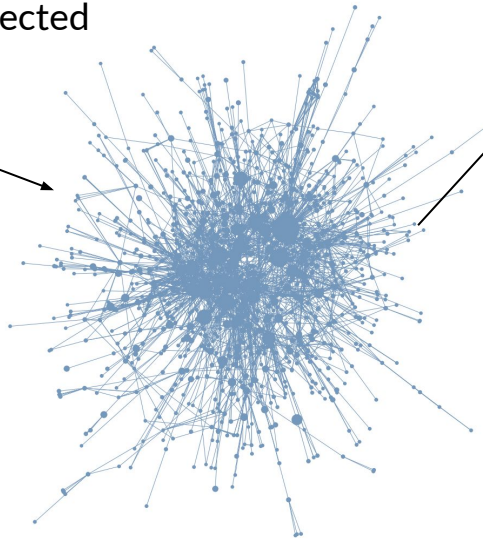
OMIM Gene Network



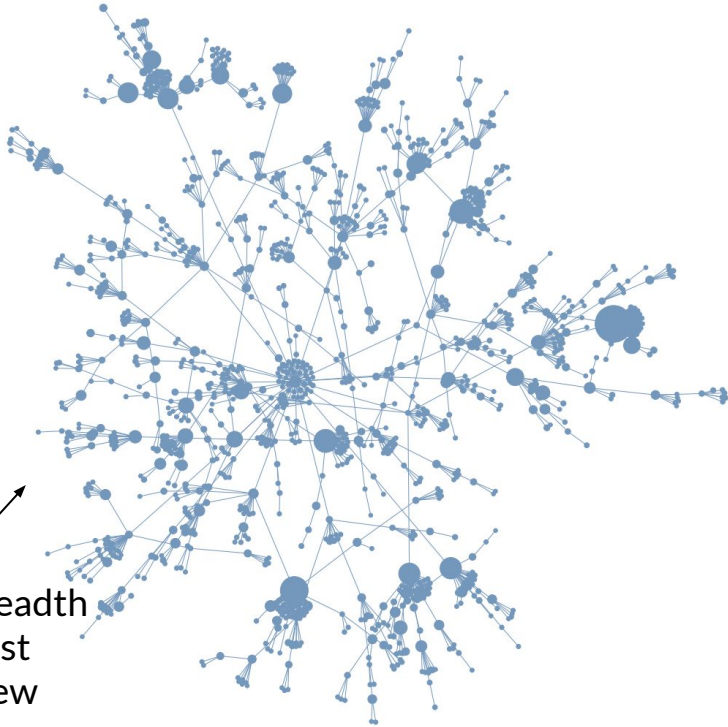
OMIM Disease Network



Largest Connected Component



Breadth First View

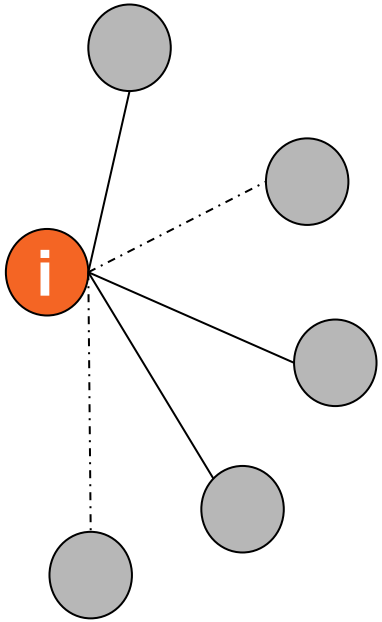


Link Prediction

Value of Link Prediction

Link prediction between disease and gene would help researchers narrow their research to relevant genes.

N-Fold Validation



1. Hide proportion of edges from i
2. Calculate similarity measures for every possible unobserved edge s.t. $(i, j) \ \forall \ j$ in G
3. Rank the unobserved edges by measure, pick top P (hyper-parameter) edges to predict as edges
4. Calculate accuracy metrics
5. Average over all i in G

Unobserved Edges

P

Similarity Measure	i	j	Pred
10	1	2	1
3	1	4	1
1	1	6	0
…			
0	1	45	0

Similarity Measures

Γ(x) - set of x's neighbors

1. Common Neighbors O(Nk²)

Common Neighbors = $|\Gamma(x) \cap \Gamma(y)|$

2. Jaccard Coefficient O(Nk²)

$$\frac{\#\text{ Common Neighbors}}{\#\text{ Neighbors}} \quad \left|\frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)}\right|$$

3. Preferential Attachment

Degree of x * Degree of y $|\Gamma(x)| * |\Gamma(y)|$

4. Adamic/Adar Index

Common Neighbors with degree penalty

$$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

5. Leicht-Holme-Newman Index

$$\frac{\#\text{ Common Neighbors}}{\text{Degree Product}} \quad \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x * k_y}$$

6. $Katz_B$ Index

Path lengths between x and y weighted by B

$$\sum_{l=1}^{\infty} \beta^l \cdot \left|\text{paths}_{xy}^{\langle l \rangle}\right|$$

https://www.hindawi.com/journals/sp/2015/172879/

Disclaimers

- OMIM network is far from fully observed
- Link prediction measures were created to be on unipartite networks, because of common neighbors

Important Metrics

Recall - TP/(TP+FN)

Proportion of actual edges that are detected

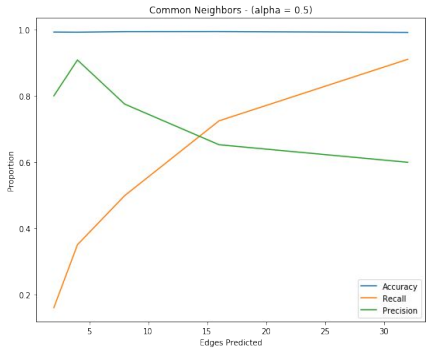
Precision - TP/(FP+TP)

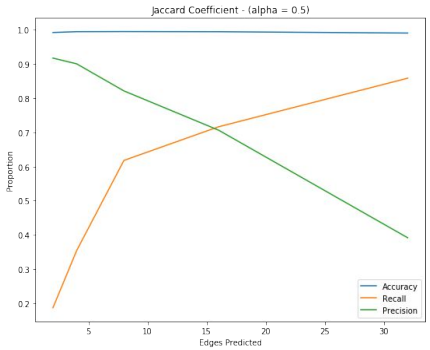
Proportion of edges detected that are actual edges

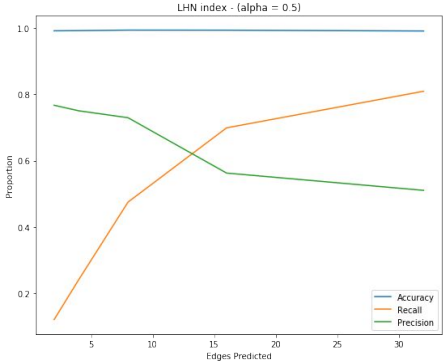
Disclaimers

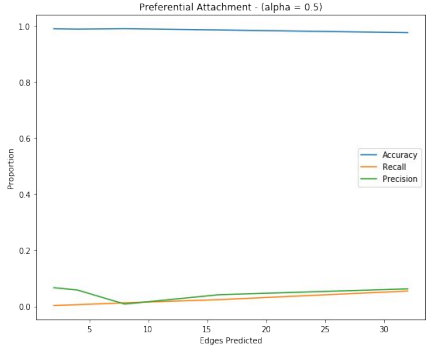
- OMIM network is far from fully observed
- Link prediction measures were created to be on unipartite networks, because of common neighbors

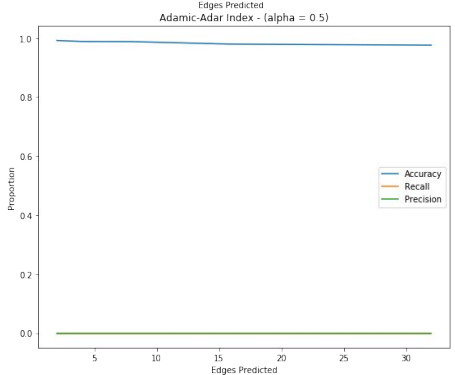
Gene Network

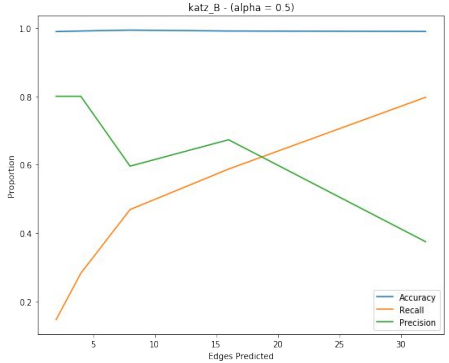




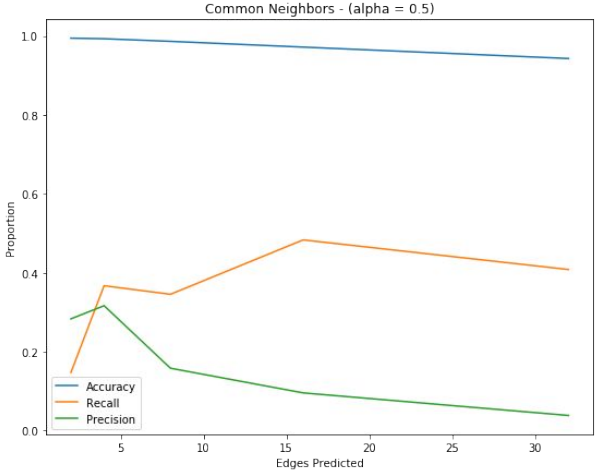


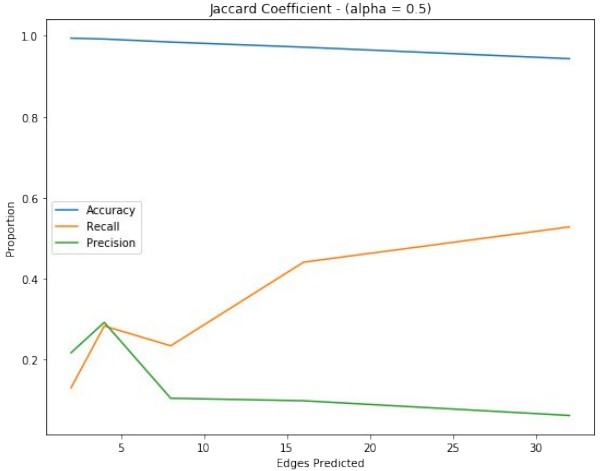


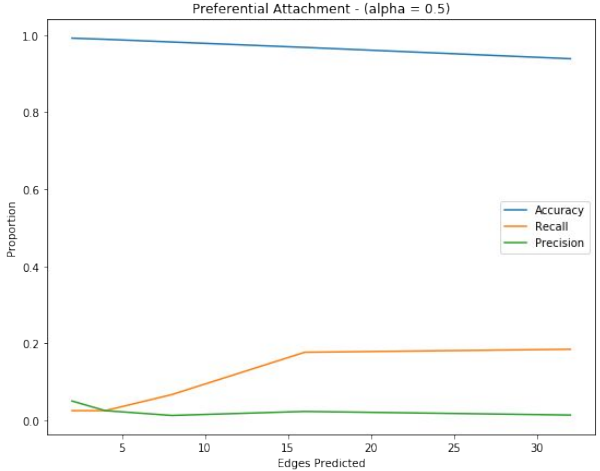


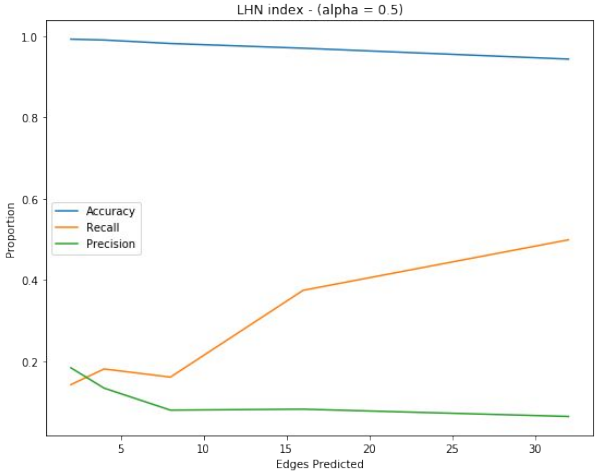


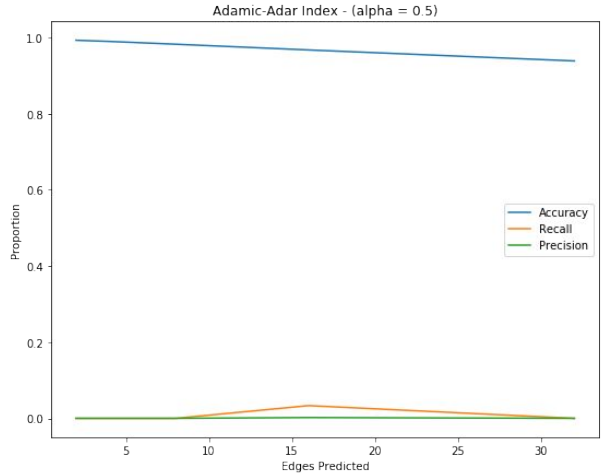
Disease Network

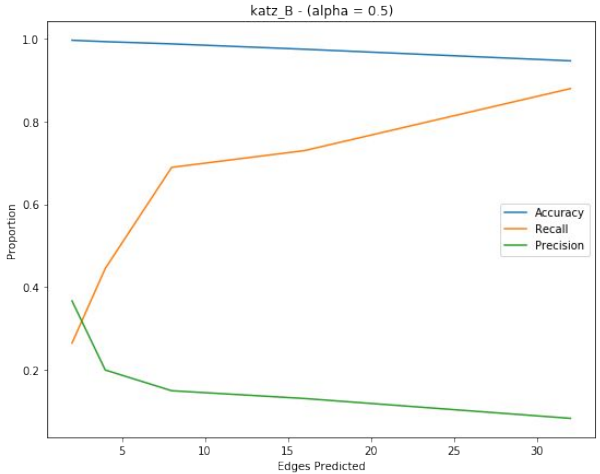












Future Directions

- Estimate class for OMIM
- Create link prediction for bipartite network