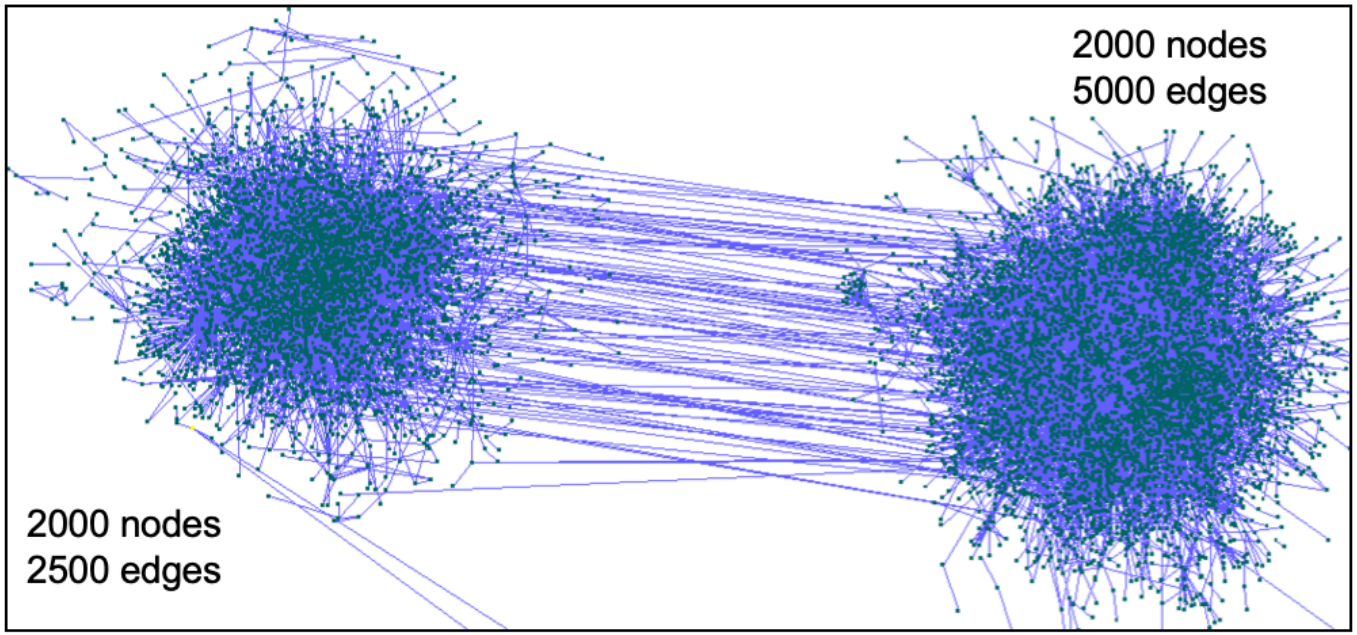
PageRank Network Aligner

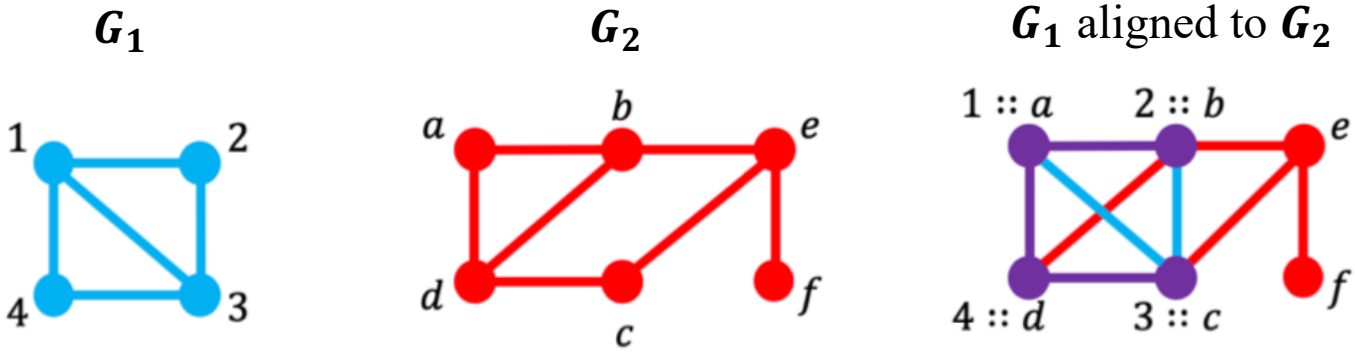
By Rishi Desai. 106X Fall 2019.



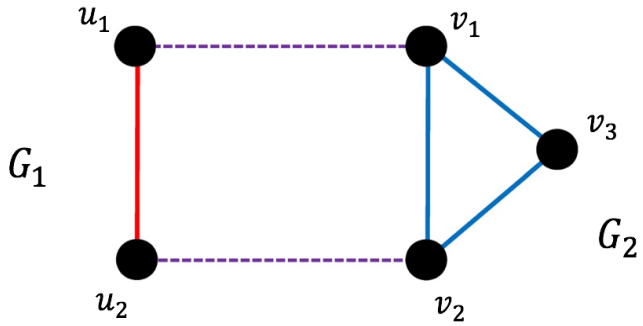
Rat PPI network (left) aligned to Yeast PPI network (right)

Network Alignment: the task of finding a mapping from the network $G_1 = (V_1, E_1)$ to the network $G_2 = (V_2, E_2)$.

- Motivation: Network alignments between undirected protein-protein interaction (PPI) networks allows one to calculate topological similarity between diverse species (e.g. mouse, plants, worms).
- One can compare topological similarity between protein networks to confirm the relative evolutionary distance between species
- Mathematically, a network alignment is a function $a : V_1 \rightarrow V_2$ where $|V_1| \leq |V_2|$



Covered Edge: An edge present in G_1 and G_2 with two aligned endpoints (e.g. aligning **1** to **a** and **2** to **b** will produce a covered edge.



EC and S^3 are two measures to calculate the topological similarity between two PPI networks with alignment a:

- **Edge Coverage (EC):** ratio of covered edges to all edges in G_1
- **Symmetric Substructure Score (S^3):** ratio of covered edges to all edges between two aligned endpoints
- In the first example diagram: EC(a) = 3/5 and S^3(a) = 3/6

Project Goal

I will utilize the PageRank algorithm to efficiently generate a network alignment and will use the topological similarity (Edge Coverage and S^3) to analyze the species that the protein networks represent.

PageRank Algorithm

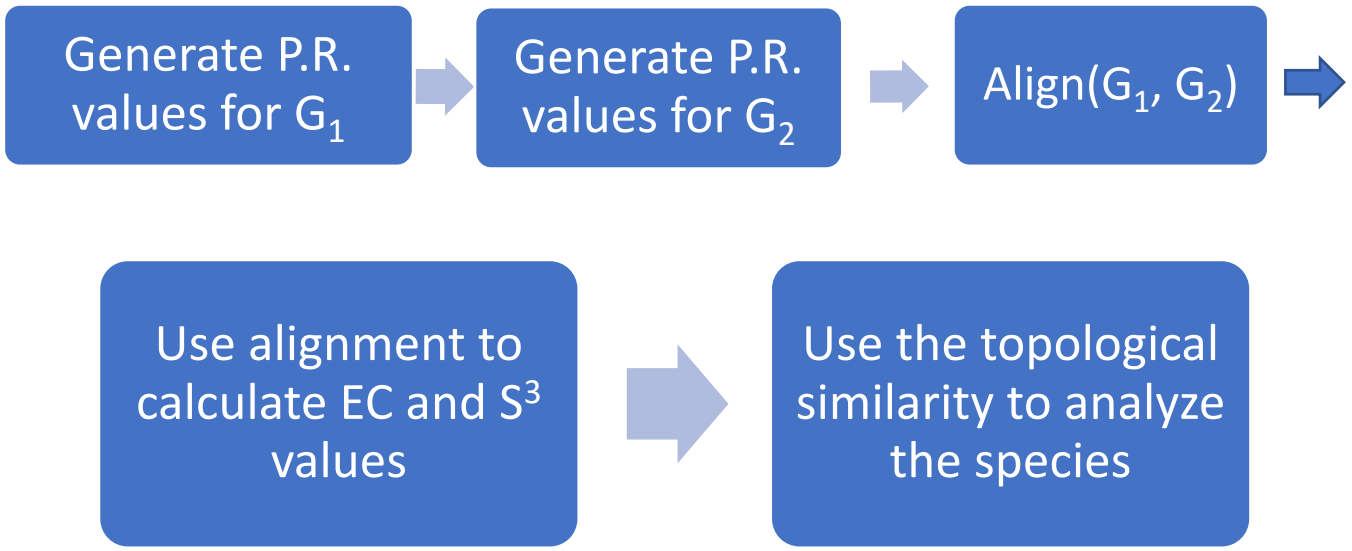
- I used an iterative method to approximate a chain of Markov matrix multiplications for $G = (V, E)$.
- The n-vector R(t) represents each node's PR values at iteration t.
- M is a steady state matrix where $M_{ij} = 1/L(v_j)$ if node v_j links to v_i and 0 otherwise. $L(v_j)$ is the number of outwards links from node v_j

$$\boldsymbol{R}(0) = \langle \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n} \rangle, \text{where } n = |V|$$

$$\boldsymbol{R}(t) = \alpha M \boldsymbol{R}(t-1) + (1-\alpha)/n\mathbf{1}$$

After testing several values of dampener α, I found $\alpha = 1$ to provide the best results. Furthermore, using only matrix-vector multiplication, my algorithm runs in $O(kn^2)$, where k = # iterations and n = # nodes. The iteration stops when the condition below is achieved with $\epsilon = .000001$:

$$\boldsymbol{R}(t) - \boldsymbol{R}(t-1) < \epsilon$$



Algorithm: creating the alignment **map** : $V_1 \rightarrow V_2$
Align(G_1,G_2):

1. sort(V_1) from greatest to least P.R. values
2. sort(V_2) from greatest to least P.R. values
3. for **i** = 0 to **i** = $|V_1|$:
   map.put(V_1[**i**], V_2[**i**])
4. return **map**

Protein Network Data Set

My project demo allows the user to align network pairs from 8 undirected protein-protein interaction (PPI) networks from the BioGrid Database.

Species	Name	Identifier	Nodes	Edges
C.Elegans	Roundworm	CE	3134	5428
S.Pombe	Fission yeast	SP	1911	4711
M.Musculus	Mouse	MM	4370	9116
A.Thaliana	Plant	AT	5897	13381
S.Cerevisiae 00	Baker's Yeast	Y00	1004	8323
S.Cerevisiae 05	Yeast + 5% noise	Y05	1004	8739
S.Cerevisiae 10	Yeast + 10% noise	Y10	1004	9155
S.Cerevisiae 20	Yeast + 20% noise	Y20	1004	9987

Featured Alignment Results

Alignment	EC (%)	S^3 (%)
SP-CE	2.53	1.35
CE-MM	1.95	0.80
MM-AT	1.34	0.59
Y00-Y05	18.31	9.81
Y00-Y10	19.50	10.24
Y00-Y20	20.55	10.30

Conclusion

- The results show how my algorithm is able to find higher similarity between the smaller Yeast species compared to the larger species.

- I confirmed that M.Musculus (mouse) is closer evolutionary to C.Elegans (roundworm) than to A.Thaliana (plant) based on similarity score in EC and S^3.

- The data show PageRank is better at maximizing similarity for edge dense networks (e.g. Yeast) rather than more sparse networks (e.g. Plant)

Challenges

- Converging the Markov matrix calculation in a timely fashion was difficult. In order to efficiently test convergence, I randomly sampled 10 nodes from the *n*-vector and averaged the successive differences rather than checking all *n* nodes in each iteration.

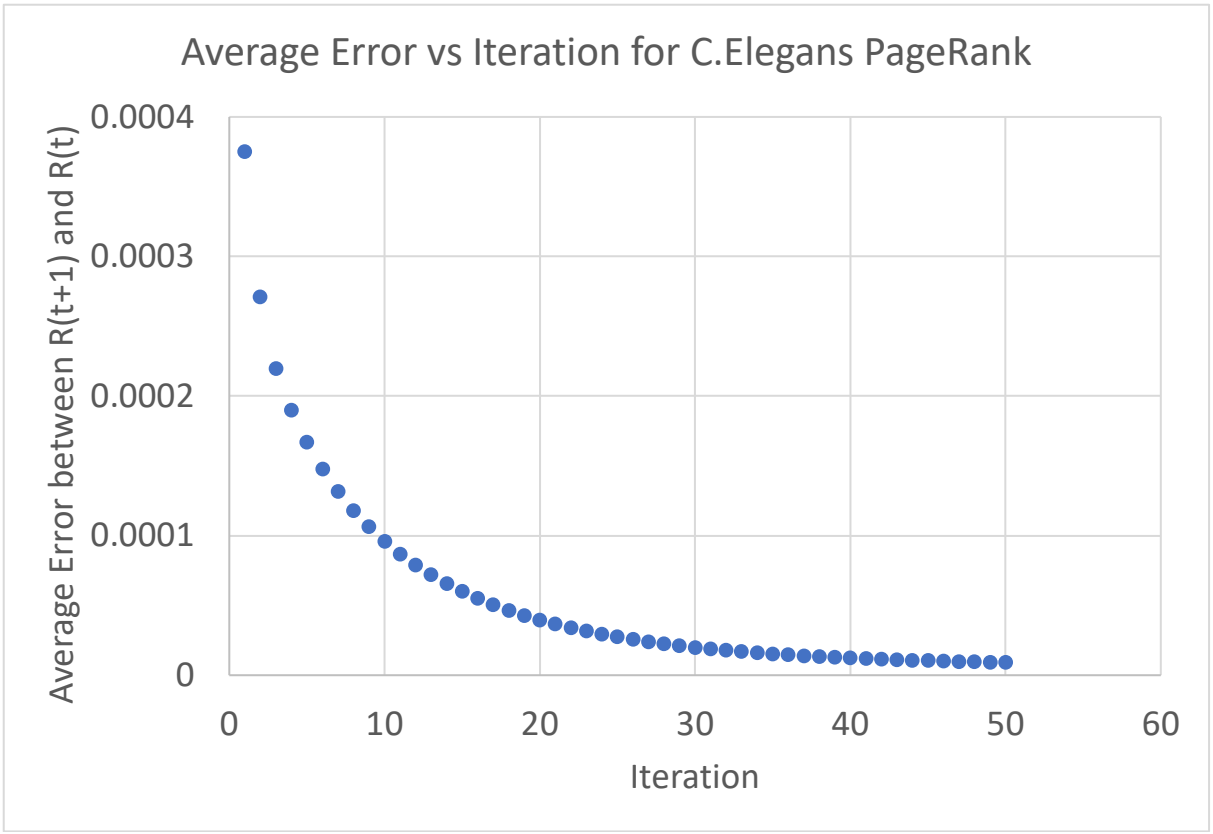
- PageRank depends on two constants, α and ϵ. Setting these constants affects convergence speed significantly. Choosing the variables without arbitrarily testing was challenging.

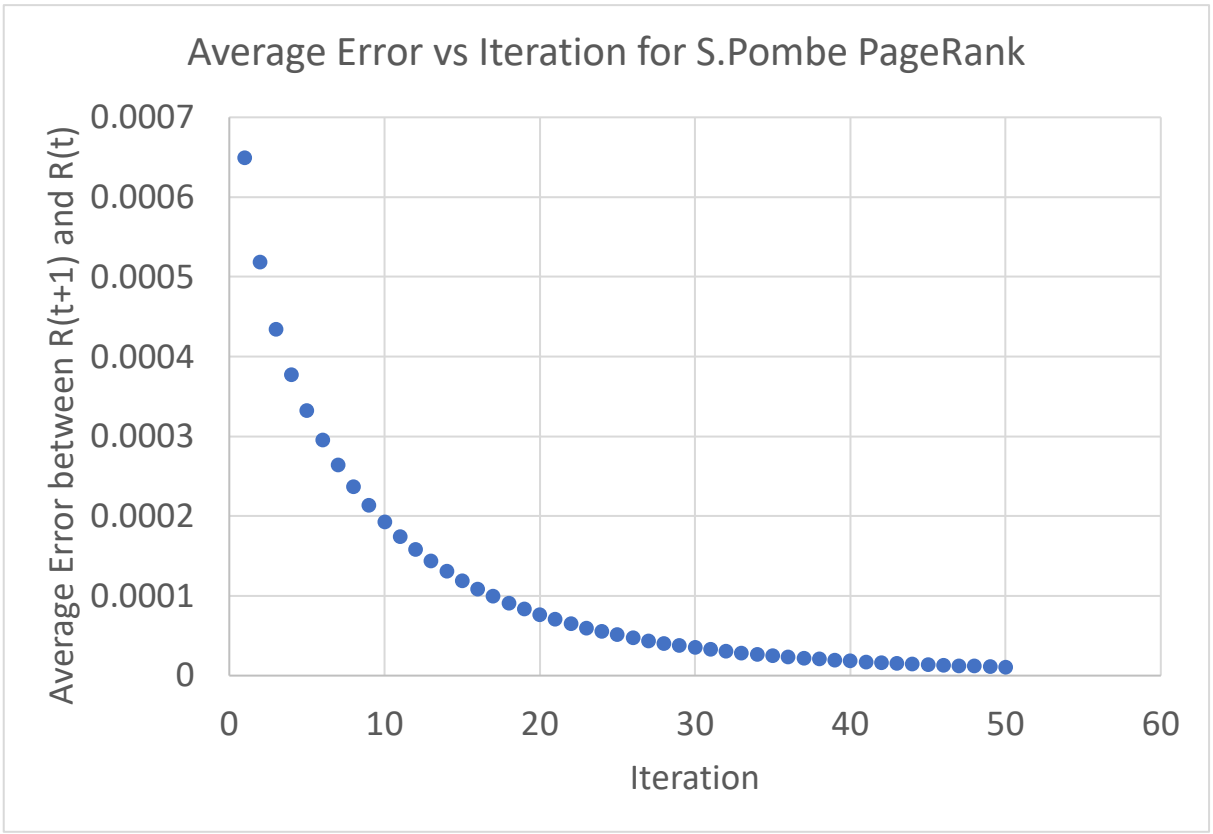
Future Work

- Many network alignment algorithms utilize protein sequence data in addition to topological data, thereby producing better alignments
- I could incorporate protein sequence data into my PageRank algorithm by editing the steady state matrix.

References

- Chatr-Aryamontri,A. et al. (2013) The biogrid interaction database: 2013 update. Nucleic Acids Res., 41, D816–D823.
- Saraph,V. and Milenkovic,T. (2014) Magna: maximizing accuracy in global network alignment. Bioinformatics, 30, 2931–2940.


Average Error vs Iteration for C.Elegans PageRank
Average Error between R(t+1) and R(t)
0.0004
0.0003
0.0002
0.0001
0
0
10
20
30
40
50
60
Iteration


Average Error vs Iteration for S.Pombe PageRank
Average Error between R(t+1) and R(t)
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
0
10
20
30
40
50
60
Iteration