Embedding Reconstruction - Experiment 27

Preface

Contents

- Embedding Reconstruction - Experiment 27
- Preface
  - Contents
  - Metadata
  - Relevant scripts
- Summary
  - TL;DR
  - Experimental Goal
  - Hypotheses (if applicable)
- Methods
  - Data
  - Procedure
- Results
- Analysis
- Future Work

Metadata

- Project ID: EMB_ex27
- Researchers: Daniel Kaiser
- Version history:

Version	Date Designed	Date Conducted	Date Analyzed	Notes
v0.1	2023-02-23	2022-02-23	2022-02-23	Prototyped refactored workflow
v0.2	2023-02-23	2022-02-23	2022-02-23	Expanded into "trouble" duplexes, tested error handling
v1.0	2023-02-23	2022-02-23	2022-02-27	Full run

Relevant scripts

The experimental simulations were run through the Python script `workflow/[operative/active]/ex27/EMB_ex27_logreg-degree.py`. A dataframe was created within the workflow and the resultant dataframe `results/dataframes/dataframe_EMB_ex27[version]_DK_[date].csv` was treated as the input data to the analysis in `notebooks/viz/analysis_EMB_ex27.ipynb`, a Jupyter notebook.

Summary

TL;DR

We appear to be recreating our prior known results for the systems detailed below with the exception of the *C. Elegans* connectome. Right-asympotitic performance does appear to converge to known values.

Experimental Goal

Reproduce Naive Bayes paper's results (under "D" classifier) for real duplexes.

Hypotheses (if applicable)

We should see quantitatively equivalent behavior to the Naive Bayes classifier paper, reproduced below (proxied by dashed lines):

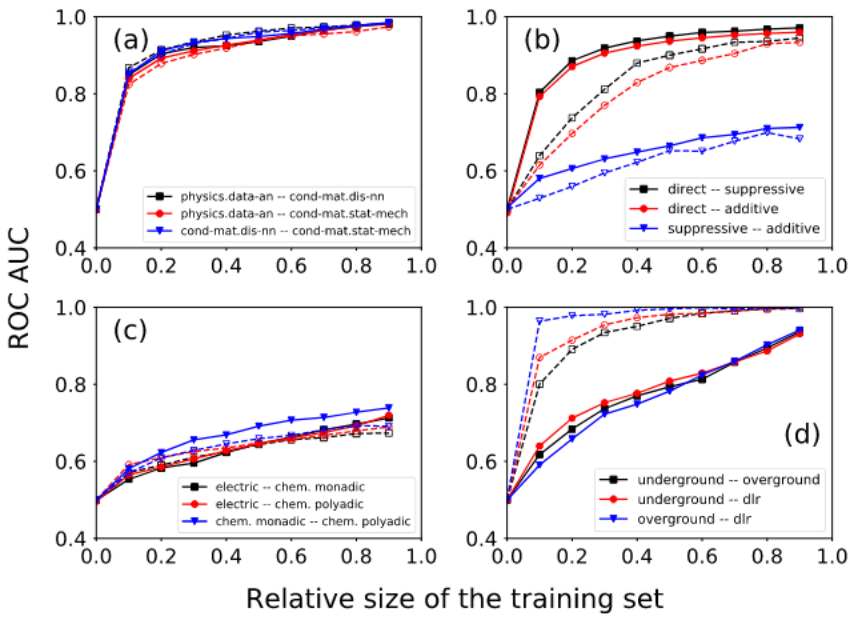

Figure 1: Previous Results

Methods

Data

We utilize four real multiplexes:

- arXiv collaboration network
- *C. Elegans* connectome
- *Drosophila* genetic interaction network
- London transportation network

Within these multiplexes, we induce a duplex and restrict our attention therein; respectively, these are:

- physics.data-an, cond-mat.dis-nn (2, 6)
- electric, chem. monadic (1, 2)
- direct, suppressive (1, 2)
- underground, overground (1, 2)

Procedure

1. [**Set-up**] Load dataset
2. [**Set-up**] Calculate total aggregate $A = \alpha \cup \beta$.
3. [**Set-up**] Observe training set $\Theta = \theta_\alpha \cup \theta_\beta$ of relative size (per layer) θ.
4. [**Set-up**] Form remnants $\mathcal{R}_\alpha, \mathcal{R}_\beta$ and aggregate $\tilde{A} = A - \Theta$.
5. [**Feature calculations**] Calculate degrees sequences k^α, k^β of $\mathcal{R}_\alpha, \mathcal{R}_\beta$.
6. [**Feature calculations**] Calculate configuration degrees

$$\forall e = (i,j) \in \tilde{A} \qquad d_e = \frac{k_i^\alpha k_j^\alpha}{k_i^\alpha k_j^\alpha + k_i^\beta k_j^\beta}$$

7. [**Model training**] Train a logistic regression classifier on $\{d_e\}$.
8. [**Reconstruction**] Reconstruct $\tilde{A}$.
9. [**Measure performance**] Measure performance using accuracy, AUROC and AUPR.
10. [**Set-up**] Repeat (3) - (9) for some range of θ.

Results

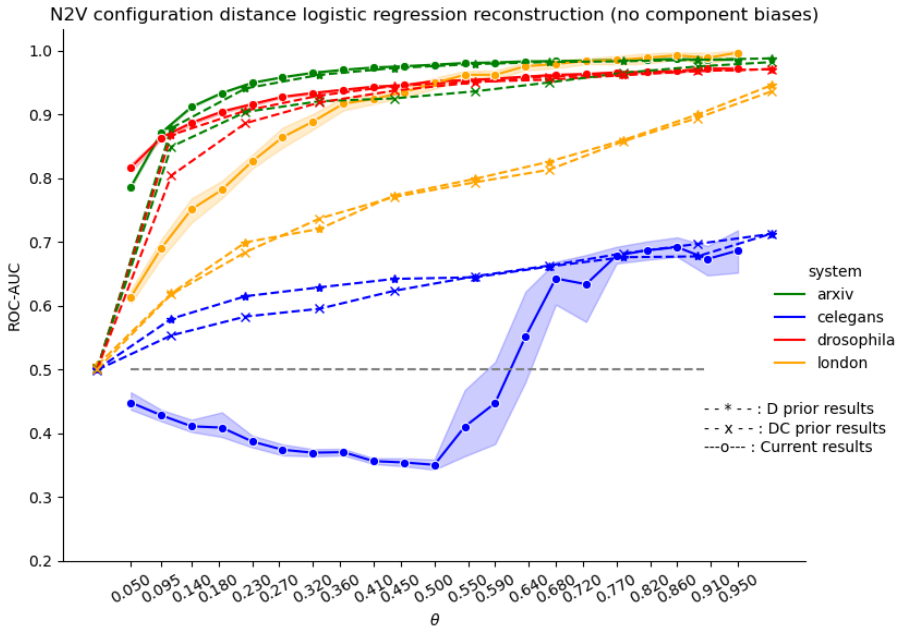

Figure 1: Comparing configuration degree as a logistic regression feature to prior Naive Bayes results.

Analysis

While I currently have ripped the DC results in Figure 1, it nonetheless does not appear that we are reproducing the behavior observed in prior work specifically for the *C. Elegans* connectome. The London transportation network, while seemingly very different, is likely explained by the DC classifier being known to perform worse than pure degree approaches on that system.

Future Work

NOTE: The D classifiers should be ran on the real networks - here I was just ripping the existant pickled results which only included the modified Wu et al. and DC classifiers.

`EMB_ex28` will expand this analysis to include N2V embedding distance as well.