# Comp790-166: Computational Biology

Lecture 18

April 6, 2021

# Announcements

- Homework 2 to be assigned tonight.
- Comments on projects? How are they going?
- Office hours today: please send me an email if you plan to visit office hours

# Today

- Another alignment problem, for single-cell multimodal alignment
- GNN vs Corrections on Simpler Approaches

Multimodal Single-Cell Profiling

- In an ideal world, one would just measure cells via a proteomic technology with flow or mass cytometry (just kidding...)
- In reality, you may want to measure multiple levels of biology with different technologies.
- You likely cannot apply these technologies to the exact set of cells.
- It would be great to be able to map cells to each other according to measurements from different modalities.

# Setup and Assumptions

- Let $X = \{x_1, x_2, \ldots x_M\}$ and $Y = \{y_1, y_2, \ldots y_N\}$ be two datasets measuring points from the some phenomenon (e.g. cells)
- Assume that $X, Y \subset \mathbb{R}^n$.
- In data matrix format, $X \in \mathbb{R}^{M \times n}$ and $Y \in \mathbb{R}^{N \times n}$

# Illustration

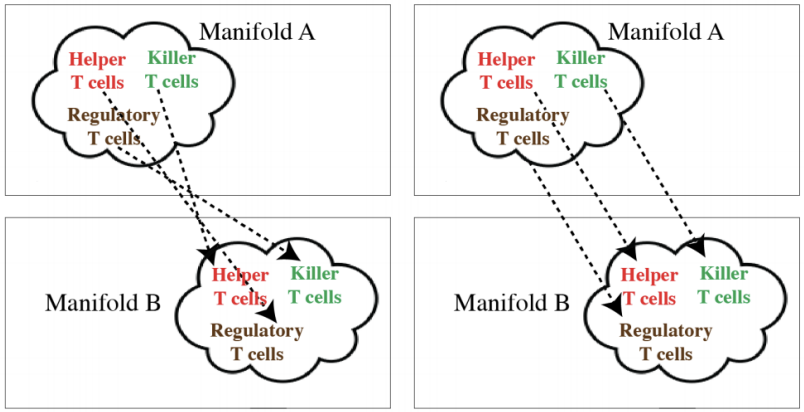



Figure: from
`http://proceedings.mlr.press/v80/amodio18a/amodio18a.pdf`

# Processing on One View

Start with $X \in \mathbb{R}^{M \times n}$ as an example.

- For a pair of rows, $x_i, x_j$, compute some notion of similarity between them.
- Create a row-stochastic probability matrix, $\mathbf{P}$ dividing each similarity bu the sum of similarities.
- Find eigenvalues of $\mathbf{P}$, such that $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$ and corresponding eigenvectors, $\{\phi_j\}_{j=1}^{M}$
- Use eigenvectors and eigenvalues to define diffusion coordinates $(\Phi_t = \{\Phi_t(x_i) : x_i \in X\})$.

# Diffusion Coordinates for one View

Recall that we can map each $x_i$ to diffusion coordinates in terms of eigenpairs as,

$$\Phi_t(x_i) = \left[\lambda_1^t \phi_1(x_i), \ldots, \lambda_N^t \phi_N(x_i)\right]^T$$

- Note that as always $t$ denotes how much we power, $\mathbf{P}$

# Graph Fourier Transform

Use $\mathbf{P}$ to compute the normalized graph Laplacian as,

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{1/2}\mathbf{P}\mathbf{D}^{-1/2}$$

.

- As we know, we can learn a lot from the eigenvectors of the graph Laplacian, and corresponding eigenvalues, (for some $j$, with the eigenvector/eigenvalue pair) $\{\Psi_j, \omega_j\}$
- GFT of a signal at the $j$-th frequency can then be computed as $\widehat{f}[j] = \langle f, \psi_j \rangle$
- Recall that the eigenvectors corresponding to lower eigenvalues ($\omega_j$) correspond to lower frequencies.

# Harmonic Alignment Overview

- Construct two diffusion maps, one for $X$ and another for $Y$.
- These are represented by eigenpairs as $\left(\lambda_i^{(X)}, \phi_i^{(X)}\right), i = 1, \ldots, M$ and $\left(\lambda_j^{(Y)}, \phi_j^{(Y)}\right), j = 1, \ldots, N$ The objective is to align $X$ and $Y$ in terms of diffusion coordinates.
- However, each diffusion component represents some notion of frequency and hence diffusion coordinates between $X$ and $Y$ should only be aligned if their frequencies are similar.

# Bandlimited Correlation

The diffusion harmonics will be binned into local frequency bands with the following window function,

$$w_\xi(\lambda) = \begin{cases} \sin\left(\frac{\pi}{2}\cos^2\left(\frac{\pi}{2}(\ell\lambda - \xi)\right)\right) & \frac{\xi-1}{\ell} \leq \lambda \leq \frac{\xi+1}{\ell} \\ 0 & \text{otherwise} \end{cases}$$

Here, $\ell$ and $\xi$ are some fixed parameters.

# Define Bandlimiting Weights

For $i = 1, \ldots, M$ and $j = 1, \ldots N$, define bandlimiting weights between the diffusion harmonics of $X$ and $Y$ as,

$$w_{ij}^{(X,Y)} = \sum_{\xi=1}^{\ell} w_\xi \left(\lambda_i^{(X)}\right) w_\xi \left(\lambda_j^{(Y)}\right)$$

These weights can then be used to construct an $M \times N$ bandlimited correlation matrix, $C$ as,

$$[C]_{ij} = w_{ij}^{(X,Y)} \operatorname{corr}\left(\phi_i^{(X)}, \phi_j^{(Y)}\right)$$

# Unpacking

$$[C]_{ij} = w_{ij}^{(X,Y)} \operatorname{corr}\left(\phi_i^{(X)}, \phi_j^{(Y)}\right)$$

- Here is ensuring correlations between diffusion harmonics that correspond to similar frequencies.

# GFT to Express Harmonics in Terms of Features

We assume that $X$ and $Y$ each have $n$ measured features, $f_s$, and $g_s$, respectively.

$$\hat{\mathbf{x}}_i = \left(\hat{f}_1[i], \ldots, \hat{f}_n[i]\right)^T$$

and

$$\hat{\mathbf{y}}_j = \left(\hat{g}_1[j], \ldots, \hat{g}_n[j]\right)^T$$

* Reminder GFT for frequency $j$ over some signal, $f$ is $\hat{f}[i] =< f, \phi_j >$

# Writing Correlation as Inner Product

Given the GFTs for $X$ and $Y$ define,

$$\text{corr}\left(\phi_i^{(X)}, \phi_j^{(Y)}\right) = \langle \hat{\mathbf{x}}_i, \hat{\mathbf{y}}_j \rangle$$

With this definition, the entire correlation matrix can therefore be defined as,

$$[C]_{ij} = w_{ij}^{(X,Y)} \langle \hat{\mathbf{x}}_\mathbf{i}, \hat{\mathbf{y}}_\mathbf{j} \rangle$$

# Rigid Alignment Based on $C$

Given $C$, the harmonic alignment can be can be computed as follows. First, define SVD of $C$ as,

$$C = U\Sigma V^T$$

Its nearest orthogonal transformation is further given by

$$\mathbf{T} = UV^T$$

# Writing a Unified Diffusion Map

A unified diffusion map can ultimately be written as,

$$\Phi_t^{(X,Y)} = \begin{bmatrix} \Phi_0^{(X)} & \Phi_0^{(X)}\mathbf{T} \\ \Phi_0^{(Y)}\mathbf{T}^T & \Phi_0^{(Y)} \end{bmatrix} \begin{bmatrix} \Lambda^{(X)} & 0 \\ 0 & \Lambda^{(X)} \end{bmatrix}^t$$

- $\Lambda^{(X)}$ is the diagonal matrix of the diffusion eigenvalues, $\{\lambda_i^{(X)}\}_{i=1}^N$
- You can use $\Phi_t^{(X,Y)}$ to build a new graph in terms of the new defined coordinates

# Pseudocode is Helpful

**Algorithm A.3** function Align$\left(\mathcal{G}^{(X)}, \mathcal{G}^{(Y)}, t, \ell\right) = {\Phi_t}^{(X,Y)}$
Compute and apply an alignment matrix to two diffusion maps.

| | | |
|---|---|---|
| **Input:** | Laplacian eigensystems and degrees | $\mathcal{G}^{(X)}, \mathcal{G}^{(Y)} : \mathcal{G}^{(Z)} = \left\{\Psi^{(Z)}, \Lambda^{(Z)}, D^{(Z)}\right\}$ |
| | Alignment diffusion | $t$ |
| | Alignment band count | $\ell$ |
| **Output:** | Aligned diffusion map | ${\Phi_t}^{(X,Y)}$ |

**for** Z=X,Y **do**
  $\Psi^{(Z)} \leftarrow \Psi^{(Z)} \setminus {\psi_1}^{(Z)}$; $\Lambda^{(Z)} \leftarrow \Lambda^{(Z)} \setminus {\lambda_1}^{(Z)}$
  ${\Phi^0}^{(Z)} \leftarrow {D^{(Z)}}^{1/2} \Psi^{(Z)}$
  $\hat{Z} \leftarrow {\Psi^{(Z)}}^T Z$ ▷ *Graph Fourier transform*
**end for**
$w^{(X,Y)} \leftarrow \text{BandlimitingWeights}\left(\Lambda^{(X)}, \Lambda^{(Y)}, \ell\right)$ ▷ *Alg. A.4*
**for** $i = 2, N_1$ **do**
  **for** $j = 2, N_2$ **do**
    $C\left(i-1, j-1\right) \leftarrow {w_{ij}}^{(X,Y)} \left\langle \hat{X}(i-1,:), \hat{Y}(j-1,:) \right\rangle$ ▷ *Bandlimited correlations*
  **end for**
**end for**
$U, S, V \leftarrow \text{SVD}(C)$
$T \leftarrow UV^T$ ▷ *Orthogonalization Sec. 4.2*

$${\Phi_t}^{(X,Y)} \leftarrow \begin{bmatrix} {\Phi^0}^{(X)} & {\Phi^0}^{(X)}\,\mathbf{T} \\ {\Phi^0}^{(Y)}\,\mathbf{T}^T & {\Phi^0}^{(Y)} \end{bmatrix} \begin{bmatrix} {\Lambda}^{(X)} & 0 \\ 0 & {\Lambda}^{(Y)} \end{bmatrix}^t$$

**return** ${\Phi_t}^{(X,Y)}$

# Recap

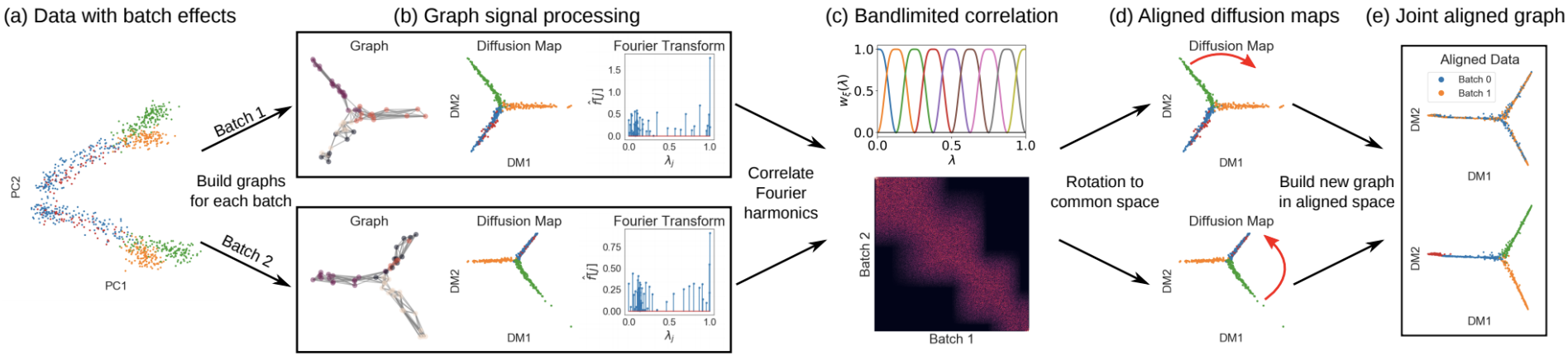



Figure: from Stanley *et al.* SIAM Data Mining. 2020.

*Note this is a different Stanley...

# Application 1: Batch Effect Correction

- CyTOF datasets of PBMCs from patients with dengue.
- In early immune response, there should be changes in IFN$\gamma$ and TFN$\alpha$.
- There should be co-upregulation of these two cytokines.

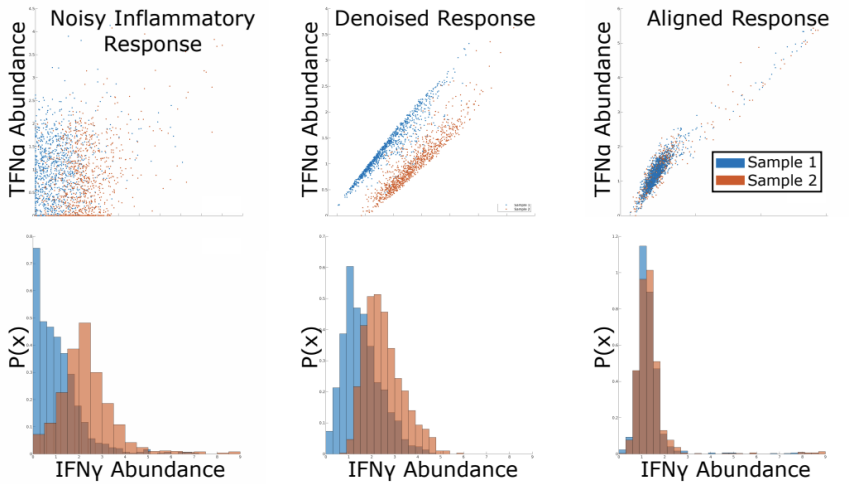



Figure: from Fig. 3. Denoising doesn't remove IFN and TNF relationship.

# Multimodal Data Fusion

A common set of cells were profiled with a gene expression assay and chromatin profiling assay. In the experiment, cells were scrambled and the objective was the quantify overlp in neighborhoods in cell graphs.

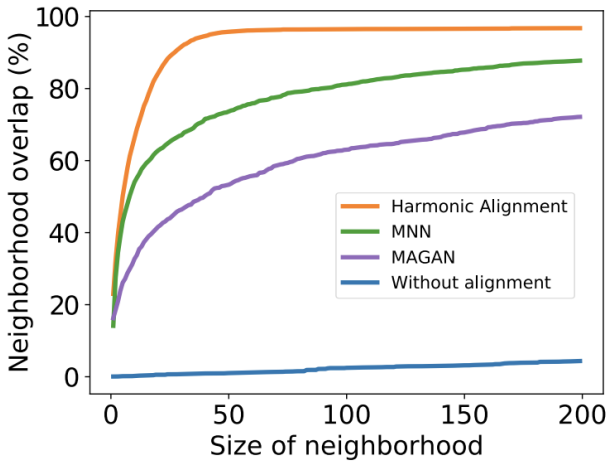



Figure: from Fig. 3

# Conclusion for Harmonic Alignment

- The authors found a set of join diffusion coordinates based on those computed in the individual datasets.
- All of this was done under the assumption that for a pair of coordinates between $X$ and $Y$ to map between datasets, they should have a similar frequency interpretation.
- It can successfully be applied for tasks such as batch effect correction and finding a correspondence between cells profiled with multiple assays.
- Geometric framework via diffusion maps + filtering through GSP

# GNN vs Simple Things

- You all love GNNs
- Recently there has been some work to study why GNNs are outperforming simpler methods and how to incorporate this intuition back to simpler methods

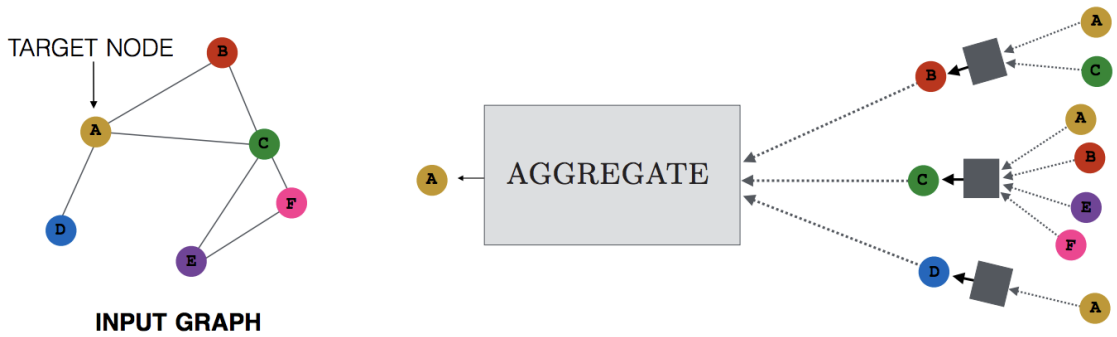


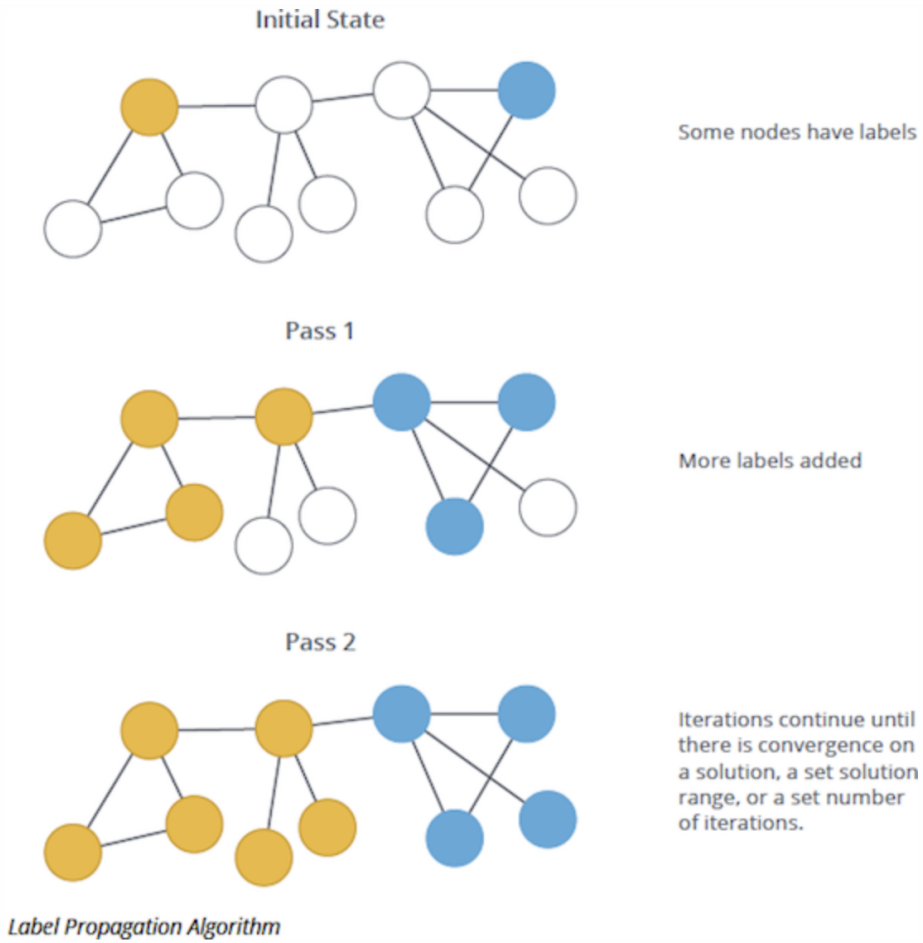

Figure: from `https://www.cs.mcgill.ca/~wlh/grl_book/files/GRL_Book-Chapter_5-GNNs.pdf`. Messages are aggregated from the neighborhood of some target node.

# Correct and Smooth Approach

- The goal is to compare how a couple of simple methods/intuition can be strung together can be used to classify nodes
- The main idea is to start with a cheap base prediction based on node features (e.g. attributes or coordinates of a spectral embedding), and clean up graph structure through label propagation (**correct and smooth**).

# What is Label Propagation?

Some of your nodes are labeled, others are unlabeled, and you predict labels of unlabeled nodes based on the structure of the graph.



# Overview of Correct and Smooth Approach

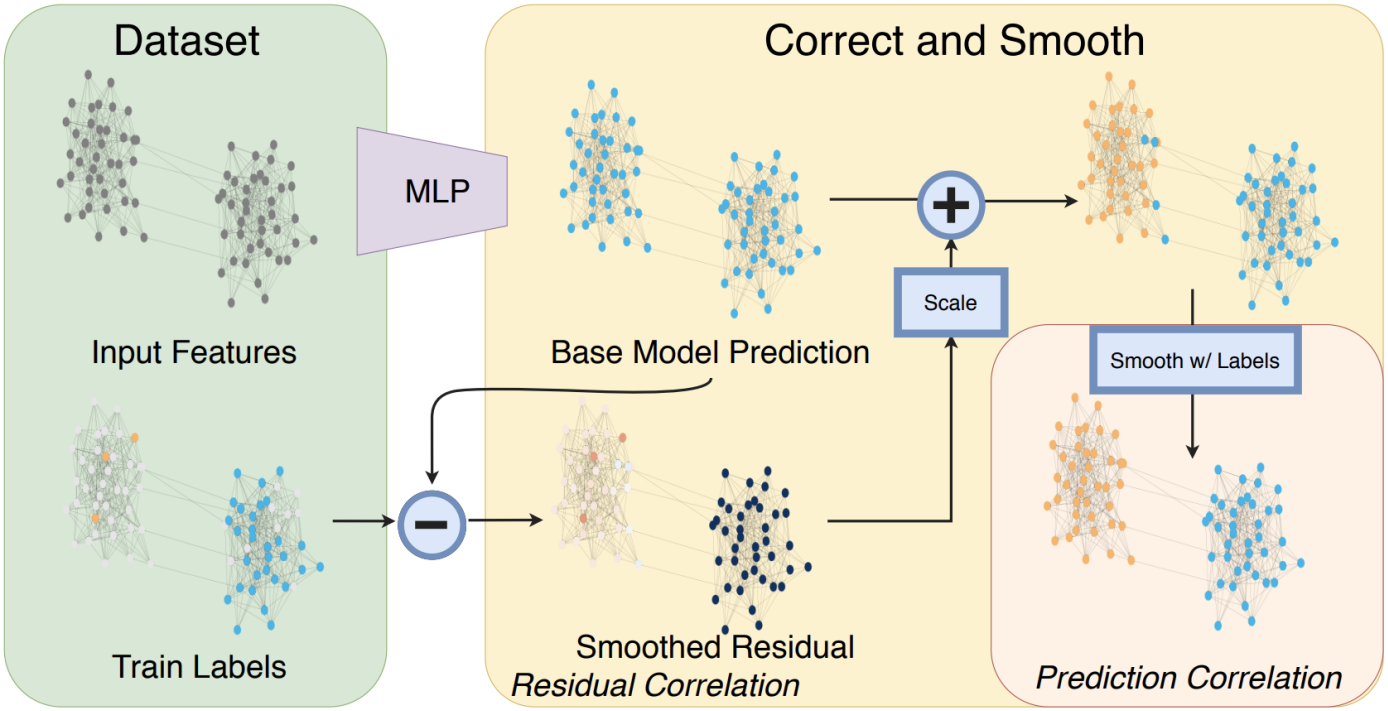



Figure: from Huang *et al.* ICLR. 2021

# Notation Preliminaries

- Let there be $n$ nodes.
- Assume we have a feature vector for each node, such that node features are encodes in an $n \times p$ matrix, $X$.
- Similarly, let $A$ be the adjacency matrix of the graph
- Split nodes into labeled ($L$) and unlabeled ($U$) sets
- Define an $n \times c$ matrix, $Y$ with a binary indicator for whether node $i$ is in class $c$.

# Simple Base Predictor

Given the matrix of features for each node, $X$ and labels, $Y$, train a simple model to minimize,

$$\sum_{i \in L_t} \ell\left(f\left(x_i\right), y_i\right)$$

- Here $L_t$ denotes the set of labeled training nodes
- Specify a matrix, $Z$ containing these base predictions.

# Error Correlation

- The intuition is that errors are expected to be correlated across edges in the graph. Hence, spread uncertainty across the edges.

Define and error matrix, $E \in \mathbb{R}^{n \times c}$ as,

$$E_{L_t,:} = Y_{L_t,:} - Z_{L_t,:}, \quad E_{L_v,:} = 0, \quad E_{U,:} = 0$$

This means that the only non-zero entries are those that correspond to labeled training nodes!

# Smooth the Error Using a Label Spreading Technique

The errors are smoothed as follow with a label spreading technique,

$$\hat{E} = \underset{W \in \mathbb{R}^{n \times c}}{\arg\min}\ \text{trace}\left(W^T(I - S)W\right) + \mu \|W - E\|_F^2$$

- $S$ is the normalized adjacency matrix, $D^{-1/2}AD^{-1/2}$
- The first term encourages smoothness of the error over the graph
- The second term keeps $W$ close to the initial estimate of error, $E$.

# Solution

Given

$$\hat{E} = \underset{W \in \mathbb{R}^{n \times c}}{\arg\min}\ \text{trace}\left(W^T(I - S)W\right) + \mu\|W - E\|_F^2$$

it was previously shown that the solution can be obtained through the following iteration,

$$E^{(t+1)} = (1 - \alpha)E + \alpha S E^{(t)}$$

The quickly converges to $\hat{E}$ and therefore gives corrected predictions as,

$$Z^r = Z + \hat{E}$$

# Smoothing Final Predictions with Prediction Correlation

- The next assumption to be used for correction is that adjacent nodes in the graph are likely to have similar labels (e.g. homophily)
- Another round of label propagation will be used to encourage smoothness over distribution of labels.

Starting with the best guess of the labels, H, with $H_{L_t,:} = Y_{L_t,:}$ and $H_{L_v \cup U,:} = Z^{(r)}_{L_v \cup U,:}$, propagate labels as,

$$H^{(t+1)} = (1 - \alpha)H + \alpha S H^{(t)}$$

# Final Prediction

The following has now been applied

- Base prediction
- Residual correction
- Label smoothing

After convergence of $H^{(t+1)} = (1 - \alpha)H + \alpha SH^{(t)}$, get a final prediction, $\hat{Y} \in \mathbb{R}^{n \times c}$, and assign node to the class with the max predicted probability.

# Results

| Datasets | Classes | Nodes | Edges | Parameter Δ | Accuracy Δ | Time (s) |
|---|---|---|---|---|---|---|
| Arxiv | 40 | 169,343 | 1,166,243 | −84.90% | +0.26 | 12 (+90) |
| Products | 47 | 2,449,029 | 61,859,140 | −93.47% | +1.74 | 171 (+2959) |
| Cora | 7 | 2,708 | 5,429 | −98.37% | +1.09 | < 1 (+7) |
| Citeseer | 6 | 3,327 | 4,732 | −89.68% | −0.69 | < 1 (+7) |
| Pubmed | 3 | 19,717 | 44,338 | −96.00% | −0.30 | < 1 (+14) |
| Email | 42 | 1,005 | 25,571 | −97.89% | +4.33 | 43 (+17) |
| Rice31 | 10 | 4,087 | 184,828 | −99.02% | +1.39 | 39 (+12) |
| US County | 2 | 3,234 | 12,717 | −74.56% | +1.77 | 39 (+12) |
| wikiCS | 10 | 11,701 | 216,123 | −84.88% | +2.03 | 7 (+11) |

Figure: from Table 1. Performance is in reported to SOTA GNN.

# Accuracy vs Number of Parameters

Higher accuracy with less parameters on one of the datasets (and training is also significantly faster)

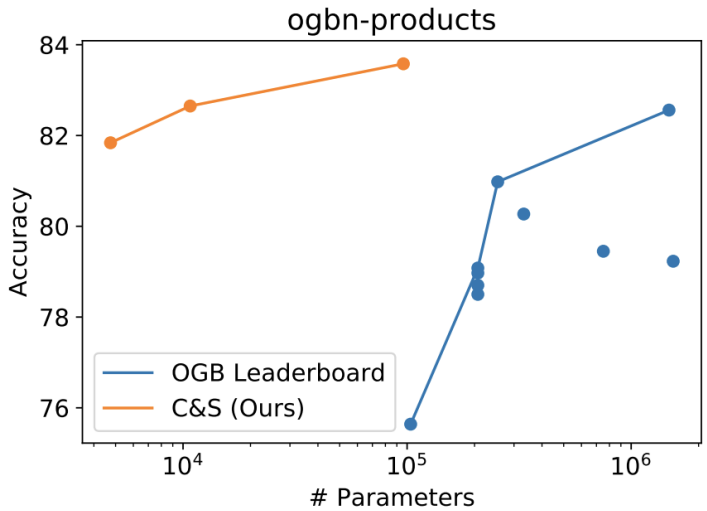
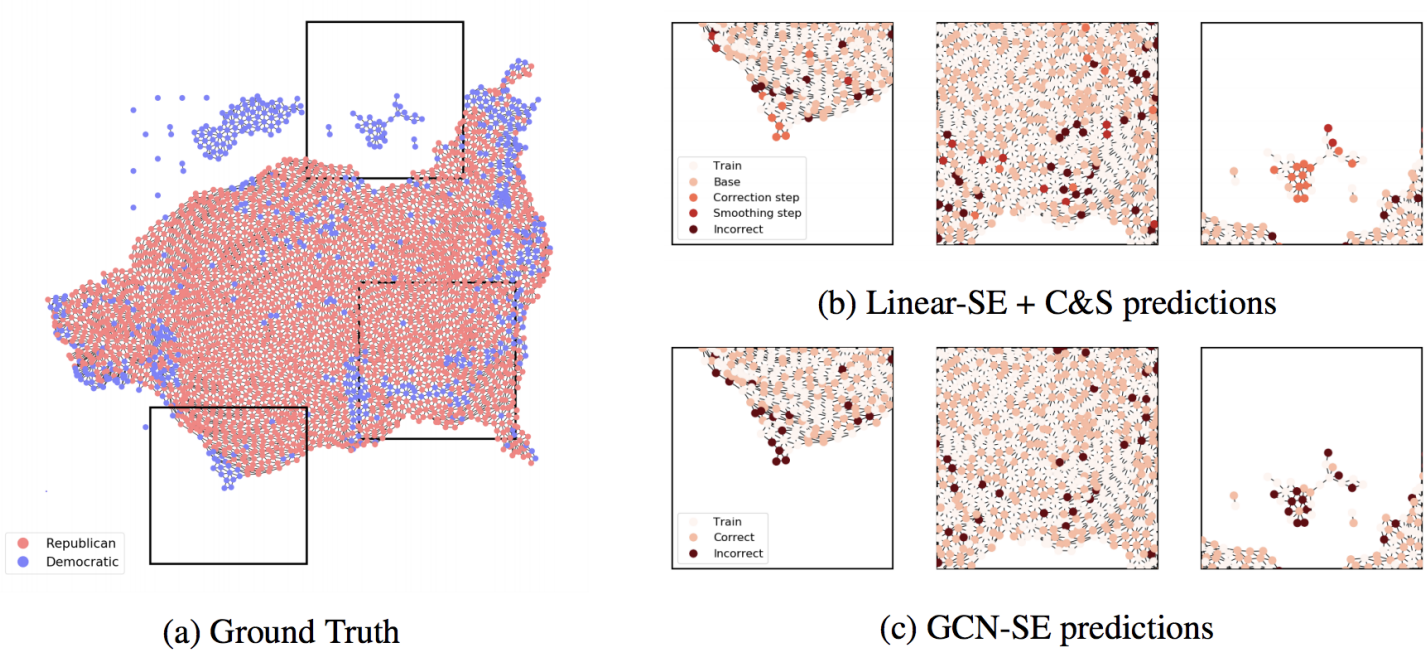



Figure: from Fig. 2

# Visualizing which correction step fixed error




(a) Ground Truth

(b) Linear-SE + C&S predictions

(c) GCN-SE predictions

Figure: from Fig. 3.

# Summary

- Simple LP, diffusion, and GNN are fundamentally related
- Augmenting graph information with attributes, spectral features, etc. can be helpful for classifying nodes
- A base prediction is corrected according to smoothing over residual errors and encouraging closely connected nodes to have similar labels.