# Topological Data Analysis

## IMDB Datasets

Jack Myrick

MATH 4V91 taught by Professor Baris Coskunuzer
The University of Texas at Dallas

December 6, 2022

# Table of Contents

# Background

"Graph classification, or the problem of identifying the class labels of graphs in a dataset, is an important problem with practical applications in a diverse set of fields."

- Bioinformatics
- Chemoinformatics
- Social network analysis
- Urban computing
- Cyber-security

Source: Graph Classification using Structural Attention (Lee et al.)

# Datasets

I will be using the IMDB-BINARY and IMDB-MULTI datasets to classify graphs of actor collaboration by genre.

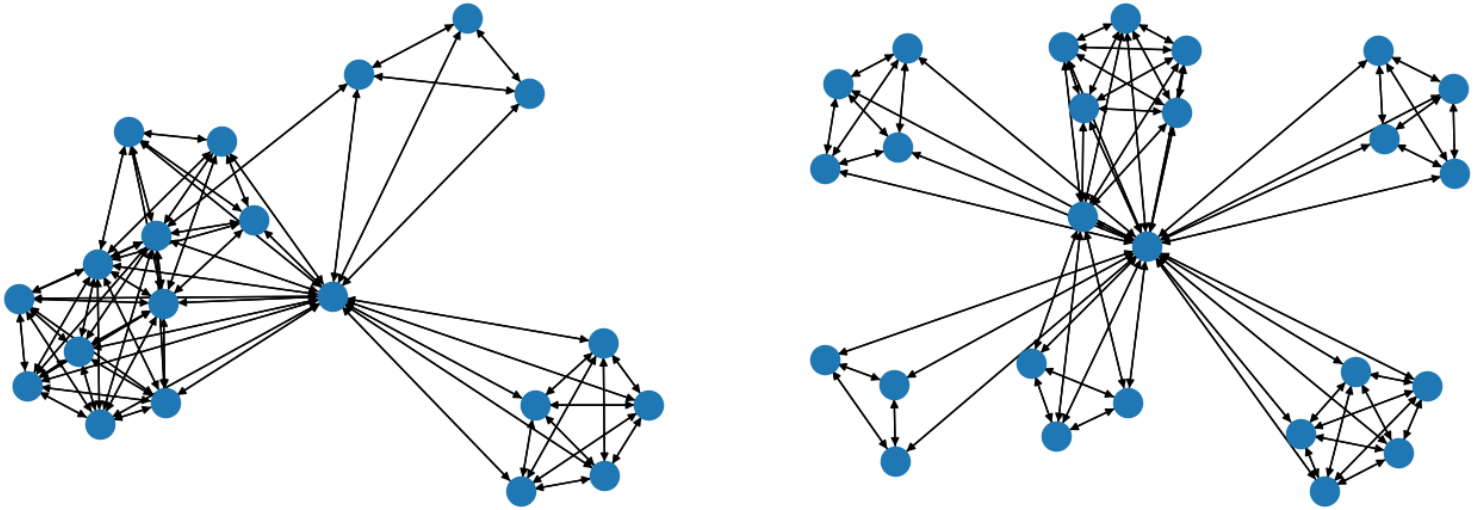

Figure: Sample graphs from IMDB-BINARY with labels 0 and 1 respectively

**Examples**

The IMDB-BINARY dataset has 2 genre classes: Action and Romance.

# Datasets

Classifying graphs is important for a variety of disciplines, but it is a difficult problem since graphs are complex objects.
The IMDB-BINARY and IMDB-MULTI datasets are benchmark graph classification datasets.

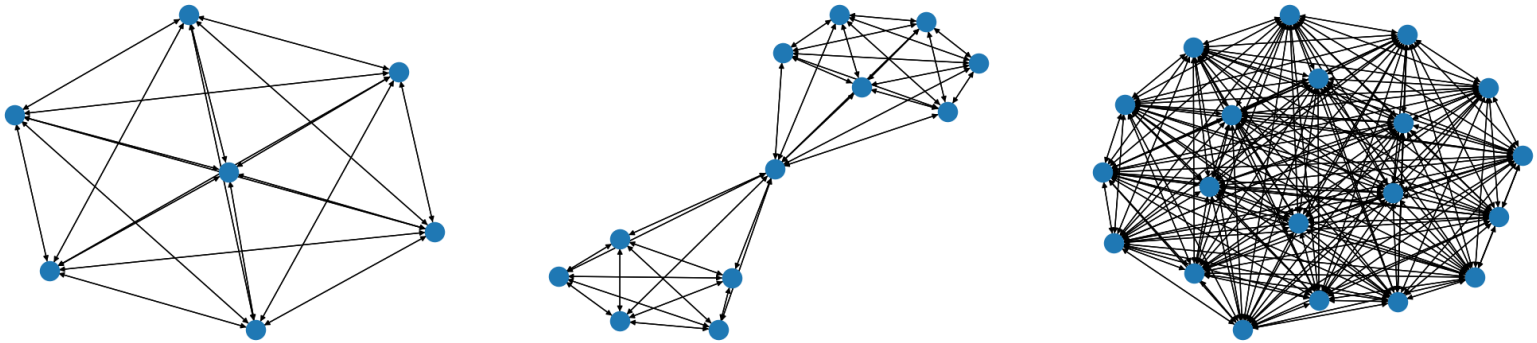

Figure: Sample graphs from IMDB-MULTI with labels 0, 1, and 2 respectively

## Examples

The IMDB-MULTI dataset has 3 genre classes: Comedy, Romance, and Sci-Fi.

# Methods

I utilized a sublevel degree filtration as described in class.

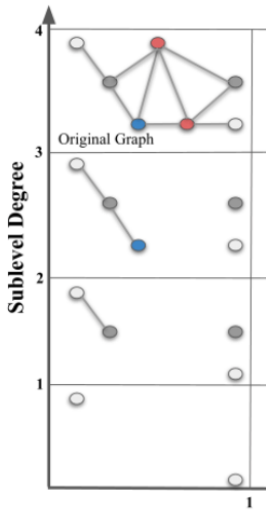


Figure: An example of the sublevel degree filtration as described in class.
Source: ToDD: Topological Compound Fingerprinting in Computer-Aided Drug Discovery (Demir, Coskunuzer, et al.)

However, I got better results by modifying it slightly so that all vertices initially appear in the filtration and only edges are added as the degree parameter increases (by setting diagonal entries to 0).

# Implementation

I used a Vietoris-Rips filtration using the Giotto-tda library and passed in a precomputed metric where each present edge has its value as the maximum of the degrees of its two endpoints. Diagonals were set to 0 and nonexistent edges were set to np.inf.

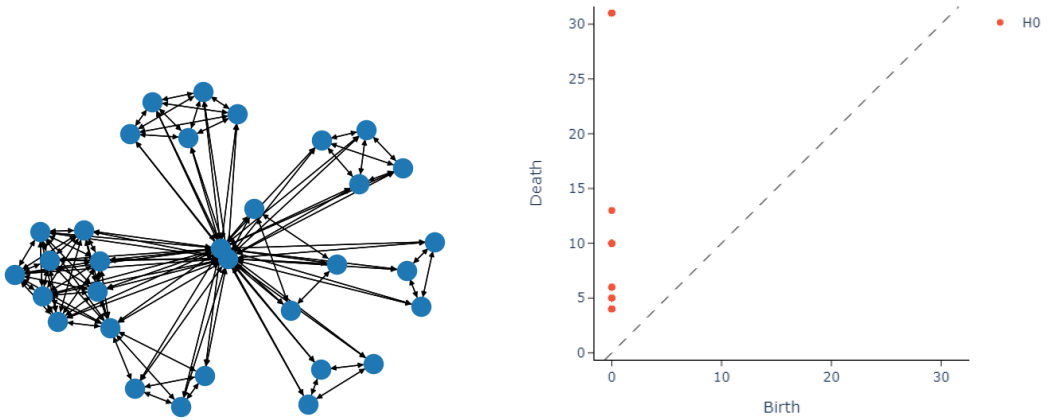


Figure: A graph from IMDB-BINARY (label 0) and its persistence diagram from a sublevel degree filtration

This resulted in a persistence diagram (H0 only) for each graph.

I generated the Betti curves from each persistence diagram using 100 bins.

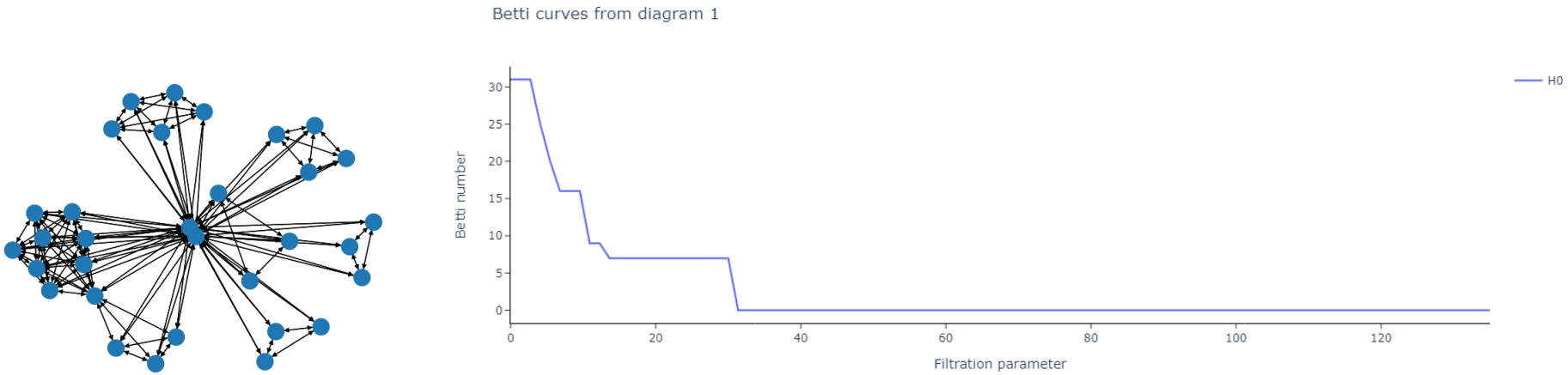


Figure: The same graph from IMDB-BINARY (label 0) and its Betti curve

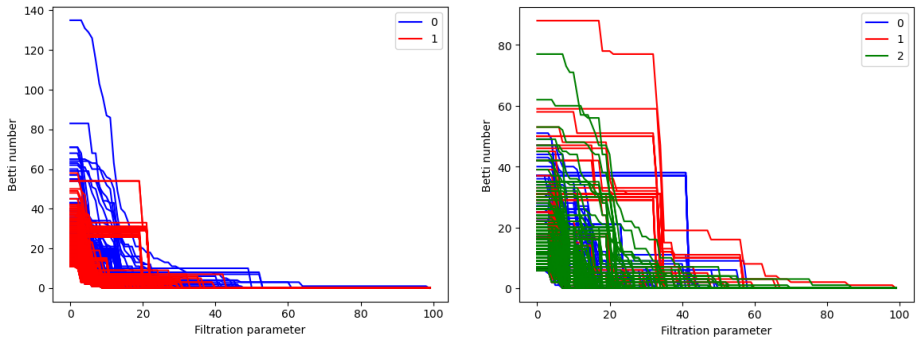


Figure: Betti curves for each graph in IMDB-BINARY and IMDB-MULTI respectively, colored by label

# Training and Results

I trained an XGBoost classifier on the Betti diagrams using an 80-20 train-test split

**Result**

On the IMDB-BINARY dataset, the model achieved an accuracy of .7250. Since the dataset is perfectly balanced between the two genre classes, this measure is appropriate.

**Result**

On the IMDB-MULTI dataset, the model achieved an accuracy of .3833. Since the dataset is perfectly balanced between the three genre classes, this measure is appropriate.

# Comparison: IMDB-BINARY

| Model | Paper | Year | Accuracy |
|---|---|---|---|
| U2GNN (Unsupervised) | Universal Graph Transformer Self-Attention Networks | 2019 | .9641 |
| G_ResNet | When Work Matters: Transforming Classical Network Structures to Graph CNN | 2018 | .7990 |
| **TDA** | **This Presentation** | **2022** | **.7250** |
| GCAPS-CNN | Graph Capsule Convolutional Neural Networks | 2018 | .7169 |
| PSCN | Learning Convolutional Neural Networks for Graphs | 2016 | .7100 |
| DGK | Deep Graph Kernels | 2015 | .6696 |

Table: Top results at their time (and my result) on the IMDB-BINARY dataset
Source: Papers with Code

# Comparison: IMDB-MULTI

| Model | Paper | Year | Accuracy |
|---|---|---|---|
| U2GNN (Unsupervised) | Universal Graph Transformer Self-Attention Networks | 2019 | .8920 |
| DUGNN | Learning Universal Graph Neural Network Embeddings With Aid Of Transfer Learning | 2019 | .5610 |
| G_ResNet | When Work Matters: Transforming Classical Network Structures to Graph CNN | 2018 | .5453 |
| DGCNN | An End-to-End Deep Learning Architecture for Graph Classification | 2018 | .4783 |
| DGK | Deep Graph Kernels | 2015 | .4455 |
| **TDA** | **This Presentation** | **2022** | **.3833** |

Table: Top results at their time (and my result) on the IMDB-MULTI dataset
Source: Papers with Code

# Discussion of Results

**Result**

On the IMDB-BINARY dataset, the model achieved an accuracy of .7250.

**Result**

On the IMDB-MULTI dataset, the model achieved an accuracy of .3833.

Attempts at improvement:

- Filtration function
- Homology dimensions
- PCA
- Number of bins for Betti curves
- Classifier
- Removing outliers

# Conclusion

Thank you for listening to my presentation! Any questions?

## Code

The code used for this project and to generate the figures in this presentation can be found on GitHub.