Deep Learning for Social Sciences

Assignment 3: Graph Neural Networks on Social Media Data

Giordano De Marzo

June 2025

To be handed in by June 25th 23:59

1 Overview

In this assignment, you will explore Graph Neural Networks (GNNs) using real social network data from Twitch. Unlike traditional machine learning approaches that treat data points independently, GNNs can leverage the network structure to improve predictions.

The assignment uses the Twitch Social Network dataset from Stanford SNAP, containing mutual follower relationships between Twitch streamers. You will predict various user attributes and compare GNN performance against traditional MLPs to understand when network structure provides additional predictive power.

Complete tasks for receiving points. Maximum points: 30. Bonus points available but will not exceed 30 total.

1.1 Dataset Information

- **Source**: Stanford SNAP dataset (benedekrozemberczki/datasets)
- **Files**: `large_twitch_features.csv`, `large_twitch_edges.csv`
- **Graph type**: Mutual follower relationships between Twitch users
- **Node features**: views, mature, life_time, created_at, updated_at, language, affiliate, dead_account

Available at: `https://snap.stanford.edu/data/twitch_gamers.html`

2 Tasks

2.1 Data Import, Visualization and Exploratory Analysis (5 points)

- Load both the node features (`large_twitch_features.csv`) and edge list (`large_twitch_edges.csv`)
- Construct the graph using PyTorch Geometric
- Perform exploratory data analysis:
  - Analyze the distribution of node features
  - Compute basic graph statistics (number of nodes, edges, average degree, etc.)
  - Visualize degree distribution
  - Create visualizations showing the distribution of your target variables
  - Analyze feature correlations

2.2 GNN Model Training and Comparison (10 points)

Train and evaluate GNN models for **two classification tasks**:

2.2.1 Task A: High Views Prediction

- **Target**: Binary classification ($\leq$10k vs >10k total views)
- **Hypothesis**: Popular streamers network with other popular streamers (success homophily)
- **Important**: Exclude the `views` feature from input to prevent data leakage
- **Features**: mature, life_time, dead_account, affiliate, language, temporal features

2.2.2 Task B: Language Prediction

- **Target**: Multi-class classification of broadcasting language
- **Hypothesis**: Users follow others who speak their language (language homophily)
- **Features**: views, mature, life_time, dead_account, affiliate, temporal features
- **Important**: The dataset is highly unbalanced, try to implement suitable corrections to mitigate this problem and/or focus on the top five languages.

For each task:

- Implement a **Convolutional GNN** architecture
- Train an **MLP baseline** using the same features (without graph structure)
- Use appropriate train/validation/test splits (e.g. 60%/20%/20%)
- Handle class imbalance (if relevant)
- Report comprehensive metrics
- Include confusion matrices and classification reports

2.3 Report Writing (15 points)

Write a comprehensive report (max 3 pages) that includes:

2.3.1 Structure:

- **Introduction**: Problem description and approach
- **Data Analysis**: Dataset description, preprocessing steps, graph statistics
- **Methods**: Model architectures, training procedures, evaluation metrics
- **Results**: Performance comparison tables, learning curves, confusion matrices
- **Discussion and Conclusions**:
  - In which task do GNNs outperform MLPs the most and why?
  - Interpret the results in the context of social network theory
  - Key findings and implications

2.3.2 Important Elements:

- Performance comparison table (GNN vs MLP for each task)
- Discussion of which feature was hardest to predict and why
- Clear methodology that allows replication

3 Bonus Tasks

You can use up to an additional half page for each bonus task.

3.1 Bonus Task 1: GraphSAGE Implementation (1 bonus point)

- Implement GraphSAGE for one of the two tasks
- Compare its performance with the convolutional GNN

3.2 Bonus Task 2: Regression Task (2 bonus points)

- Train a GNN for regression on the number of views (continuous target)
- Compare with MLP regression baseline
- Use appropriate regression metrics (RMSE, MAE, MAPE)

Note: This assignment reflects real-world social media analysis challenges. Don't expect perfect accuracy - focus on understanding when and why network structure helps predictions.