Predicting Streamer Language from Network Structure in the Twitch Social Graph

Zachary White
Tufts University
COMP 142: Network Science
Spring 2025

1. Introduction

In online platforms such as Twitch, social interactions often reflect shared interests, language, and regional identity. This project investigates whether a Twitch user's broadcast language can be predicted using only the structure of the user-user follow graph. No textual, profile, or behavioral features are used—only the network of mutual follower relationships is considered.

The dataset used is the publicly available *Twitch Gamers* graph from the SNAP repository, consisting of 168,114 nodes (streamers) and 6.8 million undirected edges (mutual follows). Each node contains metadata including broadcast language, view count, and account lifetime. This work treats language classification as a node-level supervised learning problem, evaluating a variety of models to determine how well the underlying network structure encodes language-based community information.

2. Methodology

Dataset Description: The dataset used in this project originates from the *Twitch Gamers* social network, collected in Spring 2018 via the public Twitch API. It consists of a directed graph where each node represents a Twitch user and each edge represents a mutual follow (i.e., bidirectional follower relationship). After filtering for nodes with valid metadata and converting the graph to undirected form, the final dataset contains 168,114 nodes and 6,797,557 edges.

In addition to the graph structure, each user has associated metadata including:

- `language`: The primary broadcast language (e.g., "EN", "FR", "JA")
- `views`: Total view count
- `life_time`: Account age in days
- `affiliate`: Whether the user is a Twitch Affiliate

Challenges:

- *Severe class imbalance*: Approximately 74% of users are labeled as English speakers (EN), making it difficult for standard models to learn minority class boundaries.
- *Sparse features*: The dataset lacks rich content features such as tags, text, or video metadata. Only simple scalar attributes (e.g., view count) are available.
- *Structural ambiguity*: Many users may follow others for reasons unrelated to language (e.g., shared games, content genre), so language is not guaranteed to be topologically well-separated.

Language Grouping: To reduce label fragmentation and align with broader linguistic and geographic patterns, the original language labels were grouped into six families:

- **EN**: English
- **ROMANCE**: French, Spanish, Portuguese, Italian
- **GERMANIC**: German, Dutch, Swedish, Danish, Norwegian
- **SLAVIC**: Russian, Polish, Czech, Hungarian
- **ASIAN**: Japanese, Korean, Chinese, Thai
- **OTHER**: Turkish, Finnish, and all remaining low-count labels

This mapping provided more balanced group sizes and better interpretability of class performance.

Graph Embeddings: To convert the graph structure into fixed-length vectors suitable for downstream classification, 64-dimensional **Node2Vec** embeddings were generated. Node2Vec is a scalable embedding technique that uses biased random walks to capture both homophily (nodes connected to similar nodes) and structural roles (nodes with similar connection patterns, even if far apart). Parameters were tuned to balance local and global structural capture (walk length = 10, context size = 5).

Classification Models: The classification task is to predict the `language group` of each node using only structural information (embeddings) and, in some cases, limited metadata. The following modeling approaches were implemented:

- **XGBoost (baseline):** A robust gradient-boosted decision tree model trained on the Node2Vec embeddings. Despite its non-graph nature, it performed well due to the quality of the embeddings and its ability to capture nonlinear feature interactions.
- **Two-Stage Hierarchical Classifier:** Given the dominance of English speakers, a two-step model was introduced. The first stage performs a binary classification: `EN` vs `Non-EN`. The second stage, trained only on Non-EN nodes, classifies among the remaining five language groups. This architecture mitigates the masking effect of the dominant class.
- **GCN (Graph Convolutional Network):** A graph-based neural network that learns by aggregating features from neighboring nodes. Inputs were limited to node-level scalar features (e.g., `views`, `lifetime`), and the network was trained with class-weighted loss to account for imbalance. The GCN had limited success due to sparse node features and strong class imbalance.

- **XGBoost + SMOTE:** To address imbalance in a principled way, SMOTE (Synthetic Minority Oversampling Technique) was applied to the Node2Vec embeddings. This generated synthetic samples for underrepresented classes in the training set. The balanced dataset was then used to train a new XGBoost model, which achieved better macro-level performance at a slight cost in accuracy.

Unsupervised Evaluation: To understand whether language groups naturally form structural communities, the **Louvain algorithm** was used to detect communities without labels. The resulting partitions were compared to the known language groups using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI).

Visualization: To assess clustering and separability of language groups in the embedded space, t-SNE (t-distributed stochastic neighbor embedding) was used to project the Node2Vec vectors to two dimensions. For clarity, a balanced sample (up to 800 users per class) was used in the visualization.

Evaluation: Models were evaluated using accuracy, macro F1 score, and class-wise precision/recall. Confusion matrices were generated to visualize class-specific performance. To further assess structural alignment, Louvain community detection was applied to the graph and compared to language labels using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI).

3. Results

Node2Vec + XGBoost achieved strong performance, with overall accuracy of **90.2%** and macro F1 between **0.75**. Even without node metadata, this approach successfully inferred broadcast language based on structural role in the social graph. A two-stage classifier further improved macro F1 by reducing confusion between dominant (EN) and minority classes.

Two-Stage Classifier Performance: The hierarchical two-stage classifier achieved improved macro-level performance by decoupling the dominant EN class from the minority classes. This model improved macro F1 significantly at the cost of accuracy, landing at accuracy **86.7%** and macro F1 of **0.88**

SMOTE-enhanced XGBoost improved fairness by synthetically oversampling underrepresented classes before training. While accuracy decreased slightly, the macro F1 increased significantly, demonstrating more equitable classification across all language families. Overall proved to be the best model with accuracy **90.1%** and macro F1 of **0.88**

Balanced GCN, trained with reweighted loss, improved over the baseline GCN but remained less effective than embedding-based models. Its performance plateaued at **37% accuracy** with macro F1 of **0.25**, indicating limitations due to weak input features and class imbalance.

Louvain Clustering revealed weak unsupervised alignment between communities and language groups: **NMI = 0.1586**, **ARI = 0.0411**. This suggests that language influences—but does not dominate—community formation in the Twitch network.

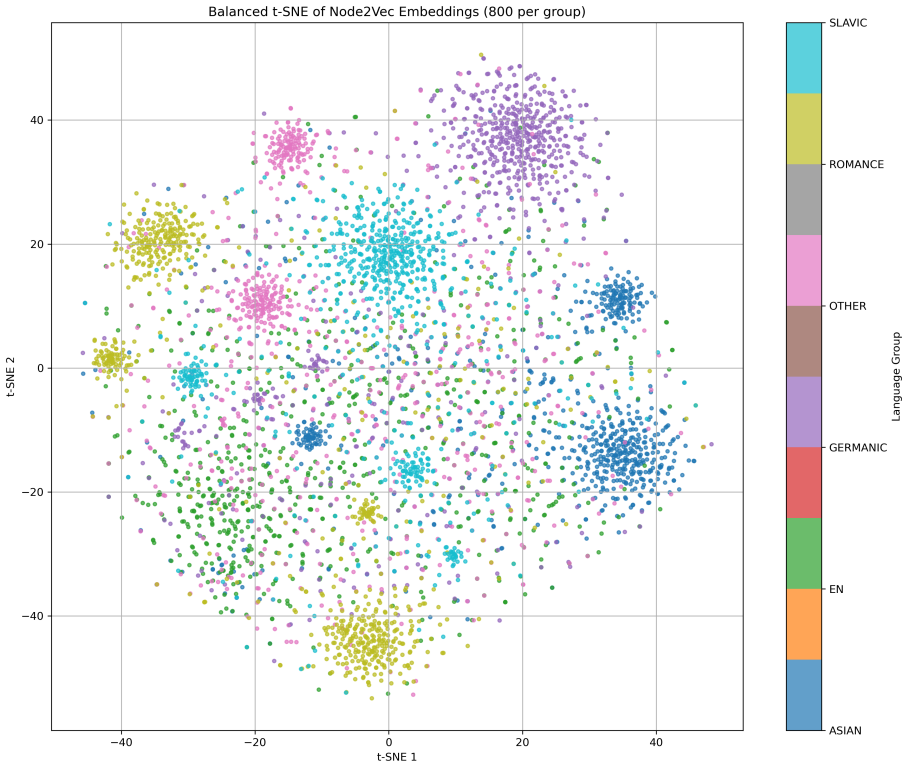

Figure 1: t-SNE of Node2Vec embeddings sampled evenly across language groups

4. Discussion and Analysis

This study confirms that broadcast language is meaningfully reflected in the structure of the Twitch social graph. Although the graph contains no explicit content or geographic metadata, supervised models trained solely on structural information—namely Node2Vec embeddings—were able to achieve strong predictive performance. The XGBoost model achieved over 87% accuracy, and performance remained robust under class rebalancing and hierarchical classification.

Language Group Structure in the Graph: t-SNE visualizations of the Node2Vec embedding space revealed a rich structural landscape. English-speaking users (**EN**), who dominate the platform, appear widely dispersed in the embedding space. Rather than forming a single dense cluster, EN nodes are spread throughout the graph, often connecting to users of other language groups. This supports the hypothesis that English users serve as *structural bridge nodes*—high-degree or well-connected users who link otherwise segregated communities. This is likely due to the global reach of English-language content and its use as a lingua franca in gaming and entertainment.

In contrast, **ASIAN language users** (e.g., Japanese, Korean, Chinese speakers) form highly compact and isolated clusters. These users tend to follow and be followed by others in the same language group, resulting in high intra-community density and low inter-group connectivity. This suggests stronger cultural or regional cohesion, and possibly language barriers that limit outward engagement.

Between these two extremes lie the **ROMANCE** and **GERMANIC** groups. These communities

are more internally cohesive than English users, yet maintain moderate connectivity to other groups. For instance, Spanish and Portuguese-speaking users may interact with English users or with each other due to regional proximity and bilingualism. These groups tend to form semi-permeable clusters—visible in the embedding space but with fuzzy boundaries.

Quantitative Results and Imbalance: While overall classification accuracy was high, initial models struggled with fairness due to class imbalance: over 74% of users in the dataset were English speakers. Without correction, classifiers defaulted to predicting EN for most users, leading to poor macro-level performance. The use of SMOTE oversampling and a two-stage classifier architecture significantly improved recall and F1 scores for minority classes such as **SLAVIC**, **ASIAN**, and **OTHER**.

Even though the Graph Convolutional Network (GCN) was able to leverage topological information directly, it underperformed compared to embedding-based approaches. This may be due to the limited node-level features and the global sparsity of class labels. However, applying class-weighted loss improved the GCN's macro F1 score from near-zero to 0.25, showing the importance of rebalancing techniques in graph-based learning.

Community Detection and Interpretation: To evaluate whether language-based communities naturally emerged in the network, Louvain clustering was applied to the undirected graph. However, the unsupervised communities showed only weak alignment with language labels, with a Normalized Mutual Information (NMI) of 0.1586 and Adjusted Rand Index (ARI) of 0.0411. This indicates that while language influences structure, it is not the sole organizing principle. Other factors—such as shared games, genre-specific fandoms, or geographic time zones—likely play a role in mutual follower dynamics.

Implications and Interpretability: The ability to infer language from structural patterns alone has meaningful implications. Platforms like Twitch could use this type of model to enhance:

- **Language-aware content recommendations**, improving user discovery without relying on potentially noisy self-declared tags.
- **Multilingual community identification**, highlighting users who act as cultural bridges or span multiple regions.
- **Cohesion and diversity metrics**, useful for analyzing engagement across linguistic boundaries and monitoring how different language groups interact over time.

It is worth noting that I was biased toward improving macro F1 over true accuracy, as in most algorithms tied to race/language and their applications, it is very important to consider the implications of a certain classification. A user in the ASIAN group may suffer disproportionately if misclassified, as ASIAN users seem to avoid users of other language groups, while misclassifying an ENGLISH user may not be as drastic, as their users tend to be more dispersed across the network.

Overall, the results suggest that network topology encodes more than just social proximity—it captures latent sociolinguistic structures that can be recovered through the right embedding and modeling strategies. These findings demonstrate the potential of graph-based machine learning to extract meaningful insight from social platforms, even in the absence of content or user-declared preferences.

Future Extensions:

- Incorporate richer metadata (stream categories, activity timestamps, user region)
- Use temporal snapshots to analyze language drift and evolving communities
- Explore dynamic GNNs and temporal node embeddings
- Apply alternative unsupervised methods (Leiden, etc.) for comparative community detection

6. Conclusion

This project confirms that broadcast language can be reliably predicted from graph structure alone. While unsupervised clustering does not fully recover linguistic communities, supervised models—especially Node2Vec with XGBoost—offer strong performance without relying on content or behavior. It was an interesting challenge to draw conclusions from such an unbalanced dataset, and I learned a lot about how to deal with these barriers. Language shapes community boundaries, and with the right modeling approach, those boundaries can be meaningfully inferred from network data.