Predicting Movie Genre Using Actor Co-Appearance Networks

Avi Herman
Department of Computer Science
Social Networks Final Project — Fall 2025

Abstract

Actor collaboration patterns form highly structured networks rather than random linkages. Certain groups of actors repeatedly appear together within specific stylistic or production niches—horror ensembles, action–adventure circles, family-film actors, prestige drama clusters. If this structure is strong enough, genre should be predictable **using network topology alone**, with no plot, script, text, or semantic features.

This paper evaluates whether actor co-appearance networks encode genre information by analyzing modularity Q, PageRank distributions, Gini coefficients G, community purity, and prediction accuracy across **three dataset scales: 250, 1,000, and 5,000 films**. Results demonstrate that genre is partially recoverable from network structure alone, achieving **61.0% accuracy** at the largest scale—over three times the random baseline of 20%. Furthermore, collaboration networks exhibit increasing inequality and realistic power-law structure as sample size grows.

1. Introduction

1.1 Motivation

The film industry is not a random system. Directors repeatedly cast the same actors; production companies specialize in particular genres; franchise ecosystems create dense collaboration clusters. These patterns suggest that **genre may be an emergent property of network structure** rather than merely a content-based label.

1.2 Hypothesis

> **H_1:** Actor collaboration graphs exhibit strong community structure, and these communities correspond to film genres at rates significantly above random assignment.

1.3 Conceptual Framework

Actors specializing in similar film ecosystems tend to co-appear repeatedly. These recurring interactions create dense subgraphs that should map onto genre clusters if genre is a **socially emergent structure** within the film industry. We test this by:

1. Constructing weighted co-appearance networks
2. Detecting communities via modularity optimization
3. Evaluating whether detected communities align with known genres

1.4 Falsification Criteria

The hypothesis is rejected if:

Criterion	Threshold
Modularity	$Q < 0.3$ across all configurations
Community purity	No dominant genre per community
Prediction accuracy	Converges to random baseline ≈ 0.20

2. Theoretical Background

2.1 Network Construction

We model the actor collaboration space as an undirected weighted graph $G = (V, E, w)$:

$$V = \{a_1, a_2, \dots, a_n\} \quad \text{(set of actors)}$$

$$E = \{(a_i, a_j) : \exists \text{ film } f \text{ where both } a_i, a_j \in \text{cast}(f)\}$$

2.1.1 Edge Weight Function

To prioritize genre-specific collaborations, we apply a **genre-purity weighting scheme**. For a film f with genre set G_f:

$$w(f) = \begin{cases} 1.00 & \text{if } |G_f| = 1 \quad \text{(single-genre film)} \\ 0.25 & \text{if } |G_f| = 2 \quad \text{(dual-genre)} \\ 0.05 & \text{if } |G_f| = 3 \quad \text{(triple-genre)} \\ 0.01 & \text{if } |G_f| \geq 4 \quad \text{(multi-genre)} \end{cases}$$

This step-function weighting strongly penalizes multi-genre films:

Genres	Weight	Rationale
1	1.00	Single-genre films are strong genre indicators
2	0.25	Dual-genre films contribute 1/4 weight
3	0.05	Triple-genre films contribute minimally
4+	0.01	Multi-genre films are nearly ignored

Intuition: A collaboration in a pure horror film is much stronger evidence of "horror actor" than a collaboration in a film tagged Horror/Comedy/Romance/Drama.

The final edge weight between actors a_i and a_j aggregates over all shared films:

$$W_{ij} = \sum_{f \in \mathcal{F}(a_i, a_j)} w(f)$$

where $\mathcal{F}(a_i, a_j)$ is the set of films featuring both actors.

2.2 Community Detection: Motivation and Method

2.2.1 Why Communities?

The central premise of this work is that **genre is encoded in collaboration structure**. But how do we extract this signal? The answer is *community detection*—the unsupervised identification of densely connected subgroups.

Key insight: If actors cluster by genre, then:

- Actors within the same community share films more often than expected by chance
- Each community should exhibit a dominant genre
- Community membership can serve as a genre *prediction*

Without community detection, we would have no way to partition the network and test whether partitions correspond to genres.

2.2.2 The Louvain Algorithm

We employ the **Louvain method** [Blondel et al., 2008], a greedy modularity optimization algorithm that iteratively maximizes:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where:

Symbol	Definition
A_{ij}	Adjacency matrix entry (edge weight between i and j)
$k_i = \sum_j A_{ij}$	Weighted degree of node i
$m = \frac{1}{2} \sum_{ij} A_{ij}$	Total edge weight in the network
c_i	Community assignment of node i
$\delta(c_i, c_j)$	Kronecker delta: 1 if $c_i = c_j$, else 0

Interpretation of Q:

- $Q = 0$: No community structure (random)
- $Q \approx 0.3–0.7$: Significant community structure
- $Q > 0.7$: Strong, well-defined communities

The Louvain algorithm proceeds in two phases:

1. **Local optimization:** Each node greedily joins the community maximizing ΔQ
2. **Aggregation:** Communities become super-nodes; repeat until convergence

2.2.3 Why Many Communities?

The algorithm typically produces **more communities than the 5 macro-genres**. This is expected and informative:

Observation	Explanation
$\lvert\mathcal{C}\rvert > 5$	Sub-genre specialization (e.g., slasher vs. supernatural horror)
$\lvert\mathcal{C}\rvert \gg 5$	Franchise isolation (MCU actors form their own cluster)
Small communities	Niche production circuits (e.g., low-budget regional films)
Large communities	Mainstream genre pools with many collaborations

The number of communities $|\mathcal{C}|$ is determined by modularity optimization, not preset. This allows the algorithm to discover natural structure at multiple resolutions.

2.2.4 What Do Communities Contain?

Each community $c \in \mathcal{C}$ is a set of actors:

$$c = \{a_{i_1}, a_{i_2}, \dots, a_{i_k}\}$$

For genre analysis, we characterize each community by its **genre distribution**:

$$P_c(g) = \frac{\sum_{a \in c} \mathbb{1}[g^*(a) = g]}{|c|}$$

where $g^*(a)$ is actor a's dominant macro-genre. This gives a probability distribution over genres for each community.

2.3 Community Purity

2.3.1 Definition

Purity measures how homogeneous a community is with respect to genre. A pure community contains actors of predominantly one genre; an impure community is a mixture.

For a single community c, purity is the fraction of members matching the dominant genre:

$$\text{Purity}(c) = \max_{g \in \mathcal{G}} P_c(g) = \max_{g \in \mathcal{G}} \frac{|\{a \in c : g^*(a) = g\}|}{|c|}$$

2.3.2 Overall Purity (Weighted)

To aggregate across all communities, we weight by community size:

$$\text{Purity}_{\text{overall}} = \frac{1}{|V|} \sum_{c \in \mathcal{C}} |c| \cdot \text{Purity}(c) = \frac{1}{|V|} \sum_{c \in \mathcal{C}} \max_{g} |\{a \in c : g^*(a) = g\}|$$

This is equivalent to **unweighted accuracy** when we assign each community its majority genre.

2.3.3 Interpretation

Purity	Interpretation
0.20	Random (5 genres, no structure)
0.40–0.50	Weak genre clustering
0.60–0.70	Moderate genre separation
0.80+	Strong genre-based communities

Why purity matters: High purity indicates that Louvain communities align with genre boundaries—the network *knows* about genre even though genre labels were never used in community detection.

2.3.4 Purity vs. Accuracy

These metrics are related but distinct:

Metric	What it measures
Purity	How homogeneous each community is
Accuracy	How well community assignment predicts individual actors

With PageRank-weighting, accuracy can exceed raw purity because central actors (who define the community) tend to have cleaner genre profiles than peripheral actors.

2.4 Centrality: PageRank

To identify influential actors within the network, we compute **PageRank** [Page et al., 1999]:

$$\mathrm{PR}(a_i) = \frac{1-d}{N} + d \sum_{a_j \in \mathcal{N}(a_i)} \frac{W_{ji}}{\sum_k W_{jk}} \cdot \mathrm{PR}(a_j)$$

where:

Symbol	Definition
$d = 0.85$	Damping factor
N	Total number of actors
$\mathcal{N}(a_i)$	Neighbors of actor a_i
W_{ji}	Edge weight from a_j to a_i

PageRank captures the notion that an actor is important if they collaborate with other important actors, weighted by the strength of those collaborations.

2.5 Inequality Measurement: Gini Coefficient

To quantify the distribution of centrality (hub formation), we compute the **Gini coefficient** over PageRank scores:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i} = \frac{2\sum_{i=1}^{n} i \cdot x_{(i)}}{n\sum_{i=1}^{n}x_{(i)}} - \frac{n+1}{n}$$

where $x_{(i)}$ denotes the i-th smallest PageRank value.

Interpretation:

- $G \approx 0$: Perfect equality (all actors equally central)
- $G \approx 0.2\text{–}0.3$: Weak hub structure
- $G \approx 0.5\text{–}0.7$: Strong power-law structure (realistic social network)
- $G \to 1$: Extreme inequality (single dominant hub)

2.6 Prediction Accuracy Metrics

We evaluate three voting schemes for assigning genres to communities:

2.6.1 Unweighted Accuracy

Each actor contributes equally:

$$\text{Acc}_{\text{unweighted}} = \frac{1}{|V|}\sum_{a \in V} \mathbb{1}\left[\hat{g}(c_a) = g^*(a)\right]$$

where $\hat{g}(c)$ is the majority genre in community c and $g^*(a)$ is actor a's true dominant genre.

2.6.2 Degree-Weighted Accuracy

Weights by collaboration volume:

$$\text{Acc}_{\text{degree}} = \frac{\sum_{a \in V} k_a \cdot \mathbb{1}\left[\hat{g}(c_a) = g^*(a)\right]}{\sum_{a \in V} k_a}$$

2.6.3 PageRank-Weighted Accuracy

Weights by network influence:

$$\text{Acc}_{\text{PR}} = \frac{\sum_{a \in V} \text{PR}(a) \cdot \mathbb{1}\left[\hat{g}(c_a) = g^*(a)\right]}{\sum_{a \in V} \text{PR}(a)}$$

Rationale: Core actors (high PageRank) more reliably represent their community's genre identity than peripheral actors.

---

3. Data and Methodology

3.1 Dataset

We use filtered IMDb data from `title.basics.tsv` and `title.principals.tsv`:

Filter	Criterion
Media type	Feature films only
Runtime	≥ 59 minutes
Release year	≥ 1960
Cast filter	Actors and actresses only
Billing	Top-3 billed cast members
Actor threshold	≥ 3 film credits
Genre whitelist	18 canonical IMDb genres

3.2 Macro-Genre Mapping

To reduce label noise, we collapse 18 genres into **5 macro-genres**:

Macro-Genre	Original Genres
ACTION	Action, Adventure, Thriller, Sci-Fi, Western, War
DRAMA	Drama, Romance, Biography, History
COMEDY	Comedy, Music, Musical
DARK	Crime, Mystery, Horror
FAMILY	Family, Animation, Fantasy, Sport

3.3 Experimental Design

For each sample size $N \in \{250, 1000, 5000\}$:

1. Select top-N films by popularity (vote count)
2. Construct weighted actor network
3. Vary genre count from 3 to 12
4. For each configuration:
    - Compute Louvain partition (weighted and unweighted)
    - Compute PageRank and Gini coefficient
    - Evaluate all three accuracy metrics
5. Generate diagnostic visualizations

---

4. Results

4.1 Network Inequality Scales with Dataset Size

PageRank distributions reveal the emergence of hub structure:

PageRank Distribution Comparison

Figure 1: PageRank distributions across sample sizes (log-log scale). Larger datasets exhibit heavier tails indicative of

power-law structure.

Sample Size	Actors	Gini G	Interpretation
250	415	**0.282**	Weak inequality; no dominant hubs
1,000	1,174	**0.413**	Moderate hub formation
5,000	4,124	**0.440**	Clear scale-free tail; strong hubs

Table 1: Gini coefficients measuring PageRank inequality.

Gini Across Samples

Figure 2: Gini coefficient increases with sample size, indicating realistic power-law structure.

Key Finding: As sample size grows, the network increasingly resembles real-world social networks with a few high-centrality actors (stars, franchise leads) dominating co-appearance patterns.

4.2 Modularity Demonstrates Strong Community Structure

Weighted modularity consistently exceeds unweighted, confirming that genre-purity weighting enhances community detection:

Modularity (Original Genres)

Figure 3: Modularity vs. number of genres (5,000-film dataset). Weighted Q remains high even with many genre categories.

Sample Size	Peak Q_{weighted}	Q at 12 genres
250	**0.919**	0.919
1,000	**0.845**	0.829
5,000	**0.755**	0.734

Table 2: Weighted modularity values. All exceed 0.7, indicating strong community structure.

Key Finding: Actor networks are highly modular at every scale. Even with 12 genre categories and 5,000 films, $Q > 0.73$.

4.3 Confusion Matrix Analysis

Confusion Matrix

Figure 4: Confusion matrix (250 films). Rows = predicted community; columns = actor's true macro-genre.

Observations:

- **ACTION** is sharply separated (high diagonal concentration)
- **DARK** overlaps with ACTION (crime–thriller–action triad)
- **COMEDY** and **FAMILY** are cleanly partitioned
- **DRAMA** is diffuse and blends with all categories (expected: Drama is a meta-genre)

4.4 Genre Co-Occurrence Structure

Genre Co-Occurrence Matrix

Figure 5: Genre co-occurrence matrix. Cell (i, j) = percentage of genre-i films also tagged with genre-j.

Key Structures:

- **ACTION–ADVENTURE–THRILLER** triad (high mutual co-occurrence)
- **FAMILY–ANIMATION** cluster
- **COMEDY** avoids **DARK** and **ACTION**
- **DRAMA** connects to nearly everything (hub genre)

This explains why adding more genres reduces modularity: multi-genre films create cross-community bridges.

4.5 Network Visualization

Actor Network

Figure 6: Actor co-appearance network (5,000 films). Node color = Louvain community. Node size = weighted degree.

Communities form visually distinct clusters with characteristic genre signatures. Bridge actors (high-degree nodes spanning clusters) often work across multiple genres.

4.6 Genre Co-Occurrence Chord Diagram

Genre Co-Occurrence Chord Diagram

Figure 7: Genre co-occurrence chord diagram (5,000 films). Circular layout showing which genres frequently appear together in the same movie. Arc thickness and color intensity indicate co-occurrence frequency.

The chord diagram provides an intuitive visualization of genre relationships. Genres positioned close together with thick connecting arcs frequently co-occur, while genres with few or no connections rarely appear together. This directly visualizes why certain genre pairs (e.g., Action–Thriller) form natural clusters in the actor network.

4.7 The Central Result: PageRank-Weighted Accuracy

Having established that the network exhibits strong community structure (§4.2), realistic inequality (§4.1), and interpretable genre patterns (§4.3–4.4), we now present the key quantitative result: **prediction accuracy**.

Accuracy Across Samples

Figure 8: PageRank-weighted accuracy vs. dataset size. Even at the largest scale, prediction quality remains far above random baseline.

Sample Size	$\mathrm{Acc}_{\mathrm{unweighted}}$	$\mathrm{Acc}_{\mathrm{degree}}$	$\mathrm{Acc}_{\mathrm{PR}}$	Random Baseline	Lift
250	0.720	0.696	**0.760**	0.200	3.8×
1,000	0.608	0.502	**0.639**	0.200	3.2×
5,000	0.524	0.553	**0.610**	0.200	3.1×

Table 3: Accuracy comparison across voting methods (best configuration per sample size).

Key Findings:

1. **PageRank-weighted accuracy is consistently highest** across all sample sizes
2. **Accuracy exceeds 3× the random baseline** even at the largest, noisiest scale
3. **Central actors encode genre identity** more reliably than peripheral actors
4. **The network knows about genre** despite never being given genre labels during community detection

> **Bottom line:** Using only network topology—no text, no plot, no semantic features—we achieve **61.0% accuracy** on a 5-class prediction problem where random guessing yields 20%.

This is the strongest single result demonstrating that genre is a network-emergent property.

---

5. Discussion

5.1 Why Does Genre Emerge from Network Structure?

The film industry operates through **collaboration microcultures**:

- Recurring casts (e.g., Christopher Nolan's ensemble)
- Director–actor partnerships
- Franchise ecosystems (MCU, horror sequels)
- Genre-bound production companies
- Low-budget horror circuits
- Rom-com ensembles
- Prestige drama clusters

These institutional structures manifest as high-modularity subgraphs that align with recognizable genres.

5.2 Why Does Accuracy Decline with Scale?

Larger datasets introduce noise:

Factor	Effect
Multi-genre actors	Blur community boundaries
Cross-genre franchises	Create inter-community bridges
High-degree bridge actors	Reduce modularity
Drama as umbrella genre	Connects disparate clusters

However, **accuracy never approaches random**—genre signal persists at all scales.

5.3 Why Is PageRank-Weighting Superior?

PageRank identifies the **core actors** of each sub-network. Consider:

$$\text{Community genre} \approx \text{weighted average of member genres}$$

Peripheral actors (low PageRank) often appear in single films across genres—they add noise. Core actors (high PageRank) have consistent genre profiles—they define community identity.

By weighting accuracy by PageRank, we **amplify signal and suppress noise**.

---

6. Conclusion

6.1 Summary of Findings

Actor co-appearance networks contain **substantial intrinsic genre information**, recoverable without textual, plot, or semantic features.

Metric	**Finding**
Modularity	$Q > 0.73$ at all scales
Gini coefficient	Increases with scale (realistic hub formation)
PageRank accuracy	**3× random baseline** at largest scale
Community purity	Majority-genre assignment succeeds

6.2 Key Result

> Even at the largest and noisiest scale (5,000 films), the model achieves **61.0% accuracy** using only network structure—over three times the random baseline of 20%.

6.3 Theoretical Implications

Genre is not merely a content-based label. It is a **network-emergent pattern** shaped by:

- Collaboration structure
- Institutional clustering
- Recurring partnerships
- Production company specialization

Genre, fundamentally, is a pattern of relationships.

References

1. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

2. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the

web. *Stanford InfoLab Technical Report*.

3. Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.

4. Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6), 066111.

Code and data available at: [github.com/aviherman/social-networks-project]