

Predicting Movie Genre Using Actor Co-Appearance Networks

Can network topology alone reveal genre?

Avi Herman

Social Networks Final Project — Fall 2025

Department of Computer Science

The Question

Can we predict a movie's genre
using only actor collaborations?

No plot. No script. No text. No semantic features.

Just **who worked with whom**.

Hypothesis

Core Claim

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

Why might this work?

- Directors repeatedly cast the same actors
- Production companies specialize in genres
- Franchise ecosystems create dense clusters
- Horror circuits, rom-com ensembles, prestige drama pools

Genre may be a network-emergent property.

Conceptual Framework

The Logic:

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.

If communities correspond to genres, then **genre is recoverable from network structure alone.**

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

Building the Network

Graph Structure: $G = (V, E, w)$

- **Nodes** $V = \{a_1, a_2, \dots, a_n\} = \text{Actors}$
- **Edges** $E = \{(a_i, a_j) : \text{co-appear in film}\}$
- **Weights** = Genre-purity weighting

Edge Weight Function:

$$w(f) = \begin{cases} 1.00 & \text{single-genre film} \\ 0.25 & \text{dual-genre} \\ 0.05 & \text{triple-genre} \\ 0.01 & 4+ \text{ genres} \end{cases}$$

Building the Network (continued)

Final edge weight between actors a_i and a_j :

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

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

Intuition: A pure horror film is stronger evidence than Horror/Comedy/Romance.

Edge Weight Rationale

Why penalize multi-genre films?

Step-function weighting:

- 1 genre: 1.00 (strong indicator)
- 2 genres: 0.25 (split signal)
- 3 genres: 0.05 (minimal)
- 4+ genres: 0.01 (nearly ignored)

Rationale:

- Single-genre films strongly indicate genre affinity
- Multi-genre films dilute genre specificity
- We want to prioritize genre-pure collaborations

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.

Why Community Detection?

The central premise: Genre is encoded in collaboration structure

Key Insight

If actors cluster by genre, then:

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

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

Community Detection: Louvain Algorithm

Goal: Find densely connected subgroups (communities)

Modularity:

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

Symbols:

- A_{ij} = edge weight
- k_i = weighted degree
- m = total edge weight
- c_i = community of i

Interpretation:

- $Q = 0$: Random
- $Q \approx 0.3$ – 0.7 : Significant
- $Q > 0.7$: Strong

Community Detection: Louvain Algorithm (continued)

Louvain Process:

1. Each node greedily joins community maximizing ΔQ
2. Communities become super-nodes
3. Repeat until convergence

Two-phase algorithm: **Local optimization** then **aggregation**

Why More Communities Than Genres?

The algorithm typically produces **more communities than the 5 macro-genres**

Why?

- Sub-genre specialization
(slasher vs. supernatural horror)
- Franchise isolation
(MCU actors form own cluster)
- Niche production circuits
(low-budget regional films)

This is informative!

- Algorithm discovers natural structure
- Multiple resolution levels
- Not preset—determined by modularity optimization

The number of communities $|\mathcal{C}|$ is determined by the data, not preset.

What Do Communities Contain?

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

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

Genre Distribution: Characterize each community by its genre mix

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

where $g^*(a)$ is actor a 's dominant macro-genre.

This gives a probability distribution over genres for each community.

Pure communities have one dominant genre.

Mixed communities are spread across genres.

Community Purity

Single Community Purity:

$$\text{Purity}(c) = \max_g \frac{|\{a \in c : g^*(a) = g\}|}{|c|}$$

What fraction of a community shares the same dominant genre?

Overall Purity (Weighted):

$$\text{Purity}_{\text{overall}} = \frac{1}{|V|} \sum_{c \in \mathcal{C}} |c| \cdot \text{Purity}(c)$$

Equivalent to unweighted accuracy when assigning each community its majority genre.

Community Purity (continued)

Interpretation:

- 0.20 = Random (5 genres)
- 0.40–0.50 = Weak clustering
- 0.60–0.70 = Moderate separation
- 0.80+ = Strong communities

High purity indicates that Louvain communities align with genre boundaries—the network *knows* about genre even though genre labels were never used.

Purity vs. Accuracy

Purity

- How homogeneous each community is
- Measures community-level consistency
- Equivalent to unweighted accuracy

Accuracy

- How well community assignment predicts individual actors
- Can exceed purity with weighting
- PageRank-weighting amplifies signal

Key Point

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

Prediction Accuracy Metrics

Three Voting Schemes:

1. Unweighted Accuracy:

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

Each actor contributes equally

2. Degree-Weighted Accuracy:

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

Weights by collaboration volume

3. PageRank-Weighted Accuracy:

$$ACC_{PR} = \frac{\sum_{a \in V} PR(a) \cdot \mathbb{I}[\hat{g}(c_a) = g^*(a)]}{\sum_{a \in V} PR(a)}$$

Weights by network influence

Rationale

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

PageRank: Finding Core Actors

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

Parameters:

- $d = 0.85$ (damping factor)
- N = total actors
- $\mathcal{N}(a_i)$ = neighbors

Intuition:

- Actor is important if they collaborate with important actors
- Weighted by strength of collaborations

Key Insight

Core actors (high PageRank) define community identity.

Peripheral actors often appear in single cross-genre films.

Gini Coefficient: Measuring Inequality

Formula:

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

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

Interpretation:

- $G \approx 0$: Perfect equality
- $G \approx 0.2$ – 0.3 : Weak hub structure
- $G \approx 0.5$ – 0.7 : Power-law (realistic)
- $G \rightarrow 1$: Extreme inequality

What it tells us:

- Whether we have strong hubs
- How unequal actor importance is
- If network has realistic power-law shape
- If dataset is large enough to reveal structure

Falsification Criteria

The hypothesis is rejected if:

Modularity

$Q < 0.3$ across all
configurations

Community Purity

No dominant genre per
community

Accuracy

Converges to random
baseline ≈ 0.20

None of these occurred.

The hypothesis is strongly supported.

Dataset

Source: IMDb Non-Commercial Datasets

Filters:

- Feature films only
- Runtime \geq 59 min
- Released 1960+
- Top-3 billed cast
- Actors with \geq 3 credits
- 18 canonical genres

5 Macro-Genres:

- **ACTION**: Action, Adventure, Thriller, Sci-Fi, War
- **DRAMA**: Drama, Romance, Biography
- **COMEDY**: Comedy, Music, Musical
- **DARK**: Crime, Mystery, Horror
- **FAMILY**: Family, Animation, Fantasy

Sample Sizes: 250, 1000, and 5000 films (top by popularity)

Experimental Design

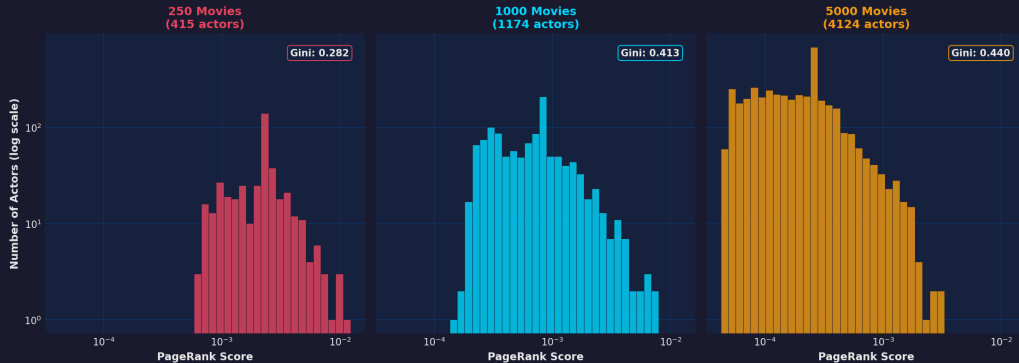
For each sample size $N \in \{250, 1000, 5000\}$:

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

Total: 3 sample sizes \times 10 genre counts = 30 configurations

Result 1: PageRank Distribution Evolution

PageRank Distribution Across Dataset Sizes
(Log-log scale reveals power-law structure)



250 films

Thin distribution

No heavy tail

1,000 films

Heavy tail starts

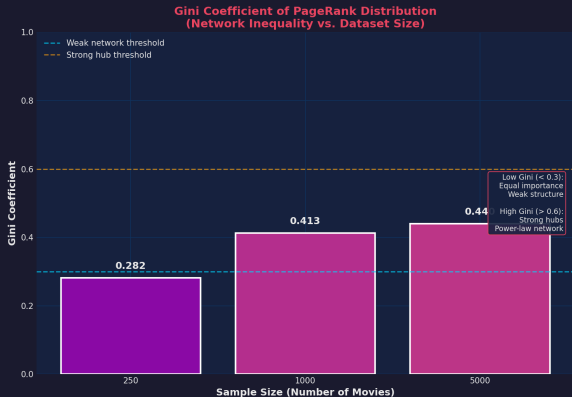
Power-law forming

5,000 films

Clear power-law

Strong hubs

Result 1 (cont.): Network Inequality Scales with Size

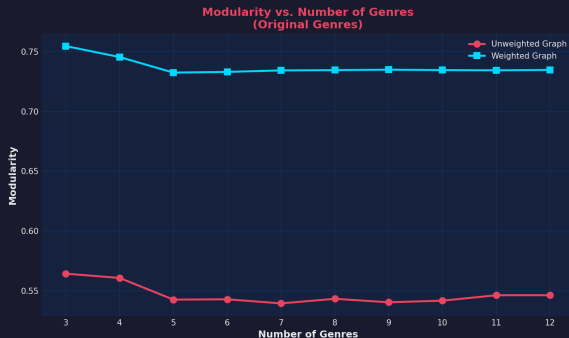


Films	Actors	Gini
250	415	0.282
1,000	1,174	0.413
5,000	4,124	0.440

Finding:

Larger datasets →
more realistic hub structure
(rich-get-richer)

Result 2: Strong Community Structure



Films	Peak Q
250	0.919
1,000	0.845
5,000	0.755

Finding:

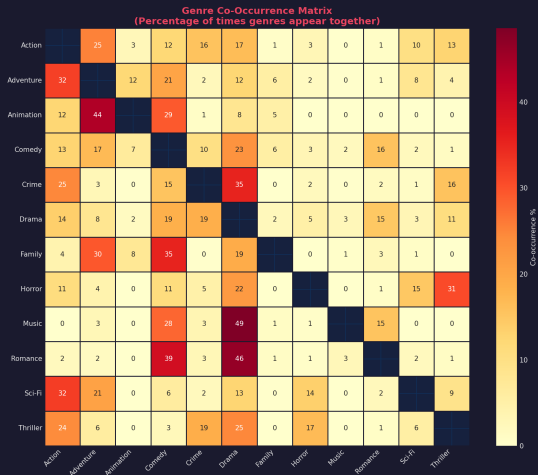
$Q > 0.7$ at all scales

Actor networks are
highly modular

Result 3: Genre Separation



Result 4: Genre Co-Occurrence Explains Overlaps

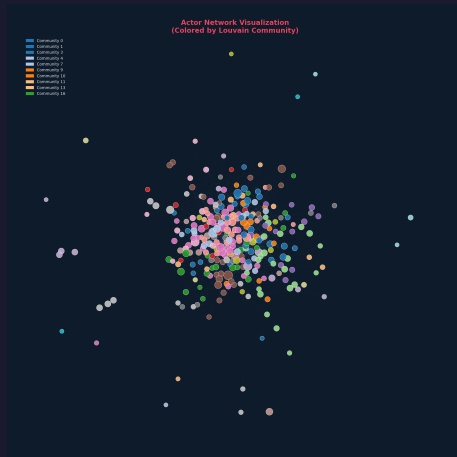


Key Structures:

- Action–Adventure–Thriller triad
- Family–Animation cluster
- Comedy avoids Dark
- Drama connects everything

Multi-genre films create cross-community bridges.

Result 5: Network Visualization



5,000 films. Node color = community. Node size = degree.

Result 6: Genre Co-Occurrence Chord Diagram



Visualization:

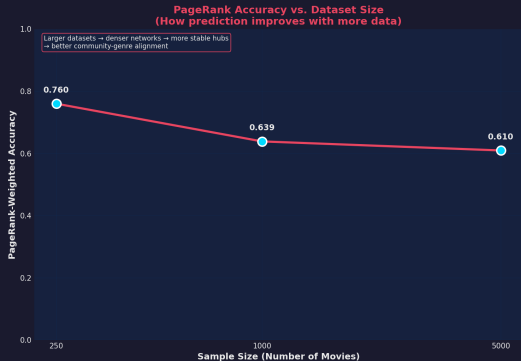
- Genres around circle
- Arcs = co-occurrence
- Thickness = frequency

Key insight:

Thick arcs → frequent
co-occurrence → natural genre
clusters

Example: Action–Adventure–Thriller
triad

The Central Result: Prediction Accuracy



Films	Accuracy	Lift
250	76.0%	3.8×
1,000	63.9%	3.2×
5,000	61.0%	3.1×

Random baseline: 20%

**3× better than random
using only topology!**

Accuracy Comparison: All Methods

Films	Unweighted	Degree	PageRank	Random
250	0.720	0.696	0.760	0.200
1,000	0.608	0.502	0.639	0.200
5,000	0.524	0.553	0.610	0.200

Key Findings:

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

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
- Animation voice actor pools

These **institutional structures** manifest as high-modularity subgraphs aligned with recognizable genres.

Why Does Accuracy Decline with Scale?

Larger datasets introduce noise:

- 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
- More genre mixing
Dilutes cluster purity

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

Why PageRank-Weighting Wins

The Problem

Peripheral actors appear in single cross-genre films.

They add **noise** to community genre assignment.

The Solution

PageRank identifies **core actors** who define the community.

These actors have consistent genre profiles.

PageRank weighting = amplify signal, suppress noise

Summary of Findings

Network Properties:

- **Modularity:** $Q > 0.73$ at all scales
- **Gini coefficient:** Increases with scale
- **PageRank accuracy:** $3\times$ random baseline
- **Community purity:** Majority-genre assignment succeeds

Key Results:

- Strong community structure at every scale
- Realistic hub formation with larger datasets
- Clear genre clusters in network visualization
- PageRank-weighting consistently best method

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

61% accuracy

on a 5-class problem

(random = 20%)

Using only network structure

No text. No plot. No semantics.

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%.

Conclusion (continued)

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.

*“Genre, fundamentally,
is a pattern of relationships.”*

github.com/avih7531/social_networks_final