Cult Classics & Community Detection: Evaluation of Film Success Using Letterboxd Social Network

Tanner Amundsen
The Johns Hopkins University Applied Physics Laboratory
Laurel, Maryland
tamunds1@jhu.edu

***Abstract*—Understanding a film's fanbase is helpful for film producers and consumers alike. Social networking sites like Letterboxd that allow users to share their reviews of films are becoming increasingly popular as users connect over sharing their reviews of the same films. Interpreting these reviewer connections as a social graph allows for community detection and analysis to evaluate and predict the success of the films being reviewed. In this paper, we identify various metrics related to communities of Letterboxd users and evaluate the strength of these metrics as predictors for film success. We find that there are weak but consistent, positive correlations between community size and transitivity and film revenue. We also find that there is a weak negative correlation between community insularity and film revenue.**

***Index Terms*—community detection, movies, Letterboxd, insularity, polarization**

I. Introduction

The film industry is changing rapidly. The rise of streaming, pandemic-era blows to movie theaters, and the amalgamation of studios into a handful of giant corporations have all altered the landscape of film production. The way people consume movies has changed with the total video streaming time by users of streaming services increasing by 18% from 2021 to 2022 [1]. Even the way people discuss film has changed. Letterboxd is a social network that allows people to connect over shared interest in films. The size of the platform has been growing rapidly from under 1.8 million users right before the pandemic to over 8 million users in February 2023 [2].

The effect of these changes on film production has been complicated but one clear trend has been a steady shift away from cult classic films toward films with more mass appeal [3] [4]. The exact definition of cult film is difficult to nail down, however they tend to be movies with a small but passionate fan base, mixed critical reviews, and poor performance at the box office. These small but passionate fanbases have joined the rest of the cinephile community in the shift online. The rising popularity of social media sites like Letterboxd offers a rich source of data to track film popularity among non-professional critics. The social component of sites like Letterboxd also provides the opportunity to analyze film data from a social networking perspective.

In this paper, we outline an approach to extract social network data from Letterboxd reviews and apply community detection and cluster analysis to evaluate the cult-factor and commercial success of films being reviewed on Letterboxd. Given the "cult" nature of cult film fanbases, we pay special attention to metrics concerned with connectedness, insularity, and polarization. We attempt to answer the question:

- RQ: What can the clustering of the social network of online film reviewers tell us about the cult-factor and success of films.

For this paper, we measure "cult-factor" using the film's total (domestic and international) box-office revenue. This is in part due to the limited time and data resources of this research and in part due to the difficulty of measuring other factors that make a film "culty". Future work should be done to factor in critic and audience sentiment to augment this social network analysis.

The results of this study could be useful to many different parties. First, Letterboxd and similar companies could use the proposed community detection and clustering methods to enrich user experience. Perhaps recommending communities with similar taste in film, or categorizing films by niche-ness or mass appeal. Second, production houses could use this revenue analysis to explore what kinds of films are performing well commercially and critically. And lastly, cinephiles with an interest in the changing landscape of film production and critique can use this analysis to evaluate the changes in film consumption over time.

II. Background

The literature review for this project consisted of three main areas of research: (1) general use of social media data to predict film success, (2) surveys of community detection algorithms, and (3) different metrics for measuring community distance, insularity, and polarization.

A. Methods for Predicting Film Success.

While there is a dearth in research related to identifying or measuring the "cult" factor of films using audience review data, there are many published methods in predicting the commercial success of films using audience review data. [5] used natural language processing on movie reviews posted to various newspapers and websites to predict opening weekend revenue for various films. [6] applied k-means clustering to various social media metrics including Twitter follower count, and sentiment analysis of YouTube viewers' comments to categorize films into either "Hit", "Neutral", or "Flop".

Graph theory has also been used to predict movie success. In [7], a graph is constructed where vertices represent movies and the distance between them is weighted according to their similarity in various meta-data features like cast, production company, genre, director, etc. They then applied graph convolutional neural networks to achieve highly accurate predictions of movie revenue. In [8], the researchers looked at forum discussions on the Internet Movie Database (IMDb) and weighted these posts according to network position to predict box office revenue and Academy Awards nominations.

B. Community Detection

Community detection is a widely studied field in graph theory and social network analysis. While there are many different methods of clustering and community assignments, the algorithm proposed by Girvan and Newman in [9] which is completed by progressively removing edges from the network in order of edge centrality, is the most widely cited and implemented. We found this algorithm to be sufficient for this application but recommend experimenting with other community detection methods in future work.

C. Cluster Metrics

1) Modularity: Modularity is a metric used to judge the quality of a community division of a graph. It measures community division and is NP-complete to optimize. For a graph with a community structure assigned to it, let e_{ij} by the fraction of edges in the network that connect vertices in group i to those in group j and let $a_i = \sum_j e_{ij}$ then [10] defines modularity Q as follows:

$$Q = \sum_i (e_{ii} - a_i^2)$$

2) Transitivity: Transitivity measures the probability that the adjacent vertices of a vertex are connected. This metric is also sometimes called the clustering coefficient. For 3 vertices i, j and k, they are considered transitive if

$$(i \to j, j \to k) \implies i \to j$$

So, globally, this measure is the ratio of the graph's triangles to its connected triplets [11]. Graphs with higher transitivity tend to be denser and more connected than those with lower transitivity.

3) Leadership Insularity: Leadership insularity is a metric applied to graphs that have been assigned a community partitioning and measures how insular the leader of each community are from one another [12]. Leadership insularity is defined as the average relative distance between leaders of different communities. The equation for leadership insularity I is

$$I = \frac{1}{(N_c - 1)N} \sum_{i=1}^{N_c} \sum_{j=i+1}^{N_c} \frac{d(L_i, L_j)}{d(i,j)} (N_i + N_j)$$

Where N_c is the number of communities identified, N is the number of nodes in the network, N_i is the number of nodes in community i, L_i is the leader of community i, $d(L_i, L_j)$ is the distance between community leaders L_i and L_j, and $d(i,j)$ is the mean distance between communities i and j. This calculation is visualized in figure 1

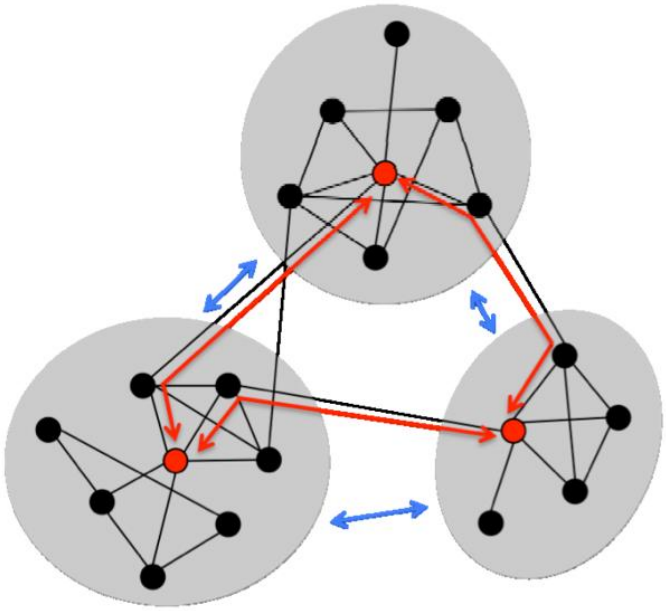

Fig. 1. A visualization of the calculation for leadership insularity. red nodes are community leaders with red lines being the distance between them. Blue lines indicate the mean distance between communities [12].

Leaders are chosen as the vertices in each cluster with the highest betweenness centrality with ties being broken by comparing vertex degree. In this study, we also isolate the term $\frac{d(L_i,L_j)}{d(i,j)}$ to define a pairwise leadership insularity between communities i and j.

4) Polarization: Polarization as a sociological concept can be measured within a social network in a variety of ways. Networks with a high modularity have been shown to also exhibit community separation along polarized lines such as the U.S. political divide [13]. However, modularity as the sole indicator of polarization has been shown to be insufficient for measuring the presence and degree of polarization in several different social networks [14]. Specifically, modularity simply measures the internal and external connectivity of two groups G_i and G_j with no distinction between homophily or antagonism being the motivator behind these connections [14]. A novel metric for polarization proposed in [14] looks at "boundary nodes" between groups G_i and G_j and compares the number of internal connections they have to the number of boundary crossing connections they have. The polarization P between two clusters is defined as

$$P = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \right]$$

where B is the set of boundary nodes, $d_i(v)$ is the number of edges that connect boundary node v to a non-boundary (internal) nodes within its own cluster, and $d_b(v)$ is the number of edges that connect boundary node v to vertices in the

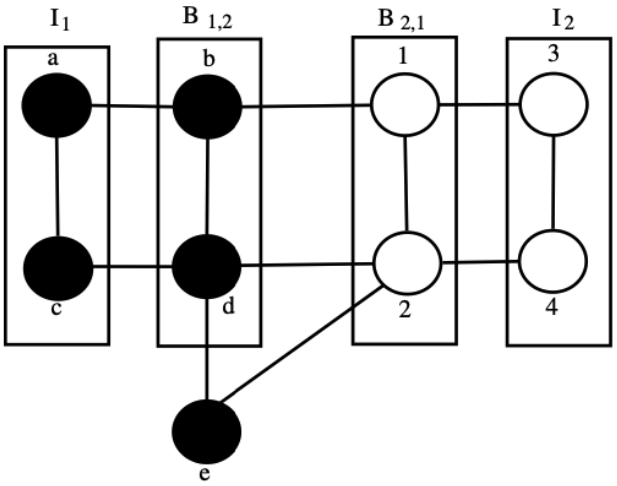

Fig. 2. An example of a graph divided into two communities (black and white). Vertex sets $B_{1,2}$ and $B_{2,1}$ contain the boundary vertices and sets I_1 and I_2 contain internal nodes which are nodes that do not connect to the opposite cluster. Source: [14].

opposite cluster. Polarization is defined in a pairwise fashion between individual clusters and is undefined for clusters that have no boundary nodes.

III. Dataset

A. Data Collection

A sample dataset of 4,596 film reviews was scraped from Letterboxd using their beta API and uploaded to Kaggle in April 2023 [15]. These reviews included the reviewer name, movie name, text review, comment count, like count, and the date the review was posted. The dataset comprised reviews from 1,113 unique titles and 1,140 unique reviewers.

The original dataset did not include box-office revenue data for the films so it was augmented with that data using a web scraper of BoxOfficeMojo's website.

B. Network Formation

We define the network $M = (reviewers \cup films, reviews)$ which is the network connecting reviewers to films using the relationship $has\_reviewed$. M is a directed network with 2,258 vertices and 4,597 edges. We define A_M as a matrix of size $[num\_reviewers \times num\_films]$. This matrix is such that index $[i, j]$ is 1 if reviewer i has reviewed film j. Then matrix A_M^T of size $[num\_films \times num\_reviewers]$ is such that index $[j, i]$ is 1 if film j has been reviewed by reviewer i.

We define the graph G to be the graph built by the adajacency matrix that results from the relational algebra

$$A_G = A_M \times A_M^T \tag{1}$$

This relational algebra creates a social graph connecting reviewer vertices by the relationship $has\_reviewed\_same\_film$ and results in the graph $G = (reviewers, common\_reviews)$. Vertices of degree 0 were removed as this analysis was on clusters of vertices and the resulting graphs had 1062 vertices and 20478 edges. Redundant edges were also removed meaning that we do not distinguish between pairs of reviewers that have reviewed many of the same films and reviewers that have only reviewed one shared film. The resulting graph is rendered in figure 3 using the Fruchterman Reingold rendering.

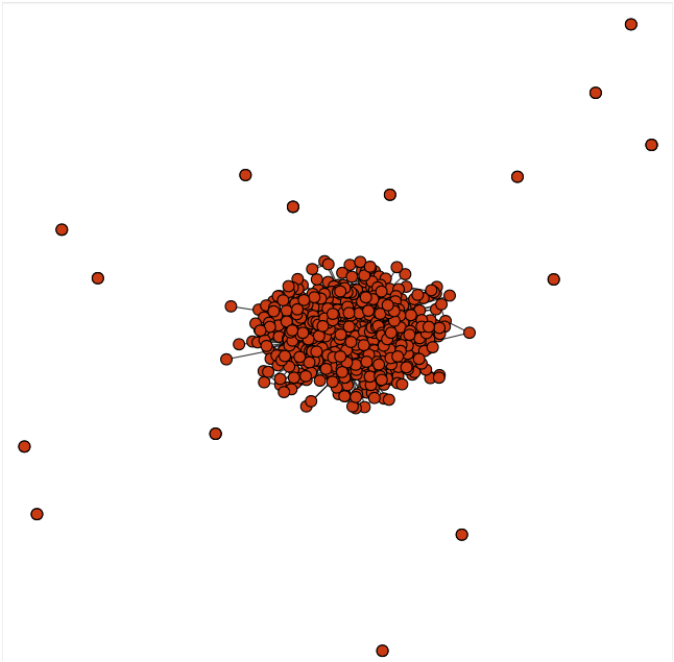

Fig. 3. Graph G of Letterboxd reviewers where vertices represent reviewers and edges represent connections where two reviewers have reviewed at least 1 of the same films. Note that the disconnected clusters that appear to be single nodes are actually all >1 nodes placed closely together

IV. Methods

The community structure detection algorithm proposed in [9] was applied to graph G. This algorithm resulted in 413 clusters over the 1062 vertices. Figure 4 shows the graph G colored by cluster membership hereafter referred to as community partitioning C.

One cluster in particular dominated the network and consisted of 434 reviewers. To better visualize the community detection assignment, the cluster graph of the resulting community assignment is shown below in figure 5. Note the highly connected central node in the center representing the largest cluster of 434 reviewers.

A. Network-Wide Metrics

Many popular metrics related to connectedness, clusterability, insularity, and polarization are network-wide metrics. These network-wide metrics were compiled for community partitioning C and tabulated in figure 6. Although we are not directly able to use these network-wide metrics to predict individual film success, we are able to use them to compare the graph G to other graphs in future work.

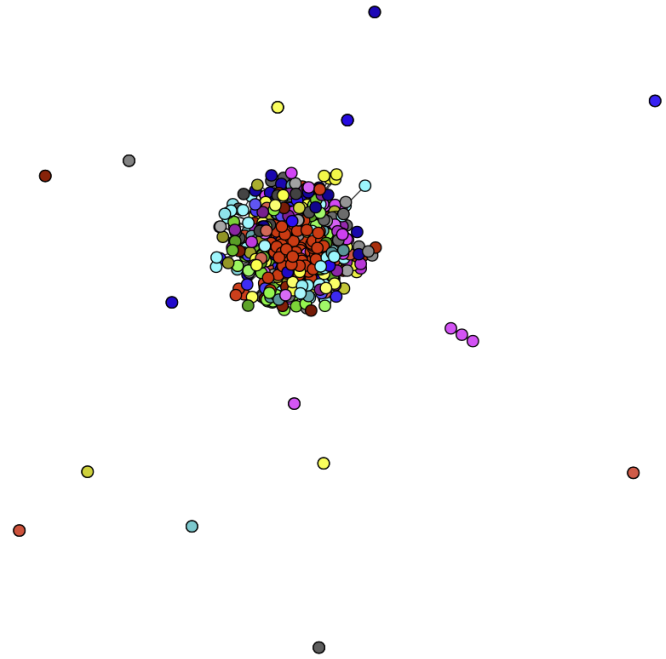

Fig. 4. Graph G colored by cluster membership. Note the large collection of red vertices in the center belong to the same large cluster.

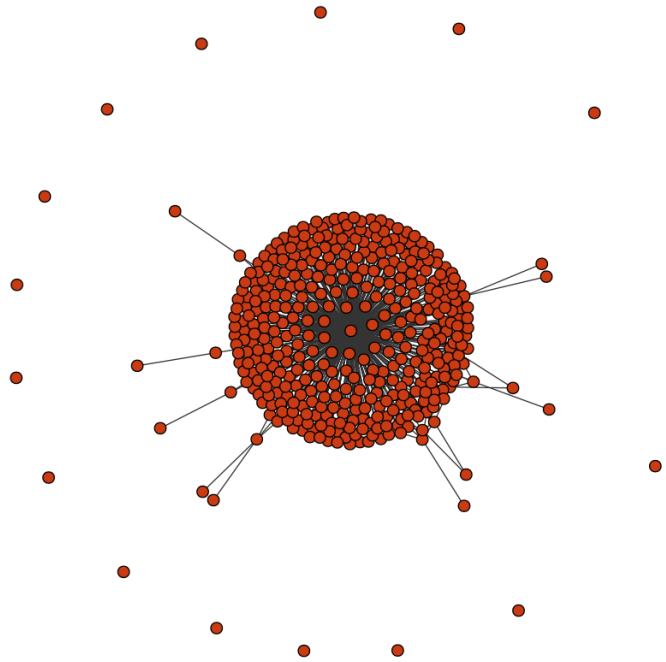

Fig. 5. The cluster graph of the community paritioning of G where vertices represent clusters of viewers connected if there is a least one vertex pair (a, b) in the original graph G such taht vertex a was in cluster i and vertex b was in cluster j

B. Cluster-Specific Metrics

Next, we focused on individual clusters to measure both the intra-cluster connectedness and inter-cluster connectedness. For each cluster i, we calculated:

- **Cluster Size**: the number of vertices assigned to cluster i. This measure gives us an idea of the size of the online fanbase for the movies that connect the vertices in cluster i.
- **Unique Movie Count**: the number of unique movies that the reviewers in cluster i have reviewed. We use the original network M (where reviewer vertices were connected to film vertices they have reviewed) to count the number of unique movies represented in each cluster in graph G. This metric is an attempt to capture the niche-ness versus mass-appeal of the films that connect the vertices of cluster i.
- **Transitivity**: the transitivity of the subgraph consisting only of vertices of cluster i to compare the connectedness within each cluster [11]. In this context, clusters with high transitivity can be interpreted as a community of reviewers that **all** tend to review **many** of the same movies.
- **Polarization**: the pairwise polarization value defined in [14] for each eligible pair of clusters, averaged over all clusters with which cluster i has a defined polarization score. In this context, clusters with a high polarization correspond to reviewers that **do** tend to review the same movies as other reviewers in their cluster and actively **don't** review films that are outside of their specific niche.
- **Leadership Insularity**: the pairwise leadership insularity between each pair of clusters, averaged to find the leadership insularity value for cluster i [12]. In this context, clusters with a high average leadership insularity may correspond to reviewers with more niche, cult interests.

.

These results (summarized in figure 7) were aggregated into a matrix we define as X of shape $[num\_clusters \times$

Network-wide metrics

Metric	Possible Range	Value
Number of Clusters	(1,1062)	413
Modularity	(-1,1)	0.0947
Transitivity	(0,1)	0.2379
Max Polarization*	(-0.5,0.5)	0.4970
Min Polarization*	(-0.5,0.5)	-0.3361
Leadership Insularity	(0,1)	0.4417

Fig. 6. *Polarization is defined between pairs of clusters

Cluster-specific metrics

Metric	μ	σ
Cluster Size	2.571	21.344
Unique Movie Count	4.143	47.515
Transitivity	0.104	0.304
Polarization	0.192	0.186
Leadership Insularity	0.922	0.181

Fig. 7. The results of cluster analysis on the community partitioning C on graph G

Film revenue by weighted average cluster metrics

Avg. Cluster Metric	r	β_1
Size	0.0895	217248
Unique Movie Count	0.0894	97315
Transitivity	0.0610	132591000
Polarization	0.100	239374000
Leadership Insularity	-0.0992	-353542000

Fig. 8. The results of correlating film revenue by the weighted average of the cluster metrics of graph G. These results were created using the data in matrix Z, the Pearson correlation coefficient, and linear least squares regression

$num\_cluster\_features]$ where the rows corresponded to clusters and the columns correspond to the values for the metrics listed above.

V. Discussion and Conclusion

A. Film Performance by Cluster Metrics

In order to relate the cluster metrics aggregated in matrix X back to film revenue, we need to map the films to their respective representation within the clusters. To do this, we define the matrix Y to be a matrix of size $[num\_movies \times num\_clusters]$ to be such that

$$Y[i,j] = \frac{|v^*_{C_j}|}{\sum_j |v^*_{C_j}|}$$

Where $v^*_{C_j}$ is the set of reviewers within cluster C_j that have reviewed movie i. Note that the denominator is just a normalizing factor over all clusters. Essentially, the ith row of matrix Y corresponds to weights of movie i's representation within each of the clusters of graph G and is normalized to sum to 1. We define matrix Z of size $[num\_films \times num\_cluster\_features]$ to be:

$$Z = Y \times X \tag{2}$$

This matrix product gives us, for each movie, an average value for each clustering metric - weighted by that movie's representation in each cluster. The final matrix Z was used as the feature vector of different predictors for film revenue. That is, we were able to create regression models for film revenue for each feature f:

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} 1 & z_{1,f} \\ 1 & z_{2,f} \\ \vdots & \\ 1 & z_{n,f} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We also calculate the correlation coefficient r. The results are tabulated in figure 8 and visualized in figure 9

B. Key Findings

1) Letterboxd social networks of reviewers who are connecting by reviewing the same film form 1 large community cluster and many smaller community clusters.
2) In terms of network-wide metrics, this graph has a very small (but positive) modularity and low transitivity. The main, large cluster displays a high average polarization with the other clusters that it borders. This means that reviewers in the largest cluster actively review the same titles as other reviewers in their cluster and actively don't review more fringe titles represented in other clusters. The network had a leadership insularity near 0.5 meaning that the leaders were neither especially insulated from one another by their respective clusters nor especially closely connected to each other on the fringes of their respective clusters.
3) **Answer to RQ:** None of the cluster metrics used had particularly strong predictive power or high-magnitude correlation coefficients with film revenue. Several cluster metrics were identified as having either a positive or negative correlation with overall film success. Cluster size, unique movie count, transitivity, and polarization were all shown to have a positive correlation with film revenue. Whereas leadership insularity was shown to have a negative correlation with film revenue.
Cluster size and unique movie count being positively correlated with revenue may be a result of the mass appeal of various films. That is, clusters of reviewers that are larger and more connected, may be connected by films that attract wider audiences. And films that attract wider audiences make more money.
Transitivity had the weakest positive correlation with revenue. The sign of the correlation may be a result of audience tastes around similar movies driving revenue. Reviewers in more tightly connected communities may review movies that have more mass appeal which in turn tend to perform better at the box office. Or movies being reviewed by these tightly connected clusters may all be similar in plot, cast, production house, etc.
Polarization having a positive correlation with revenue

Fig. 9. Dashboard of scatterplots of film revenue by various cluster metrics of graph G. Data from matrix Z. Plotted using Tableau.

was surprising. We expected clusters that were more polarized from others to correspond to lesser-known movies. However, the cluster with the highest polarization was the main, large cluster and this dominated the weighted average for most film's polarization score. Further, polarization was undefined for many of the cluster pairs making this metric unfit as a predictor for this application.

Lastly, leadership insularity having a negative correlation with revenue means that clusters with greater insularity between leader nodes tended to review less successful films. This is expected since we can interpret these highly insulated leaders as mostly reviewing films that others have not - which may likely also mean that they are reviewing films that more people have not payed money to see.

C. Limitations

The biggest limitation of this research was the lack of data from Letterboxd. Use of their API requires specific permissions so we are at the mercy of publicly available datasets scraped by other people. For the dataset used in this paper, the sampling method was not clear so we have no way of knowing how these reviews were selected or if this sample was representative of broader Letterboxd user behavior.

Another limitation of this dataset included the fact that the actual numeric rating for each film review was in a corrupted data format and was unable to be parsed. This made it impossible to factor in reviewer sentiment towards the film without doing sentiment analysis on the textual review itself.

Additionally, the dataset did not explicitly include film revenue for each film. Film revenue had to be added to the dataset after being scraped from BoxOfficeMojo.com. This resulted in several null values for revenue that hurt our analytical capabilities.

VI. Future Work

Future work in this research should be performed on larger datasets collected on reviews over a longer period of time. Review sampling should be done in a way that is representative of the true popularity of various films on these online reviewing platforms. Instead of relying solely on cluster connectivity metrics, future work should factor in vertex attributes within clusters like user ratings and review popularity. These added data points may serve as better predictors for a particular movie's "cult factor" and success. Another avenue to explore in future work is adding target variables that attempt to capture a film's "cult factor" beyond just box office revenue. This could look like predicting critical ratings or using online lists of films that are widely agreed upon to be cult films as ground truth in a categorization problem.

References

[1] A. Durrani, "The average american spends over 13 hours a day using digital media-here's what they're streaming," Mar 2023. [Online]. Available: https://www.forbes.com/home-improvement/internet/streaming-stats/#:~:text=Total%20video%20streaming%20time%20by,of%20video%20streaming%20per%20year.

[2] B. Basagre, "Letterboxd: The kiwis behind cinema's most influential platform," Feb 2023. [Online]. Available: https://idealog.co.nz/venture/2023/02/letterboxd-the-kiwis-behind-cinemas-most-influential-platform#:~:text=With%20over%20eight%20million%20users,with%20the%20traditional%20media%20form.

[3] J. Rottenberg, "What'll become of the cult movie?" Nov 2016. [Online]. Available: https://digitaledition.baltimoresun.com/tribune/article_popover.aspx?guid=f9b71a9d-ce41-4978-bfe3-90f9afcf042a

[4] E. Kohn, "Want another hollywood revolution like the '70s? make cult classics (column)," Jan 2023. [Online]. Available: https://www.indiewire.com/2023/01/hollywood-should-make-cult-classics-1234797087/

[5] M. Joshi, D. Das, K. Gimpel, and N. Smith, "Movie reviews and revenues: An experiment in text regression." 06 2010, pp. 293–296.

[6] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio, "Prediction of movies box office performance using social media," *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 1209–1214, 2013.

[7] C. M. G. N. M. C. Barbany, Oriol, "Movie grossing success prediction with convolutional neural networks on graphs."

[8] J. Krauss, S. Nann, D. Simon, K. Fischbach, and P. Gloor, "Predicting movie success and academy awards through sentiment and social network analysis," 06 2008.

[9] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, jun 2002. [Online]. Available: https://doi.org/10.1073%2Fpnas.122653799

[10] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, jun 2006. [Online]. Available: https://doi.org/10.1073%2Fpnas.0601602103

[11] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

[12] S. Arbesman and N. A. Christakis, "Leadership insularity: A new measure of connectivity between central nodes in networks," *Connect. (Tor.)*, vol. 30, no. 1, pp. 4–10, Jan. 2010.

[13] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," 01 2011.

[14] P. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, vol. 7, pp. 215–224, 01 2013.

[15] S. Learner, "Letterbox movie ratings data," 2022. [Online]. Available: https://www.kaggle.com/datasets/samlearner/letterboxd-movie-ratings-data?select=ratings_export.csv