Movies*

Stelios Michalopoulos†
Brown University, CEPR and NBER

Christopher Rauh‡
University of Cambridge, CEPR, IZA, HCEO, and PRIO

This draft: February 2024

Abstract

Why are certain movies more successful in some markets than others? Are the entertainment products we consume reflective of our core values and beliefs? These questions drive our investigation into the relationship between a society's oral tradition and the financial success of films. We combine a unique catalog of local tales, myths, and legends around the world with data on international movie screenings and revenues. First, we quantify the similarity between movies' plots and traditional motifs employing machine learning techniques. Comparing the same movie across different markets, we establish that films that resonate more with local folklore systematically accrue higher revenue and are more likely to be screened. Second, we find analogous patterns within the US. Google Trends data reveal a pronounced interest in markets where ancestral narratives among the locals align more closely with a movie's theme. Third, we delve deeper into the explicit values transmitted by films, concentrating on the depiction of risk and gender roles. Films that promote risk-taking sell more in entrepreneurial societies today, rooted in traditions where characters pursue dangerous tasks successfully. Films portraying women in stereotypical roles continue to find a robust audience in societies with similar gender stereotypes in their folklore and where women today continue being relegated to subordinate positions. These findings underscore the enduring influence of traditional storytelling on entertainment patterns in the 21st

*We would like to thank the seminar participants at Brown, Bocconi, the Paris School of Economics, Bath, Manchester, Durham, University of Piraeus - Ioannina - Macedonia - and the Athens University of Economics and Business, and the Political Economy of Development Conference at Northwestern, for their helpful comments. Eliana La Ferrara, Anna Maria Mayda, and Katia Zhuravskaya have provided insightful feedback. We are grateful to Eric Greenfeld, President of Fundamental Film Services, Inc., who has graciously shared his knowledge and expertise on the workings of the film-making industry in the US and beyond. Athiwat Thoopthong, Maria Medellin Esquerra, and Daniele Goffi provided superlative research assistance.

†Brown University, CEPR and NBER smichalo@brown.edu https://sites.google.com/site/steliosecon/

‡University of Cambridge, CEPR, HCEO, IZA, and PRIO cr542@cam.ac.uk https://sites.google.com/site/econrauh/

century, highlighting a profound connection between media consumption and deeply ingrained cultural narratives and values.

Keywords: Movies, Folklore, Culture, Values, Entertainment.

JEL Numbers. N00, N9, O10, O43, O55

1 Introduction

The entertainment industry is huge both in terms of revenue generated and individual engagement. According to Comscore (2020), the global box office revenue hit a record of 42.5 billion US$ in 2019. In the 2013–17 period, people aged 15 and older spent an average of 2 hours 46 minutes per day watching TV, videos, or movies (Krantz-Kent 2018). Given how central the entertainment and media industry is in people's lives, it is only natural that a vibrant literature has flourished exploring the impact of media exposure on (political) behavior and values.[1]

In this study, instead of looking at the consequences of media consumption, we take a step back and explore how the consumption of entertainment content varies across societies. Specifically, we ask what explains a movie's box office performance globally and in different markets. Our conceptual framework has two building blocks. First, we view a movie as a story told on the screen. Second, each society has its traditional narratives, images, tales, legends, myths, and epics. Putting these two observations together, we investigate whether entertainment content that "resonates" with local (global) ancestral storytelling is more likely to succeed.

To test these conjectures, we combine a unique catalog of oral traditions across the globe with data on international movie screenings and revenue. We utilize machine learning techniques to quantify the similarity between movies and traditional motifs and establish the following. First, blockbuster movies are those whose plots follow folkloric themes that are universal, i.e., can be found in a multitude of cultures. Second, comparing the same movie across different markets, we observe that films closer to a society's oral tradition are more likely to be screened (supply side), and when screened, they systematically accrue higher revenue (demand side). Third, we establish analogous patterns within the US. Google Trends data reveals heightened interest in Designated Market Areas (DMAs) where ancestral narratives align with the plots of specific movies.

Fourth, we delve deeper into the explicit moral values transmitted by films, concentrating on the depiction of risk-taking and gender roles. Films that promote risk-taking sell more in entrepreneurial societies today, rooted in traditions where characters pursue dangerous tasks successfully. In societies with a historical male bias in their narratives and where women today have limited attachment to the workforce, films portraying women and men in stereotypical roles continue to find a robust audience. These patterns underscore the enduring influence of traditional storytelling on entertainment consumption in the 21st century, highlighting a profound connection between media content and deeply ingrained cultural narratives and values.

Connecting narratives of the past to present film consumption allows us to bring

[1] In economics see Jensen and Oster (2009), Enikolopov, Petrova, and Zhuravskaya (2011), Adena et al (2015), Kearney and Levine (2015), Kearney and Levine (2019), Enikolopov, Makarin, and Petrova (2020), Riley (2022), Esposito, Rotesi, Saia, and Thoenig (2023), Ang (2023), and Armand, Atwell, Gomes, and Schenk (2023), among others, and Stacks, Cathy Li, and Spaulding (2015) for a literature review in media studies and cultivation research.

together two streams of research that have largely developed in parallel. The literature on the historical roots of comparative development and the works on media.

On the one hand, works on how history shapes today's outcomes often invoke cultural explanations as a mediating factor; see Voth (2021), Michalopoulos and Papaioannou (2018), Nunn (2020), and Michalopoulos and Papaioannou (2020) for a focus on Africa. But what can help explain cultural inertia in the face of media globalization? How do people maintain their cultural narratives amidst the growing influence of global media and the increasing availability of various entertainment products? Our research suggests that deeply-rooted societal narratives may endure as people choose to consume media content that is closer to the stories they (their ancestors) grew up with. This pattern paints a nuanced view of persistence and change in the world of media and culture. With the background demand for entertainment driven by our preference for images, plots, and moral themes partially rooted in our ancestral traditions.[2]

On the other hand, studies on the role of media often stress its transformative nature, including the economic, political, and social impacts of being exposed to different narratives and role models, as well as the influence of educational entertainment programs; see DellaVigna and La Ferrara (2015), and Durante, Pinotti, and Tesei (2019).[3] These studies recognize that consumer demand for media content is largely due to the demand for entertainment, with the socioeconomic impacts emerging as a by-product. We contribute to this literature by showing that films that are closer to a society's oral tradition are both more likely to be screened and generate higher revenue. Hence, consumer demand for entertainment products partially reflects deep-rooted attributes of a society's cultural fabric. Our results complement Adukia et al. (2023) who find that in the US libraries that serve predominantly white communities have fewer copies of books centered on non-dominant social identities and that people consume (children's) books that reflect their own identities.

Recognizing that people have different tastes for entertainment themes, may rationalize why backlash occurs as a result of being exposed to content that may clash with one's preferred narratives (Blumenstock, Dube, and Hussain (2022)). This finding is consistent with Bartlett's (1932) seminal experiment in memory. The study's main finding is that subjects exposed to unfamiliar stories, even after studying them repeatedly, had trouble recalling them correctly.[4] Barlett theorized that the individuals struggled to remember the plot accurately because the content deviated from the participants' preconceived notions of how stories should go.[5] This mismatch between a film's theme and the movie-goers'

[2]At the same time, the increasing globalization of media and the resulting exposure of younger cohorts to common (movie) content may help explain the observed convergence of values across the globe, see Kotschy and Sunde (2023).

[3]Recent studies focusing on the US, use computer vision tools to assess gender and race stereotypes in children's books (Adukia et al. 2023), news media articles (Ash et al. 2023) and the evolution of stylistic expressions from school yearbooks (Voth and Yanazigawa-Drott 2023)

[4]The participants were 20 English college students of the University of Cambridge, and the story they were asked to study and recall was a Native American folk tale titled "The War of the Ghosts."

[5]What and how an individual remembers is at the heart of studies in cognitive psychology, see Kahana (2012), Bordalo, Gennaioli, and Shleifer (2020), and Bordalo et al. (2024) for applications in behavioral

expectations regarding how a story should go (shaped by their cultural background) might help explain why in some markets a given film becomes a hit and in others a miss. See Atkin (2016) for a similar argument about how persistent differences in food preferences between social groups shape consumption patterns among immigrants.

Our study has implications for designing edutainment products.[6] It suggests that considering local storytelling can help structure a given message appropriately and avoid triggering an adverse reaction. More broadly, as the film industry is undergoing a period of datafication with new players including Apple entering film production and the integration of AI in filmmaking, our research suggests that the industry by incorporating motifs and themes from the folklore of hitherto underrepresented cultures may be able to reach populations whose (ancestral) stories have not been told.

The paper is structured as follows. In Section 2 we go over our data sources, describe the variation in movie consumption across the globe, and offer some preliminary evidence linking this variation to differences in local lore. Motivated by these examples, we detail how we construct the various similarity metrics between motifs and movie plots. In Section 3, we start by illustrating the global patterns and present the benchmark cross-country within-film results. In Section 4, we describe the within-US patterns. In Section 5, we quantify movies' portrayal of risk-taking and gender roles and establish that audiences favor films whose cultural message aligns with the country's norms. In Section 6, we conclude by offering some thoughts for future research.

2 Data

2.1 Motifs catalogue

Berezkin (2015) folklore and mythology catalog, introduced to economics by Michalopoulos and Xue (2021) (MX), is a compilation of folklore themes across 958 oral traditions worldwide. It categorizes a wide range of stories into 2,564 motifs, facilitating cross-cultural comparisons. Besides folklore studies, this catalog is particularly useful for researchers in anthropology, comparative literature, cultural studies, and social psychology, as it provides a structured framework for understanding the similarities and differences in storytelling traditions around the world. Some motifs, which we will return to below, are ubiquitous, while others are unique to a handful of societies. The median motif is present in 18 oral traditions. This corpus is a valuable resource for identifying universal and localized cultural narratives. We refer to MX for a detailed description of the catalog,

economics. The ongoing research on the interplay between collective and individual memory and the mediating role of culture is summarized by Wang (2021). See Fouka and Voth (2023) for a novel application on how the memories of German troops perpetrating numerous massacres in Greece during World War II, resurfaced during the 2009 Greek debt crisis, leading to a drop in German car sales in areas affected by German reprisals.

[6] Edutainment is the process of intentionally designing and implementing a media message to entertain and educate, increasing knowledge, and shifting social norms, see Anikina and Yakimenko (2015) and Kalil et al. (2024). De Graaf et al. (2012), for example, show that narrative persuasion is related to identification with the characters of the story.

including how to extract quantifiable information, caveats, and how to aggregate motifs to a country's oral tradition. The typical country in our sample has on average 250 motifs.

2.2 Movies database

The movie database we use consists of information available on Internet Movie Database (IMDb) as of 2020. IMDb is a well-known online platform that offers comprehensive film information including details on cast, genre, plot, crew, and revenue, among other aspects. The IMDb consists of hundreds of thousands of movies (about 600k as of 2023). Since the platform relies on users' and film studios' contributions and to minimize noise, the IMDb community usually focuses on the 81k films with at least 100 votes.[7] For 33,528 of these movies we have gross revenue for the original release. Since the coverage declines dramatically for movies screened before the mid-90's and to avoid the industry's disruption caused by the COVID-19 pandemic, we limit our analysis to the 1995-2019 period.

In Table A1 we present our main samples sliced in two ways. Panel A describes the sample of movies we look at when we focus on revenue as the outcome variable. Since we leverage within-movie and within country-year-month variation, we focus on movies that appear in multiple markets. 13,929 movies are screened across 85 countries for a total of 168,626 screenings. The average country revenue per movie is US$ 2,939,290 and the median is US$ 215,350, implying some large outliers. The highest-grossing movies of all time are Avengers: Endgame (2019) and Avatar (2009). Figure A1 shows the distribution of log revenue across movies(-countries). In Panel B, we present the sample we use when the outcome is whether a film is screened in a given country. We exploit within-movie within country-year-of-first-release variation for 28,810 movies in 92 countries. The average number of countries a movie is screened in is 6.42, with the median movie screened in a single market.

Table A2 presents descriptives on main actors, writers, directors, and distributors for the movies in the revenue sample. We provide two sets of statistics, one for the entire revenue sample and the other conditional on the individual/entity appearing in at least five films. The column headed by 'N' indicates the number of actors, writers, and directors in the corresponding (sub)set, while the other columns provide metrics about the number of appearances in movies. Up to three actors are listed for a movie and almost 99% of movies list exactly three actors. An average actor appears in 2.26 movies with the majority appearing only once with a long tail of famous actors featured in up to 55 films. The distributions for writers and directors display similar patterns.

Film distributors act as intermediaries between producers and exhibitors and are responsible for marketing, promoting, and placing movies in theaters. Large producers own their distribution networks worldwide, i.e., have offices in the various markets through which they conduct advertising campaigns and reach the local theaters. For example, the Buena Vista International film distributor is known as the international distribution

[7] See, for instance, the data for a Kaggle competition: `https://github.com/cckuqui/IMDb-analysis/tree/master`.

division of the Walt Disney Company. Medium and small film producers rely on independent distribution networks, such as Spentzos Films and Odeon in Greece and Gaumont and Diaphana films in France. So, the same movie may have a different distributor across markets. The average distributor is responsible for 22.35 films.

Figure A2 provides a temporal overview of our data. The left y-axis indicates the number of movies by the year of release represented by the red solid line, while the blue dashed line reflects the number of countries these movies are screened on average (right y-axis). After 2000, the number of movies in the revenue sample increases starkly, with up to 1,000 movies a year each screened at 10-15 markets on average. For the screening sample, we have up to 2,000 movies in a given year with an average screening of around 6.42 countries. Each movie may be assigned to up to three genres. The tagging of movies' genres involves users' recommendations curated by IMDb. Panel A of Table A3 shows the number of movies per genre. Figure A3 breaks down the number of movies and revenue for each genre by year. The screening of drama movies dominates across all years. Regarding total revenue per genre, two observations are in order. First, comedy movies have been a staple for the typical movie-goer, and second, there has been a meteoric rise in the box office of action, adventure, and sci-fi movies after 2012. In terms of language (reflecting the different languages spoken in a movie irrespective of length), more than half of movies feature English, followed by French, Spanish, and German as can be seen in Panel B of Table A3. Two-thirds of movies have a single language and almost 20% feature two.

We classify domestic and foreign films depending on the country of origin of financing. A movie is considered a home movie if the country of screening is listed among the countries of financing and foreign if there are foreign countries among the financiers. A movie financed both domestically and by foreign sources appears in both classifications. Around 75% of films have a single country listed as the origin of financing. 40% of the movies in our revenue and 36% of movies in our screening sample receive financing from the US. The three Panels in Figure A4 provide an overview of the screening of local and foreign movies and their success across countries. Panel A presents the log of the number of domestic movies screened on the x-axis and the log of the number of foreign movies screened on the y-axis in our screening sample. Except for the US, all countries are above the 45-degree line, indicating that only in the US there are more domestic than foreign movies screened. A similar pattern is observed in Panel B when we look at the log total revenue of domestic versus foreign movies. Here, India and China are below the 45-degree line as well, indicating that the box office of domestic films exceeds the foreign one.[8] When we plot in Panel C the log average revenue per movie for domestic versus foreign ones, it turns out that in the majority of countries, local movies on average generate more revenue. A notable exception is France where, despite the generous subsidies from the French state, the average French movie has underperformed at the box office.

Online Appendix (OA) Table C1 provides a detailed overview of the countries in the

[8]Note the IMDb revenue data for the US refer to gross box-office revenue from US, Canada, and Puerto Rico.

respective samples. OA Figure C1 shows that during our period most of the global box office came from the United States.

Figure A5 compares the genre distribution of films financed domestically versus those receiving financing only from abroad for the top-10 film financing countries. Most countries produce similar genre distributions. Some patterns that stand out are that movies receiving funding from Italy are rather unlikely to be action movies and more likely to be comedies. On the contrary, movies receiving financing from Japan are much more likely to be action, animation, and adventure movies. The Indian movie industry seems to specialize in drama, romance, and action movies.

2.2.1 How different are box office patterns across countries?

In an attempt to gauge the variation in movie consumption patterns across the globe, we construct for each country the genre-specific share of the cumulative 1995-2019 box office (dropping countries with less than 100 observations). Adventure is the most popular genre followed by action movies, with the median country spending 49% and 44% of its box office, respectively. This is not surprising considering that the most widespread motif in Berezkin's work, k27, depicts a character that needs to complete potentially deadly tasks, see OA Figure C2, and the discussion in Campbell (2008). Comedy comes in third with 38% of the revenue, whereas crime and mystery movies account for 11% and 7% in a given market on average. The cross-country variation in these patterns is significant; see Panel A of Figures A7, A8, A9, and A10. In Italy and France, for example, action movies bring about 30% of film revenue, with the corresponding statistics in Indonesia, Malaysia, and Oman exceeding 70%. Crime (Mystery) movies have a large audience in countries like Qatar, Kenya, and India (Greece and Jordan) accounting for more than 20% (10%) of the box office, whereas in countries including Chile and Bolivia (China and Kenya) similar movies account for less than 7% (3%).

In what follows, we offer suggestive evidence linking this variation to facets of the countries' folklore, motivating the use of formal text analysis tools. The idea behind our exercise is straightforward. When films incorporate elements and plots from a country's folklore, they can resonate deeply with the audience from that culture. This resonance can create a sense of familiarity and connection, making the movie more appealing and relatable to the audience.

Generally, action movies involve stunts, battles, and destructive crises. Is it the case that audiences whose oral traditions commonly feature devastating episodes find action movies more appealing? In Panel B of Figure A7 we plot the relationship between the share of devastation-related motifs in the country's folklore and the share of action movies in the country's box office during the 25-year period in our data.[9] There is a strong positive

[9]MX use ConceptNet to get at the related terms of each word and based on these construct the intensity of each concept in a country's oral tradition, i.e., the number of motifs that mention the related terms divided by the total number of motifs in the society's folklore. The term devastation, for example, is related to destruction, damage, catastrophe, ruin, and disaster, and the average country has 2.3% of its

relationship. In a similar vein, countries steeped in oral lore rich in crime, mystery, and explorer-related themes display a stronger preference for crime, mystery, and adventure movies, respectively. See Panel B in Figures A8, A9, A10.

We apply a similar logic to explore how the same film performs across different markets. Take the movie "Ice Age: The Meltdown", for example, which centers on the travel experiences of its main characters. In Panel A of Figure A11 we plot the share of the country's box office in 2006 (when the movie was released) against the share of motifs in the country's folklore with travel-related motifs. The association is evident. In countries like Hungary where travel-themes are common in the country's folklore, the film was very much embraced by the locals compared to the lukewarm performance of the same film in countries like Thailand and Vietnam where travel-related themes are much less common. A similar picture emerges when we focus on the performance of the film "Big Fish," which revolves around a son's attempts to understand his father's life. People with oral traditions where sons are central in the local story-telling flocked to the theaters.

The examples above suggest that when a movie touches upon *broad* themes that are popular in a society's folklore may generate more revenue. An alternative mapping between motifs and movies is instead of looking at general topics to directly compare movie and motif plots. This approach is motivated by the fact that motifs often take the form of a triplet involving a character, an item, and an incident. Hence, instead of exploring similarity in terms of broad terms (like sons, devastation, mystery, etc,) one may look at the similarity between plots. Panels A, B, C, and D of Figure A12 help illustrate the point. The y-axis reflects for a given film the share of the country's box office it generated in the year of its release and the x-axis shows the share of the population in the country with a specific motif (partialling out the log number of movies screened that year). Panel A plots the box office of "Final Destination", a horror movie centered on a sequence of murders, against the motif l4 that describes a character killing in succession. Panel B explores the success of "Eat Pray and Love", a film released in 2010 that features a woman traveling after her divorce and discovering love, as a function of the prevalence of motif k107, which features a woman in search of her magical husband. Panel C looks at the film "Wonder Park" released in 2019, featuring a curious and adventurous young girl going on an incredible journey, and its relationship to motif k125 which describes a fearless youth that tries frightful experiences in an attempt to know what fear is. In Panel D we show that the film "Puss in the Boots" released in 2011 found a greater audience in countries with a similar plot in their folklore, i.e., weaker animals outsmarting stronger ones (motif m109b). Across all these examples, motif plots that closely track a movie's plot seem to bring in higher revenue than otherwise.

motifs with such images.

2.3 Constructing distances between movie descriptions and motifs

The discussion above suggests that a movie's success may be partly explained by how its themes and plot compare to a society's ancestral stories. To move beyond arbitrary choices of specific concepts and motifs we leverage machine learning algorithms, to systematically evaluate the extent of this overlap and assess whether the shared elements across tales, legends, myths, and movie plots might influence a film's reception and success within different societies.

There are various NLP metrics that one might employ to construct distances between texts. We employ two that loosely follow the intuition behind the examples we constructed above. In what follows, we outline how we reduce the text dimensionality and how we use the movie and motif summaries to draw comparisons. First, we reduce the motifs and movie descriptions to embeddings using the Universal Sentence Encoder. We then compute the cosine similarity of the resulting vectors capturing the semantic resemblance between documents (see, e.g., Kenter and De Rijke 2015). Second, we summarize the documents into topics we back out using the Latent Dirichlet Allocation and then compute the dissimilarity between documents based on the Jensen-Shannon distance (see, e.g., Jing, DeDeo, and Ahn 2019). The technical details for both methods are provided in the Appendix B.

Our text database consists of 33,528 movie plot outlines (according to IMDb the latter provides a short and concise description of the storyline) and 2,564 motifs.[10] For the motifs we rely on MX. Each document representing a motif includes the title and both the English description and Russian description translated into English using the Google API.[11]

Universal Sentence Encoder

The Universal Sentence Encoder (USE) developed by Cer et al. (2018), based on the notion of attention developed in Vaswani et al. (2017), uses a neural network architecture and is a versatile model designed for various natural language processing tasks, including text classification, sentiment analysis, and semantic similarity measurements which is ideal for our setting. It encodes text into a 512-dimensional vector that captures contextual and semantic information, making it a tool for understanding the meaning of text beyond simple keyword matching. The idea is to design an encoder that summarizes any given sentence or paragraph to a 512-dimensional embedding reflecting a numeric representation of a document in the form of a vector of real numbers that encodes meaningful semantic

[10] To get at this number we take out movies with non-English summaries or outlines with less than five tokens. Tokens are single words, or two- or three-word combinations. See the text cleaning described in Appendix Section B.2.

[11] The reason for including all three elements is twofold. First, motif descriptions tend to be relatively short with a mean (median) of 27 (24) words. Second, the titles offer additional content and the separately translated descriptions often differ in their wordings as well. Thereby, we can increase the mean (median) number of words to 53 (58) achieving less sparsity in the document-term matrix which helps the LDA deal with the shorter motifs.

information. The 512 dimensions can be interpreted as latent factors. It is pre-trained on a variety of data sources, including Wikipedia, web news, web question-answer pages, discussion forums, and news corpus, encompassing a wide range of topics, languages, and writing styles.

Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan (2003), is a statistical model used to uncover the hidden thematic structure of documents. It assumes that documents are mixtures of topics and topics are mixtures of words. Through a generative process, LDA identifies the latent topics that are most likely to have generated the observed documents and the distribution of words associated with each topic. The final output is a probability distribution of topics over documents and a probability distribution of tokens over topics.

We settle on 20 topics for which the top keywords are presented in Table B1.[12] In Figure 1 we present the 1,000 most prominent tokens for two topics out of the 20, both notably featured in folklore narratives. Specifically, topic 0 is about friendships and relationships, featuring tokens like 'friend', 'meet', and 'date'. For instance, the Cinderella (2015) movie plot loads relatively highly on topic 0. Conversely, topic 1 is characterized by elements of mystique, with terms like 'mysterious', 'creature', 'giant', and 'supernatural'. A movie loading relatively strongly on topic 1 is Godzilla: King of the Monsters (2019).

The distribution of the average topic shares by genre is presented in Panel C of Figure 1. Romance and drama movies tend to load on similar topics, such as topics 3, 7, and 12. However, unlike drama, romance loads strongly on topic 0. Animation, fantasy, and sci-fi also load on similar topics, such as topic 10.

The advantage of the LDA is that it does not depend on pre-training and both the method and the output are transparent and easily interpretable. The advantages of USE are that the text requires no pre-processing, USE extracts more information from each document due to the extensive pre-training and unlike LDA, the order of words matters, and it is designed for short documents. Moreover, these embeddings can capture the nuances of sentence meaning, facilitating improved understanding and processing of natural language. Hence, in our context, USE gets closer to unpacking the plot of a narrative whereas the LDA gets at the broad themes of our movies and motifs. The USE, however, comes at the cost of being less transparent, see Ash and Hansen (2023) for a discussion about some of the trade-offs involved in using Large Language Models.[13]

[12] We further investigate the robustness of our main results to varying numbers of topics for which the top keywords can be found in OA Tables C2-C6.

[13] In essence, the technical distinction lies in USE's use of deep learning for masked prediction of sequential data to create nuanced, continuous embeddings that capture a wide array of semantic features while also considering the sequence of words, versus LDA's probabilistic modeling approach that identifies discrete topics based on word distributions and co-occurrences. In summary, while LDA categorizes the movies based on the distribution of topics derived from their content, USE captures the overall semantic and emotional essence of the motifs' or movies' descriptions. This distinction illustrates the complementary strengths of these two approaches in text analysis and natural language processing.

Figure 1: Topics



Notes: In Panels A and B the word clouds feature up to 1,000 most prominent tokens in two out of 20 topics. The size of the tokens is proportional to the probability within the topic. The location contains no information. In Panel C the x-axis displays the topic number and the y-axis the main genres. The shade of the the cells indicate the average topic shares.

In Appendix B we offer a more detailed comparison and illustration of the two approaches, take a deeper dive into the distribution of embeddings across genres, and discuss how topics and embeddings correlate with each other across motifs and movie descriptions.

Movie-motif similarity

How similar are motifs and movies based LDA and USE? Before we turn to our statistical measures we provide some illustrations. In Figure B1 we plot the 512 embeddings (left) and 20 topics (right) for all motifs and movies reduced to a two-dimensional space using T-distributed Stochastic Neighbor Embedding (t-SNE), developed by Van der Maaten and Hinton (2008).[14] We also represent the Cinderella (2015) movie with a red cross and motif

[14] t-SNE is a non-linear machine learning algorithm that is used for exploratory data analysis and visualizing high-dimensional data. The approach is comparable to principal component analysis with the advantage that it allows for non-linearities.

k57 titled also Cinderella in Berezkin's catalog through a black X.[15] On the left Panel, we see motif k57 and the homonymous movie are projected in the exactly same location according to USE despite not having the same actual underlying descriptions. On the right Panel, the Cinderella movie-motif pair is also located nearby in the two-dimensional space reduction of the 20 topics from the LDA. Both methods are able to uncover the semantic similarity of the example.

We now provide some examples of movies and motifs that are considered to be similar when comparing embeddings from USE. The movie 'Immortal' has the following description: "In the distant future, Earth is occupied by ancient gods and genetically altered humans. When a god is sentenced to death he seeks a new human host and a woman to bear his child." The motif that is considered closest is motif k25: "Woman from sky-world marries mortal man; A man gets a woman connected with the upper world she becomes his wife." For the movie the 'Furious 7' with the description "Deckard Shaw seeks revenge against Dominic Toretto and his family for his comatose brother." the closest motif is j5: "Brothers as victims; Two brothers are killed by antagonists. The avenging hero is the son of one." For the movie in 'Jumanji: The Next Level' the description reads: "In Jumanji: The Next Level, the gang is back but the game has changed. As they return to rescue one of their own, the players will have to brave parts unknown from arid deserts to snowy mountains, to escape the world's most dangerous game." The motif found to be closest is motif k10e: "Rescued People; Getting to a nest of a monstrous bird, a man finds there kidnapped people, helps them return home." For the movie 'Aladdin' with the description "A kind-hearted street urchin and a power-hungry Grand Vizier vie for a magic lamp that has the power to make their deepest wishes come true." the closest motif k163 reads: "Aladdin and his lamp; A magician orders a boy to fetch a magic object (often a lamp). The boy finds the object (but refuses to give it to the magician). A helpful genie appears and fulfills the boy's wishes." These examples provide suggestive evidence that generating similarity measures via USE provides a credible measure of plot relatedness.

We now describe in detail the two measures we mainly rely on to quantify the similarity between each movie in our database and a country's folklore.

Cosine similarity

Cosine similarity measures the angular similarity between two vectors. The cosine similarity between two vectors of embeddings $\mathbf{A}$ and $\mathbf{B}$ is defined as:

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

[15]The Cinderella's movie plot reads: "When her father unexpectedly dies, young Ella finds herself at the mercy of her cruel stepmother and her scheming stepsisters. Never one to give up hope, Ella's fortunes begin to change after meeting a dashing stranger." The description of motif k57 reads: "A girl who conceals her beauty and/or is poor and oppressed by her stepmother puts on a splendid attire and comes incognito to a feast where a man of high status falls in love with her. He marries her after identifying her by an object given to her or lost by her or (rare) seeing how she changes her clothes". This example highlights how effective is USE in identifying the similarity across short plot descriptions.

where:

$\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the Euclidean norms of the vectors. The result of the cosine similarity ranges from -1 for vectors that are diametrically opposed to 1 for vectors that are identical in orientation.

Jensen-Shannon distance

For the similarity between the topic distributions of movies and motifs we use the Jensen-Shannon distance (JSD). The JSD is a symmetrized and smoothed version of the Kullback-Leibler divergence (KLD). Unlike the KLD, the JSD is always finite and symmetric, defined for two probability distributions P and Q as the square root of the Jensen-Shannon divergence. This divergence is the average of the KLDs from a third distribution M, which is the pointwise mean of P and Q, to P and Q themselves. Mathematically, the JSD is given by:

$$\mathrm{JSD}(P \parallel Q) = \sqrt{\frac{1}{2}\mathrm{KLD}(P \parallel M) + \frac{1}{2}\mathrm{KLD}(Q \parallel M)},$$

where $M = \frac{1}{2}(P + Q)$ and the KLD is defined as:

$$\mathrm{KLD}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

Here, $\mathcal{X}$ represents the set of movies and motifs over which the distributions are defined. The JSD ranges between 0, when the distributions are identical, and 1, when they have disjoint supports, thus providing a bounded measure of the distance between two probability distributions.

In Figure A6 we show a radar plot of the average topic shares across all motifs (red) and all movie descriptions (blue) across the 20 topics. We see that topics 1 (related to mystique), 12 (related to adventures), and 15 (related to hero tasks) are more prevalent in folklore than in movies. Panel A shows unweighted averages, while in Panel B the motif topics are weighted by real GDP in the year 2010 according to the Penn World Tables (Zeileis 2023) of each country in which they appear and the movie topics are weighted by each movie's total revenue. The distances between movies and motifs calculated by JSD are 0.0181 and 0.0145, respectively. Hence, the weighted distance is 20% smaller.

The measure captures the information radius (total divergence to the average) or the amount of information lost when one distribution is used to approximate another. For topics the JSD exhibits a strong negative correlation with the cosine similarity and the Bhattacharya overlap, a measure of overlap between two probability distributions.

We residualize the similarity/distance between a motif and movie using the log number of words in the movie description, in the motif description, and the interaction of the two. We also present robustness checks to show that the results do not hinge on residualization. Our final residualized embedding-based cosine similarity and topic-based JSD of all movie-motif pairs exhibit a relatively low correlation of -0.077 suggesting that these two measures

are complementary in capturing how close a motif is to a movie plot.

Constructing the movie's similarity/distance to a country's oral tradition

Once we have the similarities and distances between each movie-motif pair, we aggregate these to have a summary measure comparing a country's folklore to a movie. In order to calculate the similarity between movie i and the j motifs in a country, k, we take the weighted average of all cosine similarities c_{ij}, while weighting each cosine similarity c_{ij} by the share of the population in k that has motif j in its folklore. Defining F_k as the set of motifs present in the folklore of a country k with associated population shares p_{jk}, the aggregated cosine similarity C_{ik} between movie i and country k is:

$$C_{ik} = \frac{\sum_{j \in F_k} p_{jk} \times c_{ij}}{\sum_{j \in F_k} p_{jk}}.$$

We apply an analogous procedure when we compute the aggregated distance based on the individual Jensen-Shannon distances. These measures reflect the proximity/distance of a movie to the *average* motif in the country's folklore. We experiment with alternative measures in the robustness section. In OA Figure C5 we show the raw and aggregated distributions of similarities/distances. In Section B.4 we describe for the main film-producing countries how the domestically financed movies compare to the foreign ones in terms of their distance to the local folklore over time.

3 The relation between global box office performance and folklore

In this section, we start by asking whether a movie's global box office may be explained by how widespread are its themes in different markets. We then examine how the similarity of a movie's plot to a country's folklore influences the revenue it generates and the probability of screening.

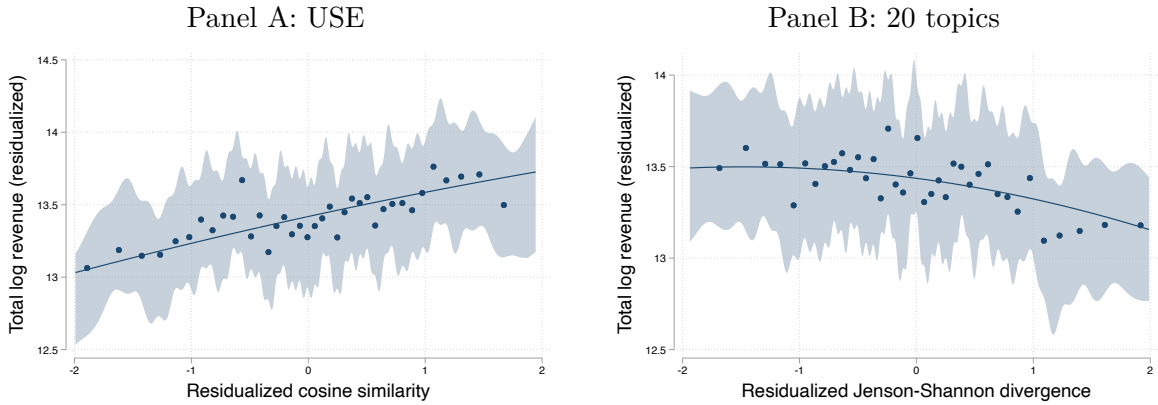
In the aggregate

Are blockbuster movies centered on universal motifs? To answer this question one needs to quantify what universal means. To answer this question, we construct for each movie its similarity/distance to the motifs across our 92 countries worldwide based on the cosine similarity of the embeddings (and JSD for the topics). We weigh each motif by the countries' population share and the country by its share of global GDP in the year 2010.[16] The resulting metric captures how much each movie resonates with the typical movie-goer in our global sample.

[16]For example, if the US produced 20% of global GDP in 2010 and the motif is shared by half of the US population, then the cosine similarity of the motif receives the weight 0.2×0.5. The aggregated cosine similarity is then scaled by the sum of all weights.

In Figure 2 we plot the relationship between the log of movie's i global revenue y_{it} released in year t and its cosine similarity to the global folklore in Panel A and the JSD in Panel B, partialling out year-of-release constants.[17] For both measures, the relationship is significant, suggesting that movies close to global folklore generate more revenue. Estimating a linear relationship suggests that a one standard deviation increase in the proximity of a film to global ancestral narratives based on the residualized cosine similarity of the embeddings translates into 0.1782 log points (standard error 0.0295) higher revenue. For the distance based on 20 topics from the LDA the corresponding coefficient is -0.0943 (0.0219).

Figure 2: Binned scatter plot of a movie's proximity and distance to global log(revenue)



Notes: Controls include year-of-release specific constants. Cosine similarity and JSD are residualized using the log(number of words) of the movie summary, motif, and their interaction. Shaded area represents 95% confidence bands and the fit is based on a polynomial of degree 2.

Across countries

We now shift our focus to the main unit of analysis, namely movie-country pairs. Doing so has several advantages which we detail below.

Our benchmark specification reads:

$$y_{ijtm} = \alpha + \beta C_{ij} + \gamma X_{ij} + \delta_{\mathbf{i}} + \eta_{jtm} + \varepsilon_{ijtm}, \tag{1}$$

where i reflects a movie, in country j, in year t and month m. The outcome variable is the log(revenue) or a binary indicator that captures whether the movie was screened in a given country. The main independent variable is the similarity/distance of a movie to the society's oral tradition. A movie is a multidimensional object that includes its budget, director, actors, setting, and genre, among other aspects. Hence, a cross-movie

[17] For all binned scatter plots we use the Stata command *binsreg* developed by Cattaneo et al (2023). We set the x-axis range within 2 standard deviations from the mean, the degree of polynomial and smoothness constraints for the construction of the 95% confidence bands are each set to 3, and the polynomial for the regression line is of degree 2.

comparison is difficult to interpret. The inclusion of a vector of film-specific constants, δ_i, allows us to absorb all film dimensions that predict a movie's success across all markets and single out the role of plot similarity to a country's ancestral narratives. Moreover, accounting for country-year-month of screening fixed effects η_{jtm}, we compare a movie's revenue to other films released locally during the same time of the year in that market, capturing the choice of films local movie-goers had access to during this window.

Table 1 presents the results. In column (1) of Panel A, we compare the revenue of films released in the same country-year-month across 85 countries between 1995 and 2019. In each of the 13,449 country-year-month bins, there are on average 13 movies. The estimated coefficient suggests that a one-standard-deviation increase in movie-folklore similarity based on the USE embeddings translates into roughly 0.20 log points higher revenue in a given period. To get some idea about the magnitude it is instructive to compare it to how much a movie's country-specific log(revenue) responds to a movie's budget. A one standard deviation increase in a film's log(budget) increases revenue by 0.97 log points.[18] This means that the additional revenue coming from a one-standard-deviation increase in movie-country folklore similarity is comparable to a 0.20 standard deviations increase in a movie's budget, i.e., approximately a 10 million US$ increase.

In column (2), we add 13,929 movie constants. Doing so, the R^2 jumps to 78% increasing by 44 percentage points (pp), and the coefficient of interest declines somewhat to 0.17 log points. This reveals that the estimated strong association between a movie's financial success and a country's folklore is not driven by comparing different movies across different markets. In column (3), we add a vector of 2,916 distributor constants. Despite the drop in sample size by 43%, because of missing information and distributors with just one film, the association remains strong and precisely estimated. Comparing films released in the same market (defined as country-year-month) by the same distributor, effectively accounting for differences in network, promotion, and advertising efforts across film distributors, and taking into account a film's diverse attributes, those closer to the local narratives generate 0.13 log points higher revenue.

In columns (4) to (6) we split the sample by the origin of financing. In column (4) we exclude films that received any domestic financing, whereas in column (6) we focus on movies that received (at least some) funding from US entities. This sample includes Hollywood movies among others that constitute the lion's share of our database. In both samples, the coefficient appears quantitatively and statistically significant. When we focus on foreign-financed movies screened locally but excluding US films, the coefficient in column (5) is similar in magnitude, 0.16 log points, however, it is also less precisely estimated. This is because, in this sample of foreign-financed, non-US movies, the typical film is shown in less than five countries limiting the movie-country variation we leverage.

Panel B mimics the structure of Panel A. The only difference is that instead of using the USE similarity between a movie and a country's folklore, we use the distance between

[18] IMDb has budget estimates for about 40% of the movies in the dataset, with a median budget of 14 million, and a standard deviation of 51 million US$.

the topic shares estimated by the topic model. We find a similar pattern across all columns, except for column (5) where now the coefficient on movie-folklore distance is larger and precisely estimated.

Table 1: Log(film revenue) explained by proximity to a country's motifs

				Origin of financing		
				Foreign	non-US	US
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Residualized cosine similarity						
Universal encoder	0.1985***	0.1662***	0.1304**	0.1798***	0.1584	0.1505**
	(0.0206)	(0.0518)	(0.0537)	(0.0511)	(0.1515)	(0.0589)
Observations	168626	168626	97081	149942	42372	123933
R^2	0.3409	0.7838	0.8499	0.8169	0.7572	0.8377
Number of movies	13929	13929	9837	10751	8164	5580
Number of countries	85	85	84	85	71	84
Country-year-month FE	✓	✓	✓	✓	✓	✓
Movie FE		✓	✓	✓	✓	✓
Distributor FE			✓			
Panel B: Residualized Jensen-Shannon distance						
20 topics	-0.0972***	-0.1313***	-0.1125**	-0.1652***	-0.2898***	-0.1291***
	(0.0266)	(0.0389)	(0.0413)	(0.0374)	(0.0836)	(0.0406)
Observations	168626	168626	97081	149942	42372	123933
R^2	0.3356	0.7838	0.8498	0.8168	0.7573	0.8377
Number of movies	13929	13929	9837	10751	8164	5580
Number of countries	85	85	84	85	71	84
Country-year-month FE	✓	✓	✓	✓	✓	✓
Movie FE		✓	✓	✓	✓	✓
Distributor FE			✓			

Notes: OLS regressions. JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction, and are standardized with a mean of zero and a standard deviation of one. Standard errors double clustered at the country and year of screening level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

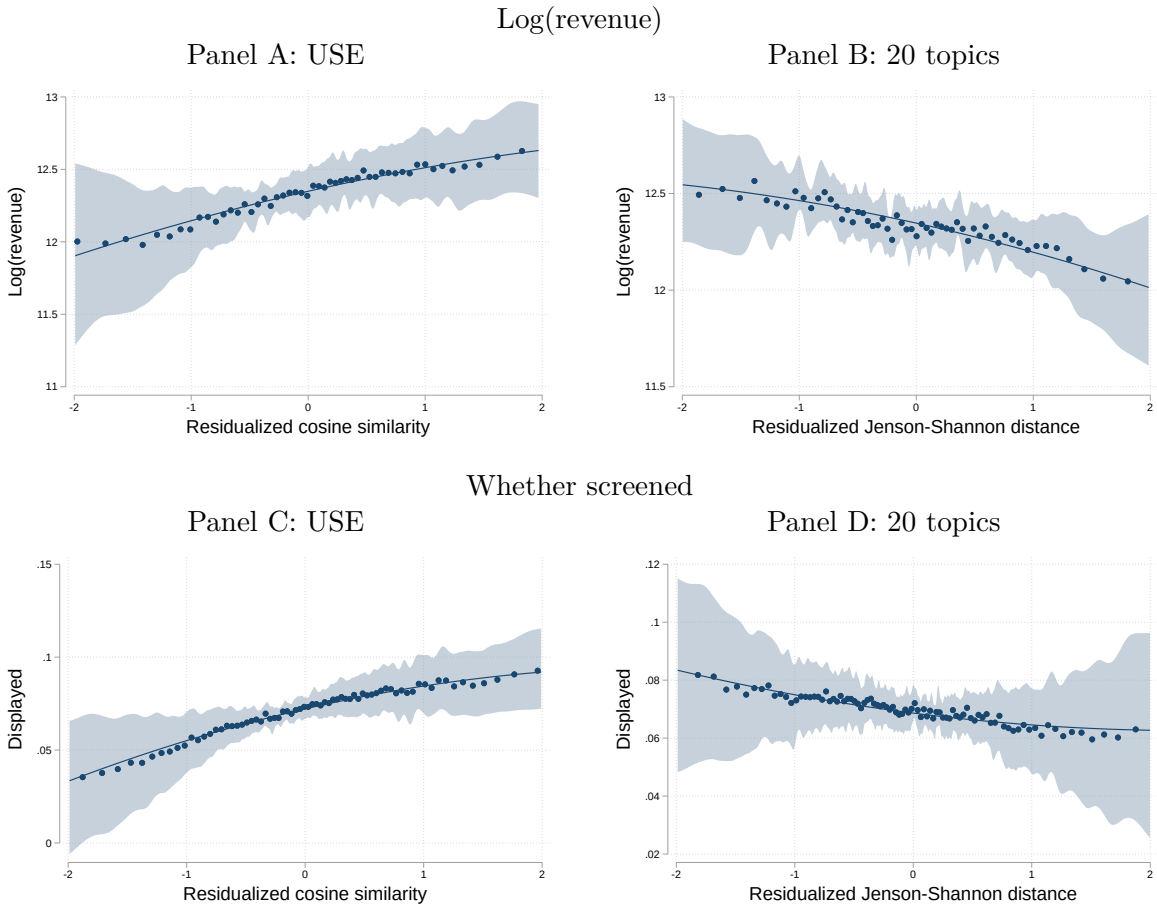
Figure 3 shows the binned scatter plots between the folklore proximity and distance measures to a film's log revenue corresponding to column (2) of Panels A and B of Table 1, respectively.

Likelihood of being screened in a country

Establishing that among screened movies those closer to the people's folklore generate systematically higher revenue, is consistent with a demand-driven interpretation, whereby film plots that resonate with people's cultural narratives garner a larger audience. Nevertheless, a movie is typically screened in a handful of countries, suggesting that we may trace whether the supply side of movies is also shaped by a movie's cultural resonance with the locals. Below, we use a binary indicator for whether a movie was screened in a given country. The benefit of this specification is that we can now look at a much larger sample of movies, 28,810, specifically, across 92 countries worldwide.

Our specification follows that of Equation (1) with two modifications. We control for

Figure 3: Binned scatter plots of a film's proximity to country's motifs and log(revenue)/ likelihood of screening



Notes: For Panels A and B controls include country-year-month and movie-specific constants. For Panels C and D controls include country-year and movie-specific constants. JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction. Shaded area represents 95% confidence bands and fit is based on polynomial of degree 2. The specification resembles that of column (2) of Tables 1 and 2 with standard errors double clustered at the country and year level.

country-year of release fixed effects for each movie and we omit the distributors' vector from the control list as there are no distributors among non-screened movies. In column (1) of Table 2 the estimated coefficient on the film-folklore similarity based on the universal encoder suggests that a movie is 1 pp more likely to be screened if it is one standard deviation closer to the local narratives. To put this magnitude into perspective, the baseline probability of a movie being screened across the 92 countries in our sample is roughly 7%. In column (2), we add a vector of 28,810 columns reflecting movie-specific constants. The R^2 increases by 28 pp (from 9.6% to 37.8%) reflecting that movies differ widely in their baseline probability of screening. Despite the large increase in explanatory power of our model, the coefficient of interest remains stable and precisely estimated.

The estimated coefficient is similar when we restrict the sample to films that did not receive any domestic financial backing, in column (3). A typical US-funded movie

is screened with a probability of 13.7% whereas for foreign, non-US-funded films the corresponding number is 3.2%. When we further split the foreign-financed movies, we find that the effect is stronger for US-funded (1.4 pp in column (5)) than for those without US support (0.6 pp in column (4)).

In Panel B, when we measure the movie-folklore gap based on the topics' distribution, we find that more distant movies are less likely to be screened, although the relationship becomes smaller in magnitude and insignificant once movie-specific differences in screening propensity are accounted for. One possible explanation is that film distributors may pay more attention to the specifics of a movie's plot deciding whether to bring it into a specific market.[19]

Table 2: Likelihood of screening and proximity to country's motifs

			Origin of financing		
			Foreign	Non-US	US
	(1)	(2)	(3)	(4)	(5)
Panel A: Residualized cosine similarity					
Universal encoder	0.0101***	0.0122***	0.0117***	0.0055**	0.0142***
	(0.0013)	(0.0034)	(0.0031)	(0.0025)	(0.0050)
Observations	2650520	2650520	2609825	1702092	948428
R^2	0.0960	0.3784	0.3907	0.1536	0.5036
Number of movies	28810	28810	28810	18501	10309
Number of countries	92	92	92	92	92
Country-year of first release FE	✓	✓	✓	✓	✓
Movie FE		✓	✓	✓	✓
Panel B: Residualized Jensen-Shannon distance					
20 topics	-0.0060***	-0.0049	-0.0048	-0.0021	-0.0072
	(0.0009)	(0.0045)	(0.0046)	(0.0014)	(0.0067)
Observations	2650520	2650520	2609825	1702092	948428
R^2	0.0950	0.3783	0.3907	0.1535	0.5036
Number of movies	28810	28810	28810	18501	10309
Number of countries	92	92	92	92	92
Country-year of first release FE	✓	✓	✓	✓	✓
Movie FE		✓	✓	✓	✓

Notes: OLS regressions (linear probability model). JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction, and are standardized with a mean of zero and a standard deviation of one. Standard errors are double clustered at the country and year of release level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Panels C and D of Figure 3 show the binned scatter plot between the similarity and distance of a movie to a country's folklore and the likelihood screening using the same controls as column (2) of Table 2.

[19] As we show in Section 5 below, film distributors tend to market movies that discourage risk-taking more frequently in countries with lower levels of entrepreneurship.

Extent of theatrical release and revenue per screening

So far we have seen that movies that resonate with local tradition are more likely to be screened and upon screening they generate stronger revenue. The latter might reflect more screenings, i.e., *more* theaters offering a given film for a *longer* period, as well as *more* people watching the movie given the extent of the theatrical release. How long a movie stays in theaters depends among other things on the film's popularity. Table 3 looks at these dimensions separately. In column (1) the dependent variable is the log of the sum of the theaters per week the movie is screened. For example, if in a given country a movie is screened in four theaters in week one and one theater in week two, the dependent variable in column (1) is log(5). In column (2) we look at the average revenue per theater-week. Both columns reveal that a film whose storyline is closer to the country's folkloric themes enjoys a wider theatrical release and generates more weekly revenue per theater.[20]

In the remaining columns, we move the analysis to the movie-country-week-of screening level. Doing so allows us to follow a film's performance during each week it is available. We use the following specification:

$$y_{ijtmw} = \alpha + \beta C_{ij} + \zeta_{jw} + \chi_{iw} + \delta_{\mathbf{i}} + \eta_{jtm} + \varepsilon_{ijtmw}, \tag{2}$$

where the unit of analysis is movie i in country j screened in year t and month m observed in week w of screening. We include ζ_{jw}, i.e. country-week fixed effects to control for the typical success any movie enjoys in its first, second, third, etc, week of screening in a given country. We further add χ_{iw}, i.e., movie-week fixed effects to flexibly control for a movie's success during each week of screening across the globe. The summary statistics for this sample are in Table A5. The median movie is shown in 104 theaters and its screening lasts for about 5 weeks.

In column (3) of Table 3 the outcome variable is the log of the total weekly revenue the film brings in a given country, while in column (4) we look at the log(revenue per theater screened). A movie that is one standard deviation closer to a country's ancestral narratives as measured by the cosine similarity between the embeddings brings in 0.15 log points more revenue, see column (3). In column (4) we show that this higher weekly revenue per film is not only because of more theaters screening the movie but the revenue generated per theater is also 0.11 log points higher. The estimated coefficients between local folklore and movies as measured by the JSD between topics in Panel B are on average smaller.

[20]Note that we do not have data on the number of daily screenings per theater.

Table 3: Extent of theatrical release, revenue per screening, and proximity to country's motifs

Unit of observation:	Movie-country		Movie-country-week of screening	
Dependent variable:	Total theater-week screened	Revenue/ total theater-week screened	Revenue	Revenue/ theaters
	(1)	(2)	(3)	(4)
Panel A: Residualized cosine similarity				
Universal encoder	0.1019**	0.0877**	0.1547**	0.1057**
	(0.0445)	(0.0319)	(0.0673)	(0.0401)
Observations	133374	133374	857788	857788
R^2	0.7909	0.7412	0.7996	0.7173
Number of movies	12667	12667	12660	12660
Number of countries	81	81	81	81
Country-year-month FE	✓	✓	✓	✓
Movie FE	✓	✓		
Week count-country FE			✓	✓
Week count-movie FE			✓	✓
Panel B: Residualized Jensen-Shannon distance				
20 topics	-0.0929**	-0.0564	-0.0826*	-0.0581**
	(0.0349)	(0.0356)	(0.0415)	(0.0268)
Observations	133374	133374	857788	857788
R^2	0.7909	0.7411	0.7995	0.7173
Number of movies	12667	12667	12660	12660
Number of countries	81	81	81	81
Country-year-month FE	✓	✓	✓	✓
Movie FE	✓	✓		
Week count-country FE			✓	✓
Week count-movie FE			✓	✓

Notes: OLS regressions. In columns (1) and (2) the unit of observation is a movie in a country. In columns (3) and (4) the unit of observation is the weekly outcome of a movie in a country. In column (1) the dependent variable is the log of the total number of theaters per week in a given country of film i. For example, if a movie is screened in four theaters in week one and one theater in week two, then the total number of theater-week screenings is equal to 5. In column (2) the dependent variable is the log of the ratio of total revenue divided by the total number of theater screenings. In column (3) the dependent variable is the log of the film's weekly revenue in a given country. In column (4) it is the log of the weekly revenue per theater in a given country in a given week. The week count variable takes values 1, 2, 3, etc. reflecting the number of weeks the movie is running. JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction, and are standardized with a mean of zero and a standard deviation of one. Standard errors double clustered at the country and year of screening level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

3.1 Robustness

In this section, we present a series of tests to gauge the robustness of the uncovered patterns. These include alternative movie-folklore similarity measures and additional movie-country characteristics. All robustness and heterogeneity checks appear in the Online Appendix (OA).

So far we have shown that the same movie performs differentially across markets depending on whether it resonates with the local narratives. One might wonder if the tastes of cinema audiences in different countries for actors, directors, and screenwriters also differ in a manner that results in the uncovered pattern. For this to be the case, it must be that the actors (least) favored in a particular country appear mainly in films that (do not) resonate with the nation's traditional stories. In OA Table D1 we present the results. In particular, we start from the movie-fixed effects specification that corresponds to column (2) in Table 1 and include interaction terms between country indicators and directors in column (2), actors in column (4), and screenwriters in column (6). To gauge the stability of our coefficient of interest in the odd-numbered columns we report the non-interacted estimates for each sample. In column (8) we also allow for movies distributed by the same film distributor to have a country-specific coefficient allowing those active in multiple markets to have a different network/penetration in each one. Buena Vista International film distributor, for example, in this specification has a country-specific coefficient for the films it distributes. Across the various permutations, the coefficient on film-folklore similarity/distance remains precisely estimated.[21]

A key consideration among people in the film industry is the week of the year a movie is screened, as popular movie-going times vary across markets. For the US this would be the time around Christmas, the Golden week in Japan, etc. To the extent that movies that resonate more locally are released during these popular weeks, this may contaminate inference. In our dataset, the vast majority of movies are released on Thursdays (12%), Fridays (48%), and Saturdays (39%). In column (1) of OA Table D2 we control for the exact calendar date a movie is screened for the first time within a country. That is we compare revenue generated across movies released in a given country at the same time of the year, for example, among those released in France on Friday, December 20th, 2013. Within these narrowly defined markets (and arguably saturated specifications), movies closer to the local narratives get a larger share of the box office.

Familiarity with a given narrative does bring more people to the theaters, but the relationship between proximity and revenue may not be linear, as people may enjoy some novelty as well. In column (2) of OA Table D2 we add a quadratic term for the distance/similarity to explore this possibility. The squared term enters negatively

[21] We interact with all countries, fixed effects for each of 629 directors in column (2), 1,815 actors in column (4), 1,113 writers in column (6), and 1,796 distributors in column (8). The criterion for inclusion is that they appear in at least five movies of the revenue sample. For the descriptive statistics we refer to Table A2. A movie can feature multiple actors, writers, or directors in which case each of the ones fulfilling the minimum threshold of five appearances is included.

suggesting some diminishing benefits of familiarity but it is not statistically significant. Note that in our sample the average distance/proximity estimated between a movie and country's oral tradition is moderate (between 0.4 and 0.7 for both measures). This might explain the absence of a strong concavity.

In the film industry, business leaders often talk about "comparables" to a given movie. This aspect is supposedly factored in the calculation of whether to bring a movie to a given market and convince local theaters to adopt it. To operationalize this concept, we compute the cosine similarity of a given movie to previously screened movies in the same market. We apply an annual discount factor of 10% and weigh each movie by the revenue it generated when we aggregate across past movies.[22] In column (3) of OA Table D2 we measure in Panel A (B) how much a movie's storyline aligns with (diverges from) the themes of films that have recently been screened in that market.[23] Two findings emerge from this exercise. First, the closer a movie's theme is to recently locally screened high-revenue generating movies the stronger its box office is. The estimated coefficients are large, with a one-standard-deviation increase in similarity to past successful movies increasing revenue by 0.71 log points. This magnitude helps explain why "comparables" are a focal subject in the film industry (although the quantification we offer is novel to the best of our knowledge). Second, the established positive relationship between revenue and proximity between a movie and a country's folklore continues to hold. This suggests that the market's movie screening history does not closely reflect the country's folkloric themes.

Finally, in OA Table D3 we show that the relationship holds for alternative measures of similarity (e.g. Bhattacharya overlap), alternative numbers of topics in the LDA model, and jointly extracted factors from the different measures. In Panels B and C, before constructing the movie-country's folklore similarity and distance measures we remove for each movie the top 1, 5, and 10 % similar motifs. This robustness assesses whether it is only the motifs closest to a movie that drive the observed association. This does not appear to be the case.

We run similar robustness checks for the probability of screening. For the other set of regressions in OA Table D4 we find little evidence of a concave relationship in column (1). When including a regressor capturing whether a movie is closer to previously screened movies in the same country, we find a strong and significant relationship in column (3). This suggests that not only the proximity to folklore makes the screening of a movie more likely, but also its similarity to previously consumed movies.

[22] In the US, documentaries such as 'Waiting for Superman' (2010), a documentary about the education system in the US, and 'Minimalism: A Documentary About the Important Things' (2016), are rather different from past consumed movies. Films estimated to be relatively similar to past consumption patterns are the American Hustle (2013) and The Bounty Hunter (2010).

[23] Carvalho, Draca, and Kuhlen (2021) rely on a similar approach to measure innovation using patent text.

3.2 Heterogeneity

In this section, we explore whether the uncovered link varies systematically depending on country and movie characteristics.

We start by looking at a movie's genre. Specifically, we run a separate regression for each genre with movie and country-year-month fixed effects as in Equation (1) for revenue and movie and country-year-of-release fixed effects for the probability of screening. In OA Figure D1 we plot for each genre the regression coefficients of the cosine similarity between a country's folklore and the movies' plot summarized into embeddings by the USE. Panels A and B plot the coefficients for log revenue and the probability of being screened, respectively. In both Panels, the estimated coefficients are positive. For screening probability, the relationship is precisely estimated within each genre whereas for revenue, it is within action, adventure, thriller, and drama movies that the link is statistically significant. In Panels C and D of OA Figure D1 we see that similar patterns hold for the JSD based on topics.

In Panel A of OA Figure D2 we plot the coefficients of the cosine similarity of the embeddings (and OA Figure D3 for the JSD) when we further split the sample by country and movie characteristics. On the top, we see that the positive relationship between similarity and revenue holds for movies screened before and after 2010 (the median year in our sample in 2011).[24] Next, it appears that the relationship is concentrated in country-year observations during which GDP growth is positive (which is the vast majority of our sample). This might seem unexpected but it has been observed that during recessions the movie industry booms, Cieply and Barnes (2009).[25] When we split our sample between democracies and non-democracies we do not find significant differences.[26]

In the last three heterogeneity checks, we look at how the relationship between a movies' storyline and its proximity to local folklore is mediated by salient movie characteristics. First, we explore the role of how prolific a movie director is. To do so, we start by dropping the movies whose directors appear only once in our sample. This reduces the number of observations by 16%. The median number of movies by directors in the remaining sample is 4. The association between movie-folklore similarity and revenue is stable above and below the median number of directors.

When looking at how the association differs for movies screened in more/less countries, we split at the median number of countries a movie is screened. While the association between similarity and revenue is positive for both splits, it is significant for movies appearing in more countries. In the latter sample, a movie is screened in at least five countries. At the bottom of Panel A of OA Figure D2 we look at whether the association between similarity and revenue depends on the movie's budget. Budget information is

[24] In OA Figure D4 we break down the coefficient for the revenue and the screening samples for each year from 2002 to 2019 separately. The coefficients are largely significant with no clear temporal pattern.

[25] To the extent that people during recessions visit the movie theaters more often would mechanically end up watching movies that they would not typically watch, producing the observed non-relationship.

[26] We use the Polity IV Project definition of democracy (Marshall and Gurr 2020). 83% of our revenue sample comes from democratic countries.

missing for most films. The plotted regression coefficients suggest the association is strong and precisely estimated for movies with missing such information and for those above median budget. For the movies below the median budget, the coefficient is smaller and less precisely estimated.

To see whether the positive relationship between revenue and movie-country folklore is driven by a subset of countries, we run one regression for each country separately. In these country-specific regressions, we naturally cannot account for movie-specific constants but we control for year-month of release, language, and film finance-origin-country-fixed effects while clustering the standard errors by year. In Panel A of OA Figure D5 we see that for all 75 countries with at least 100 observations the coefficient is positive, and for the great majority it is also significant. The starkest associations are more than 0.40 log points more revenue when the folklore and movie are one standard deviation closer to each other. Panel B of OA Figure D5 shows the coefficients from regressing the indicator of whether a movie was screened on the similarity between folklore and movies while controlling for year, language, and finance-origin-country-fixed effects. Almost all coefficients are positive and significant, reaching magnitudes of more than 3 pp for Russia and Mexico.[27] In OA Figure D6 we show that the coefficients explaining revenue and probability of screening within country are highly correlated across country.

In Panel B of OA Figure D2 (and Figure D3 for JSD) we present heterogeneity results for whether a movie was screened or not. The coefficients on movie-country folklore similarity are similar for movies screened before and after 2010. The relationship between the cosine similarity and the likelihood of screening is stronger during recession years compared to normal times, albeit being positive and significant for both subsamples. what is perhaps noteworthy is that for democracies the probability of screening for a given film is not systematically associated with its distance to the local folklore. This suggests that in democracies films that do not conform with the local ancestral narratives do not face a screening penalty. The relationship between similarity and the probability of screening is also greater for directors that make more movies, that are screened in more countries, and that have a larger budget.

4 Within US analysis

So far, the evidence reveals a strong association between a movie's box-office performance in the various international markets and its proximity to a country's ancestral narratives. In this section, we establish a similar pattern within the US.

Google Trends To overcome the fact that disaggregated movie revenue data across theaters in the US are proprietary, we use Google Trends. The Google Trends keyword research tool is publicly provided by Google. It enables users to see how often particular terms are searched on the Internet. The numbers range from 0 to 100 representing the

[27] For the JSD based on the topics the patterns in OA Figure D7 are not quite as clear.

search interest relative to the highest point on the chart for the selected region and time. A value of 100 is the peak popularity of the term, whilst a value of 50 means that the term is half as popular.[28] For the US, Google Trends offers a spatial breakdown of search interest across Designated Market Areas (DMAs). There are 210 DMAs in the United States, covering the entire continental United States, Hawaii, and parts of Alaska. DMAs are generally defined based on metropolitan areas, and the suburbs are often combined within. They usually encompass geographic areas where people receive the same local media, including radio and television stations, newspapers, and online content.

Our search parameters are the following: two alternative periods, namely, a 1-year window around the year a movie was released, and the entire 2004-2022 period, the region is set to the "United States", and the categories and search type set to "all categories" and "web search", respectively. Instead of using the movie's title, we employed the IMDb unique numerical identifier for each movie to ensure that the resulting search score reflected the interest in the specific movie rather than generic searches containing the same keywords as the movie's title.

In OA Figure C6 we show an example of a Google Trends query. The movie Wonder Woman, distributed by Warner Bros, made its global theatrical debut on June 2nd, 2017. The top Panel depicts how the movie's search intensity evolved in a 1-year window around its release year. The search interest peaked the week of June 4, 2017, the week following its global release. The bottom Panel shows how the interest varied between DMAs during this period. In the case of Wonder Woman a blockbuster movie, there is a positive interest score for 209 DMAs. Figure A13 shows the DMA-specific search intensity histograms across 5,956 films screened in the US between 2004 and 2019 according to the IMDb dataset. Armed with a measure of a film's popularity for each DMA, we now proceed to construct a measure of the local ancestral composition.

Constructing ancestry across DMAs To get at the peoples' ancestry we use the question asked by the American Community Survey (ACS) between 2005 and 2021. The question reads: "What is this person's ancestry or ethnic origin? For example, Italian, Jamaican, African American, Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, and so on." Ancestry is self-reported, and the ACS records up to two ancestries per individual. For those reporting multiple ancestries, we keep the one they mention first. We then attach the ancestral oral tradition to each respondent based on the reported country of origin or ethnic group in Berezkin's dataset.[29] Ancestry data from the ACS are at the PUMA level (Public Use Microdata Area). There are 2,351 PUMAs in the US. For our purposes, ancestry shares

[28] We omit from the analysis instances where the score equals 0 as according to Google Trends this value means that there is not enough data for this term in the specific territory.

[29] For example, we assign the Mexican folklore to those reporting ancestry "Mexican", "Chicano" and "Mexican American." To African Americans, we assign 50% of the African folklore based on the country of origin in Africa according to Putterman and Weil (2010), and the other half to the American folklore. For those with American Indian ancestry, we assign their oral tradition to the two largest tribes in the US today, the Navajo, and the Cherokee.

at the PUMA level are first calculated from micro-data using probabilistic weights. Then we intersect the two sets of boundaries and map the PUMA polygons into the DMA ones. We follow a similar procedure to construct DMA demographics using yearly IPUMS data from 2005-2021 at the PUMA level.

In Figure A14 we show the population shares across DMAs of Italian, German, Mexican, and Irish ancestry, respectively. The DMA of Hartford-New Haven has the largest share of Italian ancestry with more than 16% of its residents identifying as such. As expected, in the top 20 DMAs in terms of Italian ancestry, Boston, New York, Rochester, NY, and Providence appear. A significant portion of individuals with German ancestry is located in the northern United States, with almost half of Mankato's population in Minnesota identifying as German. People of Mexican descent are located mainly in the South and Southwestern parts of the country, appearing as the majority in at least five DMAs including Laredo, Yuma-El Centro, and El Paso-Las Cruces. Using the ancestral composition of a DMAs' population, we construct the composition of the local ancestral narratives by aggregating across the motifs of the various ancestries weighted by the share of each group.

Specification Armed with data on movie interest and ancestral motif distribution for each DMA we estimate the following specification:

$$y_{ijtm} = \alpha + \beta C_{ij} + \delta_i + \eta_{jtm} + \varepsilon_{ijtm} \tag{3}$$

where y_{ijtm} reflects the Google search interest of movie i in DMA j released in year t and month m. C_{ij} is the proximity of a movie to the DMAs' ancestral motifs based on the universal encoder. η_{jtm} captures DMA-year-month fixed effects, accounting flexibly to the overall interest in movies across DMAs across time. For example, Los Angeles registers a stronger interest in films than the average DMA, reflecting the underlying film industry. We finally include movie-specific constants, δ_i, as some movies are more popular than others across all territories. In Table 4 we present the results. Panel A looks at the search intensity in a 1-year window around the film's release, i.e. one year before and one year after the release, and Panel B looks at the average search intensity between 2004 and 2022. All variables are standardized so the estimated magnitudes are directly comparable.[30] A one-standard-deviation increase in the similarity of a film's plot to the local composition of ancestral narratives increases search intensity by 0.72 standard deviations. The estimated magnitude is large. To fix ideas, in Los Angeles, the search interest for films is 0.39 standard deviations higher than the average DMA. The estimated relationship is similar when we focus on movies released before and after 2013, respectively. When we split the sample in half by looking at high versus low population DMAs (the cutoff in our sample is

[30]In the presented results, search intensity is standardized at the aggregate level. Given that in each DMA the search intensity is relative to local other searches, one could argue for standardizing search intensity within each DMA, separately. Doing this only has minimal impact on coefficient magnitudes and does not effect precision noticeably.

roughly 1.2 million residents between 2005 and 2021) the relationship remains significant in both samples and the estimated magnitude is stronger for those that are relatively smaller. The pattern is similar when we measure interest in the 2004-2022 window in Panel B.

In Figure A15 we see the strongly positive association between cosine similarity and search intensity across DMAs using binned scatter plots related to the specifications estimated in column (1) of Panels A and B, respectively. In OA Figure D8 we present the coefficients from regressing search intensity on movie-folklore similarity for each DMA separately, while controlling for year-month, movie origin, and language fixed effects. The great majority of coefficients are positive and significant, with the largest magnitude of more than 0.2 standard deviations observed in Harlingen-Weslaco-Brownsville-Mcallen, Texas. Overall, the evidence suggests that the link between a movie's plot and its proximity to the local narratives uncovered across international markets is also found across designated market areas in the US.

Table 4: Movies' Google Trends interest across DMAs, local narratives, and values

	(1)	Since 2013 (2)	Before 2013 (3)	Low population (4)	High population (5)	(6)	(7)
Panel A: One-year window							
Residualized cosine similarity	0.7193***	0.7499**	0.6849***	0.9506***	0.4981**		
	(0.1612)	(0.2046)	(0.2007)	(0.2442)	(0.1725)		
Share self employed x Risk-taking pays off in movie						0.0096*	
						(0.0050)	
Female out of LF x Relative movie bias							-0.0031***
							(0.0010)
Observations	458368	238482	219886	230328	227631	436292	458368
R^2	0.6314	0.6462	0.6098	0.6834	0.6076	0.6359	0.6312
Number of movies	5598	2452	3146	5588	3704	5122	5598
Number of DMA	209	209	208	68	141	209	209
DMA-year-month FE	✓	✓	✓	✓	✓	✓	✓
Movie FE	✓	✓	✓	✓	✓	✓	✓
Panel B: Window 2004-present							
Residualized cosine similarity	0.8697***	0.8050**	0.9221***	1.0807***	0.6051***		
	(0.1502)	(0.2061)	(0.1452)	(0.2148)	(0.1793)		
Share self employed x Risk-taking pays off in movie						0.0117**	
						(0.0048)	
Female out of LF x Relative movie bias							-0.0024
							(0.0013)
Observations	555506	246385	309121	275839	279246	529880	555506
R^2	0.6949	0.6744	0.7089	0.7328	0.6808	0.6971	0.6947
Number of movies	5832	2469	3363	5828	3994	5337	5832
Number of DMA	208	208	208	68	140	208	208
DMA-year-month FE	✓	✓	✓	✓	✓	✓	✓
Movie FE	✓	✓	✓	✓	✓	✓	✓

Notes: OLS regressions. Google search interest scores are obtained from a one-year window around the release of a movie in Panel A (and the 2004-2022 period in Panel B) and are standardized with a zero mean and a standard deviation of one. The oral tradition of a DMA is based on the reported ancestry of the respondents (yearly average between 2005 and 20201). A movie's plot proximity to local narratives is the standardized cosine similarity of embeddings between ancestral motifs and a film's plot (residualized by the log(number of words) of the movie outline, motif description, and their interaction). In columns (4) and (5) the population cutoff that splits the estimation sample into two halves is roughly 1.2 million inhabitants. The sample restrictions are in the column headings. The variable "Risk-taking pays off in movie" takes values 0, 1, and 2, reflecting how risk is presented. The variable "Relative movie bias" is obtained by subtracting the male from the female bias. This variable ranges from -8 to 6 with higher values reflecting a more favorable representation of female over male characters. Female out of LF is the DMAs' share of women that are neither unemployed nor employed. The share of self-employed is the number of self-employed divided by those who either work on wages or are self-employed. The DMA demographics are yearly averages between 2005 and 2021 from IPUMS. Columns (6) and (7) are discussed in Section 5. Standard errors are double clustered at the DMA and year of release level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

5 Do we watch movies that reflect our values?

Estimating the distance/proximity of motifs to movies via topic modeling and sentence embeddings, respectively, has the benefit of leveraging all information available in a group's folklore. However, demonstrating a direct connection between societal norms and the consumption of films that promote similar values would clearly illustrate the relationship between (contemporary and historical) values and the consumption of entertainment products.

The array of cultural traits that can be considered is potentially large. We highlight two aspects that have received ample attention in the literature. Namely, risk-taking and gender roles. Specifically, we ask whether risk-averse societies are more likely to consume content that discourages risk and whether societies, where women have a limited role, end up watching films that depict females in similarly subordinate positions. Such a pattern would bring into the foreground a so far neglected aspect of cultural continuity, whereby historical narratives may continue to influence values and beliefs today by shaping contemporary entertainment preferences.

5.1 Risk-taking

Across countries Following MX we capture entrepreneurial activity across countries using two proxies. The number of patents filed by residents per 100,000 people (from the World Intellectual Property Organization Patent Report) and the number of new business registrations (limited liability corporations) per 1,000 people aged between 15 and 64 years (from the World Bank's Entrepreneurship Survey and database). Both measures (in logs) reflect the 2006–2018 averages. In terms of attitudes towards risk historically, we use the relative frequency of competitions and challenges that are more likely to be harmful than beneficial as coded by MX in the motif catalog. This is our measure of ancestral risk aversion.

Across films There is a lively literature in media studies and computer science on how to classify movies; see Mehal et al. (2021) and the references therein. In the absence of a systematic assessment of risk-taking behavior across films from IMDb, we resort to a Large Language Model (LLM). Specifically, we employ the ChatGPT API environment to classify movies in terms of risk-taking. The exact prompt we use reads: "How would an objective observer categorize the plot of movie 'X' directed by 'Y' in terms of risk-taking? Pick only one from the options below, and provide an explanation (max 60 tokens): 'Wins exceed Setbacks', 'Setbacks exceed Wins', 'Setbacks and Wins balance each other', 'No risk-taking behavior in the movie', 'Unknown movie'."

We use the gpt-4-1106-preview version (known as GPT-4 Turbo), see Törnberg (2023) on the advantages of ChatGPT-4 in annotation of text compared to experts and crowd workers. This is the latest version as of the writing of the manuscript and it crucially allows for the data generated to be reproducible by other researchers. Moreover, asking

for a concise justification of the classification reported allows one to interrogate the output and constrain hallucination (Ashwin, Chhabra, and Rao 2023).[31], [32]

The classification of movies in terms of risk-taking is summarized in Table A6. In 39% (41%) of movies of the revenue (screening) sample, taking risks leads to losses on average. The films' majority has a balanced representation of risk with losses and benefits balancing each other 50% of the time. It is in less than 12% of films where wins exceed losses. A movie's genre predicts how risk is represented. On the one hand, drama, horror, thriller, and crime films tend to highlight the losses that come with taking risks. On the other hand, sports, action, animation, adventure, comedy, and fantasy movies stress the upside of being bold and brave. Despite genres' predictive power, the variation in risk representation has a significant within-genre component. The R^2 of risk-taking on movie's genre is less than 30%. The portrayal of risk has been rather stable during our sample, with a small uptick in favorable representation of risk for movies produced after 2010.

Top movies in terms of global revenue where setbacks exceed wins include Avengers: Infinity War (2018), The Hunger Games: Catching Fire (2013), Jurassic World (2015), Star Wars: Episode III Revenge of the Sith (2005), King Kong (2005), Million Dollar Baby (2005), and Logan (2017).[33] Top-revenue films where risk-taking pays off feature, among others, Furious 7 (2015), The Rock (1996), Sing (2016), Olympus Has Fallen (2013), The Avengers (2012), and Mission: Impossible - Fallout (2018).[34]

Examples Before laying out the empirical specification, we offer two examples that motivate our inquiry. In Figure 4 we show the relationship between a country's willingness to take risks (reflected in its patent rate) and the fraction of the yearly box office that a particular movie generated in the year of its release. In the top panel, we look at two films: Mission Impossible - Fallout (2018) and Logan (2017). In the former, the character prevails despite the numerous challenges, whereas Logan highlights personal failings and regret, leading ultimately to the demise of the main character. More entrepreneurial societies embraced Mission Impossible but reacted adversely to Logan.

[31] We are currently working on gathering assessments on gender and risk for a subset of movies from research assistants to explore whether the GPT-generated responses fall within (or systematically deviate from) the distribution of the individual-generated ratings.

[32] Further prompting GPT to provide references that justify its assessment, the following sources are mentioned: (i) the movie's plot structure including its thematic elements and narrative arc (ii) film reviews and critiques from professionals as well as reviews from platforms including Rotten Tomatoes or Metacritic (iii) academic and analytical articles from film journals or media studies publications (iv) interviews and commentary from the director, cast and crew members, and (v) fan discussions and film forums including platforms like IMDb.

[33] The GPT-provided explanation for the classification of Logan (2017) reads: "Setbacks exceed Wins" because Logan faces numerous losses and sacrifices throughout the film, culminating in his own death. For the "Jurassic World", it reads: "the park's creation leads to chaos, destruction, and loss of life when dinosaurs escape containment."

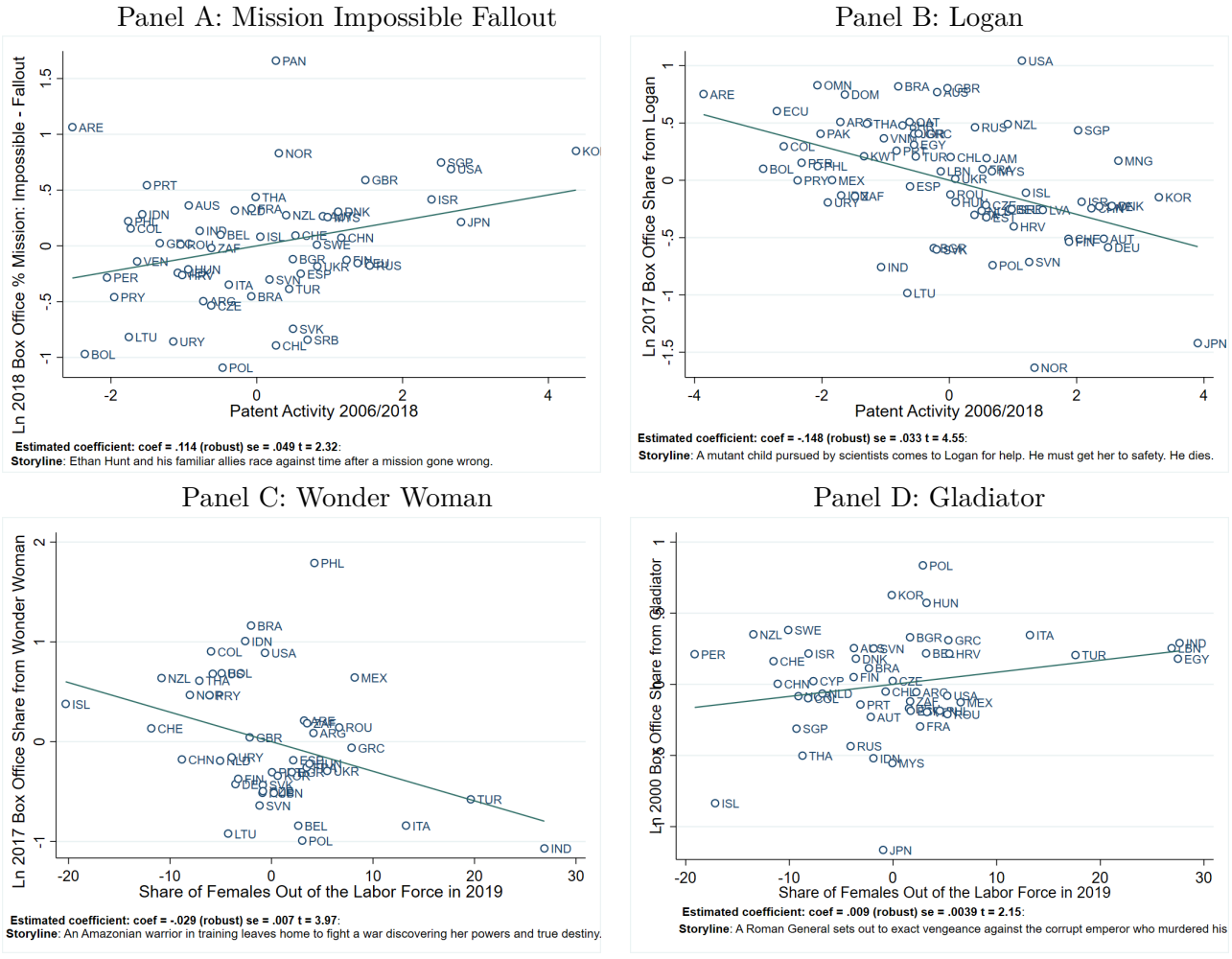
[34] The GPT-provided explanation for Sing reads: "The protagonist takes significant risks to save his theater, faces setbacks, but ultimately his risks lead to a successful show and a revitalized theater." The justification for Mission: Impossible - Fallout is: "Ethan Hunt's team ultimately prevails, thwarting nuclear threats despite numerous obstacles."

Specification We adopt the following empirical specification:

$$y_{ijtm} = \alpha + \beta Movie\ Value_i \times Country\ Value_j + \delta_i + \eta_{jtm} + \varepsilon_{ijtm}, \tag{4}$$

where y_{ijtm} is movie's i revenue (or whether it was screened) in country j in year t and month m, and $Movie\ Value_i \times Country\ Value_j$ is the value of the movie interacted with the value of the country. Country-year-month fixed effects, η_{jtm}, allow for comparisons of movies screened during the same time of the year in a given country, and a vector of movie-specific constants, δ_i, absorb all movie-specific differences in production, quality, budget, etc.

Figure 4: Box office and cultural values



Notes: The y-axis shows the log of the percentage of the yearly box office in a given country the movie in the subtitle brought in. In Panels A and B the x-axis shows the patent activity in a given country measured by the log of the total number of patents per 100,000 residents between 2006-2018 as a proxy for risk-taking. In Panels C and D the x-axis shows the share of women out of the labor force in a given country in 2019. In all plots we partial out the ln(number of movies screened in the country in the corresponding year).

Results for risk-taking In Table 5 Panel A (B) the outcome variable is a movie's standardized revenue (whether it is screened or not). Both panels have the same structure. In column (1), we introduce the film's disposition towards risk-taking. Comparing a movie with a plot that discourages risk to another that portrays risk in a favorable manner (a two-point increase in risk-taking), increases revenue by 0.57 standard deviations and the probability of being screened by 7.6 pp. The revenue (and screening) advantage of risk-encouraging films is further accentuated in markets with more entrepreneurial populations. The estimated magnitude of the interaction in column (2) suggests that an increase in patent activity by a standard deviation (about 1.79 log points), that is, moving from the patenting rate of Estonia (1.20 log points) to that of France (3.09 log points), increases the revenue of a film depicting bravery as rewarding, by an additional 0.087 standard deviations and the probability of screening by 3.7 pp. Column (3) reveals that the revenue and screening advantage of risk-promoting movies in less risk-averse markets is not driven by differences in film-specific attributes, as we compare the same film in different markets.

We find the same pattern when we interact the film's risk depiction with the rate of new business creation (standardized). A movie that centers on the rewarding aspects of risk-taking will generate systematically more revenue (and will be screened at a higher rate) in countries whose residents engage in larger numbers in the risky and potentially rewarding activity of opening a new venture (column (4)). This advantage, nevertheless, will be muted in countries where the local oral tradition depicts heroes as failing more often than not in the challenges they engage, see column (5). The magnitudes in column (4) of Panels A and B suggest that a one-standard-deviation increase in entrepreneurship (1.43 log points) increases a film's box office log revenue by 0.07 standard deviations and the probability of being screened by 2.4 pp. This increase in venture creation is equivalent to comparing Costa Rica's to Great Britain's business environment.

A similar picture emerges when we look at how the success of risk-promoting depends on the entrepreneurial activity across DMAs in the US. We modify Equation (4) replacing countries by DMAs. Tracing the same film across locations, it is in those with a higher rate of self-employment between 2005-2021, where (Google search) interest is more pronounced. The estimated coefficients in column (6) of Table 4 suggest that comparing two DMAs one-standard-deviation apart in terms of self-employment rates, a movie with a one-unit increase in its risk-portrayal will experience a boost in search intensity in the more entrepreneurial DMA by 0.0117 and 0.0096 standard deviations for the one-year and 2004-2022 search window, respectively.

5.2 Gender roles

Across countries To capture the role of women in a given country, we follow the literature and use the share of women outside the labor force (LF). Societies where females are less attached to the labor market have a limited role in other spheres, (Alesina, Giuliano, and Nunn 2013). We use the latest pre-pandemic numbers of 2019.

Table 5: Depiction of risk-taking in the movie, revenue, and likelihood of screening

	(1)	(2)	(3)	(4)	(5)
Panel A: Standardized log(revenue)					
Risk-taking pays off in movie	0.2849***	0.2848***			
	(0.0195)	(0.0181)			
Ln(# of patents per 100,000 people) x Risk-taking pays off in movie		0.0437**	0.0535***		
		(0.0168)	(0.0139)		
Ln(new businesses registered per 1,000 people) x Risk-taking pays off in movie				0.0350***	
				(0.0105)	
Risk aversion in motifs x Risk-taking pays off in movie					-0.0276**
					(0.0106)
Observations	149664	149664	149664	134529	152293
R^2	0.3843	0.3851	0.7940	0.7955	0.7927
Number of movies	10719	10719	10719	9865	10827
Number of countries	80	80	80	72	83
Movie FE			✓	✓	✓
Country-year-month FE	✓	✓	✓	✓	✓
Panel B: Likelihood of screening					
Risk-taking pays off in movie	0.0382***	0.0382***			
	(0.0038)	(0.0032)			
Ln(# of patents per 100,000 people) x Risk-taking pays off in movie		0.0185***	0.0185***		
		(0.0021)	(0.0021)		
Ln(new businesses registered per 1,000 people) x Risk-taking pays off in movie				0.0120***	
				(0.0031)	
Risk aversion in motifs x Risk-taking pays off in movie					-0.0100***
					(0.0028)
Observations	1730784	1730784	1730784	1573440	1789788
R^2	0.1275	0.1291	0.4195	0.4113	0.4136
Number of movies	19668	19668	19668	19668	19668
Number of countries	88	88	88	80	91
Movie FE			✓	✓	✓
Country-year FE	✓	✓	✓	✓	✓

Notes: OLS regressions. In Panel A standardized log(revenue) is the dependent variable, while in Panel B it is an indicator of whether a movie was screened in a given country. The variables Ln(# of patents per 100,000 people), Ln(new businesses registered per 1,000 people), and risk aversion in motifs are all standardized with a mean of 0 and a standard deviation of 1. Risk aversion in motifs is the relative frequency of motifs depicting competitions and challenges that are more likely to be harmful than beneficial in the country's oral tradition. The variable "Risk-taking pays off in the movie" takes values 0, 1, and 2, reflecting how risk is portrayed. Standard errors are double clustered at the country and year of release level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Across films MX employ M-Turks to classify gender roles into different stereotypes drawing insights from research on media. See Tobin et al. (2004) and Lauzen, Dozier, and Horan (2008) for the roles female and male characters assume in Disney movies and primetime television programs in the US.

The prompt we use is directly comparable to the one used by MX to classify gender roles in motifs. It reads: "How would an objective observer classify the portrayal of (fe)male character(s) in the movie 'X' by director 'Z'? Please choose from the following categories, and provide an explanation (max 60 tokens): Dominant/Independent, Submissive/Dependent, Physically Active, Engaged in Domestic Affairs, Sexual, Intelligent, Naive/Stupid, No Gendered Characters, Unknown Movie. Multiple categories may be chosen." We ask this prompt separately for each gender. Table A7 Panels A and B show the LLM-generated gender classification in the revenue and screening samples, respectively, focusing on films where both genders are present. Female characters compared to males are 15 pp more likely to be engaged in domestic affairs, almost 35 pp less likely to be physically active, approximately 25 pp less likely to be dominant, and by the same margin more likely to

be submissive. In terms of intelligence, the representation is balanced, whereas naive characters are almost 20 pp more likely to be males.

Adding the *dominant, physically, active*, and *intelligent* categories and subtracting the *naive, submissive*, and *engaged in domestic affairs* classifications we arrive at a measure of male and female bias, respectively.[35] Therefore, a higher value indicates a more favorable representation of each gender. By subtracting the male bias from the female bias we obtain for each movie the bias in favor of women relative to men. We call this the relative movie bias.

To fix ideas, in the movie The Avengers by Joss Whedon screened in 2012, male and female characters are portrayed with similar positive attributes, i.e., as dominant, physically active, and intelligent. Hence, the relative bias in this movie is equal to 0. Two movies, the Transformers: Age of Extinction and Transformers: Dark of the Moon, are films where the relative movie bias is -6. Women are depicted as submissive, sexual, and naive, whereas males are shown to be dominant, physically active, and intelligent. On the contrary, Run Lola Run screened in 1998 has a relative bias of 5. The female protagonist is depicted as independent, physically active, and intelligent, whereas the male character is dependent and naive.[36]

The average movie has a relative bias of -1.44 (-1.51) in the revenue (screening) sample. This skewed presentation of genders echoes that is found in the motifs across world cultures as OA Table C7 suggests. The latter reproduces Appendix Table 8 of MX and reflects the breakdown of gender roles in the folklore catalogue. The cross-tabulation of gender roles and risk-taking across movies reveals that relative gender bias is lowest (-0.97) in movies with a balanced representation of risk whereas it appears significantly more pronounced in movies either discouraging risk-taking (-1.86) or highlighting the benefits over the setbacks (-1.34).

A movie's genre classification has some predictive power in terms of how women and men are portrayed. On the one hand, relative bias is more than half a standard deviation lower in sport, western, crime, and action movies. On the other hand, women are relatively more favorably portrayed in family, animation, comedy, sci-fi, and adventure films. Nevertheless, there is significant within-genre variation. The R^2 of a regression of relative bias on the movie's genre is 11%. Over time the portrayal of women relative to

[35]More often than not the sexual categorization reflects stereotypically sexualized roles for females whereas when males are classified as sexual the explanation is more nuanced. Hence, we treat the sexual classification of female characters as a negative stereotype and for the males as neutral/positive one. This gender asymmetry in sexual representation can also be seen in the correlation structure of gender roles in Table A8. In movies where women appear as sexual, they are also systematically less likely to be independent and more submissive. For males, the pattern is less clear. They are less dependent and more dominant. Ignoring the sexual classification altogether or only assigning it as a negative stereotype to women delivers similar patterns.

[36]The GPT explanation for Run Lola Run reads: "Lola is portrayed as a strong, resourceful woman who takes action to alter the outcome of events, demonstrating physical endurance and quick thinking." and "Manni relies on Lola, makes poor choices leading to the plot." For the film Transformers: Age of Extinction the explanation provided is: "Female characters are often depicted as eye candy, reliant on male leads, and not portrayed as intelligent". "Male characters often take charge, engage in action-packed scenes, and display tactical and technical skills."

men has been improving. From 1995 to 2010 the average relative bias was -1.75 compared to -1.31 in between 2011 and 2019.

Top movies in terms of global revenue where women are more favorably depicted than men include The Hunger Games (2012), Three Billboards Outside Ebbing, Missouri (2017), Wonder Woman (2017), Frozen II (2019), Brave (2012), and Little Women (2019), among other films. Top-grossing movies with a negative relative bias feature Fight Club (1999), The Departed (2006), Rambo (2008), and The Wolf of Wall Street (2013).

Results for gender roles In the lower half of Figure 4 we illustrate the relationship between a society's gender norms and the box office success (failure) of a movie that agrees (clashes) with those norms. We focus on two films with very different gender representations. Gladiator's (2000) lopsided gender portrayal with dominating males and females playing secondary dependent roles found a robust audience in countries like Italy, Turkey, and India where women's role is limited in the labor market. The opposite pattern is true for Wonder Woman, a feminist movie, screened in 2017. Below we show that this pattern is not specific to the chosen examples but points to a generalized phenomenon.

In columns (1)-(4) of Table 6 the dependent variable is log(revenue) and from (5)-(8) whether a movie was screened. The estimated coefficient in column (1) implies that a one-unit increase in relative movie bias boosts revenue by 0.024 standard deviations. This means that for the average market in the sample, movies featuring a less biased gender representation generate stronger revenue. This revenue advantage, nevertheless, erodes in countries where women find themselves out of the labor force in larger numbers. Increasing women's out of the LF by 25 pp (a two standard deviation increase of the cross-country mean), that is moving from New Zealand's to Italy's female LF participation essentially negates this advantage, see column (2). In column (3) we add movie-specific constants, comparing the same movie across different markets. The interaction coefficient is stable and of the same magnitude as in column (2). When instead of the percentage of females out of the labor force, we use the relative role of men in the country's folklore from MX a similar picture emerges. The same movie that depicts women favorably relative to men underperforms financially in countries where ancestral stories depict women in subordinate roles. This finding reveals a remarkable consistency in narrative consumption across generations, suggesting that movies and entertainment products that do not align with local culture might encounter resistance.

In columns (5)-(8) of Table 6 we show that similar patterns, albeit less precisely estimated, are found when we explore how a movie's relative bias influences the probability of screening in a given market. Overall, movies with a higher relative movie bias (a one-unit increase in relative bias) are 0.4 pp more likely to be screened across markets but weakly less so in countries with lower female LP. In Table A9, we decompose a movie's relative bias into the female and male bias, respectively. Films portraying dominant and powerful women (men), i.e., have a higher female (male) bias, attract smaller audiences in countries where (i) women remain out of the labor force (column (1)) and (ii) with

a strong male bias in their folklore (column (2)). The opposite pattern is observed for movies with a stronger male bias. They sell more in countries where female LP is limited.

As a final exercise, we investigate whether gender roles play a role in shaping movie interest across DMAs in the US. We capture gender roles similarly by looking at the fraction of females outside the labor force (based on the IPUMS yearly data from 2005-2021) and estimate an appropriately modified version of specification (4). The coefficients in column (6) Panels A and B of Table 4 suggest that comparing two DMAs one-standard-deviation apart in terms of female labor force participation, a movie with a one-unit increase in relative movie bias will see a decline in its search intensity in the DMA with the lower female presence in the labor market by a modest 0.0023 and 0.0031 standard deviations for the one-year and 2004-2022 search window, respectively. It appears that even across local US markets today, the local populations are more interested in movies that broadly reflect their values.

Table 6: Gender representation in the movie, revenue, and likelihood of screening

	Log(revenue)				Screening			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Relative movie bias	0.0236***	0.0235***			0.0037***	0.0037***		
	(0.0058)	(0.0055)			(0.0008)	(0.0008)		
Female out of LF x Relative movie bias		-0.0100***	-0.0107***			-0.0006*	-0.0006*	
		(0.0018)	(0.0025)			(0.0003)	(0.0003)	
Male bias folklore x Relative movie bias				-0.0063**				-0.0005
				(0.0024)				(0.0003)
Observations	148279	148279	148279	150686	1611090	1611090	1611090	1628991
R^2	0.3574	0.3578	0.7947	0.7938	0.1305	0.1306	0.4203	0.4203
Number of movies	10281	10281	10281	10375	17901	17901	17901	17901
Number of countries	83	83	83	84	90	90	90	91
Movie FE			✓	✓			✓	✓
Country-year-month FE	✓	✓	✓	✓				
Country-year of first release FE					✓	✓	✓	✓

Notes: OLS regressions. Relative movie bias is obtained by subtracting the male from the female bias. This variable ranges from -8 to 6 with higher values reflecting a more favorable representation of female over male characters. In columns (1)-(4) log(revenue) standardized is the dependent variable, while columns (5)-(8) have a binary indicator for whether a movie was screened in a given country. Female out of LF is the country's share of women that are neither unemployed nor employed. The Male bias folklore variable is defined as the share of motifs where men are depicted as more dominant, less submissive, less engaged in domestic affairs, and more physically active than women in the country's oral tradition. Both the Female out of LF and Male bias folklore variables are standardized with a mean of 0 and a standard deviation of 1. Standard errors are double clustered at the country and film year of release level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

6 Conclusion

Our research bridges two fields that have evolved largely independently. The literature on historical legacies and contemporary (cultural) outcomes and the media studies program. The main insight is that societies have preferences over narratives that are rooted in their folklore and these preferences can be traced in how the same film fares in different markets. Specifically, we uncover a henceforth neglected link between ancestral narratives and the global and domestic box office. We utilize machine learning techniques to estimate the similarity between folklore motifs and movie plots and establish that films that align more closely with a society's traditional storytelling are more likely to be screened and generate higher revenue. This pattern holds when we exploit within-movie across-market variation both across countries and media markets within the US.

We further unpack the link between local ancestral narratives, contemporary values, and film consumption by exploring how a film's financial success depends on whether its moral message is in line with society's norms. We focus on risk-taking and gender roles. People with ancestral stories rooted in men's dominance and women's lack of independence and where women are less attached to the labor force today continue entertaining themselves by watching films with a similarly skewed representation of gender roles. Furthermore, less entrepreneurial societies today, with tales and legends portraying competitions and challenges as more harmful than beneficial, show a distinct preference for movies where playing it safe is the right thing to do. These findings underscore the deep-rooted impact of cultural narratives on entertainment patterns today, shedding new light on the interaction between traditional storytelling and modern media consumption.

We view our work as a springboard for further research. In light of our main finding that people consume entertainment products that reflect their cultural background, what type of narratives and at what age of exposure are more likely to be consequential for one's preferences and values? We have explored two cultural leanings of movies, namely risk representation and gender roles. Nevertheless, our AI-aided methods can be readily extended to reconstructing a comprehensive list of films' values, including how the young and the elderly are portrayed, how individualism and collectivism are represented, perspectives on diversity and tolerance, as well as the depiction of historically underrepresented groups.

Our study sidesteps film production considerations by focusing on how the same film performs in different markets. Naturally, unpacking the factors that shape movie production is vital. How do cultural entrepreneurs shape movie production? How is the film industry influenced by the prevailing institutional forces? The state, for example, is often directly involved in the industry through direct subsidies or via outright propaganda and censorship. What is the message of movies produced in autocratic versus democratic regimes? Moreover, tracing how the (domestic) film industry has evolved as a result of changes in bilateral movie flows and trade agreements, including the rise of the Korean and Turkish entertainment industries, can help quantify a country's soft power in the international arena (Antras and Padró-i-Miquel 2023).

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. (2015): "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," Software available from tensorflow.org.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015): "Radio and the Rise of the Nazis in Prewar Germany," *The Quarterly Journal of Economics*, 130(4), 1885–1940.

Adukia, A., A. Eble, E. Harrison, H. B. Runesha, and T. Szasz (2023): "What We Teach About Race and Gender: Representation in Images and Text of Children's Books*," *The Quarterly Journal of Economics*, 138(4), 2225–2285.

Alesina, A., P. Giuliano, and N. Nunn (2013): "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics*, 128, 469–530.

Ang, D. (2023): "The Birth of a Nation: Media and Racial Hate," *American Economic Review*, 113(6), 1424–1460.

Anikina, O. V., and E. V. Yakimenko (2015): "Edutainment as a modern technology of education," in *International Conference on Research Paradigms Transformation in Social Sciences 2014*, vol. 166, p. 475 – 479.

Antras, P., and G. Padró-i-Miquel (2023): "Exporting Ideology: The Right and Left of Foreign Influence," *working paper, mimeo Harvard University*.

Armand, A., P. Atwell, J. F. Gomes, and Y. Schenk (2023): "It's a Bird, it's a Plane, it's Superman! Using Mass Media to Fight Intolerance," Discussion paper, Université catholique de Louvain, Institut de Recherches Economiques.

Ash, E., R. Durante, M. Grebenshchikova, and C. Schwarz (2023): "Visual Representation and Stereotypes in News Media," *CEPR Discussion Paper 16624*.

Ash, E., and S. Hansen (2023): "Large language models for economic research: Four key questions," *VOX CEPR's Policy Portal*.

Ashwin, J., A. Chhabra, and V. Rao (2023): "Using Large Language Models for Qualitative Analysis can Introduce Serious Bias," Policy Research Working Paper 10597, World Bank Group.

Atkin, D. (2016): "The Caloric Costs of Culture: Evidence from Indian Migrants," *American Economic Review*, 106(4), 1144–81.

Bartlett, F. C. (1932): *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Berezkin, Y. (2015): "Folklore and Mythology Catalogue: Its Lay-out and Potential for Research," *The Retrospective Methods Network*, S10, 58–70.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent dirichlet allocation," *Journal of machine Learning research*, 3(Jan), 993–1022.

Blumenstock, J., O. Dube, and K. Hussain (2022): "Can Secular Media Create Religious Backlash? Evidence from Pakistan's Media Liberalization," *working paper*.

Bordalo, P., G. Burro, K. Coffman, N. Gennaioli, and A. Shleifer (Forthcoming): "Imagining the Future: Memory, Simulation, and Beliefs," *Review of Economic Studies*.

Bordalo, P., N. Gennaioli, and A. Shleifer (2020): "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 135(3), 1399–1442.

Campbell, J. (2008): *The Hero with a Thousand Faces (The Collected Works of Joseph Campbell)*. New World Library.

Carvalho, V. M., M. Draca, and N. Kuhlen (2021): "Exploration and exploitation in US

technological change," *Available at SSRN 3900479*.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2023): "Binscatter Regressions," .

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. (2018): "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*.

Cer, D., Y. Yang, S. yi Kong, N. Hua, N. L. U. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y. hsuan Sung, B. Strope, and R. Kurzweil (2018): "Universal Sentence Encoder," in *In submission to: EMNLP demonstration*, Brussels, Belgium. In submission.

Cieply, M., and B. Barnes (2009): "In Downturn, Americans Flock to the Movies," NYT - https://www.nytimes.com/2009/03/01/movies/01films.html.

Comscore (2020): "Comscore Reports Highest Ever Worldwide Box Office," .

de Graaf, A., H. Hoeken, J. Sanders, and J. W. J. Beentjes (2012): "Identification as a Mechanism of Narrative Persuasion," *Communication Research*, 39(6), 802–823.

DellaVigna, S., and E. La Ferrara (2015): "Economic and Social Impacts of the Media," *Handbook of Media Economics*, 1, 723–768.

Durante, R., P. Pinotti, and A. Tesei (2019): "The Political Legacy of Entertainment TV," *American Economic Review*, 109(7), 2497–2530.

Enikolopov, R., A. Makarin, and M. Petrova (2020): "Social Media and Protest Participation: Evidence from Russia," *Econometrica*, 88(4), pp. 1479–1514.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011): "Media and Political Persuasion: Evidence from Russia," *The American Economic Review*, 101(7), 3253–3285.

Esposito, E., T. Rotesi, A. Saia, and M. Thoenig (2023): "Reconciliation narratives: The birth of a nation after the us civil war," *American Economic Review*, 113(6), 1461–1504.

Fouka, V., and H.-J. Voth (2023): "Collective Remembrance and Private Choice: German–Greek Conflict and Behavior in Times of Crisis," *American Political Science Review*, 117(3), 851–870.

Griffiths, T. L., and M. Steyvers (2004): "Finding scientific topics," *Proceedings of the National academy of Sciences*, 101(suppl_1), 5228–5235.

Honnibal, M., and I. Montani (2021): "spaCy: Industrial-strength Natural Language Processing in Python," `https://spacy.io/`.

Jensen, R., and E. Oster (2009): "The Power of TV: Cable Television and Women's Status in India," *The Quarterly Journal of Economics*, 124(3), 1057–1094.

Jing, E., S. DeDeo, and Y.-Y. Ahn (2019): "Sameness attracts, novelty disturbs, but outliers flourish in fanfiction online," *arXiv preprint arXiv:1904.07741*.

Kahana, M. (2012): *Foundations of Human Memory*. Oxford University Press.

Kalil, A., S. Mayer, P. Oreopoulos, and R. Shah (2024): "Making a Song and Dance About It: The Effectiveness of Teaching Children Vocabulary with Animated Music Videos," Working Paper 32132, National Bureau of Economic Research.

Kearney, M. S., and P. B. Levine (2015): "Media Influences on Social Outcomes: The Impact of MTV's "16 and Pregnant" on Teen Childbearing," *The American Economic Review*, 105(12), 3597–3632.

——— (2019): "Early Childhood Education by Television: Lessons from Sesame Street," *American Economic Journal: Applied Economics*, 11(1), 318–50.

Kenter, T., and M. De Rijke (2015): "Short text similarity with word embeddings,"

in *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1411–1420.

Kotschy, R., and U. Sunde (2023): "Have Preferences Become More Similar Worldwide?," *CEPR Discussion Paper 18561*.

Krantz-Kent, R. (2018): "Television, capturing America's attention at prime time and beyond," *Beyond the Numbers: Special Studies Research*, 7(14), 3150–3182.

Lauzen, M., D. Dozier, and N. Horan (2008): "Constructing Gender Stereotypes through Social Roles in Prime-Time Television," *Journal of Broadcasting Electronic Media*, 52, 200–214.

Marshall, M. G., and T. R. Gurr (2020): "Polity5: Political Regime Characteristics and Transitions, 1800-2018," Center for Systemic Peace, Available at Center for Systemic Peace website.

Martinez Palenzuela, Y. (2014): "GeoText: Python package for geographical entities extraction," `https://pypi.org/project/geotext/`.

Mehal, A. S., K. Meena, R. B. Singh, and P. G. Shambharkar (2021): "Movie genres and beyond: An analytical survey of classification techniques," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1193–1198.

Michalopoulos, S., and E. Papaioannou (2018): "Spatial Patterns of Development: A Meso Approach," *Annual Review of Economics*, 10(1), 383–410.

——— (2020): "Historical Legacies and African Development," *Journal of Economic Literature*, 58(1), 53–128.

Michalopoulos, S., and M. M. Xue (2021): "Folklore," *The Quarterly Journal of Economics*, 136(4), 1993–2046.

Nunn, N. (2020): "The historical roots of economic development," *Science (New York, N.Y.)*, 367(6485), eaaz9986.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011): "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Putterman, L., and D. N. Weil (2010): "Post-1500 Population Flows and the Long Run Determinants of Economic Growth and Inequality," *The Quarterly Journal of Economics*, 125(4), 1627–1682.

Riley, E. (2022): "Role Models in Movies: The Impact of Queen of Katwe on Students' Educational Attainment," *The Review of Economics and Statistics*, pp. 1–48.

Stacks, D. W., Z. Cathy Li, and C. Spaulding (2015): "Media Effects," in *International Encyclopedia of the Social Behavioral Sciences (Second Edition)*, ed. by J. D. Wright, pp. 29–34. Elsevier, Oxford, second edition edn.

Towbin, M. A., S. A. Haddock, T. S. Zimmerman, L. K. Lund, and L. R. Tanner (2004): "Images of gender, race, age, and sexual orientation in disney feature-length animated films," *Journal of Feminist Family Therapy*, 15(4), 19–44.

Törnberg, P. (2023): "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning," .

Van der Maaten, L., and G. Hinton (2008): "Visualizing data using t-SNE.," *Journal of machine learning research*, 9(11).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017): "Attention is all you need," *Advances in Neural Information*

Processing Systems, 30.

Voth, H.-J. (2021): "Chapter 9 - Persistence – myth and mystery," in *The Handbook of Historical Economics*, ed. by A. Bisin, and G. Federico, pp. 243–267. Academic Press.

Voth, H.-J., and D. Yanazigawa-Drott (2023): "Image(s)," *CEPR Discussion Paper 16624*.

Wang, Q. (2021): "The Cultural Foundation of Human Memory," *Annual Review of Psychology*, 72(1), 151–179, PMID: 32928062.

Zeileis, A. (2023): *PWT10: Penn World Table (Version 10.01)*.

Appendix

A Descriptives and additional evidence

Table A1: Movie summary statistics

	N	Mean	Median	Min	Max	Std
Panel A: Revenue sample						
Countries of screenings	85					
Countries financing movies	144					
Movies screened	13929					
Movie screenings	168626					
Countries screened per movie		12.11	4	2	74	16.02
Revenue (in thousand US$)		35583	2867	1	2608844	114070
Revenue (in thousand US$) per country screened		2939.29	215.35	1	936662.2	16280.28
Year screened		2011.35	2011	1995	2019	4.9
Year of first screening		2011.19	2012	1995	2019	5.21
Panel B: Screening sample						
Countries of screenings	92					
Countries financing movies	153					
Movies screened	28810					
Movie screenings	184908					
Countries screened per movie		6.42	1	1	81	12.59
Revenue (in thousand US$)		19106	701	1	2736372	82419
Revenue (in thousand US$) per country screened		2976.83	211.26	1	936662.2	16193.91
Year screened		2011.17	2011	1995	2019	5.12
Year of first screening		2010.4	2011	1995	2019	6.07

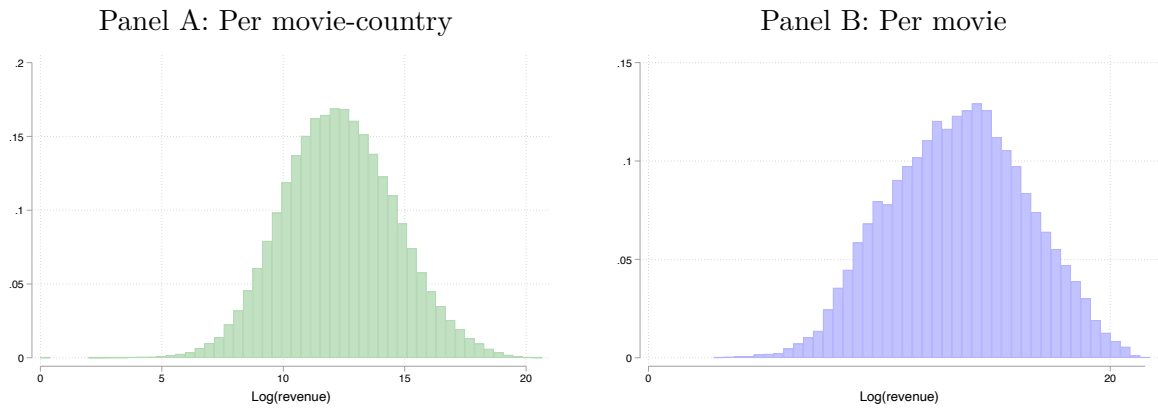
Notes: Panel A refers to the baseline sample for the revenue regressions and Panel B to the baseline sample for the probability of screening regressions. The sample in Panel A is smaller than that of Panel B because it focuses on movies screened in at least two countries and in country-year-month periods with multiple screenings. Only movies with positive revenue are included.

Table A2: Actors, writers, and directors in revenue sample

		Movie frequency				
	N	Mean	Median	Min	Max	Std
Number of actors	18331	2.26	1	1	55	3.47
Actors in at least 5 movies	1824	10.42	8	5	55	6.5
Number of writers	18021	1.79	1	1	41	1.76
Writers in at least 5 movies	1116	7.22	6	5	41	3.03
Number of directors	8106	1.93	1	1	21	1.78
Directors in at least 5 movies	629	6.84	6	5	21	2.48
Number of distributors	5256	22.35	2	1	5240	151.17
Distributors in at least 5 movies	1796	62.21	17	5	5240	253.95

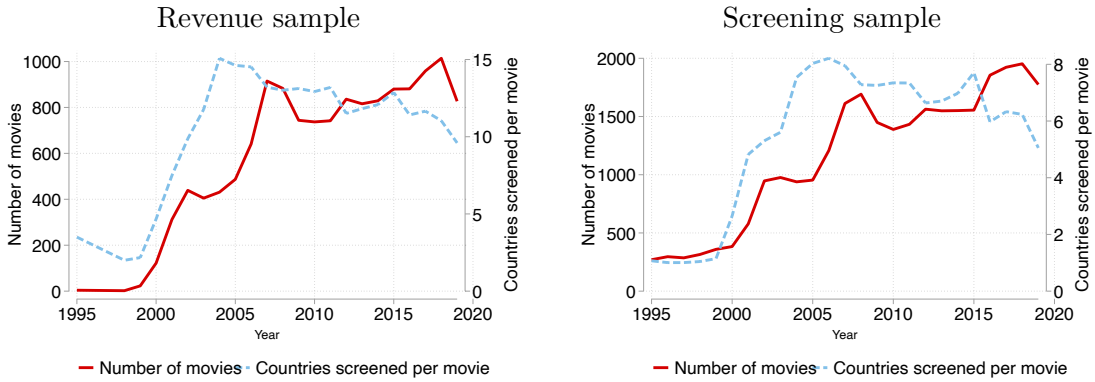
Notes: The column 'N' refers to the number of actors, writers, directors, and distributors in the 13,929 movies of the revenue sample. The remaining columns refer to the number of movies each is associated with.

Figure A1: Distribution of log(revenue)



Notes: Revenue is measured in US$.

Figure A2: Number of movies and number countries screened per movie over time



Notes: The left-hand y-axis shows the total number of movies, and the right-hand y-axis shows the mean number of countries in which these movies are screened on average.

Table A3: Number of movies by genre and language in movies

Panel A: Genres

	Samples	
	Revenue	Screening
filmnoir	1	19
western	43	99
musical	157	330
war	269	538
sport	273	563
scifi	489	842
history	570	1122
music	637	1224
documentary	651	1846
family	728	1439
fantasy	784	1335
animation	830	1221
biography	940	1627
mystery	1008	1781
horror	1121	2056
thriller	1807	3461
adventure	1822	2748
crime	2032	3803
romance	2500	4989
action	2686	4353
comedy	5155	10606
drama	7966	16502

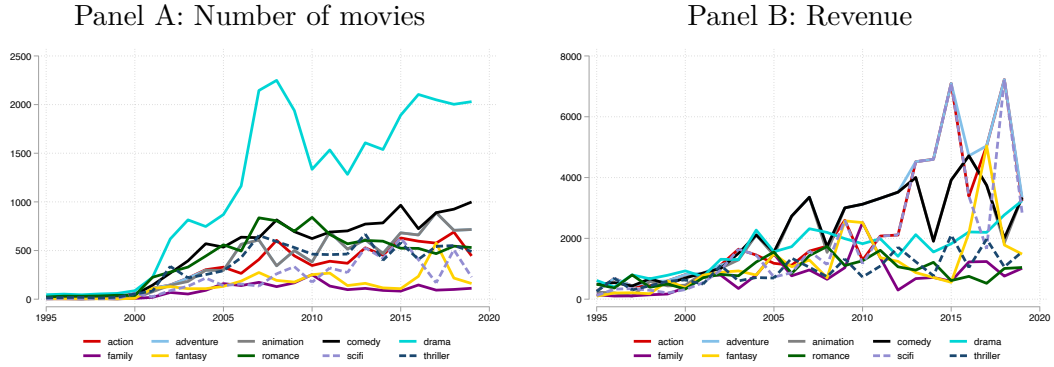
Notes: A movie can have up to three genres.

Panel B: Languages

	Samples	
	Revenue	Screening
Danish	135	332
Polish	163	406
Latin	166	241
Swedish	171	443
Dutch	176	506
Portuguese	210	688
Tamil	223	410
Turkish	301	849
Cantonese	399	631
Arabic	430	703
Korean	440	989
Italian	636	1507
Russian	732	1291
Mandarin	746	1076
Hindi	779	1233
Japanese	888	1655
German	1181	1930
Spanish	1539	3015
French	2376	3935
English	7974	14768

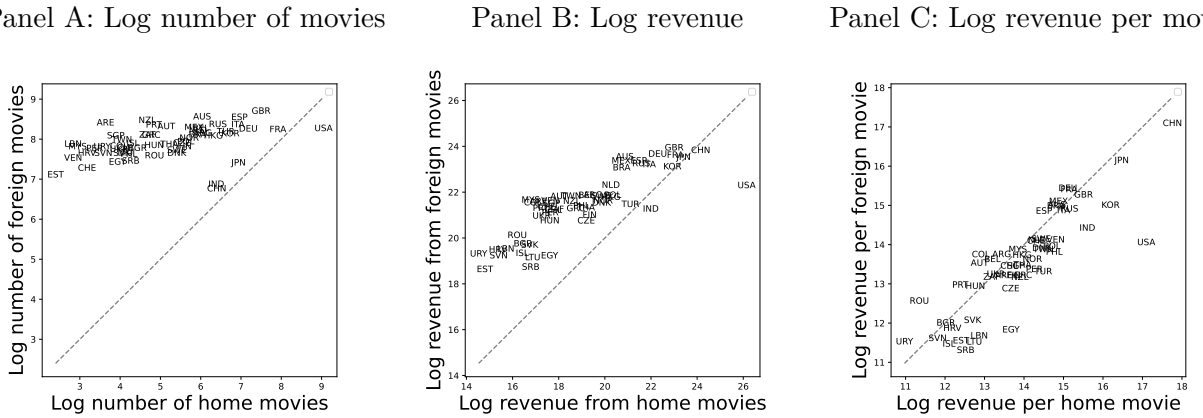
Notes: A movie can have multiple languages. The table features all languages that appear in at least 1% of movies.

Figure A3: Statistics by genre over time



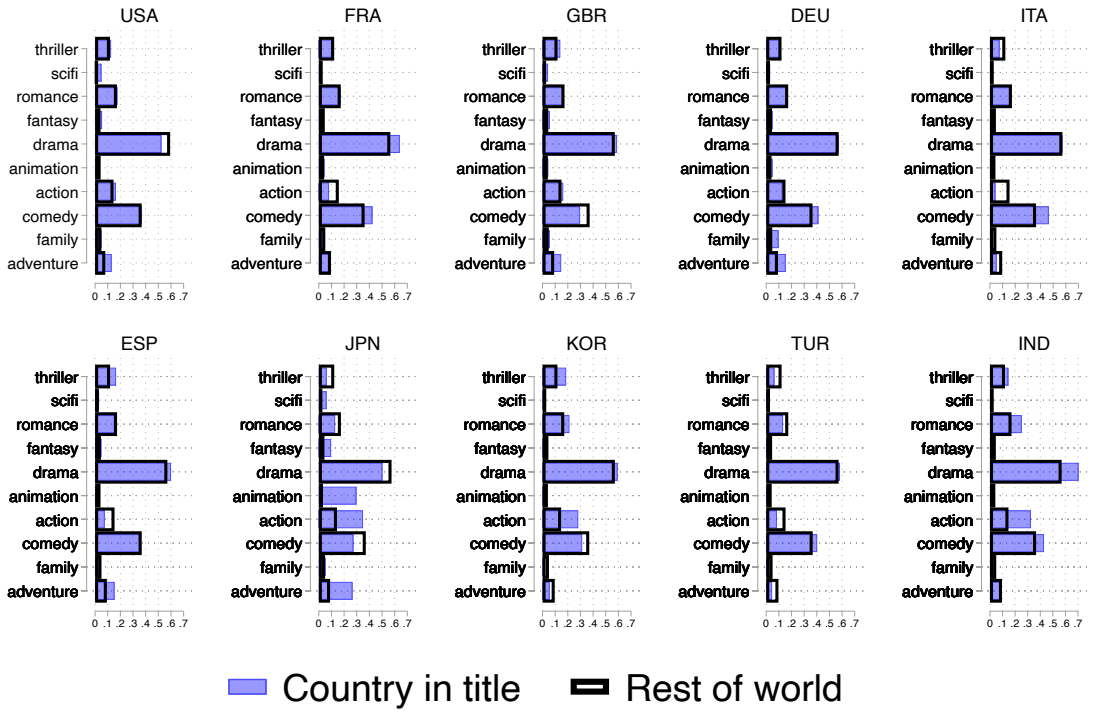
Notes: Revenue in million US$.

Figure A4: Foreign and home produced movies by country



Notes: Each point is a country with at least 10 observations of (i) domestically financed movies and (ii) foreign-origin ones. In Panel B we take the log of the total revenue across movies screened in a country by source of financing. In Panel C we take the log of the ratio of the total revenue divided by the total number of movies. A movie is considered a home movie if it receives domestic financing and foreign otherwise. A movie receiving both appears in both subsamples.

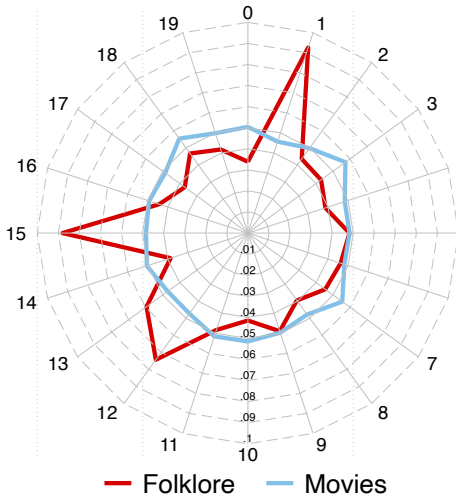
Figure A5: Movie genres by origin of financing



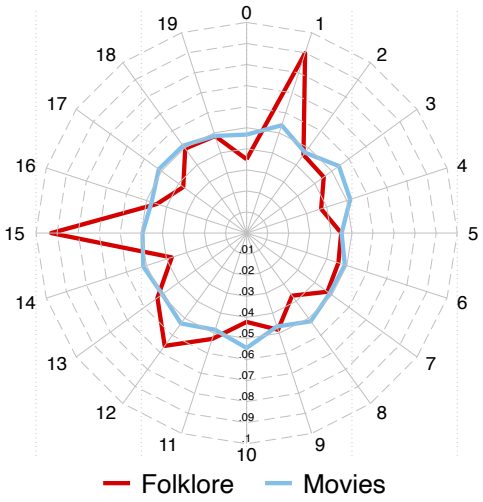
Notes: Each movie can have up to three genres. The purple bars reflect the share of movies of a given genre amongst the movies receiving financing of the country in the subtitle, while the bars with the black outline highlight the share of the representation of each genre not receiving financing of the country in the subtitle.

Figure A6: Topic distributions of movies vs folklore

Panel A: Unweighted average



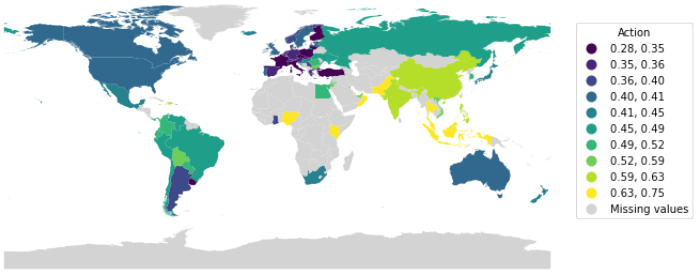
Panel B: Weighted average



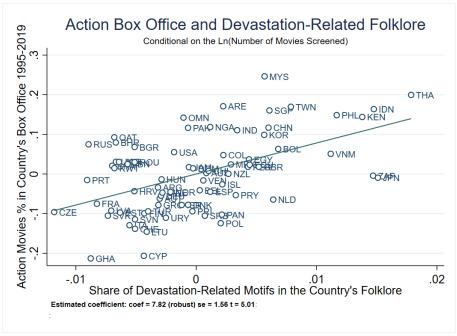
Notes: The radar plot indicates the average topic shares for motifs (red) and movies (blue). In Panel B the motifs are weighted by the GDP of the countries in which they appear and the movie topics by total movie revenue.

Figure A7: Action genre

Panel A



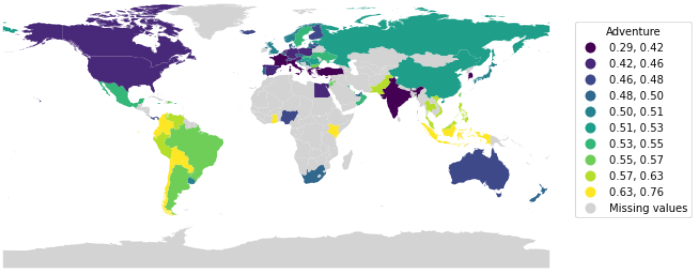
Panel B



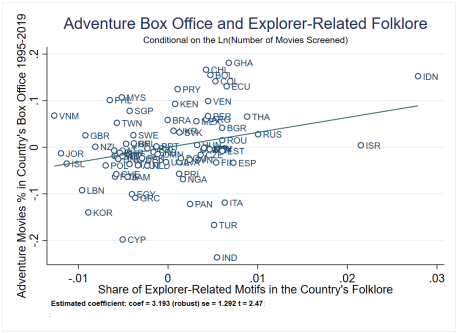
Notes: Panel A depicts for each country the share of the total box office revenue generated by action-genre movies for the period 1995-2019. Panel B plots the relationship between the former and the share of motifs that mention devastation-related terms (according to ConceptNet) in the country's folklore.

Figure A8: Adventure genre

Panel A



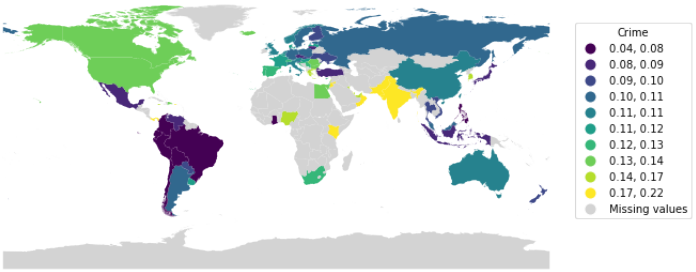
Panel B



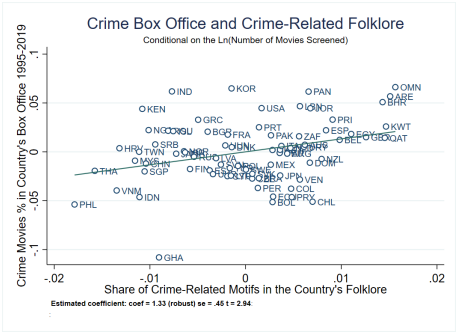
Notes: Panel A depicts for each country the share of the total box office revenue generated by adventure-genre movies for the period 1995-2019. Panel B plots the relationship between the former and the share of motifs that mention explorer-related terms (according to ConceptNet) in the country's folklore.

Figure A9: Crime genre

Panel A



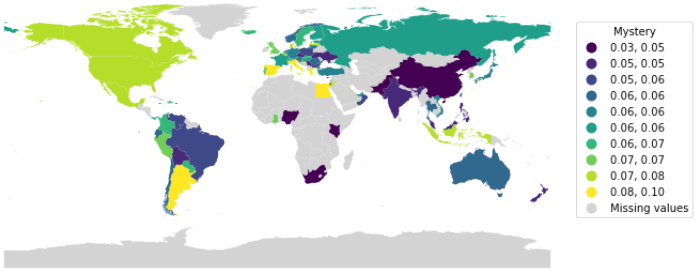
Panel B



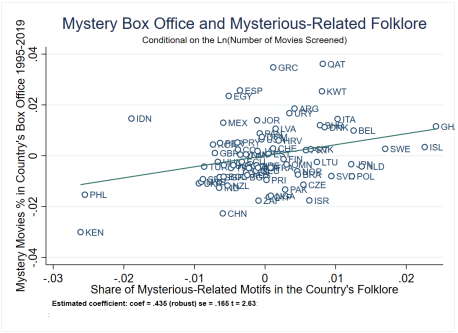
Notes: Panel A depicts for each country the share of the total box office revenue generated by crime-genre movies for the period 1995-2019. Panel B plots the relationship between the former and the share of motifs that mention crime-related terms (according to ConceptNet) in the country's folklore.

Figure A10: Mystery genre

Panel A

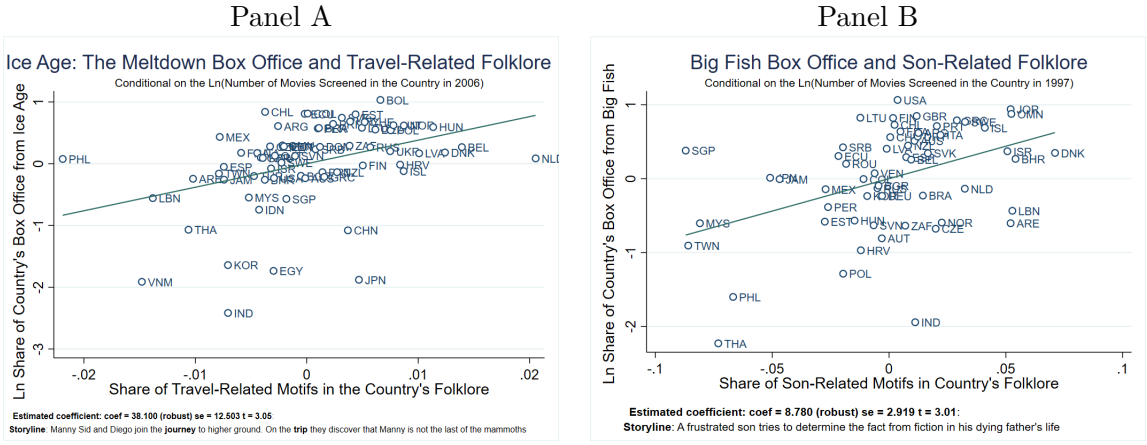


Panel B



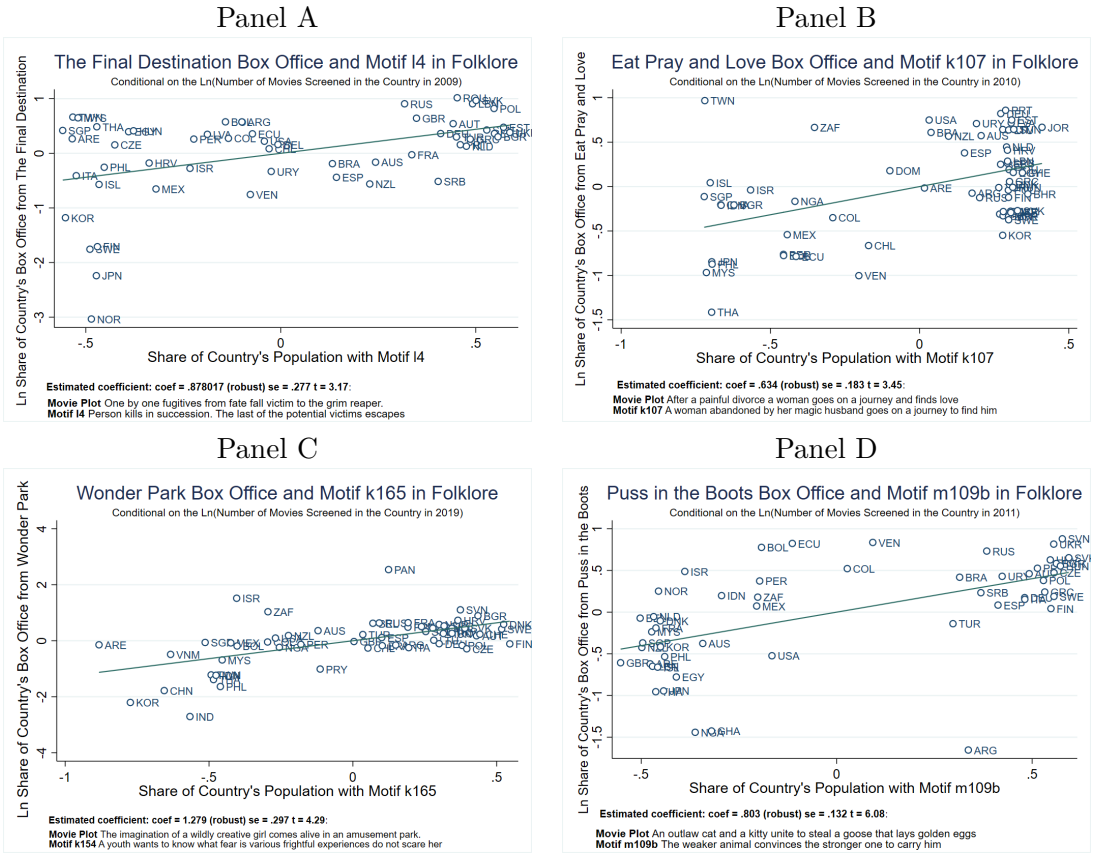
Notes: Panel A depicts for each country the share of the total box office revenue generated by mystery-genre movies for the period 1995-2019. Panel B plots the relationship between the former and the share of motifs that mention mysterious-related terms (according to ConceptNet) in the country's folklore.

Figure A11: A movie's performance and broad themes in a country's folklore



Notes: The y-axes show the log of the percentage of the total box office in a given country achieved by the movie in the subtitle. The x-axis in Panel A displays the share of travel-related motifs in a country's folklore. The x-axis in Panel B displays the share of son-related motifs in a country's folklore.

Figure A12: A movie's performance and specific motifs in a country's folklore



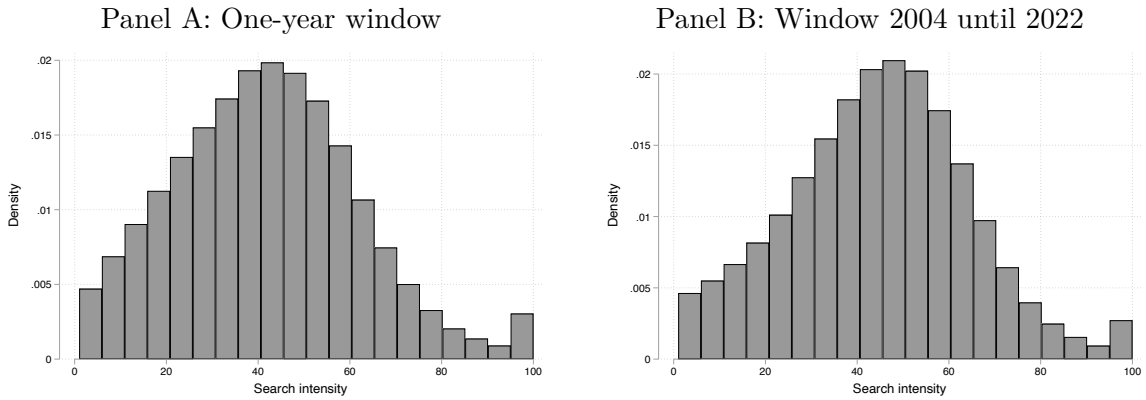
Notes: The y-axes shows the log of the percentage of the total box office in a given country achieved by the movie in the subtitle. The x-axes display the log of the share of a country's population sharing motif l4 (Panel A), k107 (Panel B), k165 (Panel C), and m109b (Panel D).

Table A5: Theaters and weeks screened summary statistics

	Mean	Median	Min	Max	Std
Theaters	816	104	1	731831	3879.12
Revenue/theaters in 1,000 USD	2.59	1.5	0	22488.71	66.15
Weeks	6.94	5	1	463	6.61

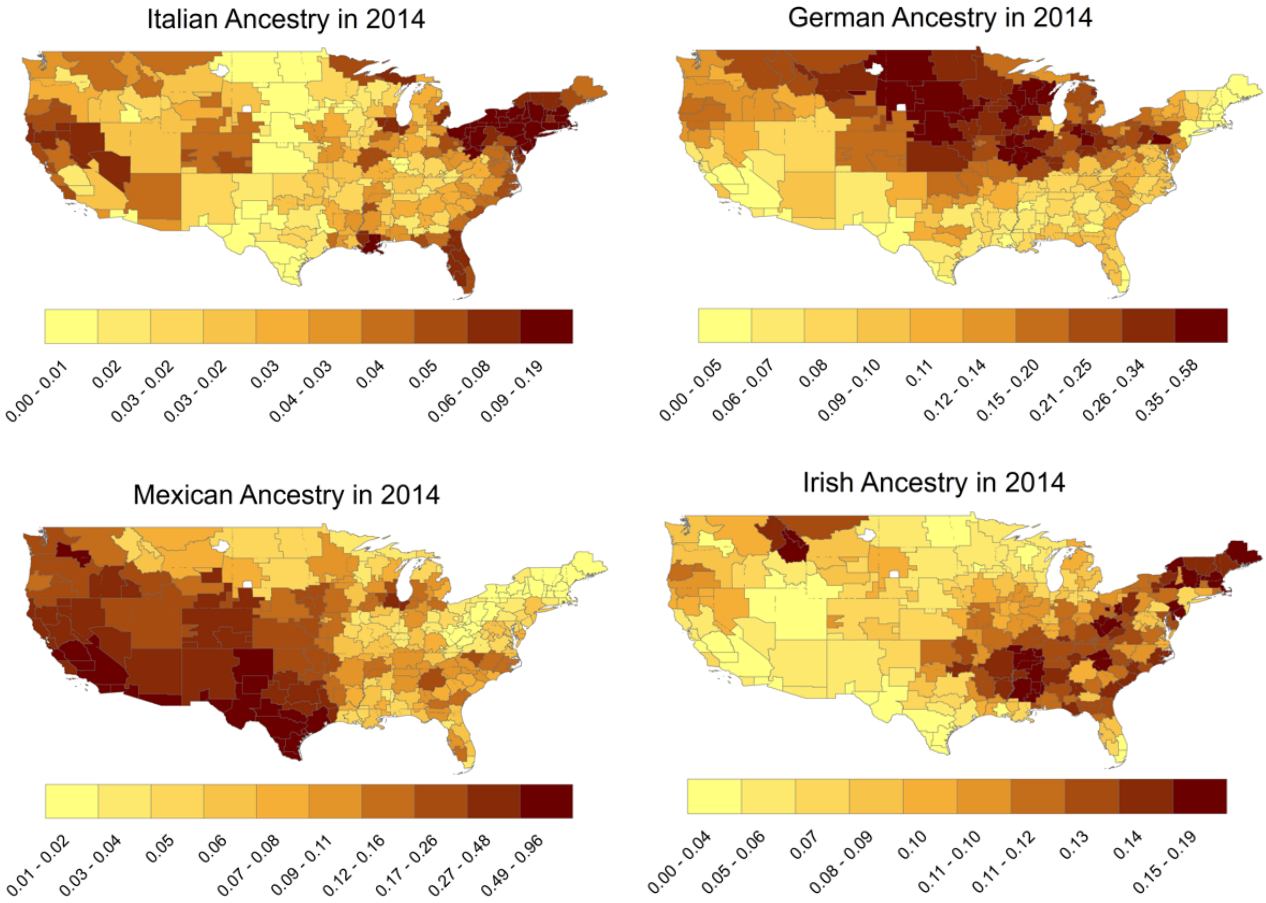
Notes: These summary stats refer to the samples used in Table 3. Zeros due to rounding.

Figure A13: Google Trends search intensity distribution



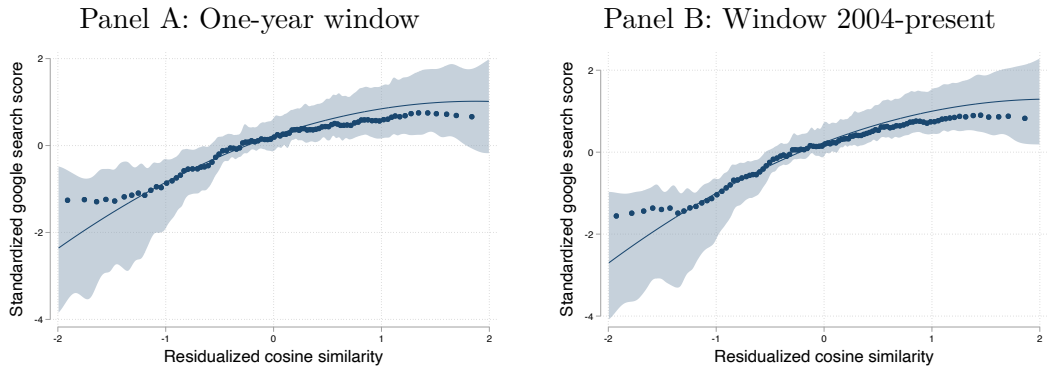
Notes: Panel A restricts the search score to a two-year window, one year before and after release of the movie, and Panel B uses the entire time period from 2004 until 2022.

Figure A14: Ancestry across DMAs



Notes: The panels show the distribution of Italian, German, Mexican, and Irish ancestry across DMAs. Ancestry is computed using responses from the ACS aggregated across DMAs.

Figure A15: Binned scatter plot of a film's proximity to local narratives and Google search score



Notes: The left Panel A restricts the search score to one year before and one year after the theatrical release of the movie, and the right Panel B uses the entire period; namely from 2004 until 2022. Controls include DMA-year-month and movie-specific constants and standard errors are double clustered at the year and DMA level and closely relates to the specification in column (1) in Table 4. The proximity of a movie plot to local narratives is based on the cosine similarity of embeddings between ancestral motifs and a film's plot residualized using the log(number of words) of the movie summary, motif description, and their interaction.

Table A6: Risk-taking in movies

	Movies	Share
Panel A: Revenue sample		
Setbacks exceed wins	4140	.39
Setbacks and wins balance each other	5337	.5
Wins exceed setbacks	1242	.12
Total	10719	
Panel B: Screening sample		
Setbacks exceed wins	8128	.41
Setbacks and wins balance each other	9747	.5
Wins exceed setbacks	1793	.09
Total	19668	

Risk classification descriptives

	Mean	Min	Max	St.d.
Panel A: Revenue sample:	.73	0	2	.66
Panel B: Screening sample:	.68	0	2	.63

Notes: Panels A and B report the risk classification for movies in the revenue and screening samples, respectively. Setbacks exceed wins' is classified as 0, 'setbacks and wins balance each other' as 1, and 'wins exceed setbacks' as 2.

Table A7: Gender roles in movies

	Female		Male	
	Movies	Share	Movies	Share
Panel A: Revenue sample				
Sexual	4054	.39	3266	.31
Engaged in domestic affairs	3117	.3	1563	.15
Submissive	2724	.26	411	.04
Naive	2602	.25	4946	.47
Intelligent	7689	.73	7559	.72
Physically active	2758	.26	6453	.61
Dominant	5257	.5	7938	.75
Total	10523		10523	
Panel B: Screening sample				
Sexual	7400	.41	6108	.34
Engaged in domestic affairs	5531	.31	2760	.15
Submissive	4770	.27	681	.04
Naive	4889	.27	8892	.5
Intelligent	12521	.7	12277	.69
Physically active	3997	.22	10303	.58
Dominant	8322	.46	13052	.73
Total	17901		17901	

Movie bias descriptives

	Mean	Min	Max	St.d.
Revenue sample: Mean movie bias (female)	.3	-4	3	1.92
Revenue sample: Mean movie bias (male)	1.74	-3	4	1.49
Revenue sample: Mean relative movie bias	-1.43	-8	6	2.29
Screening sample: Mean movie bias (female)	.13	-4	3	1.92
Screening sample: Mean movie bias (male)	1.64	-3	4	1.52
Screening sample: Mean relative movie bias	-1.52	-8	6	2.31

Notes: Panels A and B report statistics in the revenue and screening samples, respectively. To arrive at a measure of male and female bias, we add the dominant, physically active, and intelligent categories and subtract the naive, submissive, and engaged in domestic affairs classifications. We treat the sexual classification as a negative stereotype for female characters and as a neutral/positive one for males. This is because, unlike males, when females are classified as sexual they often appear in stereotypically sexualized roles. The relative bias is obtained by subtracting the male from the female bias. Hence, higher values reflect a more favorable representation of female over male characters.

Table A8: Correlations of gender roles in movies

	Women								Men								
	Bias	Domi.	Act.	Int.	Naive	Subm.	Dome.	Sex.	Bias	Domi.	Act.	Int.	Naive	Subm.	Dome.	Sex.	Rel.
Panel A: Revenue sample																	
Female																	
Mean bias	1.00																
Dominant	0.78	1.00															
Active	0.56	0.24	1.00														
Intelligent	0.75	0.52	0.17	1.00													
Naive	-0.55	-0.33	-0.18	-0.53	1.00												
Submissive	-0.70	-0.58	-0.25	-0.53	0.21	1.00											
Domestic	-0.37	-0.32	-0.31	-0.10	-0.12	0.16	1.00										
Sexual	-0.48	-0.24	-0.23	-0.32	0.20	0.21	-0.23	1.00									
Male																	
Mean bias	0.12	0.07	0.13	0.08	-0.20	0.09	-0.19	0.08	1.00								
Dominant	0.04	0.05	-0.00	0.05	-0.14	0.14	-0.08	0.01	0.74	1.00							
Active	0.05	-0.10	0.36	-0.12	-0.03	0.15	-0.09	-0.10	0.52	0.32	1.00						
Intelligent	0.30	0.19	0.04	0.36	-0.36	-0.07	-0.01	-0.22	0.60	0.39	0.10	1.00					
Naive	-0.20	-0.11	-0.08	-0.18	0.37	0.02	-0.05	0.14	-0.62	-0.34	-0.16	-0.51	1.00				
Submissive	-0.01	0.03	-0.04	-0.00	-0.00	0.01	0.03	0.01	-0.42	-0.32	-0.21	-0.20	0.13	1.00			
Domestic	-0.00	-0.01	-0.15	0.11	-0.13	-0.12	0.42	-0.19	-0.52	-0.36	-0.34	-0.08	-0.03	0.09	1.00		
Sexual	-0.23	-0.02	-0.20	-0.12	0.17	-0.01	-0.15	0.58	0.16	-0.03	-0.28	-0.19	0.08	-0.02	-0.13	1.00	
Relative bias	0.76	0.61	0.38	0.58	-0.33	-0.65	-0.18	-0.46	-0.56	-0.45	-0.30	-0.14	0.24	0.26	0.34	-0.29	1.00
Panel B: Screening sample																	
Female																	
Mean bias	1.00																
Dominant	0.79	1.00															
Active	0.55	0.25	1.00														
Intelligent	0.77	0.53	0.17	1.00													
Naive	-0.57	-0.35	-0.18	-0.56	1.00												
Submissive	-0.68	-0.55	-0.23	-0.52	0.21	1.00											
Domestic	-0.33	-0.31	-0.29	-0.06	-0.14	0.14	1.00										
Sexual	-0.48	-0.24	-0.23	-0.33	0.22	0.20	-0.27	1.00									
Male																	
Mean bias	0.11	0.07	0.13	0.07	-0.19	0.11	-0.19	0.09	1.00								
Dominant	0.03	0.05	0.01	0.04	-0.13	0.16	-0.07	0.02	0.75	1.00							
Active	0.02	-0.10	0.34	-0.13	-0.03	0.17	-0.06	-0.10	0.52	0.35	1.00						
Intelligent	0.30	0.19	0.04	0.36	-0.36	-0.06	0.01	-0.22	0.60	0.38	0.10	1.00					
Naive	-0.22	-0.12	-0.09	-0.21	0.38	0.02	-0.06	0.16	-0.61	-0.33	-0.16	-0.52	1.00				
Submissive	-0.01	0.03	-0.03	-0.00	0.01	0.01	0.03	0.01	-0.39	-0.30	-0.19	-0.18	0.11	1.00			
Domestic	0.03	-0.00	-0.13	0.14	-0.15	-0.14	0.42	-0.22	-0.53	-0.38	-0.33	-0.08	-0.04	0.09	1.00		
Sexual	-0.22	-0.02	-0.19	-0.13	0.18	-0.01	-0.20	0.59	0.16	-0.03	-0.29	-0.20	0.09	-0.02	-0.16	1.00	
Relative bias	0.76	0.61	0.36	0.59	-0.34	-0.64	-0.15	-0.45	-0.57	-0.47	-0.32	-0.14	0.22	0.25	0.37	-0.29	1.00

Notes: The Table shows the bilateral correlations of male and female categories across movies in the revenue (Panel A) and screening sample (Panel B).

Table A9: Gender roles versus revenue and likelihood of screening

	Log(revenue)				Screening			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female out of LF x Female bias movie	-0.0097***				-0.0018*			
	(0.0032)				(0.0010)			
Male bias folklore x Female bias movie		-0.0088***				-0.0021**		
		(0.0023)				(0.0009)		
Female out of LF x Male bias movie			0.0080***				-0.0014	
			(0.0022)				(0.0010)	
Male bias folklore x Male bias movie				-0.0014				-0.0021*
				(0.0049)				(0.0012)
Observations	148279	150686	148279	150686	1611090	1628991	1611090	1628991
R^2	0.7945	0.7939	0.7943	0.7937	0.4204	0.4205	0.4203	0.4204
Number of movies	10281	10375	10281	10375	17901	17901	17901	17901
Number of countries	83	84	83	84	90	91	90	91
Movie FE	✓	✓	✓	✓	✓	✓	✓	✓
Country-year-month FE	✓	✓	✓	✓				
Country-year of first release FE					✓	✓	✓	✓

Notes: OLS regressions. The female bias in the movie is obtained by adding the dominant, physically active, and intelligent categories and subtracting the naive, submissive, sexual, and engaged in domestic affairs classifications for females. This variable ranges from -4 to 3 with higher values reflecting a more favorable representation of female characters. The male bias in the movie is obtained by adding the dominant, physically active, sexual, and intelligent categories and subtracting the naive, submissive, and engaged in domestic affairs classifications for males. This variable ranges from -3 to 4 with higher values reflecting a more favorable representation of male characters. The dependent variables in columns (1)-(4) is log(revenue) standardized, while columns (5)-(8) it is whether a movie is screened in a given country. Female out of LF is the country's share of women that are neither unemployed nor employed, i.e., classified as out of the labor force. The male bias folklore variable is defined as the share of motifs where men are depicted as more dominant, less submissive, less engaged in domestic affairs, and more physically active than women in the country's oral tradition. Both the female out of LF and male bias folklore variables are standardized with a mean of 0 and a standard deviation of 1. Standard errors are double clustered at the country and year level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

B Details on text analysis and distance/similarity measures

B.1 Universal Sentence Encoder

The Universal Sentence Encoder (USE) by Cer et al. (2018) is a machine learning model developed by Google AI, designed to convert textual data into high-dimensional vector embeddings. These embeddings capture semantic meanings of sentences, enabling various natural language processing tasks. Even though it is called 'sentence encoder', the model is trained and optimized for greater-than-word length text, such as sentences, phrases, and short paragraphs. We rely on the 4th vintage available on Tensorflow which is based on Abadi et al. (2015)).

USE employs a deep averaging network (DAN) where input word embeddings are averaged together and passed through a feedforward deep neural network (DNN) to produce sentence embeddings. Alternatively, it uses a transformer-based architecture, which leverages attention mechanisms to capture contextual information from all words in a sentence.

The encoder is trained using a variety of data sources and tasks, including unsupervised data and supervised tasks like textual similarity and entailment. This diverse training allows the model to generalize well across different domains and tasks.

The distribution of the average loading on each of the 512 embeddings by genre is presented in OA Figure C3. While some entries load neither positively nor negatively on any genre, e.g. embedding 35, others load negatively on all (embedding 16), positively on all (embedding 99), or both negatively and positively (embedding 465). Drama and romance tend exhibit similar patterns, as do sci-fi, fantasy, and animation.

B.2 LDA description

Latent Dirichlet Allocation (LDA) is a generative statistical model for discovering the abstract topics that occur in a collection of documents. The LDA model is defined as follows:

1. For each document d, choose a distribution over topics $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. For each word w in document d:
   (a) Choose a topic $z_{d,w} \sim \text{Multinomial}(\theta_d)$.
   (b) Choose a word $w \sim \text{Multinomial}(\beta_{z_{d,w}})$,

   where β represents the distribution of words for topic z.

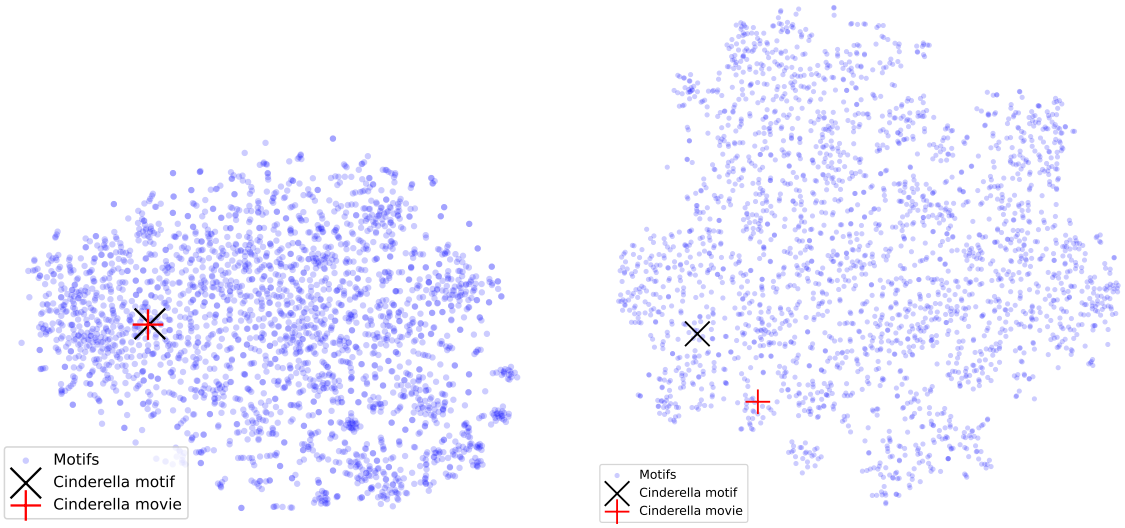
The Dirichlet distribution, parameterized by α, influences the mixture of topics in each document. The goal of LDA is to infer the set of topics θ_d and their distribution β that are likely to have generated the observed collection of documents.

For the implementation we use the Python package by Pedergosa et al (2011). The priors we set to $\alpha = 0.9$ and $\beta = 0.1$, which are standard values (Griffiths and Steyvers 2004). Before performing the LDA, we clean the text data. This involves tokenizing the text into individual words, converting all characters to lowercase for consistency, and removing common stop words that offer little meaningful information. Additionally, punctuation and special characters as well as locations identified using the geotext Python package (Martinez Palenzuela 2014) are eliminated. Then words are reduced to their base or root form through the lemmatizer of Honnibal and Montani (2021). Lemmatization converts words to their meaningful base form considering the context. We then form and add two-word (bigrams) and three-word (trigrams) combinations to the bag of words. Lastly, very rare terms appearing in less than five documents or very frequent terms appearing in more than 80% of all documents are removed to enhance the performance of the model.

Table B1: Top 15 keywords of each of 20 topics

Topic 0: friend old meet good want good_friend grow beautiful party boyfriend lonely date fun unexpectedly crush
Topic 1: mysterious event human catch road far path cross appear earth gets inside forest creature near
Topic 2: find house night happen wedding strange stay miss stand trouble disappear prove stranger prepare pass
Topic 3: life time relationship follow change village age win heart challenge enter chance race test risk
Topic 4: lead city head body future mission killer prison encounter king special question drive track driver
Topic 5: turn child couple husband parent leave marriage single revenge lover situation abandon affair married baby
Topic 6: force town hope girlfriend join teacher arrive kid steal evil bear female spirit partner mean
Topic 7: young love girl fall brother sister marry money rich young_girl separate reunite poor condition twin
Topic 8: world group war fight deal survive soldier community hunt wake happy conflict witness light blind
Topic 9: woman learn different break young_woman experience summer unexpected visit feel choose book society write youth
Topic 10: work look face drug stop threaten create control free attack fear action rescue fate deadly
Topic 11: man dream journey true young_man realize childhood need come think hard worker share self sex
Topic 12: people run escape away adventure power battle island dangerous mountain club military sea survival throw
Topic 13: start save star company bad cause owner wrong hit bond rise fail apart water accidentally
Topic 14: death secret murder plan attempt crime dead criminal involve accident investigate protect reality commit suspect
Topic 15: know tell kill send believe member lie receive ask hero army doctor come master enemy
Topic 16: try lose bring play job real music game successful suffer accept offer train player able
Topic 17: discover boy begin struggle past travel seek truth dark memory suddenly confront powerful haunt come
Topic 18: family father live mother son home daughter return die care leave law dog raise orphan
Topic 19: decide help wife order search spend problem teenager hide business personal right reveal social kidnap

Figure B1: t-SNE plot of Cinderella example



Notes: The figures show two-dimensional representations of the 512-dimensional summaries provided by the USE (left) and 20 topics provided by the LDA (right) using T-distributed Stochastic Neighbor Embeddings. Each blue dot represents one motif. The black X's represent the Cinderella motif and the red cross the Cinderella movie.

B.3 Comparing USE and LDA

To illustrate the workings of the Universal Sentence Encoder (USE) and its contrast with Latent Dirichlet Allocation (LDA), let's consider two movies: one a comedy and the other a drama. We'll explore how USE might encode descriptions or reviews of these movies and how LDA would categorize their content.

Imagine we're reducing USE's complexity to a hypothetical two-dimensional model for simplicity. In this model, one dimension may represent sentiment, with negative values for more serious or sad tones and positive values for humorous or light-hearted content. The second dimension may represent thematic depth, with negative values for surface-level, lighthearted themes (typical of many comedies) and positive values for in-depth, serious themes (often found in dramas).[1]

For a comedy movie, descriptions or reviews might highlight its humor, quirky characters, and entertaining plot twists. USE could encode this movie with a positive sentiment score, reflecting the overall light-heartedness, and a negative thematic depth score, indicating the less serious content. This positioning effectively captures the movie's essence in the

[1]This two dimensional representation would allow for only a limited characterization of stories compared to the 512 dimensions provided by the USE. Another convenient feature of embeddings is that mathematical operations, such as vector addition and subtraction, can reveal and manipulate semantic relationships between different pieces of text. For instance, consider embeddings for two different words, "king" and "woman." In a simplified example, we can add the embedding vectors of these two words together to produce a new vector that may be closer in the embedding space to the vector for "queen."

hypothetical two-dimensional semantic space.

Conversely, a drama movie, known for its complex characters, emotional depth, and serious themes, might be encoded with a negative sentiment score due to its serious tone. The thematic depth score would be positive, reflecting the movie's deeper, more nuanced exploration of its subjects. This encoding allows USE to represent the movie's semantic nuances in the same two-dimensional space, but positioned quite differently from the comedy.

Turning to LDA, this model would approach the content differently by identifying prevalent topics within the scripts or reviews of these movies without considering the order of words or the nuanced semantics captured by USE.

For the comedy, LDA might identify topics such as "humor and satire," "romantic misadventures," and "comedic conflicts," characterized by frequent words like "laughter," "mistake," "funny," "date," and "unexpected." The model would assign a high proportion of "humor and satire" to this movie, reflecting its comedic nature.

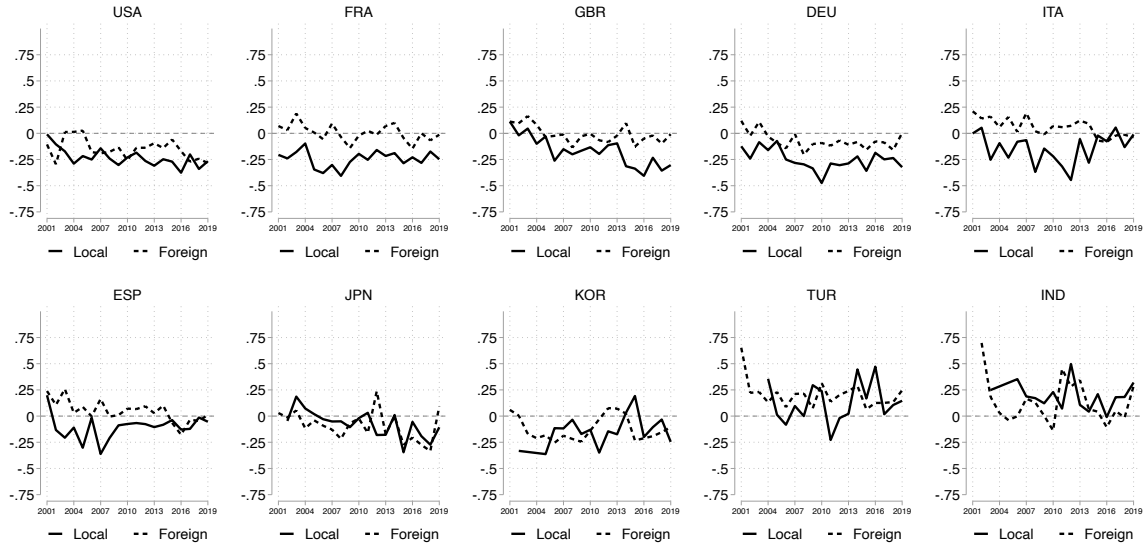
For the drama, LDA might find topics like "personal struggles," "social commentary," and "emotional journeys," with words like "challenge," "society," "tearful," "growth," and "conflict" being prominent. The drama would have a significant proportion of "personal struggles" and "emotional journeys," highlighting its focus on in-depth storytelling and character development.

B.4 Similarities/distances across countries and time

Figure B2 shows the average distance between folklore and movie description embeddings over time for each of the main movie finance countries. Each panel focuses on movies screened in the country indicated in the heading. The solid line indicates the distance between movies that received domestic financing and local folklore, while the dashed line represents the same statistic for movies that did not receive local financing. Some interesting patterns emerge. In India, domestically financed films (mainly Bollywood ones) are closer to Indian folklore than foreign-financed films screened in India. In Japan, Turkey, and Korea, this distinction is less sharp but still the domestically financed films seem to be closer to the local folklore. The opposite pattern is observed for movies financed in the UK, Italy, and Spain with the most pronounced difference observed for French-financed films. It seems that the latter are further away from French folklore than foreign movies screened in France. The corresponding gap between US and non US-financed films screened in the US is less pronounced.

Looking back at the cinematic revenue patterns in Figure A4, French films on average accrue less box office revenue per film compared to Indian/Korean/Japanese films in India/South Korea/Japan. This pattern may be attributed to the cultural proximity of the domestically-financed films. French movies tend to be further apart from French folklore compared to foreign films screened in France, whereas in India, South Korea, and Japan, domestic movies more closely align with local folklore narratives, exceeding the

Figure B2: Distance between folklore and movies over time and across countries depending on source of financing



Notes: Country-year-origin of financing cells with less than 10 movies are excluded. Only movies that are in the revenue dataset are included. The residualized and standardized JSD capturing distance between folklore and movies is on the y-axis. The solid line indicates the distance between consumed movies that received financing from the country of screening indicated the heading, while the dashed line represents the distance for those that did not receive financing from the country.

cultural resonance of foreign films screened locally.

Online Appendix

C Other figures and tables

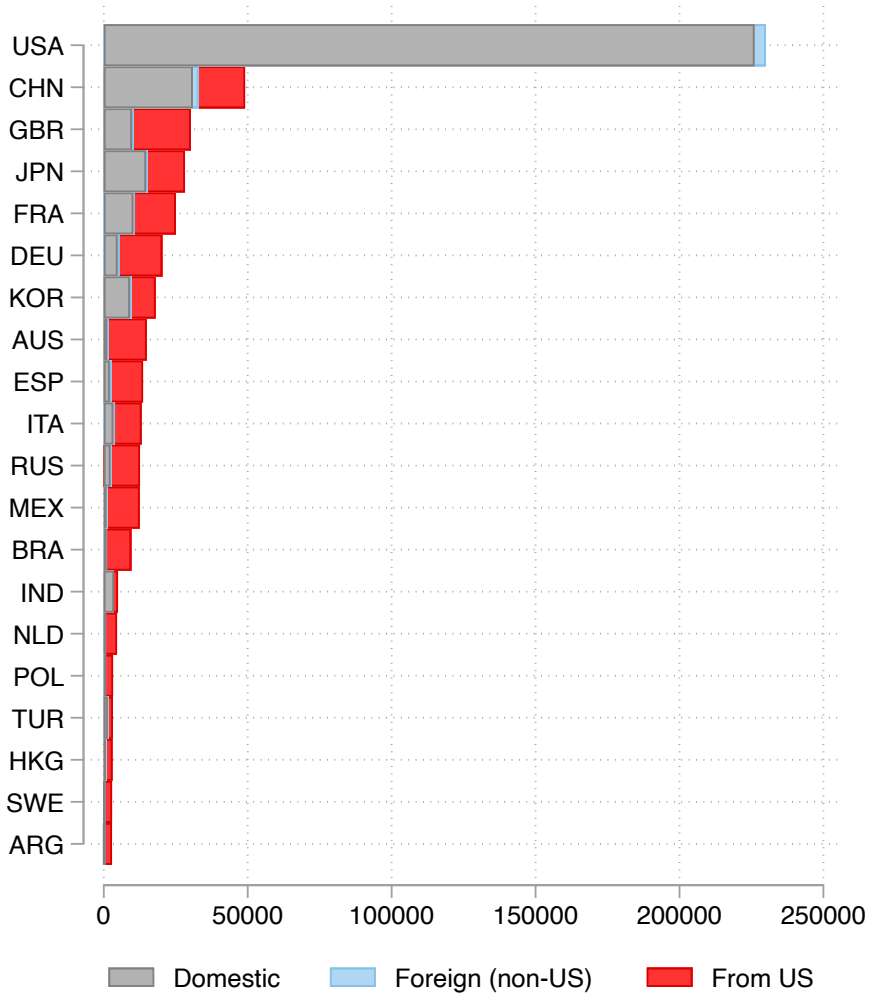
Table C1: Displayed movies by country

		Revenue sample		Screening sample	
		Foreign	Domestic	Foreign	Domestic
Albania	ALB	2	0	5	0
Argentina	ARG	3049	254	3092	423
Australia	AUS	4590	272	4657	418
Austria	AUT	3619	117	3654	172
Bahrain	BHR	516	0	548	0
Bangladesh	BGD	0	0	3	0
Belgium	BEL	3432	313	3480	406
Bolivia	BOL	1858	1	1893	8
Brazil	BRA	2933	81	2973	373
Bulgaria	BGR	2106	30	2124	84
Cambodia	KHM	10	0	14	0
Chile	CHL	2132	34	2151	63
China	CHN	722	441	772	600
Colombia	COL	2194	19	2237	53
Costa Rica	CRI	51	0	64	0
Croatia	HRV	1852	10	1879	24
Cyprus	CYP	153	0	203	0
Czech Republic	CZE	2392	164	2414	252
Denmark	DNK	1858	96	1876	223
Dominican Republic	DOM	259	1	296	1
Ecuador	ECU	1460	0	1475	0
Egypt	EGY	1486	22	1503	51
El Salvador	SLV	0	0	3	0
Estonia	EST	1078	5	1093	11
Ethiopia	ETH	74	0	90	0
Finland	FIN	2201	62	2211	271
France	FRA	3142	2074	3387	2739
Germany	DEU	3414	1082	3443	1313
Ghana	GHA	225	0	253	0
Greece	GRC	2912	28	2941	118
Guatemala	GTM	0	0	2	0
Honduras	HND	36	0	52	0
Hong Kong	HKG	2819	410	2894	552
Hungary	HUN	2263	53	2280	127
Iceland	ISL	2371	29	2394	75
India	IND	840	363	863	594
Indonesia	IDN	750	0	772	2
Iraq	IRQ	2	0	6	0
Israel	ISR	1121	2	1141	2
Italy	ITA	3753	407	3834	1016
Jamaica	JAM	210	0	247	0
Japan	JPN	1464	607	1475	1031
Jordan	JOR	387	1	422	1
Kenya	KEN	156	1	205	1
Kuwait	KWT	327	0	364	0
Laos	LAO	0	0	1	0

Table C1: Displayed movies by country (continued)

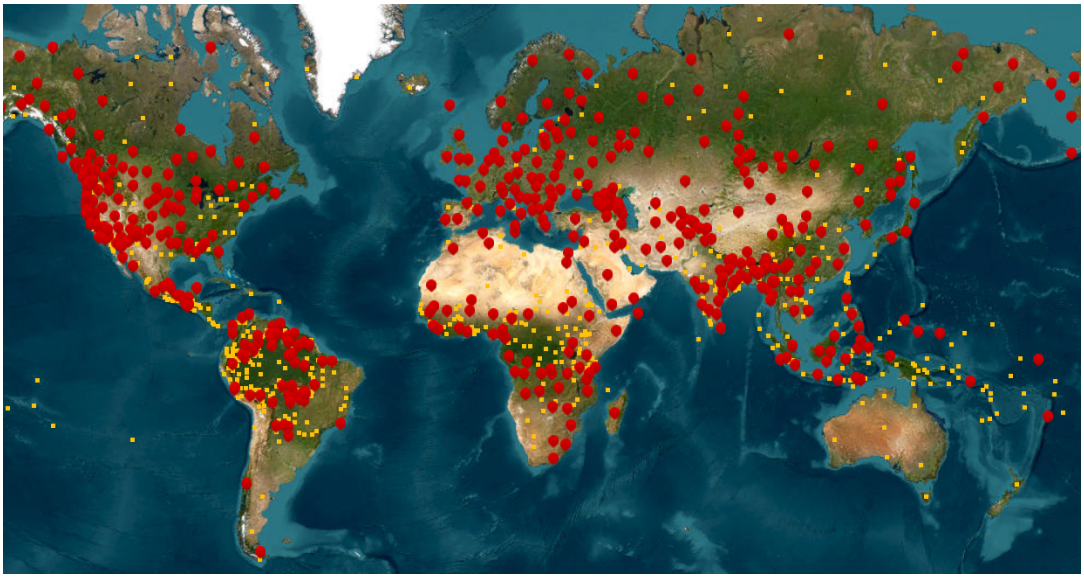
		Revenue sample		Screening sample	
		Foreign	Domestic	Foreign	Domestic
Latvia	LVA	1144	5	1162	8
Lebanon	LBN	2331	8	2355	17
Lithuania	LTU	1858	19	1892	62
Malaysia	MYS	2130	8	2206	19
Malta	MLT	0	0	1	0
Mexico	MEX	3523	186	3564	341
Mongolia	MNG	56	0	70	0
Netherlands	NLD	3299	138	3324	376
New Zealand	NZL	4184	67	4255	107
Nicaragua	NIC	33	0	47	0
Nigeria	NGA	718	0	749	0
Norway	NOR	2684	116	2733	303
Oman	OMN	549	0	580	0
Pakistan	PAK	108	1	134	2
Panama	PAN	140	0	150	0
Paraguay	PRY	466	3	474	3
Peru	PER	2070	16	2108	29
Philippines	PHL	1796	14	1827	68
Poland	POL	2486	104	2503	279
Portugal	PRT	3719	38	3801	126
Puerto Rico	PRI	176	1	206	1
Qatar	QAT	339	1	373	1
Romania	ROU	1724	47	1760	128
Russia	RUS	3736	332	3848	618
Saudi Arabia	SAU	0	0	4	0
Serbia	SRB	1521	22	1542	71
Singapore	SGP	2863	38	2900	49
Slovakia	SVK	1864	44	1875	57
Slovenia	SVN	1845	11	1872	36
South Africa	ZAF	2943	34	2959	108
South Korea	KOR	2888	365	3055	844
Spain	ESP	4498	513	4591	1055
Sri Lanka	LKA	0	0	15	0
State of Palestine	PSE	2	0	7	0
Sweden	SWE	1998	99	2016	223
Switzerland	CHE	1268	21	1292	24
Taiwan	TWN	2567	55	2617	57
Thailand	THA	2310	89	2336	185
Turkey	TUR	3128	242	3207	752
Ukraine	UKR	2048	22	2067	53
United Arab Emirates	ARE	3627	29	4007	38
United Kingdom	GBR	4936	1409	5362	1806
United States	USA	2601	4710	3479	8508
Uruguay	URY	2190	26	2209	35
Venezuela	VEN	1639	8	1660	17
Vietnam	VNM	578	4	609	9
Movies		13929		28810	

Figure C1: Total revenue in top 20 consuming countries



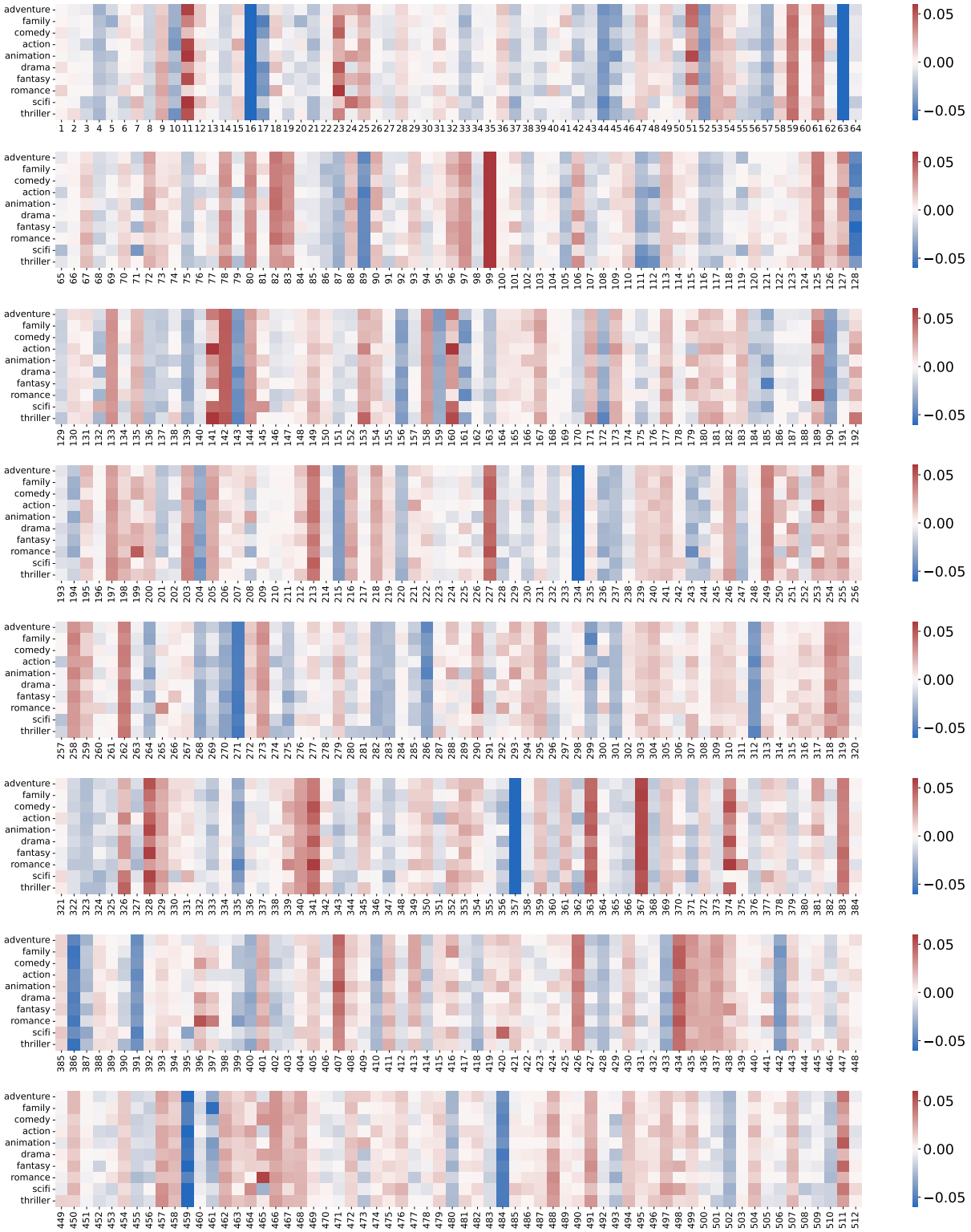
Notes: In million US$.

Figure C2: Motif k27 - The Hero with a Thousand Faces, Joseph Campbell



Notes: The red dots denote the presence of motif k27 in the group's oral tradition. Yellow are groups without this motif. Motif k27 reads: The character receives deadly tasks that can only be completed with supernatural abilities or assistants; the hero completes tasks and/or miraculously stays alive. The confrontation of the characters unfolds as a game or competition in which the loser loses his life or status.

Figure C3: Average loadings of USE by genre of movie



Notes: The x-axis displays the dimension of the embedding (from 1 to 512) and the y-axis displays the genre. The shade of the cell indicates the loading onto that dimension by genre.

Table C2: Top 15 keywords of each of 5 topics

Topic 0:	young love family friend woman man life old father girl mother meet find son live
Topic 1:	world war save mysterious fight force journey group power adventure secret battle human man event
Topic 2:	time come house people find night run know home live return different turn tell couple
Topic 3:	life school student dream team play town world follow struggle group win career star work
Topic 4:	life police work murder find job money crime try plan officer drug city lead criminal

Table C3: Top 15 keywords of each of 10 topics

Topic 0:	man friend woman old meet relationship good year life friendship spend childhood good_friend beautiful decide
Topic 1:	mysterious adventure turn human embark island road dangerous evil create attack hero journey rescue travel
Topic 2:	find come home return village house run night know time away happen turn help dead
Topic 3:	school student team play win documentary company teacher game follow challenge director college car race
Topic 4:	city police murder lead crime officer drug criminal involve detective killer investigate cop personal french
Topic 5:	wife child son leave start discover life begin couple husband past decide parent time marriage
Topic 6:	girl town plan dream star music head real teenage artist kid career steal famous girlfriend
Topic 7:	love work fall life brother want look find sister job money try live marry teenager
Topic 8:	world people group death war face fight force save kill deal battle survive attempt protect
Topic 9:	young family life father mother boy struggle daughter learn true journey die live change different

Table C4: Top 15 keywords of each of 30 topics

Topic 0: friend good spend good_friend party bad boyfriend test forever fun crush cheat super extraordinary courage
Topic 1: event adventure catch road dangerous path cross appear inside trap forest creature animal sea figure
Topic 2: house learn marriage happen wedding strange stay disappear partner prepare pass room surprise complicated word
Topic 3: village age win enter chance race risk remote build perfect success competition opportunity crash unlikely
Topic 4: lead city body future mission killer prison encounter reality question track leader quest kind watch
Topic 5: turn head single revenge lover stop owner wrong affair haunt ancient neighbor expect news convince
Topic 6: discover town girlfriend hope teacher arrive kid situation think gets bank pair strong prevent teach
Topic 7: help brother want sister marry money need trouble rich offer month pay condition debt african
Topic 8: world war fight soldier far happy conflict light blind surround corruption minister unusual river warrior
Topic 9: time follow different break beautiful tell visit feel choose society youth separate desperate rule loss
Topic 10: company battle survive threaten attack fear action rescue special fate explore continue task guard line
Topic 11: man relationship childhood community worker free self sell accept near study idea difficult musician let
Topic 12: group escape island fail pregnant club military throw pursue walk flee shoot magic threat sexual
Topic 13: try start save star hard king hit apart water accidentally capture wait living queen huge
Topic 14: secret murder crime criminal involve investigate protect commit suspect witness draw survival suicide victim tragic
Topic 15: send member receive ask share hero doctor bond master enemy patient fire claim agree charge
Topic 16: old lose play job parent game train player stranger having cat grandfather rival slowly engage
Topic 17: boy death begin struggle mysterious past search truth dark memory cause suddenly identity driver upside
Topic 18: family father mother son home daughter young_man care law dog raise lonely orphan return_home aged
Topic 19: order plan deal teenager hide business steal right reveal social army female kidnap tell grandmother
Topic 20: find meet face travel seek heart accident experience suffer role perform prince song ill lot
Topic 21: young woman girl young_woman evil book spirit young_girl write unfold gain global pretty regain possess
Topic 22: love fall know true challenge deadly state release plot support forget lady horse thief allow
Topic 23: force kill attempt grow hunt powerful deep prove rise wish mountain promise destroy danger innocent
Topic 24: decide return look drug real music join dead sex act matter reunite dance talk wild
Topic 25: child people die human personal bear abandon confront baby ghost decision choice chase hear birth
Topic 26: journey bring power summer create control earth remain mean divorce date alive supernatural alien disease
Topic 27: come leave dream night realize stand wake singer final reach overcome portrait demon buy hot
Topic 28: wife couple husband run away unexpected believe miss lie married male run_away spy married_couple adopt
Topic 29: life live work change problem successful drive solve desire farm poor private simple issue extreme

Table C5: Top 15 keywords of each of 40 topics

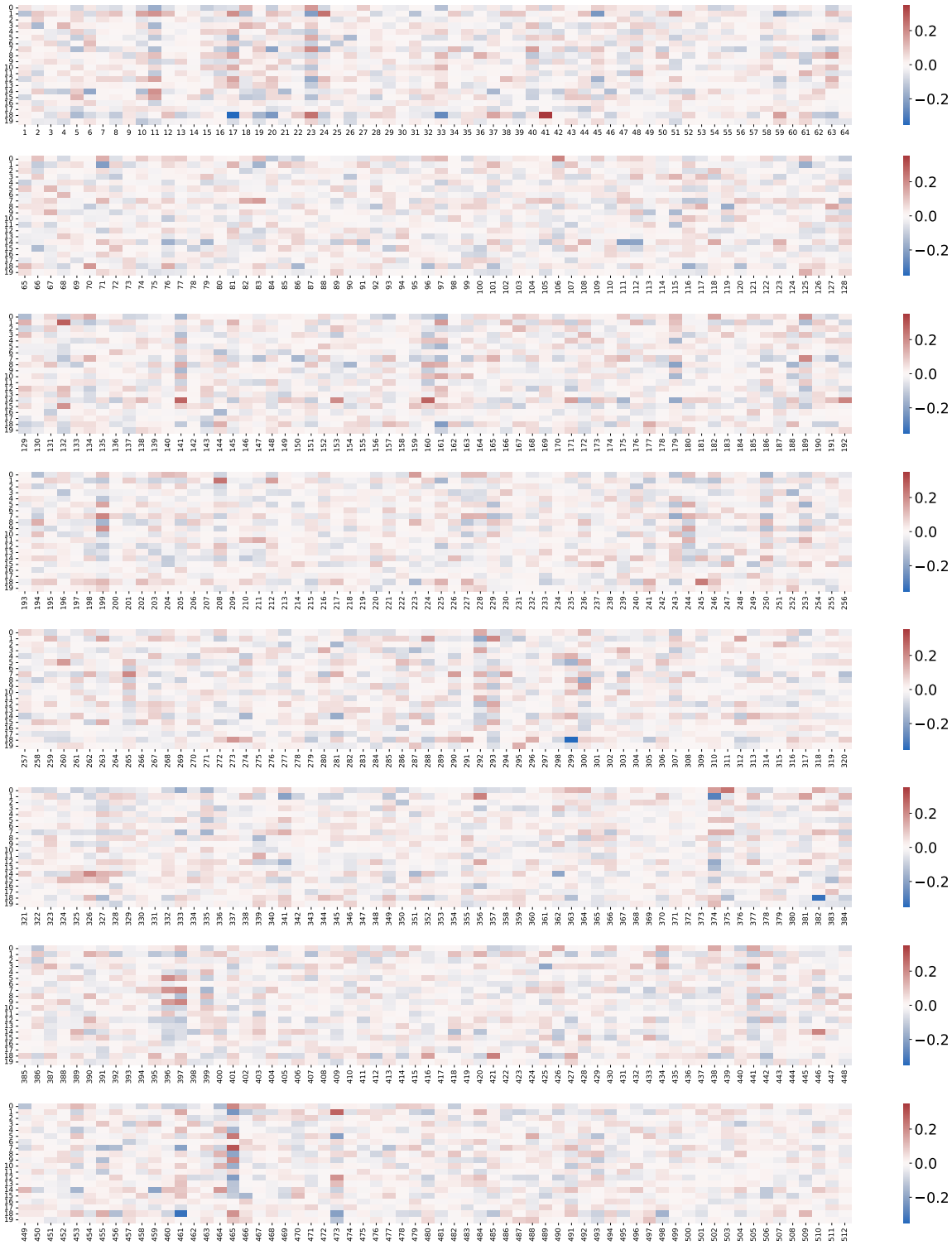
Topic 0: friend group good spend good_friend party boyfriend forever change intimate crush cheat super extraordinary courage
Topic 1: event adventure action cross appear trap forest animal sea figure fly outside hunter space unable
Topic 2: find learn marry marriage wedding strange worker disappear prepare surprise complicated word incident provide cold
Topic 3: village age win stop enter race test remote build near competition opportunity crash unlikely beginning
Topic 4: lead travel revenge killer prison encounter reality question leader ghost quest kind watch entire recover
Topic 5: head dark lover wrong affair haunt ancient neighbor expect news border allow door target stick
Topic 6: town girlfriend hope teenager arrive situation owner baby gets bank strong unique winter underworld performance
Topic 7: leave brother want money bad confront condition debt african interest origin princess hidden impossible fighter
Topic 8: world war kill hunt powerful happy light blind rival surround corruption minister unusual warrior spanish
Topic 9: time different break visit feel society youth desperate loss holiday connect manage beloved fan sense
Topic 10: work begin company threaten attack fear rescue fate explore suspect guard line god violence hell
Topic 11: relationship dream struggle star childhood experience community self musician refuse strike avoid direct participate examine
Topic 12: run escape away island club throw pursue walk shoot sexual accuse strength run_away expose rest
Topic 13: come free hit inside water master accidentally capture alive carry wait beauty bird huge red
Topic 14: secret murder crime criminal investigate commit witness act draw plot survival suicide victim tragic camp
Topic 15: hide member lie receive share hero ask doctor bond enemy patient claim freedom agree food
Topic 16: start lose play job parent game train player stranger having teach cat engage settle introduce
Topic 17: boy death mysterious past memory cause suddenly driver upside overcome unknown answer young_boy cousin underground
Topic 18: man home young_man care dog lonely return_home aged happiness guide handsome meaning sibling sick drag
Topic 19: force tell protect right reveal army state grandmother process charge magic public horror trial justice
Topic 20: find face sister seek heart accident role song ill lot illegal uncle voice elderly enjoy
Topic 21: young girl young_woman truth evil book spirit pregnant young_girl write unfold gain regain possess sight
Topic 22: love know true challenge stand soldier release support forget lady prince horse ride employee sacrifice
Topic 23: fall attempt grow dangerous risk rise deep wish mountain task destroy blood await bury cut
Topic 24: decide return look join dead success matter reunite dance dancer lord consider gift normal ultimately
Topic 25: child people bring die catch bear abandon choice talk flee chase supernatural hear culture soul
Topic 26: summer create earth divorce date large threat alien disease ordinary connection cover magical extreme understand
Topic 27: night happen realize suffer wake singer final continue reach portrait demon fun hot approach meeting
Topic 28: wife couple husband believe miss female married male spy married_couple adopt construction behavior ruin stake
Topic 29: family meet problem successful think social drive able solve desire poor private simple issue attract
Topic 30: life change involve kid personal chance rich fail apart perfect study tear service focus nature
Topic 31: power send human future mission special deadly creature prevent compete low global emerge moon disturb
Topic 32: live house road stay far path accept month decision perform invite slowly destiny common medium
Topic 33: father mother son daughter single law kidnap raise promise let unexpectedly grandfather pick value intention
Topic 34: turn order survive king choose prove rule remain mind wild serve twin birth existence determine
Topic 35: woman old military pair pay room living curse store earn dump lack old_woman invade silence
Topic 36: discover city drug real teacher business sex partner conflict track separate disappearance organize guest official
Topic 37: try plan music deal need steal hard sell offer mean fire idea convince thief buy
Topic 38: follow journey save search beautiful body unexpected control identity farm orphan kingdom responsible terrible model
Topic 39: help fight battle trouble pass difficult danger important consequence adult letter gather accompany relative achieve

Table C6: Top 15 keywords of each of 100 topics

Topic 0:	woman fate soul intimate rest inner hate circle companion taste suffering acquire dimension lift infant
Topic 1:	age dangerous inside creature sea tear fly flight collide fisherman niece fishing rabbit mistaken ash
Topic 2:	tell wedding believe word provide survivor earn wing seduce shadow find wash entirely guilty young
Topic 3:	win enter race continue competition crash ready edge imprison contest chain norwegian gap kick abel
Topic 4:	boy revenge quest surprise condition young_boy watch recover convict link grand ice lost wilderness provoke
Topic 5:	begin affair suspect ancient neighbor news door demand level love_affair quick crown sort climb hiding
Topic 6:	people bear cross beginning river wood mark reject rape earthquake strip everybody shortly alternative time
Topic 7:	brother confront sexual interest origin pull princess young impossible young_brother fairy bed network heel instantly
Topic 8:	world dark happy forever light red performance life tie execute dragon descend june patriarch dawn
Topic 9:	search break desperate life happiness fan sense responsibility fix kiss rogue alternate everyday_life man_travel repeatedly
Topic 10:	escape fear rescue master guard security major wall bird finish arise dress knock torture eliminate
Topic 11:	dream childhood hard self funny priest church transformation bite got severe sexually maid opposition life
Topic 12:	group power club accuse gold difficulty ultimate life restore fatal round embrace man silver scary
Topic 13:	home prepare capture divorce wait attention cancer ally ocean hostile location speech come_home rip perceive
Topic 14:	send create shoot effort space complete deceased fake fish immediately transfer election confess revive listen
Topic 15:	kill hide hero animal freedom food eat cut original image wound jump leg garden demonic
Topic 16:	job parent promise having thief existence introduce old_man ram sleep trick disorder source ritual spoil
Topic 17:	turn past suddenly disappear month upside overcome emerge mistake subject tired betray darkness acquaintance deception
Topic 18:	father lonely carry incident meaning belong drag inhabitant appearance constantly fierce angry outcast myth medicine
Topic 19:	follow magic simple public europe inherit value obstacle arm mirror wander persuade assignment eastern combat
Topic 20:	find owner entire uncle life position award occur roof loan man frozen young direction iron
Topic 21:	sister evil book spirit write aged gain deliver regain sight read honor status young_sister evil_spirit
Topic 22:	look challenge share state forget horse allow sacrifice golden ring trust shelter kidnapper totally participant
Topic 23:	grow destroy connect refuse alien mayor tree ground sweet expert maintain announce shepherd tunnel expensive
Topic 24:	deal path partner act matter dance life incredible audience suppose highly fresh previously fade mourn
Topic 25:	child bring ask choice hear culture peace voice speak language sound winner alcohol forge deaf
Topic 26:	daughter future connection extreme understand plant bright anymore eld young_daughter green hood sum absent humorous
Topic 27:	mysterious accident prison singer release serve witch gambling cruel flat stalk fox anger expedition sharp
Topic 28:	happen star head chase compete male morning adopt construction european remember produce endure victory frog
Topic 29:	meet issue attract message resort ultimately contact cope terror thank drop intrigue imaginary oppose jealous
Topic 30:	life problem involve personal chance perfect document host thought oil pupil initially chicken gambler intent
Topic 31:	start soldier let curse ship web disturb lock life site find shift undertake fragile sail
Topic 32:	house social slowly medium dinner pretty life effect newborn valuable dare young domestic content deny
Topic 33:	mother company single boyfriend unexpectedly grandfather pick rare number broken list neglect extremely nice malevolent
Topic 34:	live seek hell twin surgeon life block check aim twin_brother wrestle mercy quality find descendant
Topic 35:	lead lie pair official life cook lack hill trigger mile dig armed folk accidental unhappy
Topic 36:	teacher care life conflict adult responsible family young bitter charm similar reserve old_sister shell terrify
Topic 37:	plan money sell convince agree buy handsome exchange burn heir sing hatch pool sheep permanent
Topic 38:	journey save control camp kingdom root hang property destination tiger switch harsh loose spin interested
Topic 39:	son danger transform relative warrior achieve foot collapse young_son purpose scare finnish servant ahead horn
Topic 40:	work join final large living illegal trial extraordinary locate confrontation outlaw coastal custom rainy prayer
Topic 41:	help miss trouble forest nature boat individual safe spell friendly express hole rain intend bone
Topic 42:	town wrong choose female manage reason push terrible resident intertwine life limit rocky holy supposedly
Topic 43:	decide village decision retire fun rob hot lake wonder surface vicious spring landscape frame appoint
Topic 44:	love draw poor unknown deeply gradually secretly devil sure christ mainly man_fall vessel example flock
Topic 45:	time spend visit society accidentally grandmother cold elderly unusual sport snow august material heaven warn
Topic 46:	real law stand kidnap success perform walk song unfold fill collect stake secure ransom supply
Topic 47:	adventure far baby expect service strike guide approach aunt young slave narrative life combine reign
Topic 48:	come true action wake figure teach quiet underworld hidden behavior come_true prominent absurd undergo writing
Topic 49:	king attack bank hunter queen kill monster model robber current season upper man conversation handle
Topic 50:	man reveal community loss answer engage await notice invasion find explain resolve intelligent weight exploit
Topic 51:	face different criminal stay system era cheat especially storm old_brother touch unite inventor speed life
Topic 52:	try learn heart mission plane define find industrial infidelity wife_leave newcomer commercial interior significant life
Topic 53:	know struggle right rule strong disappearance succeed central household angel upset retreat height flower mixed
Topic 54:	marry invite justice corruption chief organize duty man disaster launch young happily presence welcome mistress
Topic 55:	fight realize survival lady charge target cast design physical ray rush recognize heal mortal float
Topic 56:	stop bond train separate study reunite unable injustice passenger distance young humor find monkey man
Topic 57:	world appear earth rise water opposite giant sun sky mate advice bug invent suggest freak
Topic 58:	force threaten near kind consequence circumstance unique destiny prisoner gift underground unravel life trail dear
Topic 59:	change young_woman summer fail talk young winter fast direct failure mount silence pregnancy drunk life
Topic 60:	human truth protect wish support cover abduct unit air fit press beast drinking married_man knight
Topic 61:	want girlfriend farm orphan ill pretend enjoy clean paint celebration life respect ethnic labor descent
Topic 62:	dead body receive question alive actually smart wolf inform wave castle swear laugh pack dead_body
Topic 63:	attempt party trap throw young luck plunge sign life jealousy find junior endanger loop overthrow
Topic 64:	city discover hunt dancer cure recent depict isolated man_try ceremony wear bridge recognition artifact woman_find
Topic 65:	order young violence gather spy employee occupy western describe assume candidate life find opponent asian
Topic 66:	wife husband lover pass birth normal evening young_people report table length housewife crack mask arrange
Topic 67:	return young_man army plot god pursue supernatural demon return_home ability gun disguise man channel man_find
Topic 68:	rich raise claim wild focus cat ride possible find trace transport mouse pursuit require humble
Topic 69:	fall remote apart holiday reach letter treat fall_apart everyday lay remind regard theft pole entrance
Topic 70:	old memory explore spark blue clan judge old_woman double potential fund grandson add triumph wealth
Topic 71:	girl identity player young_girl outside young spread adapt psychological shark gray lazy stab knife pattern
Topic 72:	friend good good_friend goal health mental spanish willing perspective stone propose practice bullet relation voyage
Topic 73:	leave kid fire low investigator evidence tradition shape blow repair expectation leave_home adjust erupt pit
Topic 74:	need catch gets process african operation easy stick pain find bit guardian bag dirty vital
Topic 75:	lose steal track date beloved blind beach match abuse murderer box poverty item illusion dive
Topic 76:	play game road hit task role rival cultural lure staff deserted ball ensure bee apple
Topic 77:	doctor sex pregnant young patient fulfill sinister intention find ruin man divide heavy fool thrust
Topic 78:	relationship drug experience special accept portrait quickly expose court method crook treatment address phenomenon shaman
Topic 79:	cause powerful enemy prevent magical meeting blood exist weapon count venture aid request replace minor
Topic 80:	die beautiful bad suffer unlikely lord seemingly store assign balance destruction loser invade split beautiful_girl
Topic 81:	find married offer pay debt married_couple possess rid man feed clash inhabit nephew lion young
Topic 82:	discover dog abandon ghost desire idea innocent general defend distant brave bloody information doll hurt
Topic 83:	young able life prince daily assistant field man beat bind sick find wind getting sink
Topic 84:	battle reality risk build disease life treasure world complex historical establish palace cup legal wonderful
Topic 85:	family event survive unexpected member movement grant tournament tribe seriously temporary family_member diamond fianc importance
Topic 86:	situation worker test life senior huge find young nation universe false entrust gender elaborate possibility
Topic 87:	travel time musician private accompany fortune funeral nearly young life possession discuss injure find manner
Topic 88:	youth man remain opportunity ordinary life participate nurse young fiance dump pose previous concern plague
Topic 89:	run away music mountain run_away crush illness defeat shake weather refuge natural roll pre detail
Topic 90:	war strange island surround avoid skill fighter straight rank bargain pacific frightening unstable man_meet necessary
Topic 91:	young marriage feel horror bride man corpse grave groom snake insist elder find quietly insult
Topic 92:	death encounter deep driver lot minister bury moon spot grief contain grieve tiny select shed
Topic 93:	life night stranger tragic important strength cousin common humanity version unaware consist exactly vehicle prey
Topic 94:	secret teenager investigate deadly beauty determine settle protagonist discover missing painting conceal man_return butcher remedy
Topic 95:	man drive mean mind consider drink afraid brain easily awkward extra blame mix regularly branch
Topic 96:	murder crime commit haunt solve suicide room complicated summon commit_suicide attractive reflect measure cool delicious
Topic 97:	hope killer free prove witness line victim threat avenge innocence declare maker villain multiple predator
Topic 98:	arrive successful think leader flee find guest clear nearby board conduct luxurious life marine isolation
Topic 99:	couple business young military difficult border east global sibling young_couple contract string peasant find smoke

OA Figure C4 shows how each of the 512 embeddings and 20 topics correlate with each other. While some entries show no clear correlation with any topic, e.g. embedding 14, others show both strong negative and positive correlations with many topics, e.g. embedding 465.

Figure C4: Correlation between topic share and embedding loading

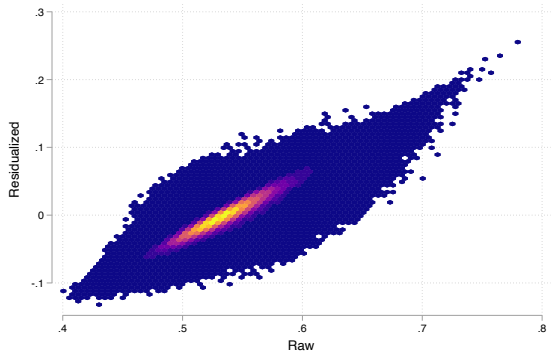


Notes: The x-axis displays the dimension of the embedding (from 1 to 512). The y-axis indicates the topic number. The shade of the cell indicates the correlation between the two across all motifs and movies.

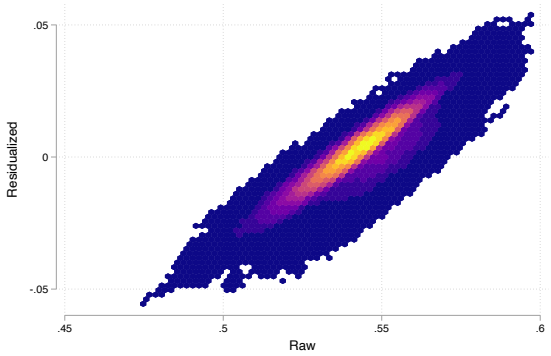
Figure C5: Heatmaps of cosine similarity and Jensen-Shannon distance

Cosine similarity of USE

Panel A: Motifs to movies

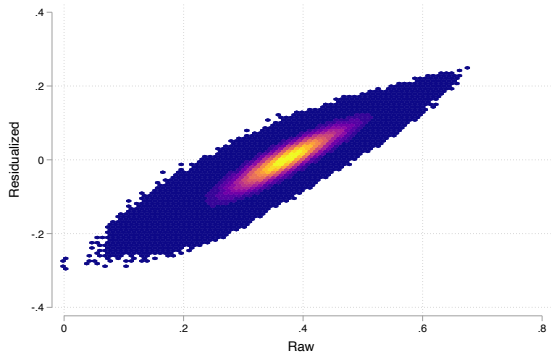


Panel B: Countries to movies

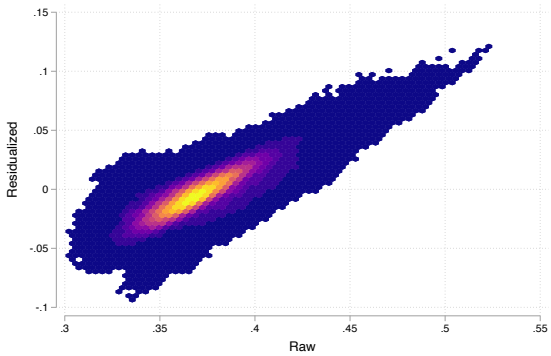


JSD of 20 topics (LDA)

Panel C: Motifs to movies

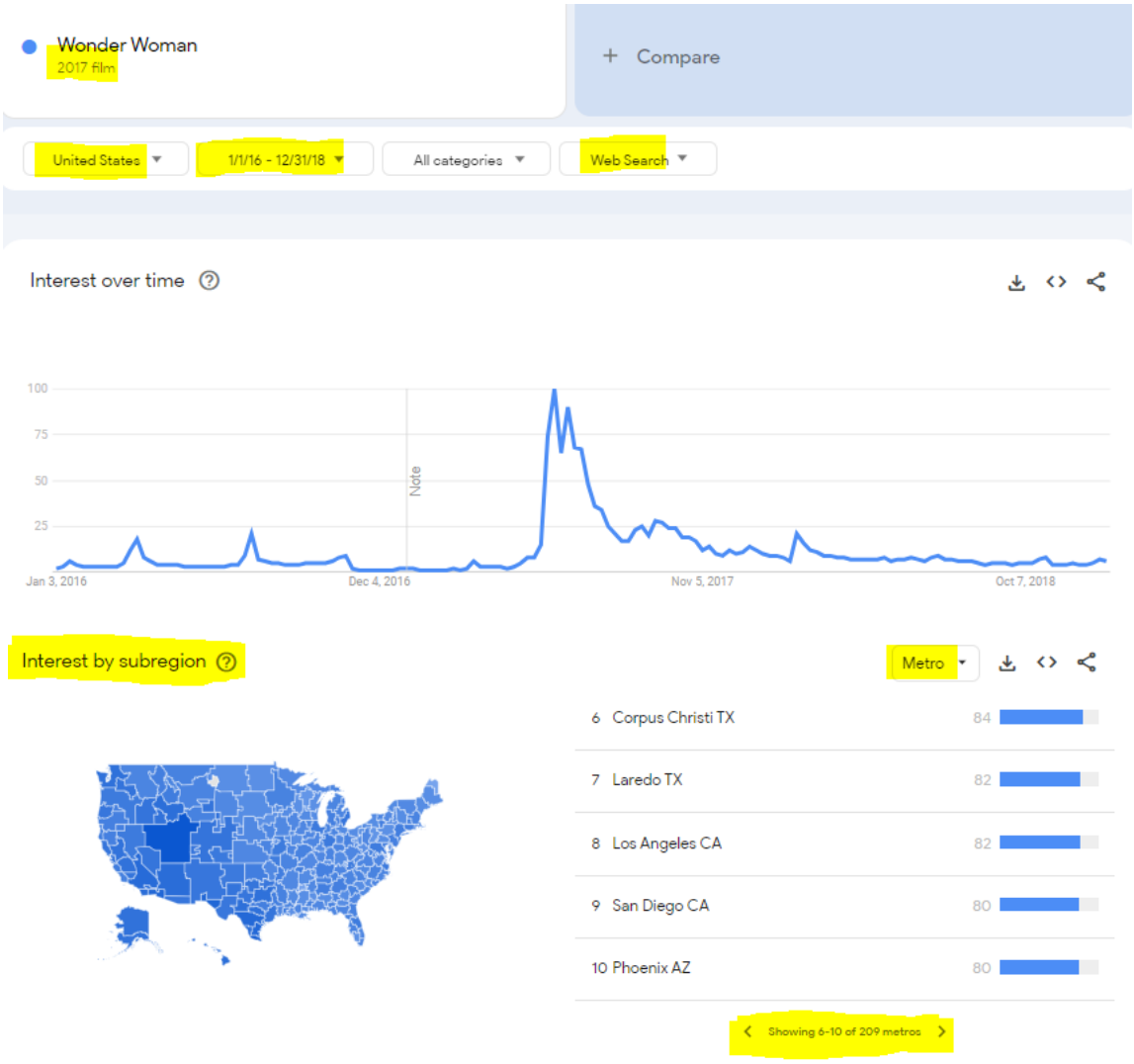


Panel D: Countries to movies



Notes: The top Panels show the heatmaps of the cosine similarity of the embeddings from the USE of motifs and movies (Panel A) and countries and movies (Panel B). The bottom Panels show the heatmaps of the JSD of 20 topics of motifs and movies (Panel C) and countries and movies (Panel D).

Figure C6: Example from Google Trends



Notes: The figure shows the Google Trends data for the movie Wonder Woman which was theatrically released worldwide on June 2, 2017. The top Panel shows the score in a one-year window, and the bottom Panel shows the search interest score during this time across DMAs.

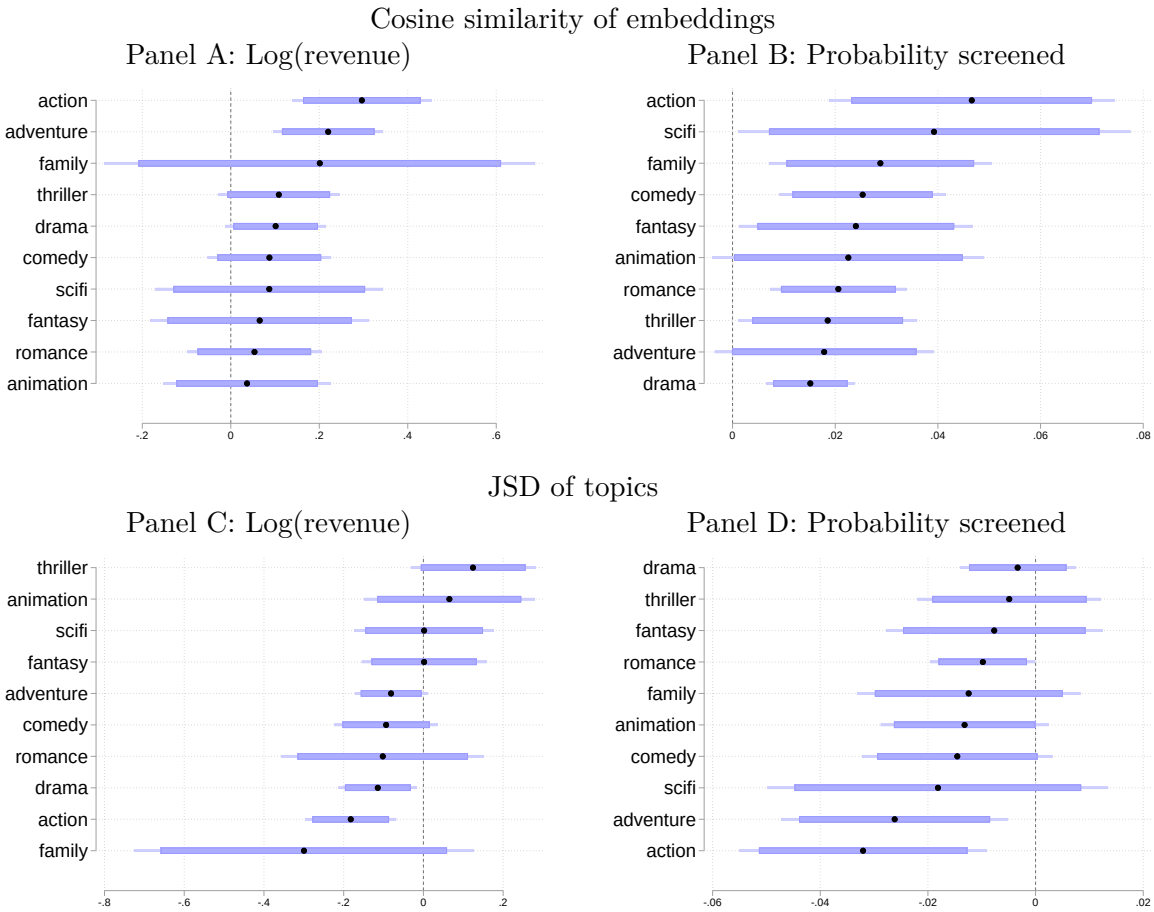
Table C7: Gender roles in motifs

	Male character is:		Female character is:	
	N	mean	N	mean
Engaged in domestic affairs	801	0.0724	748	0.1297
Intelligent	801	0.1511	748	0.1096
Naïve	801	0.1024	748	0.0816
Other	801	0.2185	748	0.2072
Physically active	801	0.1511	748	0.0842
Submissive	801	0.1336	748	0.3035
Dominant	801	0.3283	748	0.1457

Notes: This table reproduces Appendix Table 8 of MX and reports summary statistics on how male and female characters are depicted in Berezkin's motif catalogue. The classification of each motif into stereotypes is done by 9 Mturks on average.

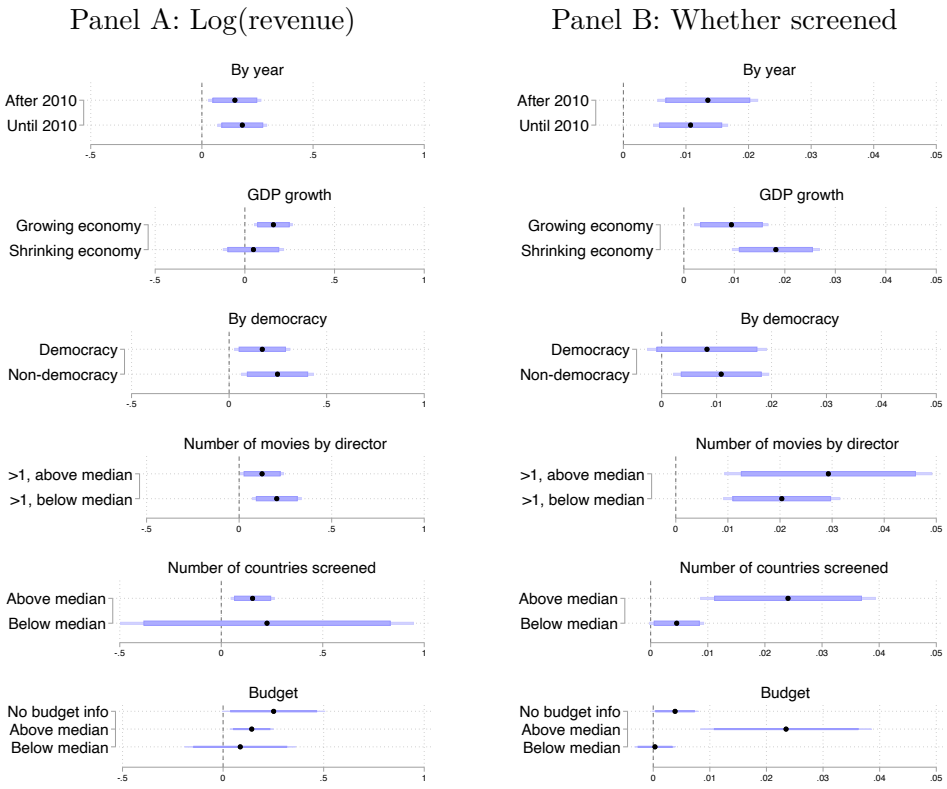
D Heterogeneity and robustness

Figure D1: Coefficient of similarity/distance by genre for foreign movies



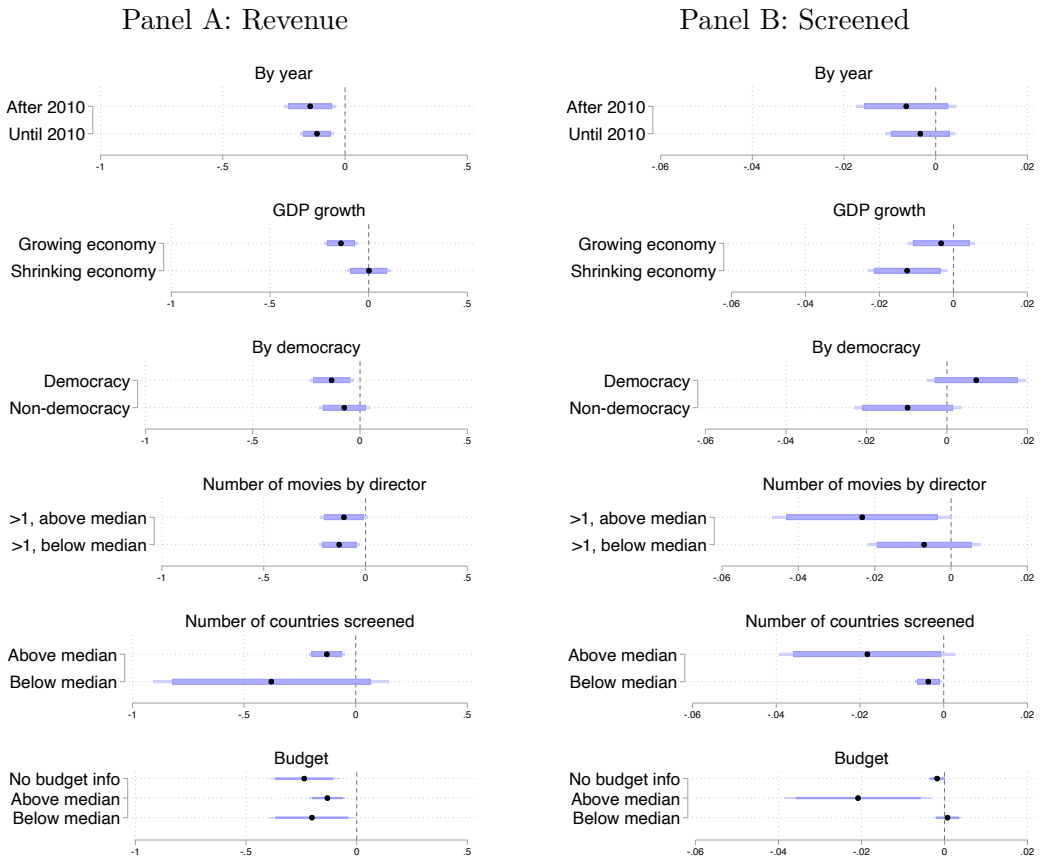
Notes: In Panels A and C controls include movie, and country-year-month fixed effects. In Panels B and D controls include movie and country-year-of-release fixed effects. Standard errors are double clustered at the year and country level. Cosine similarity and JSD are residualized using the log(number of words) of the movie summary, motif, and their interaction. The thick bars represent 90% and the thin bars 95% confidence intervals.

Figure D2: Heterogeneity of cosine similarity of embeddings explaining outcomes for foreign movies



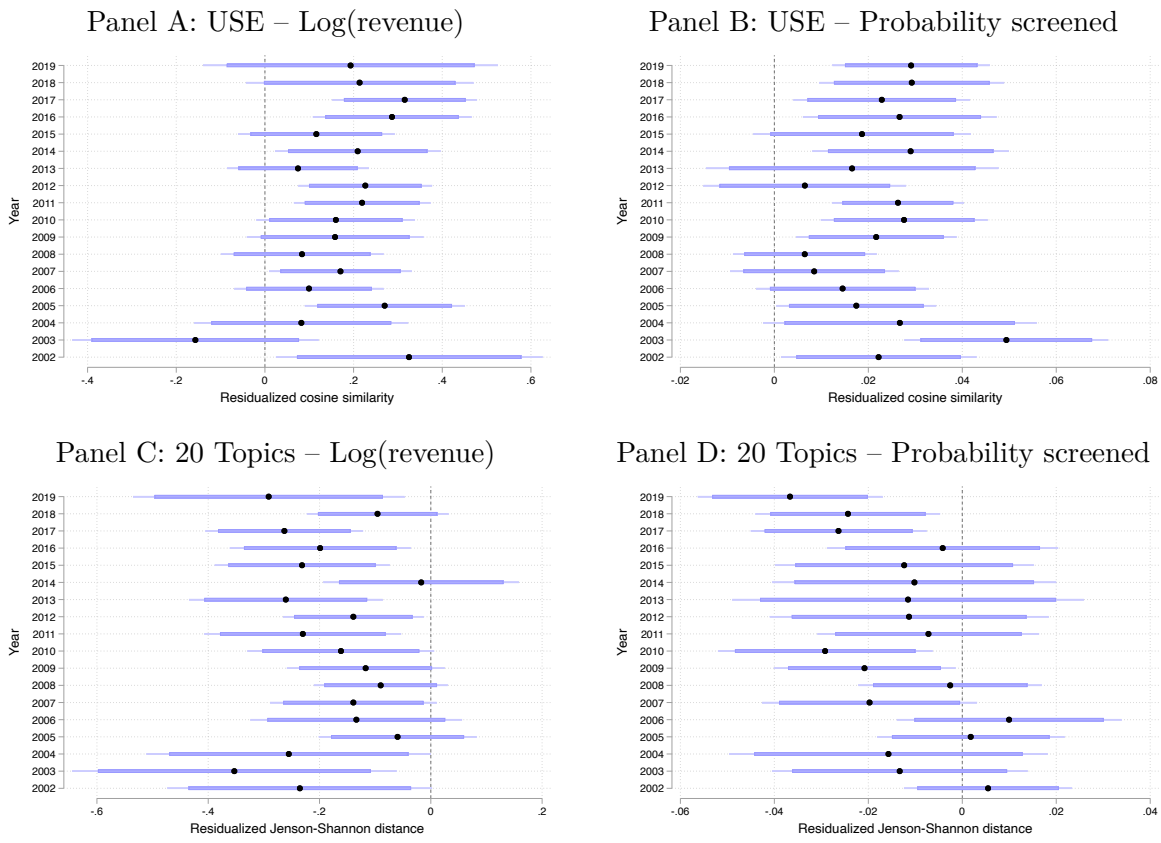
Notes: Controls include movie fixed effects, and country-year-month fixed effects for Panel A and country-year of release fixed effects for Panel B. Standard errors are double clustered at the country and year level. Cosine similarity is residualized using the log(number of words) of the movie summary, motif, and their interaction. The thick (thin) bars represent the 95% (90%) confidence intervals. For the number of movies by director we consider the director that directs the most movies and drop 4,249 movies (27,093 observations) with directors that have only directed one movie. Amongst the remaining sample we split at the median.

Figure D3: Heterogeneity of coefficients of Jensen-Shannon distance of topics explaining outcomes for foreign movies



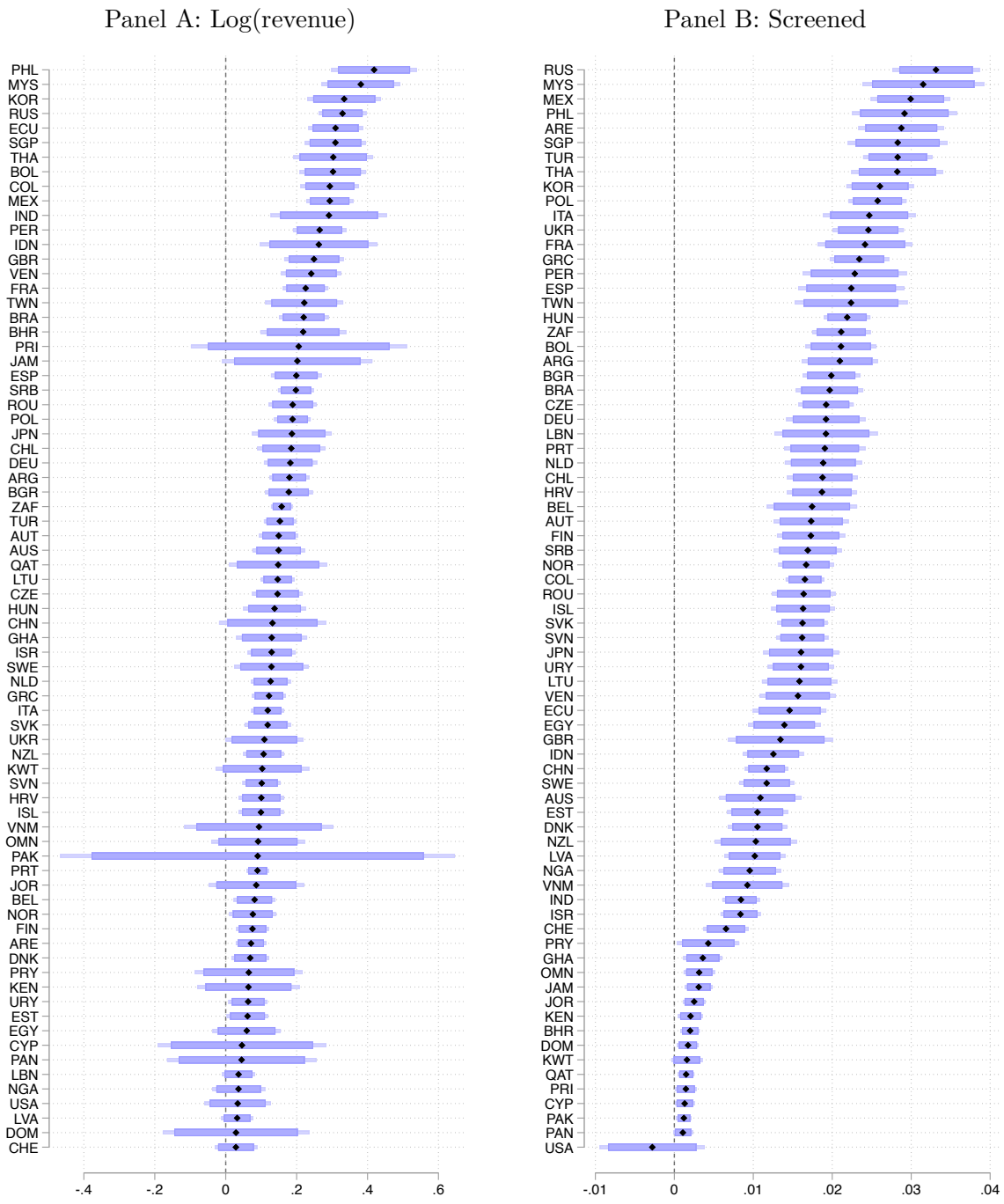
Notes: Controls include movie, and country-year-month fixed effects for Panel A and country-year of production fixed effects for Panel B. JSD is residualized using the log(number of words) of the movie summary, motif, and the interaction.

Figure D4: Coefficient of distance/similarity by year for foreign movies and outcomes



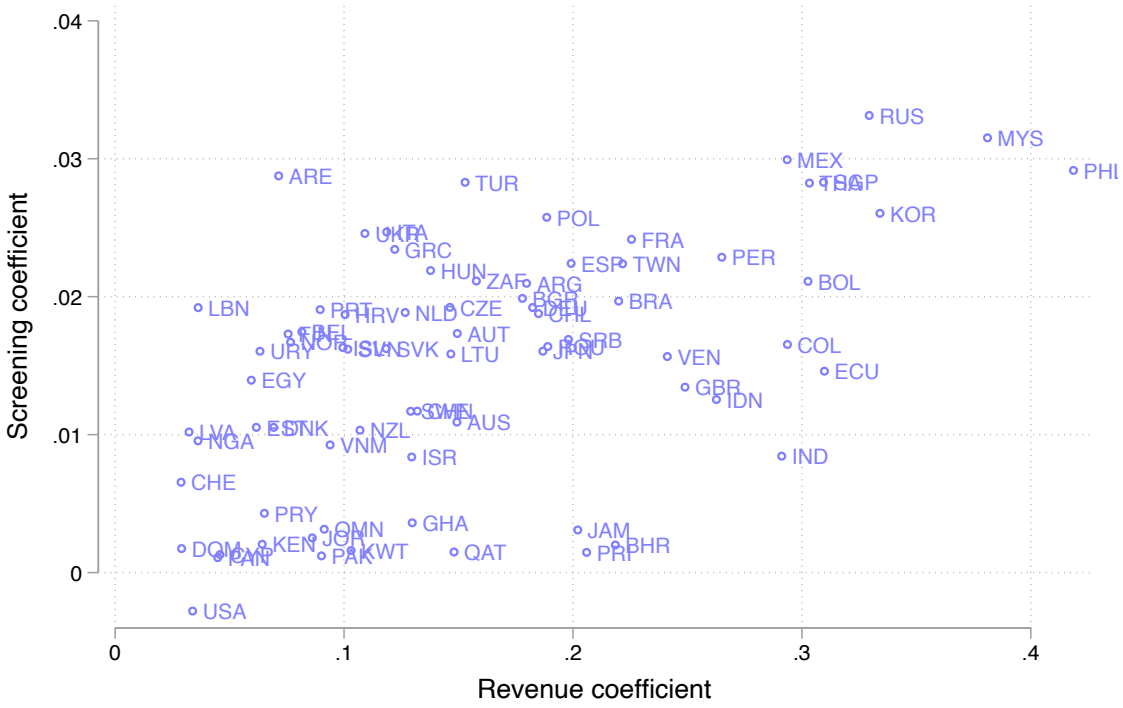
Notes: In Panels A and C controls include movie, and country-year-month fixed effects. In Panels B and D controls include movie and country-year of release fixed effects. Standard errors are clustered at the country level. Cosine similarity and JSD are residualized using the log(number of words) of the movie summary, motif, and their interaction.

Figure D5: Cosine similarity of embeddings explaining outcomes of foreign movies within countries



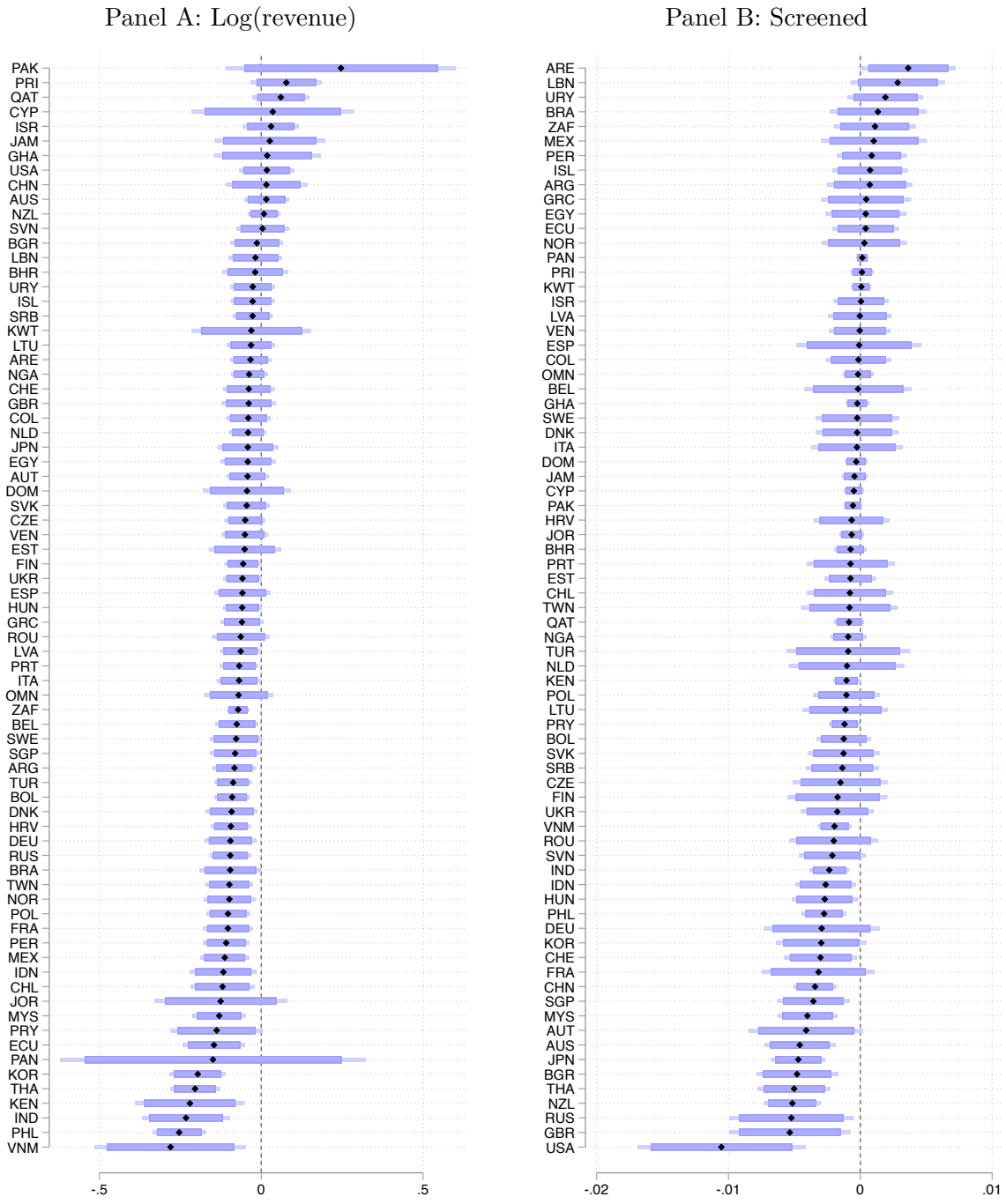
Notes: Controls include year-month, language, and country of origin of financing. Standard errors are clustered at the year level. Cosine similarity is residualized using the log(number of words) of the movie summary, motif, and their interaction. The sample is restricted to countries with at least 100 observations. The thick (thin) bars represent the 95% (90%) confidence intervals.

Figure D6: Scatter plot of coefficient of cosine similarity of embeddings explaining outcomes of foreign movies within countries



Notes: This figure plots the coefficients presented in Figure D5 with the coefficients from Panel A on the x-axis and from Panel B on the y-axis.

Figure D7: JSD of 20 topics explaining outcomes within countries



Notes: Controls include year-month, language, and country of origin of financing. Standard errors are clustered at the year level. JSD is residualized using the log(number of words) of the movie summary, motif, and their interaction. The sample is restricted to countries with at least 100 observations. The thick (thin) bars represent the 95% (90%) confidence intervals.

Table D1: Log(revenue) explained by distance to country's motifs controlling for directors, actors, writers and distributors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Residualized cosine similarity								
Universal encoder	0.1409**	0.1088*	0.1702**	0.2118***	0.2291***	0.1368***	0.1983***	0.1163**
	(0.0666)	(0.0524)	(0.0644)	(0.0639)	(0.0620)	(0.0430)	(0.0564)	(0.0443)
Observations	68500	68500	119017	119017	79313	79313	89994	89994
R^2	0.8310	0.9286	0.8027	0.9272	0.8269	0.9355	0.8138	0.8737
Number of movies	3962	3962	9041	9041	5175	5175	9437	9437
Number of countries	67	67	62	62	64	64	81	81
Country-year-month FE	✓	✓	✓	✓	✓	✓	✓	✓
Movie FE	✓	✓	✓	✓	✓	✓	✓	✓
Country-director FE		✓						
Country-actor FE				✓				
Country-writer FE						✓		
Country-distributor FE								✓
Panel B: Residualized Jensen-Shannon distance								
20 topics	-0.1009	-0.0953*	-0.1129**	-0.1249***	-0.1429***	-0.1471***	-0.1138**	-0.1193***
	(0.0586)	(0.0502)	(0.0527)	(0.0361)	(0.0387)	(0.0369)	(0.0511)	(0.0384)
Observations	68500	68500	119017	119017	79313	79313	89994	89994
R^2	0.8310	0.9286	0.8026	0.9272	0.8268	0.9355	0.8137	0.8737
Number of movies	3962	3962	9041	9041	5175	5175	9437	9437
Number of countries	67	67	62	62	64	64	81	81
Country-year-month FE	✓	✓	✓	✓	✓	✓	✓	✓
Movie FE	✓	✓	✓	✓	✓	✓	✓	✓
Country-director FE		✓						
Country-actor FE				✓				
Country-writer FE						✓		
Country-distributor FE								✓

Notes: OLS regressions. JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction, and are standardized with mean zero and a standard deviation of one. Directors, actors, writers, and distributors with at least five appearances are included as dummy variables and interacted with country fixed effects. Standard errors double clustered at country and year level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table D2: Log(revenue) explained by distance to country's motifs (robustness checks)

	(1)	(2)	(3)	(4)
Panel A: Residualized cosine similarity				
Cosine similarity folklore to movie	0.1709***	0.1820***	0.1657***	0.1392**
	(0.0593)	(0.0612)	(0.0519)	(0.0526)
Cosine similarity folklore to movie squared		-0.0203		
		(0.0283)		
Cosine similarity to past movies				0.7015***
				(0.1862)
Observations	149506	168626	168512	168512
R^2	0.8273	0.7838	0.7837	0.7853
Number of movies	13485	13929	13925	13925
Number of countries	81	85	83	83
Movie FE	✓	✓	✓	✓
Country-year-month-day FE	✓			
Country-year-month FE		✓	✓	✓
Panel B: Residualized Jensen-Shannon distance				
JSD folklore to movie (20 topics)	-0.1168**	-0.1323***	-0.1302***	-0.1248***
	(0.0471)	(0.0389)	(0.0391)	(0.0376)
JSD folklore to movie squared (20 topics)		-0.0169		
		(0.0133)		
Surprise relative to past movies				-0.1442*
				(0.0814)
Observations	149506	168626	168512	168512
R^2	0.8273	0.7838	0.7837	0.7837
Number of movies	13485	13929	13925	13925
Number of countries	81	85	83	83
Movie FE	✓	✓	✓	✓
Country-year-month-day FE	✓			
Country-year-month FE		✓	✓	✓

Notes: OLS regressions. JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction, and are standardized with a mean of zero and a standard deviation of one. Surprise for topics is calculated as the Kullback-Leibler distance relative to the revenue weighted and time discounted mean of all past topics in the respective country. Similarity for the USE is calculated as the cosine similarity to the revenue weighted and time discounted mean of all past movie embeddings in the respective country. The yearly discount factor is 10%. Surprise and cosine similarity relative to past movies are also standardized. Standard errors double clustered at country and year level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table D3: Log(revenue) explained by distance to country's motifs

				Origin of financing		
				Foreign	non-US	US
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Residualized Jensen-Shannon distance						
5 topics	-0.1786***	-0.3993***	-0.4165***	-0.4700***	-0.5226***	-0.4355***
	(0.0196)	(0.1039)	(0.1018)	(0.1019)	(0.1501)	(0.1021)
10 topics	-0.2570***	-0.4159***	-0.4640***	-0.5175***	-0.6596***	-0.4444***
	(0.0213)	(0.1177)	(0.1194)	(0.1097)	(0.2043)	(0.1091)
30 topics	-0.0890***	-0.1269***	-0.2042***	-0.1647***	-0.2989**	-0.1212***
	(0.0220)	(0.0403)	(0.0490)	(0.0410)	(0.1069)	(0.0427)
40 topics	-0.0780***	-0.0654	-0.0922*	-0.1045*	-0.0229	-0.0921*
	(0.0207)	(0.0556)	(0.0516)	(0.0531)	(0.0915)	(0.0537)
100 topics	-0.0987***	-0.1830**	-0.2943***	-0.2744***	-0.0714	-0.2446***
	(0.0185)	(0.0744)	(0.0831)	(0.0746)	(0.2064)	(0.0781)
Panel B: Residualized cosine similarity of USE removing top X% of closest motifs						
1	0.2017***	0.1545***	0.1163**	0.1707***	0.1712	0.1432**
	(0.0207)	(0.0531)	(0.0553)	(0.0518)	(0.1596)	(0.0609)
5	0.2103***	0.1495***	0.1204**	0.1690***	0.1658	0.1443**
	(0.0212)	(0.0526)	(0.0548)	(0.0500)	(0.1716)	(0.0596)
10	0.2191***	0.1475***	0.1200**	0.1692***	0.1105	0.1504**
	(0.0215)	(0.0517)	(0.0551)	(0.0471)	(0.1770)	(0.0567)
Panel C: Residualized Jensen-Shannon distance of 20 topics removing top X% of closest motifs						
1	-0.0989***	-0.1362***	-0.1216***	-0.1739***	-0.2803***	-0.1396***
	(0.0267)	(0.0403)	(0.0426)	(0.0375)	(0.0837)	(0.0420)
5	-0.1027***	-0.1402***	-0.1165**	-0.1744***	-0.2979***	-0.1438***
	(0.0266)	(0.0442)	(0.0503)	(0.0419)	(0.0808)	(0.0455)
10	-0.1055***	-0.1717***	-0.1362**	-0.1962***	-0.3383***	-0.1647***
	(0.0268)	(0.0463)	(0.0551)	(0.0430)	(0.1077)	(0.0485)
Panel D: Residualized Bhattacharya overlap (LDA)						
20 topics	0.0982***	0.1411***	0.1181**	0.1752***	0.3085***	0.1359***
	(0.0262)	(0.0409)	(0.0458)	(0.0398)	(0.0905)	(0.0432)
Panel E: Residualized cosine similarity (LDA)						
20 topics	0.0912***	0.1442***	0.1233**	0.1744***	0.2907***	0.1370***
	(0.0278)	(0.0397)	(0.0451)	(0.0385)	(0.0904)	(0.0424)
Panel F: Non-residualized Jensen-Shannon distance (LDA)						
20 topics	-0.0703**	-0.1406***	-0.1197**	-0.1772***	-0.3063***	-0.1396***
	(0.0264)	(0.0421)	(0.0446)	(0.0405)	(0.0889)	(0.0440)
Panel G: Non-residualized cosine similarity (embeddings)						
Universal encoder	0.1839***	0.1607***	0.1197**	0.1742***	0.1249	0.1500**
	(0.0199)	(0.0549)	(0.0567)	(0.0531)	(0.1606)	(0.0616)
Panel H: Factor of Jensen-Shannon distance of 5, 10, 20, 30, 40, 100 topics						
First factor	-0.0962***	-0.1388***	-0.1180**	-0.1715***	-0.2969***	-0.1339***
	(0.0269)	(0.0397)	(0.0438)	(0.0384)	(0.0877)	(0.0420)
Panel I: Factor of JSD, cosine, Bhattarchya with 20 topics						
First factor	-0.1141***	-0.1696***	-0.1740***	-0.2246***	-0.3077***	-0.1816***
	(0.0247)	(0.0554)	(0.0588)	(0.0527)	(0.0920)	(0.0569)
Panel J: Factor of JSD, cosine, Bhattarchya of 5, 10, 20, 30, 40, 100 topics						
First factor	-0.1259***	-0.1865***	-0.1917***	-0.2429***	-0.3095***	-0.2001***
	(0.0253)	(0.0584)	(0.0631)	(0.0557)	(0.0984)	(0.0599)
Panel K: Factor of JSD, cosine, Bhattarchya of 5,10,20,30,40,100 topics & Encoder cosine similarity						
First factor	-0.1216***	-0.1838***	-0.1895***	-0.2404***	-0.3075***	-0.1978***
	(0.0253)	(0.0583)	(0.0631)	(0.0555)	(0.0982)	(0.0595)
Observations	168626	168626	97081	149942	42372	123933
R^2	0.3366	0.7838	0.8499	0.8169	0.7573	0.8377
Number of movies	13929	13929	9837	10751	8164	5580
Number of countries	85	85	84	85	71	84
Country-year-month FE	✓	✓	✓	✓	✓	✓
Movie FE		✓	✓	✓	✓	✓
Distributor FE			✓			

Notes: OLS regressions. Each cell is one separate regression and the displayed regressor is standardized with mean zero and standard deviation of one. The statistics in the bottom refer to the last row of regressions. Panel A displays the same regressions as in Panel B of Table 1 with the difference that JSD is calculated based on different number of topics from the LDA. In the first row of Panels B and C the 1%, in the second row the 5%, and in the third row the 10% closest motifs of a country to the respective movie are dropped. Panel D uses the residualized Bhattacharya overlap which between topic distributions P and Q is $BC(P,Q) = \sum_{k=1}^{20} \sqrt{P(k)Q(k)}$. Panel E uses the cosine similarity of the 20 topics from the LDA. Panel F uses the JSD without residualizing by length of documents. Panel F uses the cosine similarity of the embeddings from the USE without residualizing by length of documents. Panels H, I, J, and K are based on the first factor extracted using principal component analysis. In Panel H we extract the factor from the residualized JSD of 5, 10, 20, 30, 40, and 100 topics. In Panel I we extract the factor from the residualized JSD, cosine similarity, and Bhattarchya with 20 topics. In Panel I we extract the factor from the residualized JSD, cosine similarity, and Bhattarchya with 5, 10, 20, 30, 40, and 100 topics each. In Panel J we use the same variables as in Panel I and add the residualized cosine similarity of the embeddings. Standard errors double clustered at country and year level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

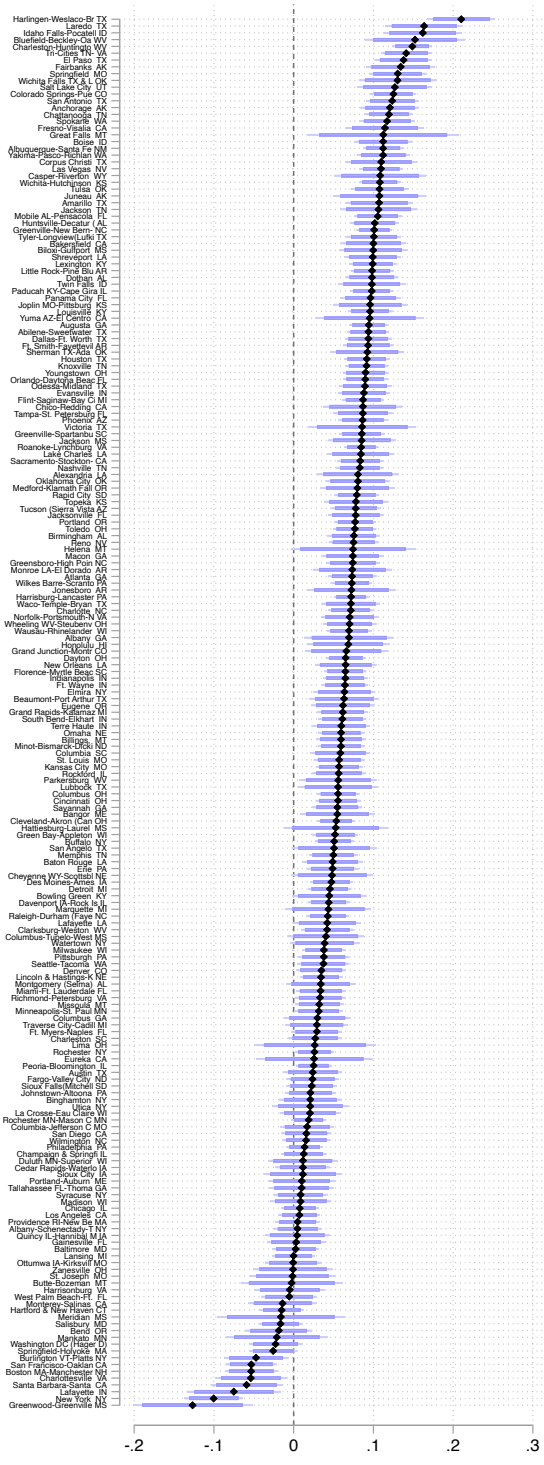
Table D4: Whether screened explained by distance to country's motifs (robustness checks)

	(1)	(2)	(3)
Panel A: Residualized cosine similarity			
Cosine similarity folklore to movie	0.0147***	0.0132***	0.0143***
	(0.0044)	(0.0033)	(0.0036)
Cosine similarity folklore to movie squared	-0.0024		
	(0.0015)		
Cosine similarity to past movies			0.0179***
			(0.0057)
Observations	2650520	2313824	2313824
R^2	0.3784	0.3933	0.3936
Number of movies	28810	28543	28543
Number of countries	92	91	91
Movie FE	✓	✓	✓
Country-year of production FE	✓	✓	✓
Panel B: Residualized Jensen-Shannon distance			
JSD folklore to movie (20 topics)	-0.0052	-0.0064	-0.0063
	(0.0045)	(0.0045)	(0.0046)
JSD folklore to movie squared (20 topics)	0.0011*		
	(0.0006)		
Surprise relative to past movies			-0.0038
			(0.0026)
Observations	2650520	2313824	2313824
R^2	0.3783	0.3932	0.3932
Number of movies	28810	28543	28543
Number of countries	92	91	91
Movie FE	✓	✓	✓
Country-year of production FE	✓	✓	✓

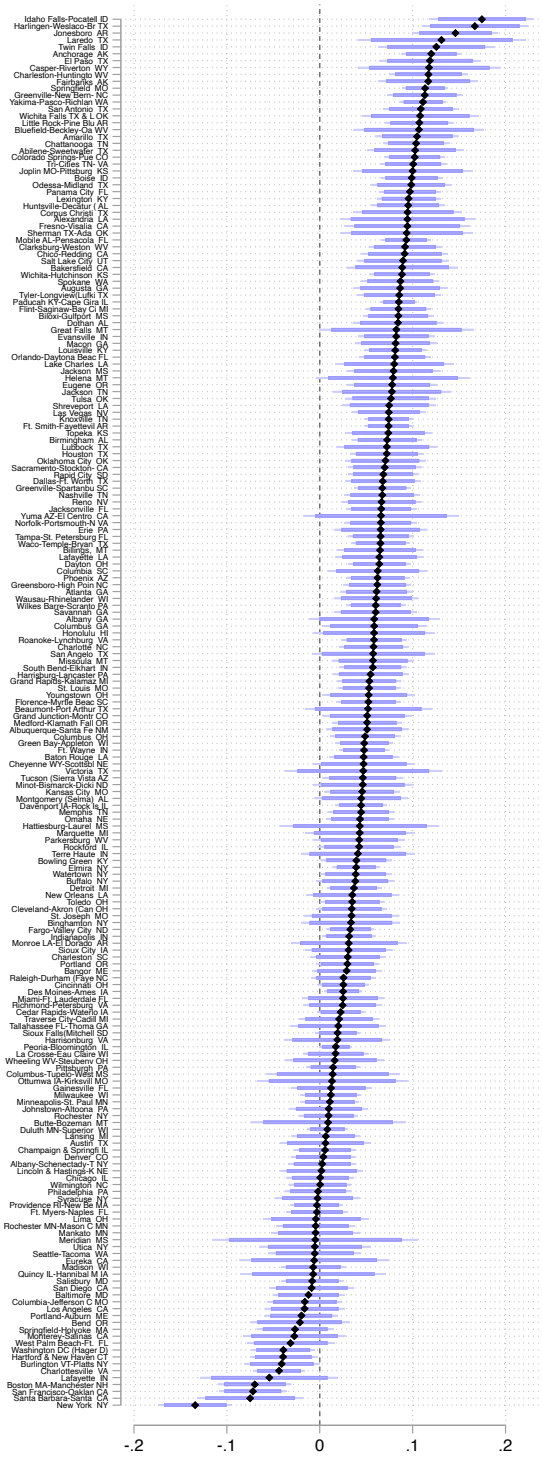
Notes: OLS regressions. JSD and cosine similarity are residualized using the log(number of words) of the movie summary, motif, and their interaction, and are standardized with mean zero and a standard deviation of one. Surprise for topics is calculated as the Kullback-Leibler distance relative to the revenue weighted and time discounted mean of all past topics in the respective country. Similarity for the USE is calculated as the cosine similarity to the revenue weighted and time discounted mean of all past movie embeddings in the respective country. The yearly discount factor is 10%. Surprise and cosine similarity relative to past movies are also standardized. The dependent variable is a binary indicator whether a movie has been screened. Standard errors double clustered at country and year level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Figure D8: Cosine similarity explaining outcomes within DMAs using Google Trends

Panel A: One-year window



Panel B: 2004 until present



Notes: Controls include year-month, language, and country of origin of financing fixed effects. Standard errors are clustered at the year level. Cosine similarity is residualized using the log(number of words) of the movie outline, motif's description, and their interaction. The sample is restricted to DMA with at least 500 observations. The thick (thin) bars represent the 95% (90%) confidence intervals.