9th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Genres, Subgenres, and Storytelling Tropes: a Data Science Approach

Keywords: cultural evolution, cultural analytics, genre, NLP, network science

Extended Abstract

What's the difference between a first-person shooter and an action-adventure video game? How are Korean dramas different from Japanese dramas? How are genres defined? Whether they are top-down impositions such as "Dark Suspenseful Gangster Dramas" (an actual genre on Netflix), or emerging musical categories such as hyperpop, genres are fundamental to how we navigate the world, and to the role that popular media plays in shaping and reflecting society.

In this study, we analyse computational formulations of genre and the differences in these 'inferred genres' through time and geography through a novel methodology incorporating methods from data science, cultural analytics, and natural language processing. Notably, here we explore the genre classification of artworks based only on crowdsourced metadata about them, i.e. based on the public perception of each artwork, rather than directly examining the raw content of each piece. In that, we show how computational studies of popular media can contribute to scholarship on genre and the public perception of art, in addition to having applications to content recommendation and content generation.

Traditionally, genres have been defined in three ways. First, by genre scholars and literary critics who look at narrative structure and symbolic patterns [1]. Second, by artists and consumers, through labels that are either self-assigned or assigned to others (arguably, Netflix or Spotify fall here). Third, by examining the content outside of its original context, whether through close reading or through machine learning. In this system, crowdsourced databases such as IMDb have emerged somewhere between the second and third category. In this paper, we investigate the extent to which some of these methods of genre definition align.

To do this, we collected data from databases of media tropes, including TVTropes and RAWG, and clustered works based on their tropes and metadata tags. We then compared these clusters with actual genre classifications provided by users. Our analysis sought to identify patterns in the use of common tropes, such as the "mad scientist" or "villain wearing a dark palette", and to study differences in media across geography and time.

Our research aims to answer the question: do the genres as defined by people match the genres emerging from the data? Through our data-driven analysis, we hope to provide new insights into the nature of genre and its role in shaping and reflecting culture.

To investigate the relationship between cultural tropes and genres, this study utilizes three main datasets. The first dataset is collected from TVTropes, a crowdsourced website that describes tropes in popular media. This wiki-format website contains a vast collection of entries describing common storytelling devices, such as the "mad scientist" or "villain wearing a dark palette" mentioned above. We also collected data from IMDb (Internet Movie Database), an online database of information related to films and other media which includes information on cast, crew, plot summaries, and most importantly for this project, genre metadata. The third dataset analyzed in this study is a corpus of over 600,000 video games from RAWG, a video game database. This dataset spans from 2011 to 2021 and contains user-annotated metadata, including tags on the game's genre and gameplay elements.

To cluster tropes and the works they feature in, we apply a combination of approaches from cultural evolution and stylometry, along with NLP methods such as topic modelling, sentence embedding, and stochastic block models. These methods allow us to identify recurring patterns in the use of tropes and the works they appear in, as well as the emergence of new subgenres. In particular, we use topSBM, a nonparametric stochastic block model-based topic model introduced by Gerlach et al. [3], to identify a hierarchy of genres and subgenres, as well as a hierarchy of storytelling tropes present in the data over time. Our analysis reveals the applicability of this method for identifying and tracking the emergence of new subgenres and tropes within popular media. Finally, in our analysis of the RAWG video game database, we construct a network of video game tags, such as "first person" and "shooter", which we use to identify emerging subgenres within the video game corpus. We identify a total of five subgenres, and study their evolution over time. By clustering works by their tropes and tags, we are able to identify recurring patterns in the use of storytelling devices within each subgenre, and track how these patterns change over time.

Figure 1 shows the evolution of TVTropes and RAWG genres over time, in the top and bottom panels respectively. The changes in topic distributions at level 1 of the TopSBM hierarchy are shown in Figure 1a for TV shows and Figure 1d for video games.

Figures 1b and 1e show the effective number of genres for both datasets. This is done by measuring Camargo et al's Effective Number of Issues (ENI) in each cluster. ENI intuitively represents the number of topics receiving attention, and captures both capacity (the number of issues at a time) and diversity (how evenly attention is spread) [2]. In practice, an effective number of genres close to k indicates that a genre cluster is mainly composed of k topics.

Finally, we compare our conclusions regarding temporal trends in media to the changes in the fraction of works annotated with IMDb and RAWG.io genres, shown in Figures 1c and 1f. Interestingly, many of the features exhibiting notable changes over time, such as the use of storytelling devices in both video games and TV shows, do not correspond to predefined genres of media, rather the deeper characteristics in the narrative. This shows the advantage of our methodology, particularly the use of nonparametric models such as TopSBM, in the analysis of cultural data, as the model can draw new inferences compared to traditional means.

Our analysis highlights a discrepancy between official genre classifications in IMDb and RAWG and genres emerging from the data. Our findings suggest that official genre classifications are often inadequate for capturing the full range of subgenres that exist within popular media, and that a data-driven approach can provide a more accurate and nuanced understanding of the landscape of cultural genres.

Overall, our results from TopSBM have enabled us to determine changes in popular media over time, including the increase in narrative complexity for both video games and TV shows. Our study also reveals some limitations of data science and crowdsourced databases for stylometry and cultural evolution. These limitations include issues of bias and representativeness in the data, as well as the challenge of interpreting the results of automated analyses without human input and expertise. Overall, our results demonstrate the power and potential of data science methods for understanding the evolution of cultural genres, while also highlighting the need for careful interpretation and critical reflection on the limitations of these methods.

References

[1] Roland Barthes. Mythologies: The complete edition. *New York: Hill and Wang*, 2012.

[2] Chico Q Camargo, Peter John, Helen Z Margetts, and Scott A Hale. Measuring the volatility of the political agenda in public opinion and news media. *Public Opinion Quarterly*, 85(2):493–516, 2021.

[3] Martin Gerlach, Tiago P Peixoto, and Eduardo G Altmann. A network approach to topic models. *Science advances*, 4(7):eaaq1360, 2018.

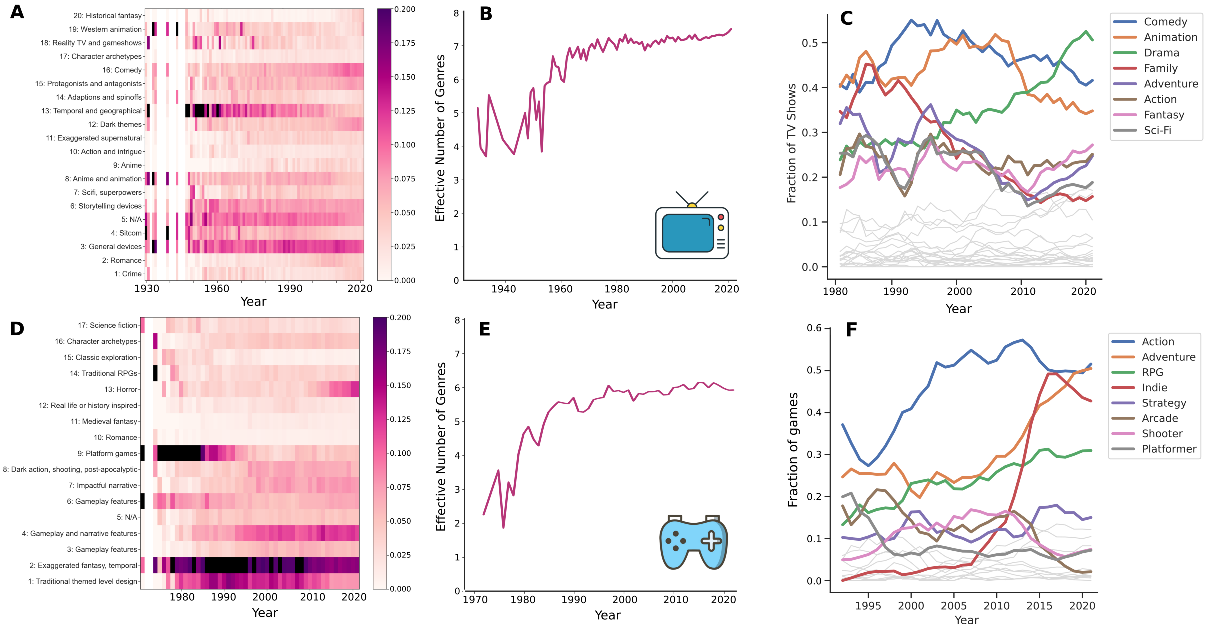

Figure 1: Genres over time, for TV (top) and video games (bottom). Panels (a) and (d) show heatmaps indicating topic frequency over time, panels (b) and (e) show the effective number of issues over time, and panels (c) and (f) show changes the frequencies of IMDb and RAWG genres.