Analyzing Movie Script Similarity: A Weighted Multilayer Network Approach

Majda Lafhel, Hocine Cherifi, Mohammed El Hassouni, Benjamin Renoust

FRCCS

Analyzing Movie Script Similarity: A Weighted Multilayer Network Approach

Majda Lafhel[1✓], Hocine Cherifi[2], Mohammed El Hassouni[1], and Benjamin Renoust[3]

[1] *Mohammed V University in Rabat, Morocco; majdalafhel1@gmail.com*
[2] *ICB UMR 6303 CNRS - Univ. Bourgogne - Franche-Comté, Dijon, France;*
[3] *Median Technologies, France;*

[✓] *Presenting author*

Abstract. Measuring similarity between multilayer networks is a challenging task that involves capturing various layers and relationships using distance measures. Existing techniques have limitations in comparing layers with the same number of nodes and ignoring inter-relationships. In this research, we propose a novel approach for measuring similarity between multilayer networks while considering inter-relationships and weighted networks of various sizes. Our approach is applied to multilayer movie networks, consisting of layers of different entities (character, keyword, and location) and inter-relationships between them. Our proposed method captures both intra-layer and inter-layer relationships, providing a comprehensive overview of the multilayer network. This approach has potential applications in analyzing movie story structures and social network analysis. Our results demonstrate that our approach improves the accuracy of similarity measurement between multilayer networks.

Keywords. *Movie Similarity; Multilayer network; Network Distance*

1 Introduction

Over the last few years, the analysis of movies has become a challenging issue in complex network analysis. The process involves identifying different components of a movie story, converting them into networks, and then selecting appropriate approaches to measure and evaluate these networks. Markovic et al. (2019) [13] constructed the character network of Slovene belles-lettres by analyzing character interaction structures. They indexed sentences containing main characters and computed distances between characters based on frequency of occurrence. Lv et al. (2018) [12] introduced StoryRoleNet, utilizing video and subtitle analysis to identify main characters in movies and construct character networks. Chen et al. (2019) [4] proposed a method using minimum span clustering on community structures and centrality to determine character distances in the novel Dream of the Red Chamber. These studies aim to quantify character relationships and identify critical characters within narratives. However, character interactions are not enough to analyze the story. To address this issue, Mourchid et al. [15] proposed a multilayer network model to capture various aspects of the movie story. Their model includes characters, keywords, and locations.

Although multilayer network analysis has gained significant attention in recent years, comparing multilayer networks is challenging. Indeed, there are countless entities and relationships to consider. Brodka et al. (2018) [3] proposed a property matrix that represents a multiplex network. This matrix maps layers and nodes into structures. To compare multiplex networks, the authors used three methods: aggregations (min, max, entropy), layer distributions (Jensen-Shannon Divergence), and similarity functions (Jaccard, cosine, correlation). Giordano et al. (2019) [7] used factorial methods for quantifying visually multiplex networks. Ghawi et al. (2022) [6] used community detection to quantify the similarity between multilayer networks.

In previous studies [10, 8, 9], we measured the similarity between movie stories by analyzing various components and computing distances between their corresponding movie networks using graph measures. We focused on ordering and classifying movies by genre. Initially, we used a multilayer network model [14] to represent story elements. However, we only compared monolayers of the same entities. In a separate work [11], we compare multilayer networks considering intra and inter-relationships. The current work aims to enhance the methodology by incorporating weighted network properties.

2 Methodology

In a previous work [11], we presented an algorithm that maps intra-layers $\mathcal{G}_{intra}$ and inter-layers $\mathcal{G}_{inter}$ into one matrix $\mathcal{P}$. We now incorporate properties related to weighted networks.

We convert unweighted movie networks into weighted ones by assigning a weight, denoted w_{ij}, to an edge based on the frequency of connections between two nodes i and j. In other words, we transform character, keyword, location networks, and their interactions into weighted networks as follows: (1) For intralayers ($\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, and $\mathcal{G}_{intra_{LL}}$), we calculate the weight of the edge connecting two nodes i and j of the same entity based on the number of times the connection occurs between them. (2) For interlayer ($\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, and $\mathcal{G}_{inter_{CL}}$), we calculate the weight of the edge connecting two nodes i and j of different entities based on the number of times the connection occurs between them. Conversely, we use incidence and Laplacian matrices to extract weighted network properties via the following factorization (eq. 1).

$$\mathcal{M} = I_{incid}^{T} \times L \times I_{incid} \tag{1}$$

We then integrate the eigenvalues of the output M into the property matrix $\mathcal{P}$ by adding a new column. The property matrix $\mathcal{P}$ is then defined as follows:

- (1) Six rows, where the first three rows represent the three intralayers ($\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, and $\mathcal{G}_{intra_{LL}}$), and the last three rows represent the three interlayers ($\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, and $\mathcal{G}_{inter_{CL}}$).
- (2) Seven columns, each one representing a network feature: max degree, max centrality, density, adjacency, Laplacian, network portrait, and Incidence.
- (3) Each cell c_{ij} encodes a network feature j of the network type i.

Algorithm 1 Matrix property extraction

input: $\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, $\mathcal{G}_{intra_{LL}}$, $\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, $\mathcal{G}_{inter_{CL}}$
output: matrix property $\mathcal{P}$

for i **in** $\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, $\mathcal{G}_{intra_{LL}}$, $\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, $\mathcal{G}_{inter_{CL}}$ **do**
  $\mathcal{D} \leftarrow \max((\deg(1), \deg(2), ..., \deg(\mathcal{N})))$ //return the max node degree of i.
  $\mathcal{BC} \leftarrow \max(\sum_{s \neq t \in V} \frac{\sigma_{st}(\mathcal{N})}{\sigma_{st}})$ //return the max node betweenness centrality of i
  //σ_{st}: total shortest paths passing from a node s to a node t
  //$\sigma_{st}(\mathcal{N})$: total number of σ_{st} that passing through a node n
  $Dens \leftarrow \mathcal{E}/(\mathcal{N} * (\mathcal{N} - 1))$ //return density of i
  $\mathcal{A} \leftarrow Extract\_Adjacency\_matrix(i)$
  $\mathcal{L} \leftarrow Extract\_Laplacian\_matrix(i)$
  $\mathcal{B} \leftarrow Extract\_NetworkPortrait\_matrix(i)$
  $\mathcal{M} = I_{incid} \times L \times I_{incid}^{T}$
  $s_A \leftarrow sum(eigenvalues(\mathcal{A}))$
  $s_L \leftarrow sum(eigenvalues(\mathcal{L}))$
  $s_B \leftarrow sum(\mathcal{B})$
  $s_M \leftarrow sum(eigenvalues(\mathcal{M})$
  $v_i \leftarrow [\mathcal{D}, \mathcal{BC}, Dens, s_{\mathcal{A}}, s_{\mathcal{L}}, s_{\mathcal{B}}, s_{\mathcal{M}}]$
end for
$\mathcal{P} \leftarrow [v_{\mathcal{G}_{intra_{CC}}}, v_{\mathcal{G}_{intra_{KK}}}, v_{\mathcal{G}_{intra_{LL}}}, v_{\mathcal{G}_{inter_{CK}}}, v_{\mathcal{G}_{inter_{KL}}}, v_{\mathcal{G}_{inter_{CL}}}]$

3 Experimental Results

We conduct experiments using three popular romance movie scripts: Titanic, Twilight Episode 1, and Episode 2. For each movie, we extract three weighted layers (character, keyword, and location) along with their intra- and inter-relationships. Individuals ranked the similarities between the romance movies to build the ground truth. Based on their evaluation, we obtained the following ranking (Majority rule): Episodes 1 and 2 of Twilight rank first, Titanic and Episode 1 of Twilight are second, as well as Twilight and Episodes 2. Table 1 displays the results using weighted and unweighted networks. Weighted networks appear to give better results according to the ground truth data. Indeed, the distance between Twilight Episode 1 and 2 (43,705.53) is smaller than the distance between Twilight Episode 1 and Titanic (725,451.89) or between Twilight Episode 2 and Titanic (683,244.27). In contrast to unweighted networks, which show high similarity between episode 1 of Twilight and Titanic, Future work will concentrate on introducing perceptual features in the model [1, 2] as well as more sophisticated network structural properties[5, 17, 16, 18].

Table 1 Distance between romance multilayer networks for weighted and unweighted networks. Values are normalized by dividing each value by the maximum value in its corresponding column.

Romance movies	Distance	
	Unweighted networks	Weighted networks
Twilight (Episode 1) & Twilight (Episode 2)	11 997.74 (0.96)	**43 705.53 (0.06)**
Twilight (Episode 1) & Titanic	**8 731.8 (0.7)**	725 451.89 (1)
Twilight (Episode 2) & Titanic	12 405.22 (1)	683 244.27 (0.94)

References

[1] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. A curvature based method for blind mesh visual quality assessment using a general regression neural network. In *IEEE SITIS*, pages 793–797. IEEE, 2016.

[2] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. A convolutional neural network framework for blind mesh visual quality assessment. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 755–759. IEEE, 2017.

[3] Piotr Bródka, Anna Chmiel, Matteo Magnani, and Giancarlo Ragozini. Quantifying layer similarity in multiplex networks: a systematic study. *Royal Society op. science*, 5(8), 2018.

[4] RH-G Chen, C-C Chen, and C-M Chen. Unsupervised cluster analyses of character networks in fiction: Community structure and centrality. *Knowledge-Based Systems*, 2019.

[5] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Extracting backbones in weighted modular complex networks. *Scientific Reports*, 10(1), 2020.

[6] Raji Ghawi and Juergen Pfeffer. A community matching based approach to measuring layer similarity in multilayer networks. *Social Networks*, 68:1–14, 2022.

[7] Giuseppe Giordano, Giancarlo Ragozini, and Maria Prosperina Vitale. Analyzing multiplex networks using factorial methods. *Social Networks*, 59:154–170, 2019.

[8] Majda Lafhel, Lylia Abrouk, Hocine Cherifi, and Mohammed El Hassouni. The similarity between movie scripts using multilayer network laplacian spectra descriptor. In *2022 IEEE Workshop on Complexity in Engineering (COMPENG)*, pages 1–5. IEEE, 2022.

[9] Majda Lafhel, Hocine Cherifi, Benjamin Renoust, and Mohammed El Hassouni. Comparison of graph distance measures for movie similarity using a multilayer network model. *Entropy*, 26(2):149, 2024.

[10] Majda Lafhel, Hocine Cherifi, Benjamin Renoust, Mohammed El Hassouni, and Youssef Mourchid. Movie script similarity using multilayer network portrait divergence. In *Complex Networks & Their Applications IX: Volume 1*, pages 284–295. Springer, 2021.

[11] Majda Lafhel, Mohammed El Hassouni, Benjamin Renoust, and Hocine Cherifi. Measuring movie script similarity using characters, keywords, locations, and interactions. In *French Regional Conference on Complex Systems*, 2023.

[12] Jinna Lv, Bin Wu, Lili Zhou, and Han Wang. Storyrolenet: Social network construction of role relationship in video. *IEEE Access*, 6:25958–25969, 2018.

[13] Rene Markovič, Marko Gosak, Matjaž Perc, Marko Marhl, and Vladimir Grubelnik. Applying network theory to fables: complexity in slovene belles-lettres for different age groups. *Journal of Complex Networks*, 7(1):114–127, 2019.

[14] Youssef Mourchid, Benjamin Renoust, Hocine Cherifi, and Mohammed El Hassouni. Multilayer network model of movie script. In *Complex Networks and Their Applications VII: Volume 1 COMPLEX NETWORKS 2018 7*, pages 782–796. Springer, 2019.

[15] Youssef Mourchid, Benjamin Renoust, Olivier Roupin, Lê Văn, Hocine Cherifi, and Mohammed El Hassouni. Movienet: a movie multilayer network model using visual and textual semantic cues. *Applied Network Science*, 4:1–37, 2019.

[16] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. Complex network and source inspired covid-19 fake news classification on twitter. *IEEE Access*, 9:139636–139656, 2021.

[17] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Scientific reports*, 11(1):10088, 2021.

[18] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Comparative evaluation of community-aware centrality measures. *Quality & Quantity*, 57(2), 2023.