Introduction to Social Networks

Assignment 2

Siyu Jiang, Roman Kracht, Zhengting He, Seokwon Choi

Task 1: Network Hypotheses

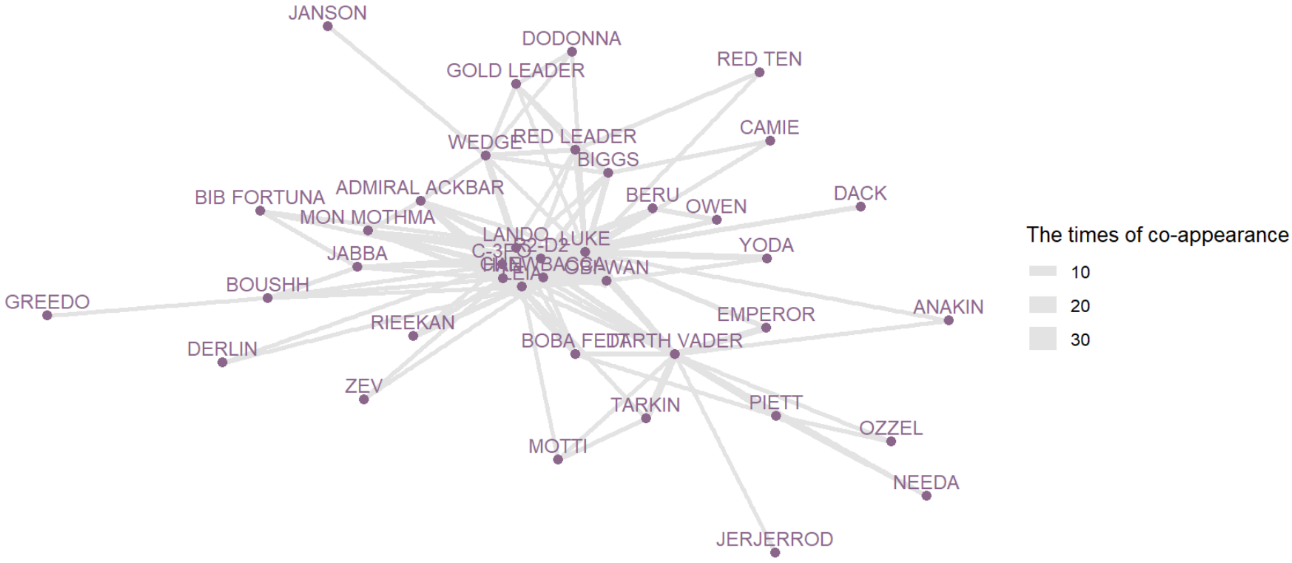
(a) network visualization and description

For our analysis, we chose the “Star Wars Character co-emergence” dataset from the networkdata package[1], mostly because all members of our homework group are big fans of the franchise. Another reason for our choice was that individual node data provided by the star wars dataset is very rich since many characteristics of the nodes are captured.

We first chose especially the dataset for episode V - «The Empire Strikes Back» because we thought that this episode has the distinct sides of the characters well balanced. In addition, we merged the three episodes of the Original Trilogy IV - VI in order to enhance our network with overlapping nodes along the episodes.

In the dataset, characters that occur in Star Wars episodes IV to VI are represented by 39 nodes with various attributes (e.g., home world, weight, etc.). These nodes are connected by 170 undirected weighted edges, indicating whether the characters appear in movie scenes together. The weight of the edge is given by the number of scenes in which they appear together. The complete network is shown in Figure 1. The network density is 22.94%.

Figure 1 The co-appearance network of characters in Star Wars IV-VI episodes



Firstly, we notice that the network is relatively small compared to others we discussed in the lecture so far (e.g., the needle-sharing example). This difference is to be expected, given that the number of main characters in movies has to be limited to keep the plot

[1] https://github.com/schochastics/networkdata

understandable. Snowball-sampled social network studies, however, often rely on having a large sample size from which empirical findings can be derived. Moreover, the Star Wars networks are generated by direct observational data from the movie scenes rather than self-reported data as discussed in class.

Another interesting observation is that all the nodes in the Star Wars network lie in one connected component, and in particular, there are no isolated nodes. This also stands in contrast to the needle-sharing example that we have seen in the lecture. From a story-telling perspective, isolated nodes would correspond to characters that never appear in scenes together with other characters, thus making them irrelevant to the movie's plot. Furthermore, having separate connected components in the network would mean that the film has multiple plots that never interfere with each other. It might be the case for some movies, but not for the Star Wars franchise.

Of course, these differences to real-world networks that we have studied before arise from the fact that movies have to be entertaining for the audience in the first place.

(b) research hypotheses

The first network theory we chose is **homophily**, which states that individuals with similar characteristics or attributes tend to form social relationships more than others. The Star Wars character interactions network provides 10 attributes, including their physical characteristics and background information, and we hypothesize that *"Characters from the same home world are more likely to appear together in the same scenes (i.e., be connected and have a higher weight in the network given)."* The node weight represents the number of scenes that the two linked characters appear together. Therefore, our hypothesis states that characters from the same home world tend to appear together in the scenes. This hypothesis is reasonable because they have a higher chance of living within a short distance and staying together in the movies. We think this hypothesis would show a higher accuracy, particularly in Episode IV, since the characters didn't travel far from their home world in that episode.

Our second hypothesis states that *"Characters with higher betweenness centrality tend to be the main characters,"* meaning they appear more often in the scenes. The second hypothesis is related to the **structural hole** theory, proposing that brokers in a network gain more advantages, such as valuable information. Structural holes are indicated by high betweenness centrality, meaning that they are in a bridging and brokering position. We have thought that the advantage for the film's characters is becoming the main characters, which can be quantified by their screen time. Since this network gives data only about the interactive scenes, each character's appearance time would need to be additionally collected to test this hypothesis.

Introduction to Social Networks

Assignment 2

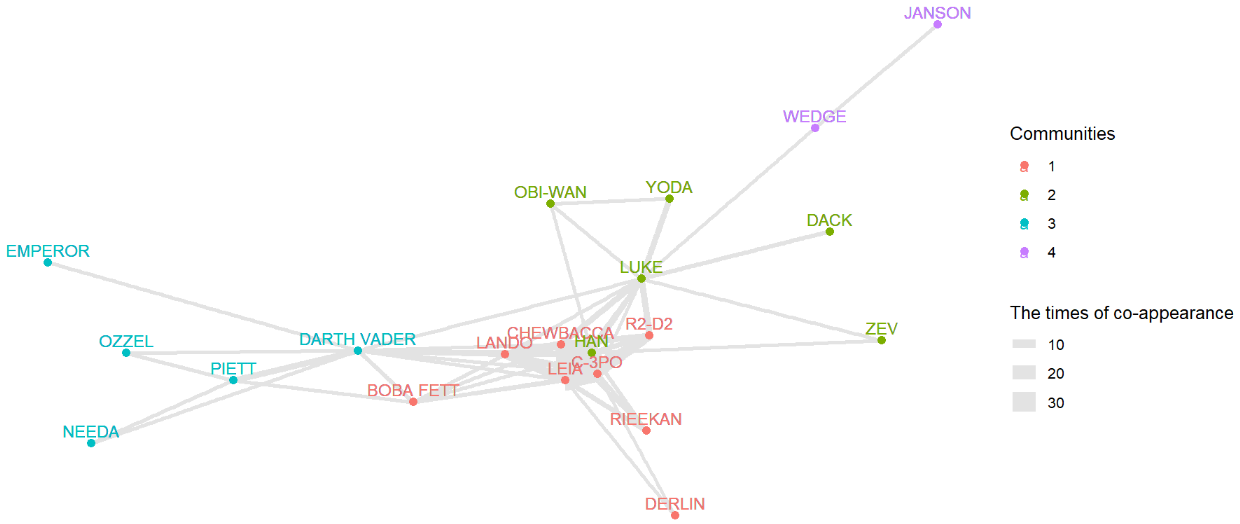
For both hypotheses, we can merge the networks from various Star Wars episodes into one network to enlarge the network data. Specifically, (as we did in task 1a) combining only the data of the original trilogy (episodes IV to VI) would be reasonable since the prequel trilogy (episodes I to III) takes place in a totally different era. But this may lead to a more inaccurate result because some characters play a central role during the whole series while others mainly appear in one or two episodes. This imbalance may occur as an unreliable result of hypothesis testing. Therefore, regarding task 2a in the next section, we only use the network data of Star Wars V.

Task 2: Data Analysis

(a) community detection

We chose the edge-betweenness algorithm to detect the potential communities in the movie Star Wars: Episode V — The Empire Strikes Back. The plot of fictional movies often involves the protagonists (e.g., the Jedi) fighting against the antagonists (e.g., the Sith). Characters from different sides usually do not often appear in the same scenes. Instead, they show up more frequently with their group peers. Because the mechanism of the edge-betweenness algorithm is to assess and diminish the edge with the highest edge-betweenness one by one to decide the partition with the highest modularity finally, this algorithm is more advantageous in finding out the movie characters that best link the distinct groups (i.e., communities). Considering the meaning of the network ties in our data set (i.e., characters' co-appearance), the most important characters from different communities may be the significant "bridges" between communities; their ties can be detected by the algorithm and removed earlier to differentiate the distinct communities. Therefore, we chose the edge-betweenness algorithm, and the result indicates *four communities* in the data set (see Figure 2). The modularity score is 0.0579.

Figure 2 The community detection in Star Wars V by Edge-Betweenness algorithm



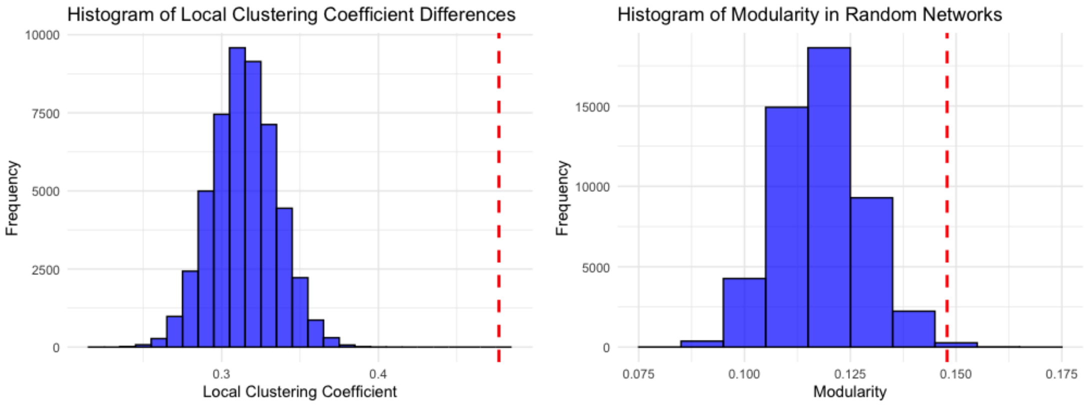
Introduction to Social Networks

Assignment 2

Based on the knowledge of the movie fans from our group, we could roughly suggest that community 4 in purple is the rebellion pilots, community 3 in blue is the dark side — the Galactic Empire with Darth Vader as the central figure, community 2 in green is the good side — Rebel Alliance, and community 1 in red mainly includes some characters from Rebel Alliance (e.g., Princess Leia), robots and spies. Community 1 is mixed on all sides but can be explained by the relationship between the spies (Boba Fett and Lando Calrissian) and their victims (Leia, Chewbacca, and C3-PO). This is why the employer of the spies, Darth Vader, also has many connections to the characters in community 1.

(b) hypothesis testing - conditional uniform graph tests

Figure 3 The plots of conditional uniform graph tests



The p-values calculated for the local clustering coefficient (p < 0.001, see the distribution on the left of Figure 3) and modularity (p = 0.00264, see the distribution in the right of Figure 3) suggest that the co-appearance network of characters in Star Wars episodes IV-VI is significantly different from random Erdős-Rényi networks.

A p-value close to 0 for the local clustering coefficient implies that none of the generated random networks had a local clustering coefficient as high as the original network. This indicates that the Star Wars network has a much stronger tendency for characters to form tightly-knit groups or communities than random networks. This result aligns with our understanding of the Star Wars storylines, centered around a small group of core characters (e.g., Luke Skywalker, Princess Leia, and Han Solo) who frequently interact with each other and share many common connections.

The modularity p-value of 0.00264 also suggests a significantly higher community structure in the Star Wars network than in random networks. The high modularity indicates that the network is organized into multiple communities where characters within a community interact more frequently with each other than with characters in other communities. This is consistent with the narrative structure of Star Wars movies, where groups of characters often have their own storylines that are distinct from other groups (e.g., the Rebel Alliance, the Galactic Empire, and the Jedi Order).

Introduction to Social Networks

Assignment 2

In conclusion, the local clustering coefficient and modularity metrics demonstrate that the Star Wars episodes IV-VI co-appearance network exhibits a stronger community structure than expected in random networks. This finding reflects the nature of the Star Wars storyline, which revolves around distinct groups of characters who often interact closely within their own communities.