Adithya Raghu Ganesh
200314659

CSC555-Assignment P1

Task1:

Hypothesis 1: *Characters can be characterized as on the light side or the dark side by comparing their homophily to the subnetwork they belong to.*

Metrics:
I approached the given hypothesis in two different ways.

1. Analysis was made based on the **degree centrality** of the characters in each episode and the **weight** of the links between the characters. After research about the characters, a complete list of characters that fall on the light side and the dark were generated.

   First analysis involved the selection of 5 top light side characters and 5 top dark side characters to determine how every other character in the network interacts with these top 5 from each side. Five instead of one character was chosen as the data gets insufficient and unreliable when just considering a single character.

   After generating the number of interactions each character had with the ten characters mentioned above, I decided to take the top ten character interactions with each side. In doing so, I noticed a slight discrepancy in the data as there may be some main characters in the light side that overpower the interactions of the characters in the dark side.

2. This made me use a new metric to calculate the homophily of each subnetwork.

   I generated an exhaustive list of all the characters that appear in the star-wars universe and annotated them based on which side they belonged to. I used **average cohesion** of each character as a metric to get the interaction of each character with the dark side and the light side. Since there exists an uneven split between the characters on each side, I felt they could be averaged out to determine a score for the light and dark side.

   Now, depending on which score is higher, the characters are placed in their respective subnetworks and the total number of characters in their correct network was determined.

   This metric proved to support the hypothesis that the interactions each character has with each other determines which side they belong to.

Hypothesis 2: *The character of a series, developed to be the main character will have the highest degree centrality, however the most integral character of a series will have the highest betweenness.*

Metrics:
As per the hypothesis given, two main metrics have been considered - **degree centrality and betweenness**.

1. The degree centrality has been calculated by calculating the **out-degree** for each character. Since the main characters in the series tend to change through time I felt the necessity to determine a list of main characters for each episode.
2. The **betweenness centrality** of the characters can be determined by generating the interactions graph for each network and leveraging the networkx library in Python to compute the betweenness centrality of each character**.**

Task 2:

Hypothesis Analysis:

Hypothesis 1:

1. The first set of metrics involved the generation of two sets of characters - one for the light side and one for the dark.
   When calculating the top interactions of each character with each set mentioned above, the following results were obtained.

```
[[[('PADME', 81), L
   ('OBI-WAN', 65), L
   ('QUI-GON', 44), L
   ('ANAKIN', 24), L
   ('JAR JAR', 21), L
   ('YODA', 16), L
   ('CAPTAIN PANAKA', 14), L
   ('R2-D2', 11), L
   ('MACE WINDU', 6), L
   ('SHMI', 6)], L
  [('NUTE GUNRAY', 24), D
   ('EMPEROR', 23), D
   ('RUNE', 11), D
   ('DARTH MAUL', 6), D
   ('SIO BIBBLE', 3), L
   ('CAPTAIN PANAKA', 3), L
   ('PADME', 2), L
   ('GENERAL CEEL', 2), D
   ('JAR JAR', 2), L
   ('VALORUM', 2)]], D
 [[('OBI-WAN', 51), L
   ('PADME', 50), L
   ('ANAKIN', 30), L
```

```
   ('YODA', 24), L
   ('MACE WINDU', 14), L
   ('COUNT DOOKU', 7), D
   ('R2-D2', 7), L
   ('LAMA SU', 6), L
   ('C-3PO', 5), L
   ('TAUN WE', 5)], L
  [('EMPEROR', 15), D
   ('NUTE GUNRAY', 8), D
   ('COUNT DOOKU', 2), D
   ('POGGLE', 2), D
   ('SUN RIT', 2), D
   ('JAR JAR', 2), L
   ('OBI-WAN', 2), L
   ('BAIL ORGANA', 2), L
   ('YODA', 2), L
   ('MACE WINDU', 2)]], L
 [[('OBI-WAN', 82), L
   ('ANAKIN', 42), L
   ('PADME', 40), L
   ('YODA', 31), L
   ('BAIL ORGANA', 18), L
   ('R2-D2', 17), L
   ('C-3PO', 9), L
   ('EMPEROR', 8), D
   ('MACE WINDU', 7), L
   ('CLONE COMMANDER CODY', 3)], L
  [('EMPEROR', 34), D
   ('ANAKIN', 14), L
   ('PADME', 4), L
   ('OBI-WAN', 4), L
   ('BAIL ORGANA', 3), L
   ('NUTE GUNRAY', 2), D
   ('C-3PO', 2), L
   ('R2-D2', 2), L
   ('DARTH VADER', 1), D
   ('MACE WINDU', 1)]], L
```

In the above block **L** depicts Light side and **D** depicts dark side. But as one can see, Anakin's interaction for instance in Episode 3 shows how a main character tends to have more interactions with the Emperor than other characters in the dark side. This led me to fine-tune the analysis to remove such biases.

2. I created a characters.json file of the following format where 1 depicts the character actually belongs to the light side and 2 if the character belonged to the dark after some research.

```
{
        "R2-D2": "1",
        "CHEWBACCA": "1",
        "QUI-GON": "1",
        "NUTE GUNRAY": "2",
        "YODA": "1",
        "RABE": "2",
        "BAIL ORGANA": "1",
}
```

Ref: https://swgoh.gg/characters/f/Dark%20Side/ & https://starwars.fandom.com/wiki/Main_Page

There were a total of 112 characters that appeared and amongst them 81 belonged to the light side and 31 belonged to the dark. This was proof for the uneven nature of the sides split and the need for more robust analysis. I thus decided to take the average and obtain a single interaction cohesion score for each side.

```
[(23, 3),
 (23, 0),
 (18, 1),
 (17, 0),
 (15, 1),
 (14, 1),
 (20, 3)]
```

The above list of sets depicts the characters that belonged in the right subnetwork and the ones in the wrong subnetwork. Even after taking into consideration some human error that may have come up while classification, it can be seen that more than 90% of the characters have a higher interaction with their corresponding side. Such an outcome strongly suggests that the above hypothesis is in fact true for all the episodes in the series.

Series as a whole

When we consider the series as a whole and separate the characters according to which side they have a higher affinity to, we notice that 73 out of the 80 characters were correctly classified giving an accuracy for homophily in subnetworks to be 91.25%.

This brings light to some new evidence that ***“lesser individuals change their stance on which side they belong to through time”***

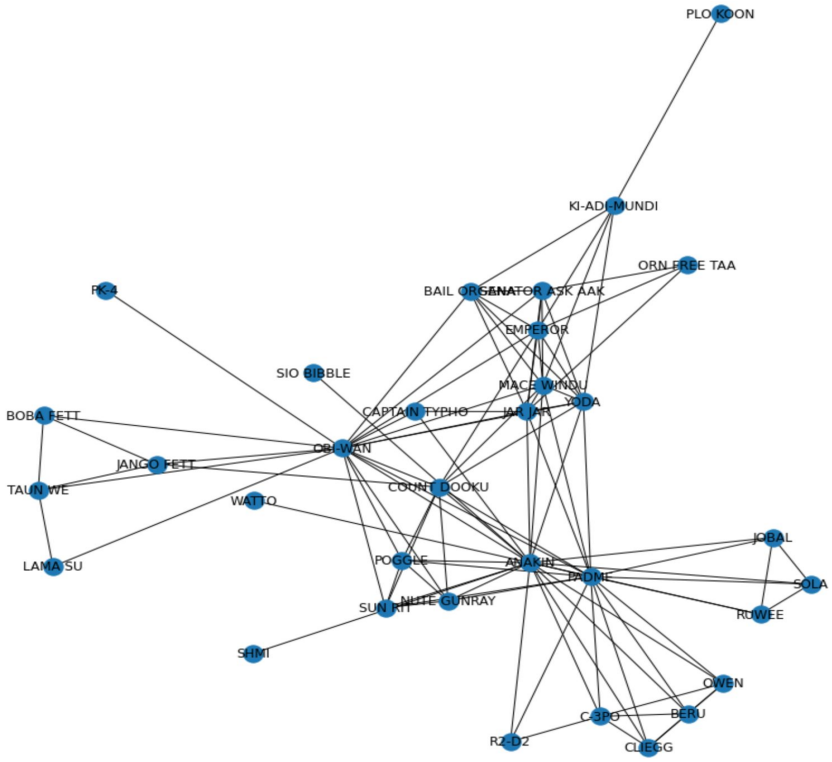
True - The hypothesis holds true

Adithya Raghu Ganesh
200314659

Hypothesis 2:

1. **Degree centrality**
   The degree centrality of the nodes in a network have been calculated by creating a hashmap of the characters and incrementing the value of the map each time a new link for a character is present.

2. **Betweenness centrality**
   In order to calculate the betweenness centrality, I first generated a graph for all the characters along with their links. A sample graph for episode 3 has been displayed below.



After the generation of such a graph, it is passed to the betweenness_centrality library in python to get the value for the integral characters in each movie. Given below are the top five characters based on degree centrality and betweenness centrality for each episode. Since the bottom five characters aren't essential, they have been included only in the jupyter notebook and python scripts present in the github repository.

```
Degree Centrality
[[('QUI-GON', 26), ('ANAKIN', 23), ('JAR JAR', 18), ('PADME', 18), ('OBI-WAN', 13)],
[('ANAKIN', 21), ('OBI-WAN', 18), ('PADME', 17), ('COUNT DOOKU', 10), ('JAR JAR', 10)],
[('ANAKIN', 14), ('OBI-WAN', 13), ('BAIL ORGANA', 12), ('EMPEROR', 11), ('PADME', 10)],
[('LUKE', 15), ('LEIA', 12), ('C-3PO', 10), ('R2-D2', 9), ('HAN', 8)],
[('LUKE', 12), ('DARTH VADER', 11), ('HAN', 11), ('LEIA', 10), ('C-3PO', 10)],
[('LUKE', 15), ('C-3PO', 11), ('LEIA', 9), ('HAN', 9), ('LANDO', 9)],
[('POE', 16), ('HAN', 14), ('FINN', 14), ('BB-8', 12), ('CHEWBACCA', 12)]]
```

```
Betweenness
[[('QUI-GON', 0.32), ('ANAKIN', 0.16), ('JAR JAR', 0.14), ('NUTE GUNRAY', 0.12), ('PADME', 0.10)],
[('ANAKIN', 0.38), ('OBI-WAN', 0.31), ('PADME', 0.17), ('KI-ADI-MUNDI', 0.06), ('JAR JAR', 0.06)],
[('OBI-WAN', 0.18), ('ANAKIN', 0.18), ('EMPEROR', 0.18), ('BAIL ORGANA', 0.18), ('YODA', 0.11)],
[('LUKE', 0.32), ('LEIA', 0.23), ('HAN', 0.17), ('BIGGS', 0.04), ('C-3PO', 0.04)],
[('LUKE', 0.41), ('DARTH VADER', 0.32), ('HAN', 0.11), ('WEDGE', 0.09), ('C-3PO', 0.07)],
[('LUKE', 0.59), ('DARTH VADER', 0.20), ('C-3PO', 0.06), ('LANDO', 0.05), ('ADMIRAL ACKBAR', 0.04)],
[('POE', 0.29), ('KYLO REN', 0.22), ('REY', 0.11), ('BB-8', 0.08), ('HAN', 0.08)]]
```

It can be seen that the main character in all the series is actually even the most integral character in most cases. However, when a series tends to have multiple main characters and when we look at the top 5 main characters in each episode, we see that there are a few discrepancies. One important thing to note in episode 6 is that Darth Vader is not at all present in the top 5 main characters but when we look at the betweenness centrality for the character, we notice that he in fact was integral in that episode. Another example to back this claim is present in Episode 7 where Han is portrayed as a main character but his “integralness” in the same seems to me low.

We can even claim that ***“through the series, the main character's integrity or importance decreases as more and more plots are created and newer characters come into the picture”.***

Series as a whole

The following are the degree centrality and betweenness values for the characters when combining the entire star-wars series as one.

```
{
"Degree Centrality" :
[('ANAKIN', 42), ('OBI-WAN', 37), ('C-3PO', 36), ('PADME', 34), ('LUKE', 27)],
"Betweenness" :
[('OBI-WAN', 0.20), ('C-3PO', 0.16), ('ANAKIN', 0.15), ('LUKE', 0.14), ('HAN', 0.09)]
}
```

When we look at the entire star-wars universe instead of episode by episode, we see a bigger difference between who the main character was and who the integral character was. Padme - who seemed to be a main character through the initial few episodes doesn’t seem to play an integral part in the movie and **Han** who has had an impact in most of the episodes as an integral character connecting multiple story lines seems to not be part of the main character list that has been displayed.

We can thus say that ***even though the main character “may” seem to be the integral character when looking at each episode in isolation, we notice the same is not true when looking at the entire star-wars universe as a whole*** **- the hypothesis holds true**

Adithya Raghu Ganesh
200314659

Task 3

1. Compute the cliquishness and characteristic path lengths of the social graph

   Cliquishness in other words **clustering coefficient** for a node may be defined as the the ratio between the number of triangles the node is part of and the product of degree(node) and degree(node) - 1

$$c_u = \frac{2T(u)}{deg(u)(deg(u)-1)}$$

Ref: https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.cluster.clustering.html

We leverage the networkx library on the braphs previously generated for betweenness calculation to determine what the chiquishness value is.

```
Cliquishness for Episode 7:
{
'LUKE': 0,
'R2-D2': 1.0,
'CHEWBACCA': 0.56,
'BB-8': 0.46,
'LOR SAN TEKKA': 0.66,
'POE': 0.39,
'KYLO REN': 0.4,
'CAPTAIN PHASMA': 0.61,
'FINN': 0.54,
'UNKAR PLUTT': 1.0,
'REY': 0.48,
'GENERAL HUX': 0.3,
'LIEUTENANT MITAKA': 1.0,
'HAN': 0.54,
'BALA-TIK': 1.0,
'SNOKE': 1.0,
'MAZ': 1.0,
'C-3PO': 0.71,
'LEIA': 0.75,
'SNAP': 0.78,
'ADMIRAL ACKBAR': 1.0,
'ADMIRAL STATURA': 1.0,
'YOLO ZIFF': 1.0,
'COLONEL DATOO': 0,
'ELLO ASTY': 1.0,
'JESS': 0.7,
'NIV LEK': 1.0
}
```

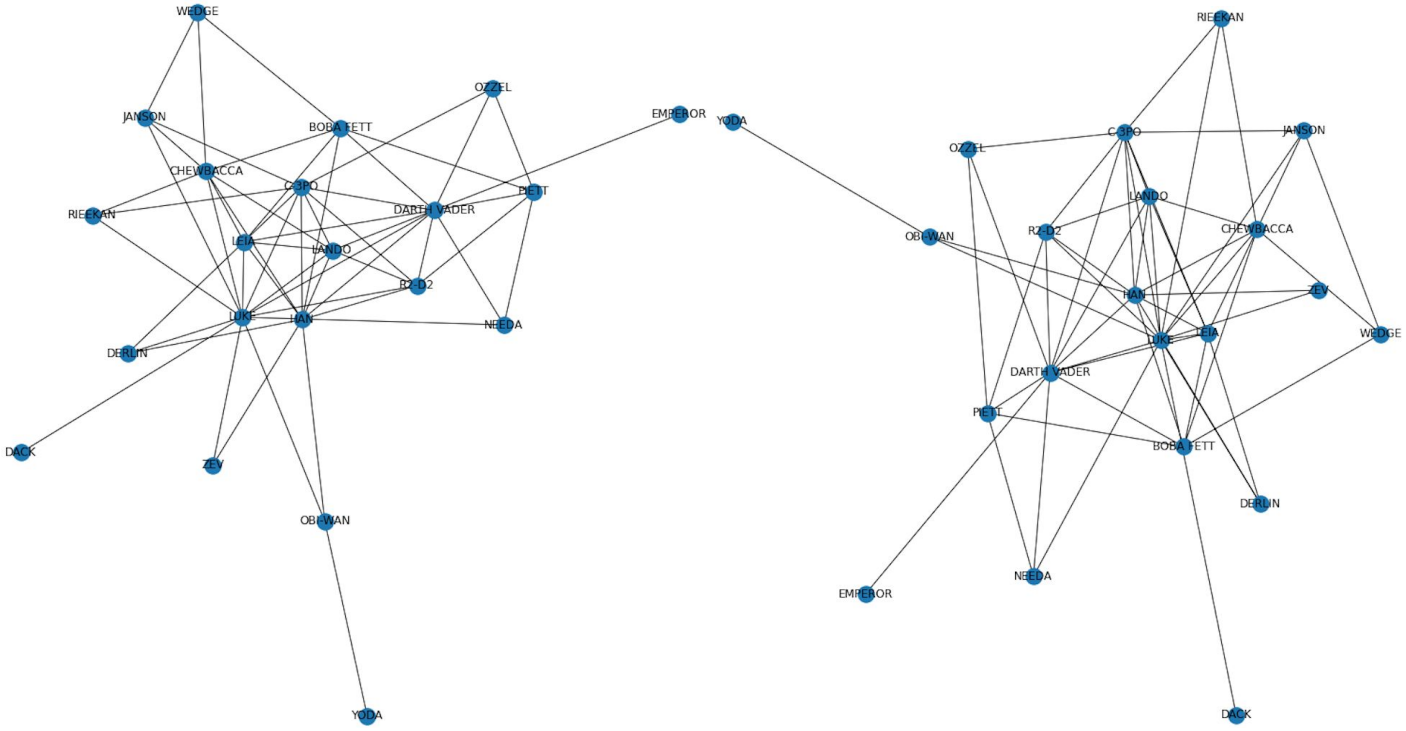
The values generated for the remaining episodes can be seen as parts of the output generated

from the python script.
Path lengths have also been determined on these graphs using the networkx library. A sample output path length for episode 7 has been displayed below.

Graph Generation and Link Randomization

From the graph created using the links and nodes from the dataset, I decided to follow the process used by Watts and Strogatz to randomize links and see how the structure changes through multiple iterations.

If the graphs created for each episode was to be considered as a small-world network, we would see a change in average path length (the average path length would increase). However, upon analysis of the two structures before and after the addition of randomness, we notice that not a huge difference in the graph structure can be seen as there already exists random links already exists



From the two random graphs generated, we notice that the degree centrality of Bobba Fett is 6 and 8. Even after adding the randomness, we notice that the average path lengths of both the graphs are similar. This is because the initial graph is already highly randomized.
Below is a numeric representation of how the average path length changed with each randomness.

```
[2.0619047619047617, 2.0619047619047617, 2.038095238095238, 1.9714285714285715, 1.9857142857142858,
1.9714285714285715, 1.9714285714285715, 1.9714285714285715, 1.9619047619047618, 1.9857142857142858,
1.9666666666666666, 1.9619047619047618, 1.938095238095238, 1.938095238095238, 1.938095238095238,
1.9142857142857144, 1.957142857142857, 1.9761904761904763, 1.9714285714285715, 1.9714285714285715]
```

Adithya Raghu Ganesh
200314659

Task 4:

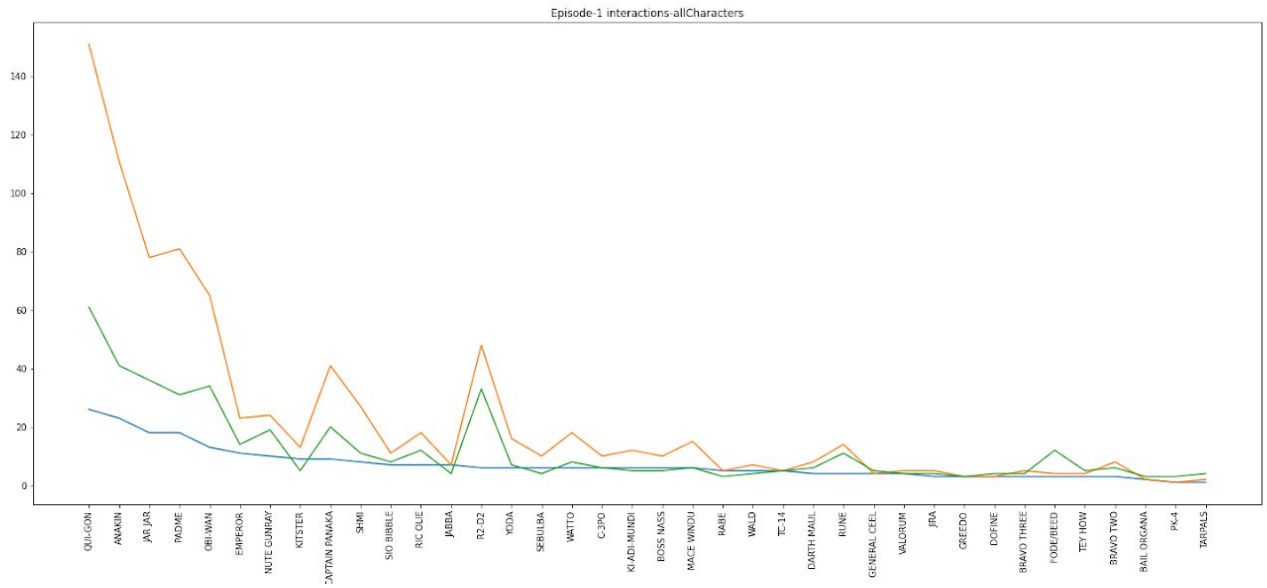
In order to identify weak ties, I am taking a relation between the number of interactions the character makes and the number of connections the character has. While determining the ties, I have assumed that the number of connections the character has is more important than the interactions he makes. It can be seen from the graphs below that the lesser connections a character has, the lesser is the interactions.

I decided to determine the weak ties by looking at the character vs interaction plot for each episode to determine when a tie gets weak. An assumption that I considered while determining a set of weak ties is that characters that exhibit very small connections are irrelevant as they tend to change episode by episode and nothing can be deduced from those results. So the approach I considered while determining the weak ties is by checking where the interactions and connections converge. This happens when each character has close to a single interaction with another character making the relationship weak. (the place where the orange and blue lines converge)

The following were the weak ties that were determined in each episode.

Key:

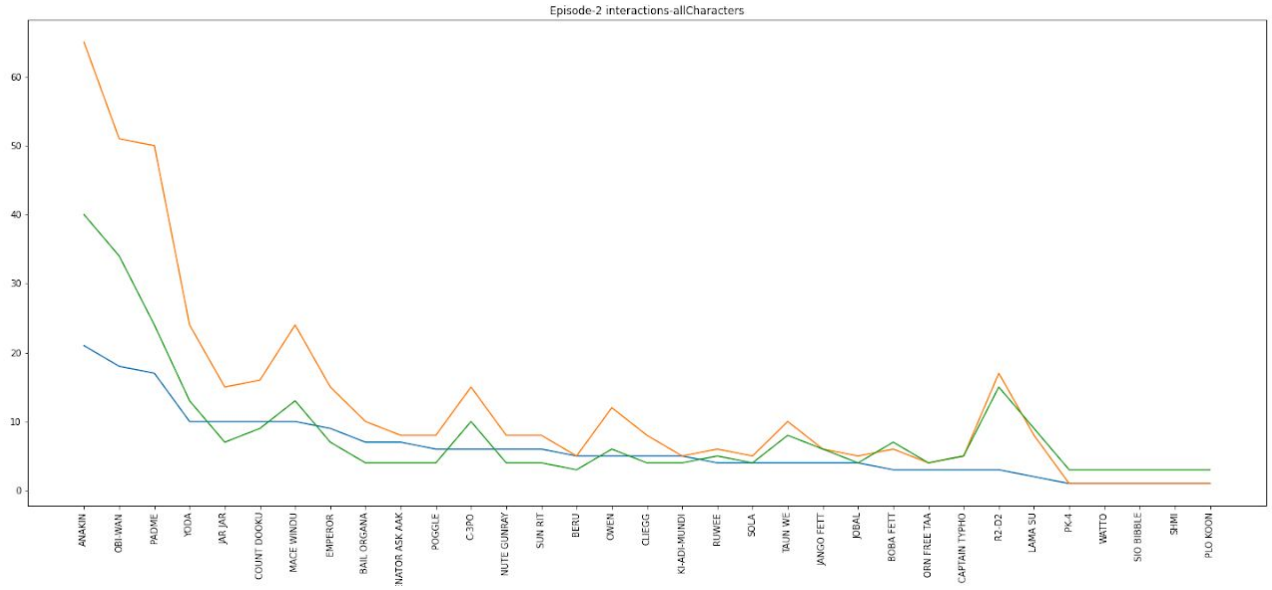
Orange line→ Number of character interactions, Green line→ Number of character appearances, Blue line→ Number of character connections



Weak ties → Jabba > Kitster > Rabe

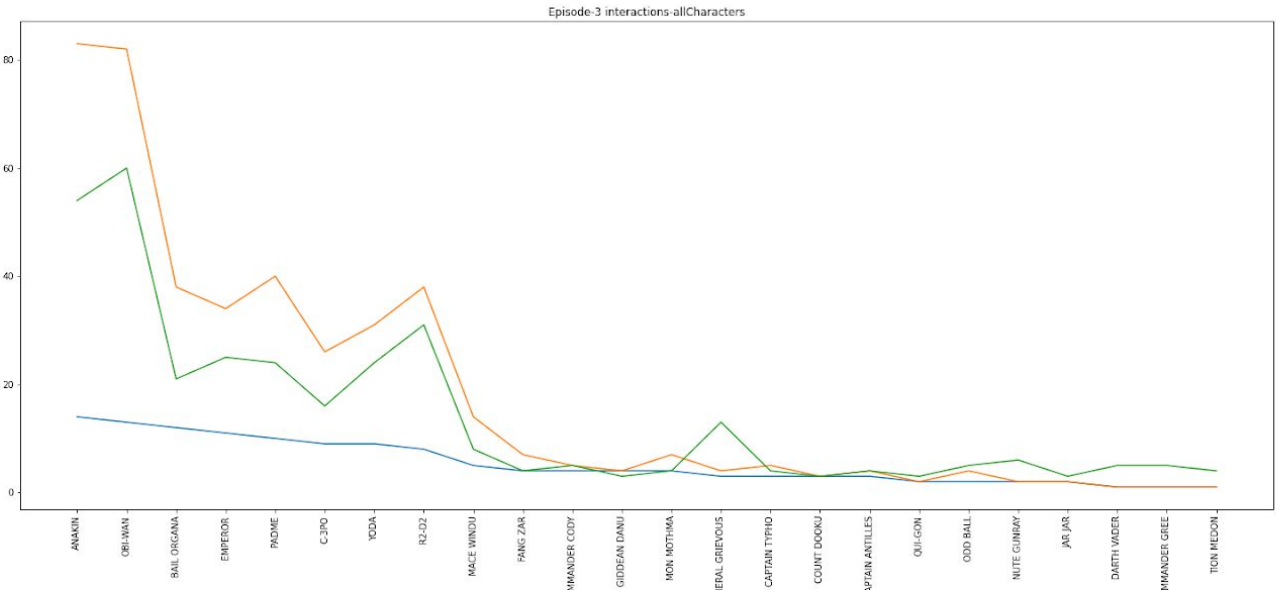
Through the series we notice that Jabba takes up more prominent roles. In this episode he is just being introduced and he has only weak ties. Kitster and Rabe on the other hand have weak ties because they come only in this episode making them less important.

Adithya Raghu Ganesh
200314659



Weak ties → Nute Gunray > Sio Bibble > Poggle

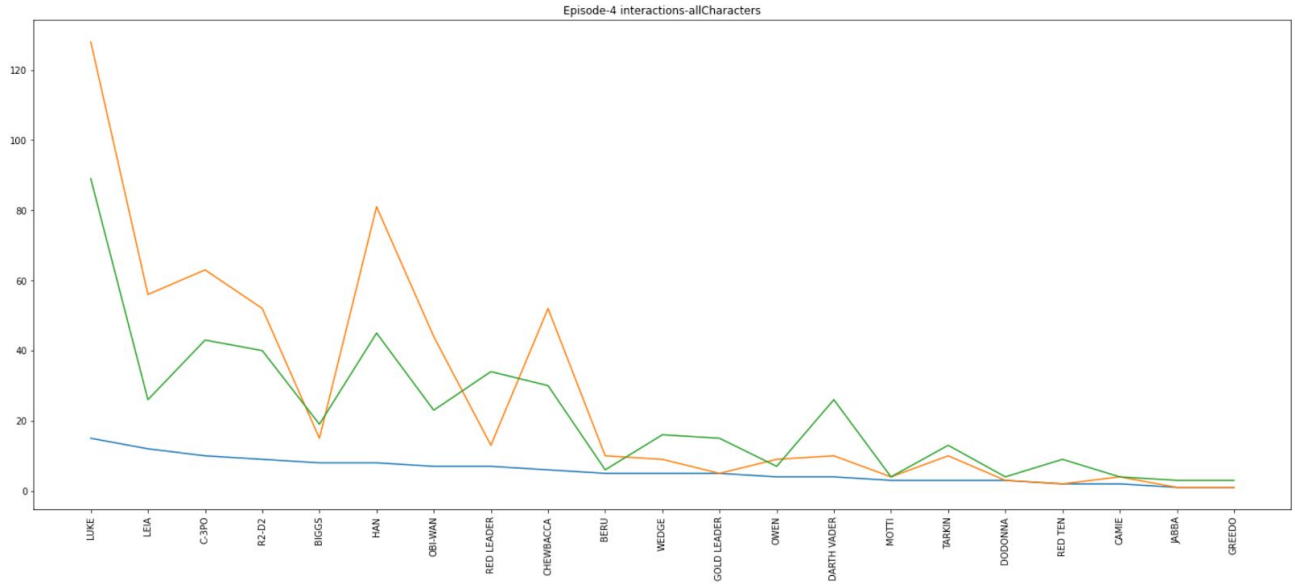
From this episode we can see that the orange and the blue curves start converging at Sio Bibble, Poggle and Nute Gunray. It looks like Nute Gunray and Sio Bibble had a more important role in Episode 1 in comparison to this one as their interactions decreased from the previous episode. Poggle on the other hand seems less important as it seems to be a temporary character



Weak ties → Darth Vader> Qui-Gon > Jar Jar

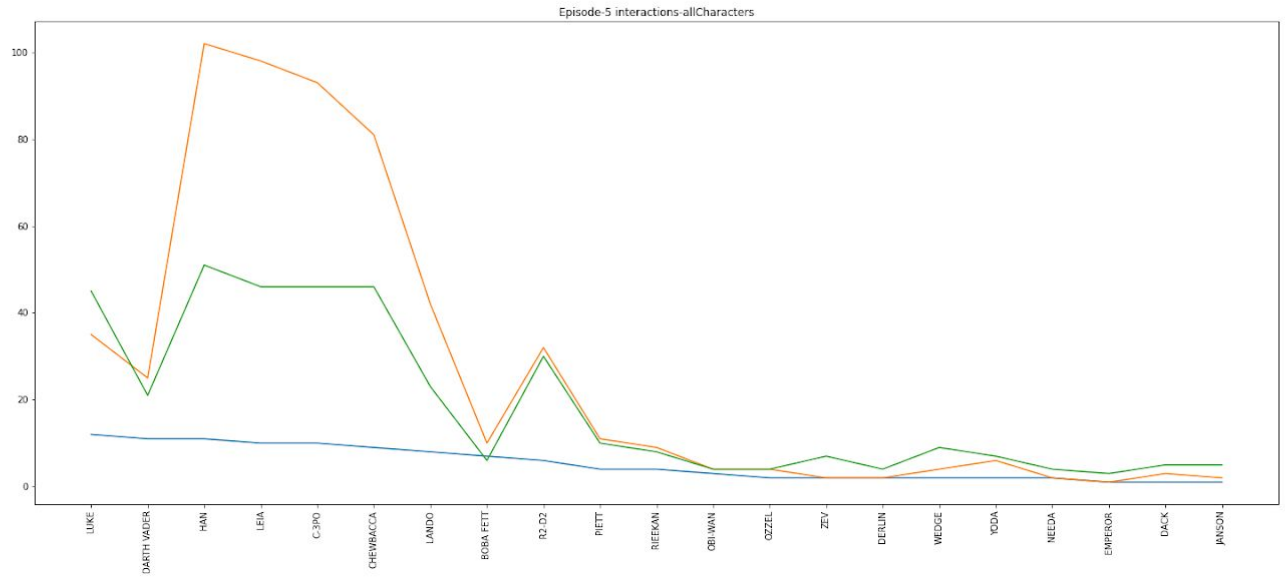
In this episode, we see the introduction of Darth Vader. He is being set up with weak ties as this episode acts as a character build-up. If you look at the importance of Jar Jar in this episode, you tend to notice that he was part of the top 5 characters in the previous episode and his presence is slowly decreasing. Finally, Qui-Gon - the most important character in Episode one has become the one of the least important ones here.

Adithya Raghu Ganesh
200314659



Weak ties → Darth Vader > Gold Leader > Jabba

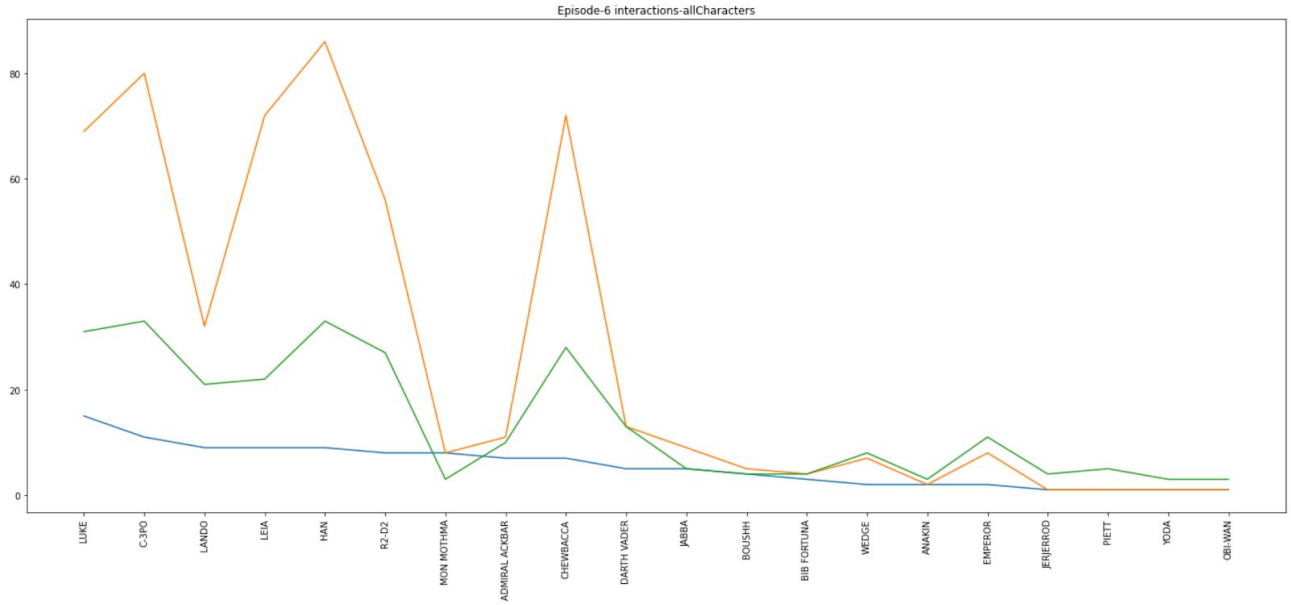
In this episode, Jabba has made his first appearance and is the weakest link in this episode. The viewers now have a glimpse of the character and in the episodes that follow, he seems to make an important presence making him a valuable character in spite of the weak ties in this episode. Darth Vader who seemed to have made a tiny appearance in the previous episode is again shown in this episode but his ties are still considered weak as his value is present after the two curves converge. Gold leader seems to be a less important temporary character as he appears only in this episode with a tiny role.



Weak ties → Needa = Janson > Emperor

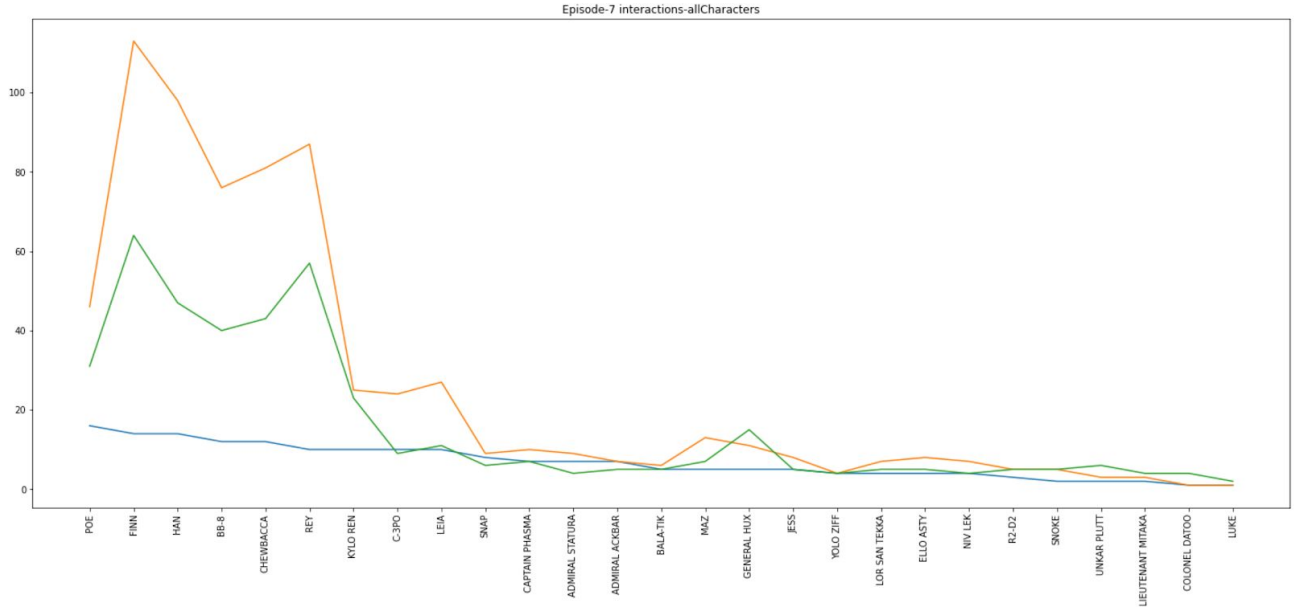
This episode shows us how the Emperor, an important character from the dark side, was getting phased out. Two other temporary characters that exhibited weak ties were Needa and Janson who seem unimportant

Adithya Raghu Ganesh
200314659



Weak ties → Anakin > Obi-Wan > Yoda

This episode in the series shows us how the three most important characters from the light side in the star wars series were phased out of the movie. All three of them exhibited weak ties as they are present towards the end of the curve above. This leads us to believe a change in plot with the integral characters making way.



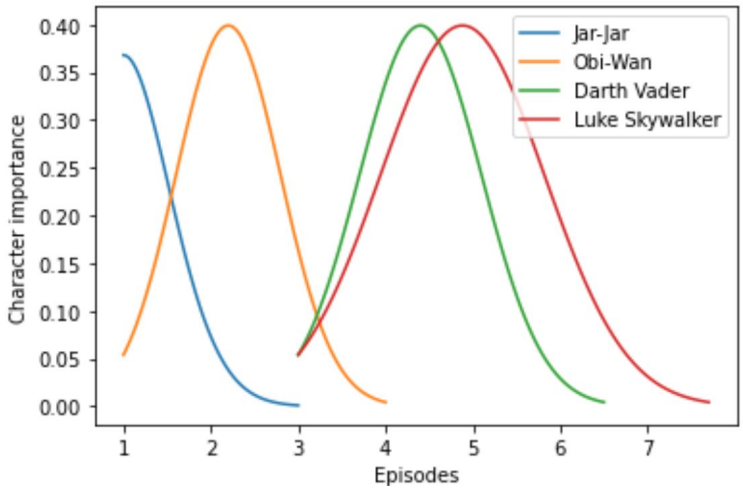
Weak ties → Admiral Ackbar > R2-D2 > Luke

Just like in the previous episode where primary characters from the first three episodes were being phased out, we notice a similar trend where the main characters - Luke and R2-D2 are being phased out. This is a classic example of how characters are introduced and how characters are removed in a series. Admiral Ackbas on the other hand has been making guest appearances throughout the series and he continues to do so in this episode by showing weak ties as depicted above.

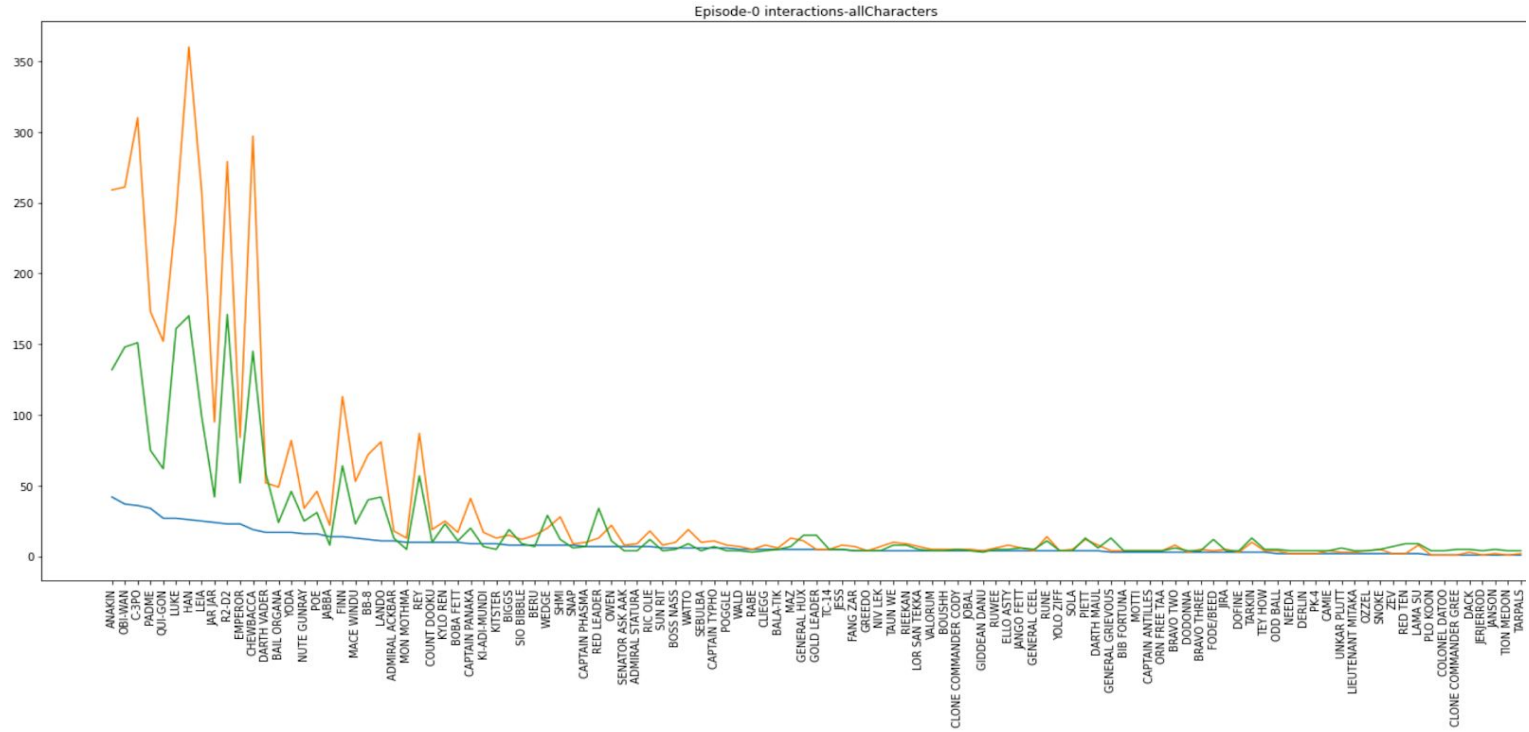
Adithya Raghu Ganesh
200314659

Task 5: The overall star-wars universe

The first major talking point from the above depictions is how a series has been kept alive with the appearance and disappearance of many major characters. The main characters follow a normal curve similar to the one given below



The above plot shows how 4 key characters overlapped and how one character took over the main role from another through the series. We can even confidently say important characters do appear in weak ties in some episodes as it may be at a time when the character is either being introduced or being faced out. Darth Vader is a good example for this.



One important thing to notice in the graph is the existence of R2-D2 before Darth Vader and Yoda. Even though the two characters had a much more essential role in certain episodes, it can be seen that R2-D2 however less essential seems to have had more interactions through the series making it more important when looking at the star wars universe as a whole.

Even though there were a lot of great insights, there are always a few shortcomings. From the global play it can be seen that over the series of 7 episodes, there were a few characters that were termed as main characters by looking at the universe graph. However when we look at the individual episodes, we notice that the number of characters in the initial few episodes are much greater than the final few.

This leads to an anomaly where older characters tend to have more interactions than the newer ones. Consider the case of Luke and Qui-Gon. We notice that Qui-Gon was one of the main characters in the first two episodes and Luke was the main character in episodes 4,5 and 6. Even though Luke had a main role to play in one extra episode, we see that Qui-Gon is placed ahead of Luke making him more important than Luke - which doesn't seem to be correct.

Link to the source code for detailed analysis and plots:
https://github.com/adithyarganesh/CSC555_Star_Wars_Network_Analysis