Social Computing

Name: Karan Rakesh
Unity ID: krakesh

Task 1

Hypothesis 1 : The metrics used to validate the first hypothesis are the median interactions of specific known characters that represent the light and the dark side. The median will give an unbiased estimation of who these particular characters interact with the most. This in turn will represent the sub-network that they are part of, which will help in validation of the hypothesis.

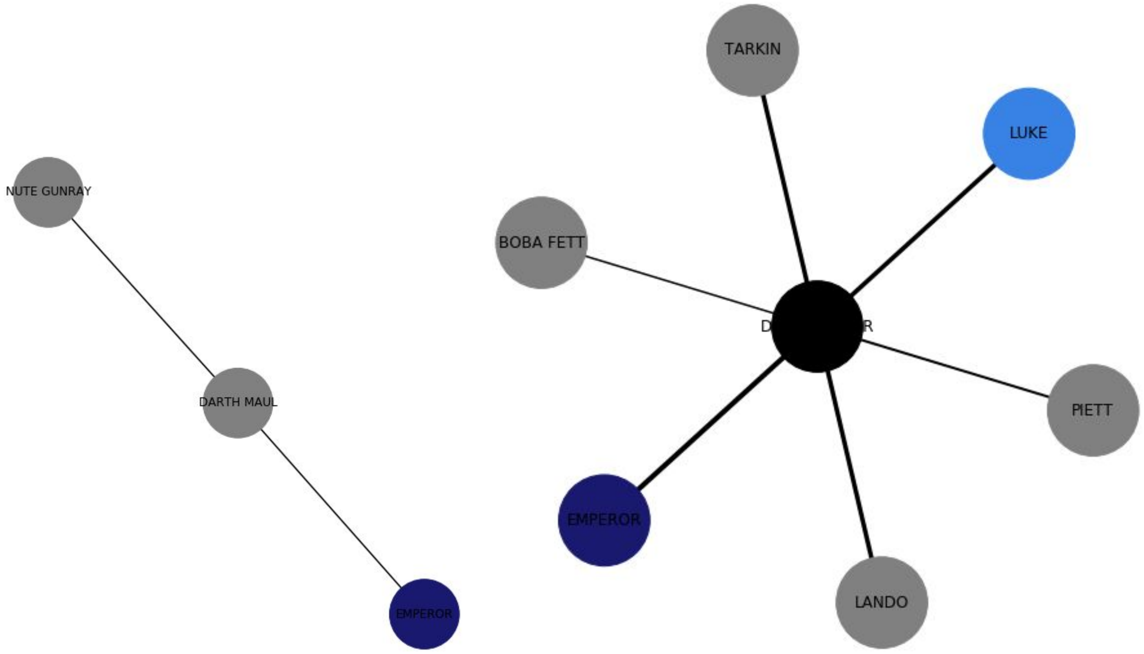
Hypothesis 2 : The metrics used to validate the second hypothesis are degree centrality and betweenness centrality. Degree Centrality gives us an understanding of the main character since it checks for the characters with the highest degree whereas betweenness centrality gives us an understanding of the most integral characters since it looks for the characters with the most number of shortest paths involving itself.

Task 2

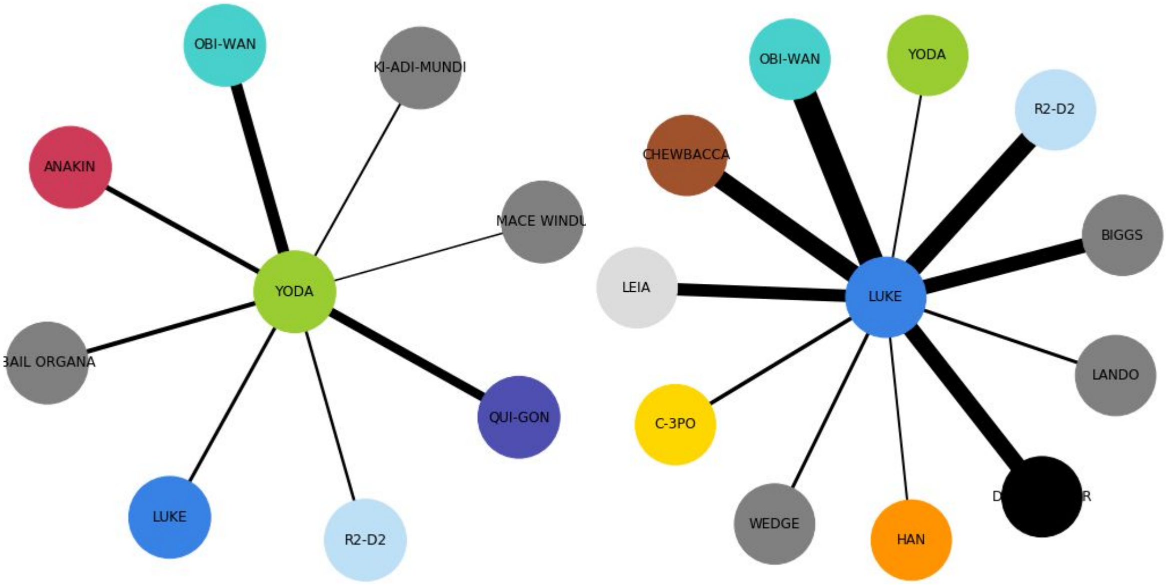
Hypothesis 1: To validate the hypothesis of the good side and dark side having their separate sub-networks that they communicate within, I identified two characters from the light side (Luke and Yoda) as well as two characters from the dark side (Darth Vader and Darth Maul). I then found all the links to these characters along with their weights, which represent the number of interactions these characters have had.

I then computed the median of the edge list of the character. This helped in providing an unbiased estimate of the threshold of interactions which can be eliminated as non integral. A statistic like mean would not work as well, since it will get affected by outliers. So once the median number of interactions were computed, I built a list of the characters that talk to the chosen character strictly more than the median. This will ensure only the important sub-network of that character are chosen. There are only a handful of these characters, which can be easily identified as the light side or dark side. Here are the results for the above analysis:

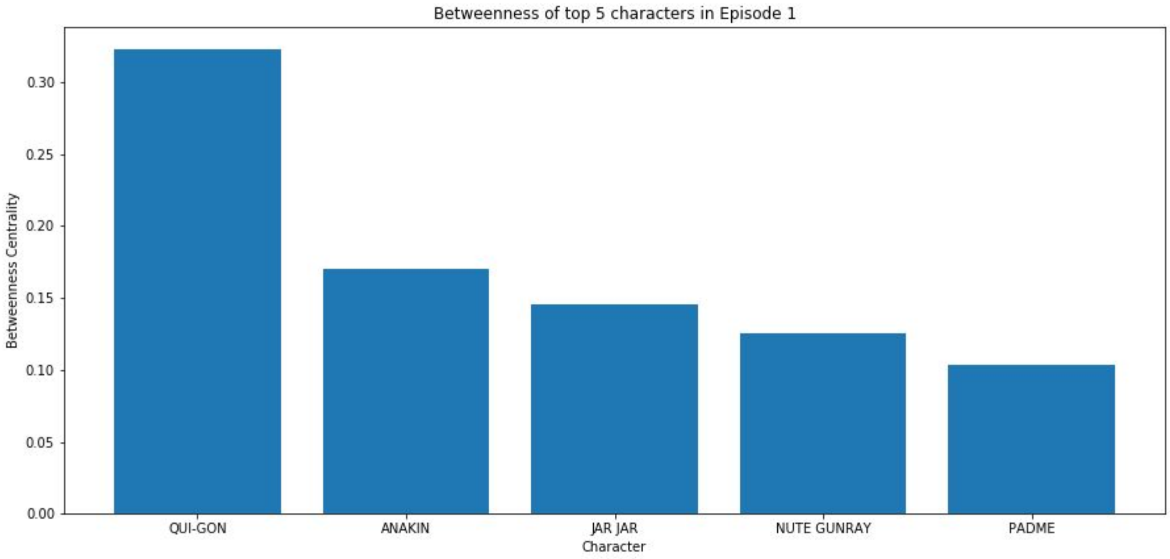
Dark Side: (A) Darth Maul (B) Darth Vader

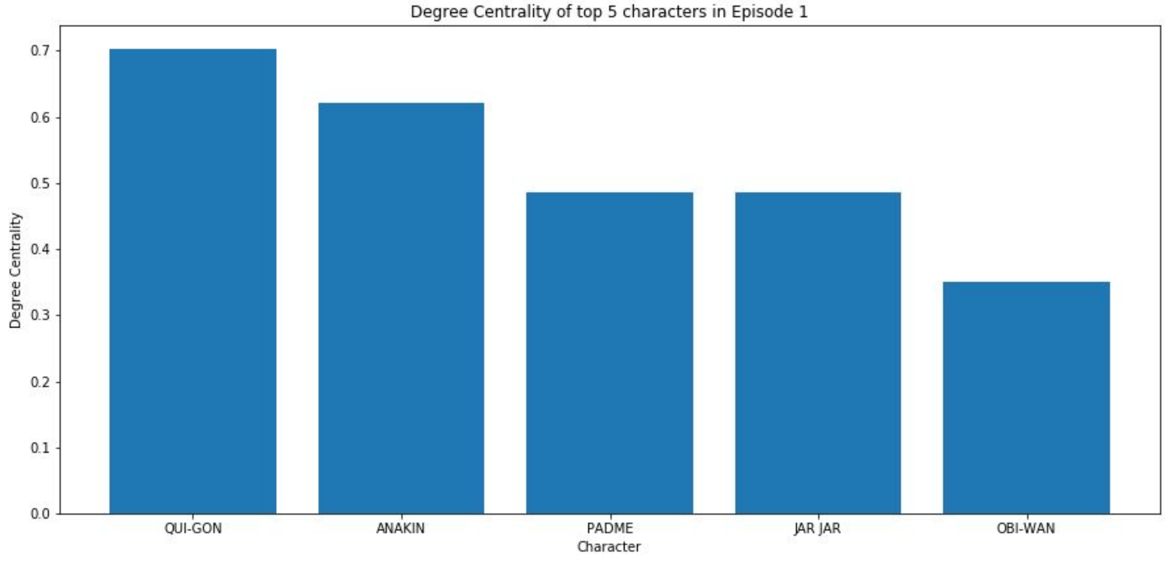


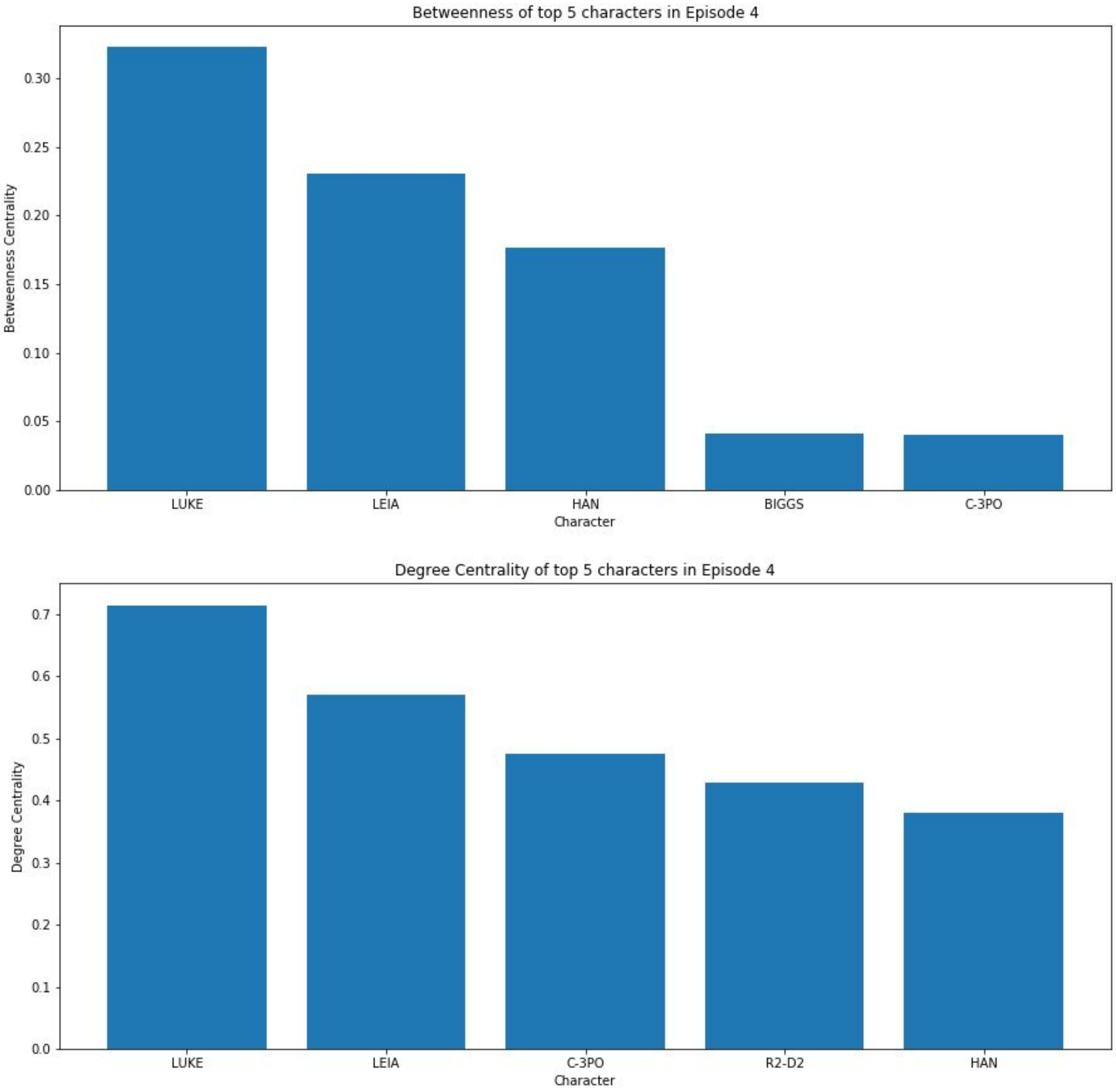
Good Side: (A) Yoda (B) Luke

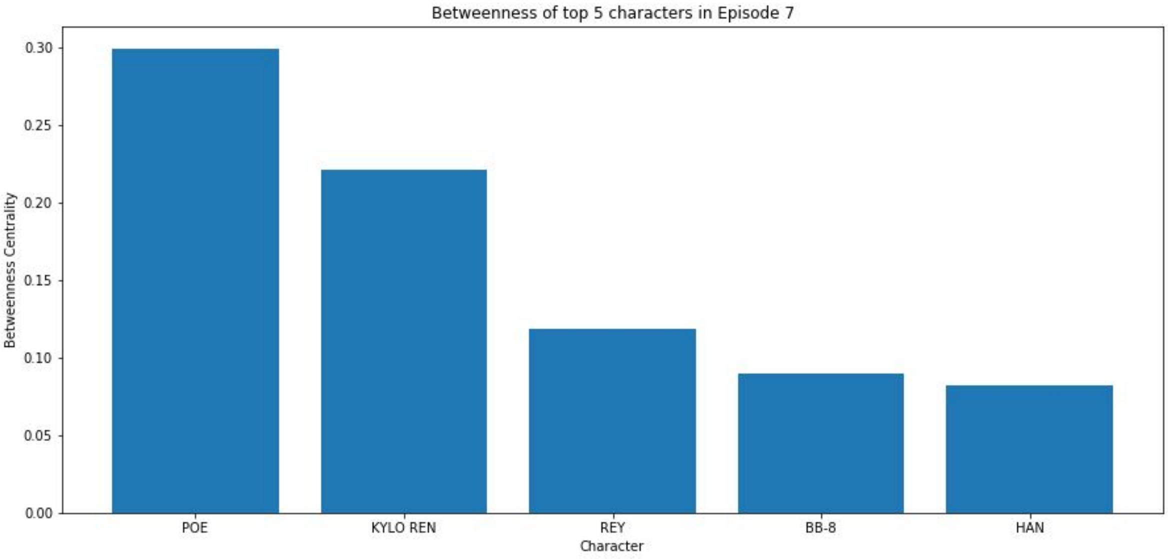


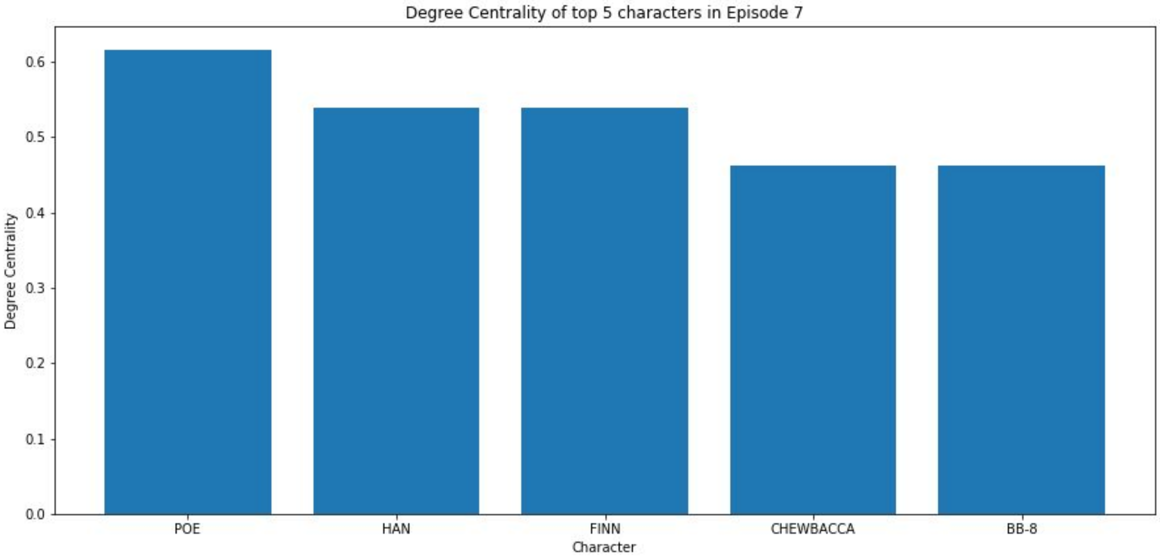
Hypothesis 2: If we perform an episode by episode analysis, we can see that the hypothesis holds for the characters that are not the most main or most integral i.e, the characters in the 3rd to 5th position in the 2 charts (betweenness and degree centrality). But the hypothesis doesn’t hold for the most important/main character as they also usually top the most integral characters as well. This can be seen from the graphs presented below.





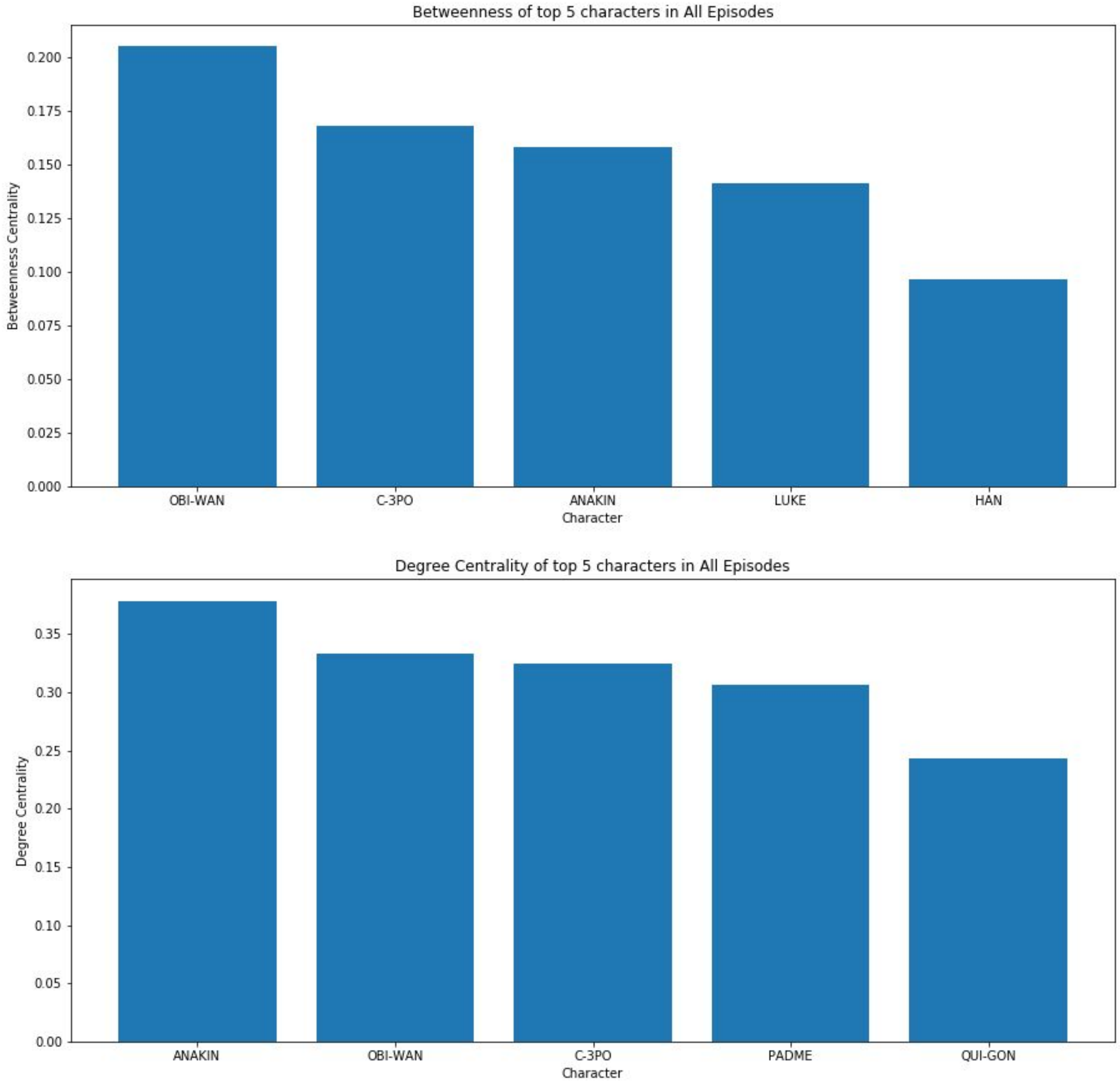

Betweenness of top 5 characters in Episode 4
Betweenness Centrality
0.00
0.05
0.10
0.15
0.20
0.25
0.30
LUKE
LEIA
HAN
BIGGS
C-3PO
Character
Degree Centrality of top 5 characters in Episode 4
Degree Centrality
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
LUKE
LEIA
C-3PO
R2-D2
HAN
Character





As we can see in the graphs above, while the position of the most integral characters (Qui-gon, Luke and Poe respectively) is the same as their position on the betweenness list, the other positions can vary significantly. A possible reason for this tendency can come down to the fact that the most integral character is an anomaly to the rule since they have many interactions with characters that do not talk to any other characters and hence require to have their shortest path to pass through the main character and hence it increases the betweenness of the main character. Also, as a consequence of the main character having so many connections, it is easier for most characters to have a shortest path that simply goes through the main character, hence invalidating the actual significance of the second main character. Since, if their network also includes the main character, all the shortest paths will just pass directly through them instead of passing through this character. As a result, in general there is no correlation between the main characters and the integral characters as per the observations from all the episodes.

If we look at the overall universe, we get the following metrics for betweenness and degree centrality :



From the graph of all the episodes combined, I drew the following conclusions:

- The graphs of betweenness do not correlate with the graphs on an episode by episode basis. This is likely because the characters integral to a particular episode need not be integral to the greater story being portrayed.
- The betweenness graph of all the episodes correlates with the actual integral characters of the star wars storyline as a whole. All the integral characters are characters that have been part of all the episodes. Hence these characters connect a majority of the characters that featured in either the first or the second trilogy.
- The degree centrality graph is dominated by characters that appear in the first trilogy versus the second trilogy. This is because the first trilogy has fewer characters overall and hence each character can have more interactions through a single episode thus causing their degree to rise as compared to characters that appeared solely in the second trilogy, where the higher number of characters gave each character less time to interact as a result.

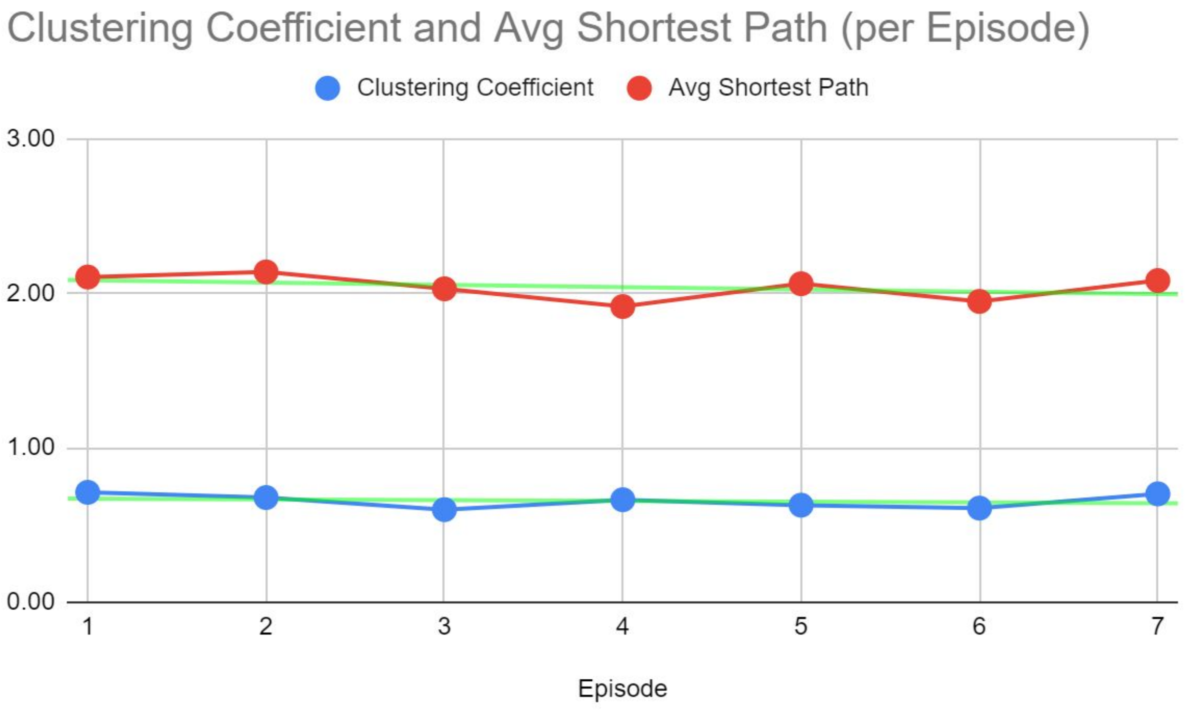
Task 3

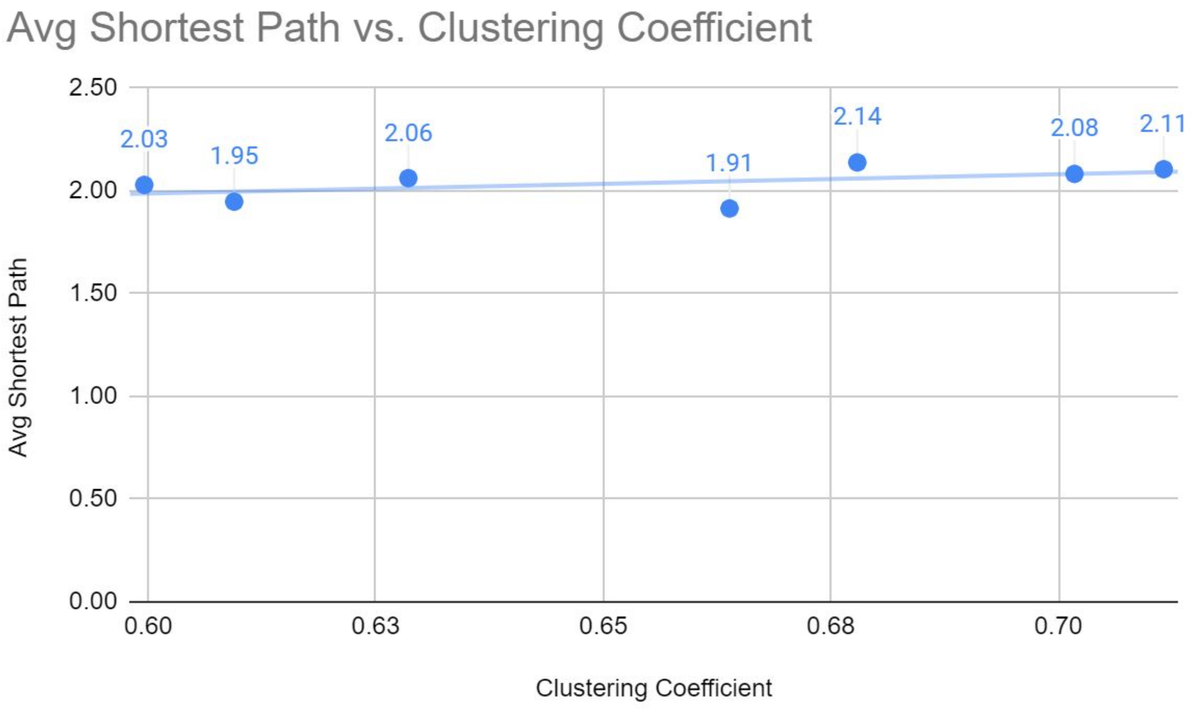
To find the cliquishness and the characteristic path length, I referred to a post by a professor which explains these terms in context of a small world network. A small world network is defined as a network where most pairs of nodes are only a few steps away from each other. As quoted from the source, a network is a small world network if it has

1. a small average shortest path length (scaling with logn, where n is the number of nodes), and
2. a high clustering coefficient.

Following this post, I found Networkx functions that perform all the calculations performed in the document and I carried those calculations out on each episode individually as well as on the all episodes json as well. I have attached a table as well as a few graphs to report my findings:

Cliquishness and Characteristic Path Length		
Episode	Clustering Coefficient	Avg Shortest Path
1	0.71	2.11
2	0.68	2.14
3	0.60	2.03
4	0.66	1.91
5	0.63	2.06
6	0.61	1.95
7	0.70	2.08
Mean from 1-7	0.66	2.04
All Episodes	0.69	2.66





From the above data, the following observations can be inferred:

- The clustering coefficient and the avg shortest path are both about the same for the entire series. If we divide the series into 2 trilogies ,i.e.,1-3 and 4-6, we can see that the clustering coefficient and the average shortest path of the first trilogy (parts 4-6) are both nearly always lesser than the mean values for the entire series.
- The clustering coefficient and avg shortest path are both positively correlated, albeit very slightly, as can be seen from the trendlines plotted in the second graph.
- When calculating the following values for all the episodes combined, as shown in the table, we see that the clustering coefficient is higher than the mean while the avg shortest path is longer. This can be explained since the clustering coefficient of the entire graph will be higher as they all form larger clusters with the main characters. While on the other hand, since there are completely disjointed characters being brought from different timelines, it will be harder to form short paths rather than when all the characters are from a single episode. Hence the avg shortest path is higher than the mean value calculated from the episodes.

For the random graph part, I tried two approaches:

- For the first approach, I used the mappings for the nodes and created a completely random graph. This resulted in a graph with a significantly lower correlation coefficient once I adjusted the probability of generating the edge to the final amount. This is understandable since the truly random graph will be sparse connected and not have many cliques since it is truly random.
- The second method was to take the graph from the first episode and then with a randomly probability chance remove and replace edges. This did not adverse affect the cliquishness or the average shortest path of the graph. But if this is done at a large scale, it would start to approach the truly random graph.

Task 4

To define a weak tie, I carried forward the definition that I used from hypothesis 1 to calculate the strong sub-network of particular characters. But instead, I tweaked it to calculate the median of all the interactions within an episode and then only displayed the strictly weak ones, i.e., the interactions that had a value strictly lower than the median were classified weak ties and then I picked 3 which I thought might be interesting as per their general importance or other factors. I have mentioned them in a per episode basis below:

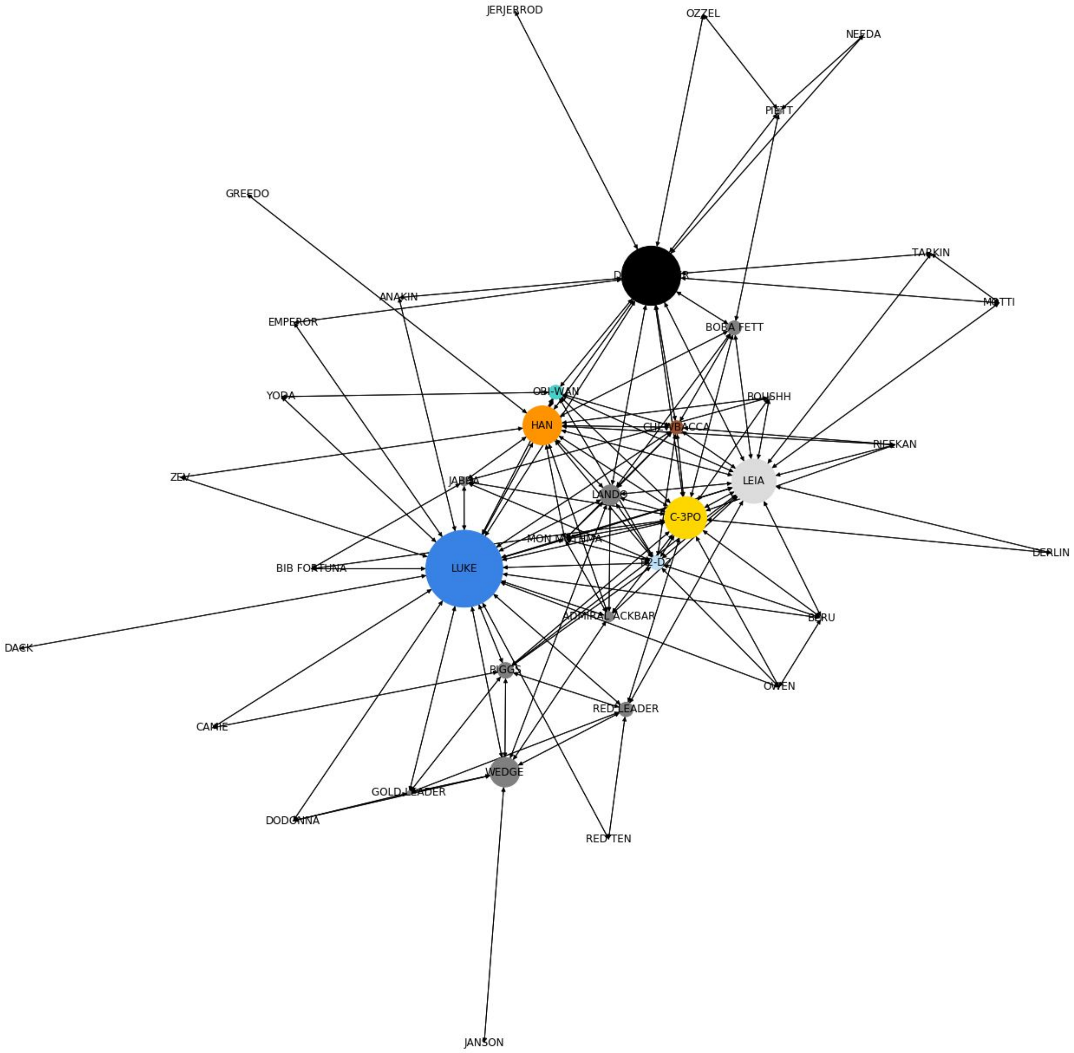
- Episode 1
    - OBI-WAN and PADME >= NUTE GUNRAY and
PADME > DARTH MAUL and GENERAL CEEL
    - The first two interactions are surprisingly classified as weak ties as they occur between main characters in the first episode. The third one was interesting since the dark side in general seem to have very few interactions, resulting in weak ties.
- Episode 2
    - Episode 2 has no weak ties as per the given definition. This is because the median interaction value is 1. And hence it looks for strictly less than 1, hence nothing shows up.

- Episode 3
  - EMPEROR and DARTH VADER > YODA and
    QUI-GON >= ANAKIN and COUNT DOOKU
  - It is interesting that the emperor and darth vader have a weak tie despite being very important allies on the dark side. It is also surprising to see important characters like Qui-gon and Yoda have weak ties. Finally Anakin and Count Dooku probably battle before Anakin turn into Darth Vader and hence share a relevant weak tie.
- Episode 4
  - DARTH VADER and OBI-WAN >= DARTH VADER
    and LEIA > LEIA and OBI-WAN
  - It is interesting that Darth Vader and his original master Obi-wan have a weak tie. It is also surprising that darth vader and his daughter Leia have a weak tie despite being important characters. Finally Leia and Obi-wan also have a weak tie, signifying obi-wan's diminishing importance as the series unfolds.
- Episode 5
  - DARTH VADER and EMPEROR > LUKE
    and LEIA > LEIA and DARTH VADER
  - For the second episode, Darth Vader and the emperor have a weak tie despite their importance on the dark side. Also siblings Luke and Leia have a weak tie, while the weak tie from the previous episode between father-daughter Leia and Darth Vader continues.
- Episode 6
  - DARTH VADER and ANAKIN > LUKE and
    OBI-WAN > C-3PO and LANDO
  - It is funny that Darth Vader and Anakin have a weak tie, despite being the same person. Probably because the conversion happens over the course of a single scene. Following that, the extremely important characters from the previous parts, Luke and Obi-wan form a weak tie. Finally, there is the weak tie between C-3PO and Lando.
- Episode 7
  - REY and LEIA > LUKE and
    REY > FINN and C-3PO
  - While Rey is a very important character in the 7th episode, she has weak ties with important characters from previous episodes like Luke and Leia. Lastly I chode Finn and C-3PO as it's not very important.

Task 5

For task 5, I created 2 separate graphs corresponding to the two trilogies. First being 1-3 and second being 4-6. To ensure the graph was displayed well, I only considered the biggest connected component. This leaves out only a negligible amount of disconnected components and helps while rendering the graph. To create this merged graph, I first stored all the episodes' graphs in separate variables. I then made the name of the character into the node identifier since the IDs have different mapping in different episodes. I then considered the interactions between characters over the 3 episodes and added them up. This gives me a true merged graph. I then went ahead and plotted this graph with dynamic node sizes as per their betweenness. Hence the most integral characters will have the largest node sizes. Finally, I also performed analysis to find out the links who have over 25 interactions, which is an arbitrarily high threshold. I also computed the total number of characters, edges and interactions to compare over the two trilogies.

EPISODES 4-6


JERJERROD
OZZEL
NEEDA
PIETT
GREEDO
TARKIN
DARTH VADER
MOTTI
ANAKIN
EMPEROR
BOBA FETT
YODA
OBI-WAN
BOUSHH
HAN
CHEWBACCA
RIEEKAN
ZEV
JABBA
LEIA
LANDO
C-3PO
MON MOTHMA
BIB FORTUNA
LUKE
R2-D2
DERLIN
ADMIRAL ACKBAR
BERU
DACK
BIGGS
OWEN
CAMIE
RED LEADER
WEDGE
GOLD LEADER
DODONNA
RED TEN
JANSON

EPISODES 1-3

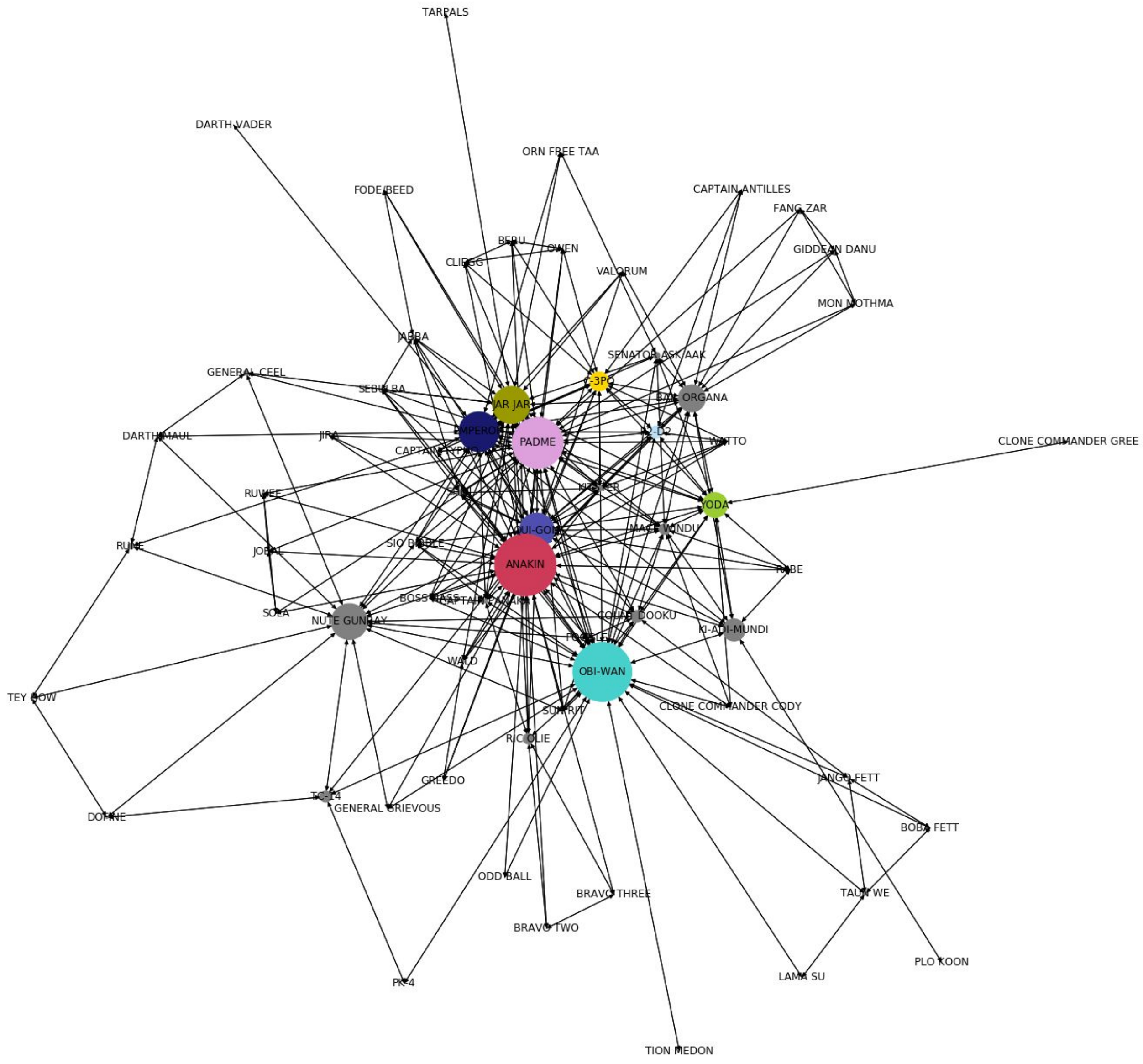

TARPALS
DARTH VADER
ORN FREE TAA
FODE/BEED
CAPTAIN ANTILLES
FANG ZAR
GIDDEAN DANU
BERU
OWEN
CLIEGG
VALORUM
MON MOTHMA
JABBA
SENATOR ASK AAK
GENERAL CEEL
SEBULBA
C-3PO
JAR JAR
BAIL ORGANA
DARTH MAUL
JIRA
EMPEROR
PADME
R2-D2
WATTO
CLONE COMMANDER GREE
CAPTAIN TYPHO
SHMI
KITSTER
RUWEE
YODA
RUNE
QUI-GON
MACE WINDU
JOBAL
SIO BIBBLE
ANAKIN
RABE
BOSS NASS
CAPTAIN PANAKA
SOLA
NUTE GUNRAY
COUNT DOOKU
KI-ADI-MUNDI
POGGLE
WALD
OBI-WAN
TEY HOW
SUN RIT
CLONE COMMANDER CODY
RIC OLIE
JANGO FETT
GREEDO
TC-14
GENERAL GRIEVOUS
DOFINE
BOBA FETT
ODD BALL
BRAVO THREE
TAUN WE
BRAVO TWO
PLO KOON
LAMA SU
PK-4
TION MEDON

Table comparing key metrics over the two trilogies

Metrics	1-3	4-6
No of Characters	65	40
No of Edges	251	125
No of Interactions	728	700

Pairs with over 25 Interactions

1-3	4-6
R2-D2 and ANAKIN QUI-GON and OBI-WAN	R2-D2 and C-3PO R2-D2 and LUKE CHEWBACCA and C-3PO CHEWBACCA and LUKE CHEWBACCA and HAN CHEWBACCA and LEIA C-3PO and LUKE LUKE and LEIA LUKE and HAN

Some of the key findings from this task are as follows:

- The latter trilogy (4-6) is much simpler at first glance and has around 5-6 clear nodes that are integral to the plot while every other character plays a supporting role. Whereas in the other trilogy, it is much more distributed. There are about 10 significant characters that are integral to it and hence requires deeper analysis to gather insights.
- As shown in the first table, there are only 45 characters in the second trilogy versus 65 in the second. The amount of edges also has a significant gap with 125 vs 251. This implies that there are more unique interactions in the first trilogy. But the most surprising fact is that despite more characters and edges, the overall number of interactions in both trilogies are similar at 700 vs 728. This implies that there are more number of interactions between the characters in the second trilogy versus the former, and this is verified by the second table. There are 9 pairs with over 25 interactions in the second trilogy vs 2 in the first.
- We can see that despite R2-D2 having such a high number of interactions with integral characters in the story, it does not show up on either the betweenness or the degree centrality graphs. This must mean that there needs to be a parameter that incorporates who is interacting with you to decide the main characters. And we should weigh the characters interactions according to their importance to the plot rather than merely counting the number of instances.
- Despite CHEWBACCA showing up in the maximum number of 25+ interactions, it doesn't show in any of the betweenness or degree centrality graphs. This is again because, there is no weight given to the number of interactions instead it is merely the number of unique interactions a character has.