A Hive of Scum and Villainy - A Report on the Social Networks in the Original Star Wars Film Trilogy

Advanced Social Network Analysis

Declan Kehoe

Introduction

The Star Wars universe has grown since the original 1977 film into one of the largest and most recognised intellectual property's in the world. The networks developed for this analysis are based on the Star Wars Social Network dataset (Bhatia, 2020), which uses the films' scripts to identify ties between characters. Specifically, characters that talk in the same scene were identified as having a connection which, while overall a good methodology, has the key limitation of not capturing connections even if characters are in the same scene but don't converse. While the dataset contains connections for each of the first seven films in the franchise, this analysis uses data on the originally released trilogy, somewhat confusingly named Episodes 4: A New Hope, Episode 5: The Empire Strikes Back, and Episode 6: Return of the Jedi. These were selected as a more reasonable amount of information to interpret and analyse. A baseline level of familiarity with some of the characters and themes of the films is assumed, although is certainly not necessary for the report to be coherent.

This report will first introduce the research questions studied, followed by a more thorough description of the data. Next, the methods used to conduct the study will be explained, before an analysis of the results that were produced as related to some social network theories. A final summary of the findings, as well as identification of the weaknesses and potential improvements further research could investigate, finalises the report.

Hypotheses

The creation of the hypotheses arose from both the nature of the Star Wars narrative, and the information that was captured about the characters comprising the three networks. For context regarding the story, the series of films are examples of the space opera genre, where forces of good and evil struggle to prevail on a galactic scale. The narrative of the three featured films focuses on a few conflicting factions, and each character is primarily aligned with one of these groups. Underlying everything is also the concept of 'the Force', a magical power which influences everything, and to which some people are sensitive. The three networks will each have the same three hypotheses tested, and from these results a fuller picture of the dynamics of the networks can hopefully be seen evolving throughout the story. The hypotheses are:

- Belonging to the same faction increases the likelihood of characters interacting
- Being Force-sensitive increases the likelihood of making connections
- If characters meet, it is more likely that they will have a mutual acquaintance than not

The first hypothesis relates to homophily effects, the observation that similar people tend to interact with each other (McPherson et al., 2001). The second explores the effect the Force-sensitivity attribute of a character has on their amount of interaction, a measure of those nodes' degree. The final hypothesis relates to the likelihood of a new interaction creating a mutually connected sub-group of three people in the network, known as transitive closure. Overall the theme of the research questions can be seen as a measure of balance, homophily, and transitivity.

Describing the Data

The dataset itself is as a set of JSON files, each of which contain a zero-indexed list of character names and ids which correspond to a separate zero-indexed list of 'source' and 'target' integers containing the ties. Both lists also have a 'value' integer, which for the ties represents how many times that interaction occurred, and the value for characters representing the amount of times the character appears in the relevant film. Neither of these value measures were used in this analysis, although follow up studies could certainly employ them, as discussed further in the conclusion.

From the original dataset, three .CSV files were generated containing the list of ties as an undirected and unweighted edge list, and three .txt files containing character id's and names. The attribute files were then extended to include the character's Force-sensitivity (either None or Sensitive), and their primary faction. The attributes were derived from information about each character available on Wookiepedia, one of the largest and most authoritative sources of Star

Wars data (Garcia, 2018). The networks representing the chosen films and their properties are summarised in the figures on the following pages. Please note the colours in each episode's table serve as a legend for the network plot that follows it, and the numbers show how many nodes have that attribute.

Network for Episode 4 (A New Hope)

	Descriptive statistics					
Nodes	20					
Ties	58					
Density	0.305					
Force	**None** 16			**Force Sensitive** 4		
Faction	Bounty Hunter 4	Galactic Empire 2	Jedi 2	Neutral 2	Rebel Alliance 9	Sith 1

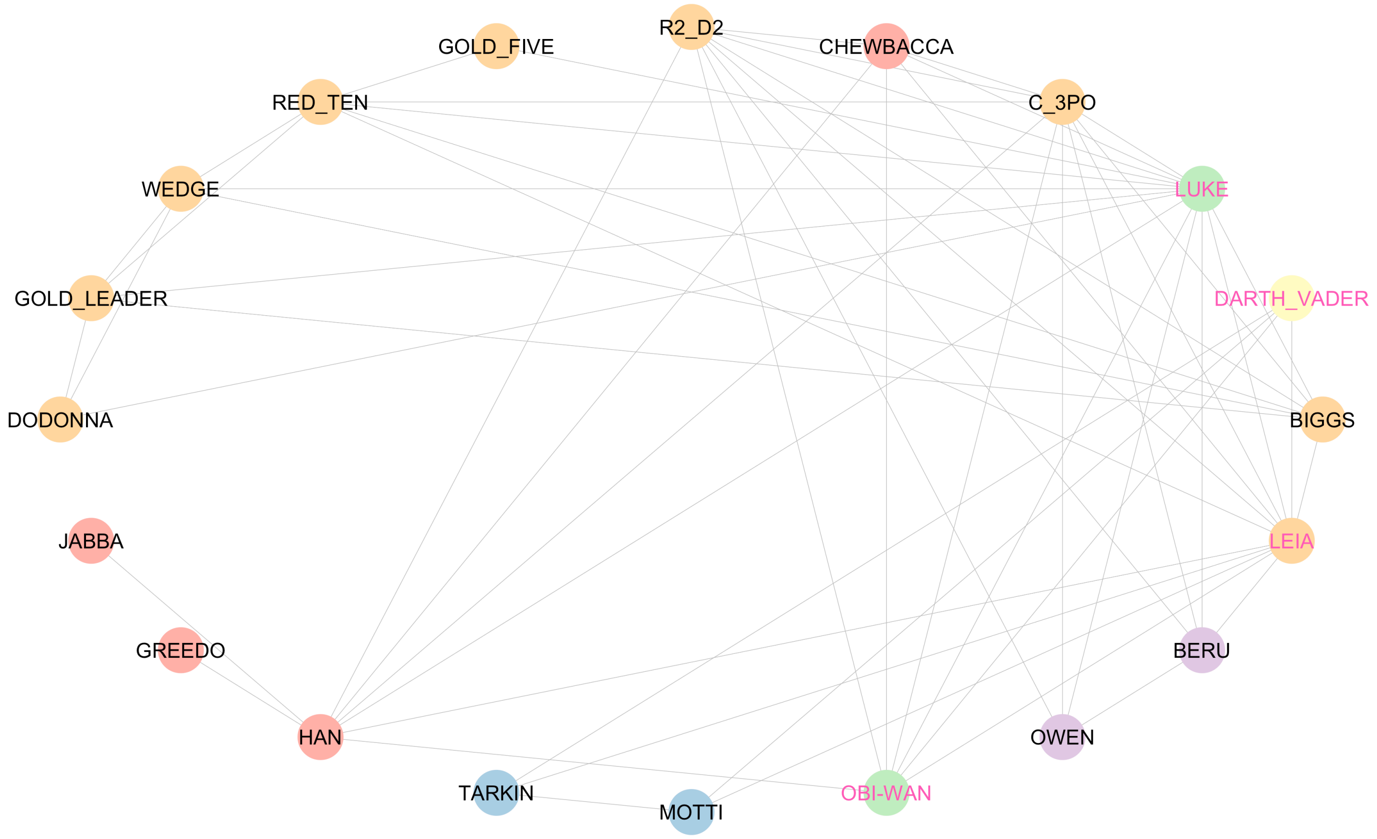


Figure 1 – Table and plot showing the network for Episode 4, where node background colour denotes faction and text colour denotes Force-sensitivity, following the colours in the table. Ties are undirected and unweighted.

Network for Episode 5 (The Empire Strikes Back)

	Descriptive statistics				
Nodes	20				
Ties	54				
Density	0.284				
Force	**None** 14			**Force Sensitive** 6	
Faction	Bounty Hunter 2	Galactic Empire 3	Jedi 3	Rebel Alliance 10	Sith 2

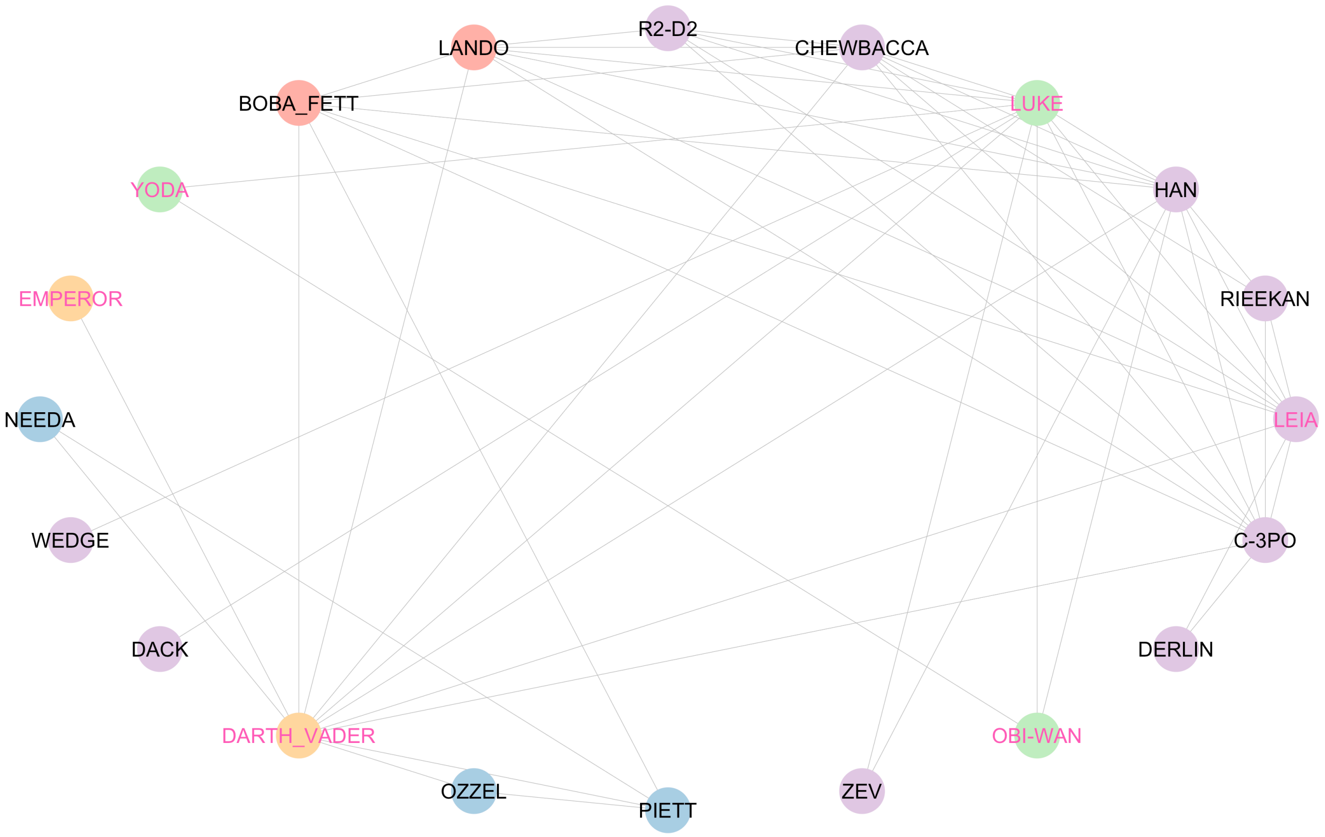


Figure 1 – Table and plot showing the network for Episode 5, where node background colour denotes faction and text colour denotes Force-sensitivity, following the colours in the table. Ties are undirected and unweighted.

Network for Episode 6 (Return of the Jedi)

	Descriptive statistics				
Nodes	19				
Ties	53				
Density	0.310				
Force	None 13			Force Sensitive 6	
Faction	Bounty Hunter 3	Galactic Empire 2	Jedi 3	Rebel Alliance 9	Sith 2

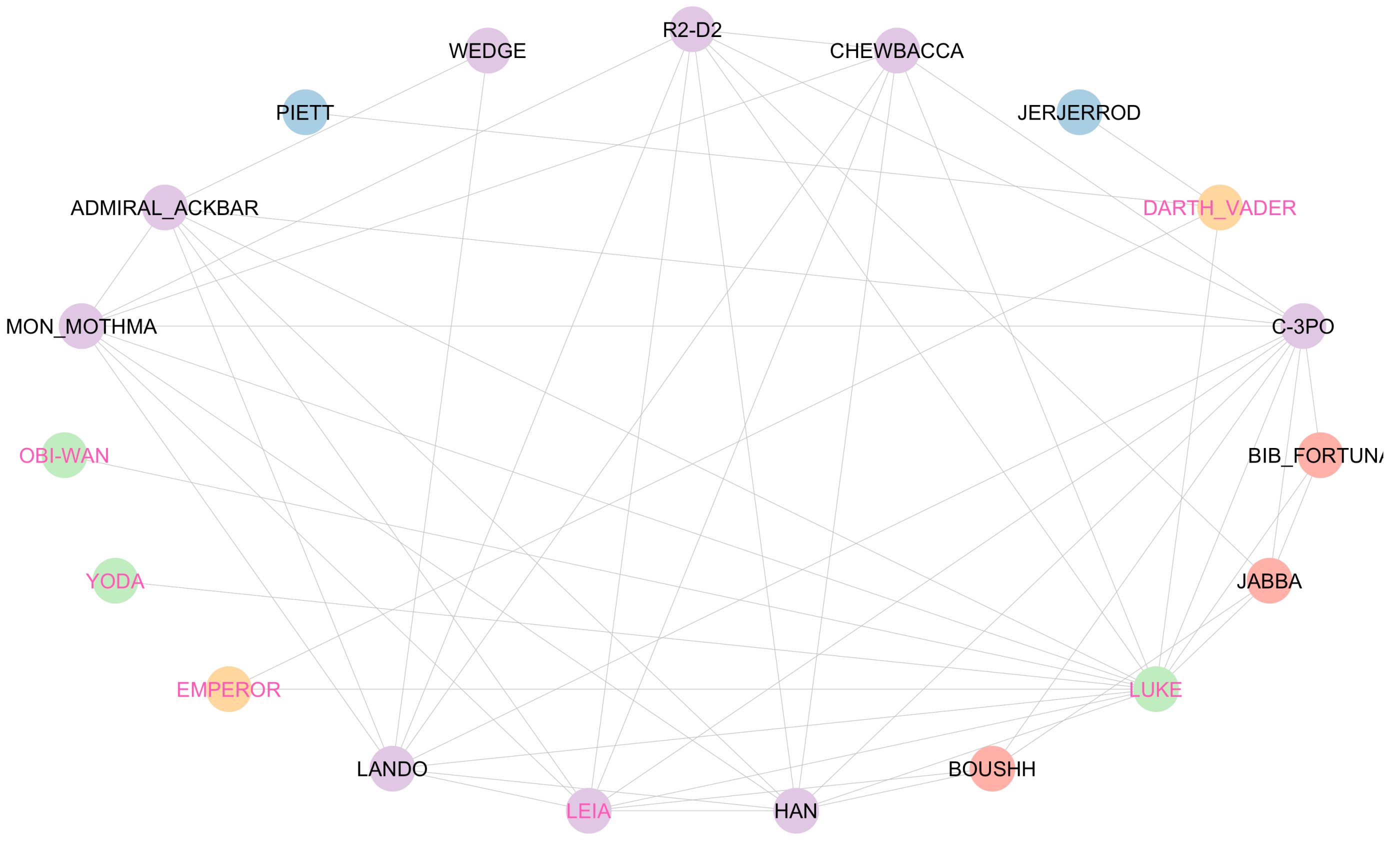


Figure 1 – Table and plot showing the network for Episode 6, where node background colour denotes faction and text colour denotes Force-sensitivity, following the colours in the table. Ties are undirected and unweighted.

Hypothesis Testing Method

All of the hypotheses were tested using exponential random graph models (ERGMs). ERGMs use some chosen empirical statistics of an observed network to generate models of similar networks. These model networks can subsequently be employed to ascertain the likelihood of the original network's properties having occurred by chance, and thus help answer research questions about the forces which influenced that original network's shape (Robins et al., 2007).

To test the hypothesis that belonging to the same faction increases the likelihood of characters interacting, a statistical measure, or a θ, was added to the ERGM which uses the effect that a node's faction attribute has on the likelihood of a tie forming between them. In the R code this used the method `nodematch`, which calculates how often a chosen attribute is the same for the nodes at both ends of a tie. For example, if an interaction was between two characters who were both in the Rebel Alliance, this would be recorded by the nodematch. This hypothesis and method are a way of measuring the network's homophily; specifically the uniform homophily was measured, that is, the overall amount of homophily across all the factions, rather than the specific amount within each faction, which would be differential homophily (statnet team, 2017).

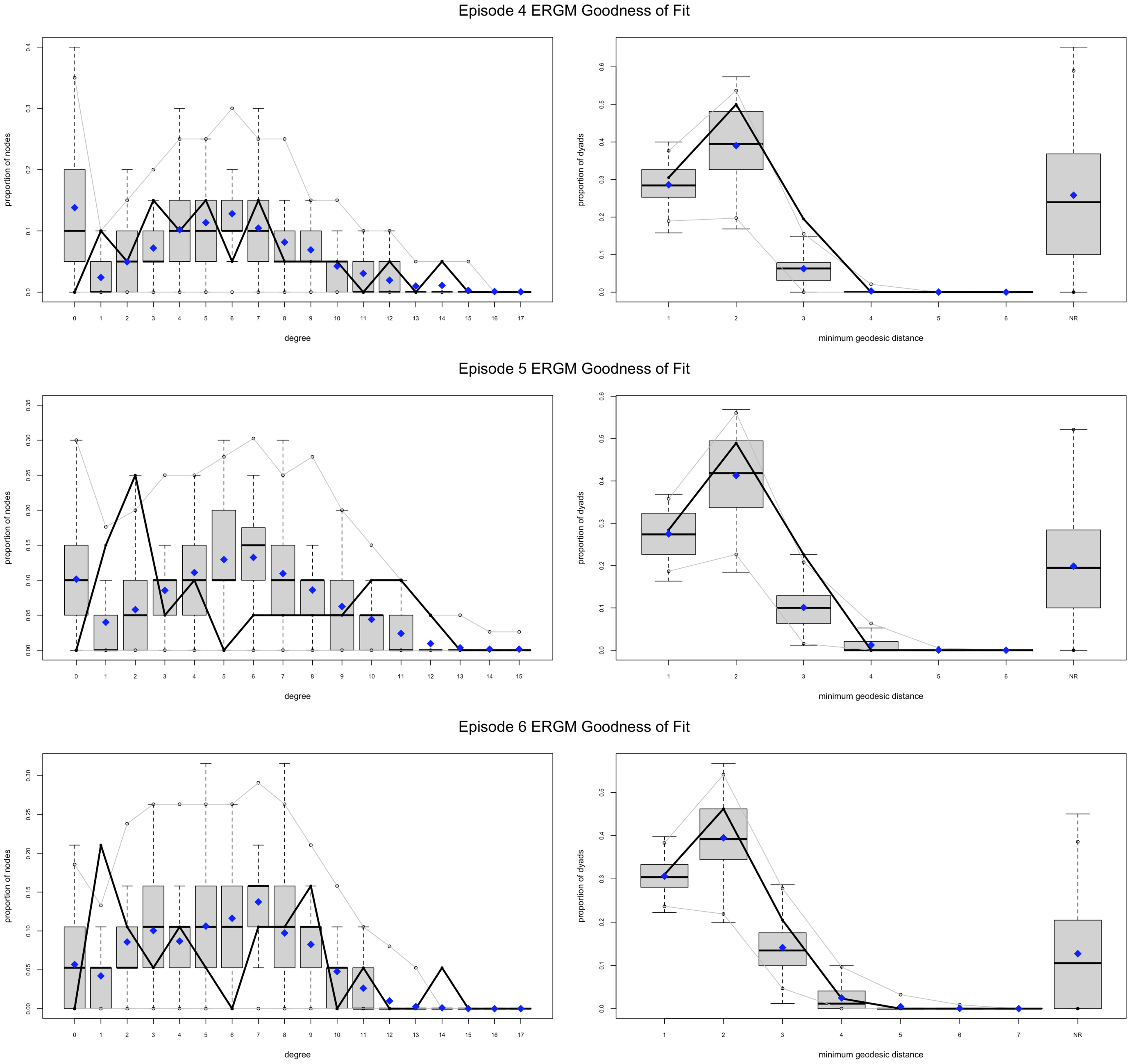
The hypothesis that being Force-sensitive increases the likelihood of being connected was tested with another θ added to the ERGM, this time measuring the effect that Force-sensitivity had on the probability of forming ties. Traditionally, this kind of θ would use `nodecov` in R, a measure of covariance between a given node's attribute and the number of ties it has. As there is only a binary possibility for this attribute - the character either 'has the Force', or does not - the `nodefactor` method was employed, which is similar to `nodecov`, but measures the "factor attribute effect" instead (statnet team, 2017).

The final hypothesis tested, that if characters meet, it is more likely that they will have a mutual acquaintance than not, relates to the idea of triadic closure, or transitivity. This means the θ added to the ERGM for this hypothesis is essentially a measure of character interaction forming nesting triangles in the network, which is analogous to groups of mutual acquaintances (Borgatti et al., 2018). In terms of these film networks, the measure is concerned with the likelihood that a new interaction between characters will result in the closing of a triangle of mutual interactions. This is measured using the geometrically weighted edgewise shared partner (GWESP) statistic, and the GWESP θ added to the ERGM is the final measure. Unlike the previous two however, the `gwesp` method in R also requires an argument for the decay weight (statnet team, 2017). As GWESP uses the amount of shared nodes to predict the likelihood of a tie forming a triad, each subsequent common node must be weighted less so as not to cause the model to degenerate into being fully connected. Equally, these common nodes should not be so severely discounted as to result in no triadic closures, and so a balance must be struck in order to have an acceptable ERGM. In this study a fixed decay rate of 0.5 proved successful for all three networks, rather than either a variable decay rate, or the more standard fixed 0.693.

With the additional θs and the mandatory `edges` - equivalent to the network's ties - parameter, the resulting ERGM summaries were used to look at whether the hypotheses were supported, or should be rejected in favour of the null, for each of the films. Before analysing these results though, the goodness of fit of the models should be verified to see how well the ERGMs have stayed within the bounds of the network they were attempting to model.

There are three terms to assess goodness of fit for ERGMs, and they are either at the node, edge, or dyad level. Importantly, only terms that have not been used as parameters to build the model are valid, as assessing the goodness of fit on an explicitly defined value would not give useful information. As the GWESP was used as one of the θs to specify the model, it is not possible to use the edge level term of edgewise share partners as a measure of goodness of fit, however the node and dyad level terms are usable.

The node level term for goodness of fit is the degree of the model networks, and the dyad level term is the geodesic distance. Both of these were used to verify the films' ERGMs, and the resulting box plots are seen in figure 4, below. In summary, the goodness of fit gave results that were quite acceptable. For degree, the original network observation was mostly either within, or close, to the values found in the model networks; sometimes it was even aligned with the median of those values (the blue diamonds in the box plots). The geodesic distance was also generally good, however, the final value in the box plots is the proportion of non-reachable nodes, which never featured in any of the original networks, but are sometimes present in the ERGMs. As such, the goodness of fit on the dyad level is not particularly strong.


Episode 4 ERGM Goodness of Fit

Episode 5 ERGM Goodness of Fit

Episode 6 ERGM Goodness of Fit

Figure 4 – Box plots showing the goodness of fit analysis of the three ERGMs, with degree on the left and minimum geodesic distance on the right. The bold line shows the original network's observed values, and the blue diamonds show the median value from the ERGMs

Results and Analysis

The first noteworthy finding is that the hypotheses do not hold equally among the films, and that there is both increasing and decreasing significance in terms of the tie-creating probabilities of these attributes. Also of note is that although all ERGMs are specified first with a measure of their edges - modelling the network's density, as in a Bernoulli graph - and that this was not related to any specific hypothesis, there are significantly fewer connections in all of the networks than were probable by chance. This may simply be a reflection of the realities of the films requiring characters to move within relatively limited circles, or could be representative of the sparsity found in real-life networks. Despite the fantastical setting of the films, it seems at least this aspect of the networks align with reality. Before analysing the results pertaining to the hypotheses though, the complete table of results from the ERGMs can be found in figure 5, below.

```
Episode 4 ERGM Results:
                      Estimate Std. Error MCMC % z value Pr(>|z|)
edges                  -6.7698     0.9946      0  -6.807  < 1e-04 ***
nodematch.Faction       1.1709     0.3419      0   3.425 0.000616 ***
nodecov.ForceSensitive  0.8358     0.2910      0   2.872 0.004074 **
gwesp.fixed.0.5         1.8824     0.4919      0   3.827 0.000130 ***
---------------------------------------------------------------------
Episode 5 ERGM Results:
                      Estimate Std. Error MCMC % z value Pr(>|z|)
edges                  -4.9517     0.8313      0  -5.956   <1e-04 ***
nodematch.Faction       0.8111     0.3326      0   2.439   0.0147 *
nodecov.ForceSensitive  0.3869     0.2288      0   1.691   0.0908 .
gwesp.fixed.0.5         1.4689     0.3354      0   4.380   <1e-04 ***
---------------------------------------------------------------------
Episode 6 ERGM Results:
                      Estimate Std. Error MCMC % z value Pr(>|z|)
edges                  -4.3997     0.9102      0  -4.834  < 1e-04 ***
nodematch.Faction       2.5912     0.4747      0   5.458   <1e-04 ***
nodecov.ForceSensitive  0.3276     0.2536      0   1.292   0.1964
gwesp.fixed.0.5         1.0480     0.3653      0   2.869  0.00412 **
---------------------------------------------------------------------
```

Figure 5 - Complete table of results from the three ERGMs

The first parameter of the ERGMs related to one of the hypotheses shows the effect that being in the same faction has on the likelihood of a tie forming. An increased likelihood would be an example of homophily, and was suspected to be present as it is something that would be intuitively unsurprising, given the frequency with which it occurs in real world networks (Borgatti et al., 2018). This effect was indeed found in all three films' networks, but although all of the positive correlations were significant, the correlation found for the second film (Episode 5) is less significant. This is something actually paralleled by the plot of the film, in which many of the characters actively choose to separate and question their loyalty to the factions they were previously pledged to. While many ultimately do not change factions, the lower significance of the effect in this film specifically makes sense, even if it doesn't result in an outright switch to a preference for heterophily.

Beyond recognising the significant influence of the homophily effect, it is somewhat difficult to categorise these ties as demonstrations of social selection or social effect, that is, if the driver behind the formation of ties is either: the character's interacted because they were similar, or they became similar because they interacted. As these networks are representations of interactions which are entirely the creations of filmmakers, the influences are likely to be both; bringing likeminded individuals into the factions throughout the first film is akin to social selection, and then choosing to stick with their factions in the final film after a journey of doubt and discovery is more a demonstration of social influence.

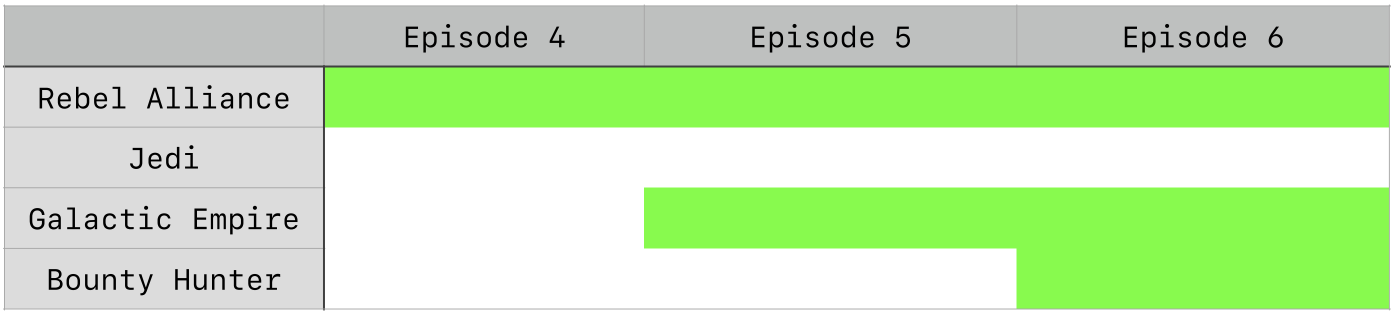
The next statistic relevant to the hypotheses is the influence of the character's Force-sensitivity on their likelihood of interacting with others. Recall from earlier that there are only two possible states of Force-sensitivity, and a positive correlation would indicate that being Force-sensitive increases the likelihood of a person making connections. This measure is the most variable of all the parameters, with the hypothesis only holding in the first film, as there is a continuous decrease in the significance of this parameter's explanatory power throughout the series.

The decreasing influence of a character's Force-sensitivity on their likelihood of making connections is again a corollary to the narrative of the films, and can be somewhat explained by them. In Episode 4 battle lines are drawn, alliances are formed, and characters are established. With many of the primary protagonists and antagonists falling into the small group of Force-sensitive characters, their almost inevitable intermingling with the wider cast at this point explains the Force's subtle influence on their connectedness. As the story progresses, the loyalties and factionalism of the character's becomes more emphasised, with most characters barely interacting outside of their factions, and with most of the subset of Force-sensitive characters mostly engaged in conflicts exclusively with each other, invested in their own tangential plot.

Finally, the GWESP statistic does show that there is an increased likelihood for connections to be transitive. The results are significant for all three films, which on reflection should not be hugely surprising given the general nature of both storytelling, and the specific Star Wars story itself. Essentially, there are a limited amount of characters possible to introduce throughout the course of the films, and from the point of view of an audience it makes sense to have characters that are previously known to the viewers make connections with other previously known characters. In this way the chemistry and group dynamics are consistently shifted and explored, creating a more interesting viewing experience. It should be noted, however, that as this is an undirected network there is no way to measure the reciprocity in these connections, only if there is a tendency towards transitivity or not (Borgatti et al. 2018).

Summary and Limitations

These networks are made up of only a subset of the characters that they could possibly be made up of due to the nature of ERGMs easily becoming degenerate. As such, some potentially interesting measures were not included. One which was not fully suitable was that of using differential homophily measures on the factions, to see if some factions were more homophilous than others. When trying this, there was a warning as to the potential lack of accuracy in the results, but nonetheless the significant results of the factions which featured more than two nodes are shown in figure 6 below.



	Episode 4	Episode 5	Episode 6
Rebel Alliance	🟩	🟩	🟩
Jedi			
Galactic Empire		🟩	🟩
Bounty Hunter			🟩

Figure 6 - Significant differential homophily results highlighted in green for relevant factions across the three films

It would also be interesting to use a stochastic actor oriented model for some characters in future studies, so that the individual journeys of some of the characters could be better explored and explained by the available methods. Also, expanding the existing method and hypotheses to the films that make up the rest of the saga could also provide illuminating results, as could generating more detailed and novel hypotheses.

The dataset itself could also benefit from some expansion. A more robust means of identifying the complete list of actors in the networks would be useful, perhaps including not just those characters which converse, but those which feature in the same location. The ‘value’ data in the original dataset (as described in the introduction) could also be incorporated into future analysis for further hypotheses formulation. But perhaps a better use of these values for the edges would be in creating a signed and/or directed network. As it stands it is not really possible to empirically test the balance of the network without having some measure of the various antagonistic and collaborative inter-factional interactions. With these additions it may well be possible to identify some influences akin to Heider’s theories on structural balance (Cartwright and Harary, 1956), as found within the networks generated from the narratives of these films.

Although this report chose an ultimately not entirely serious subject to use as the topic of its analysis, I believe it does do a reasonable job of demonstrating and illustrating some fundamental network effects. Whether it is the role that the writer’s decisions regarding the character’s social selection and influence plays on the homophilous nature of the networks, the correlation between a character’s attributes and their propensity towards making ties, or the frequency with which groups of characters create subgroups of mutual acquaintances - all of these phenomena are present in the Star Wars story to some extent.

Word Count: 2875

References

Bhatia, R. (2020) ‘Star Wars Social Network’, *Kaggle* [online] Available at: https://www.kaggle.com/ruchi798/star-wars?select=starwars-full-interactions-allCharacters-merged.json (Accessed: May 2021)

Borgatti, S.P., Everett, M.G. and Johnson, J.C. (2018). ‘Testing hypotheses’. *Analyzing social networks*. Sage. pp.125-148

Cartwright, D. and Harary, F. (1956). ‘Structural balance: a generalization of Heider's theory’. Psychological review. 63(5), p.277, available at: https://psycnet.apa.org/record/1957-06811-001 (Accessed: May 2021)

Garcia, A. (2018) ‘Fitting Into the Franchise: Texts, World Building, and the Possibilities of Creative Expansion’, Journal of Adolescent & Adult Literacy, 61(5), pp.585– 588, available online: https://ila.onlinelibrary.wiley.com/doi/abs/10.1002/jaal.722 (Accessed: May 2021)

McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001). ‘Birds of a feather: Homophily in social networks.’, Annual review of sociology, 27(1), pp.415-444.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007). ‘An introduction to exponential random graph (p*) models for social networks.’, Social networks, 29(2), pp.173-191, available online: https://doi.org/10.1016/j.socnet.2006.08.002 (Accessed: May 2021)

Statnet Team. (2017) ‘ergm-terms’, *ERGM R Lab* [online], available at: https://zalmquist.github.io/ERGM_Lab/ergm-terms.html (Accessed: May 2021)

Appendix - R Code

```
library('igraph')
library('GGally')
library('RColorBrewer')
library('statnet')
library('intergraph')
library('network')

#### Setup ####
# Read in the edge lists of the three films
edgL4 <- read.csv('SW4_links.csv')
edgL5 <- read.csv('SW5_links.csv')
edgL6 <- read.csv('SW6_links.csv')

# Turn them into networks
sw4Net <- as.network(edgL4, matrix.type = 'edgelist',
                    directed = FALSE, vertex.names = NULL)
sw5Net <- as.network(edgL5, matrix.type = 'edgelist',
                     directed = FALSE, vertex.names = NULL)
sw6Net <- as.network(edgL6, matrix.type = 'edgelist',
                     directed = FALSE, vertex.names = NULL)

# Read in the character attributes
charAtr4 <- as.matrix(read.table('SW4.txt', header = TRUE))
charAtr5 <- as.matrix(read.table('SW5.txt', header = TRUE))
charAtr6 <- as.matrix(read.table('SW6.txt', header = TRUE))

# Attach the attributes to the networks
sw4Net %v% 'Name' <- charAtr4[,2]
sw5Net %v% 'Name' <- charAtr5[,2]
sw6Net %v% 'Name' <- charAtr6[,2]

sw4Net %v% 'Faction' <- charAtr4[,3]
sw5Net %v% 'Faction' <- charAtr5[,3]
sw6Net %v% 'Faction' <- charAtr6[,3]

sw4Net %v% 'FactionId' <- as.numeric(charAtr4[,4])
sw5Net %v% 'FactionId' <- as.numeric(charAtr5[,4])
sw6Net %v% 'FactionId' <- as.numeric(charAtr6[,4])

sw4Net %v% 'Force' <- charAtr4[,5]
sw5Net %v% 'Force' <- charAtr5[,5]
sw6Net %v% 'Force' <- charAtr6[,5]

sw4Net %v% 'ForceId' <- as.numeric(charAtr4[,6])
sw5Net %v% 'ForceId' <- as.numeric(charAtr5[,6])
sw6Net %v% 'ForceId' <- as.numeric(charAtr6[,6])

sw4Net %v% 'ForceCol' <- charAtr4[,7]
sw5Net %v% 'ForceCol' <- charAtr5[,7]
sw6Net %v% 'ForceCol' <- charAtr6[,7]

# Make an igraph and matrix of all this for potential later use
sw4Ig <- asIgraph(sw4Net)
sw5Ig <- asIgraph(sw5Net)
sw6Ig <- asIgraph(sw6Net)

sw4Mat <- as.matrix(sw4Net)
sw5Mat <- as.matrix(sw5Net)
sw6Mat <- as.matrix(sw6Net)
#### Plotting ####
# Summarise and plot the networks with force and faction colours
summary(sw4Net)
ggnet2(sw4Net, mode = 'circle', color.palette = 'Pastel1', edge.color = 'grey',
       node.color = 'Faction', node.size = 20, legend.position = 'none',
       label = 'Name', label.color = 'ForceCol', label.size = 7)
```

```
summary(sw5Net)
ggnet2(sw5Net, mode = 'circle', color.palette = 'Pastel1', edge.color = 'grey',
       node.color = 'Faction', node.size = 20, legend.position = 'none',
       label = 'Name', label.color = 'ForceCol', label.size = 7)

summary(sw6Net)
ggnet2(sw6Net, mode = 'circle', color.palette = 'Pastel1', edge.color = 'grey',
       node.color = 'Faction', node.size = 20, legend.position = 'none',
       label = 'Name', label.color = 'ForceCol', label.size = 7)

#### Hypothesis Test ####
# ERGMs Let's Go, add diff=TRUE to nodematch for differential homophily
sw4ERGM <- ergm(sw4Net ~ edges
                + nodematch('Faction')
                + nodefactor('Force')
                + gwesp(decay = 0.5, fixed = TRUE))

sw5ERGM <- ergm(sw5Net ~ edges
                + nodematch('Faction')
                + nodefactor('Force')
                + gwesp(decay = 0.5, fixed = TRUE))

sw6ERGM <- ergm(sw6Net ~ edges
                + nodematch('FactionId')
                + nodefactor('Force')
                + gwesp(decay = 0.5, fixed = TRUE))

#### Goodness of fit ####
# Do GoF for each and then plot them, simples
mcmc.diagnostics(sw4ERGM)
gof4 <- gof(sw4ERGM ~ degree + distance)
par(mfrow = c(2, 2))
plot(gof4, main = 'Episode 4 ERGM Goodness of Fit')

mcmc.diagnostics(sw5ERGM)
gof5 <- gof(sw5ERGM ~ degree + distance)
par(mfrow = c(2,2))
plot(gof5, main = 'Episode 5 ERGM Goodness of Fit')

mcmc.diagnostics(sw6ERGM)
gof6 <- gof(sw6ERGM ~ degree + distance)
par(mfrow = c(2,2))
plot(gof6, main = 'Episode 6 ERGM Goodness of Fit')

#### Get the results, we done ####
summary(sw4ERGM)

summary(sw5ERGM)

summary(sw6ERGM)
```