Nabeel Tariq

Game of Thrones Network Analysis

The book series *A Song of Ice and Fire* and the television show *Game of Thrones* based on the books have become very popular in the last couple of years. They feature an enthralling storyline and, perhaps more importantly, a very large and rich cast of characters. The characters are interconnected in many ways with constantly changing relationships and allegiances that drive the story and keep us reading and watching. The story was first created in book form, with the tv series created later based on the books. Since the tv series exploded in popularity, and the books take so long to write, the tv series has lately moved beyond the storyline of the books in existence and had to come up with its own plot. With seven seasons already released, the series has announced that the eighth season will be the final one and thus is expected to wrap up much of the story.

We chose to explore these character relationships using the tools of graph theory and network analytics to see what the numbers would tell us about the nature of the character network and how it ties in with the story. One of the more interesting aspects of the stories in this series is propensity for main characters to die unexpectedly. This is a very different approach from most TV series', and possibly one of the reasons for the show's strong following. One question in particular that we thought would be fascinating to answer was whether or not the character network could be used to predict which of the main characters were likely to live and die in the final season of the tv show. Many of the main characters are competing for power in the final season and it is highly likely some of them will meet a violent end. There is much speculation on the internet as to how this might play out, so we tried to use network analysis to make our own predictions. In the end we were able to make a few projections, and learned a lot about the characters' interactions along the way.

Data and Strategy

To make our predictions we utilized data gathered from the books and compiled by Professor Andrew Beveridge at Macalester College. The character connections were based on having their names mentioned within 15 words of each other in the text, with weighting to account for the number of repeat mentions. One effect of these weights is that they also help denote main characters as some characters weights are high across most of their edges simply by being generally mentioned more in the text than other characters. An important thing to remember is that we are drawing parallels between the books and events that occur in the TV series. For the most part the TV series follows the first 3 books, but diverges a bit for the 4th and 5th books. Thus, the different plotlines (and hence interactions) may result in model and metric ambiguity. We believe, however, that the analysis value added is far greater than the data limitation.

Since the TV series reaches a much wider audience than the books, it might be interesting to see if any onscreen metrics are impacted by the networks and interactions in the books. More importantly, it may be extremely interesting to see how network metrics relate to on-screen time and how that relates to a character's likelihood to die. The major metric we were able to source from the TV series was character time on-screen, in minutes, with the prediction of how much screen time a character would be expected to get as a driver of whether or not they were likely to be killed off.

Another way we looked at predicting character death was to fit a binary classifier to the data set. This posed an interesting challenge as the nodes in our network are not all necessarily equal or interchangeable. There were certain characters we were particularly interested in making predictions on and certain characters that were very important that the classifier was fit to, making pulling out a random validation or test sample impractical. Pulling any main character out of the network would significantly affect the number of nodes with their similar characteristics and the fact that live characters are still alive is salient to how the model should fit to feature sets similar to theirs. The size of the dataset was also somewhat limited since we wanted to prune the network down to a size practical for some manual data entry to integrate outside data. In the end, we decided to fit a parametric model to the entire set and simply evaluate the fitted values for characters of interest who were still alive.

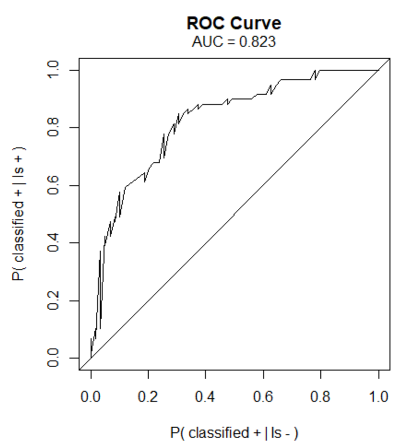
We generated a variety of feature sets to use in evaluating the death of characters. Features could be divided into two categories: network-level variables and non-network variables. For network-level variables we used node-focused variables including closeness, betweenness, and degree centrality, as well as the Burt's Constraint value for structural holes, the value of the maximal k-core the character belongs to, and the eigenvector value for the character. Edge-level values like weight were also taken into consideration. Non-network variables were pulled from various sites on the internet. We utilized screen time, gender, and loyalty to the various factions in the show. Using screen time was valuable for certain analyses but couldn't be used everywhere as the number of characters that we had a record of screen time for was limited to about 80.

Analysis & Findings

The parametric classification model produced interesting results that we believe are useful for predicting who might live and die. We used a logistic regression curve to fit the dataset with a binary value for dead as the target. The model produced the following summary statistics:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.51299  -0.88033   0.00003   0.88929   1.88582

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.596e+00  3.083e+00   0.842   0.3999
dgr              -1.289e+00  5.043e-01  -2.555   0.0106 *
btw               5.427e-03  4.389e-03   1.236   0.2163
cls              -2.610e+03  1.670e+03  -1.563   0.1180
burts             5.117e-02  2.019e+00   0.025   0.9798
loyaltyCrown     -1.164e+00  1.119e+00  -1.040   0.2983
loyaltyFrey       1.645e+01  2.785e+03   0.006   0.9953
loyaltyGreyjoy   -1.914e+01  1.894e+03  -0.010   0.9919
loyaltyLannister -3.581e-01  1.405e+00  -0.255   0.7989
loyaltyMartell   -1.228e-01  1.757e+00  -0.070   0.9443
loyaltyminor     -2.220e+00  1.352e+00  -1.642   0.1006
loyaltyother     -1.768e+00  1.249e+00  -1.415   0.1570
loyaltyStark      4.961e-01  1.221e+00   0.406   0.6846
loyaltyTargaryen  1.161e-01  1.211e+00   0.096   0.9236
loyaltyWatch     -5.027e-01  1.259e+00  -0.399   0.6898
genderM          -5.452e-02  5.626e-01  -0.097   0.9228
core              1.035e+00  7.684e-01   1.347   0.1780
eigen             1.424e+01  5.683e+00   2.506   0.0122 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 163.58  on 117  degrees of freedom
Residual deviance: 122.22  on 100  degrees of freedom
AIC: 158.22

Number of Fisher Scoring iterations: 16
```

As you can see only one of the predictors produced a statistically significant coefficient, but removing values did little to improve significance. Plotting the ROC curve and AUC value produced the following:



The curve suggests that the model is a decent predictor, though imperfect. An AUC value of 0.823 was actually better than we expected given the variance in the data and small sample size. Most importantly, the model produced fitted values for each character, allowing us to produce a list of characters that are still alive ranked by their likeness to dead characters which we used as a prediction for likeliness to die.

table.names	table.fitted
Varys	0.957
Edmure-Tully	0.921
Beric-Dondarrion	0.728
Bronn	0.714
Sansa-Stark	0.690
Wyman-Manderly	0.676
Missandei	0.604
Daario-Naharis	0.591
Illyrio-Mopatis	0.591
Jorah-Mormont	0.591

table.names	table.fitted
Gendry	0.145
Gerris-Drinkwater	0.141
Bran-Stark	0.130
Belwas	0.123
Cersei-Lannister	0.062
Tyrion-Lannister	0.031
Samwell-Tarly	0.007
Daenerys-Targaryen	0.000
Jon-Snow	0.000
Euron-Greyjoy	0.000
Victarion-Greyjoy	0.000
Theon-Greyjoy	0.000
Asha-Greyjoy	0.000

While it seems unlikely that Daenerys, Jon Snow, and Cersei will all survive the last season it is interesting to see how they landed. The high likelihood deaths may be more interesting as the list contains some characters one might not expect, but whose death would certainly keep the story interesting.

We also explored other relationships between our data and characters and studied the differences between living and dead characters. One thing we evaluated was TV screen time as a proxy for character lifetime to see if anything interesting came out. We pruned the network to around 70 characters by filtering out low edge weights to focus our data. This removed the effect of minor characters from the analysis as we were only concerned with our ability to predict outcomes for main characters. We then separated our sample into living and dead characters and used a linear regression to evaluate our metrics' impact on screen time.

Name	closeness	degree	betweenness	eigen	coreness	Allegiances	minutes	Average.time.per.season
Jon-Snow	0.000560224	114	6.06E+04	0.57314669	12	Night's Watch	338.25	48.321429

Example of the feature vector for one of the most important characters

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.18427   26.57085  -2.265  0.03990 *
closeness    -8.54122    3.51714  -2.428  0.02923 *
degree        2.01235    0.54052   3.723  0.00227 **
betweenness  -0.07165    0.11311  -0.633  0.53667
eigen         0.94353    0.55910   1.688  0.11364
coreness     -2.82837    1.36651  -2.070  0.05745 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5186 on 14 degrees of freedom
Multiple R-squared:  0.644,	Adjusted R-squared:  0.5168
F-statistic: 5.065 on 5 and 14 DF,  p-value: 0.007383
```

Model summary for dead characters

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.2419    22.7528   0.582   0.5635
closeness     1.4815     3.0204   0.491   0.6262
degree        1.8289     0.3679   4.971 1.01e-05 ***
betweenness  -0.1849     0.0981  -1.884   0.0660 .
eigen        -0.2039     0.3829  -0.533   0.5970
coreness     -1.2761     0.5409  -2.359   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6287 on 45 degrees of freedom
Multiple R-squared:  0.668,	Adjusted R-squared:  0.6311
F-statistic: 18.11 on 5 and 45 DF,  p-value: 8.234e-10
```

And for living characters

Degree is a strong predictor for screen time for all characters, but closeness has a distinctly different effect for dead characters than living ones. This may suggest that characters that have racked up significant screen time but have low closeness are more likely to die. Another takeaway is that these metrics are more significant when predicting screen time for living characters, possibly indicating that the characters who deviate most from the model of screen time following network metrics are easier to kill off. These factors so far support the idea that Jon Snow and Tyrion are likely to live, and Varys is highly at risk.

A key part of network analysis and modelling is analyzing the change in ties between nodes over time. For the Game of Thrones book series, the fast-moving plotlines and world-changing events may have significantly impacted ties and relationships as we move through the books, which makes it interesting to measure any changes in factors influencing death over time. To this end, we looked at edge lists separated by book and treated each book as a separate time period (though combining books 4 and 5 based on Professor Beveridge’s data. We again filtered edges based on weights to come up with the key characters in each book, with roughly 90 characters in our subset. In this analysis it is sometimes easy to see how events in the story translate to network level metrics.

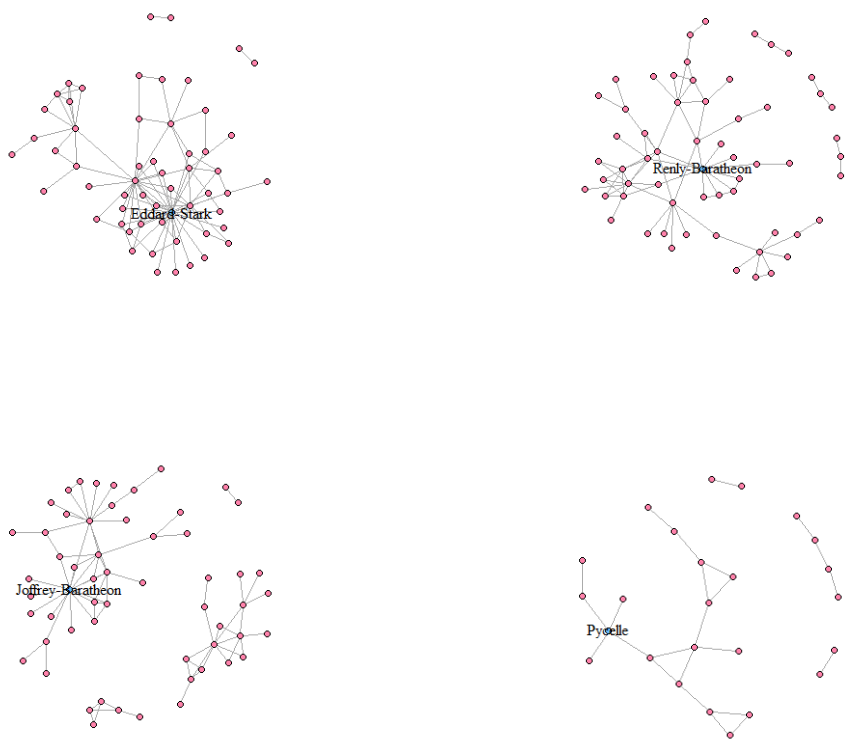
Parameter	Book 1	Book 2	Book 3	Book 4 & 5	Books Combined
EigenVector Centrality	-6.16E+00	1.49E+01	-7.70E-01	1.19E+01	-3.83E-01
P - Value	0.03	0.142	0.7164	0.192	0.7987
Closeness Centrality	-6.33E+03	5.23E+02	-3.04E+03	-8.52E+03	-4.93E+03
P - Value	0.0957	0.88	0.3067	0.189	0.1102

For example, the eigenvector and closeness centrality measures for each character are the only significant predictors of death in the first book, which makes sense given the deaths of Ned Stark and King Robert. Both were well connected characters with many ties to other well connected characters. We can also see that this relationship does not hold up through the rest of the books.

Parameter	Book 1	Book 2	Book 3	Book 4 & 5	Books Combined
Burt's Coefficient	-4.14E+00	-3.61E-01	-5.11E+00	2.16E+00	-2.98E+00
P - Value	0.179	0.888	0.1126	0.709	0.3049

Burt’s coefficient indicates structural holes. It is an interesting factor because while certain centrality measures change as the network grows (average closeness declines and average degree rises naturally) the average value of Burt’s across the network stays relatively consistent from book to book. We can see here it is generally negatively correlated with likelihood of death, especially throughout the first 3 books. Being a structural hole (having a lower Burt’s value) may result in a higher probability of death, especially given that this could often apply to leaders and armies. An interesting case could be the Red Wedding, where Robb Stark may have been the only point of connection to his army. This would seem to contradict our earlier list of people likely to survive that included multiple Lannisters as well as Jon Snow and Daenerys, but one interesting character high on that list and with a low Burt’s value is Bran Stark. Perhaps his power as the Three-Eyed-Raven won’t be able to save him in the end.

An interesting final analysis focused on the networks of already deceased characters and the performance of simple visual analysis. We graphed the network of dead characters and noted the character with the highest degree centrality among the dead. The names that come up are Ned Stark for book 1, Renly Baratheon for book 2, Joffrey in book 3, and Maester Pycelle in book 4. The first three characters were some of the most prominent to die in each book, suggesting that perhaps the best indicator of who will die within a season (analog for a book) is their number of connections to other dead in that season. Perhaps we can predict which major character will die by the end of the final season by analyzing their connections to minor characters who pass away during the season.



Results and conclusions

In summary, using network analytics provides an interesting way to attempt to get ahead of the speculation around the future of everyone’s favorite TV series in the last few years. While there is no single simple relationship between a network factor and who will die on the show, the network metrics can provide interesting insights. We used one comprehensive parametric model to predict how likely living characters are to join the dead, but this is only one way to speculate about who might be killed off. We also looked at a variety of relationships between individual factors and a character's longevity to further reinforce predictions.

Based on our research, it seems likely that two principle characters few would have expected, Varys and Sansa Stark, will meet their end. They pop up high in the prediction model and also share a low Burt's value, though they diverge in their closeness values. We also believe that it is unlikely we will see too many deaths in some of the major warring characters, specifically the Lannisters, Daenerys, and Jon Snow. While it's unlikely they will all escape alive given their moves toward confrontation, they each pop up high on the survival probability. Perhaps this indicates that defeat won't necessarily represent a bloodbath, though we're sure the producers are looking forward to killing off some less important characters along the way.