Analysis of Death Across Shakespeare's Plays and Genres

Nathan Bickel

4/19/2022

Introduction

William Shakespeare’s 37 plays fall into three generally distinct categories: tragedies, histories, and comedies. Tragedies usually involve the main characters suffering great misfortune, while comedies are more light-hearted and usually feature happy endings. Since histories are based on historical events, they can combine elements of both, with characters experiencing good and bad fortune. As such, death is treated quite differently across the genres, as it is common in tragedies for the main character to die and quite uncommon in comedies. This report aims to explore the differences in death among the three genres of Shakespeare’s plays by comparing the characters who die to the characters who do not using various metrics. Specific attention will be paid to whether death in histories is more comparable to tragedies or comedies, and whether there are outliers within genres that deviate from how the genre treats death as a whole. Investigating the structure of death in Shakespeare’s plays will likely reveal patterns that reflect underlying similarities and differences between genres, and paying particular attention to outliers may show how those plays compare to the genre they are written in.

Data and Methods

In order to compare the characters and deaths across different categories, they first need to be sorted into genres. This was done using the character, play, and speech metadata: each play is categorized as a comedy, tragedy, or history, each character is associated with a unique play, and each speech is associated with a character. While there are some small issues with this (not every character appears in only one play, some speeches are said by more than one character, etc.), they only affect a small portion of the corpus and should not significantly affect the results.

The first way the data were analyzed was creating a pie graph that displays the breakdown of dead and alive characters in each genre for a total of six slices. Then, within each genre, the percent of characters that die was calculated. Next, the raw counts of number of characters that die in each play was calculated (for example, there are eight characters that die in *Hamlet*) and the variation between genres was considered. This aims to examine the difference between volume of death in the three genres, and to begin to consider whether any plays seem to be outliers for their genre.

Next, a bar plot was created using the character metadata to track the percentage of deaths over time in each genre. For each act, the percentage of deaths that occur in a given genre in that act is displayed. For example, 0/59 deaths in tragedies occur in the first act while 3/42 deaths in histories occur in the first act, so 0% and 7.14% are displayed for tragedies and histories respectively in Act 1. Since the beginning acts serve a different function than the ending acts, this chart can give insight into how deaths progress in different genres and how they may serve different functions.

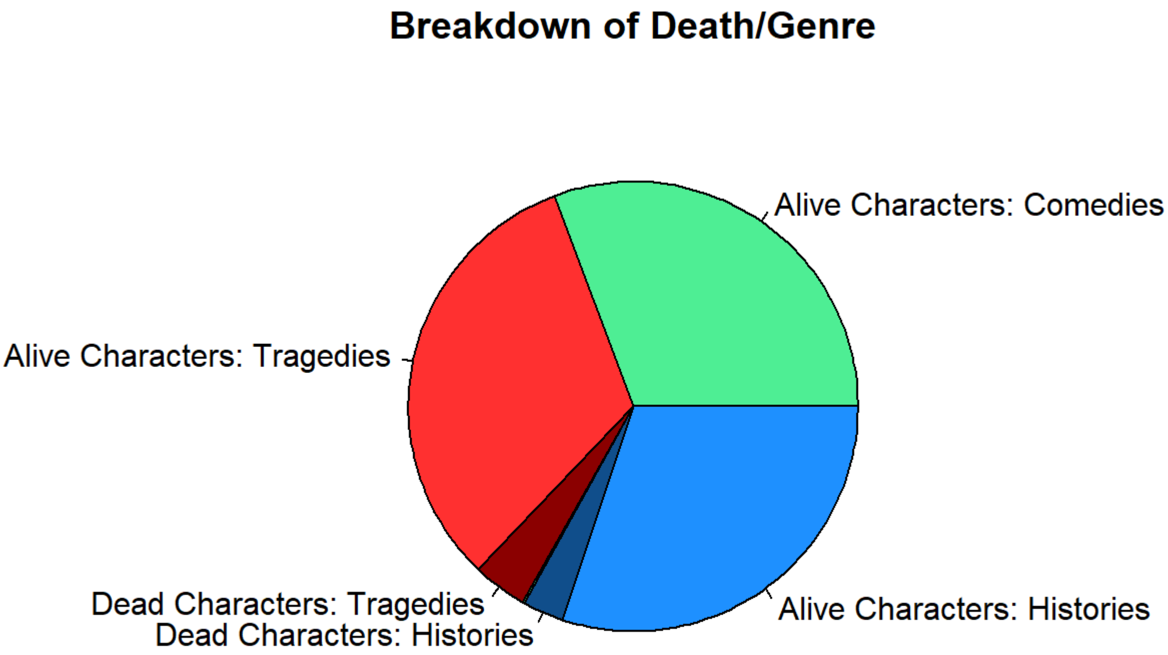
After calculating this fairly basic information about the deaths themselves, two networks was created to learn more information about the dead characters’ function in the plot throughout the plays. The first network was called **g_spch**, and consisted of 1431 vertices and 29812 edges. The vertices were all associated with characters, and the edges were related to speaking: two vertices were connected with an edge if the character associated with the first vertex spoke and then the character associated with the second vertex spoke directly after in the same scene. Thus, an edge usually represents a conversation with a character, and thus a connection of sorts (albeit a rather tenuous one). The second network was called $\mathbf{g\_relt}$, and consisted of 602 vertices and 792 edges. The vertices were again associated with characters, but the edges were related to a more well-defined relationship. Using the character metadata, a relationship between character A and character B is included as an edge in $\mathbf{g\_relt}$ if one or more of the following is true: A is B's spouse, A is B's sibling, A is B's servant, A is the child of B, or A kills B. While these are obviously quite different relationships, it represents a more substantial and important relationship than simply talking to another character, which is why there are so many fewer edges than in $\mathbf{g\_spch}$.

To compare how these networks vary across genres, six undirected subgraphs were made for each genre that included only connections from plays in the genre called $\mathbf{g\_spch\_trag}$, $\mathbf{g\_spch\_hist}$, $\mathbf{g\_spch\_comd}$, $\mathbf{g\_relt\_trag}$, $\mathbf{g\_relt\_hist}$, and $\mathbf{g\_relt\_comd}$. In order to examine how characters who die in these networks are different from characters in general, the degree for each vertex was calculated. Degree is simply defined as the number of edges connected to a vertex, so a character having a high degree in $\mathbf{g\_spch}$ indicates that they talk to many other characters, while a high degree in $\mathbf{g\_relt}$ indicates that they have a relationship with many other characters. In order to compare characters to the norm, the mean degree in each subgraph was found and then each degree was divided by the mean. Thus, a vertex having a value of 1.0 would mean that their degree is typical for the characters in the genre, while a vertex having a value of 5.0 would mean that their degree is 5 times as high as the typical character. Then, for both $\mathbf{g\_spch}$ and $\mathbf{g\_relt}$, a box plot was created for these ratio degrees only for dead characters in each genre. Since the plot shows the distribution of degree ratio for each dead character, if the median is greater than 1.0, it would suggest that the dead characters tend to have a higher degree than characters in general in the subnetwork, and thus tend to be more important than the average character. Additionally, since the box plots are side by side for each network, the distributions for dead characters can be compared across genre, which can give insight into whether dead characters tend to be more important compared to the mean in one genre than they are in another genre.

A similar technique was then conducted for the measure of betweenness. For a given vertex, betweenness is the number of shortest paths between any two vertices in the network that contain the vertex. So if a vertex has a high betweenness, it acts as a sort of “middle man” between many other points and is important in the network staying together. So in $\mathbf{g\_spch}$, a vertex with high betweenness may be someone who relays information between many different people, and in $\mathbf{g\_relt}$ it may be someone who connects two families. Thus, characters with high betweenness can be quite important, so it makes sense to do a similar calculation as with betweenness. The same procedure was carried out–the mean betweenness was found and then divided by the betweenness of every vertex, and then boxplots were made for each genre in each network. Similar reasoning then also applies here: if the median in a genre of the dead characters as shown in the boxplot is greater than 1.0, then it suggests that characters who end up dying tend to serve more important roles in connecting the network than the average character in the network, and the boxplots being shown side by side can show the differences in how this disparity plays out over genre.

Analysis

```
tragedies = as.matrix(play_metadata)[which(play_metadata$GENRE == "TRAGEDY"),1]
histories = as.matrix(play_metadata)[which(play_metadata$GENRE == "HISTORY"),1]
comedies = as.matrix(play_metadata)[which(play_metadata$GENRE == "COMEDY"),1]
dead_chars = character_metadata[which(character_metadata[,14]=="Y"),2]
alive_chars = character_metadata[c(1:1516)[-which(character_metadata[,14]=="Y")],2]
dead_chars_trag = dead_chars[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% tragedies)]
dead_chars_hist = dead_chars[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% histories)]
dead_chars_comd = dead_chars[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% comedies)]
alive_chars_trag = alive_chars[which(character_metadata[c(1:1516)[-which(character_metadata[,14]=="Y")],1] %in% tragedies)]
alive_chars_hist = alive_chars[which(character_metadata[c(1:1516)[-which(character_metadata[,14]=="Y")],1] %in% histories)]
alive_chars_comd = alive_chars[which(character_metadata[c(1:1516)[-which(character_metadata[,14]=="Y")],1] %in% comedies)]
pie(c(length(alive_chars_comd),length(alive_chars_trag),length(dead_chars_trag),length(dead_chars_comd),length(dead_chars_hist),length(alive_chars_hist)),labels=c("Alive Characters: Comedies","Alive Characters: Tragedies","Dead Characters: Tragedies","","Dead Characters: Histories","Alive Characters: Histories"),main="Breakdown of Death/Genre",col=c("seagreen2","firebrick1","darkred","seagreen4","dodgerblue4","dodgerblue1"))
```



```
cat(paste("Percentage of Characters that Die in Tragedies:",round(100*length(dead_chars_trag)/(length(dead_chars_trag)+length(alive_chars_trag)),2),"\nPercentage of Characters that Die in Histories:",round(100*length(dead_chars_hist)/(length(dead_chars_hist)+length(alive_chars_hist)),2),"\nPercentage of Characters that Die in Comedies:",round(100*length(dead_chars_comd)/(length(dead_chars_comd)+length(alive_chars_comd)),2)))
```

```
## Percentage of Characters that Die in Tragedies: 10.81
## Percentage of Characters that Die in Histories: 8.98
## Percentage of Characters that Die in Comedies: 0.64
```

The pie graph shows the percentage of alive and dead characters across genres. As expected, the highest percentage of deaths occurs in the tragedies while an almost negligible percentage of deaths occurs in comedies (the unlabelled green slice that represents dead characters in comedies between the dark blue and dark red slice is hardly visible). Also as expected, the death percentage in histories is somewhere in the middle between the two, but it seems to more closely resemble the death toll in tragedies than in comedies: the percentage difference is 1.83% between tragedies and histories and 8.34% between histories and comedies. This makes some sense—death is a large part of history, but feels out of place to include often in light-hearted comedies.

```
trag_deaths = vector("integer",length(tragedies))
hist_deaths = vector("integer",length(histories))
comd_deaths = vector("integer",length(comedies))
names(trag_deaths) = tragedies
names(hist_deaths) = histories
names(comd_deaths) = comedies
for (r in 1:1516) {
  if (!is.na(character_metadata[r,14]) & character_metadata[r,14]=="Y") {
    play = as.matrix(character_metadata)[r,1]
    if (play %in% tragedies)
      trag_deaths[play] = trag_deaths[play] + 1
    else if (play %in% histories)
      hist_deaths[play] = hist_deaths[play] + 1
    else if (play %in% comedies)
      comd_deaths[play] = comd_deaths[play] + 1
  }
}
print(trag_deaths)
```

```
## Ant Cor Ham  JC  Lr Mac Oth Rom Tim Tit Tro
##   6   1   8   7   6   7   4   5   1  12   2
```

```
print(hist_deaths)
```

```
## 1H4 1H6 2H4 2H6 3H6  H5  H8  Jn  R2  R3
##   2   3   5   5   8   3   0   6   3  10
```

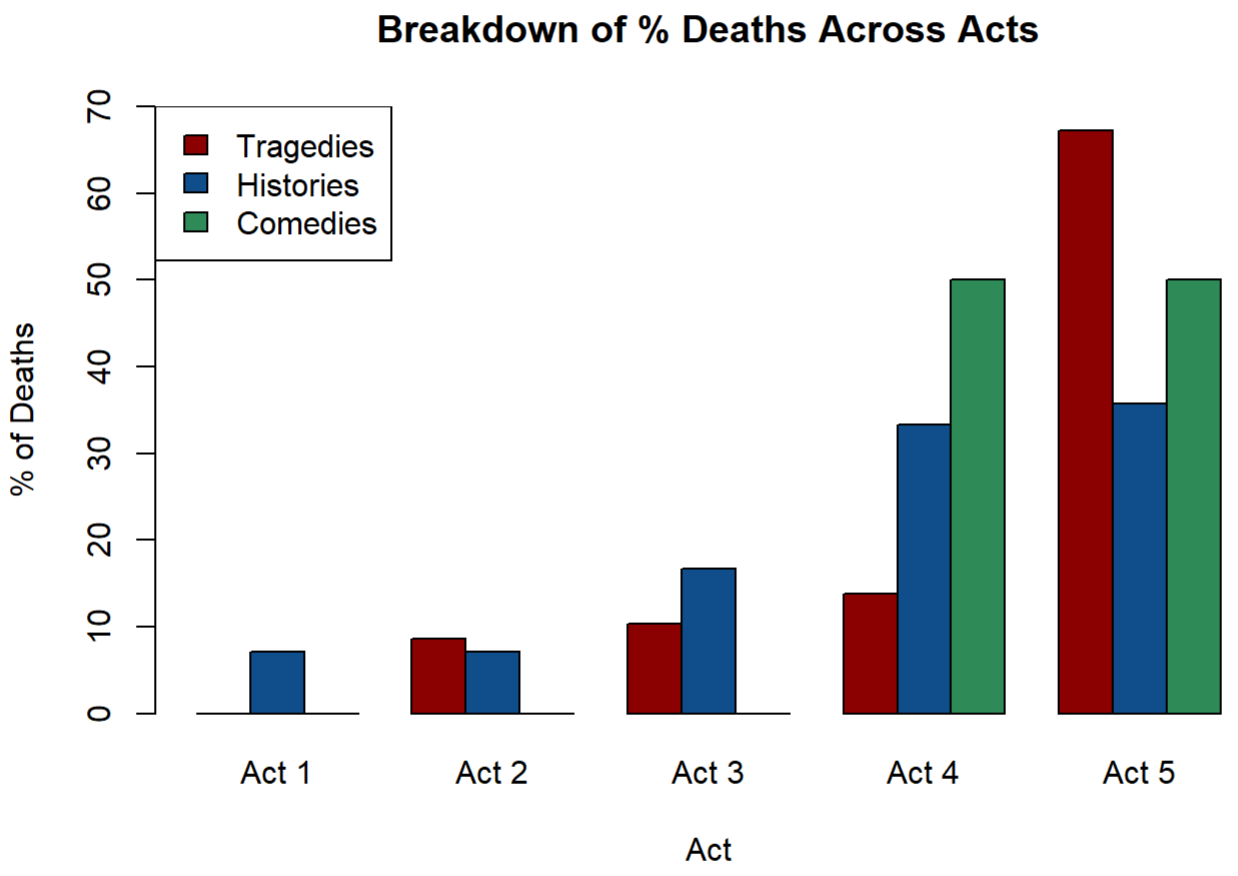
```
print(comd_deaths)
```

```
## Ado AWW AYL Cym Err LLL  MM MND  MV Per Shr TGV Tmp  TN Wiv  WT
##   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   1
```

In order to examine specific plays in each genre, the raw death counts were calculated for each play: the first row is tragedies, the second is histories, and the third is comedies. Notably, every tragedy contains at least one death, and most contain multiple with the median being six deaths. Most histories contain at least one death, but there seem to be fewer than tragedies: the median deaths is four rather than six, and some plays like *Henry the 8th* and *Henry the 4th, Part I* contain relatively few deaths. Thus, these two plays are more similar to the death toll in comedies than in tragedies, whereas plays like *Richard III* and *Henry the 6th, Part III* have a more similar death toll to tragedies. Finally, almost no deaths occur in comedies, with the exception of *Cymbeline* with two deaths and *Winter's Tale* with one.

These borderline cases in terms of death toll seem to be indicative that the plays are rather unique in their genres in general. For example, *Cymbeline*, the comedy in which two thirds of the deaths across the genre occur, was included in the First Folio under the title *The Tragedy of Cymbeline*. After one of the characters, Cloten, is killed, Belarius says “What hast thou done? […] We are all undone,” (IV.ii.153,161), which rather resembles the dismay characteristic of the end of tragedies. While most today consider the play to generally be a comedy, it is sometimes called a “tragicomedy” to acknowledge the tragic aspects, and it is somewhat unique in the genre of comedy. Thus, it is important to remember that since the majority of deaths in the genre come from this play, the results for comedies may be somewhat uncharacteristic of comedy since they come from a very small sample size and largely from a play that has tragic aspects. Another example is *Henry the 4th, Part I*: while the play is a history that describes the death of important characters, it also has many comedic aspects with the relationship between Prince Hal and Falstaff, and this comic aspect is mirrored by its lower death count than most of the rest of the histories.

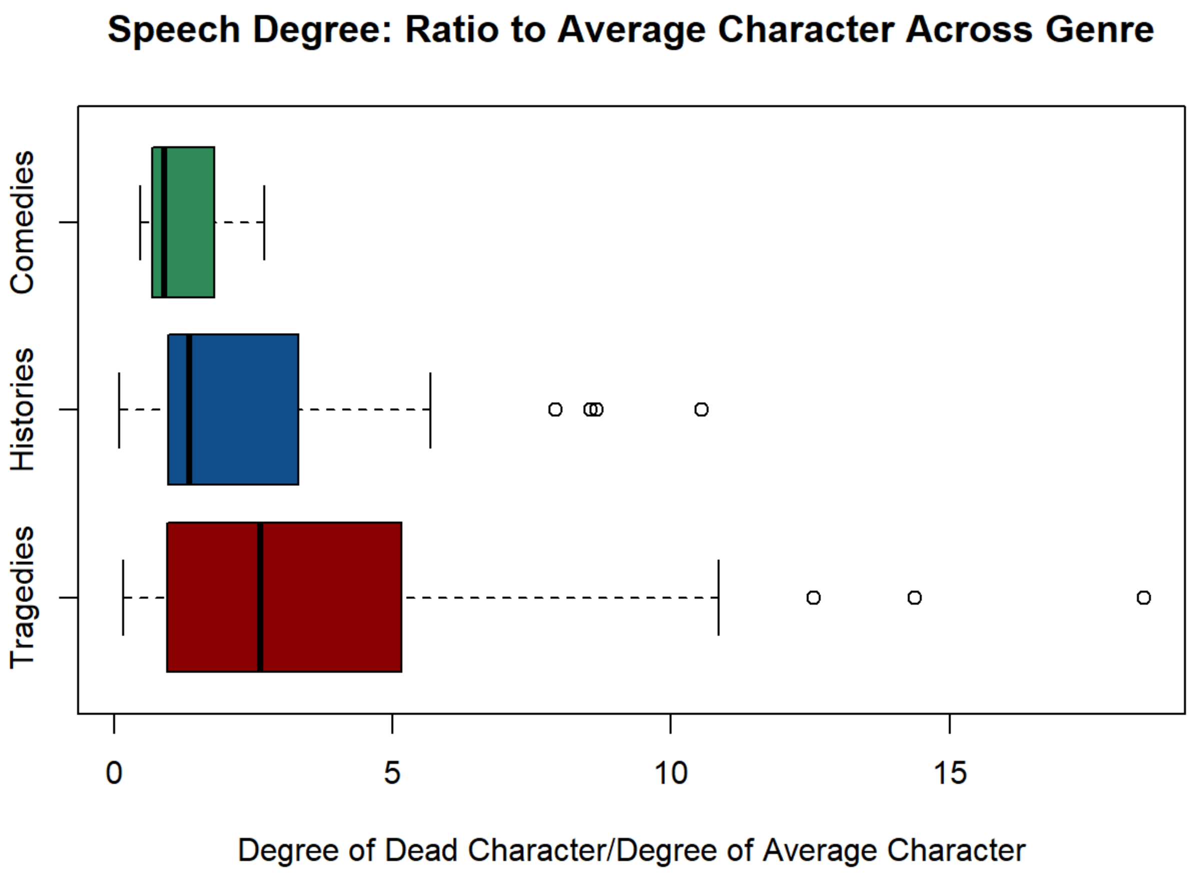
```
death_acts = matrix(0,3,5)
for (i in 1:1516) {
  if (character_metadata[i,2] %in% dead_chars_trag)
    death_acts[1,character_metadata[i,15]] = death_acts[1,character_metadata[i,15]] + 1
  else if (character_metadata[i,2] %in% dead_chars_hist)
    death_acts[2,character_metadata[i,15]] = death_acts[2,character_metadata[i,15]] + 1
  else if (character_metadata[i,2] %in% dead_chars_comd)
    death_acts[3,character_metadata[i,15]] = death_acts[3,character_metadata[i,15]] + 1
}
death_acts[1,] = 100*death_acts[1,]/sum(death_acts[1,])
death_acts[2,] = 100*death_acts[2,]/sum(death_acts[2,])
death_acts[3,] = 100*death_acts[3,]/sum(death_acts[3,])
barplot(death_acts,main="Breakdown of % Deaths Across Acts",xlab="Act",ylab="% of Deaths",names.
arg=c("Act 1","Act 2","Act 3","Act 4","Act 5"),col=c("darkred","dodgerblue4","seagreen4"),beside
=T,ylim=c(0,70))
legend("topleft",c("Tragedies","Histories","Comedies"),fill=c("darkred","dodgerblue4","seagreen
4"))
```



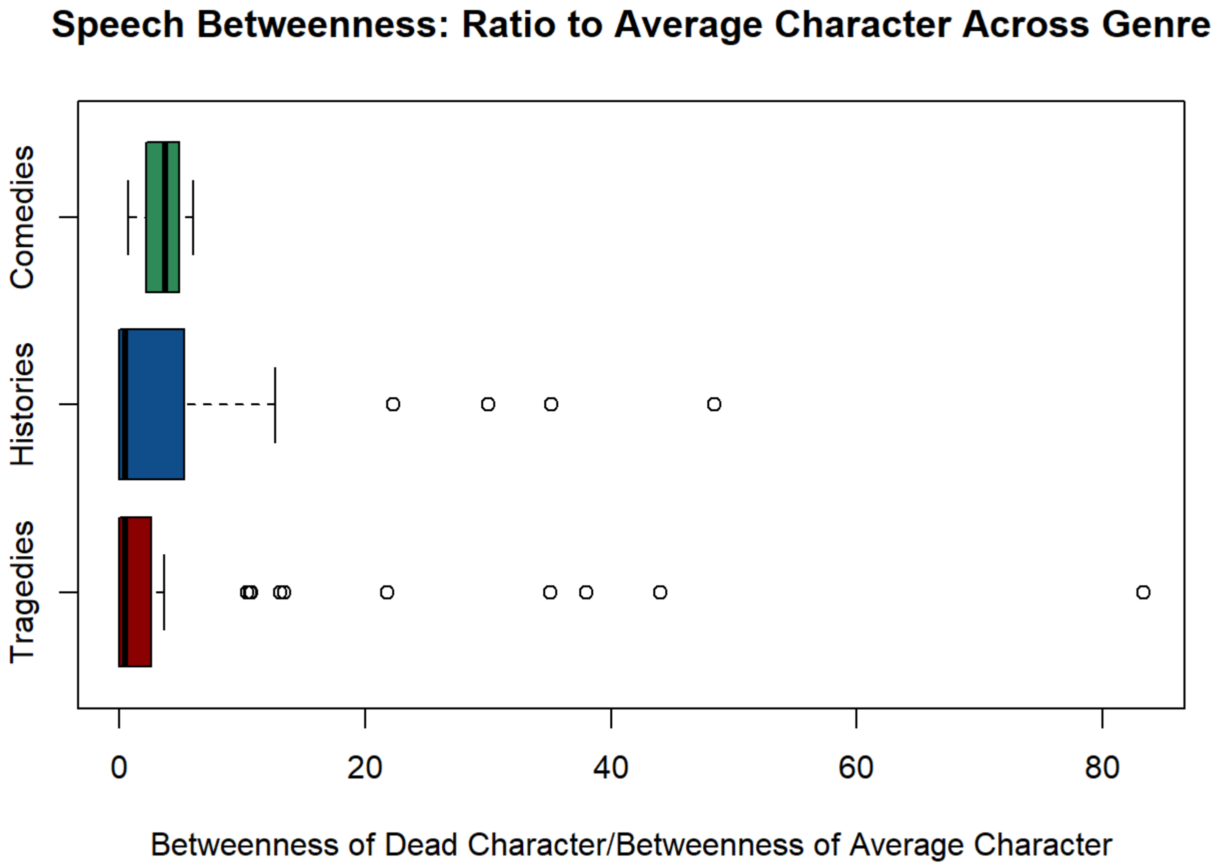
This barplot shows the distribution of death across the acts of each play. As shown, deaths in histories tend to be more spread out across a play while more than two thirds of deaths in tragedies occur in the last act. The comedies bar is split across Act 4 and Act 5 because one death occurs in Act 4 and one in Act 5. Since Act 5 is usually the resolution or catastrophe resulting from the climax, it is telling that so many more deaths occur in Act 5 for tragedies than for histories. It suggests that the deaths are more central to the plot of tragedies and what the narrative has been building toward, whereas deaths are more commonly part of the backdrop of a history play rather than a central focus. It makes sense that few characters die in early acts in tragedies, because the audience hasn't had enough time to get to know that character for it to have much of an impact. Thus, more characters dying earlier in histories suggest that the deaths aren't as central to the plot and that the loss of the characters isn't meant to affect the audience to the same extent.

Degree for speech network

```r
data(network_by_speech)
g_spch_trag = subgraph.edges(g,which(E(g)$speech_num %in% as.character(which(speech_metadata$PLAY %in% tragedies))))
g_spch_hist = subgraph.edges(g,which(E(g)$speech_num %in% as.character(which(speech_metadata$PLAY %in% histories))))
g_spch_comd = subgraph.edges(g,which(E(g)$speech_num %in% as.character(which(speech_metadata$PLAY %in% comedies))))
# Names
dead_chars_spch = character_metadata[which(character_metadata[,14]=="Y"),3]
dead_chars_trag_spch = dead_chars_spch[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% tragedies)]
dead_chars_trag_spch = dead_chars_trag_spch[which(dead_chars_trag_spch %in% names(V(g_spch_trag)))]
dead_chars_hist_spch = dead_chars_spch[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% histories)]
dead_chars_hist_spch = dead_chars_hist_spch[which(dead_chars_hist_spch %in% names(V(g_spch_hist)))]
dead_chars_comd_spch = dead_chars_spch[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% comedies)]
dead_chars_comd_spch = dead_chars_comd_spch[which(dead_chars_comd_spch %in% names(V(g_spch_comd)))]
# Degrees (expressed as ratio to average)
deg_spch_trag = degree(g_spch_trag,mode="total")
deg_spch_hist = degree(g_spch_hist,mode="total")
deg_spch_comd = degree(g_spch_comd,mode="total")
boxplot((deg_spch_trag/mean(deg_spch_trag))[dead_chars_trag_spch],(deg_spch_hist/mean(deg_spch_hist))[dead_chars_hist_spch],(deg_spch_comd/mean(deg_spch_comd))[dead_chars_comd_spch],main="Speech Degree: Ratio to Average Character Across Genre",names=c("Tragedies","Histories","Comedies"),xlab="Degree of Dead Character/Degree of Average Character",col=c("darkred","dodgerblue4","seagreen4"),horizontal = T)
```



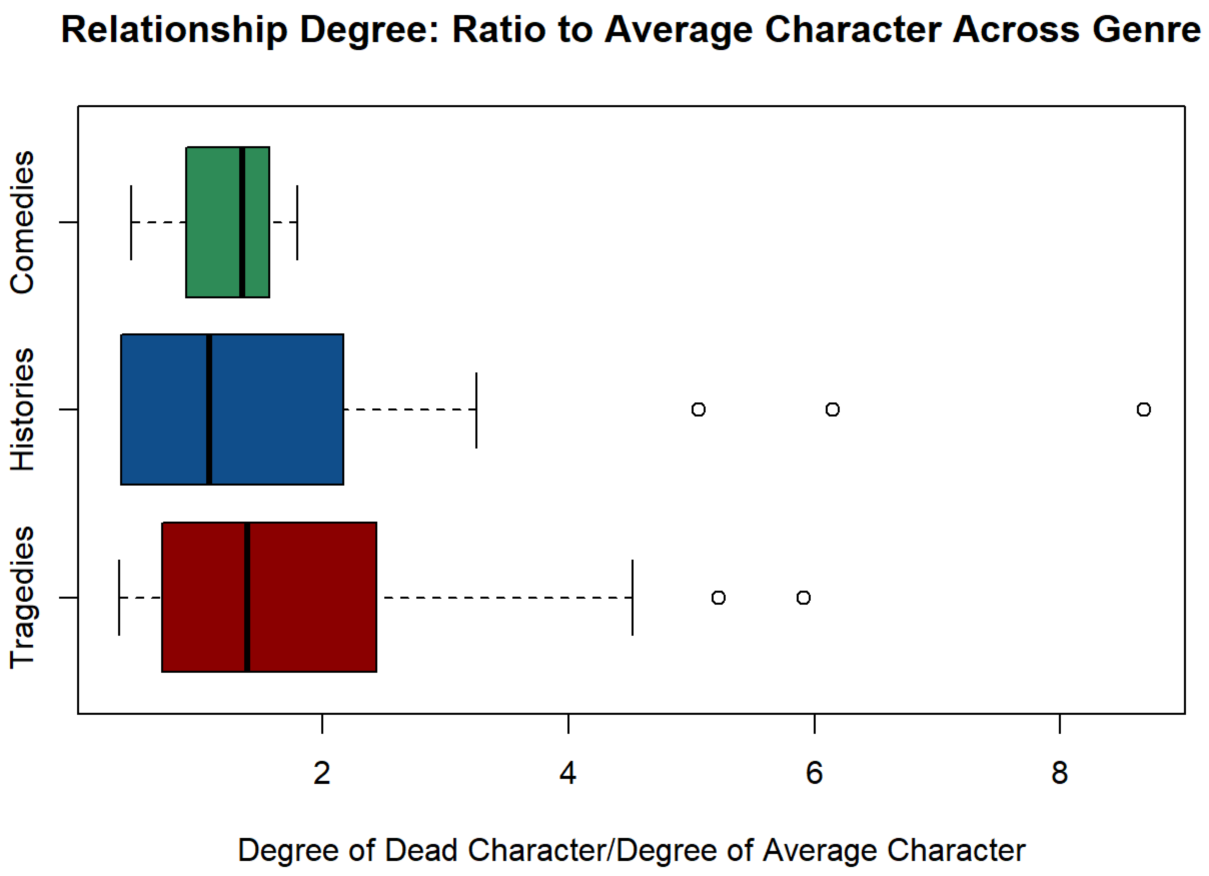
```
# Betweenness
btw_spch_trag = betweenness(g_spch_trag,directed = F)
btw_spch_hist = betweenness(g_spch_hist,directed = F)
btw_spch_comd = betweenness(g_spch_comd,directed = F)
boxplot((btw_spch_trag/mean(btw_spch_trag))[dead_chars_trag_spch],(btw_spch_hist/mean(btw_spch_h
ist))[dead_chars_hist_spch],(btw_spch_comd/mean(btw_spch_comd))[dead_chars_comd_spch],main="Spee
ch Betweenness: Ratio to Average Character Across Genre",names=c("Tragedies","Histories","Comedi
es"),xlab="Betweenness of Dead Character/Betweenness of Average Character",col=c("darkred","dodg
erblue4","seagreen4"),horizontal = T)
```



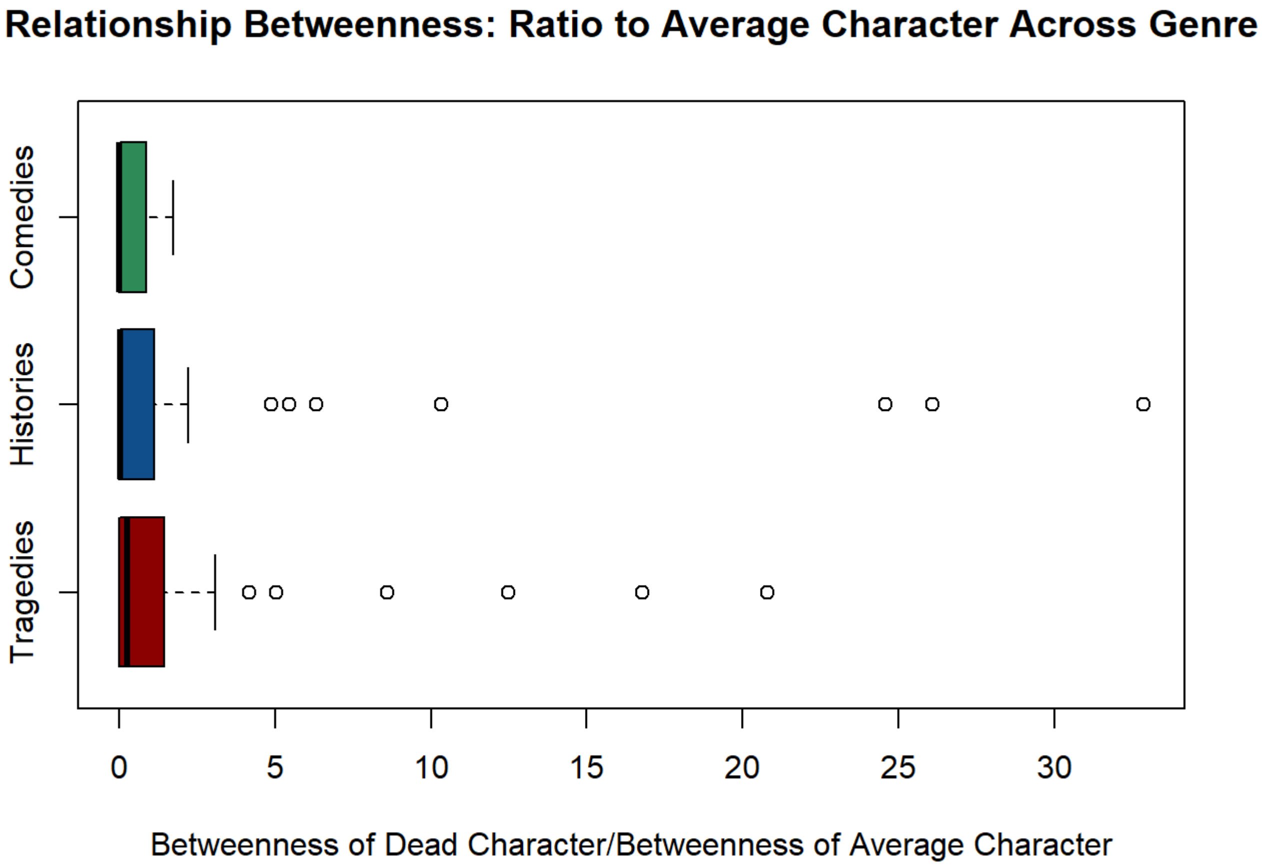
Both of these graphs support the claim that characters that die in tragedies are more important than those who die in histories. The median ratio degree for tragedies is 2.62, meaning roughly that the average character that dies in a tragedy will have a degree 2.62 times that of an average character in tragedy. The median for histories is 1.35, suggesting that dead characters still tend to be a bit more important than usual in the genre, but to a lesser extent than in tragedies. (The median is 0.900 for comedies, but as mentioned this is probably not a very reliable measure since it is based only on three characters). The highest degree is Hamlet with 18.5 times the average degree in tragedies; most of the outliers are main characters, so it makes sense that they would have a very high degree. It is also worth noting that the range for degree is the highest by far for tragedies, which suggests that death is something that happens to characters of many different levels of importance—this makes sense, as death is a significant part of every tragedy.

Interestingly, these patterns are fairly different for betweenness. The median for ratio of betweenness in histories is 0.473 and in tragedies is 0.514, which suggests that roughly on average, dead characters have a lower betweenness than the average character. However, there are a number of outliers in both histories and tragedies that are very high, with Mark Antony from *Antony and Cleopatra* having a betweenness 83.3 times that of the average character and Gloucester from *Henry the 6th, Part II* having a betweenness ratio of 48.4. Thus, the betweenness ratio is extremely left-skewed, with most characters that die not offering much betweenness in the network while a few characters offer an extremely high amount. This skew is even more so the case for tragedies than comedies. This suggests that many deaths in tragedies are overshadowed by the deaths of the main character, and the deaths of the characters with high betweenness (the outliers) will have severe effects on the network while the deaths of the other characters may be less impactful.

```
data(network_by_relationship)
g_relt_trag = subgraph(graph = g, v = which(V(g)$id %in% character_metadata[which(character_metadata[,1] %in% tragedies),2]))
g_relt_hist = subgraph(graph = g, v = which(V(g)$id %in% character_metadata[which(character_metadata[,1] %in% histories),2]))
g_relt_comd = subgraph(graph = g, v = which(V(g)$id %in% character_metadata[which(character_metadata[,1] %in% comedies),2]))
# Names
dead_chars_relt = character_metadata[which(character_metadata[,14]=="Y"),3]
dead_chars_trag_relt = dead_chars_relt[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% tragedies)]
dead_chars_trag_relt = dead_chars_trag_relt[which(dead_chars_trag_relt %in% names(V(g_relt_trag)))] # Get rid of nameless characters
dead_chars_hist_relt = dead_chars_relt[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% histories)]
dead_chars_hist_relt = dead_chars_hist_relt[which(dead_chars_hist_relt %in% names(V(g_relt_hist)))]
dead_chars_comd_relt = dead_chars_relt[which(character_metadata[which(character_metadata[,14]=="Y"),1] %in% comedies)]
dead_chars_comd_relt = dead_chars_comd_relt[which(dead_chars_comd_relt %in% names(V(g_relt_comd)))]
# Degrees (expressed as ratio to average)
deg_relt_trag = degree(g_relt_trag,mode="total")
deg_relt_hist = degree(g_relt_hist,mode="total")
deg_relt_comd = degree(g_relt_comd,mode="total")
boxplot((deg_relt_trag/mean(deg_relt_trag))[dead_chars_trag_relt],(deg_relt_hist/mean(deg_relt_hist))[dead_chars_hist_relt],(deg_relt_comd/mean(deg_relt_comd))[dead_chars_comd_relt],main="Relationship Degree: Ratio to Average Character Across Genre",names=c("Tragedies","Histories","Comedies"),xlab="Degree of Dead Character/Degree of Average Character",col=c("darkred","dodgerblue4","seagreen4"),horizontal = T)
```



```
# Betweenness
btw_relt_trag = betweenness(g_relt_trag,directed = F)
btw_relt_hist = betweenness(g_relt_hist,directed = F)
btw_relt_comd = betweenness(g_relt_comd,directed = F)
boxplot((btw_relt_trag/mean(btw_relt_trag))[dead_chars_trag_relt],(btw_relt_hist/mean(btw_relt_hist))[dead_chars_hist_relt],(btw_relt_comd/mean(btw_relt_comd))[dead_chars_comd_relt],main="Relationship Betweenness: Ratio to Average Character Across Genre",names=c("Tragedies","Histories","Comedies"),xlab="Betweenness of Dead Character/Betweenness of Average Character",col=c("darkred","dodgerblue4","seagreen4"),horizontal = T)
```



The results in these graphs from the relationships suggest much the same things as from the speech network. Characters that die in both tragedies and histories are more likely to have a higher degree than average, with this effect generally being more pronounced in tragedies than histories. The graph with betweenness is also similar, with the median being lower than 1.0 in both genres while there are numerous outliers with much higher betweenness than average. Despite the networks being created in different ways, the trends for dead characters in the genres seem to be consistent across both, which gives evidence that the trends are reliable.

Conclusion

These data seem to indicate that in tragedies and histories, characters that die generally have a greater impact on the plot than those that do not die. While some characters that die have a lower betweeness than average, the main characters that die do, and the more resounding results with degree show that these characters are important to the network. This is not the case in comedies, because death is extremely infrequent except in *Cymbaline*, which is different in some ways than a typical comedy. The characters that die in tragedies have the most impact out of the three genres, suggesting that death is most central to this genre, but histories also have many characters that die that are influential to the plot and network. So while death is less central to histories than to tragedies, the structure of death in histories is generally more similar to that in tragedies than in comedies, which indicates that tragedies and histories have some structural similarities.