Social Network Analysis of Literary Fiction

Hamid Pirnia

1. Executive Summary

In this project we studied the characteristics of literary fiction in the context of social networks. Our goal in this project was to study various facets of network science within a piece of literature. There are three main areas we wanted to tackle. These were: Identifying similarities between real and artificial social networks, identifying the main characters and protagonists of the story, and identifying the formation of character ties and communities.

The way in which we set out to achieve the goals of this project was to use multiple metrics of a network diagram. Utilizing various network metrics we were able to analyse different characteristics of the fictional networks in the stories of Harry Potter and Game of Thrones. One of the main points of our project was finding communities and main characters throughout the fictional network using only the network metrics. We were able to successfully find communities within the story's social network. In the case of the Harry Potter network we saw communities largely based around some main character, house, or morals. By this many of the communities were formed based on a character's house affiliation or weather they were on the side of the protagonist or antagonist. In terms of determining the protagonist we found an interesting observation regarding the metric used for protagonist identification. As we know the protagonist of Harry Potter is Harry himself. As we compiled the data we found that if we based our protagonist identification on the more commonly used centrality metric it would not return the protagonist. We found that the more accurate metric to use to identify a protagonist, contrary to common thought, is betweeness. The last of our main goals through this project was to find similarities between a real social network and an artificial social network. An artificial social network refers to one that based on fiction such as the networks we were basing our project on. Throughout this analysis we found that Harry Potter resembled a real social network more than some other fictional stories as the narrative provides more realistic social interactions in the school type setting. Throughout this report we will detail our method and assess the results.

2. Motivation

What inspired our pursuit of this project was very much based on research and studies done on networks of literary works from other sources. The different sources used in doing our initial research provided us with many different perspectives to analyse networks in literature. These differing perspectives provided a broad base of motivation to take on this project and allowed us to achieve a wider base of results and analysis.

The first of our motivations for this project was to find how an artificial social network from a work of fiction can relate to that of a real life social network. The similarities between a real life social network and one based on fiction have had little to no academic exploration (Albreich, et al., 2002). The lack of research into this topic showed us that comparing artificial social networks to a real one would provide great insight on the topic of social networks in literature. The insight that we were looking for in this topic was to find what differentiates a social network to be real and how were certain fictional networks more real than others.

Our second motivation in this project was to analyse the communities and character ties within the community. Community detection is something that has been done very often

in many different networks. Community detection is also something that we have explored in a different context but we wanted to find how communities were formed in popular pieces of work that are so often analysed in pop-culture. In the context of this project we wanted to study how ties are formed according to different attributes of each node as well as the nature of these ties. In turn, to reveal the effect of some node characteristics in the formation of ties inside and outside their social network group. Then study the communities formed in our networks and try to explain the structure by the given characteristic. We hoped that this would provide great insight into the motivations of community formation within the stories used.

One of our three main goals was to not only identify the protagonist and main characters but to also find if there is any relationship between the identification process and the point of view of a story. As mentioned earlier, we know that insight into real life networks can be gained by studying the structure of artificial networks. Thus we have a purpose in developing means of retrieving the most important set of characters via certain quantitative network properties of these characters. Furthermore, the same metrics that reveal the most important characters can indicate the protagonist of a literary fiction as well as the perspective from which the story is being told.

When reading a book the plot and narrative is the driving motivation of the reader. In this project one of our motivations was to use the character network to analyse the plot dynamics of the story. Skorinkin's paper he analyses Tolstoy's work *War and Peace* in an effort to find moments of war and peace based on network analysis (Skorinkin D. A, 2017). In this paper Skorinkin establishes a connection between plot dynamics and network density, as a result of moments of was having lower character interactions and therefore lower network density (Skorinkin D. A, 2017). Utilizing this information we hoped to also see if the correlation detailed by Skorinkin persisted. Another goal of this type of analysis is to detail animosity between characters.

3. Background

We want to see how a narrative develops over time and how the narrative dynamics can be analysed via network science. We may observe the visual changes to graph or use a heatmap to see the growth of the network as the story develops, in turn utilising network analysis for computational research of fictional narrative. However, Skorinkin's article *Extracting Character Networks to Explore Literary Plot Dynamics* provides insights in network analysis of literature implementing network metrics to achieve this goal. Skorinkin's hypothesis in the report is "to prove was that the parts of the novel describing war (i.e. those where the battlefield or military units are the primary settings), have statistically lower density of interaction between characters, resulting in lower network density, higher network diameters and lesser average node degrees. By showing this correlation we mean to demonstrate the applicability of network analysis to computational research of fictional narrative (e.g. detection of tension changes in the plot)." (Skorinkin D.A, 2017). A main piece of information that was relevant of our project, from Skorinkin's paper was his correlations data. These metrics are used to support his hypothesis but we may also compare out

metrics to his and make appropriate interpretations. For example, here Skorinkin points out that moment of war in Tolstoy's *War and Peace* will correlate positively with a higher edge density because the scenes of the book have less character interactions (Skorinkin D.A, 2017). This correlation is show in data from Skorinkin's paper:

('war' (0), 'peace' (1) and 'a mixture of both' (0.5))

Parameter	Correlation with 'war or peace' value
Density	0.650
Diameter	−0.533
Average degree	0.730
Average weighted degree	0.714

(Skorinkin D.A, 2017)

Volume	1	1	1	2	2	2	2	2	3	3	3	4	4	4	4
Part	1	2	3	1	2	3	4	5	1	2	3	1	2	3	4
Peace/ War	1	0	0,5	1	1	1	1	1	0	0	0	0,5	0	0	0
Density	0.15	0.16	0.11	0.31	0.25	0.24	0.36	0.21	0.17	0.13	0.13	0.13	0.14	0.18	0.18
Average degree	3.85	2.38	2.64	4.00	2.55	2.67	2.50	3.29	2.00	2.57	2.32	2.00	1.50	1.60	1.60
Average weighted degree	11.41	5.63	7.44	12.86	7.82	6.50	13.00	10.35	5.38	4.76	4.42	3.88	1.67	9.40	5.00

(Skorinkin D.A, 2017)

Similar network metric are studied by Dimitrio Kydros in *Social network analysis in literature.* These metrics can be used for comparison with our results. *The case of The Great Eastern by A. Embirikos*. Kydros presents us with this "Topological comparison to other literature networks":

	the *Iliad*	*Les Misérables*	GEN
Nodes	538	77	572
Links	1557	254	1764
Density	0.001	0.087	0.008
Average Degree	5.78	3.299	3.084
Average shortest path	3.33	2.641	3.12
Diameter	9	5	7
Average eccentricity	6.56	4.32	2.342
Average clustering coefficient	0.41	0.736	0.766
Assortativity	-0.08	0.01	-0.07

(D. Kydros, 2014)

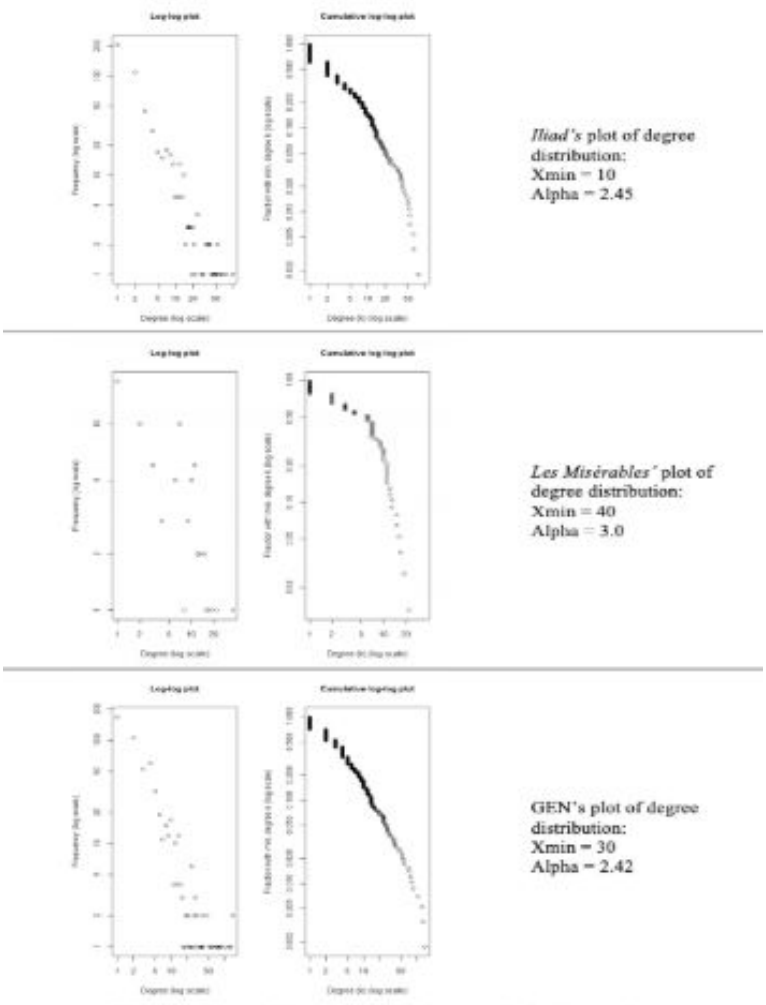


Figure 3: Plots of degree distributions and relative results

(D. Kydros, 2014)

In Alberich's paper, written alongside other authors, *Marvel Universe looks almost like a real social network*, he analyses the Marvel Universe comic books social network and proposes that it is like a real social networks in every way except for one. In his conclusion he states that "[a]lthough to some extent the Marvel Universe tries to mimic human relations, and in particular it is completely different from a random network, we have shown that it cannot completely hide its artificial origins.". Alberich argues that all real social networks must satisfy the following properties: "(a) on average, every pair of nodes can be connected through a short path within the network; (b) the probability that two nodes are linked is greater if they share a neighbor; and (c) the fraction of nodes with k neighbors decays roughly as a function of the form k^{-alpha} for some positive exponent alpha , with perhaps a cutoff for large values of k." (R. Alberich, 2002). "Plots of degree distributions and relative results" above from Kydros' study support Alberich's paper and in particular demonstrate the networks alignment with (c), we observe a power law distribution with an exponential cut-off. In that all "real" social networks present this property and in their case artificial neural network of their respectively studied literary fictions also do(R. Alberich, 2002), (D. Kydros, 2014) .

Franco Moretti, in his paper *Network Theory, Plot Analysis*, defines the protagonist as the character with the highest centrality for every network (Moretti, 2011). Apoorv Agarwal and his collaborators in a 2012 paper titled *Social Network Analysis of Alice in Wonderland* build on this definition. Using Social Network Analysis to extract vertex properties and treating interaction and observation events in isolation were "able to conclude that Alice is the only perspective holder in the story" (A. Agarwal 2012). This was unprecedented and challenging since point-of-view and the protagonist are not necessarily the same person and a method for extracting the perspective holder had not been previously examined (A. Agarwal 2012).

Dimitrio Kydros in *Social network analysis in literature. The case of The Great Eastern by A. Embirikos* assesses some of the same vertex and graph metrics as Moretti and Agarwal. He lists top five highest characters by their centrality variables and other metrics such as closeness and betweenness as well as define these terms. He proceeds to interpret these lists and makes remarks about characters importance locally and globally. Furthermore he shares insight that he gains about author's intentions. Kydros states: "It looks that he definitely had a consistent sociological universe in his mind, neither chaotic nor completely utopian, reflecting real social structures composed of persons with real social relations." (D. Kydros, 2014).

Dimitrio Kydros in the same paper proceeds to detect communities of the social network of *The Great Eastern by A. Embirikos* using the techniques described by Girvan & Newman in their 2012 paper: *Community structure in social and biological networks*. Kydros concludes his analysis of the communities detected by noting that "the main actors in each community hold a very high degree centrality ranking… Main actors operate as hubs and other actors group around hubs, forming these communities" (Kydros, 2014). Thus he demonstrates the applicability of Girvan & Newman's Community detection techniques to literary fiction character network analysis.

4. Data and Analysis

4.1 Data

Dataset 1:

- 65 nodes.
- All books combined.
- Given attributes: "type" that describes if a relationship is an friendly (+) or not (-).
- Relationships extracted by comparing with the Harry Potter Wikia.

Dataset 2:

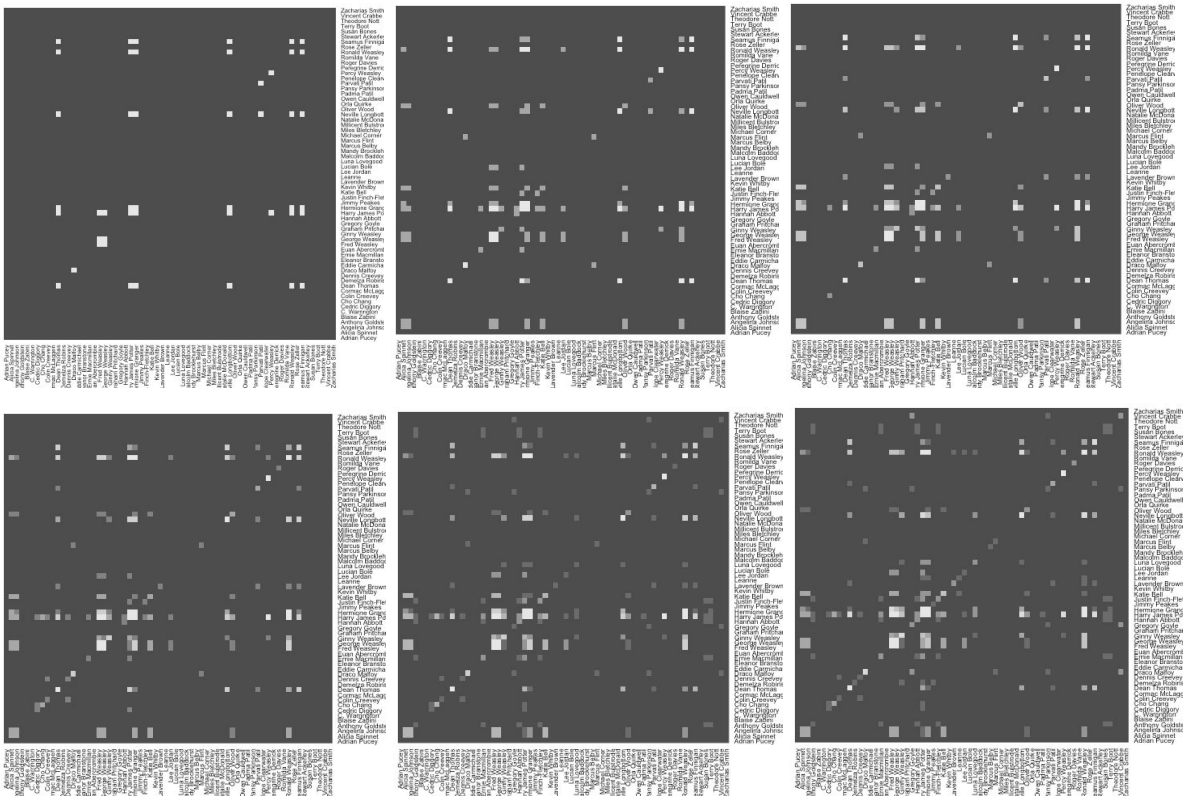
- 64 nodes.
- Adjacency Matrix for each book.
- Given attributes: "schoolyear" - year attended Hogwarts, "house" - (1) Gryffindor (2) Slytherin (3) Hufflepuff (4) Ravenclaw, and "gender" - (1) Male (2) Female.
- Relationships extracted by considering a relationship if two characters appear in the same sentence.

4.2 Heatmaps

Below each heatmap represents the adjacency matrix of the character network of each Harry Potter book. The brightness is proportional to the weight of the relationship. As in each book the relationships accumulate, the existing relationships get reinforced (since we are literally adding the adjacency matrices). The names need to be zoomed-in to see but a development in the network is clear to the naked eye. We can see generally more relationships defined as the story progresses as well as many relationships getting reinforced. This demonstrates a visual technique to possible detect plot developments and interpret story timeline. Some stories do not progress in a linear fashion. Stories such as Star Wars may be presented in a reverse order however this technique may find the trend of

the plot development and possible the correct time-line by implementing network visualization.

Dataset 2 (Relationship defined according to occurrences of names within the same sentence, Ordered books 1 to 6 accordingly):



Method:

Import Books, then run:
> dimnames(book1) <- list(as.vector(names.df$name), as.vector(names.df$name))
> heatmap(book1, symm=TRUE, col=gray.colors(10), Colv=NA, Rowv=NA, scale='none')
Repeat for all books, add books as such:
> book12<-book1+book2

4.3 Graph Summaries

These Results will be discussed in depth in the conclusion. However we can say that our network has a fairly large density compared to literary novels presented in the paper by Kydros and Skorilkin with the tables above in the “background” section of this paper. However these results will be compared to other real and artificial networks in the conclusion section for further interpretation.

Dataset 1 (Relationship defined according to official Wiki, All books:

Nodes	**Vertices**	**Density**	**Diameter**	**Average Degree**
65	513	0.247	4	15.69

Dataset 2 (Relationship defined according to occurrences of names within the same sentence, Ordered books 1 to 6 accordingly):

Each book is assessed as an individual network. “All” is the binding of all networks.

Book	**Nodes**	**Vertices**	**Density**	**Diameter**	**Average Degree**
1	64	20	0.015	3	0.625
2	64	55	0.037	2	1.719
3	64	52	0.034	3	1.625
4	64	22	0.018	3	0.688
5	64	73	0.050	2	2.281
6	64	35	0.028	4	1.094
All	64	116	0.182	4	8.031

Method:

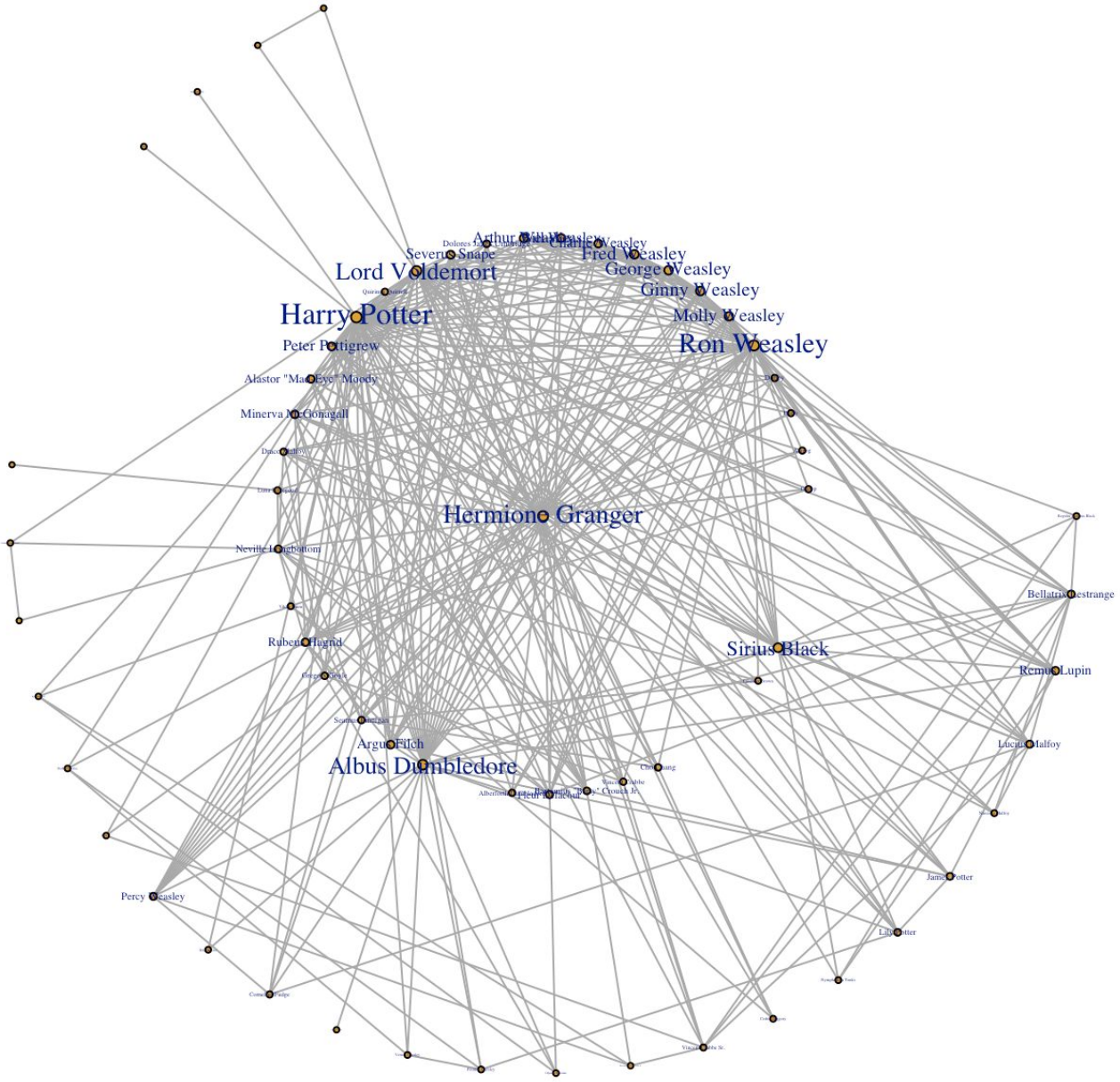
```
> summary(graph) #For Nodes and Vertices
> edge_density(graph, loops = FALSE) #For Edge Density.
> mean(degree(graph, v = V(graph), mode = "all", loops = FALSE, normalized = FALSE))
#For Avg. Degree.
> diameter(book1.graph, directed = FALSE, unconnected = TRUE, weights = NULL) #For
Diameter.
```

4.4 Graph Plots (Vertex Size Proportional to Centrality Score).

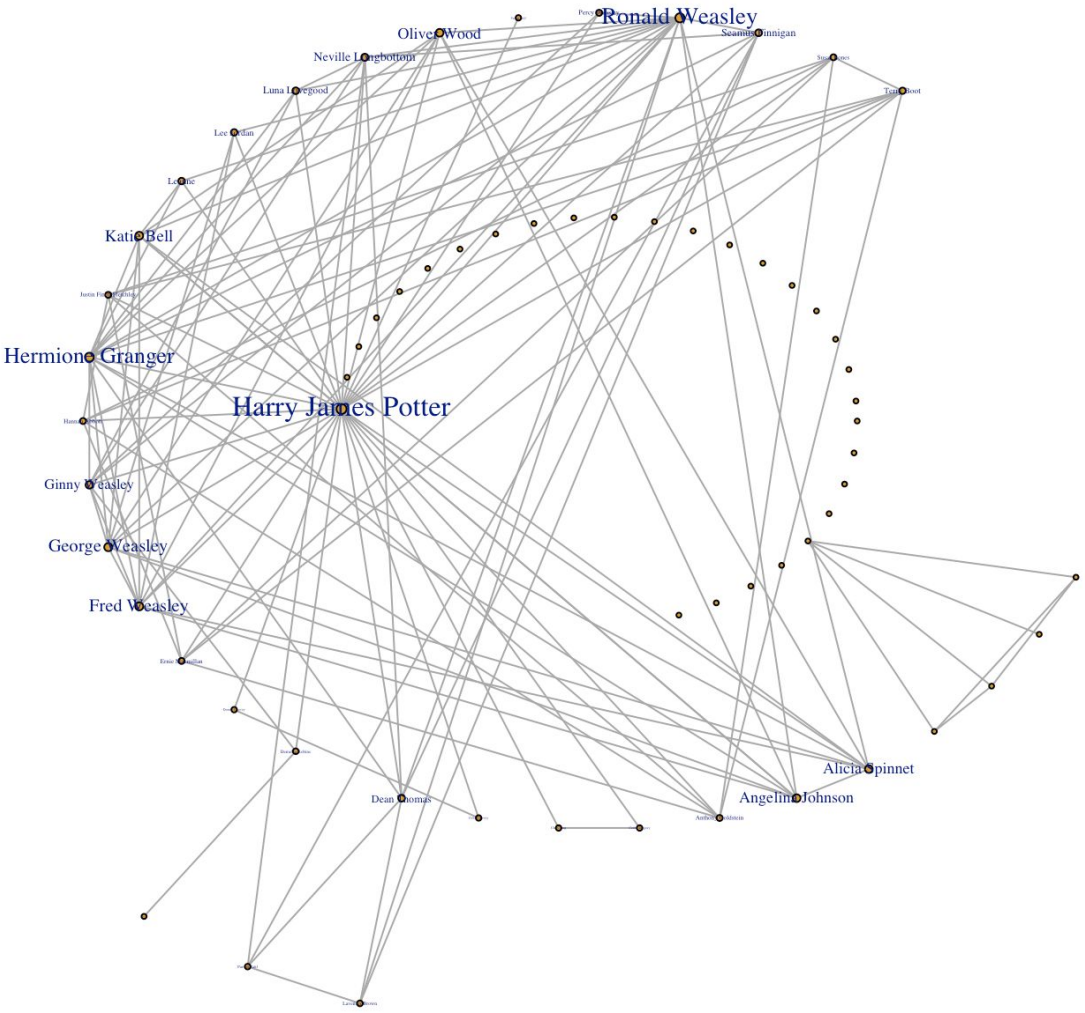
Method:

```
> plot(simplify(graph), layout=layout.reingold.tilford(graph, circular=T),
vertex.size=1+eigen_centrality(simplify(graph))$vector+0.01,
vertex.label.cex=eigen_centrality(simplify(graph))$vector+0.01)
```

Dataset 1 (Relationship defined according to official Wiki, All Books):



Dataset 2 (Relationship defined according to occurrences of names within the same sentence, All Books):



4.5 Degree Distribution

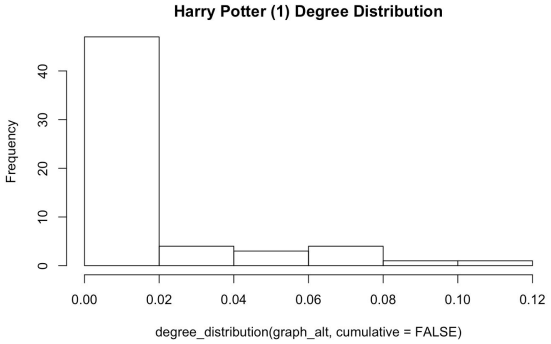
According to R. Alberich, “An interesting statistical datum that can be used to distinguish random networks from non-random networks is the distribution P(k) of degrees in the network.”. In his paper he states that random networks follow a binomial degree distribution whereas in non-random networks “the distribution P(k) has a tail that follows either a power law for some constant, positive exponent…”, alpha, “...or a power law form with an exponential cutoff.” (R. Alberich, 2002). Thus we set out to assess the degree distribution of our networks.

Using “fit_power_law()” function from the igraph library we can fit our degree distribution into a power law distribution. The output of the function will be the alpha that is described by R. Alberich and another variable labeled “KS.p”. According to official igraph manual: KS.p is a “[n]umeric scalar, the p-value of the Kolmogorov-Smirnov test. Small p-values (less than 0.05) indicate that the test rejected the hypothesis that the original data could have been drawn from the fitted power-law distribution.”. So we have:

Null Hypothesis: The original data could have been drawn from the fitted power-law distribution.

Alt. Hypothesis: The original data not could have been drawn from the fitted power-law distribution.

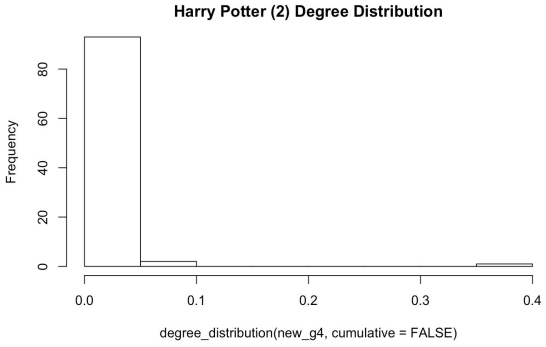
Dataset 1 (Relationship defined according to official Wiki, All Books):



- Alpha: 3.050
- p-Value: 0.314

We fail to reject the null hypothesis that the original data could have been drawn from the fitted power-law distribution.

Dataset 2 (Relationship defined according to occurrences of names within the same sentence, All Books):



- Alpha: 2.651
- p-Value: 0.969

We fail to reject the null hypothesis that the original data could have been drawn from the fitted power-law distribution.

Thus we can say that our findings support both Kydros' and Alberich's as we observe a power law distribution with an exponential cut-off in both data-sets after failing to reject the null hypothesis. We can say now that the results satisfy a part of the definition given by R. Alberich by "(c) the fraction of nodes with k neighbors decays roughly as a function of the form k^{-alpha} for some positive exponent alpha , with perhaps a cutoff for large values of k." (R. Alberich, 2002). However these results will be compared to other real and artificial networks in the conclusion section for further interpretation.

4.6 Average Path Length

According to R. Alberich, "The distance between two connected nodes in a network is defined as the length (the number of links) of the shortest path connecting them" (R. Alberich, 2002). Since non-random real social networks tend to have lower average distances between vertices, we move on to examine this value for our networks. We use the igraph library function "mean_distance()" that returns this value.

Dataset 1 (Relationship defined according to official Wiki, All Books):

- Average Path Length: 2.028

Dataset 2 (Relationship defined according to occurrences of names within the same sentence, All Books):

- Average Path Length: 1.895

We can say now that the results satisfy a part of the definition given by R. Alberich by "(a) on average, every pair of nodes can be connected through a short path within the network." (R. Alberich, 2002). However these results will be compared to other real and artificial networks in the conclusion section for further interpretation.

4.7 Local and Global Clustering Coefficients

According to R. Alberich, "In most social networks, two nodes that are linked to a third one have a higher probability to be linked between them: two acquaintances of a given person probably know each other. This effect is measured using the clustering coefficient" (R. Alberich, 2002). Using igraph library function "transitivity()" we can calculate global and local clustering coefficients by setting the "type" parameter accordingly.

Dataset 1 (Relationship defined according to official Wiki, All Books):

- Local Clustering: 0.652
- Global Clustering: 0.413

Dataset 2 (Relationship defined according to occurances of names within the same sentence, All Books):

- Local Clustering: 0.836
- Global Clustering: 0.528

We can see immediately that in all our cases we get a clustering coefficient higher than that of Marvel Universe studied by R. Alberich (which was 0.012 Global) (R.Alberich, 2002). These results will be compared to other studies in the conclusion section. However we can say now that the results indicate a more "real" social network than Marvel universe following a part of the definition given by R. Alberich by satisfying the statement "(b) the probability that two nodes are linked is greater if they share a neighbor" (R. Alberich, 2002).

4.8 Vertex Metrics

Dataset 1 (Relationship defined according to official Wiki, All Books):

Eigen Centrality	Betweenness	Closeness	Hubs	Authority	Transitivity
Hermione Granger	Harry Potter	Harry Potter	Hermione Granger	Hermione Granger	Lavender Brown
George Weasley	Lord Voldemort	Ron Weasley	George Weasley	George Weasley	Alice Longbottom
Ron Weasley	Ron Weasley	Hermione Granger	Ron Weasley	Ron Weasley	Quirinus Quirrell
Albus Dumbledore	Albus Dumbledore	Lord Voldemort	Albus Dumbledore	Albus Dumbledore	Tom Riddle Sr.
Arthur Weasley	Hermione Granger	Albus Dumbledore	Arthur Weasley	Arthur Weasley	Mary Riddle

Dataset 2 (Relationship defined according to occurrences of names within the same sentence, All Books):

Eigen Centrality	Betweenness	Closeness	Hubs	Authority	Transitivity
Harry Potter	Harry Potter	Harry Potter	Harry Potter	Harry Potter	Dean Thomas
Ronald Weasley	Ronald Weasley	Ronald Weasley	Ronald Weasley	Ronald Weasley	Draco Malfoy
Hermione Granger	Demelza Robins	Hermione Granger	Hermione Granger	Hermione Granger	George Weasley
Ginny Weasley	Hermione Granger	Ginny Weasley	Ginny Weasley	Ginny Weasley	Hermione Granger
Leanne	Ginny Weasley	Katie Bell	Katie Bell	Katie Bell	Katie Bell

In both cases of the datasets our findings align with the papers of Newman, Kydros and Agarwal. The highest eigen-centrality character, and in fact all other metrics in most cases as well (except transitivity), is the protagonist (D. Kydros, 2014).

Method:

#Centrality Scores
> eigen_centrality(graph)

#Betweeness Score
>betweenness(graph)

#Closeness Score
>closeness(graph)

#Hubs Scores
>hub_score(graph, scale = TRUE, weights = NULL, options = arpack_defaults)$vector

#Authority Scores
>authority_score(graph, scale = TRUE, weights = NULL, options = arpack_defaults)$vector

#Transitivity
>transitivity(simplify(graph), type = "local")

4.9 Group Ties
Dataset 2 (Relationship defined according to occurances of names within the same sentence, All Books):

Attribute	Within Group Ties %	Between Group Ties %
School Year	72.79%	27.21%
House	89.52%	10.48%
Gender	63.42%	35.58%

```
, , year = Harrys Year

            SexString
group        Female Male
  Gryfindor       3    5
  HufflePuff      3    3
  Ravenclaw       2    3
  Slytherin       2    5
```

```
, , year = Not Harrys Year

            SexString
group        Female Male
  Gryfindor       7   10
  HufflePuff      2    3
  Ravenclaw       4    4
  Slytherin       0    8
```

We can see that the Within Group Tie% for School Year and House is significantly high. This can be confirmed by reading the books or watching the movies as different houses are almost like different “fraternities”. In that term comes an explanation for observation we make, the high though not prominent Within Group Ties % of Gender. Since the sleeping rooms are separated by gender in the series, this makes sense.

Method:

```
hp$year[hp$schoolyear == 1991] <- 'Harrys Year'
hp$year[hp$schoolyear != 1991] <- 'Not Harrys Year'

hp$SexString[hp$gender == 1] <- 'Male'
hp$SexString[hp$gender == 2] <- 'Female'

hp$group[hp$house == 1] <- 'Gryfindor'
hp$group[hp$house == 4] <- 'Slytherin'
hp$group[hp$house == 2] <- 'HufflePuff'
hp$group[hp$house == 3] <- 'Ravenclaw'

with(hp, table(group, SexString, year))

n = 64 # count of vertices
withinGroupTies <- 0 # within group ties
betweenGroupTies <- 0 # between groups ties
for (i in 1:n) {
  for (j in 1:n) {
    if (hp$gender[i] == hp$gender[j]) {
      withinGroupTies = withinGroupTies + add_books[i,j] }
    if (hp$gender[i] != hp$gender[j]) {
      betweenGroupTies = betweenGroupTies + add_books[i,j]}}}
totalTies = withinGroupTies + betweenGroupTies
withinGroupsPercentGender = format(round(withinGroupTies/totalTies * 100, 2), nsmall = 2)
betweenGroupPercentGender = format(round(betweenGroupTies/totalTies * 100, 2), nsmall=2)

n = 64 # count of vertices
withinGroupTies <- 0 # within group ties
betweenGroupTies <- 0 # between groups ties
for (i in 1:n) {
  for (j in 1:n) {
    if (hp$house[i] == hp$house[j]) {
      withinGroupTies = withinGroupTies + add_books[i,j] }
    if (hp$house[i] != hp$house[j]) {
      betweenGroupTies = betweenGroupTies + add_books[i,j]}}}
totalTies = withinGroupTies + betweenGroupTies
withinGroupsPercentHouse = format(round(withinGroupTies/totalTies * 100, 2), nsmall = 2)
betweenGroupPercentHouse = format(round(betweenGroupTies/totalTies * 100, 2), nsmall=2)

n = 64 # count of vertices
withinGroupTies <- 0 # within group ties
betweenGroupTies <- 0 # between groups ties
for (i in 1:n) {
  for (j in 1:n) {
    if (hp$year[i] == hp$year[j]) {
      withinGroupTies = withinGroupTies + add_books[i,j] }
    if (hp$year[i] != hp$year[j]) {
      betweenGroupTies = betweenGroupTies + add_books[i,j]}}}
totalTies = withinGroupTies + betweenGroupTies
withinGroupsPercentYear = format(round(withinGroupTies/totalTies * 100, 2), nsmall=2)
betweenGroupPercentYear = format(round(betweenGroupTies/totalTies * 100, 2), nsmall=2)
```

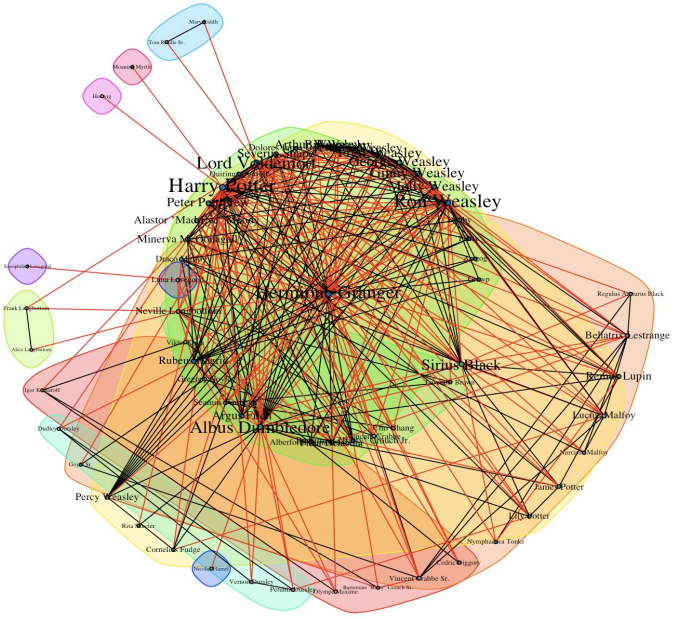
4.10 Community Detection

Using “walktrap.community(simplify(graph))” from the “igraph” library. Concentrating on Communities with 2 or more members. As we can see most of the communities are formed in a similar way as described by D. Kydros in his study of *The Great Eastern*. We see

that most communities are formed by having a high centrality degree “leader” and their “friends”. We See that both datasets produce a community with Harry Potter, Ron Weasley, Hermione Granger and other minor Harry allies (Community #5 for dataset 1 and community #3 for dataset 2). Other communities follow the same trend. Voldemort and his allies, Dumbledore and his allies and so on. Though in the first dataset Voldemort appears in the same community as Harry this show minor flaws of walk-trap method that we used.

Dataset 1 (Relationship defined according to official Wiki, All Books):

- Community #3 (17 Members): Sirius Black, Albus Dumbledore, Cornelius Fudge, Remus Lupin, Alastor "Mad-Eye" Moody, Peter Pettigrew, James Potter, Lily Potter, Rita Skeeter, Arthur Weasley, Bill Weasley, Charlie Weasley, Fred Weasley, George Weasley, Ginny Weasley, Molly Weasley.



- Community #5 (16 Members): Bartemius "Barty" Crouch Jr., Fleur Delacour, Aberforth Dumbledore, Argus Filch, Hermione Granger, Rubeus Hagrid, Minerva McGonagall, Harry Potter, Quirinus Quirrell, Lord Voldemort, Severus Snape, Dolores Janes Umbridge, Ron Weasley, Fluffy, Aragog, Grawp.

- Community #2 (11 Members): Dobby, Nymphadora Tonks, Lucius Malfoy, Draco Malfoy, Narcissa Malfoy, Bellatrix Lestrange, Gregory Goyle, Goyle Sr., Vincent Crabbe, Vincent Crabbe Sr., Regulus Arcturus Black.

- Community #1 (5 Members): Bartemius "Barty" Crouch Sr., Cedric Diggory, Igor Karkaroff, Viktor Krum, Olympe Maxime.

- Community #6 (4 Members): Lavender Brown, Cho Chang, Seamus Finnigan, Neville Longbottom.

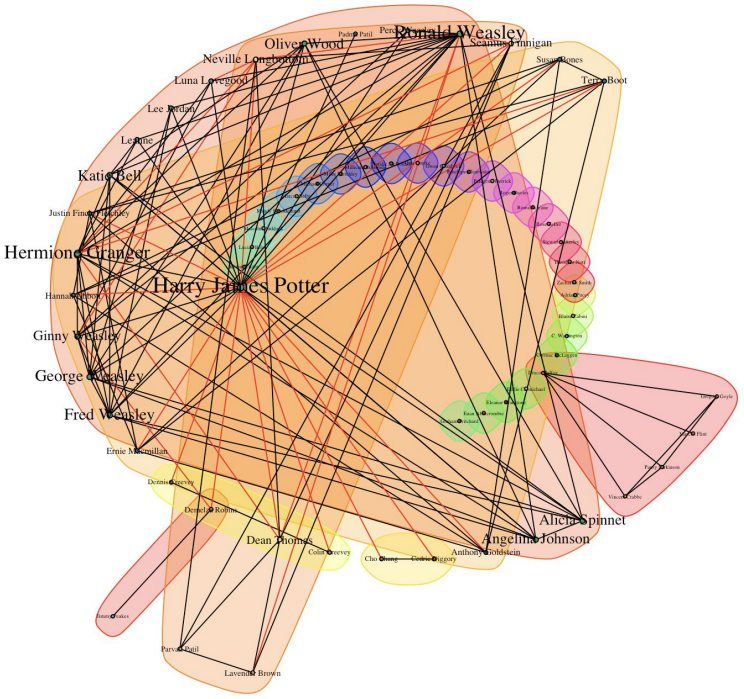
Dataset 2 (Relationship defined according to occurances of names within the same sentence, All Books):

- Community #3 (15 Members): Ginny Weasley, Harry James Potter, Katie Bell, Lee Jordan, Oliver Wood, Alicia Spinnet, Hermione Granger, Angelina Johnson, Fred Weasley, Luna Lovegood, George Weasley, Leanne, Padma Patil, Percy Weasley, Ronald Weasley.

- Community #5 (6 Members): Terry Boot, Ernie Macmillan, Susan Bones, Hannah Abbott, Anthony Goldstein, Justin Finch-Fletchley.

- Community #1 (5 Members): Draco Malfoy, Vincent Crabbe, Marcus Flint, Gregory Goyle, Pansy Parkinson, Lord Voldemort.

- Community #4 (5 Members): Parvati Patil, Seamus Finnigan, Dean Thomas, Neville Longbottom, Lavender Brown.



4215. Conclusion

Network	Density	Diameter	Average Degree	Alpha	(Local Clustering) or (Local, Global Clustering)	Average Path Length
1. The Iliad	0.001	9	5.780	2.45	0.41	3.33
2. Les Miserables	0.087	5	3.299	3.0	0.736	3.641
3. The Great Eastern	0.008	7	0.766	2.42	0.766	3.12
4. War and Peace	At Peak: 0.36	NA	At Peak: 4.00	NA	NA	NA
5. Marvel	NA	5	51.88	0.72	0.012	2.63
6. DBLP	NA	NA	8.457	NA	0.7231, 0.1868	NA

7. Orkut	NA	NA	76.28	NA	0.1794, 0.0413	NA
8. Flickr	NA	NA	20.92	NA	0.3616, 0.1076	NA
9. Live Journal	NA	NA	17.69	NA	0.3508, 0.1179	NA
10. Harry Potter	0.247	4	15.69	3.050	0.652, 0.413	2.028

Legend:

1. The Iliad, 2. Les Miserables and 3. The Great Eastern data from Kydros' paper (D. Kydros, 2014).

4. War and Peace data from Skorilkin's paper (D.A Skorinkin, 2017).

5. Marvel data from Alberich's paper (R. Alberich, 2002)

Estimating Clustering Coefficients and Size of Social Networks via Random Walk by Stephen J. Hardiman et al., presents us with the following data for some "real" social networks for comparison:

6. "DBLP In the "Digital Bibliography and Library Project" (DBLP[18]) dataset each entry is a reference to a paper which contains a title and a list of authors. In the corresponding network each node is an author and an edge between two authors represent co-authorship of one or more papers."

7. "Orkut is a general purpose social network...In this social network the friendship connections (edges) are undirected. "

8. "Flickr is an online social network with focus on photo sharing...In this social network the friendship connections (edges) are directed."

9. "LiveJournal is an on-line social network with focus on journals and blogs....In this social network the friendship connections (edges) are directed." (S. J. Hardiman, 2015)."

10. Harry Potter dataset 1 is selected covering all the books since it was based on the wikia and relationships seem more accurate.

We can observe that we have the second highest network density out of given data following Dostoyevsky's *War and Peace.*

The Diameter for Harry Potter's network is the smallest, which can be interpreted as a closely knit network.

The average vertex degree is higher than most novels but significantly lower than

Marvel Universe. Though the Harry Potter network average vertex degree is a close match to the four social networks we have in the table above.

We have an alpha value above 2 for all fictional character networks except Marvel universe which has alpha below 1. In the analysis section we demonstrated that our degree distribution fits into a power law distribution with an exponential tail cut-off the results satisfied a part of the definition given by R. Alberich: “(c) the fraction of nodes with k neighbors decays roughly as a function of the form k^{-alpha} for some positive exponent alpha , with perhaps a cutoff for large values of k.” (R. Alberich, 2002).

Average path length was the lowest out of the data at hand. We can say that the results satisfy a part of the definition given by R. Alberich by “(a) on average, every pair of nodes can be connected through a short path within the network.” (R. Alberich, 2002).

Regarding the clustering coefficients of artificial networks, Alberich’s paper stated that: “Although to some extent the Marvel Universe tries to mimic human relations, and in particular it is completely different from a random network, we have shown that it cannot completely hide its artificial origins. As in real-life collaboration and, in general, social networks, its nodes are on average at a short distance of each other, and the distribution of collaborators shows a clear power-law tail with cutoff. But its clustering coefficient is quite smaller than what’s usual in real-life collaboration networks” (R. Alberich, 2002). However in our case, we have a clustering coefficient far larger than that of the Marvel Universe. This large clustering coefficient as well as a small average path length and a power law degree distribution with an exponential cut-off leaves our Harry Potter network indistinguishable from any “real” social network, contradicting R. Alberich’s finding with the Marvel Universe Network and demonstrating the futility of the definition of a “real” social network when analysing such realistic fictional networks such as Harry Potter’s.

6. Discussion

Throughout this report we demonstrated a technique to visualize plot development using a heatmap of the character network. This shows promises in detecting a linear timeline and plot development in fictional narratives. We proceeded to use R. Alberich’s paper to define a real social network and show that the Harry Potter character network meets all the criteria of a real social network. Furthermore, many other topological network metrics were very much alike real social networks as we compared 10 social networks in our conclusion section. We also found communities and showed how they are formed according to “group leaders” with high centrality scores as shown by D. Kydros as well. We demonstrated the formation of ties and how house and school year had a very significant effect on the formations. We supported D. Kydros’ papers by showing that the highest vertex metrics belong to significant characters in the story and Newmans definition of a protagonist was validated as the highest centrality scores belonged mostly to Harry Potter, the protagonist.

Our next steps would be to:

- Extract Relationship information to define a relationship as observed or interacted to then replicate A. Agarwal’s Study and detect the point-of-view. (A. Agarwal 2012).
- Apply the same method to a larger artificial social network such as Game of Thrones or Lord of The Rings.
- Derive a quantitative variable describing the “peace level” like in D. A. Skorilkin’s paper to create a correlation table and predict plot dynamics.
- Create a predictive model for plot narrative.

7. References

Dataset 1: https://github.com/efekarakus/potter-network

Dataset 2: http://www.stats.ox.ac.uk/~snijders/siena/siena_datasets.htm

Agarwal, A., Corvalen, A., Jensen, J., & Rambow, O. (2012). Social Network Analysis of Alice in Wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*(pp. 88-96). Montreal: Association for Computational Linguistics.

Alberich, R., Miro-Julia, J., & Rossello, F. (2002). Marvel Universe looks almost like a real social network. ArXiv E-prints.

Skorinkin, D. A. (2017). Extracting Character Networks to Explore Literary Plot Dynamics. In Computational linguistics and intellectual technologies: According to the materials of the annual international conference "Dialogue" (pp. 257-270). Publishing House of the RSUH.

Katzir, L., & Hardiman, S. J. (2015). Estimating Clustering Coefficients and Size of Social Networks via Random Walk. *ACM Transactions on the Web, 9*(4), 1-20. doi:10.1145/2790304

Kydros, D., & Anastasiadis, A. (2014). Social network analysis in literature. The case of The Great Eastern by A. Embirikos. *Social Networking, 06*(02), 164-180. doi:10.4236/sn.2017.62010

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E,69*(2). doi:10.1103/physreve.69.026113

Fothergill, A. (2011). Moretti, Franco. *The Encyclopedia of Literary and Cultural Theory*. doi:10.1002/9781444337839.wbelctv2m009