Network Theory and Text Analysis

Christina West, Taylor Martins, Yihe Hao

with help from: Jerry Luo and Dr. Gabitov

Introduction

- Network Theory has been widely used and broadly researched to analyze various networks.
- Category of networks: Social, Fictional, Biological, electrical and Financial
- With a proper analysed mathematical model, we are able to compare important parameters between networks from a different field
- We are also able to conclude whether there’s any similarities in between
- Successful implementation of this project allows the qualitative components of network to be determined quantitatively

What we have done

- Develop network tree in order to aid visualization of the networking system
- Plot Degree Distribution within the networking system.
- Construct, modify and analyze the model in which we've chosen
- Calculate coefficients within each network
- Stories Picked: Harry Potter & Star Wars (Episode IV)

Igraph Package in R

- Igraph is a convenient software built within R
- With the use of Igraph, we are able to plot the degree distribution and visualize network tree
- Command to summon Igraph in R-Script

```
install.packages("igraph")
install.packages("poweRlaw")
library(igraph)
library(poweRlaw)
```

Sample Look of the CSV file (Edges)

	A	B	C
1	source	target	weight
2	C-3PO	R2-D2	17
3	LUKE	R2-D2	13
4	OBI-WAN	R2-D2	6
5	LEIA	R2-D2	5
6	HAN	R2-D2	5
7	CHEWBACCA	R2-D2	3
8	DODONNA	R2-D2	1
9	CHEWBACCA	OBI-WAN	7
10	C-3PO	CHEWBACCA	5
11	CHEWBACCA	LUKE	16
12	CHEWBACCA	HAN	19
13	CHEWBACCA	LEIA	11
14	CHEWBACCA	DARTH VAD	1
15	CHEWBACCA	DODONNA	1
16	CAMIE	LUKE	2
17	BIGGS	CAMIE	2
18	BIGGS	LUKE	4
19	DARTH VAD	LEIA	1
20	BERU	LUKE	3
21	BERU	OWEN	3
22	BERU	C-3PO	2
23	LUKE	OWEN	3
24	C-3PO	LUKE	18
25	C-3PO	OWEN	2
26	C-3PO	LEIA	6

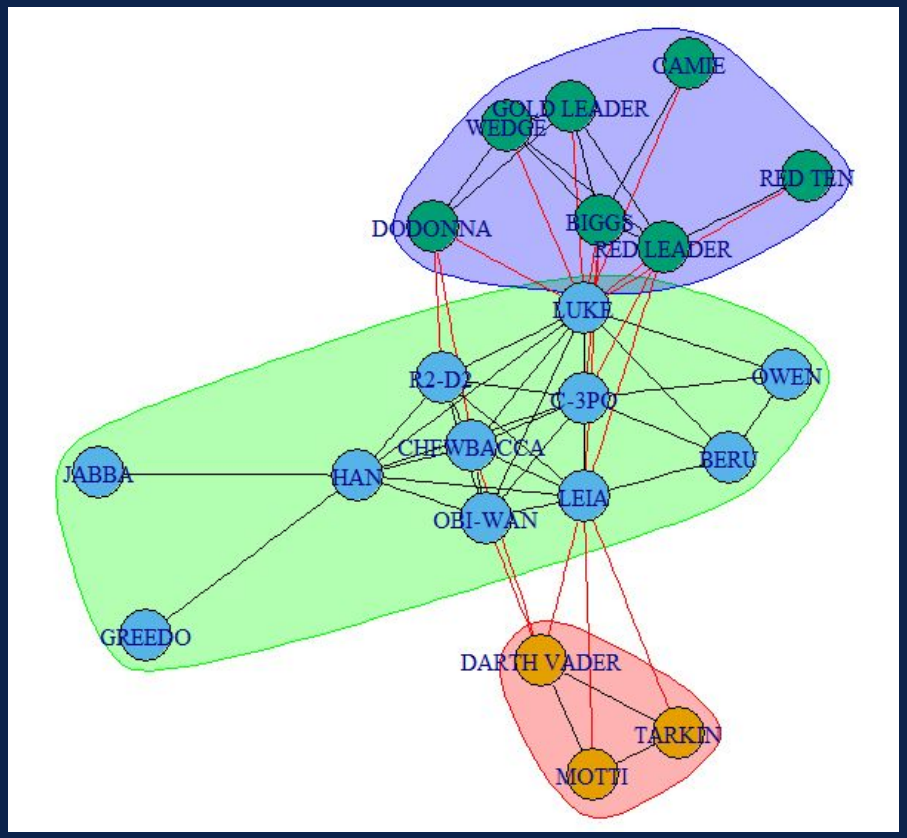
P. (2016, January 20). ablobarbera/data-science-workshop.

	A	B
1	Label	link
2	Hannah Abbott	Hannah_Abbott
3	Ludo Bagman	Ludo_Bagman
4	Bathilda Bagshot	Bathilda_Bagshot
5	Katie Bell	Katie_Bell
6	Cuthbert Binns	Cuthbert_Binns
7	Regulus Black	Regulus_Black
8	Sirius Black	Sirius_Black
9	Amelia Bones	Amelia_Bones
10	Susan Bones	Susan_Bones
11	Terry Boot	Terry_Boot
12	Lavender Brown	Lavender_Brown
13	Millicent Bulstrode	Millicent_Bulstrode
14	Charity Burbage	Charity_Burbage
15	Frank Bryce	Frank_Bryce
16	Alecto Carrow	Alecto_Carrow
17	Amycus Carrow	Amycus_Carrow
18	Reginald Cattermole	Reginald_Cattermole
19	Mary Cattermole	Mary_Cattermole
20	Cho Chang	Cho_Chang
21	Penelope Clearwater	Penelope_Clearwater
22	Michael Corner	Michael_Corner
23	Vincent Crabbe, Sr.	Crabbe
24	Vincent Crabbe	Vincent_Crabbe
25	Colin Creevey	Colin_Creevey

D.(2014,November5).Dpmartin42/Networks

Network Graph

Star Wars Communities (Episode IV)

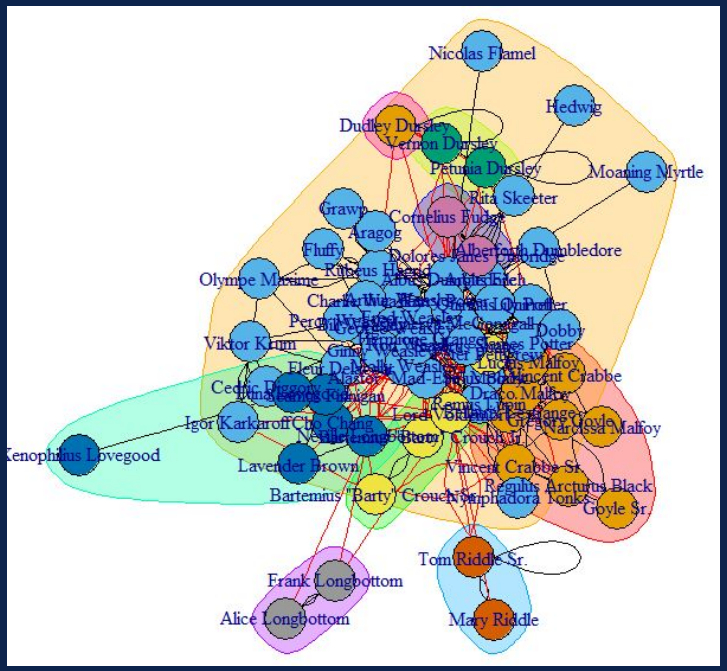


Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

Network Graph

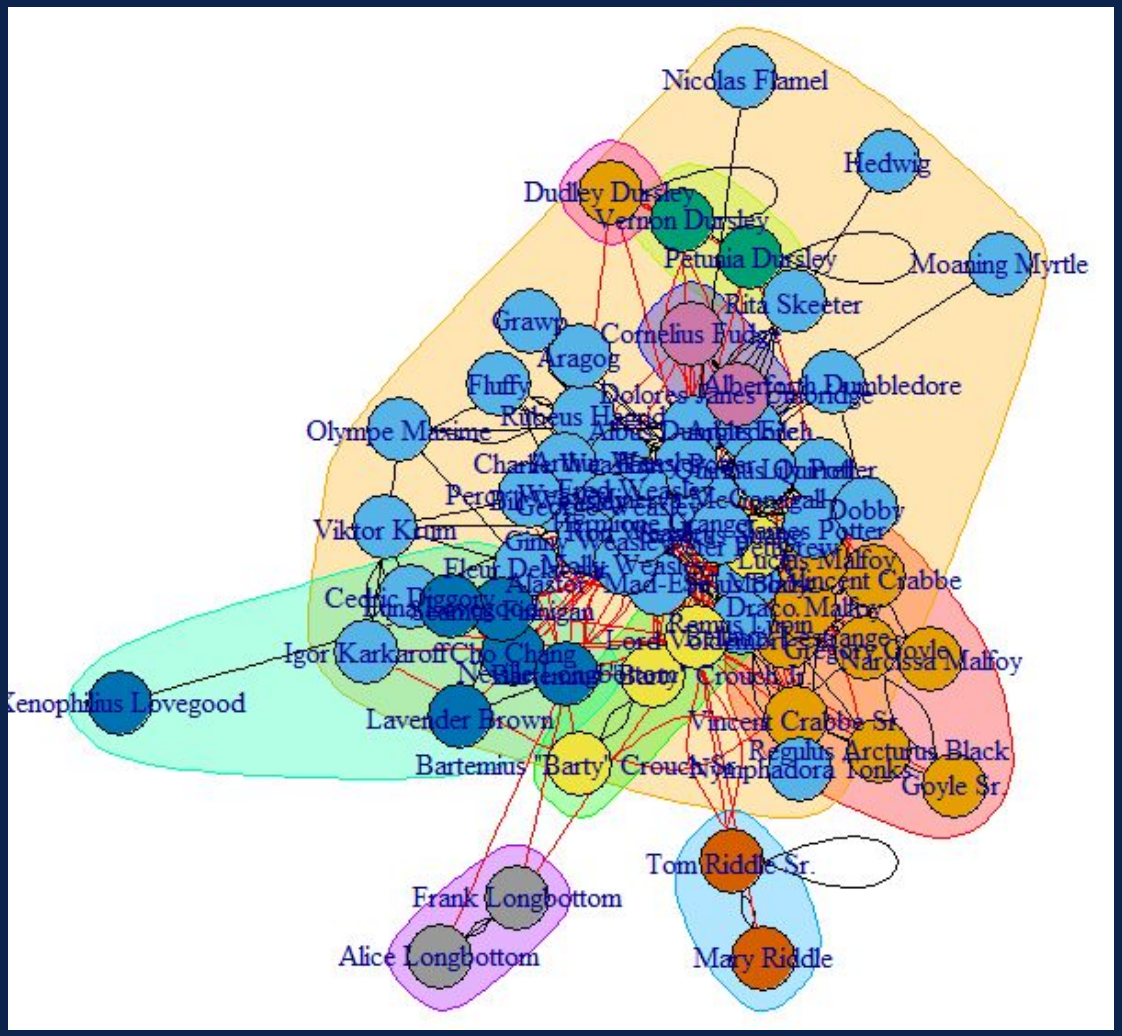
Harry Potter Communities

- communities formed using short random walks (10)



Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

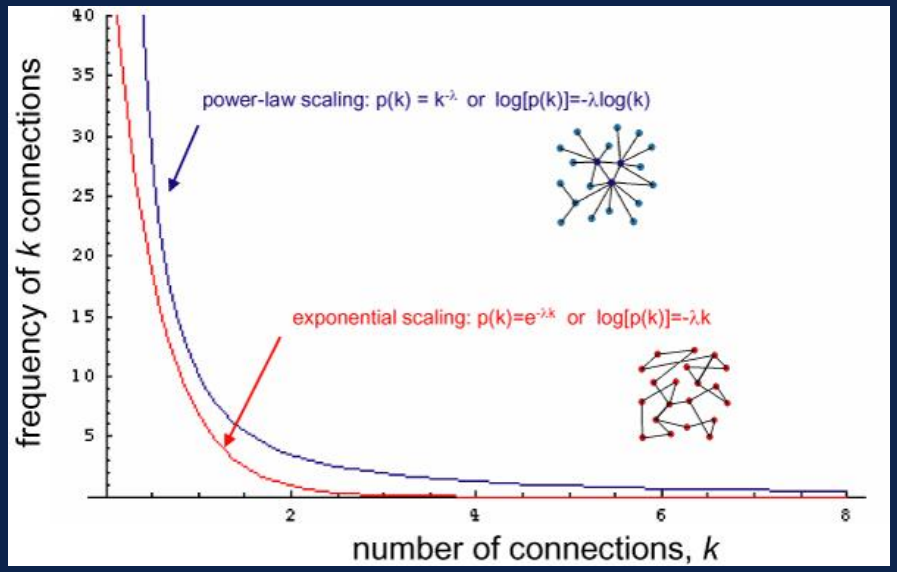
Network Graph



Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

PoweRlaw Package to Fit Distribution

- Contains functions for fitting power laws and other heavy tailed distributions. Overall provides an approach for fitting power laws to data sets.
- Bootstrapping
  - randomly sampling with replacement from a dataset
  - the randomly generated resample is used to make an inference about the original sample
- Vuong Closeness Test
  - a ratio of the log likelihood functions of each distribution
  - the sign of the statistic indicates which model is better

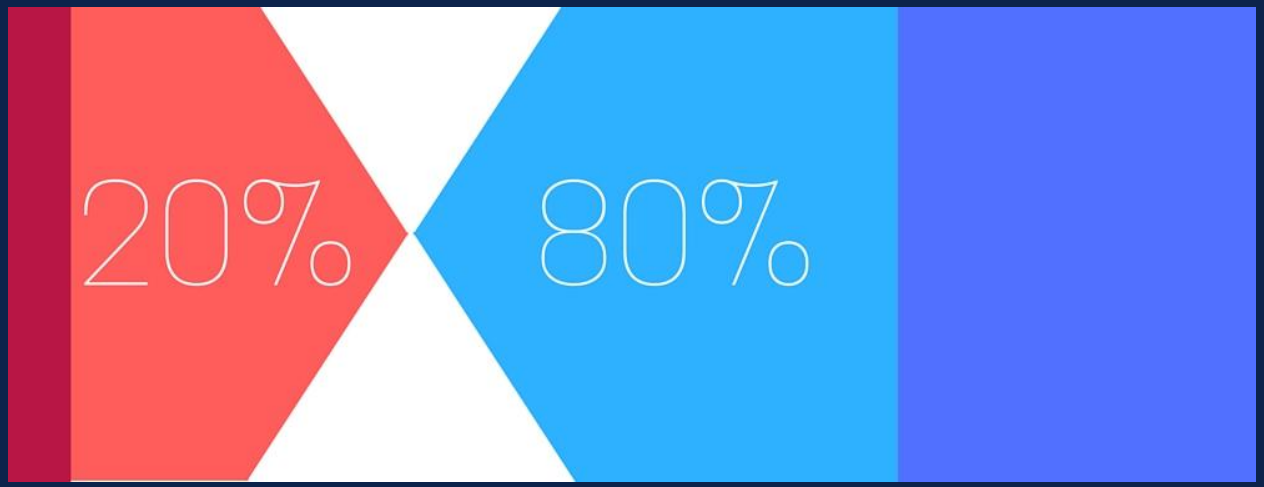


Pagel, Mark & Meade, Andrew & Scott, Daniel. (2007). Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes. BMC evolutionary biology. 7 Suppl 1. S16. 10.1186/1471-2148-7-S1-S16.

Why Power Law?

A network fit to a power law distribution is more representative of a "real world network."

A network fit to an exponential distribution is more representative of a "fictional network."



Pareto Principle - 80% of effects come from 20% of causes. Power law shows long tail and few who dominate.

How We Calculate Coefficients

- Assortativity - Measures Homophyly

$$a = \frac{\sum(jk(e(j,k) - q(j)q(k)), j, k)}{\sigma(q)^2}$$

- Clustering Coefficient - Measures Transitivity

$$c = \frac{|e_{jk} : v_j, v_k \epsilon N_i, e_{jk} \epsilon E)}{k_i(k_i - 1)}$$

- Betweenness Centrality - Number of shortest paths

$$r = \sum(\frac{\sigma_{st}(v)}{\sigma_{st}})$$

- Giant Component - Role of influential vertices

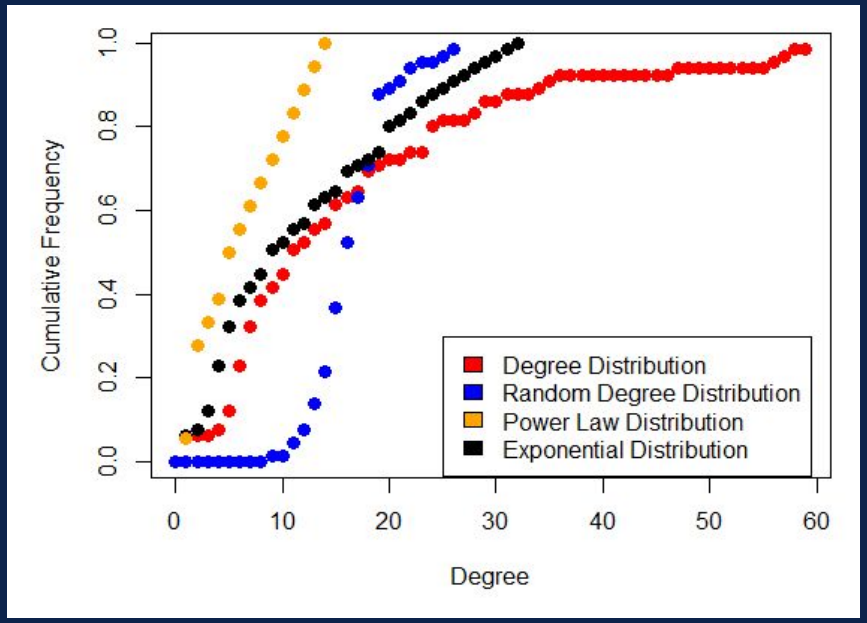
$$g = \frac{max(components)}{N}$$

Model

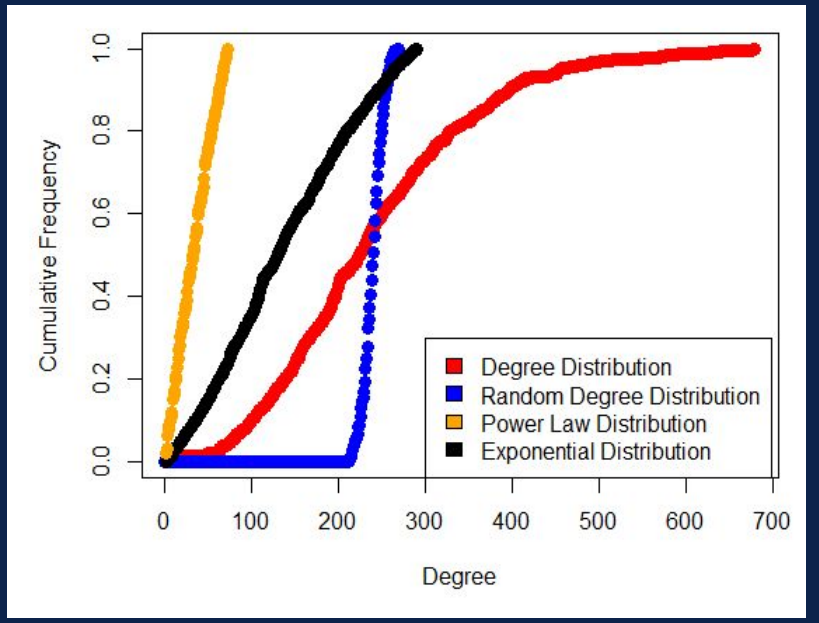
	Star Wars Episode IV	Harry Potter	Congress Twitter Network	Random Networks (Star Wars | HP | Congress)
Number of nodes	22 (60 edges)	65 (513 edges)	517 (61766 edges)	22 | 65 | 517
Diameter	3	4	3	3 | 3 | 2
Characteristic path length	1.9095	2.0284	1.66	1.8381 | 1.7654 | 1.5357
Clustering Coefficient	0.5375	0.4134	0.57	0.2492 | 0.2468 | 0.4629
Degree Distribution	power law (R = 0.1697)	exponential (R = -1.0166)	exponential (R = -1.541)	N/A
Giant Component	0.9545	1.0	0.99	1.0 | 1.0 | 1.0
Assortativity	-0.1934	-0.2069	-0.1027231	0.0730 | -0.0792 | -0.0107

Degree Distribution Graph

Harry Potter Network



Congress Twitter Network



Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

Model

Star Wars Network
- small world ✔
- disassortative ✔
- power law ✔
- giant component > 90% ✔
- but… small sample size

Harry Potter Network
- small world ✔
- disassortative ✔
- exponential ✘
  - small sample size (18)
- giant component > 90 % ✔

Congress Twitter Network
- not small world ✘
- disassortative ✘
- exponential ✔
- giant component > 90 % ✘

Conclusion

Star Wars and Harry Potter Networks shared properties of fictional networks

Unique properties of the Congress Twitter Network
- Not exactly social
- Not exactly fictional
- Artificial

Su, Sharma & Goel (2016)
- In 2010, Twitter's "Friend of a friend" feature caused popular accounts to get more popular

Areas for more investigation
- investigation into twitter networks v. other social networks
- further analysis of congressional data set (party affiliation, chamber, etc.)

References

Colin S. Gillespie (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. Journal of Statistical Software, 64(2), 1-16. URL http://www.jstatsoft.org/v64/i02/.

D.(2014,November5).Dpmartin42/Networks.From https://github.com/dpmartin42/Networks/tree/master/Harrypotter/data

Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

P.(2016, January 20). ablobarbera/data-science-workshop. Retrieved from https://github.com/pablobarbera/data science-workshop/blob/master/sna/data

Padraig Mac Carron and Ralph Kenna, Universal Properties of Mythological Networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002

Su, J., Sharma, A., & Goel, S. (2016, April). The effect of recommendations on network structure. In Proceedings of the 25th international conference on World Wide Web (pp. 1157-1167). International World Wide Web Conferences Steering Committee.