Network Theory and Text Analysis

Christina West, Taylor Martins, Yihe Hao

The University of Arizona

Introduction

Network Theory -

The study of complex interacting systems that are represented as graphs with different structures.

Graphs -

A set of vertices representing some data connected by edges.

Text Analysis -

Text analysis is a type of content analysis involving the systematic reading of text to indicate the presence of meaningful patterns.

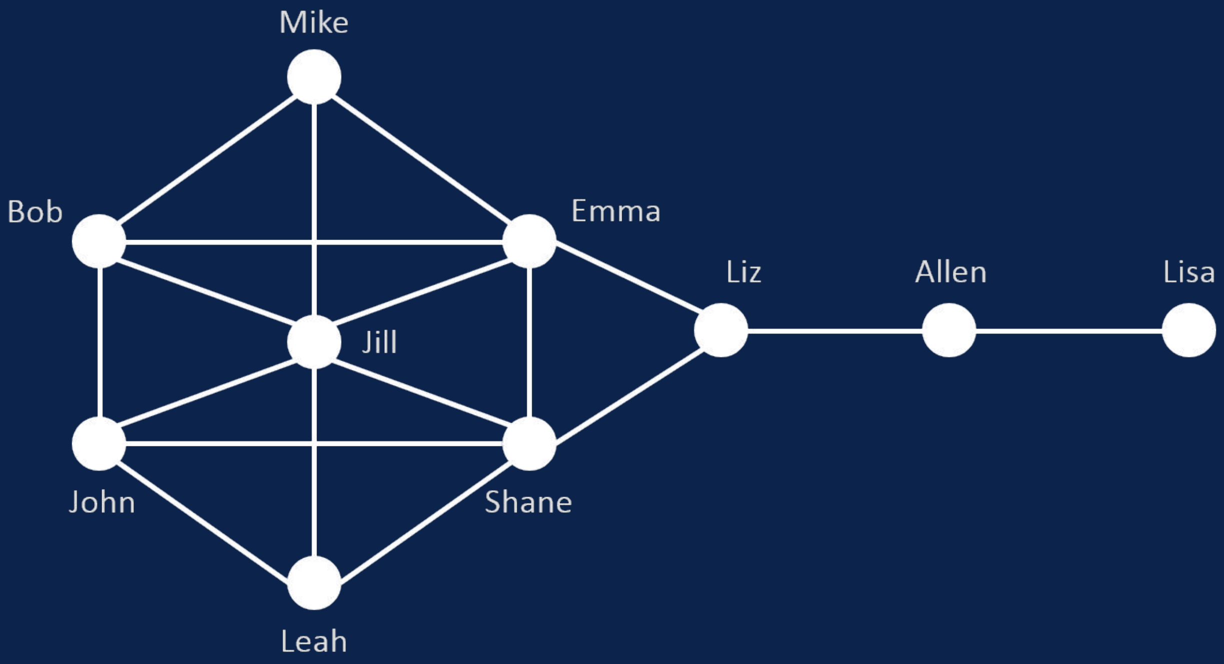
Motivation -

Network theory is used to analyze various systems that can be represented in graph format. This includes social networks, biological interactions, resource flow, electrical systems and more. For text analysis we can compare the patterns in text to other existing networks.

Description of the Project

Main Goals -

- Utilize various tools of network theory to analyze the network of characters in mythological stories.
- Compare the fantasy character networks to other networks such as a real world social network.



Analyzation Method

- We have data regarding the connections in each of three different types of mythological stories
  - Folktales, Legends, Epic

- Examples from each category:
  - Folktales: *Tain Bo Cuailnge (12 - 14 century)*
  - Legends: *The Iliad (8 century B.C.)*
  - Epic: *Beowulf (8 - 11 century)*

- We can use the methods employed for analysis of these stories to determine the relationships present in other networks.
  - Comparisons of important parameters.
  - Data is gathered from online resources (csv) of existing networks or otherwise compiled from reading a story.

Applications of the Project

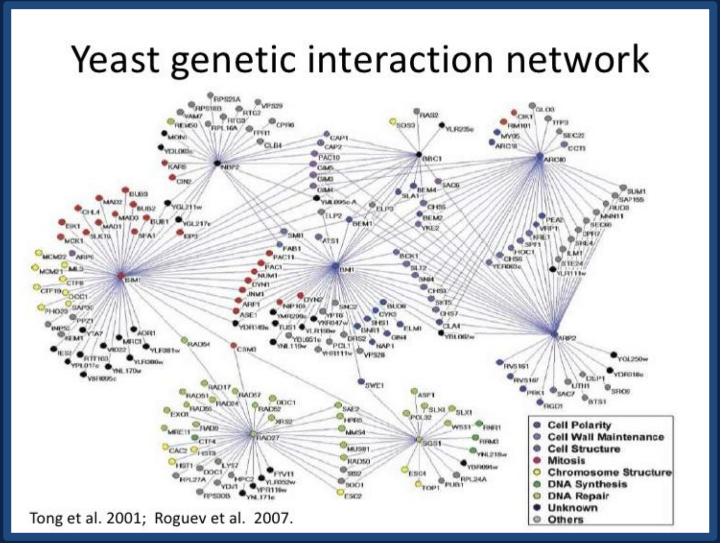
Numerous measures can be employed to investigate the characteristics of each network.

- path length - distance between vertices
- clustering coefficient - measures transitivity
- giant component - role of influential nodes
- degree distribution - connections between nodes

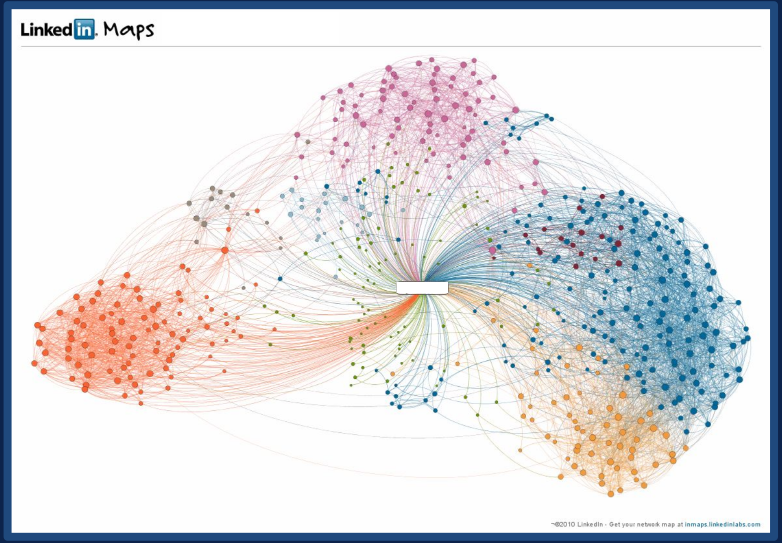
Investigating these traits allows the qualitative components of networks to be determined quantitatively.

Applications of the Project

We can calculate these parameters in order to determine the characteristics of various networks. These can point to similarities and differences between different networks.



Tong et al. 2001; Roguev et al. 2007.



©2010 LinkedIn - Get your network map at inmaps.linkedinlabs.com

What is the Model

- Statistical coefficients
    - Global
        - *N*: number of nodes
        - *l*: characteristic path length
        - *lmax*: longest geodesic
    - local
        - *C*: clustering coefficient
    - mean degree <*k*>
    - *p*(*k*) degree distribution
    - *Gc:* Giant Component
    - *gl*: betweenness centrality

- Network characteristics of social networks
    - Small World
    - Scale free
    - Hierarchical
    - Giant Component < 90%
    - Vulnerable to targeted attack
    - Assortativity

Small World

- more clustered than a random network of the same size

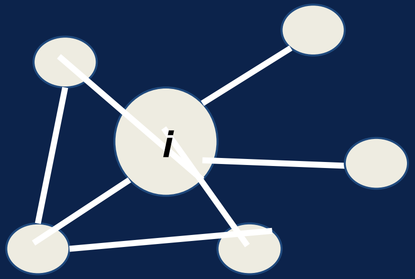
characteristic path length *l*

- average minimum separation between pairs of nodes
- the maximum shortest path is called the diameter

clustering coefficient C

- average over all the *N* nodes

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i, \ where\ C_i = \frac{2n_i}{k_i(k_i+1)}$$



$k = 5$, $n = 2$, $C = 0.2$

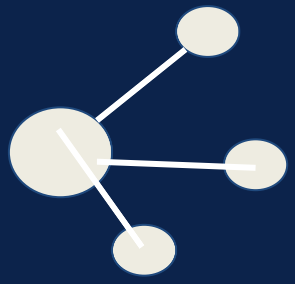
Small World

- $l \approx l_{rand}$
- $C \gg C_{rand}$
- with l_{rand} and C_{rand} coming from a random network of same number of nodes N

Degree Distribution

- A node with *k* connections has degree *k*.
- Degree distribution is given by the probability that any given node in the network has degree k.
- can plot degree *k* vs. probability p(*k*)
- a network is **scale free if**
  $p(k) \sim k^{-\gamma}$, with $2<\gamma\leq 3$

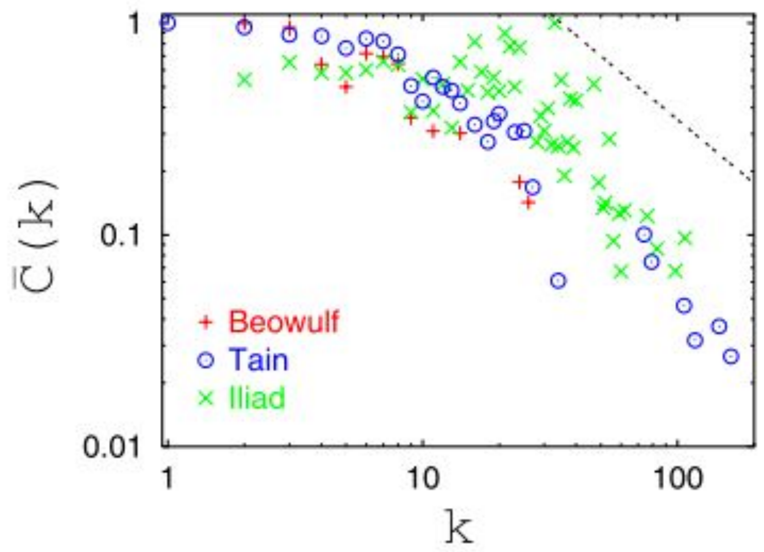
  a random network has
  $p(k) \sim e^{-k}$



k = 3

Hierarchical Structure

- A network has hierarchical structure if the correlation coefficient follows the power law, C(k) ~ 1/k

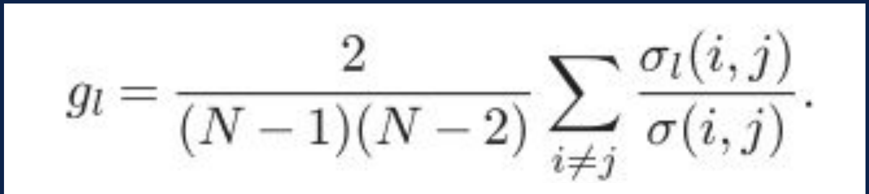


Padraig Mac Carron and Ralph Kenna, Universal properties of mythological networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002

Betweenness Centrality

- $\sigma(i,j)$ is the number of shortest paths between node i and node j, with $\sigma_l(i,j)$ of these passing through node l

- the betweenness centrality of node l is given by



$$g_l = \frac{2}{(N-1)(N-2)} \sum_{i \neq j} \frac{\sigma_l(i,j)}{\sigma(i,j)}.$$

Padraig Mac Carron and Ralph Kenna, Universal properties of mythological networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002

- normalization guarantees that g_l=1 if every shortest path goes through node l.

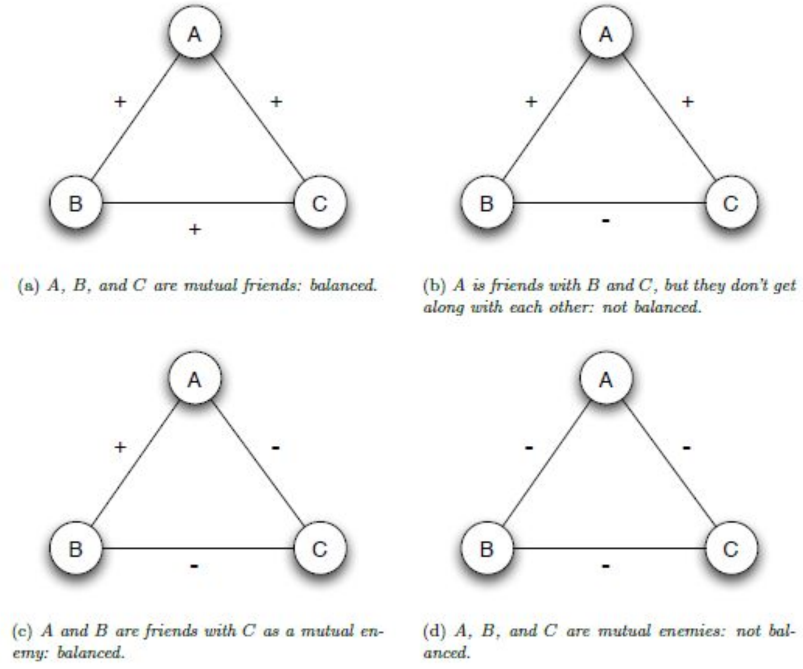
Giant Component

- Giant component: What percentage of the total nodes are connected to each other?
  - e.g., In *Tain* and *Iliad,* 98%, In *Beowulf,* 68%

- If removing the top 5% of nodes with the highest betweenness centrality significantly reduces the giant component, the network is vulnerable to targeted attack

- The network is robust if the giant component is unaffected the removal of nodes

Structural Balance



(a) *A, B, and C are mutual friends: balanced.*

(b) *A is friends with B and C, but they don't get along with each other: not balanced.*

(c) *A and B are friends with C as a mutual enemy: balanced.*

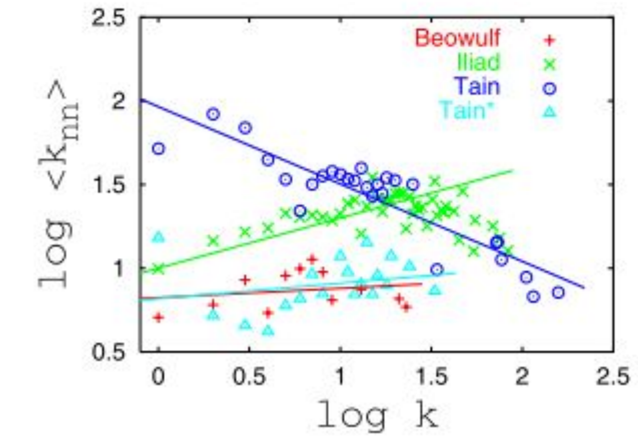
(d) *A, B, and C are mutual enemies: not balanced.*

Figure 5.1: Structural balance: Each labeled triangle must have 1 or 3 positive edges.

David Easley and Jon Kleinberg, "Positive and negative relationships," in Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010

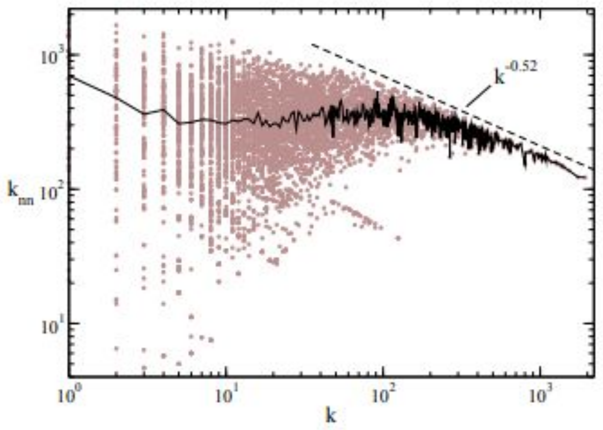
Assortativity

- Assortativity by degree: highly connected notes associate with other highly connected nodes
- Pearson correlation coefficient, *r*
- plot the degree of each node against the mean degree of its neighbors
- social networks tend to be assortative, while fictional networks tend to be disassortative



Padraig Mac Carron and Ralph Kenna, Universal properties of mythological networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002

Negative slope -> disassortativity in the Marvel Universe



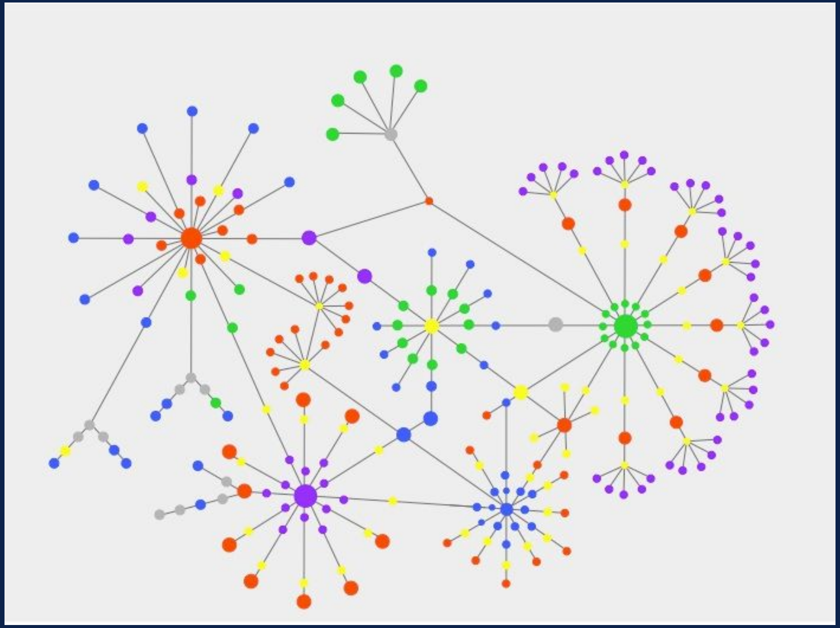
Gleiser P. M., J. Stat. Mech. (2007) P09020.

What is the Model

Social Networks	Mythological Networks	Fictional Networks
Small World	?	Small World
Hierarchical	?	Hierarchical
Power Law dist.	?	Exponential dist.
Scale Free	?	Not scale free
Assortative	?	Not assortative
Giant component < 90%	?	Giant component > 90%
Vulnerable to targeted attack	?	Resilient against targeted attack

Network Tree

- A network tree is the most direct way for us to find out the connections
  - Pros: Able to identify network relationship virtually
  - Cons: Hard to run Statistical Analysis



What are we going to do?

- We are going to analyze the connection in between only one of the mythological stories
    - Division: Folktales, Legends, Epic
- Idea for the story to choose from:
    - Charlie and his chocolate factory
    - King Arthur in Combat
    - Star Wars

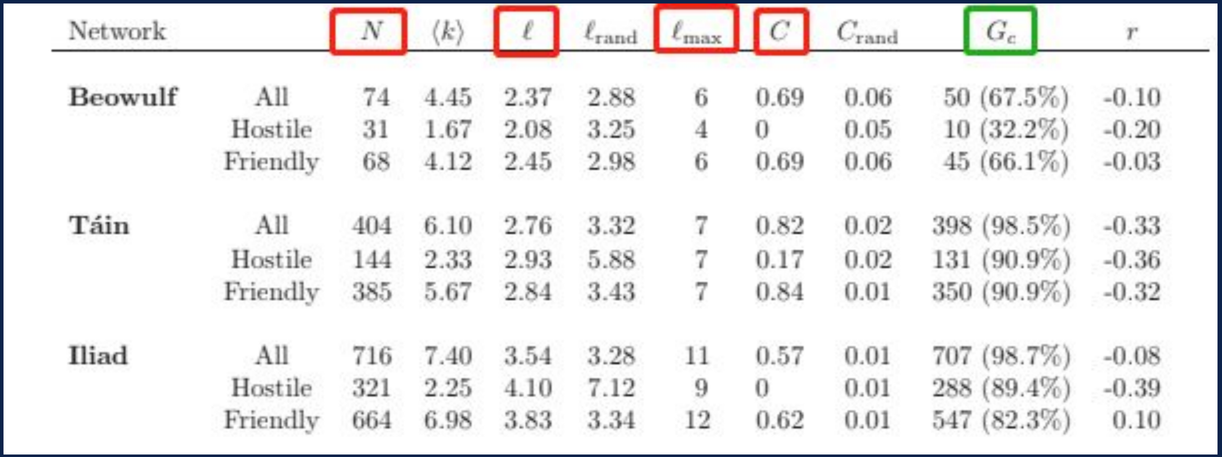
Sample Input

Network	Network Description	Global			Local	All other Components		
		N	L	L_max	C	p(k)	G_c	gl
Story	All Hostile Friendly							

N: Number of Nodes
L: Characteristic path length
L_max: Longest geodesic
C   : Clustering Coefficient
P(k): Degree Distribution
G(c): Giant Component
gl: Centrality

- Gather Data for All Parameters
- Calculate G(c), gl
- Plot Degree Distribution in R

Sample Input



Network		N	$\langle k \rangle$	ℓ	ℓ_{rand}	$\ell_{\max}$	C	C_{rand}	G_c	r
Beowulf	All	74	4.45	2.37	2.88	6	0.69	0.06	50 (67.5%)	-0.10
	Hostile	31	1.67	2.08	3.25	4	0	0.05	10 (32.2%)	-0.20
	Friendly	68	4.12	2.45	2.98	6	0.69	0.06	45 (66.1%)	-0.03
Táin	All	404	6.10	2.76	3.32	7	0.82	0.02	398 (98.5%)	-0.33
	Hostile	144	2.33	2.93	5.88	7	0.17	0.02	131 (90.9%)	-0.36
	Friendly	385	5.67	2.84	3.43	7	0.84	0.01	350 (90.9%)	-0.32
Iliad	All	716	7.40	3.54	3.28	11	0.57	0.01	707 (98.7%)	-0.08
	Hostile	321	2.25	4.10	7.12	9	0	0.01	288 (89.4%)	-0.39
	Friendly	664	6.98	3.83	3.34	12	0.62	0.01	547 (82.3%)	0.10

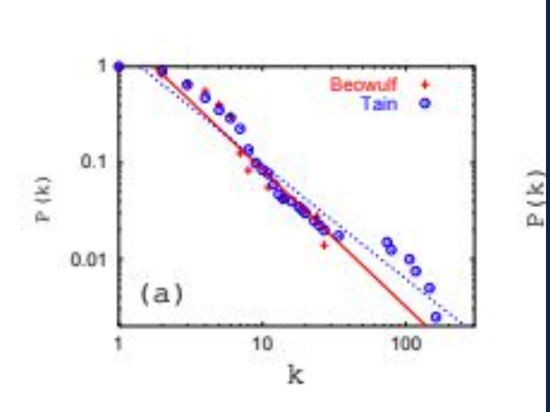
Input Data

N: Number of Nodes
L: Characteristic path length
L_max: Longest geodesic
C : Clustering Coefficient
P(k): Degree Distribution
G(c): Giant Component
gl: Centrality

- Gather Data for All Parameters
- Calculate G(c), gl
- Plot Degree Distribution in R

Expectations

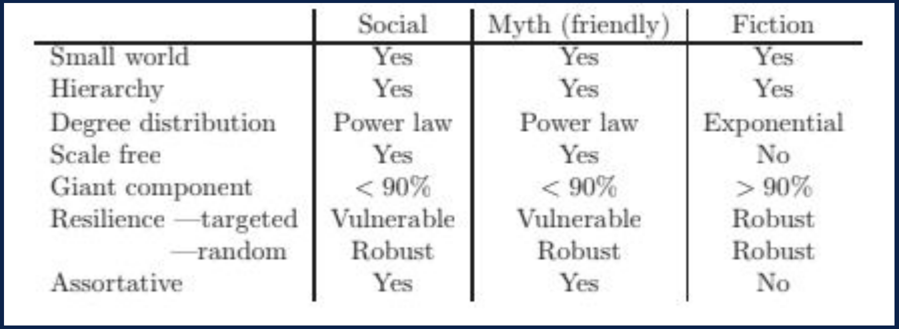
- After gathering all Connection data, we are going to see the relationship in between Social, Mythological and Fictional Networks
  - Analyzation Method: Degree Distribution by Using R (Chi-Sq Test)



Degree Distributions Graph

Expectations

- We are also going to create a summary of properties and compare it with the other networks
  - By doing so would help us to understand the connections in between different network categories.
  - The data for Social as well as Fiction Networks will be given



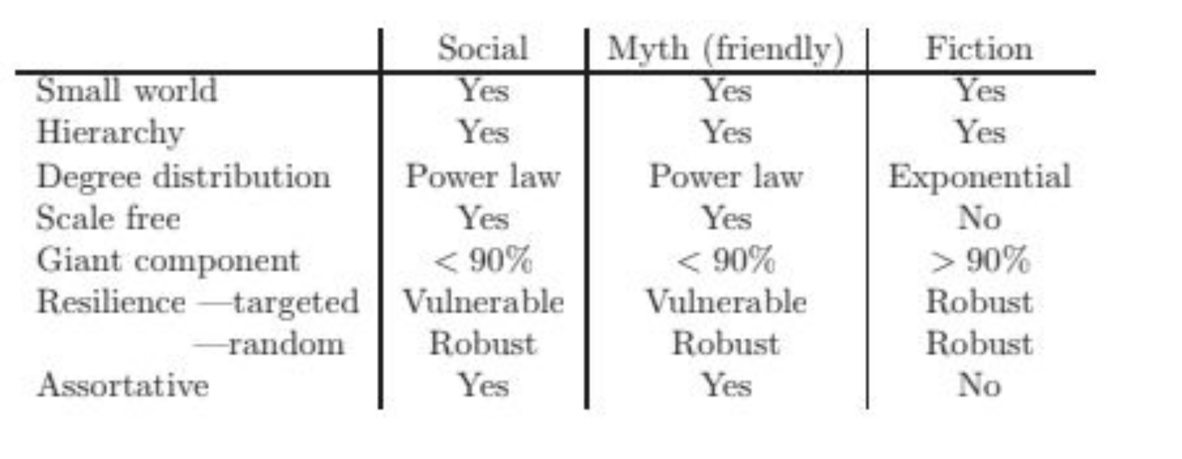
	Social	Myth (friendly)	Fiction
Small world	Yes	Yes	Yes
Hierarchy	Yes	Yes	Yes
Degree distribution	Power law	Power law	Exponential
Scale free	Yes	Yes	No
Giant component	$< 90\%$	$< 90\%$	$> 90\%$
Resilience —targeted	Vulnerable	Vulnerable	Robust
—random	Robust	Robust	Robust
Assortative	Yes	Yes	No

In conclusion

- Analyze one type of mythological stories’ network
- Compute the degree of Distribution and Plot it in R
- Compare the summary with the other two categories of network
  - See if there are any similarities in between each other

In the future

- Political Network
  - Allies, Enemies
- Biological Network
  - Food Chain



	Social	Myth (friendly)	Fiction
Small world	Yes	Yes	Yes
Hierarchy	Yes	Yes	Yes
Degree distribution	Power law	Power law	Exponential
Scale free	Yes	Yes	No
Giant component	$< 90\%$	$< 90\%$	$> 90\%$
Resilience —targeted	Vulnerable	Vulnerable	Robust
—random	Robust	Robust	Robust
Assortative	Yes	Yes	No

Distribution of Work

- Christina: Build a model and compare it with the other two networks
- Taylor: Gather data from a mythological story
- Yihe: Construct, organize the CSV file and build the code base in R for the Chi-Sq test and Distribution Plot

Reference

- Baez J. “Network Theory” for the Azimuth Project. http://math.ucr.edu/home/baez/econ.pdf. (2012)
- David Easley and Jon Kleinberg, "Positive and negative relationships," in Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010
- Gleiser P. M., J. Stat. Mech. (2007) P09020
- “Innovation Networks.” *Gilburg Leadership Incorporated*, ilburgleadership.com/services/innovation-networks.
- Padraig Mac Carron and Ralph Kenna, Universal properties of mythological networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002