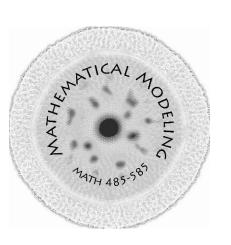# Network Theory & Text Analysis

Project Description

- Network Theory has been widely used and broadly researched to analyze various networks.
- The main goal of this project is to utilize various tools of network theory to analyze the network of characters in fictional stories.
- With the mathematical model, we are able to compare fictional networks to networks from different fields.
- Networks included are, but not limited to, social, biological, resource and electrical systems.
- Network trees have also been implemented in order to aid visualization of the networking system.

Scientific Challenges

- It is important to develop a model for this specific research project. By using the model we've developed, we are able to compare important parameters between social and fictional networks. Furthermore, we are able to conclude whether there's any similarities between them.

Potential Applications

- Successful implementation of this project allows the qualitative components of network to be determined quantitatively.
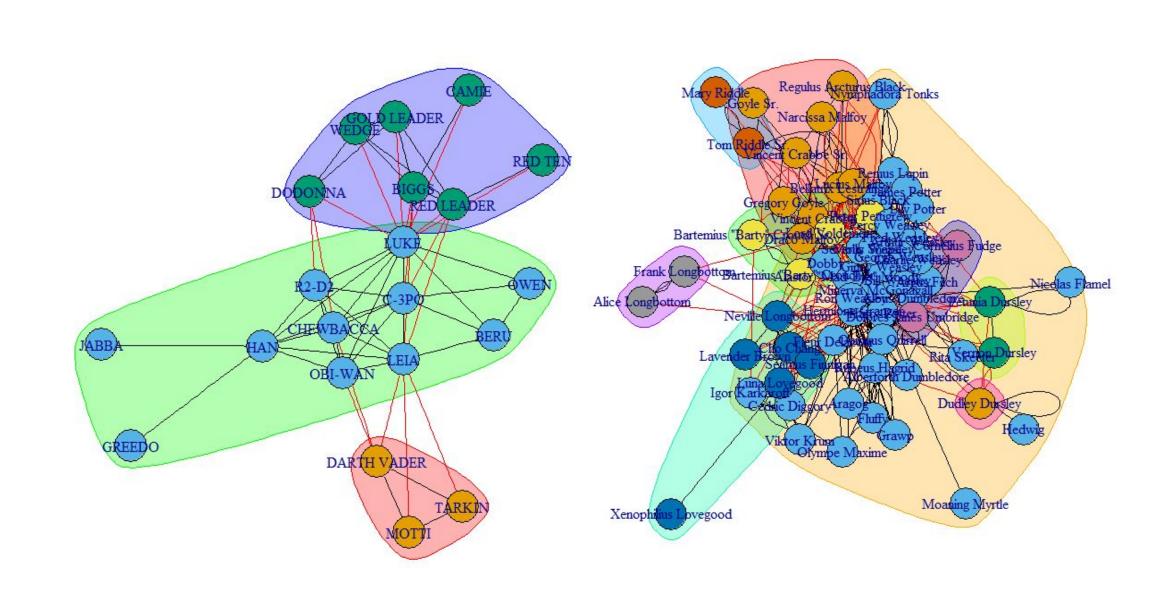

Figure 1: Network Connection in between Star-War [4] & Harry Potter [2] Characters

Team Members:
Christina West, Taylor Martins, and Yihe Hao

Methodology

Calculation of coefficients: (Using the igraph package)

- assortativity - measures homophyly

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

e-Edges connecting vertex j and k.
q-sum of edges

- clustering coefficient - measures transitivity

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

v-vertex, N-neighboring vertices, E-all edges, k-vertex degree

- diameter - longest path length in the graph
  - Breadth-first search algorithm
- characteristic path length, l - average distance between nodes
- betweenness centrality - number of shortest paths

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

sigma-shortest paths s to t,
sigma(v)-shortest paths through v

- degree distribution - probability of node with particular degree
  - Calculated as empirical cdf

- giant component - role of influential vertices

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

components-depth-first search for connected vertices, N-total number of vertices

Random networks were computed using *sample_gnm* (constant probability with number of nodes and edges equal to comparison network)

Results

	Star Wars Episode IV [3]	Harry Potter[4]	Congress Twitter Network	Random Networks (Star Wars | HP | Congress)
Number of nodes	22	65	517	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 (KS = 0.1697, p = 0.8653)	exponential (KS = -1.0166, p = 0.3094)	exponential (KS = -1.541, p = .1234)	N/A
Giant Component	0.9545	1.0	0.99	1.0 | 1.0 | 1.0
betweenness centrality	15.4932	Error	168.88	9.7727 | 24.4923 | 168.882
Assortativity	-0.1934	-0.2069	-0.1027231	0.0730 | -0.0792 | -0.0107

Results (Continued):

1. As expected, the Star Wars network is small world ($l \cong l_{rand}$ and $C >> C_{rand}$), disassortative, has power law distribution and has a giant component > 90% (all characteristics of fictional networks)
2. Surprisingly the twitter network is not small world ($C \cong C_{rand}$), is disassortative, and has giant component > 90%. (unusual for social networks)

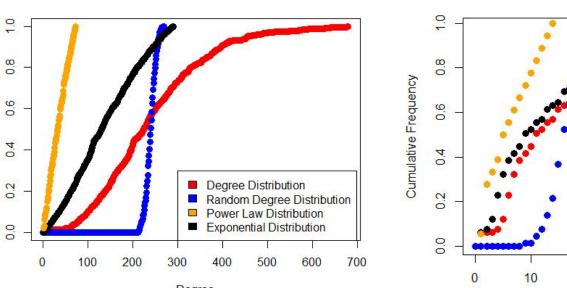## Glossary of Technical Terms
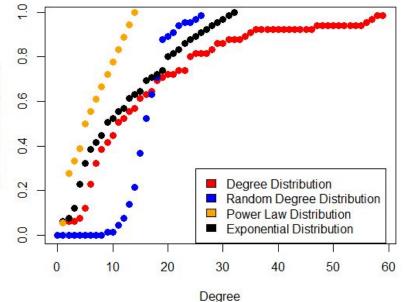
Vertex: Particular person or character in a network
Edge: Line denoting related characters
Degree: Number of edges incident to a vertex
Homophyly: Relatedness from common source vertex
Transitivity: Strength of grouping for connected vertices
Component: Subgraph with path between vertices

Dynamics of the Degree Distribution (Congress Twitter[4] & Harry Potter [2] , plotted with the software R Studio [3].

References

1. 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/.
2. D.(2014,November5).Dpmartin42/Networks.From https://github.com/dpmartin42/Networks/tree/master/Harrypotter/data
3. Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.
4. P.(2016, January 20). ablobarbera/data-science-workshop. Retrieved from https://github.com/pablobarbera/data science-workshop/blob/master/sna/data
5. Padraig Mac Carron and Ralph Kenna, Universal Properties of Mythological Networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002