Mining and modeling character networks*

Anthony Bonato[1], David Ryan D'Angelo[1],
Ethan R. Elenberg[2], David F. Gleich[3], and Yangyang Hou[3]

[1] Ryerson University
[2] University of Texas at Austin
[3] Purdue University

Abstract. We investigate social networks of characters found in cultural works such as novels and films. These *character networks* exhibit many of the properties of complex networks such as skewed degree distribution and community structure, but may be of relatively small order with a high multiplicity of edges. Building on recent work of [4], we consider graph extraction, visualization, and network statistics for three novels: *Twilight* by Stephanie Meyer, Steven King's *The Stand,* and J.K. Rowling's *Harry Potter and the Goblet of Fire.* Coupling with 800 character networks from films found in the `http://moviegalaxies.com/` database, we compare the data sets to simulations from various stochastic complex networks models including random graphs with given expected degrees (also known as the Chung-Lu model), the configuration model, and the preferential attachment model. Using machine learning techniques based on motif (or small subgraph) counts, we determine that the Chung-Lu model best fits character networks and we conjecture why this may be the case.

1 Introduction

Complex networks lie at the intersection of several disciplines and have found broad application within the study of social networks. In social networks, nodes represents agents, and edges correspond to some kind of social interaction such as friendship or following. For more on complex networks and on-line social networks, the reader is directed to the book [5] and the survey [6].

In the present paper, we consider social networks arising in the context of cultural works such as novels or movies. In these *character networks*, nodes represent characters in a specified fictional or non-fictional work such as a novel, script, biography, or story, with edges between characters determined by their interaction within the work. We also consider character networks as weighted graphs, where the weights are positive integers specifying the co-appearance or co-occurrence of character names within a specified range of the text or scenes (such as being within fifteen words of each other; see [4]). Not surprisingly, character networks are typically of smaller order than many other types of complex networks. Nevertheless, they still exhibit many of the interesting features of complex networks including clearly defined community structure, with communities centered on the various protagonists of the story, skewed degree distributions, focused on the most important characters, and dynamics. Character networks defined over larger fictional universes, such as the Marvel Universe, even grow to over 10,000 nodes [1,12].

There is an emerging approach using the tools of graph theory and big data to mine and model character networks. This new topic reflects the ease of access of cultural works in electronic formats, and the efficacy of big data-theoretic algorithms. Our approach in this work is study new networks with these tools to replicate some of the findings as well as study network models of these data.

First, we wish to study the complexity of these character networks through graph mining. Our approach here is more a microscopic view of an individual work's network.

* Research supported by grants from NSERC and Ryerson University; Gleich and Hou's work were supported by NSF CAREER Award CCF-1149756, IIS-1546488, CCF-093937, and DARPA SIMPLEX.

We focus on three well known novels: *Twilight* by Stephanie Meyer, Steven King's *The Stand*, and J.K. Rowling's *Harry Potter and the Goblet of Fire*. Various complex network statistics, such as diameter and clustering coefficient, are presented along with centrality metrics (such as PageRank and betweenness, paralleling the approach of [4]) that predict the major characters within each book. See Section 2 for the methodology used, and Section 3 has a summary of our results.

The second part of our approach is to compare and contrast the character networks with several well known stochastic network models. Hence, in this approach, we take a broader, macroscopic view of the structure of a larger sample of character networks. Using motifs (that is, small subgraph counts), eigenvalues, and machine learning techniques, we develop an approach for model selection for character networks. The models considered were the configuration model, preferential attachment model, the Chung-Lu model for random graphs with given expected degree sequences, and the binomial random graph (as a control). The parameters of the models were chosen as to equal the number of nodes and average degree of the character network data sets. Model selection was conducted for the three novels described above, and also for a set of 800 networks arising from movies in the `http://moviegalaxies.com/` database. Our results show consistent selection of the Chung-Lu model as the most realistic, with a clear separation between the models. We will discuss possible interpretations and implications of our results in the final section.

We consider undirected graphs throughout the paper. For background on graph theory, the reader is directed to [21]. Additional background on machine learning can be found in [13,19].

1.1 Previous work

Quantitative methods have now emerged as a modern tool for literary analysis. Literary theories are now supported, debated, and refuted based on data [10]. In recent work, Reagan et al. [16] implement data mining techniques inspired by Kurt Vonnegut's theory of the shape of stories. Vonnegut suggested graphing fictional works based on the fortune of the main character's experiences over the passage of time in the story. Using text sentiment analysis, Reagan et al. scored the emotional content over the course of a novel based on the occurrence of select words in the labMT data set for 1,737 books from the Project Gutenberg database. They found the majority of emotional arcs resided in six classes. Through the analysis of 60 different novels, including Jane Austin's *Pride and Prejudice*, Dames et al. determined that the narrative is a good predictor for network structure within the novels.

In [4], Beveridge and Shan applied network algorithms on the social network they generated from *A Storm of Swords*, the third novel in George R.R. Martin's *A Song of Ice and Fire* series (which is the literary origin of the HBO drama *Game of Thrones*). Metrics such as PageRank, closeness, betweenness centrality, and modularity provided an empirical approach to determine communities and key characters within the network. Work done by Ribeiro et al. [17] focuses on examining structural properties, such as assortivity and transitivity, of communities in the social network of J.R.R. Tolkien's *The Lord of the Rings* (which included that unabridged novel, along with text from *The Hobbit* and *The Silmarillion*). Beyond static networks, Agarwal et al. [2] analyze the dynamic network for *Alice in Wonderland*, defined by the mining of the ten chapters independently of each other. Such analysis may be important in determining importance characters whose metrics score low over the passage of time in the novel, but are significantly important for parts of the story. Deviating from the extraction of character networks, Sack [18] provides

a social network generation model for narratives through the concept of *structural balance theory* using signed edges between characters.

2 Experimental design and methods

The twin goals of our experiments are to highlight some of the complexities present in character networks via their network properties and to determine a possible synthetic model of the character networks.

2.1 Network properties

We use the Gephi open source software package to extract communities and compute various network statistics from character networks. These analyses are all done on weighted, undirected, graphs. For community analysis, we use modularity and the Louvain method. Centrality measures are a classic tool in social network analysis to determine the important individuals. They have been found to also serve the same role in character networks. We consider weighted degree, closeness, betweenness, eigencentrality, and PageRank centrality. We briefly review these methods; see [5] for more background on complex network properties. The *closeness* of a node u is the average distance between u and all other nodes (here distance is the standard shortest path metric in graph theory). The *betweenness* of u is the proportion of shortest paths that transit through an u as an intermediate node. The *eigencentrality* of u is its corresponding coordinate in the largest eigenvector of the weighted adjacency matrix. PageRank centrality is based on the stationary distribution of a random walk on the network that periodically teleports to a node chosen uniformly at random.

2.2 Model selection

The goal of our model selection experiments is to determine a random graph model that matches empirically observed properties of character networks. Our methodology is to create a compact summary of the network statistics that is invariant to the labeling of the nodes of the network. In other words, we would derive the same statistics if we permuted the adjacency matrix. The summaries we use are the 3-profile, 4-profile, and eigenvalue histogram. The *k-profile* of a graph G counts the number of times each graph on k nodes appears as an induced subgraph of G; see [8]. An *eigenvalue histogram* is a histogram of the eigenvalues of the normalized Laplacian matrix, which all lie between 0 and 2, with equally spaced bins. These techniques are well established in model selection for various types of biological and social networks [6,14].

In contrast to the previous section we use undirected, unweighted graphs for this experiment. This choice reflects our goal to model the connectivity of the networks, rather than their joint connectivity and weight structure.

We use the algorithm in [9] to compute a global graph 4-profile for each character network. This is a generalization of graphlets [15,20], a similar method of motif counting for connected subgraphs. One difference is that the 4-profile includes disconnected graphs as well. We use standard algorithms for computing all eigenvalues of the normalized Laplacian where we treat the normalized Laplacian as a dense matrix. We compute a histogram based on five equally spaced bins.

We examine the following random graph models on n nodes, with parameters chosen to match those of the original character network:

1. *Preferential Attachment (PA).* In the PA model, at each step, a node is added to the graph and m edges are placed from the new node to existing nodes. These edges are chosen with probability proportional to the degree of each node before the new node arrived. If m is chosen such that

$$\frac{2}{n} + 2m = \frac{2|E|}{n},$$

then the number of edges will match that of the original graph in expectation.
2. The *Binomial Random Graph* $G(n, p)$, or *Erdős -Rényi (ER)* model. Each of the $\binom{n}{2}$ edges is connected according to an independent binomial random variable with probability p proportional to the expected average degree. We use $p = |E|/\binom{n}{2}$ to match the average degree of the original network.
3. The *Chung-Lu (CL)* model. The CL model generalizes the binomial random graph model to non-uniform edge probabilities. Graphs in this model are parameterized by an expected degree distribution (the character network's true degree distribution) rather than a scalar average degree. Each edge is connected with probability proportional to the product of the expected degrees w_i of its endpoints:

$$p_{ij} = \frac{1}{C} w_i w_j.$$

4. The *Configuration Model (CFG).* In the CFG model, we select a graph uniformly from the set of graphs which exactly match the target degree distribution. In practice, the degree distribution may vary slightly from the target since we disregard self loops and multi-edges created during this process.

Our method to determine the best random graph model fitting the data is to generate samples and train a machine learning algorithm to identify each model. We then we ask the algorithm to classify the real graph. First, 100 random graphs from each model are used to train a machine learning classifier. Then in the test step, the classifier predicts a class label for the original character network. This provides a measure of which random graph model best fits the character network. We study the following machine learning algorithms: two variants of linear classifiers (SVMs) and two ensemble methods based on decision trees (Random Forests and Boosted Decision Trees). For more about these models, see [13].

1. *Support Vector Machines (SVM).* The SVM algorithm is a simple way to classify points in Euclidean space. Geometrically, the binary SVM classifier is defined by a hyperplane $\mathbf{w}$ that maximally separates points from both classes on either side. This problem can be formulated as a quadratic program with either ℓ_1 or ℓ_2 regularization. Since our application involves more than two classes, a "one-versus-the-rest" classifier is trained for each random graph model. Then we select the model corresponding to the highest confidence score during classification.
2. *Random Forest.* In this algorithm, classifiers combine many weaker decision trees, each working on a random subset of the feature space, to reduce variance and increase robustness. The output is simply a sum of the scores given by each tree.
3. *Boosted Decision Trees.* This algorithm gives another approach to combine several weak learners. We use a popular boosting algorithm called AdaBoost [11] in which new trees are constructed sequentially to correct mistakes made by the previous trees. As before, the final prediction is decided by summing across trees.

2.3 Data

Novels: Our method for extracting character networks from novels begins with the tokenization of an input of text. Character names and aliases are then gathered by the parser, coupled with manual addition and subtraction as needed. Names and aliases representing one character are assigned to it's main name. The main names represent the nodes in the network. The parser runs through the text recording the occurrence of two names within a certain number of words apart. For our results, we set the distance parameter to 15 words apart. In the instance where two names share the same keystone and are both within the specified distance along with another name, the parser will record one occurrence between the unique key names. The number of occurrences between two key names represents the weighted edge between the corresponding nodes in the character graph. The node and adjacency lists are recorded via two separate CSV files, which are imported to Gephi, an open source software platform for network analysis and visualization.

The following books were selected for the experiment: *Twilight* by Stephanie Meyer, *Harry Potter and the Goblet of Fire* by J.K. Rowlings, and *The Stand* by Stephen King. We summarize basic network statistics for the novels in Table 1. The results support the view of character networks as complex networks that are dense and small world.

Table 1. Global metrics of character networks from the novels.

Novel	# Nodes	Avg. Degree	Avg. Weighted Degree	Diameter	Edge Density	Avg. Distance	Clust. Coeff.
The Stand	39	14.36	335.33	3	0.378	1.66	0.718
Goblet	62	18.55	305.29	2	0.304	1.69	0.746
Twilight	27	9.11	76.37	4	0.35	1.74	0.783

Moviegalaxies: The website `http://moviegalaxies.com/` has assembled a large number of character networks based on movie scripts. There are over 800 networks available. Each network is weighted, although we discard the weights as we only use this for the model selection problem. Some of the properties of these networks are shown in Figure 1.

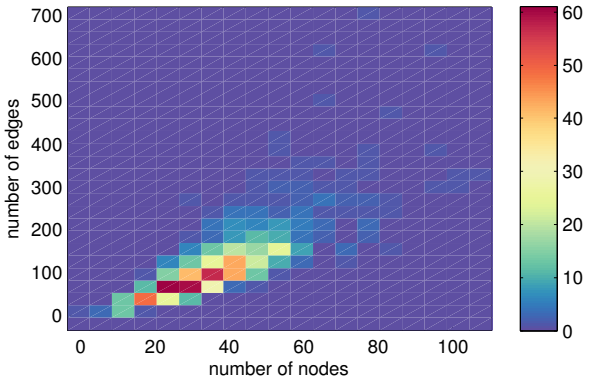

Fig. 1. Order versus number of edges in the *Moviegalaxies* network data.

3 Results

3.1 Analysis of novel character networks

Main characters from each of the novels analyzed scored consistently high in each of the six centrality measures. We present the centrality measures for the top twelve characters from the novel character networks in the figures below. Characters are ranked by increasing PageRank. For example, Harry, Ron and Hermione are identified as the top characters in *Harry Potter and the Goblet of Fire*. Further, our methods accurately predict the community structure for each of the three novels. Visualizations of the character networks and their community in the novels is found below.

For *Harry Potter and the Goblet of Fire* the communities were: Hogwarts, the Dursleys, the Weasleys, Sytherin, and the inseparable friends Seamus and Dean. See Figure 2.

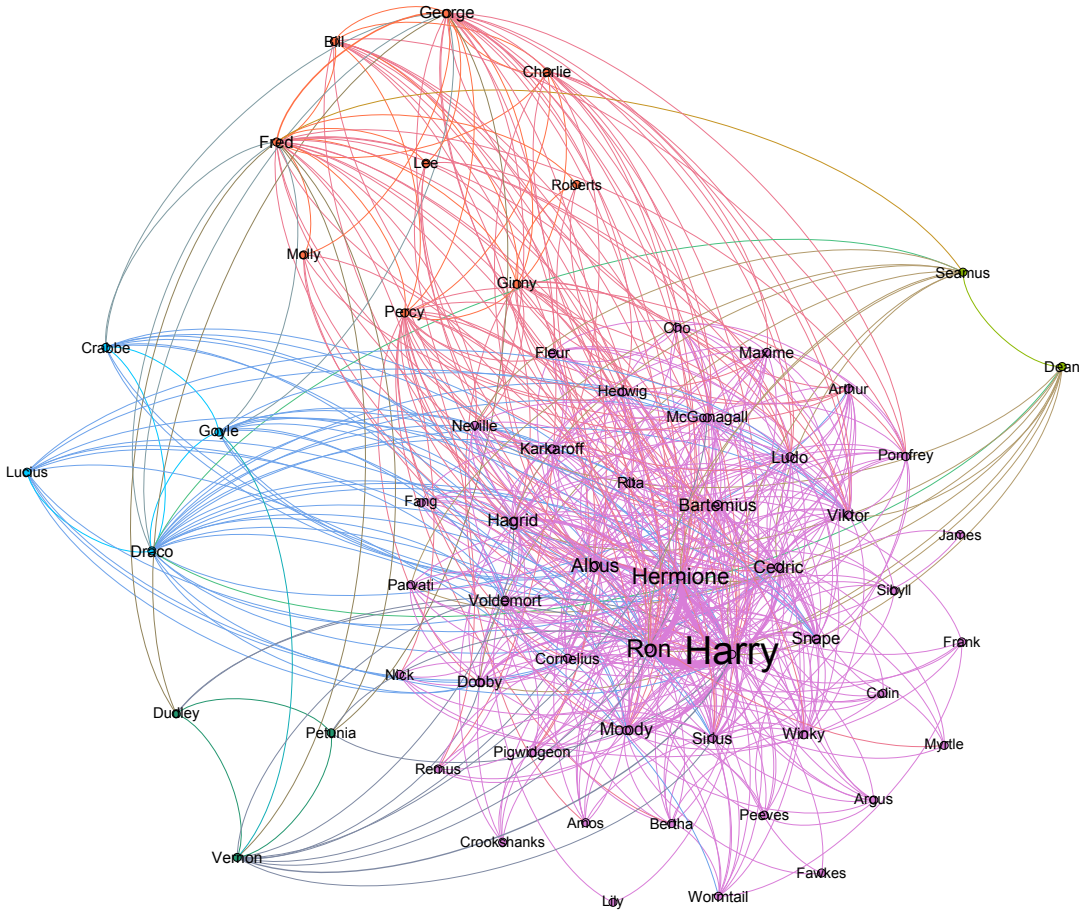

Fig. 2. The character network for *Harry Potter and the Goblet of Fire*. Each community is represented by a distinct color. The thickness of an edge is scaled to its weight, and the size of a name is scaled to the Pagerank score.

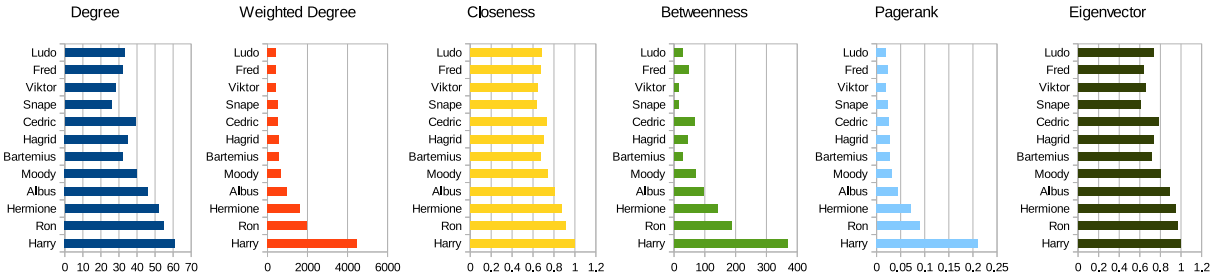

Fig. 3. Centrality measures for *Harry Potter and the Goblet of Fire*.

For *Twilight*, the three communities can be labeled as: vampires, high school students, and characters close to Charlie. See Figure 4.

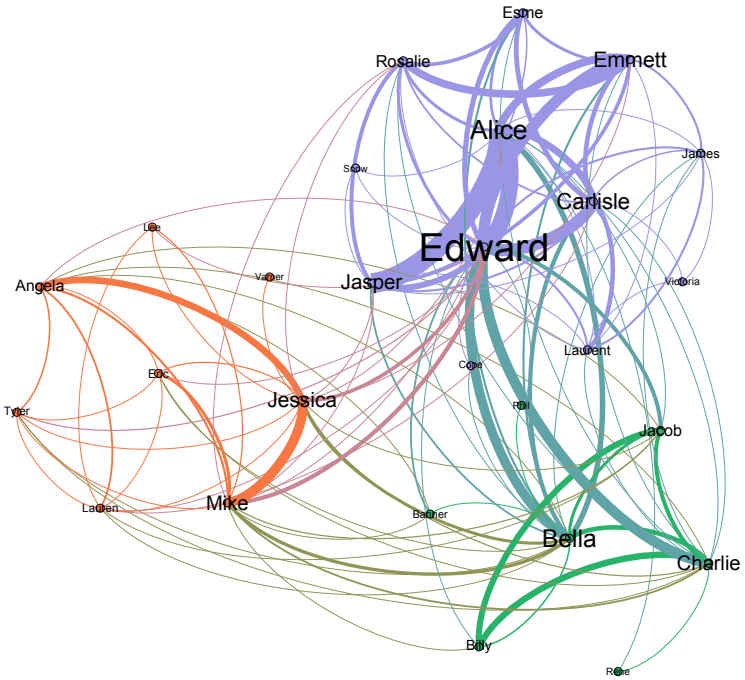

Fig. 4. The character network for *Twilight*.

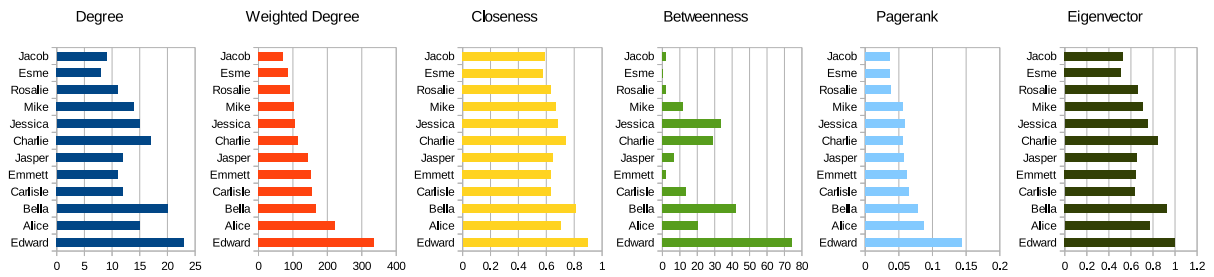

Fig. 5. Centrality measures for *Twilight*.

For *The Stand*, the government and the evil Las Vegas group emerged as separate communities. The *free zone society* was divided into three groups based on their relation to the main characters, Stu, Larry, and Nick. See Figure 6.

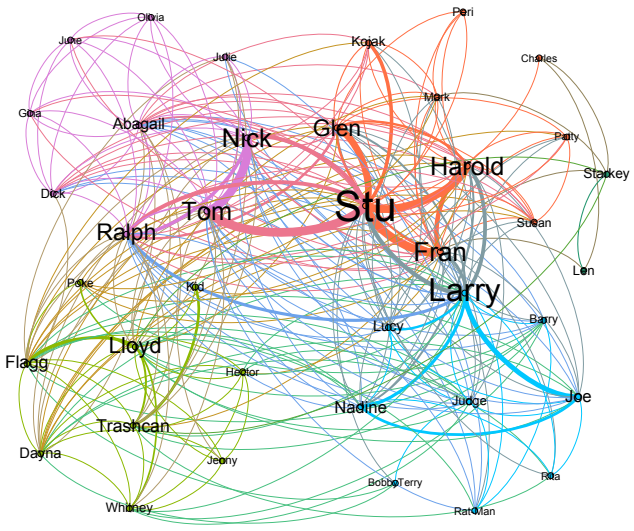

Fig. 6. The character network for *The Stand*.

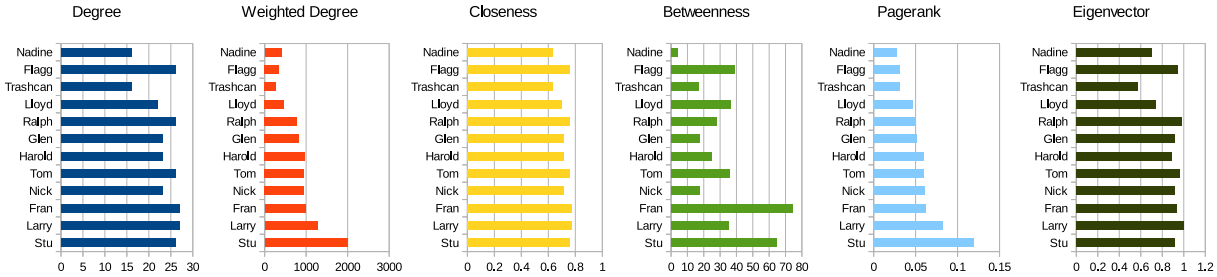

Fig. 7. Centrality measures for *The Stand*.

3.2 Model selection results

Hyperparameters for each classifier were selected using stratified, 5-fold cross validation. All features were normalized to have zero mean and unit variance before training. First, the random graph data was split into half training and half holdout. Classification performance on the holdout set verified both the choice of hyperparameters and the separability of classes in our chosen feature space. All classifiers achieved nearly perfect classification on the holdout set, with over 0.98 precision and recall in nearly all cases (often exactly 1). The F1 score was at least 0.97. Thus, our four random graph models represent distinct classes.

Table 2 shows model selection scores for the setup described in Section 2.2 (we train on the entire random graph data, and test on the original character network). These scores were calculated differently depending on the classifier. For the SVM algorithms, distance

to the separating hyperplane was used. For AdaBoost, we use the final decision function, and soft decision probabilities were used for the random forests; see Figure 8. For all classifiers, a more positive (negative) score indicates more confidence the original graph does (not) belong to the model. Clearly, CL is the best random graph model for all three novels, with each remaining model taking a distant second place in at least one case.

Figure 9 shows our naming convention for the motifs used in our graph profile features. The most important features for the CL SVM hyperplanes were predominantly cliques: induced subgraphs H_3, F_5, F_9, and F_{10}. For the tree-based classifiers, the most important motifs for distinguishing among graph models include some disconnected subgraphs: H_0, H_2, F_2, F_5, and F_{10}. The eigenvalue histograms generally had low importance for all machine learning classifiers. Thus, similar results were obtained using only graph profile features. See Table 3. Figure 10 shows similar aggregate results for the 800 character networks in the *Moviegalaxies* data set, with CL as the best random graph model for the overwhelming number of character networks.

Table 2. Model selection scores for random graph models using graph profiles and eigenvalue histograms as features. CL is selected by all machine learning classifiers as the best model.

Novel	Classifier	PA	CL	ER	CFG
Goblet	SVM-ℓ_2	2.78	**4.59**	−1.10	−10.65
	SVM-ℓ_1	−0.66	**3.81**	−1.55	−10.80
	Forest	0.00	**0.91**	0.094	0.0011
	AdaBoost	−47.2	**47.4**	25.5	−25.7
Twilight	SVM-ℓ_2	−0.671	**4.49**	−2.98	−9.39
	SVM-ℓ_1	−3.08	**5.19**	−2.06	−12.21
	Forest	0.00083	**0.800**	0.0248	0.175
	AdaBoost	−43.06	**32.30**	10.74	0.0205
The Stand	SVM-ℓ_2	−1.52	**2.65**	−1.24	−3.87
	SVM-ℓ_1	−2.32	**2.87**	−1.14	−4.97
	Forest	0.00	**0.946**	0.00	0.0544
	AdaBoost	−47.04	**50.03**	37.83	−40.82

Table 3. Model selection scores for random graph models using graph profiles alone as features. Once again, CL is selected by all machine learning classifiers as the best model.

Novel	Classifier	PA	CL	ER	CFG
Goblet	SVM-ℓ_2	3.18	**4.44**	−1.15	−10.64
	SVM-ℓ_1	−0.68	**3.81**	−1.53	−10.81
	Forest	0.000	**0.998**	0.002	0.000
	AdaBoost	−47.2	**47.4**	25.5	−25.7
Twilight	SVM-ℓ_2	−0.54	**5.51**	−2.73	−9.52
	SVM-ℓ_1	−2.78	**5.25**	−2.02	−12.24
	Forest	0.00	**1.00**	0.00	0.00
	AdaBoost	−39.72	**34.51**	−7.44	12.66
The Stand	SVM-ℓ_2	−1.18	**2.58**	−1.33	−4.02
	SVM-ℓ_1	−2.35	**2.86**	−1.14	−4.99
	Forest	0.00	**0.94**	0.00	0.06
	AdaBoost	−46.49	**50.32**	38.36	−42.19

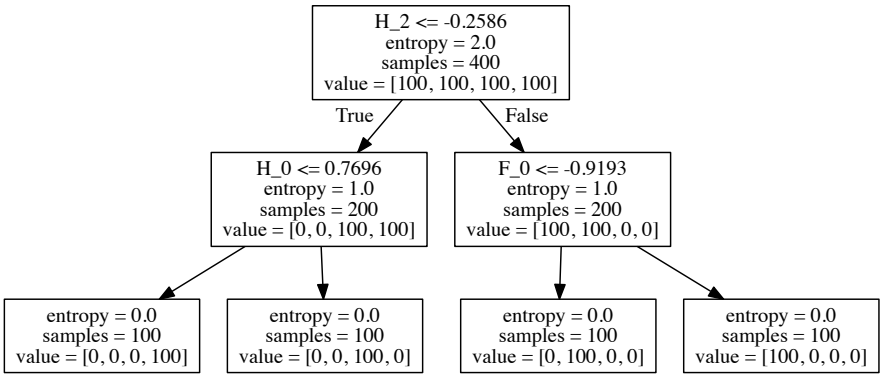

Fig. 8. Example decision tree for the *Goblet* graph.

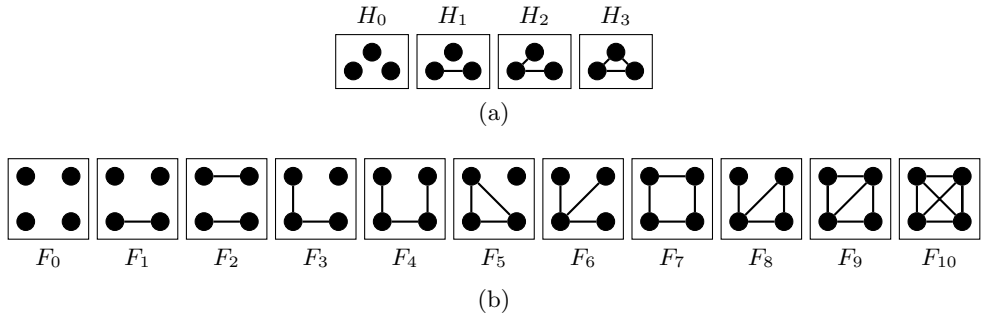

Fig. 9. (a) The four non-isomorphic graphs on 3 nodes that comprise the graph 3-profile. (b) The eleven non-isomorphic graphs on 4 nodes that comprise the graph 4-profile.

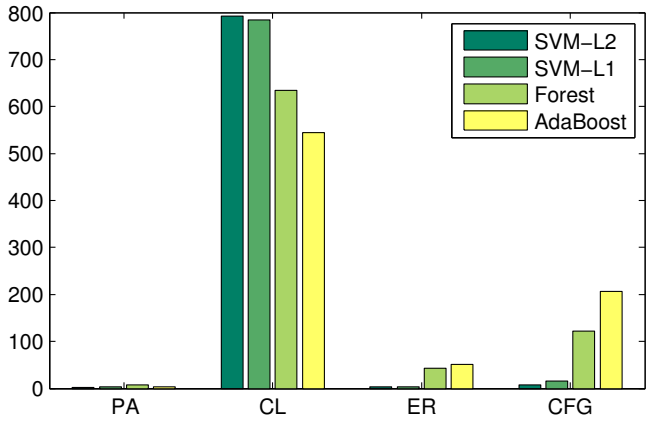

Fig. 10. Summary of *Moviegalaxies* model selection using graph profiles and eigenvalue histograms as features.

4 Discussion and future work

We presented a comparative and quantitative analysis of character networks arising from various novels and films. In particular, we analyzed the weighted social networks from the novels *Twilight The Stand*, and *Harry Potter and the Goblet of Fire*, along with social networks from 800 films catalogued by `http://moviegalaxies.com/`. For each of the character networks from the three novels, we extracted the social network from co-occurrence of character names. Community structure was extracted for each network,

and statistics such as PageRank and various centrality measures were computed for the characters. In each case, our methodology extracts accurate literary conclusions from the data sets, and successfully identifies the influential characters and the constellations of lesser characters in the books. As pointed out first in [4], the analysis provided of these texts was done algorithmically, without resort to conventional literary analysis.

For both the novel and `http://moviegalaxies.com/` data sets, we employed machine learning techniques to compare and contrast the models against simulated data from popular complex network models. The models considered were the Chung-Lu (CL) model, the configuration model, the PA model, and binomial random graphs. Our methodology used small subgraph counts or motifs as classifiers for the Support Vector Machine (SVM) and other machine learning algorithms. For all the data sets, SVM and the other algorithms clearly separated the models, and indicated that the CL model provided the best alignment with the data.

There are various explanations for the conclusions derived from the model selection experiments. As the character networks we consider have relatively few nodes, they are less likely to exhibit various properties such as power law degree distributions or dimensionality found in various on-line social networks such as Facebook. Hence, preferential attachment (an early and successful adopted model for complex networks) or geometric models may be less relevant for character networks. The CL model has a number of properties amenable to modeling character networks. From a literary perspective, an author may intuit a hierarchy of character influence (separated by the degrees of the nodes representing characters), then randomly generate the social ties in the fictional work to complete the network. For instance, Rowlings may have decided in the Harry Potter series that the main triad was Harry, Hermione and Ron, and then gradually added lesser characters revolving around this triad. In terms of the various models, the CL model has 4-node subgraph counts that more accurately model character networks. This is likely due to the property of CL graphs that they have a more diverse set of dense subgraph structures that are more closely related to those that appear in character networks. We plan to continue investigating this finding that CL graphs are good matches for character networks.

In future work, we plan on expanding our analysis of literary works using Project Gutenberg and other sources. We will also explore other models such as random geometric graphs and Kronecker graphs. More broadly, our approach and those of other recent works [2,4,16,17], represents a trend towards the algorithmic and big data-theoretic analysis of cultural works. Such a direction may lead to new models for the evolution and construction of character networks, and a broader view of such networks as complex and evolving.

References

1. R. Alberich, J. Miro-Julia, F. Rossello. Marvel Universe looks almost like a real social network, *arXiv preprint arXiv:0202174* 2002.
2. A. Agarwal, A. Corvalan, J. Jensen, O. Rambow, Social network analysis of "Alice in Wonderland", In: *Proceedings of the Workshop on Computational Linguistics for Literature*, 2012.
3. M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, In: *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2009.
4. A. Beveridge, J. Shan, Network of Thrones, *Math Horizons Magazine* **23** (2016) 18–22.
5. A. Bonato, *A Course on the Web Graph*, American Mathematical Society Graduate Studies Series in Mathematics, Providence, Rhode Island, 2008.
6. A. Bonato, D.F. Gleich, M. Kim, D. Mitsche, P. Pralat, A. Tian, S.J. Young. Dimensionality matching of social networks using motifs and eigenvalues, *PLOS ONE*, 9(9):e106052, 2014.
7. A. Bonato, A. Tian, Complex networks and social networks, invited book chapter in: *Social Networks*, editor E. Kranakis, Springer, Mathematics in Industry series, 2011.

8. E.R. Elenberg, K. Shanmugam, M. Borokhovich, A.G. Dimakis, Beyond triangles: a distributed framework for estimating 3-profiles of large graphs , In: *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.
9. E.R. Elenberg, K. Shanmugam, M. Borokhovich, A.G. Dimakis, Distributed estimation of graph 4-profiles, In: *Proc. International World Wide Web Conference*, 2016.
10. D. Elson, N. Dames, K. McKeown, Extracting social networks from literary fiction, In: *Proceedings of the 48th annual meeting of the association for computational linguistics* 2010.
11. Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer & Systems Science* **55** (1995) 119-139.
12. P.M. Gleiser, How to become a superhero, *Journal of Statistical Mechanics: Theory and Experiment*, 2007, P09020, 2007.
13. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag. 2009.
14. J. Janssen, M. Hurshman, N. Kalyaniwalla, Model selection for social metworks using graphlets, *Internet Mathematics* **8** (2012) 338-363.
15. N. Pržulj, Biological network Ccomparison using graphlet degree distribution, *Bioinformatics* **23** (2007) 77–183.
16. A.J. Reagan, L. Mitchell, D. Kiley, C.M. Danforth, P.S. Dodds, The emotional arcs of stories are dominated by six basic shapes, *arXiv preprint arXiv:1606.07772* 2016.
17. M.A. Ribeiro, R.A. Vosgerau, M.L.P. Andruchiw, S. Ely de Souza Pinto, The complex social network from the Lord of the Rings, *Rev. Bras. Ensino Fs.* (2016) **38** 1304.
18. G. Sack, Character networks for narrative generation, In: *Intelligent Narrative Technologies: Papers from the 2012 AIIDE Workshop, AAAI Technical Report WS-12-14*, 2012.
19. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.
20. N. Shervashidze, S.V.N. Vishwanathan, T.HṖetri, K. Mehlhorn, K.M. Borgwardt, Efficient graphlet kernels for large graph comparison, In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2009.
21. D.B. West, *Introduction to Graph Theory, 2nd edition*, Prentice Hall, 2001.