Knowledge-Based Systems 163 (2019) 800–810

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Unsupervised cluster analyses of character networks in fiction: Community structure and centrality

R.H.-G. Chen [a,b], C.-C. Chen [c,d], C.-M. Chen [a,*]

[a] Department of Physics, National Taiwan Normal University, Taipei, Taiwan
[b] Department of Computer Science, University of British Columbia, BC, Canada
[c] Minnan Cultural Research Institute, Minnan Normal University, Zhangzhou, PR China
[d] Research Institute of the Central Soviet District, Longyan University, Longyan, PR China

ARTICLE INFO

Article history:
Received 9 January 2018
Received in revised form 5 September 2018
Accepted 3 October 2018
Available online 10 October 2018

Keywords:
Unsupervised clustering
Computer-aided visualization
Social network
Community structure detection
Centrality measures

ABSTRACT

We present an integrated approach to cluster and visualize character networks in fiction with the aid of computational and statistical methods. An unsupervised clustering algorithm, minimum span clustering (MSC), was applied to cluster fictional characters at various characteristic resolutions based on their activities in the novel. As a demonstration, we study the character network in *Dream of the Red Chamber*, the greatest novel in Chinese literature. The character network of the novel is found to exhibit properties of scale-free and small-world networks. Based on unsupervised cluster analyses, we construct and visualize the community structure of the network, and find a three-tiered structure of core, secondary, and peripheral characters. By treating the network as a weighted graph, we further analyze the centralities of characters to determine their importance in the network, and find that betweenness centrality, as a measure of characters' control over the flow of the narrative, is differentiated from other centrality measures for *Dream of the Red Chamber*. We believe that these analytic methods provide beneficial tools for applications such as autonomous novel writing.

1. Introduction

Computer-assisted techniques in the humanities have developed rapidly with the increasing accessibility of computers and the internet since the 1980s. Accordingly, a subfield in computational linguistics has emerged to address questions in literature using computational approaches [1]. In particular, advancements in the efficiency of databases and automated data retrieval have incited interest in the computational analysis of literary data [2]. A recent publication on "culturomics" analyzed a corpus of digitized texts containing approximately 4% of all books ever printed, enabling a quantitative investigation of cultural trends [3]. In spite of major hurdles, computing technologies in the humanities have shown promising successes.

In the academic literature, a disparity between scientific and humanistic research is generally perceived in terms of their methods and assumptions. The rise of the digital humanities offers an opportunity to reconcile these estranged domains by applying scientific methods to solve problems in the humanities. Computational literary analysis and modeling, for instance, may facilitate the examination of hypotheses by providing concrete evidence through a more efficient and extensive means of reading [4]. High-level "distant readings" of texts facilitated by natural language processing, machine learning, and statistical modeling can offer alternative perspectives complementary to those acquired from "close readings" [5].

In the following sections, we review related work on the extraction of character networks in fiction as well as the cluster analysis of complex networks. We then describe our methods for extracting characters and their relationships from a novel, for clustering the characters based on their relationships, and for visualizing the character network. As a case study, we demonstrate these methods using the novel *Dream of the Red Chamber*. After presenting our analytical results, we discuss their significance from both statistical and literary perspectives.

2. Background

2.1. Related work on the analysis of character networks in fiction

Computational and statistical methods have become increasingly useful for literary studies, with applications including author identification and plagiarism detection. However, due to difficulties in generating automatic interpretations of textual semantics, computational analyses of literary works rarely focus on the theme, the plot, and the inter-relationships of characters. Nevertheless, in

* Corresponding author.
E-mail address: cchen@phy.ntnu.edu.tw (C.-M. Chen).

https://doi.org/10.1016/j.knosys.2018.10.005
0950-7051/

recent years, researchers have started to extract and investigate the character networks within novels using natural language processing [6–11]. Elson et al. [6] derived the character networks for 60 nineteenth-century British novels from dialogue-based interactions, and used structural properties of the networks to verify the validity of existing literary hypotheses. Lee and Yeung [7] created a network linking characters and scenes by investigating the narrative structure in the Old Testament. Agarwal et al. [8] extracted the character network from Carroll's *Alice in Wonderland* by detecting social events. Ardanuy and Sporleder [10] extracted character networks from 46 novels to quantitatively classify these novels in terms of genre and authorship. More recently, Skorinkin [11] demonstrated a possible application of character network extraction in the exploration of literary plot dynamics.

Various measures have been proposed to define the scope of an inter-character interaction. These include co-occurrences [12] or dialogues between characters in the narrative [6,13]. Using these interactions, a character network can be constructed to provide readers with a high-level overview of the novel's organization and plot. The community structure and the centrality indices of the character network can also be derived to further characterize its properties [12,14]. Additionally, the network's time evolution can be visualized by mapping each chapter of the novel to a network with multi-dimensional scaling [15,16]. Temporal analysis of character networks has been performed by finding eigenvector-based centrality measures [17].

2.2. Related work on the cluster analysis of complex networks

The character networks discussed in Section 2.1 are an example of complex networks. An important subject in complex network analysis that has received considerable attention is the detection of community structure [12]. A deeper understanding of a given complex system can be obtained by identifying characteristic structural patterns within a network; this has been applied to real-world social communities [18] as well as to networks of related topics on the World Wide Web [19]. The identification of network communities – i.e. cluster analysis – requires an algorithm to efficiently partition a network into clusters of densely-connected nodes that are sparsely connected to nodes in other clusters. Cluster analyses perform unsupervised machine learning on unlabeled network data; the ability of these analyses to model unlabeled data holds enormous potential to expand the applicability of unsupervised clustering.

Several clustering methods have been developed to handle large, heterogeneous networks. These include hierarchical clustering (HC) [20,21], affinity propagation (AP) [22,23], minimum spanning tree [15,24], *k*-means clustering [20,25], and self-organizing maps [26]. The quality of clustering results is often evaluated by the modularity or the silhouette coefficient of the partition. Modularity measures the density of intra-cluster links relative to inter-cluster links [27]; the optimization of modularity is also used by greedy algorithms [27], spectral methods [28], and extremal optimization [29] to find an optimal network partition. However, a resolution limit exists for modularity optimization, below which it may even fail to identify complete graphs [30]. Meanwhile, the silhouette coefficient is a measure of how proximal a node is to its assigned cluster compared to other clusters [31]; for this study, we used the silhouette coefficient to assess the quality of our clustering results. For comparative purposes, in the Supporting Information, we provide definitions of the silhouette coefficient as well as Jaccard's similarity measure, and we utilize both measures to evaluate clustering results (Table S1 and Fig. S2) of the character network derived by several clustering algorithms.

2.3. Contributions

In this paper, we propose an integrated approach to investigating the social network of literary characters based on their activity patterns in the text. The main innovation of our methodology is the application of the minimum span clustering (MSC) algorithm for the identification of the character network's community structure. Using only the distance matrix of the character network as input, MSC is able to recursively determine the number of clusters and the community structure at various characteristic resolutions. We then applied computational visualization and statistical analysis to the identified network structures and used measures of character centrality to determine the relative significance of characters in the network. Such an integrated approach may offer an automated methodology for obtaining a high-level perception of the plot and themes of a literary work. A panoramic understanding of character relationships obtained through network clustering provides a useful foundation for developing systems capable of automatic narrative comprehension and generation.

3. Materials and methods

3.1. Dataset preparation

To demonstrate our approach to the analysis of character networks in literary fiction, we selected a Chinese long-form novel, *Dream of the Red Chamber*, as our subject. Composed by the 18th-century Chinese writer Cao Xueqin, *Dream of the Red Chamber* is considered to be the greatest novel in Chinese literature and contains a sophisticated social network. It has given rise to the field of Redology, which is devoted exclusively to this work. The novel is set during the early Qing dynasty in the Rongguo and Ningguo Houses, which are two large, adjacent family compounds that serve as residences for two branches of the wealthy, noble Chia clan. Supplementary Fig. S1 shows the relationships between the characters in the narrative. The plot centers on the romantic rivalry and affection between three main characters (Chia Pao-yu, Lin Tai-yu, and Hsueh Pao-chai), set against the milieu of the family's diminishing fortunes. Chia Pao-yu, born with a magical piece of jade in his mouth and thus treasured by his mother and grandmother, is strongly attached to his cousin Lin Tai-yu. However, he has been arranged to be married to another cousin, Hsueh Pao-chai, in whom he has no romantic interest. A number of secondary conflicts pervade the social network of the Chia household, including conflicts between masters and servants, between wives and concubines, and between legitimate and illegitimate children.

It is generally believed that *Dream of the Red Chamber* is a semi-autobiographical account of Cao's life, with the Chia family's fall from dignity due to wanton greed paralleling that of the author's own family. As a reflection of the social superstructure of the Qing dynasty, the work details the economics, politics, culture, education, law, ethics, religion, and marriage customs of the time. Many manuscripts of the novel are known; in this study, we use the text of the Cheng-B edition, which contains 120 chapters and is the most widely-circulated edition.

3.2. Node and character extraction

Identifying literary characters is a fundamental task in the computational literary analysis. Many analysts assume that automated detection of characters in texts, alongside the synonyms of their names, can be reliably and accurately performed using standard algorithms such as Named Entity Recognition (NER). However, state-of-the-art methods for automatically detecting and distinguishing occurrences and mentions of characters are still unsatisfactory [6,9,32–34], and substantial improvement is still necessary for NER.

Table 1
List of the 100 most frequently occurring characters in *Dream of the Red Chamber*.

1. Chia Pao-yu	2. Wang Hsi-feng	3. Lady Dowager	4. Lin Tai-yu
5. Hsi-jen	6. Hsueh Pao-chai	7. Lady Wang	8. Chia Cheng
9. Chia Lien	10. Ping-erh	11. Aunt Hsueh	12. Tzu-chuan
13. Chia Tan-chun	14. Yuan-yang	15. Shih Hsiang-yun	16. Chia Chen
17. Li Wan	18. Madam Yu	19. Ching-wen	20. Granny Liu
21. Lady Hsing	22. Hsueh Pan	23. Hsiang-ling	24. Sheh-yueh
25. Chia Jung	26. Chia Sheh	27. Chia Yun	28. Chia Hsi-chun
29. Mrs. Chou	30. Chia Yu-tsun	31. Fang-kuan	32. Miao-yu
33. Chia Huan	34. Hsueh-yen	35. Chia Ying-chun	36. Ming-yen
37. Concubine Chao	38. Ying-erh	39. Pao-chan	40. Chin Chung
41. Chia Chiao	42. Hsueh Ko	43. Chiu-wen	44. Yu Erh-chieh
45. Chia Lan	46. Wu-erh	47. Ssu-chi	48. Mrs. Lin
49. Lai Ta	50. Chen Shih-yin	51. Chin Ko-ching	52. Chia Jui
53. Chia Chiang	54. Lin Chih-hsiao	55. Feng Tzu-ying	56. Tsai-yun
57. Hu-po	58. Pao Yung	59. Chin-chuan	60. Yu-chuan
61. Feng-erh	62. Chou Jui	63. Tsui-lu	64. Mrs. Liu
65. Men-tzu	66. Li Kuei	67. Pao Erh	68. Yu San-chieh
69. Hsiao-hung	70. Chang Hua	71. Ni Erh	72. Pan-erh
73. Chui-erh	74. Nanny Li	75. Chen Pao-yu	76. Hsing-erh
77. Wang Jen	78. Wang Chi-jen	79. Ou-kuan	80. Chiu-tung
81. Chin Jung	82. Chia Chin	83. Chun-yen	84. Wang Shan-pao
85. Mrs. Wang	86. Chiang Yu-han	87. Ssu-erh	88. Shih-shu
89. Li Shih-erh	90. Chih-neng	91. Tsai-ping	92. Li Wen
93. Chen-chu	94. Wang Tzu-teng	95. Tsui-mo	96. Abbot Chang
97. Tsai-hsia	98. Abbess Ma	99. Jui-kuan	100. Ling-kuan

Compared with English literature, a much smaller volume of analyses have focused on applying NER to Chinese literature [35]. The lack of properly annotated data and other relevant resources has become a major challenge in the application of NER to Chinese literature [36].

In this study, we extracted a list of character names from *Dream of the Red Chamber* using natural language processing. The implementation of name extraction primarily utilized the Stanford NER algorithm, using linear chain Conditional Random Field (CRF) models [37,38]. Manual inspection was applied following automated extraction, during which false positives were expurgated, synonyms of character names were merged with and replaced by the primary names, and occurrences of minor characters overlooked by the algorithm were re-inserted. Fig. 1 shows the occurrence frequencies for the characters in *Dream of the Red Chamber*. The line of best fit for the frequencies follows a power law relationship $f \propto A^{-\gamma}$, where f denotes the number of characters with A appearances in the novel. To condense the visualizations, we considered a partial social network containing only a subset of the 100 most frequently occurring characters. In Table 1, characters are listed by their ranking in terms of the frequency of their appearances in the novel; we Romanized the Chinese names based on the Wade–Giles system. Characters who lack surnames are servants.

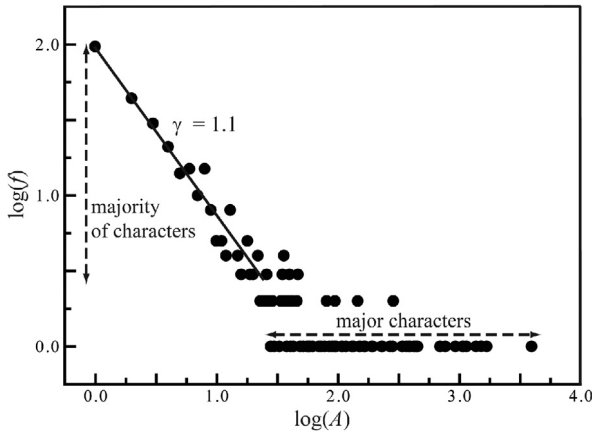


Fig. 1. Frequency (f) in the number of character appearances (A) in *Dream of the Red Chamber*. Here A denotes the counting of appearances of each character in the novel, and f denotes the number of characters with A appearances. The solid line is a line of best fit, given by the equation $f \propto A^{-\gamma}$.

3.3. Edge and relationship extraction

The first element of our analysis involves inter-character relationships, which we investigated based on the quantities and relative locations of their appearances in the novel. Since *Dream of the Red Chamber* contains 120 chapters, we considered each chapter as an appropriately-sized subdivision of the text. We calculated a vector of appearances in the novel for each of the 100 most frequently occurring characters; based on this vector, we calculated the pattern similarities for pairs of characters using the cosine measure. Specifically, if a_k^i denotes the number of appearances of character i in chapter k, we defined the pattern similarity between two characters (i, j) as

$$w_{i,j} = \frac{\sum_{k=1}^{120} a_k^i a_k^j}{\sqrt{\sum_{k=1}^{120} \left(a_k^i\right)^2} \sqrt{\sum_{k=1}^{120} \left(a_k^j\right)^2}}, \tag{1}$$

where the value of $w_{i,j}$ (the weight of the edge between i and j) is in the range [0, 1]. Supplementary Fig. S3 displays the normalized appearance patterns of the central trio of protagonists in the novel. For mapping and visualization, we converted similarity into distance by expressing the distance as

$$d_{i,j} = \sqrt{\frac{1}{\max(t,\, w_{i,j})} - 1}, \tag{2}$$

where t sets the upper bound of $d_{i,j}$. Using this definition of distance, closely-related characters are separated by a short distance, while remotely-related characters are separated by a long distance. When $t = 0$, the distance between two unrelated characters is infinite, which is undesirable for visualization. We used $t = 0.001$ to minimize distortions in our visualizations of the character network. For chaptered novels, the frequency of co-occurrences in each chapter is generally a valid measure of the social distance between a given pair of characters, since we derive their relationship from the similarity of their activity patterns. Although time confusion (i.e., inconsistency in the speed of the passage of time) has

been recognized in the narrative of *Dream of the Red Chamber*, our proposed similarity measure still effectively quantifies character relationships and yields consistent predictions for the community structure of the character network.

3.4. MSC algorithm

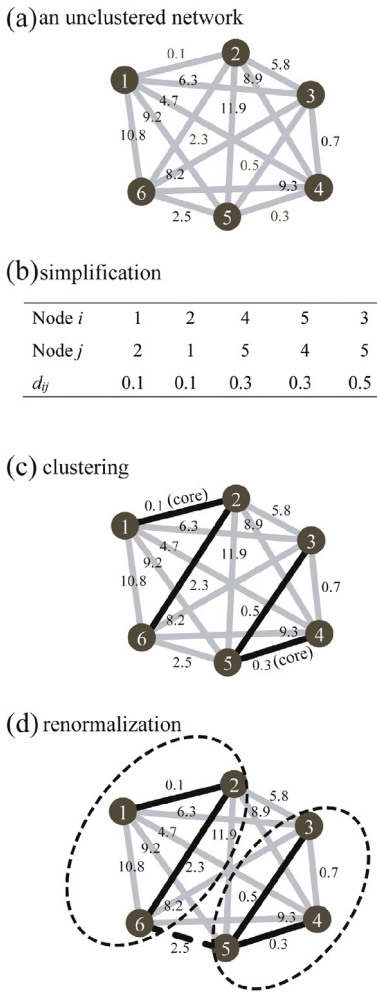
Many clustering methods suffer from two major drawbacks: the necessity of parameterization, and the high computational cost for large datasets. To overcome these obstacles, we have developed the MSC algorithm for the unsupervised clustering of complex networks; it provides a hierarchical approach for clustering the structure of a complex network at various levels of resolution. In this work, we used a lightweight version of MSC to cluster the character network of *Dream of the Red Chamber* based on the distance measure defined in Eq. (2). Here, we describe the three steps of the algorithm:

Step 1 (simplification). The algorithm identifies the closest neighbor of each node and records their distances in ascending order. For a network of N nodes, instead of operating on a distance matrix of N^2 elements, MSC only processes a list of shortest distances that is N elements long after the initial simplification step.

Step 2 (clustering). MSC constructs the first cluster by starting from the shortest node pair and then adding additional pairs from the list in order of increasing distance. For newly-added pairs, if one of the two nodes is involved in an existing cluster, the size of this cluster increases, but the number of clusters remains unchanged. If neither node is involved in the existing clusters, a new cluster is identified and the number of clusters increases. All clusters of the network have been found once all of the distances in the list have been considered. In each cluster, the pair of nodes with the shortest distance is considered to be the core of the cluster; for a character network, the core denotes the most closely-related characters in a cluster. Upon the completion of the first iteration, the identified clusters collectively constitute the first-level clustering, which has the highest resolution.

Step 3 (renormalization). Clusters constructed in step 2 are considered as renormalized nodes. The distance between the clusters ($\{\tilde{d}_{i,j}\}$, where i, j represent the constructed clusters) is calculated by finding the inter-cluster node pairs with the shortest distances. By iterating repeatedly through these steps, sub-groupings within the network consisting of these renormalized components are identified at increasingly coarse resolutions. We note that $\{\tilde{d}_{i,j}\}$ could be calculated using a different definition; we chose the present definition such that the minimum spanning tree diagrams for the network visualization are invariant with changes in scale. In other words, given a visualization of the network at the coarsest resolution, this definition allows us to enlarge any region of the network to finer resolutions without compromising the relative positions of linked characters.

An example of implementing MSC for a simple six-node network is shown in Fig. 2. Fig. 2(a) displays an un-clustered network, and the corresponding list of shortest distances is shown in Fig. 2(b). In step 2, the first cluster is created with nodes 1 and 2 (the pair with the shortest distance), while the second cluster begins with nodes 4 and 5. Nodes 6 and 3 are then respectively added to the two clusters. The primary clustering is complete after all of the links in the list have been considered. As shown in Fig. 2(c), the network is then decomposed into two MSC clusters, each being a



Node i	1	2	4	5	3	6
Node j	2	1	5	4	5	2
d_{ij}	0.1	0.1	0.3	0.3	0.5	2.3

Fig. 2. A demonstration of the MSC algorithm for clustering a simple six-node network: (a) the un-clustered network, (b) the list of the shortest distance pairs for network nodes, (c) the clustered network showing the nodes of each cluster and the intra-cluster links (black solid edges), and (d) the renormalized network showing clusters (i.e. renormalized nodes, enclosed by a dashed loop) and their links (black dashed edge).

minimum spanning tree. In step 3, each cluster (shown in a dashed loop) is considered as a renormalized node for the next-level clustering. As shown in Fig. 2(d), the shortest distance between the two renormalized nodes is 2.5. Notice that the minimum spanning tree diagrams are invariant with respect to scale.

Utilizing MSC for clustering complex networks is associated with a number of inherent advantages. MSC can efficiently cluster large networks, and its speed also outperforms that of other clustering algorithms. It takes < 0.1 s to cluster a network of 10,000 nodes in the Matlab software on Intel Core i7 desktop computer, compared to 18 s using HC and >12 h using AP or K-means. Additionally, compared to results of other clustering methods, MSC produces clusters with higher intra-group similarity and lower inter-group similarity; this observation has been validated for the social science network [39], the stock network [40], and various protein similarity networks [41,42]. MSC also does not require hyper-parameterization, which eliminates the sensitivity of other major clustering algorithms to parameterization. In particular, the experimenter does not need *a priori* knowledge of the number or size of the clusters. Furthermore, the implementation of MSC is straightforward, and it enables users to recursively analyze the structure of a network at various characteristic resolutions.

3.5. Visualizing the character network

In this study, MSC found two distinct resolution levels for the community structure of the 100 most frequently occurring characters. To visualize these community structures, the graph layout algorithm ForceAtlas2 was applied in the Gephi software [43] to visualize the network structure at the finer resolution level, while the Kruskal algorithm [44] was used to construct the minimum spanning tree diagram at the coarser resolution level. ForceAtlas2 is a force-directed algorithm that spatializes networks using a simulation of a physical system; in this simulation, nodes are represented as charged particles repelling each other, while edges are represented as springs that attract their associated nodes. Forces move nodes on a two-dimensional plane until the network converges to an equilibrium imposed by the distance matrix of the network. As the simulation could be trapped at a local minimum, we perturbed the resulting configuration by arbitrarily repositioning hub nodes to verify its stability. To reduce the number of local minima, we deleted edges with distances longer than a cut-off value (d_{cut}); to ensure that each node has at least one link, we chose d_{cut} to be 1.86 (corresponding to a weight of $w_{\text{cut}} = 0.22$). We used the following parameters for the ForceAtlas2 algorithm to produce network structures that are consistent with the MSC results: Tolerance = 1.0, Approximate repulsion = False, Approximation = 1.2, Scaling = 10.0, Stronger gravity = False, Gravity = 1.1, Dissuade hubs = False, LinLog = True, Prevent overlap = True, and Edge weight influence = 1.2.

At the coarser resolution level, we constructed the minimum spanning tree diagram using an ascending list of shortest distances $\{d_{i,j}\}$ between clusters obtained from the first iteration of MSC (i.e. renormalized nodes). We iterated through the edges, adding links between clusters to the seeding graph if no loops were present; this process generated a complete minimum spanning tree diagram of the social network. To integrate geometric information of the social networks obtained using the ForceAtlas2 algorithm into the minimum spanning tree diagrams, we computationally repositioned all clusters on a two-dimensional plane. In doing so, we attempted to preserve the $\{\tilde{d}_{i,j}\}$ of connected nodes and the relative orientations of the clusters (which we defined as the relative orientations of their cores), by minimizing the cost function

$$E = \sum_{i,j\in\Omega} \left\{ c_1 \frac{\left[\tilde{d}_{ij} - d\left(x^i, x^j\right)\right]^2}{\tilde{d}_{ij}^2} + c_2 \left[1 - \hat{d}\left(x^i, x^j\right) \cdot \hat{o}\left(x^i, x^j\right)\right]^2 \right\}, \tag{3}$$

where the summation encompasses all connected pairs in the minimum spanning tree (i.e. Ω), $c_1 = 1$ and $c_2 = 0.1$ are weighting coefficients, $d(a,b)$ is the edge distance between nodes a and b, $\hat{d}(a,b)$ is the unit vector of $d(a,b)$, and $\hat{o}(a,b)$ is the unit vector of their relative orientation on a two-dimensional plane, beginning at the core of cluster a and ending at the core of cluster b. This cost function sums the discrepancies between the distance matrix and the visualization of the network in the length (the first term) and the orientation (the second term) of connected edges, and its minimization yields a more faithful two-dimensional representation of the character network. When minimizing the above cost function, we prohibited both node overlaps and edge crossings. The final minimum spanning tree diagram was subject to manual inspection and minor modification. In the diagram, the distance between two connected nodes is proportional to the length of the edge connecting them; however, the distance between two unconnected nodes may be subject to substantial distortion.

4. Results and discussion

4.1. Properties of the character network

We found that the character network in *Dream of the Red Chamber* exhibits properties of scale-free and small-world networks. As shown in Fig. 1, the frequency (f) of characters with a given number of appearances (A) follows a power-law decrease as A increases. In the power-law equation $f \propto A^{-\gamma}$, we found the exponent γ to be 1.1, suggesting that the novel lacks a preferred scale when it comes to the frequency of a character's appearances. Major characters in the novel have a large number of appearances, ranging from 30 to 3000, while the remaining characters are largely peripheral and appear less than 10 times in the narrative.

Although the number of appearances can measure the importance of a character to the narrative, it does not convey information about inter-character interactions. To focus on the interactions between important characters, we utilized a simplified character network containing only the 100 most frequently occurring characters. We found that this simplified network satisfies the definition of a small-world network. Small-world networks are mathematical graphs in which the average shortest path length (L) between any pair of nodes is relatively small while the average level of clustering (C) is relatively high. As applied to real-world social networks, the small-world property suggests that a given pair of strangers in the network is linked by a short chain of acquaintances. Typically, network "small-worldness" is quantified by the small-world coefficient $\sigma = (C/C_{\text{r}})\cdot(L/L_{\text{r}})^{-1}$, where L_{r} and C_{r} are respectively the average shortest path length and the average level of clustering as calculated from equivalent random networks with the same degree. The definition and derivation of L, C, L_{r}, and C_{r} are described in the Supporting Information. A network is considered to exhibit small-world properties if $\sigma > 1$ (i.e. $C \gg C_{\text{r}}$ and $L \approx L_{\text{r}}$). For the character network in *Dream of the Red Chamber*, we obtained $\sigma = 5.72$ ($C/C_{\text{r}} = 7.26$ and $L/L_{\text{r}} = 1.27$), suggesting that this network can be classified as a small-world network.

4.2. Community structure of the character network

To further investigate the community structure of *Dream of the Red Chamber*, we conducted unsupervised cluster analyses of the simplified network containing the 100 most frequently occurring characters using the MSC algorithm. MSC detected two characteristic resolution levels, as shown in Table 2. In the first-level clustering, the character network is decomposed into 28 clusters. We used statistical methods to identify two clusters in the first-level clustering related to overarching themes of the narrative; other clusters formed as a consequence of isolated events. In the second-level clustering, the 28 first-level clusters are merged into 5 groups. We consider group I to be the primary group, again related to the development of narrative themes, and groups II–V to be auxiliary groups that represent sequences of events. As an example of the latter, cluster 18 mainly describes an event at the school which connects Ming-yen (36), Li Kuei (66), Chin Jung (81), and Chin Chung (40); this event is followed by the affair of Chin Chung and Chih-neng (90), as depicted by cluster 25. As group V consists of clusters 18 and 25, it is a combination of these temporally proximal events.

The central importance of cluster 1 and group I to the narrative can be seen through a close examination of their members. The main characters of *Dream of the Red Chamber* form cluster 1, the largest cluster within the network; they include the central protagonists, Chia Pao-yu (1), Lin Tai-Yu (4), and Hsueh Pao-chai (6), alongside 11 other characters who are their close family members or servants. As the bulk of the primary narrative in *Dream of the Red Chamber* details the love affair of the central trio, the composition

Table 2
MSC results for the community structure of *Dream of the Red Chamber* at two characteristic resolution levels. Characters are numbered as in Table 1.

Level 1 MSC clusters														
Cluster 1	1	3	4	5	6	7	8	11	21	22	26	59	74	89
Cluster 2	37	65	94	97	98									
Cluster 3	16	18	25	51	96									
Cluster 4	27	62	67	69	71									
Cluster 5	31	46	48	56	64									
Cluster 6	19	24	73	87										
Cluster 7	28	32	75	91										
Cluster 8	33	41	45	77										
Cluster 9	12	34	78	88										
Cluster 10	9	44	70	80										
Cluster 11	52	53	100											
Cluster 12	15	63	92											
Cluster 13	2	10	61											
Cluster 14	23	39	42											
Cluster 15	54	58	93											
Cluster 16	30	50	55											
Cluster 17	49	82	86											
Cluster 18	36	66	81											
Cluster 19	79	83	99											
Cluster 20	20	29	72											
Cluster 21	35	47												
Cluster 22	43	95												
Cluster 23	14	57												
Cluster 24	13	17												
Cluster 25	40	90												
Cluster 26	38	60												
Cluster 27	68	76												
Cluster 28	84	85												

Level 2 MSC clusters							
Group I	Cluster 1	Cluster 2	Cluster 6	Cluster 8	Cluster 9	Cluster 11	Cluster 12
	Cluster 13	Cluster 14	Cluster 17	Cluster 20	Cluster 22	Cluster 23	Cluster 24
Group II	Cluster 5	Cluster 16	Cluster 19	Cluster 21	Cluster 26		
Group III	Cluster 3	Cluster 4	Cluster 10	Cluster 27			
Group IV	Cluster 7	Cluster 15	Cluster 28				
Group V	Cluster 18	Cluster 25					

of cluster 1 is thus consistent with the plot. We note that Wang Hsi-feng (2), who manages the operations of the Chia household, is not included in cluster 1. Instead, she forms cluster 13 along with close maidservants, Ping-erh (10) and Feng-erh (61). In a further-simplified character network with $d_{\text{cut}} = 1.86$, Wang Hsi-feng and Chia Pao-yu share the highest number of connections at 38, indicating that they both interact strongly with many other characters. In the narrative, Wang Hsi-feng is, in fact, central to the downfall of the Chia household. Early critics of *Dream of the Red Chamber* identified two major themes within the narrative, namely romantic love and the ephemeral nature of worldly possessions; Chia Pao-yu and Wang Hsi-feng respectively embody these two themes. Both cluster 1 and cluster 13 belong to the central group I at the coarser resolution level, suggesting that the plot and themes of the narrative were primarily developed based on the interactions between characters in this group. In summation, our cluster analysis of the character network in *Dream of the Red Chamber* predicts a community structure that is consistent with observations which can be made during a close reading of the novel.

It is worthwhile to compare the quality of the MSC algorithm's clustering results with the predictions of other clustering algorithms. In this study, we compared the performance of MSC with that of the affinity propagation and hierarchal clustering methods. Table S1 compares the clusters identified in the character network by MSC, AP, and HC. In general, MSC's clustering results are more similar to those of AP than those of HC. The Jaccard's similarity measure is 0.86 between the clustering results of MSC and AP, 0.66 between the results of MSC and HC, and 0.68 between the results of AP and HC.

To further compare these clustering results, we use the silhouette coefficient to validate the consistency within clusters of data. As shown in Fig. S2, the silhouette coefficients are mostly positive in the clustering results of MSC, AP, and HC, suggesting the validity of these results. However, HC has greater silhouette coefficients than MSC and AP, at the expense of isolating six characters in the network. By comparison, there are three isolated characters in AP's result and none in MSC's result. As these single-node clusters possess a maximal average intra-cluster similarity, their presence increases the proportion of nodes with large, positive silhouette coefficients, and accordingly reduces the proportion of nodes with negative silhouette coefficients; this phenomenon can be observed in Fig. S2, which shows silhouette plots for the three clustering methods. As we aimed to identify the most significant relationship for each character in the network, and also to visualize the interrelationships of characters, the lightweight implementation of the MSC algorithm that we used for the present study does not allow for single-node clusters (note that isolated nodes can be detected by the full algorithm [45]). For instance, while both AP and HC identify Li Shih-erh (89) as a self-enclosed cluster, MSC includes this character in cluster 1 due to his close relationship with Chia Cheng (8) as described in Chapter 99 of the novel. Chapter 99 provides an illustration of the political corruption that was rampant during the Qing dynasty; in this chapter, Li Shih-erh persuades a reluctant Chia Cheng to allow the latter's subordinates to extort money from the public, with Chia Cheng acquiescing to maintain order in the local government. From a narrative perspective, through Chia Cheng's assimilation into the corrupt political climate, the events of this chapter also contribute to the eventual downfall of the Chia household. Considering Li Shih-erh as an isolated character, as one would be led to do based on the clustering results of AP and HC, is evidently a reductionist approach that fails to consider the deeper nuances of his interactions. Therefore, although the calculated

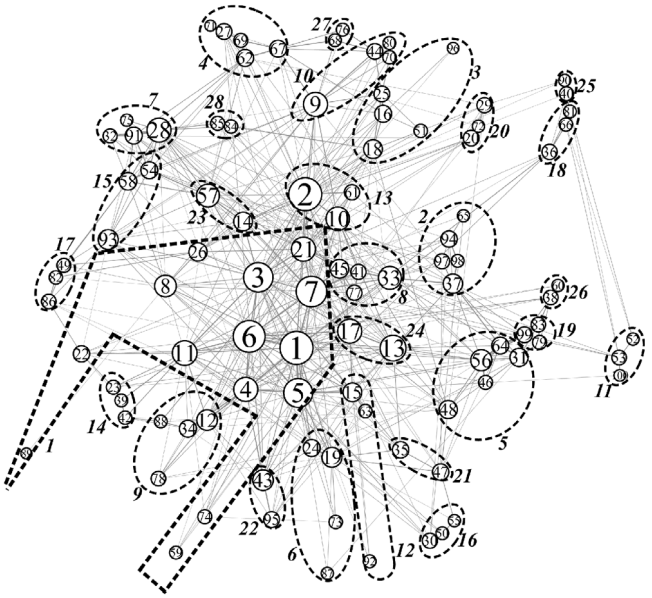


Fig. 3. Visualization of the character network in *Dream of the Red Chamber* with $w_{\text{cut}} = 0.22$. Characters are represented by circles and numbered as in Table 1. The size of a node is proportional to its degree centrality. Level 1 MSC clusters are enclosed by dashed contours. Each inter-character relation is represented as a grey edge.

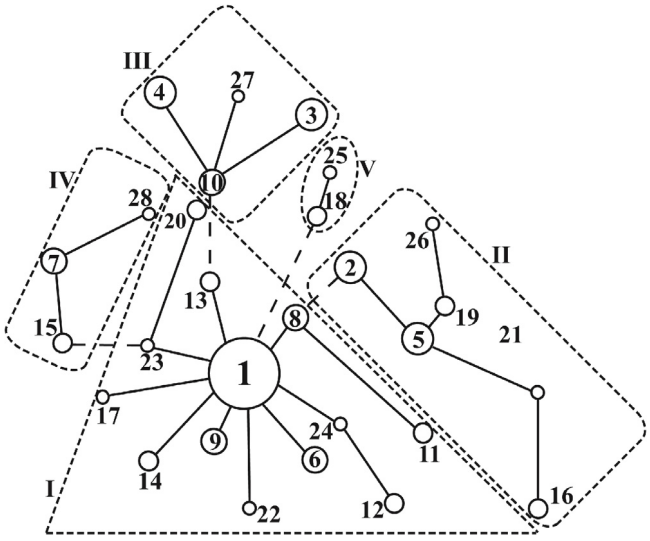


Fig. 4. A minimum spanning tree diagram of the coarse-grained character network in *Dream of the Red Chamber*. Each node represents a level 1 MSC cluster as numbered in Table 2, and its size is proportional to the number of characters in the cluster. Each group of clusters is enclosed by a dashed contour. Solid lines represent intra-group connections, and long dashed lines represent inter-group connections.

silhouette coefficients of the three clustering results justify their overall quality, only MSC is able to effectively cluster the character network in *Dream of the Red Chamber* without isolating characters. Based on our two criteria – overall quality of network clustering, and the prevention of isolated characters in the community structure – we conclude that MSC is more suitable for investigating character networks in fiction.

4.3. Visualizations of the character network

To visualize the network structure of *Dream of the Red Chamber* at the finer resolution level, we constructed a two-dimensional network map representing the characters and their interactions using Gephi, as shown in Fig. 3. Each node in the network represents a character, and is labeled by its index in Table 1. The radius of each node is proportional to the corresponding character's degree, i.e. the number of connections that the character possesses in the network (plus a constant for the visibility of small nodes). Grey lines denote inter-character interactions, with shorter lines representing closer bonds. Each cluster of characters is enclosed by a dashed contour. Cluster 1 is the largest cluster, and it forms the core of the network. Within this cluster, the primary protagonist Chia Pao-yu (1) is surrounded by his close relatives: Lin Tai-yu (4; his lover), Hsueh Pao-chai (6; his wife), Hsi-jen (5; his closest maidservant), Lady Wang (7; his mother), and Lady Dowager (3; his grandmother). Hseuh Pao-chai occupies a central position in the cluster, consistent with her status as a universally-beloved "Renaissance woman". On the other hand, Lin Tai-yu is more isolated; her connections with characters in other clusters are generally remote, except for those in cluster 9 – her maidservants Tzu-chuan (12) and Hsueh-yen (34), as well as the doctor who treats her pulmonary illness, Wang Chi-jen (78).

Compared to the core cluster, cluster 13 – which contains Wang Hsi-feng (2) and two close maidservants (10, 61) – is relatively small. However, Wang Hsi-feng strongly interacts with characters in other clusters. For instance, she is closely connected to characters in cluster 10, including Chia Lien (9; her husband), Yu Erh-chieh (44; Chia Lien's concubine), Chang Hua (70; Yu Erh-chieh's ex-husband), and Chiu-tung (80; Wang Hsi-feng's maidservant). This cluster describes an event where, fueled by vengeance due to Chia Lien's affair with Yu Erh-chieh, Wang Hsi-feng manipulates the latter into committing suicide by leveraging her relationships with Chang Hua and Chiu-Tung.

Fig. 4 shows a two-dimensional visualization of the network at the coarser resolution level as a minimum spanning tree. Each node in the minimum spanning tree represents a cluster obtained from the first iteration of the MSC algorithm, numbered according to its index in Table 2, and the radius of each cluster node is proportional to the number of characters that the corresponding cluster contains. Intra-group connections are represented by solid lines between nodes, while inter-group connections are represented by dashed lines. We only attempted to preserve the distance between cluster nodes that are connected by inter- or intra-group links in the network. Second-level groups of cluster nodes are enclosed by dashed contours. As seen in Fig. 4, group I is connected to group II through clusters 2 and 8. Cluster 2 mainly describes the event where Chia Pao-yu and Wang Hsi-feng are struck down with illness through the witchcraft of Abbess Ma (98) and Concubine Chao (37). Concubine Chao is also connected to her son Chia Huan (33) and Tsai-hsia (97) by the role that she plays in arranging their relationship. Meanwhile, cluster 8 describes Chia Chiao's (41) betrayal and attempted sale into concubinage by her relatives Chia Huan and Wang Jen (77). Although Chia Lan (45) is not involved in this event, he is also included in cluster 8 due to his presence for another event within the same chapter. Group I is also connected to group III through clusters 10 and 13, which, as previously delineated, represent Wang Hsi-feng's plot of revenge against her husband's infidelity. Groups I and IV are linked by clusters 15 and 23, which are both clusters of servants; this link is created by a scene where Pao Yung (58) fends off robbers who have invaded Rongguo House. Finally, groups I and V are connected by the event involving Chia Pao-yu, his servants, and his classmates at the school. Thus, judging by the inter-cluster connections in the network visualizations of Figs. 3 and 4, the community structure detected by the MSC algorithm is congruent with the storyline and overarching themes of *Dream of the Red Chamber*.

Our network visualizations presented can provide a high-level model of the character network's structure. Figs. 3 and 4 suggest

Table 3
Centrality measures of the top 20 characters in *Dream of the Red Chamber*. Characters are numbered according to their indices in Table 1.

Degree	Weighted degree	Closeness	Betweenness	Eigenvector	PageRank
1	1	1	1	1	1
2	6	3	2	6	2
6	7	7	3	7	6
7	2	6	7	3	7
3	3	2	6	5	3
5	5	5	11	2	5
13	4	4	28	4	4
11	13	17	9	17	11
4	11	10	5	13	13
9	28	13	56	21	28
17	17	11	4	11	9
21	21	21	8	10	21
28	9	33	36	12	17
10	10	45	10	43	33
33	33	57	13	57	10
57	57	9	33	28	57
8	56	28	27	33	56
56	12	8	53	15	8
12	37	14	37	45	12

a three-layered hierarchy of characters: the main characters (clusters 1 and 13) form the core of the network, which is surrounded by a secondary layer of closely-allied servants (other clusters in group I), while other servants and relatives –the peripheral characters – constitute the outmost layer (clusters in groups II–V). The visualizations can also be used to elucidate details in the narrative and provide potential answers to questions. For instance, it can be inferred based on Fig. 3 that the most central servants in the plot are Hsi-jen (5), Ping-erh (10), Tzu-chuan (12), and Yuan-yang (14), with Hsi-jen being the most important of the four. These central servants can be identified by their number of appearances, their degree in the network, and their proximity to their masters.

To further interpret our clustering results, we note that the MSC algorithm hinges upon identifying the closest relationship for each character. From a literary perspective, main characters (e.g. Chia Pao-yu and Wang Hsi-feng) are closely related to a large number of characters and are central to the flow of the narrative. Meanwhile, supporting characters (e.g. Chia Cheng and Li Shih-erh) tend to have close relationships with only a few other characters, and their involvement in the plot is likewise limited. Our clustering results, as shown in Figs. 3 and 4, demonstrate that clusters of characters in *Dream of the Red Chamber* are distributed throughout a multi-layer framework, with the main characters occupying a focal position. In most of the peripheral clusters, inter-character relationships are developed over the course of a few chapters; however, the relationships between the main characters are established and augmented throughout the novel. The various interactions between the characters drive the development of the plot as well as the exploration of overarching themes — conflict, threat, death, and defeat, among others. Our cluster analysis provides a useful high-level approach to resolving and analyzing these complex relationships. For example, while the love triangle between Chia Pao-yu, Lin Tai-yu, and Hsueh Pao-chai is readily apparent from the text, some analyses have also postulated that Chia Pao-yu and Shih Hsiang-yun (15) share a close romantic bond throughout the narrative, based on their close interactions as well as the fact that they were childhood sweethearts. However, quantitative data does not suggest that such a relationship was intended by the writer. Although the two characters have proximal positions in Fig. 3, they belong to different clusters, and the link between them does not appear among the five closest relationships for either character.

4.4. Centrality measures of the character network

An important objective of network analysis is to identify nodes that occupy more central roles in the network; such nodes can be identified using measures of centrality [46,47]. To analyze the importance of each character to the narrative of *Dream of the Red Chamber*, we calculated six different centrality measures – the degree, weighted degree, betweenness, closeness, eigenvector, and PageRank centralities – for the novel's character network. The degree centrality is proportional to the number of links (and their weights, in the case of weighted degree centrality) held by a given node, and measures the involvement of a node in the network. Closeness centrality is given by the inverse sum of the shortest distances from the node in question to all other nodes; it measures the accessibility of the node via its neighbors. Betweenness centrality assesses the involvement of a node in the shortest paths between other nodes in the network and indicates the node's ability to control the flow through the network. Eigenvector centrality measures the influence of a node upon a network based on the principle that a connection to a high-scoring node contributes more than a connection to a low-scoring node. A variant of eigenvector centrality is PageRank centrality; the PageRank centrality of a node tends to be high if it is strongly linked or linked to important nodes. We provide mathematical definitions of these centrality measures in the Supporting Information.

Fig. 5 shows the eigenvector (a) and PageRank (b) centrality measures of the character network in *Dream of the Red Chamber*, while Fig. 6 shows the closeness (a) and betweenness (b) centrality measures. The 20 highest-ranking characters for each centrality measure are listed in Table 3. Based on the centralities of the character network, the aforementioned three-tiered structure of the network can be further substantiated and quantified. In Figs. 5 and 6(a), the core layer (enclosed by the solid loop) is formed by the 7 characters with the highest weighted degree, closeness, eigenvector, and PageRank centrality measures. The immediately-adjacent secondary layer consists of 36 characters (enclosed by the dashed loop). Compared to the core layer, the average closeness centrality of the secondary layer tends to be lower (∼80% of the core layer's average), while the eigenvector and PageRank centralities are significantly lower (respectively ∼40% and ∼50% of the core layer's average). The remaining characters in the outermost layer have much lower closeness, eigenvector, and PageRank centralities on average (respectively ∼60%, ∼10%, and ∼25% of the core layer's average). Note that the encircling loops in Figs. 5 and 6 are included only for visualization, and do not reflect the actual shape of each layer's boundaries. The number of characters in the second layer was determined by maximizing the consistency in the constituent characters of the second and the third layers for various centrality measures.

It can be observed in Fig. 6(b) that the betweenness centrality of the character network differs qualitatively in its variation compared to the other centrality measures. For instance, Chia Lien (9) and Chia His-chun (28) have greater betweenness centralities than Lin Ta-yu (4) and Hsin-jen (5), whereas the opposite is true for the other centralities. As characters with large betweenness centralities control the flow of the narrative, the larger betweenness centralities of the former two characters reflect their importance in connecting group I to groups III and IV. Similarly, Chia Huan (33), Ming-yen (36), and Concubine Chao (37) have higher betweenness centralities than other characters in their respective clusters, as a consequence of their role in connecting group I to groups II and V. In general, betweenness centralities are highest for characters in the core layer of the network, and they diminish in the outer layers. We note that the centrality of a given character is not directly related to the frequency of their occurrence since their appearance pattern has been normalized. However, it can be seen that characters with a large number of appearances are also likely to be situated in the core layer, while characters who rarely appear tend to be distributed throughout the outermost layer.

To further explore the properties of the character network, we calculated the Pearson correlation coefficients between different

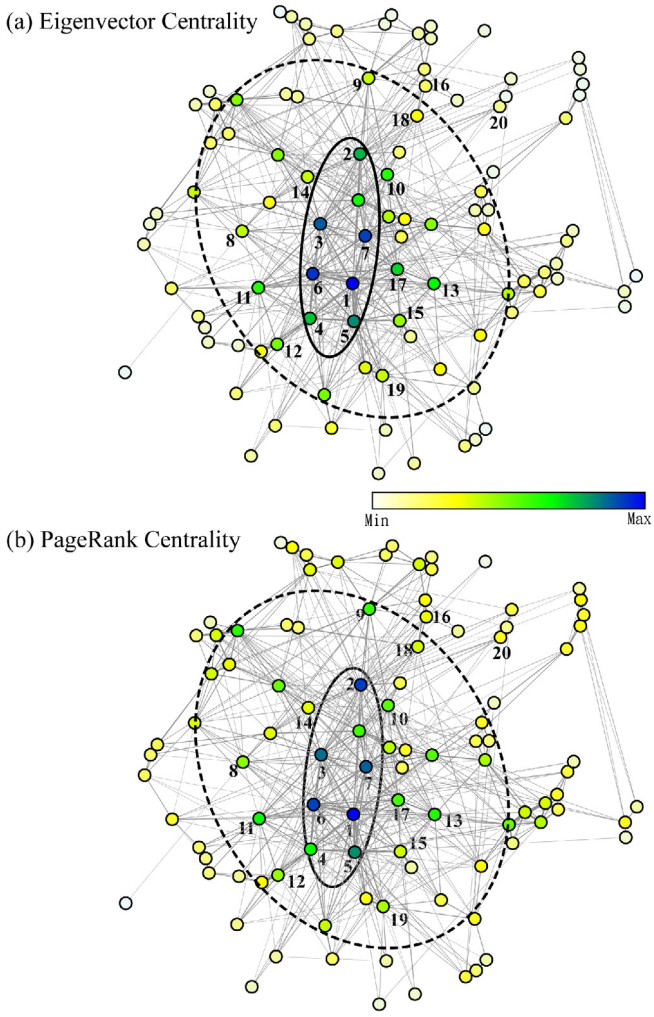


Fig. 5. Eigenvector (a) and PageRank (b) centralities for the character network of *Dream of the Red Chamber*. The character network is the same as that in Fig. 3, except nodes are equally-sized and colored by their centrality measures according to the colorbar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

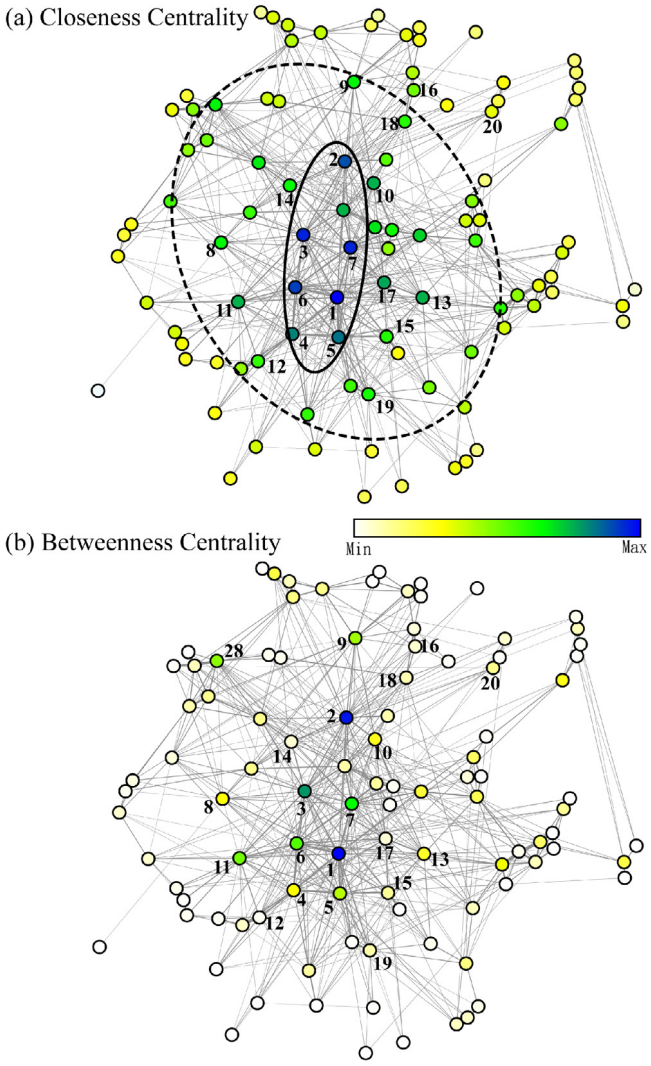


Fig. 6. Closeness (a) and betweenness (b) centralities for the character network of *Dream of the Red Chamber*. The character network is the same as that in Fig. 3, except nodes are equally-sized and colored by their centrality measures according to the colorbar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

centralities. We found that the degree, weighted degree, closeness, eigenvector, and PageRank centrality measures are strongly correlated; the Pearson coefficients between any two of these centralities range from 0.93 to 0.99. However, the Pearson coefficients of the above centralities with betweenness centrality are substantially lower, with an average of 0.84. Similarly, the distributions for all of the centrality measures except for betweenness centrality yield R^2 coefficients of determination between 0.95 and 0.99 when fitted to a logarithmic function. Betweenness centrality, however, is a poorer fit to a logarithmic function, having an R^2 coefficient of 0.89, as shown in Fig. S4. Evidently, betweenness centrality can be differentiated from other centralities as a unique property of the character network, which may be associated with its more dynamic role in representing the flow of the narrative.

5. Conclusions

The present work is primarily concerned with computationally-aided literary analysis. As demonstrated through our analysis of *Dream of the Red Chamber*, the MSC algorithm can be applied to identify, cluster, and visualize the community structure of the character networks in literary works, as well as to calculate centrality measures for individual characters. Our analysis suggests that the character network of *Dream of the Red Chamber* exhibits properties of scale-free and small-world networks. Two characteristic resolution levels were detected by MSC within a reduced character network consisting of the 100 most frequently occurring characters. At the finer resolution level, most peripheral clusters form as consequences of singular events in the narrative. At the coarser resolution level, these peripheral clusters merge into groups, which describe sequences of related events. The central clusters (clusters 1 and 13) in the main group (group I) can be differentiated from other clusters in that they correspond to two overarching themes that permeate the narrative, and they are consequently of higher importance. Network visualizations and centrality measures suggest that the community of characters in the novel forms a three-layered structure. The core layer is formed by 7 main characters with the highest centralities. They are surrounded by 36 secondary characters, with centrality measures approximately half of the core layer's average. Finally, the minor characters in the peripheral-most layer have very small centralities. We note that betweenness centralities for characters differ qualitatively from other centrality measures; characters who bridge discrete events, and thus exert control over the narrative, have large betweenness centralities.

Our integrated analytical approach identifies the most significant inter-character relationships in a literary work; in doing so, it offers an automated means of obtaining a high-level perspective of the work's plot and overarching themes. The ability of network clustering to provide such a panoramic understanding of a work can be leveraged to develop methods for automatic narrative comprehension and generation, which is currently a field of active research [48–50]. A more extensively-automated iteration of our method would facilitate the construction of a database of plots, themes, and inter-character relationships extracted from a corpus of existing novels; such a database can be used to train artificially intelligent agents in autonomous novel writing.

Acknowledgments

RHGC carried out the study of the character network in *Dream of the Red Chamber* and drafted the manuscript. CCC compared existing versions of *Dream of the Red Chamber* and participated in the preparation of data and figures and the interpretation of results. CMC conceived of the study and wrote the main manuscript text. This work was supported in part by the Ministry of Science and Technology of Taiwan under grant no. MOST105-2112-M-003-003-MY3. CMC thanks R. Ng for the hospitality at the Department of Computer Science, University of British Columbia.

Conflict of interest

CMC declares that there are no competing financial interests that might have influenced the performance or presentation of the work described in this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2018.10.005.

References

[1] I. Mani, Computational Modeling of Narrative, Morgan & Claypool Publishers, 2013.
[2] R. Siemens, S. Schreibman, J. Unsworth, A Companion to Digital Humanities, Blackwell, Oxford, 2004.
[3] J.-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, E.L. Aiden, Quantitative analysis of culture using millions of digitized books, Science 331 (2011) 176–182.
[4] S. Hockey, Electronic Texts in the Humanities, Oxford University Press, Oxford, 2000.
[5] F. Moretti, Graphs, Maps, Trees: Abstract Models for a Literary History, Verso, 2005.
[6] D.K. Elson, N. Dames, K.R. McKeown, Extracting social networks from literary fiction, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 138–147.
[7] J. Lee, C.Y. Yeung, Extracting networks of people and places from literary texts, in: Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation, 2012, pp. 209–218.
[8] A. Agarwal, A. Kotalwar, O. Rambow, Automatic extraction of social networks from literary text: A case Study on Alice in Wonderland, in: Proceedings of the 6th International Joint Conference on Natural Language Processing, Nagoya, Japan, 2013.
[9] M.C. Ardanuy, C. Sporleder, Structure-based clustering of novels, in: Proceedings of the 3rd Workshop on Computational Linguistics for Literature, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 31–39.
[10] M.C. Ardanuy, C. Sporleder, Clustering of novels represented as social networks, Linguist. Issues Lang. Technol. 12 (2015).
[11] D.A. Skorinkin, Extracting character networks to explore literary plot dynamics, in: Proceedings of the International Conference Dialogue 2017, Moscow, Russia, 2017.
[12] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[13] A. Celikyilmaz, D. Hakkani-tur, H. He, G. Kondrak, D. Barbosa, The actor-topic model for extracting social networks in literary narrative, in: Proceedings of the NIPS 2010 Workshop – Machine Learning for Social Computing, 2010.
[14] Y. Rochat, Character Networks and Centrality, Université de Lausanne, 2014.
[15] I. Samoylenko, T.C. Chao, W.C. Liu, C.M. Chen, Visualizing the scientific world and its evolution, J Am Soc Inf Sci Tec 57 (2006) 1461–1469.
[16] R.H.G. Chen, C.-M. Chen, Visualizing the world's scientific publications, J. Assoc. Inf. Sci. Technol. 67 (2016) 2477–2488.
[17] S.D. Prado, S.R. Dahmen, A.L.C. Bazzan, P.M. Carron, R. Kenna, Temporal network analysis of literary texts, Adv. Complex Syst. 19 (2016) 1650005.
[18] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (2002) 7821–7826.
[19] G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, Self-organization and identification of web communities, Computer 35 (2002) 66–70.
[20] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, second ed., 2008.
[21] J.H. Ward Jr, Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc. 58 (1963) 236–244.
[22] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[23] C.M. Chen, Classification of scientific networks using aggregated journal-journal citation relations in the journal citation reports, J Am Soc Inf Sci Tec 59 (2008) 2296–2304.
[24] J. Gower, G. Ross, Minimum spanning trees and single linkage cluster analysis, Appl. Stat. 18 (1969) 54–64.
[25] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 881–892.
[26] C.-Y. Liou, W.-P. Tai, Conformality in the self-organization network, Artificial Intelligence 116 (2000) 265–286.
[27] D.B. Vincent, G. Jean-Loup, L. Renaud, L. Etienne, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (2008) P10008.
[28] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (2006) 8577–8582.
[29] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[30] S. Fortunato, M. Barthélemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. 104 (2007) 36–41.
[31] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[32] D. Bamman, T. Underwood, N.A. Smith, A bayesian mixed effects model of literary character, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 370–379.
[33] P.A. Jayannavar, A. Agarwal, M. Ju, O. Rambow, Validating literary theories using automatic social network extraction, in: Proceedings of the NAACL-2015 Workshop on Computational Linguistics for Literature, Association for Computational Linguistics, Denver, Colorado, USA, 2015, pp. 32–41.
[34] H. Vala, D. Jurgens, A. Piper, D. Ruths, Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 769–774, 2015.
[35] J. Gao, M. Li, A. Wu, C.-N. Huang, Chinese word segmentation and named entity recognition: a pragmatic approach, Comput. Linguist. 31 (2005).
[36] Y. Long, D. Xiong, Q. Lu, M. Li, C.-R. Huang, Named entity recognition for chinese novels in the ming-qing dynasties, in: M. Dong, J. Lin, X. Tang (Eds.), Chinese Lexical Semantics: 17th Workshop, CLSW 2016, Springer International Publishing, Cham, Switzerland, 2016, pp. 362–375.
[37] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Ann Arbor, Michigan, USA, 2005, pp. 363–370.
[38] R. Levy, C. Manning, Is it harder to parse chinese, or the chinese treebank? in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Sapporo, Japan, 2003, pp. 439–446.
[39] Y.F. Chang, C.M. Chen, Classification and visualization of the social science network by the minimum span clustering method, J. Am. Soc. Inf. Sci. Technol. 62 (2011) 2404–2413.
[40] C.M. Chen, Y.F. Chang, Visualizing the clustering of financial networks and profitability of stocks, J. Complex Netw. 3 (2014) 303–318.
[41] T.-L. Mai, G.-M. Hu, C.-M. Chen, Visualizing and clustering protein similarity networks: sequences, structures, and functions, J Proteome Res 15 (2016) 2123–2131.
[42] G.-M. Hu, T.-L. Mai, C.-M. Chen, Clustering and visualizing similarity networks of membrane proteins, Proteins 83 (2015) 1450–1461.
[43] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software, PLoS One 9 (2014) e98679.

[44] J.B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, Proc. Amer. Math. Soc. 7 (1956) 48–50.
[45] G.-M. Hu, T.-L. Mai, C.-M. Chen, Visualizing the GPCR network: Classification and evolution, Sci. Rep. 7 (2017) 15495.
[46] L.C. Freeman, Centrality in social networks conceptual clarification, Social Networks 1 (1978) 215–239.
[47] T. Opsahl, F. Agneessens, J. Skvoretz, Node centrality in weighted networks: generalizing degree and shortest paths, Social Networks 32 (2010) 245–251.
[48] J. Porteous, F. Charles, M. Cavazza, NetworkING: Using Character Relationships for Interactive Narrative Generation, in: T. Ito, C. Jonker, M. Gini, O. Shehory (Eds.) Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems, Saint Paul, Minnesota, USA, 2013.
[49] M.O. Riedl, R.M. Young, Narrative planning: Balancing plot and character, J. Artificial Intelligence Res. 39 (2010) 217–268.
[50] S. Chaturvedi, M. Iyyer, H. Daumé, III, unsupervised learning of evolving relationships between literary characters, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2017.