Complicating the Social Networks for Better Storytelling: An Empirical Study of Chinese Historical Text and Novel

CHENHAN ZHANG, Southern University of Science and Technology

Digital humanities is an important subject because it enables developments in history, literature, and films. In this paper, we perform an empirical study of a Chinese historical text, Records of the Three Kingdoms (*Records*), and a historical novel of the same story, Romance of the Three Kingdoms (*Romance*). We employ natural language processing techniques to extract characters and their relationships. Then, we characterize the social networks and sentiments of the main characters in the historical text and the historical novel. We find that the social network in *Romance* is more complex and dynamic than that of *Records*, and the influence of the main characters differs. These findings shed light on the different styles of storytelling in the two literary genres and how the historical novel complicates the social networks of characters to enrich the literariness of the story.

CCS Concepts: • **Networks** → *Topology analysis and generation.*

Additional Key Words and Phrases: natural language processing, social network analysis

ACM Reference Format:
Chenhan Zhang. 2020. Complicating the Social Networks for Better Storytelling: An Empirical Study of Chinese Historical Text and Novel. 1, 1 (August 2020), 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital humanities is a transdisciplinary subject between information technologies and humanities, such as literary classics. For instance, Google makes a contribution to digital humanities by promoting the "Google Books Library Project" which includes millions of paper books scanned into electronic text [33]. Digital text is easier for researchers to explore than printed books, since the development of information technology has provided numerous effective tools [35]. In the past decade, overwhelming data science techniques have advanced the research on digital humanities; thus, components can be extracted and analyzed from literature.

A review of previous research reveals that some areas in digital humanities remain unexplored. First, mainstream studies are limited to the humanities works on the background of the Western world [19]. It is both interesting and constructive to investigate humanities works with oriental backgrounds. Second, only a few comparative studies on literature with different styles of the same story are conducted [39]. Particularly, previous researches focused more on longitudinal studies, wherein researchers usually adopt a story series, such as Harry Potter Books 1âĂŞ7, as the object of study [12]. A potential research interest about the same story that discovers varied features (narrative levels, characters, events) or sentiments can arise from different literature, which may be driven by literary genres or authorsâĂŹ opinion,

Author's address: Chenhan Zhang, zhangch@mail.sustech.edu.cn, Southern University of Science and Technology, 1099 Xueyuan Ave., Shenzhen, Guangdong, 518052.

among others. Third, network study is essential for the social network of a story and any network that possesses a topological structure, which can help gain an insight into the storyâĂŹs characters based on its grand narration [30].

To fill the gap, this paper introduces a social network and sentimental analysis work on two different texts of one of the most famous Chinese story, The Three Kingdoms. In particular, we leverage the state-of-the-art natural language processing (NLP)-based model to extract the social networks in the narratives of two books. A series of descriptive statistical analysis on the extracted networks is conducted, and we discover the homogeneity and heterogeneity in terms of topological features in these networks. Additionally, we adopt the sentimental analysis to compare the evaluations on some of the main characters. The results reveal that the social network is more complicated in the narrative of the novel (*Romance*) than that of the historical text (*Records*). Consequently, it can be concluded that the literariness of stories has a tight relationship with the complexity of the social networks they entail.

The main contribution of this paper is as follows:

- We integrate the latest NLP and network science techniques to extract and analyze the social networks of historical text and novel.
- We depict the difference in the dynamic social networks of the Records and the Romance, the classic historical text, and novel of the same story.
- A series of comprehensive case studies are performed, and we find that the historical novel complicates the social networks of characters to enrich the literariness of the story.

The remainder of this paper is organized as follows. In Section 2, the backgrounds of text mining and social network analysis researches are presented. Section 3 elaborates on the network extraction approach. We perform a series of empirical studies in Section 4 to demonstrate the thesis of this work. Finally, this paper is concluded in Section 5 with a summary of potential future studies.

2 LITERATURE REVIEW

2.1 Social Network Analysis (SNA)

Previous studies have demonstrated the importance of network analysis in different domains, such as complex network in supply chains [5] and risk identification in electric industries [3]. For networks that possess a social structure, SNA can be used to study social structures by analyzing the relationships, communities, and activities through topology graph theory [17, 26]. Initially, the study of SNA focuses on the network that actually exists, such as mobile social networks [18] (Table 1 categorizes the metrics of SNA according to various features of social network). The development of NLP enables the extraction of the latent social network in narratives, such as literary text and news text (narrative network analysis). Recently, studies focus on narrative networks in literary works such as novels. For example, studies on Harry Potter find salient, small-world, and scale-free features in its social network [38, 41], and these features reveal that the story is penetrated by compact character relationships.

Table 1. Metrics of SNA

Connections	Homophily [22], Multiplexity [10], Reciprocity [14], Network Closure [16], Propinquity
Distributions	Bridge, Centrality [6], Density, Distance [28], Structural holes [9], Tie Strength
Segmentation	Cliques, Clustering Coefficient, Cohesion

2.2 Text Mining and Natural Language Processing

2.2.1 Named Entity Recognition. Named Entity Recognition (NER) is among the core tasks in NLP. In story-oriented text ming, the NER task requires that the characters and sentiment representatives are treated as entities and can be identified in the texts [7]. A bulk of computational linguistic-based NER methods are developed, which plays vital roles in NER tasks, especially the token-level tasks [21, 34].

2.2.2 Part-of-speech tagging. Part-of-speech (POS) tagging is the process of tagging a token (a word) for a particular part of speech according to its context [20]. Table 2 shows each type of tag with its corresponding meaning. POS tagging helps form related grammatical rules for different language patterns.

Table 2. POS tags [27]

Tag	Meaning
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential ÃŠthereÃŞ
FW	Foreign word
IN	Preposition of subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
TO	Infinitive marker ÃŠtoÃŞ
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-third person singular present
VBZ	Verb, third person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb
XNOT	Not and nÃŢt

2.3 Deep Learning-based NLP Models

To extract the social network of a story, characters and their connections among one another must be identified. The distribution of characters in a story is scattered and sometimes connotative. Natural language processing (NLP) technologies automate the identification of this specific information in texts, which can be a useful weapon [24].

The popularity of deep learning has facilitated designing a number of related models to handle the subtasks of NLP, such as NER. Google proposed BERT [13], which substantially overcomes the limitations of existing models. BERT is based on Open AI GPT and performs attention mechanism on its model [31, 37]. It can predict the correct textual ID according to its entire context without a single directional limitation. In actual cases, BERT distinctly outperforms existing models in various metrics. For reference of the readers, Table 3 compares the capacities of the most widely adopted NLP models.

Table 3. NLP models

Model	Level of long-distance semantic obtaining	Parallel	Bidirectional context
Word2vec [23]	1	✓	
Unidirectional LSTM [40]	2		
ELMo [29]	2		✓
OpenAI GPT [31]	3	✓	
BERT	**3**	✓	✓

3 PRELIMINARY

3.1 The Story of the Three Kingdoms

Recently, the popular video game âĂIJTotal War: Three KingdomsâĂİ captured the attention of numerous fans, and the story of the Three Kingdoms has been explained in detail to the audience. The story of the Three Kingdoms depicts the splendid and complex plot across the three kingdoms that emerged from the remnants of the Han Dynasty in the 14th century. The two most famous books based on this story, *“Records of the Three Kingdoms”* and *“Romance of the Three Kingdoms”*, are chosen as the research objects. The former, written by Chen Shou (B.C. 233âĂŞ297), who was an official and a historian, is a biographical, historical text that chronicles the events in the three Kingdoms era by combining the respective histories of the three kingdoms. The latter is a couplet historical novel, written by the famous novelist Luo Guanzhong (B.C. 1320âĂŞ1400). Its narrative is part historical, part legend, and part mythical, wherein historical facts are combined with personal opinion and folklore.

Both booksâĂŹ original versions are written in classical Chinese. Text mining may lead to biases due to the complicated syntactic rules of classical Chinese and the various entities of vocabulary (i.e., similar wordings can have different meanings). In comparison, text mining in English texts is simpler, and in addition provides more straightforward intuition to non-Chinese audience. In this context, we choose a version of *Romance* by CH Brewitt-Taylor [8] and a version of *Records* by Wilt L. Idema et al. Although they are essentially similar to the original literature, differences such as the number of characters and the framework of the story indeed exist. For readers who intend to investigate the original literature, the results of this paper are for their reference only.

3.2 SQuAD Corpus

To train deep learning-based NLP models in supervised or semi-supervised manners, a text corpus is required as the training dataset. In this work, we employ Stanford Question Answering Dataset (SQuAD) as the text corpus. The design of SQuAD is inspired by answering questions from reading comprehension [32]. Unlike the previous datasets, the

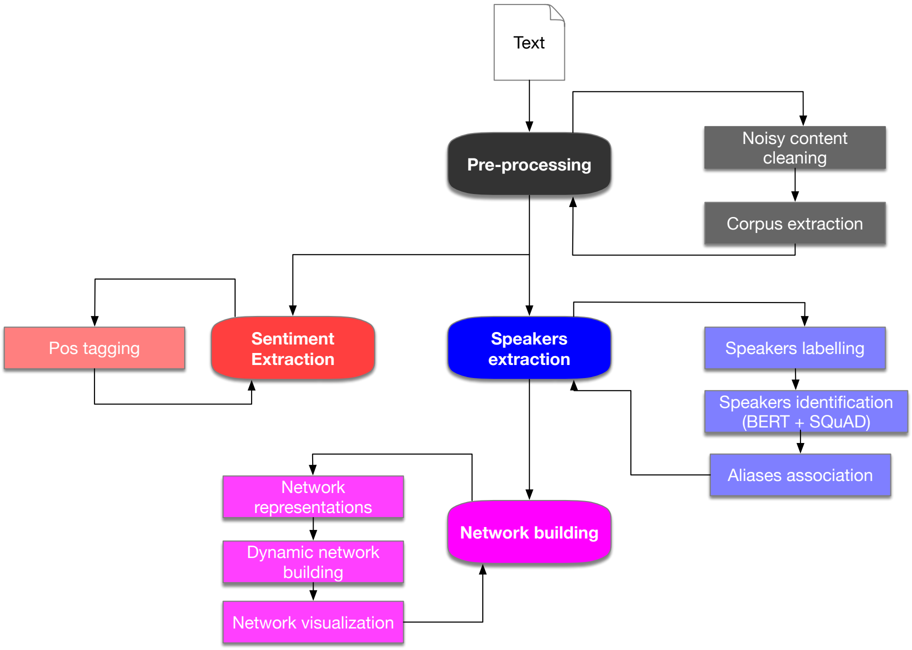


Fig. 1. Overview of the entire text mining process.

mechanism of SQuAD requires machines to select the answer from all possible candidates in the contexts rather than from a list of possible answers for each question. The answer is sometimes not a single word but a phrase, which makes the answer difficult to predict. Therefore, the robustness of models can be improved through such rigorous learning.

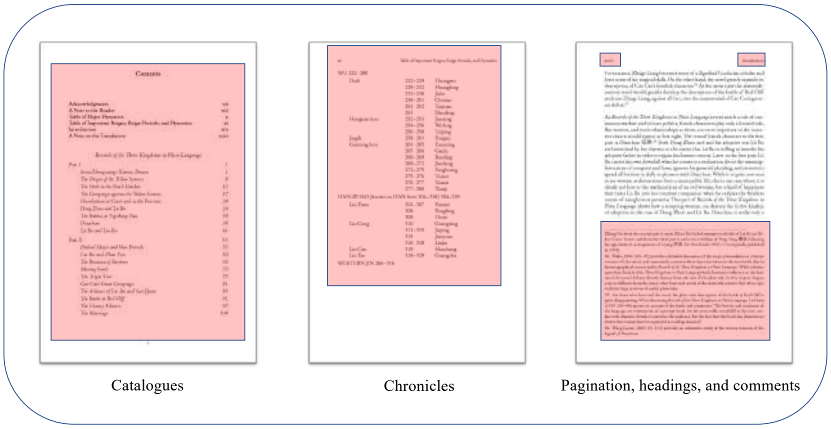
In this paper, data from literary texts are limited. Therefore, we use a BERT + SQuAD method that can substantially address the problem because it can considerably improve prediction accuracy despite limited data [13]. Furthermore, some traditional methods are still adopted in such situations.

3.3 Text Mining

In this work, we propose a text mining algorithm to extract the social networks in the narratives. We first pre-process the raw text to clean out the noise in the text and extract the accessible text from the narrative as the corpus. Then, we identify the characters from the corpus. Meanwhile, we also achieve sentiment extraction. Finally, the extracted characters are utilized to construct the social network. The schematic view of our text mining algorithm is illustrated as Figure 1.

3.3.1 Pre-processing. Raw text is required to be cleaned and further normalized to a specific format (i.e., corpus) for the processing of the algorithm. Pre-processing work is relatively simplified in this work since the adoption of the deep learning-based tool enables a loose format of the text in the corpus.

Regular expression in data cleaning. Noisy contents are expected to be adjusted or eliminated because they are mixed with useful data, which may mislead results. Such content usually includes tables of contents, titles, headers. Fortunately, most of these noises usually follow a specific format. For example, as a translation of historical records, the âĂIJ*Records*

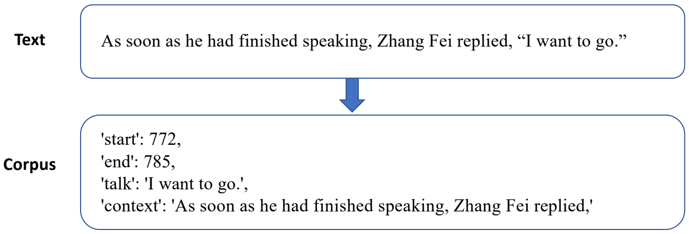


Catalogues

Chronicles

Pagination, headings, and comments

Fig. 2. Noisy content in text mining.



Text

As soon as he had finished speaking, Zhang Fei replied, “I want to go.”

Corpus

```
'start': 772,
'end': 785,
'talk': 'I want to go.',
'context': 'As soon as he had finished speaking, Zhang Fei replied,'
```

Fig. 3. Mapping from the text to the corpus.

*of the Three Kingdoms in Plain Language*âĂİ includes a multitude of notes (See Figure 2), where they follow the same format that starts with a serial number that leads the content. A similar phenomenon is also observed in broken words, which are all split by a hyphen or a space. Regular expressions can be used to effectively eliminate this type of noisy text by developing corresponding rules.

Corpus Extraction. Independently building a character-oriented corpus instead of basing on the existing corpus is essential for the character extraction task in this work. We assume that in a narrative, characters usually perform in conversations; hence, their identification is focused on such conversations. Each conversation consists of several dialogs, which usually follow a specific double quotation mark format, that is, one paragraph starts with the double opening quote (âĂIJ) and ends with the double closing quote (âĂİ). Following this rule allows all dialogs to be extracted from where conversations are located. In addition, the context in which a conversation occurs commonly contains the characters (a.k.a. speakers). The context usually follows the dialog, which is easy to identify. Therefore, each item of the corpus consists of two parts, âĂIJcontextâĂİ and âĂIJtalk,âĂİ which map to their corresponding content (See Figure 3).

3.3.2 Speakers Extraction.

Labelling the speakers. ConversationsâĂŹ portrayal varies in the storytelling. Similarly, the location of the speaker in a dialogical context differs considerably, thereby making it difficult to identify in an automated way. Therefore, a manual

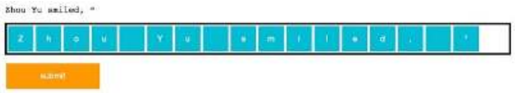
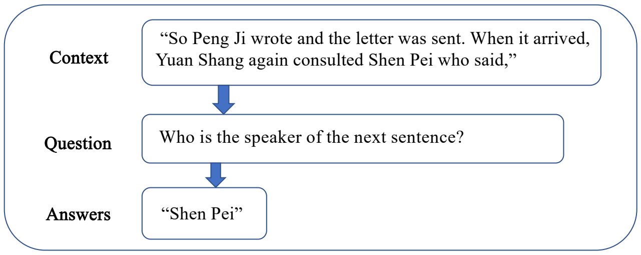

Zhou Yu smiled, "

Fig. 4. A visual resolution for data-labeling.


Context

"So Peng Ji wrote and the letter was sent. When it arrived, Yuan Shang again consulted Shen Pei who said,"

Question

Who is the speaker of the next sentence?

Answers

"Shen Pei"

Fig. 5. An example of SQuAD.

labeling process is required to locate the speaker in each context accurately. Given that this process is time-consuming, a GUI-labeling tool based on Jupyter Notebook is developed, and the visual operation substantially facilitates manual work (See Fig. 4). In this work, a total of 1,702 items from *Romance* can be labeled within just three hours.

Data augmentation. The size of data extracted from books is usually insufficient to reach a promising number of training samples, and it may result that the deep learning-based models cannot achieve a satisfactory prediction result [36]. In this work, the speaker corpus of Records collected only 1,248 items, and a measly portion of 806 samples (64.5%) are labeled after transcribing the entire text. To address this issue, a data augmentation approach is introduced to generate a sufficient number of new annotated data.

How to generate new data and how much data should be generated are essential questions to answer. All speakers are assumed to be included in all the contexts. Supposing a total of $\mathcal{S}$ labelled speakers and $\mathcal{M}$ contexts, we can generate $\mathcal{D}_A = \mathcal{S} * \mathcal{M}$ new data samples. In this work, we use this data augmentation method and obtain over a million new data samples, as shown in Table 4.

Table 4. Data augmentation

	$\mathcal{S}$	$\mathcal{M}$	$\mathcal{D}_A$
Romance	664	1702	1130128 (approx.)
Records	806	1248	1005888 (approx.)

Speakers Identification. A BERT + SQuAD algorithm is used to build a speaker prediction model in this work. SQuAD provides a structure to answer the question (prediction) by comprehending the context. Referring to the structure of SQuAD, we structured a ternary dataset (i.e., context, answer, question). (See Figure 5 for an example of the dataset).

BERT provides a contextual prediction algorithm, and we use this model to predict the speakers from the text. Specifically, we employ GoogleâĂŹs `BERT-Base-Multilingual-Cased` model as the pre-trained model, which incorporates 12-layer, 768-hidden, 12-heads, 110 million parameters; the pre-trained model is then used to fine-tune our dataset. Note

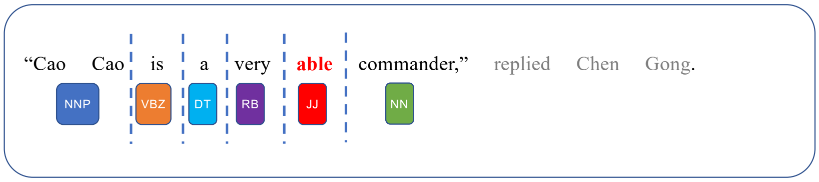


Fig. 6. Sentence tokenization with POS tags.

that we omit the training procedure of BERT since it is not among the main focuses, interested readers can refer to our codes available at https://github.com/GreatWizard9519/Social-network-extraction-and-analysis-of-Three-kingdoms. No related baseline evaluates the training effort because this study is an individual project. The predicted results after manual proofreading covers the vast majority of characters in the books (approximately 93% on *Romance* and 91 on *Records*). In this way, the appearing characters can be obtained from the prediction result.

Aliases association. More than a few characters in the two books possess one or more aliases. For example, âĂIJXuan-deâĂİ, âĂIJLord LiuâĂİ, and âĂIJThe First RulerâĂİ all refer to the character âĂIJLiu BeiâĂİ. To overcome this problem, an alias-matching mechanism is built to map aliases of the characters. A flaw of this mechanism in practical use is that the shared family name or title may be mapped to multiple characters. For example, âĂIJSimaâĂİ can be mapped to âĂIJSima YiâĂİ and âĂIJSima ZhongxiangâĂİ. We develop two solutions that can solve this problem. First, the aliases mapping is classified according to the chapters of the story. For instance, âĂIJSima ZhongxiangâĂİ is a character who simply appears at the beginning of Records; hence, the mapping: âĂIJSimaâĂİ to âĂIJSima ZhongxiangâĂİ should solely be applied at the first few chapters. Second, the context is considered when mapping aliases. For example, when âĂIJLiu BeiâĂİ appears, the closest âĂIJLordâĂİ should be âĂIJLord LiuâĂİ (i.e., âĂIJLiu BeiâĂİ) with a high possibility.

3.3.3 Sentiment Extraction. While the extraction and analysis w.r.t. sentiment is not a main focus of this paper, we still conduct related simple studies on some key characters to make the audience gain a deeper understanding of the story. Sentiments toward a character can be differently described. In this work, our sentiment analysis focuses on evaluative words. Other characters who comment about a certain character is a good entry to extract evaluations. Figure 6 shows one of âĂIJChen GongâĂİ's evaluations on âĂIJCao CaoâĂİ in *Romance*.

Therefore, the extraction of such evaluative words is applied. First, all conversations involving a specific character are extracted and tokenized, and each token is tagged with the corresponding part of speech (i.e., POS tag). Subsequently, following the example shown in Figure 6, words that possess an adjective POS (tagged with âĂIJJJâĂİ) and collected since we consider them as "evaluative words" to characters, which can be utilized in sentiment analysis.

3.4 Network Building

3.4.1 Representations. Upon the collection of characters and the interaction that represents the nodes and edges, we can construct social networks. The essential representations for our extracted social networks are defined below:

- **Nodes:** For each character coming on stage, a node is built. As aforementioned, all characters are from the identified speakers; hence, the social network merely describes the relationship between characters who have monologues or dialogs. It is worth noticing two phenomena when using this node representation. First, the number of nodes is less than the actual number of characters that appear in the books. Second, there appear

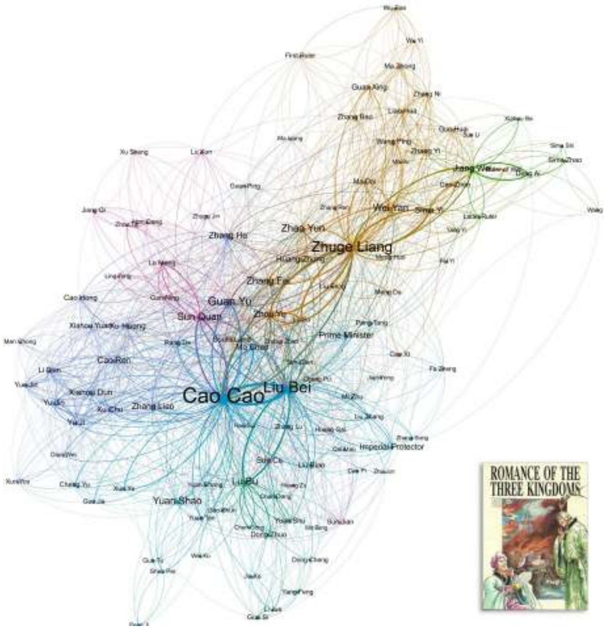


Fig. 7. Network extracted from *Romance* (only show nodes whose degree is larger than 6).

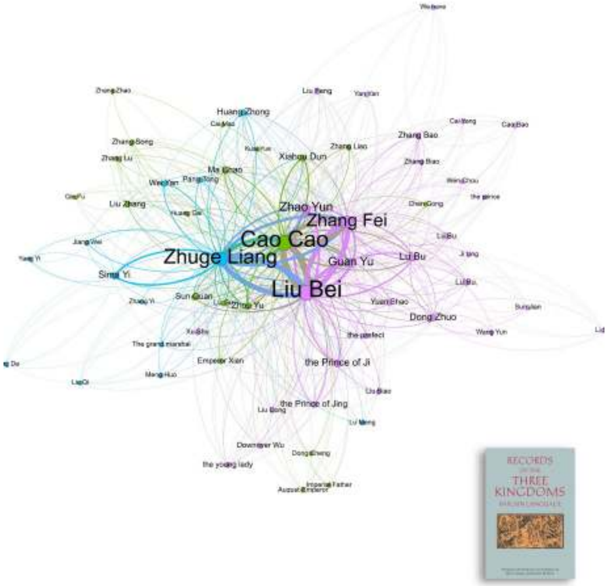


Fig. 8. Network extracted from *Records* (only show nodes whose degree is larger than 6).

some characters are isolated without any interactions with other characters (i.e., nodes whose degree is 0) in our networks. To ensure the completeness of the social network, we manually append some of the missing characters and meanwhile include the isolated nodes when constructing the networks.

- **Edge:** To correlating the nodes, namely, construct edges, we establish an assumption that the adjacent appearance of characters will serve as the basis for creating interactions. Such an assumption is seemingly a coarse-grained solution. However, the outcome will have a high degree to match the actual situation when the size of the involved data is large enough. Based on this assumption, an algorithm established that an interaction (edge) is built when two adjacent characters are detected in the same context. Furthermore, on the account that the representation of edge describes a reciprocal relationship, the network is thereby considered as a bidirectional graph, wherein the values of in-degree and out-degree of every single node are equivalent.

3.4.2 Dynamic Network. Unlike others, the social network extract from narrative will grow as the story carries on. Investigation of network dynamics can help us gain a better insight into the story. To this end, the texts of the two books are chronologically split into five stages, and their corresponding networks are extracted though the same method introduced above. Some key events are set aside as separate markers to normalize the distribution of each stage due to the difference in the chapter settings of the two books, for instance, âĂIJthe death of Dong ZhuoâĂİ and âĂIJthe death of Liu BeiâĂİ. Moreover, these five stages represent the five most prominent periods in the story of the Three Kingdoms. Joining the five separate networks, a dynamic network with evolving growth across the five stages is obtained.

3.4.3 Network Visualization. In this work, we use Gephi [4] to visualize the extracted social networks. To present a clear visual effect, the two demonstrated networks (See Figure 7 for *Romance* and Figure 8 for *Records*) have been filtered to only include the characters (nodes) whose degree is greater than 6. Nodes are classified by using different colors and sizes, wherein the size of nodes is ranked from its value of degree. Moreover, the color of nodes is determined by their communities categorized by modular algorithms for aesthetic needs.

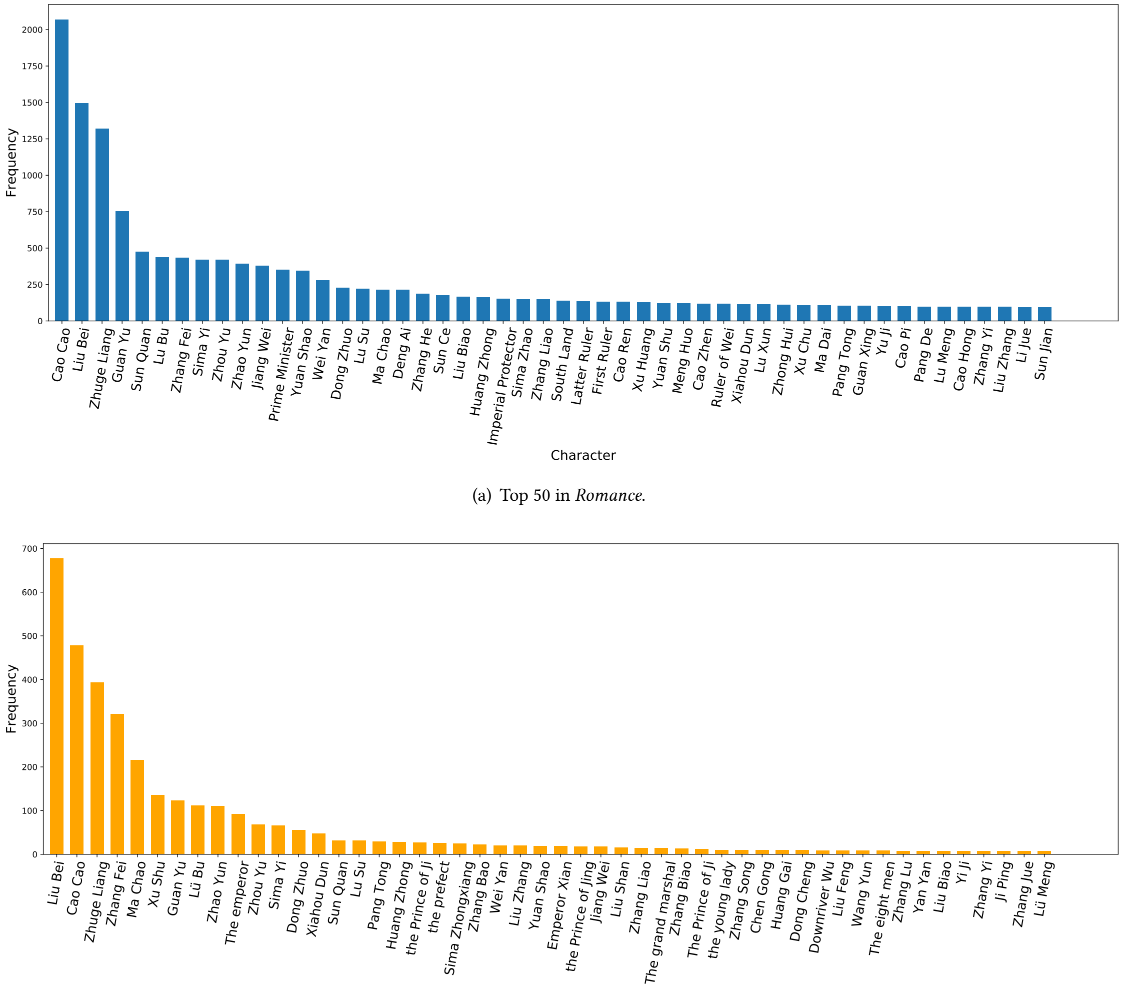
4 RESULTS AND DISCUSSION

This study focuses on exploring the discrepancy or similarity of multiple dimensions between the two books of the Three Kingdoms by employing social network analysis, and further gain an insight into the storytellings entailed in the two books. Our analysis incorporates two dimensions. First, a holistic analysis on the social networks extracted from the two books is introduced wherein global properties are emphatically considered. Subsequently, the observation on some protagonists will be discussed. To present the research logic, in the following investigations, we will first raise some interesting questions and approach them with rational explanations from the analysis results.

4.1 Global Network Analysis and Comparison

4.1.1 History vs Romance: Which is Grander? The framework of a great story is grand, which generally involves numerous characters, intricate relationships, and thus entails a vast social network. In this work, we measure the "grandness" of the two social networks by using the three metrics as below:

- $\mathcal{N}$: The number of characters who appear in the story (i.e., number of nodes).
- $\mathcal{E}$: The number of interactions that occur in the story (i.e., number of edges).
- d: The shortest distance between the two most distant nodes in the network (i.e., the diameter of network).



(a) Top 50 in *Romance*.

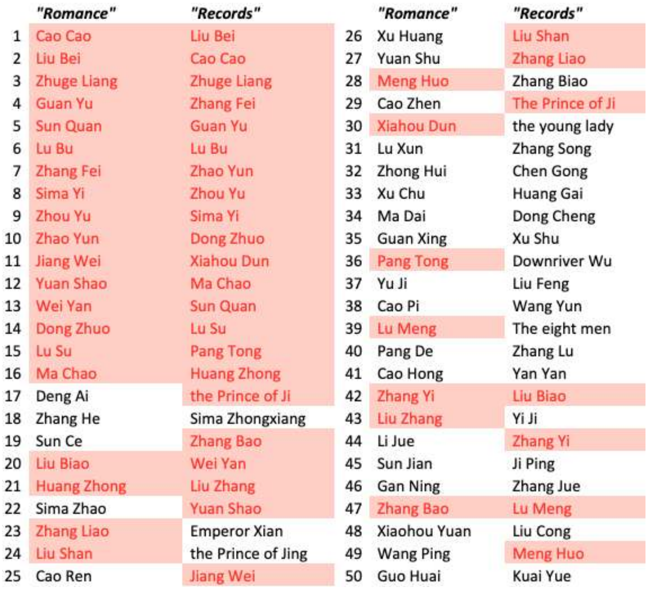
(b) Top 50 in *Records*.

Fig. 9. The frequency of characters occurences.

The results of the comparison is shown in Table 5, where the number of nodes of *Romance* is four times of that of *Records*. Additionally, the diameter of *Romance* is 9 while the one of *Records* is only 3. Consequently, there are more interactions (i.e., edges) in *Romance*. It implies that the social network of Romance is grander than that of Records.

Table 5. The number of nodes and edges, and the diameter of the two networks.

	$\mathcal{N}$	$\mathcal{E}$	d
Romance	510	13,348	9
Records	128	8,986	3



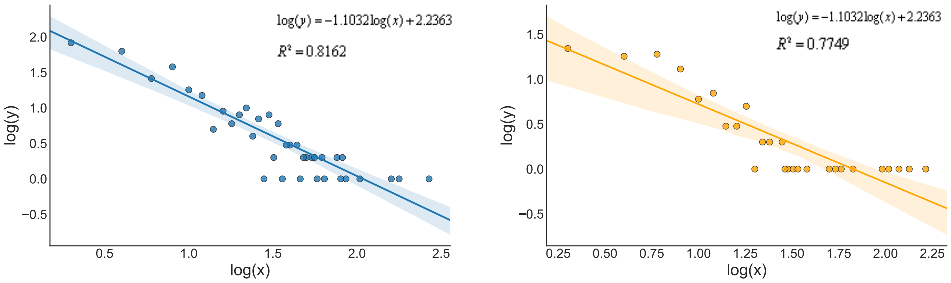
	"Romance"	"Records"		"Romance"	"Records"
1	Cao Cao	Liu Bei	26	Xu Huang	Liu Shan
2	Liu Bei	Cao Cao	27	Yuan Shu	Zhang Liao
3	Zhuge Liang	Zhuge Liang	28	Meng Huo	Zhang Biao
4	Guan Yu	Zhang Fei	29	Cao Zhen	The Prince of Ji
5	Sun Quan	Guan Yu	30	Xiahou Dun	the young lady
6	Lu Bu	Lu Bu	31	Lu Xun	Zhang Song
7	Zhang Fei	Zhao Yun	32	Zhong Hui	Chen Gong
8	Sima Yi	Zhou Yu	33	Xu Chu	Huang Gai
9	Zhou Yu	Sima Yi	34	Ma Dai	Dong Cheng
10	Zhao Yun	Dong Zhuo	35	Guan Xing	Xu Shu
11	Jiang Wei	Xiahou Dun	36	Pang Tong	Downriver Wu
12	Yuan Shao	Ma Chao	37	Yu Ji	Liu Feng
13	Wei Yan	Sun Quan	38	Cao Pi	Wang Yun
14	Dong Zhuo	Lu Su	39	Lu Meng	The eight men
15	Lu Su	Pang Tong	40	Pang De	Zhang Lu
16	Ma Chao	Huang Zhong	41	Cao Hong	Yan Yan
17	Deng Ai	the Prince of Ji	42	Zhang Yi	Liu Biao
18	Zhang He	Sima Zhongxiang	43	Liu Zhang	Yi Ji
19	Sun Ce	Zhang Bao	44	Li Jue	Zhang Yi
20	Liu Biao	Wei Yan	45	Sun Jian	Ji Ping
21	Huang Zhong	Liu Zhang	46	Gan Ning	Zhang Jue
22	Sima Zhao	Yuan Shao	47	Zhang Bao	Lu Meng
23	Zhang Liao	Emperor Xian	48	Xiaohou Yuan	Liu Cong
24	Liu Shan	the Prince of Jing	49	Wang Ping	Meng Huo
25	Cao Ren	Jiang Wei	50	Guo Huai	Kuai Yue

Fig. 10. Similar characters that appear in both booksâĂŹ top 50 ranks are highlighted in red.

4.1.2 Similar Casts? Since the two books narrate the same story, we assume the casts of them (appearing characters) are similar. According to the statistical results, we find that while the number of characters involved in the *Romance* is far larger than that of *Records*, the former covers approximately 71.4% characters of the latter. It indicates the similarity of casts between the two books is very high. In addition, we rank the top 50 most frequently appearing characters in the two books (See Figure 9) and find a 50% coincidence referring to Figure 10. The protagonists (i.e., the top 20 characters) are notably similar; the top 3 most frequently appearing characters in both books are Liu Bei, Cao Cao, and Zhuge Liang.

4.1.3 Complex Network Features of the Story. Previous studies indicate that novelistic literature usually involves a social network that is more complex and thus the literariness and the dramaticism can be greatly enriched [15, 38]. In this paper, two main topological features of the complex networks are considered, and related investigations are conducted on the two extracted social networks.

Small-world. Small-world is a complex network feature that describes a random network with a highly clustered structure. In a small-world network, most nodes are not neighbors of each other, yet the neighbors of some random nodes are probably going to be neighbors of one another, and most nodes can be reached from each other node by few jumps or steps. We can find out more homogeneity in the social structure when its social network possesses such a small-world feature. To measure the small-world feature in our extracted social networks, we introduce the two key metrics, namely, average clustering coefficient and average path length. Small-world networks are usually recognized as having large average path value length and low average clustering coefficient value. Moreover, an advanced metric, Small-World Index (SWI) [25], is introduced. SWI is capable of quantifying the small-world feature, which can provide



(a) Degree distribution of the network in *Romance*.

(b) Degree distribution of the network in *Records*.

Fig. 11. Degree distribution.

a more straightforward recognition. The calculation of SWI is

$$\text{SWI} = \frac{(L - L_l)(C - C_r)}{(L_r - L_l)(C_l - C_r)}, \tag{1}$$

where C and L are the clustering coefficient and average path length respectively, which are derived from the observed network (note that we compute them by Gephi in this work); C_l and L_l refer to the clustering coefficient and mean path length in a lattice reference network characterized by a high C and L; Similarly, C_r and L_r refer to the clustering coefficient and mean path length in a random reference graph characterized by a low C and L.

Table 6. Average clustering coefficient, average path length, and small-world index of the two networks.

	Avg. Clustering coefficient	Avg. Path length	SWI
Romance	0.326	2.977	0.8624
Records	0.647	2.127	1.56247

From the results shown in Table 6, we can observe that the *Records* has a significantly higher average path value and smaller average clustering coefficient compared to those of *Romance*. Especially, the results of the calculated SWI indicate that the SWI of Records (1.562) is higher than that of Romance (0.862), thereby quantifiably confirming our assumption. Literature that focuses on a single character or a group of characters presents a higher SWI than those focused on a mass of characters. *Romance* focuses on a few protagonists, features a much higher SWI than the *Records*, where the story follows several protagonists. It implies that *Romance* focuses more on storytelling around several characters rather than epic depiction.

Scale-free. Scale-free describes a network whose degree distribution follows a power law. It reveals the Pareto principle that 20% of individuals commonly hold 80% of the total resources in a society, a.k.a., “the rich get richer” [11]. To investigate the scale-free feature of the two networks, we demonstrate the degree distribution of nodes in them, as shown in Figure 11, where x is the degree of a node and y is the number of nodes which possess this degree.

From the results, we can observe a salient power-law distribution in the diagrams of both networks. The satisfaction of the power-law indicates that networks of the two books are both scale-free. However, the distribution of *Romance* has a more significant coefficient of determination (R^2: 0.8162 > 0.7749) than that of Records, which means that Romance is relatively more in line with this law.

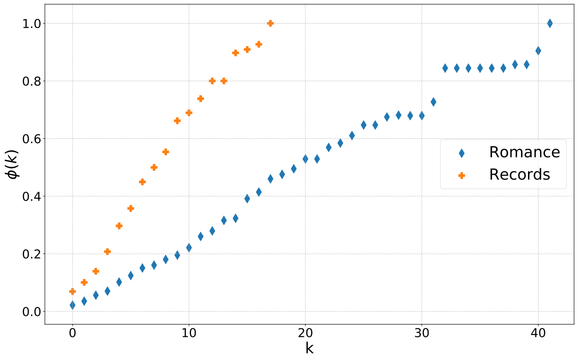


Fig. 12. Rich-club coefficients of the two networks.

Rich-club Coefficient. A number of scale-free networks exhibit a âĂIJrich-clubâĂİ feature, indicating that a small number of nodes possessing a large number of edges also connect well to one another [1, 42]. The Rich-Club Coefficient is used to measure this feature, which can be computed by

$$\phi(k) = \frac{2\mathcal{E}_{>k}}{\mathcal{N}_{>k}(\mathcal{N}_{>k} - 1)} \tag{2}$$

where $\mathcal{N}_{>k}$ is the number of nodes whose degree is not less than k, and $\mathcal{E}_{>k}$ is the actual number of edges among the nodes whose degree are not less than k; $\phi(k)$ is the ratio between the number of edges that exist among the nodes that have a degree larger than k and the total possible number among them. Considering the different sizes of the two networks, we compare the ratio of nodes that can form a fully connected network ($\phi(k) = 100\%$) deduced by the cut-off degree observed, which can be calculated by

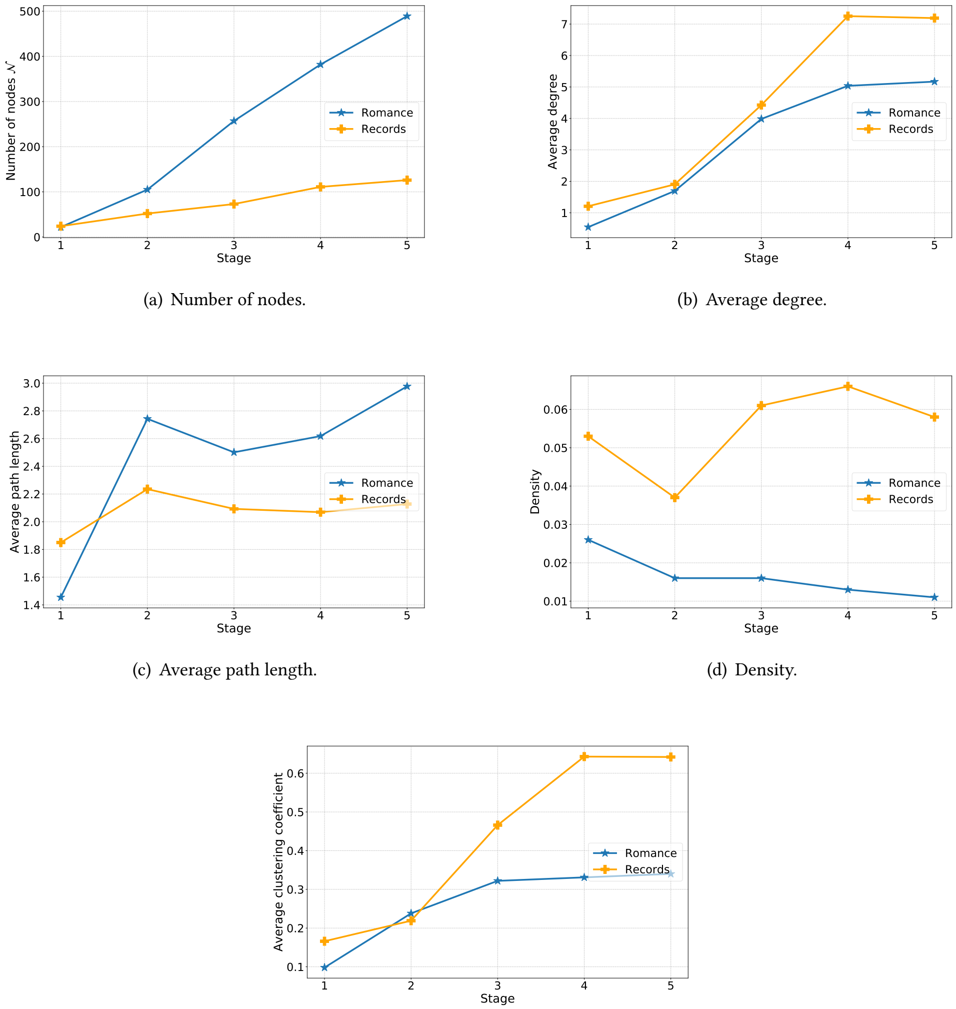
$$r_{fc} = \frac{k_{\phi(k)=100\%}}{\mathcal{N}}, \tag{3}$$

where $k_{\phi(k)=100\%}$ is the minimum k which make $\phi(k) = 100\%$, and N is the number of nodes in the network.

The calculated r_{fc} are 5.09% (*Romance*) and 20.31% (*Records*), respectively. It reveals that the top 5% rich nodes in *Romance* can approximately form a fully connected network, whereas the number has to be approximately the top 20% in *Records*. It can be concluded that both networks have a rich-club feature, which is more significant in *Romance*. These results reveal that despite more characters appearing in *Romance* than in *Records*, the story always revolves around a few protagonists in *Romance*.

4.1.4 History or Romance: Which is More Dramatic? Dramatic changes make stories splendid. The rise and fall of warlords constantly change the social structure of the story of the Three Kingdoms across all stages. To study the growth of the social structure, we investigate the social network according to the idea introduced in Section 3.4.2. As shown in Figure 13, five metrics are adopted to observe the dynamic change of the networks.

Interesting phenomena are found in the results as below. The average node growth rate in *Romance* and *Records* are 147% and 63%, respectively. This suggests that *Romance* has dramatic changes in terms of the number of characters, and the appearance of characters on each stage is overwhelming. The density comparison indicates that *Romance* has a larger network size through all the stages, yet its density is lower, where the gap in the last three stages is especially notable. The change of the average degree of two networks follows a similar pattern, demonstrating that they both increase at the beginning and then reach a plateau. *Records* has a considerably larger average path length



(a) Number of nodes.

(b) Average degree.

(c) Average path length.

(d) Density.

(e) Average clustering coefficient.

Fig. 13. Dynamic growth of the two networks (5 stages).

and lower average clustering coefficient in the majority of five stages except in the first two, which match its better performance in small-worldliness. Overall, we can observe that the growth of the social network in *Romance* is more rapid. Comparatively, Records has a tight, and gradually clustered network.

4.2 Network Feature on Specific Characters

In this subsection, we assess the network feature on specific characters. While the story of the Three Kingdoms involves numerous forces, the main focus is the three force blocs, i.e., Wei, Shu, and Wu. Therefore, their respective sovereigns, namely, Cao Cao (Wei), Liu Bei (Shu), and Sun Quan (Wu), are chosen as the targets for our character-centric investigation.

4.2.1 Who is the most influential? In the story of the Three Kingdoms, the personal influence of each sovereign considerably represents the influence of the forces they possess. Given this kind of influence in a social network, the sovereignsâĂŹ interactions with other characters reflect their influence. Three related metrics are introduced to compare their influence:

- **Degree:** Degree or degree centrality is a basic measure that counts the number of neighbors that a node (character) has. The weighted degree is additionally considered, which is calculated by considering the number of interactions that occur between two characters.
- **Closeness centrality:** Closeness centrality measures the extent of closeness of a node to a network. It is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Its formula is expressed as

$$C(i) = \frac{1}{\sum_j d(i,j)}, \tag{4}$$

  where $C(i)$ is the closeness centrality of node i, and $d(i,j)$ denotes the distance between node i and node j.
- **Betweenness centrality:** For each pair of nodes in a network, at least one shortest path exists between nodes, wherein either the number of edges that the path passes through (for unweighted networks) or the sum of the weights of the edges (for weighted networks) is minimized. Betweenness centrality is a measure of the number of the shortest path that passes through a node. Denoted by $g(v)$, the betweenness centrality of node v can be calculated by

$$g(v) = \sum_{i \neq v \neq j} \frac{\sigma_{ij}(v)}{\sigma_{ij}}, \tag{5}$$

  where σ_{ij} is total number of shortest paths from node i to node j, and $\sigma_{ij}(v)$ is the number of those paths that pass through node v. A characterâĂŹs property of "bridge" can be measured by betweenness centrality.

Table 7 presents the results of *Romance* and Table 8 shows the results of *Records*. In *Romance*, Cao Cao exhibits the highest measures of the four metrics, followed by Liu Bei, whereas Sun Quan has the lowest measures. In *Records*, Liu Bei leads the performance instead of Cao Cao, and Sun Quan is far behind them. This can support us to conclude that Cao Cao is the most influential of the three sovereigns in *Romance*, and Liu Bei is the one in *Records*, and the influence of Sun Quan is lower than the other two lords in both books.

Table 7. The degree and centrality of Cao Cao, Liu Bei, and Sun Quan in the network of *Romance*.

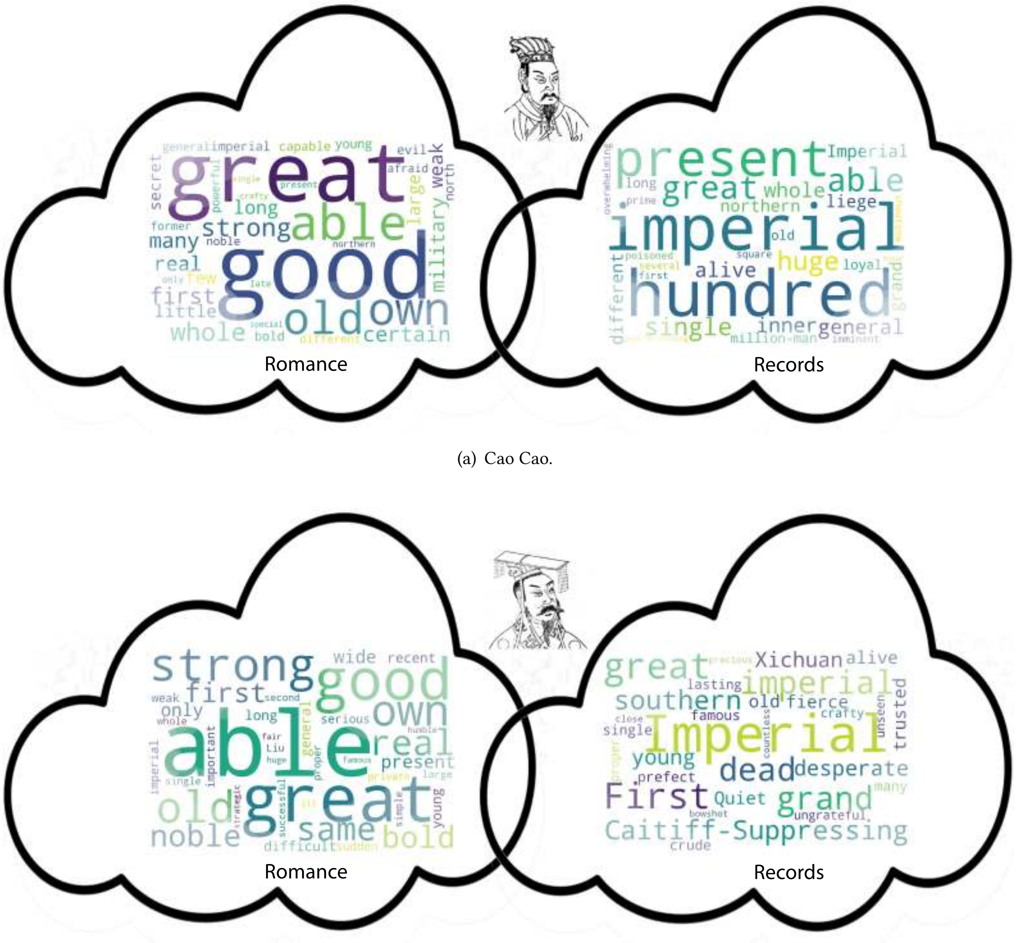
	Degree	Weighted degree	Closeness centrality	Betweenness centrality
Cao Cao	268	3,442	0.565371	18,765.21
Liu Bei	178	2,270	0.506329	7,891.51
Sun Quan	80	908	0.448808	1,688.14

4.3 Sentiment Analysis on Characters

Generally, historical records tend to lean toward objectivity, whereas fictional novels contain subjective emotions. The creation of Romance began approximately toward the end of the Yuan Dynasty, which was a dark era for common

Table 8. The degree and centrality of Cao Cao, Liu Bei, and Sun Quan in the network of *Records*.

	Degree	Weighted degree	Closeness centrality	Betweenness centrality
Liu Bei	164	2,753	0.778523	4,731.94
Cao Cao	134	1,920	0.707317	2,764.84
Sun Quan	24	326	0.527273	20.007937



(a) Cao Cao.

(b) Liu Bei.

Fig. 14. Word clouds of the evaluative words on Cao Cao and Liu Bei.

people. The dissatisfaction with the ruling class can be reflected by the impressionable attitude of people to some forces (e.g., Shu) in the story of the Three Kingdoms. In this context, the author of *Romance* emotionally depicted a series of characters who are different in actual history. This phenomenon substantially occurs to Liu Bei and Cao Cao, which are

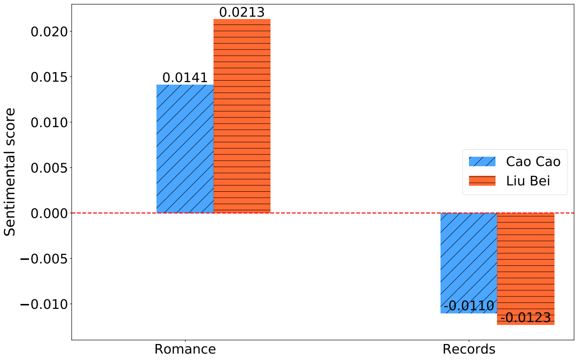


Fig. 15. SentiWordNet scores of Cao Cao and Liu Bei.

lords of Shu and Wei, respectively. Our investigation focuses on these two characters from the point of their evaluating words.

4.3.1 A "Taste" of the Character Sentiment. We commence by collecting and ranking the evaluative words to Liu Bei and Cao Cao. Specifically, we present the results by adopting the word cloud, as shown in Figure 14. As the word cloud shown in Figure 14, A sketchy sentimental opinion on Cao Cao and Liu Bei can be obtained. For example, in *Romance*, evaluative words such as âĂIJgreatâĂİ and âĂIJableâĂİ are mentioned for both two lords. However, we in addition find words such as âĂIJcraftyâĂİ and âĂIJevilâĂİ on Cao Cao and âĂIJhumbleâĂİ on Liu Bei, which reveals the difference. Moreover, more negative words are obviously found about Cao Cao in *Romance* than in *Records*. While this observation cannot bring us to the conclusion that the authors of the two books have an evident preference to a character, we can at least find the there exists differences regarding the depiction of the same character in the two books.

4.3.2 Sentimental Quantification. For a better understanding of the evaluative words to , we conduct a quantitative comparison . Particularly, we introduce a sentimental scoring metric, SentiWordNet [2]. SentiWordNet score can be calculated by subtracting both polarities (positive and negative) of each token and subsequently calculating them:

$$score = \frac{\sum_{i=1}^{n}(posScore_i - negScore_i)}{n} \tag{6}$$

where n denotes the number of involved evaluative words, $posScore_i$ and $negScore_i$ are the positive and negative scores of word i provided by SentiWordNet. The criterion of SentiWordNet gives Negative (i.e., ËŮ1), Neutral (i.e., 0) and Positive (i.e., 1) for users to classify the word.

Figure 15 implies that Cao CaoâĂŹs score is lower than that of Liu Bei in *Romance*. Nonetheless, the score of Cao Cao is higher in *Records*. This finding is consistent with the subjective perception obtained in Section 4.3.1. In addition, the scores of the two characters are both higher in *Romance* than in *Records*. This possibly reveals the different sentimental tones of the authorsâĂŹ wording in the narrative.

5 CONCLUSION

Surrounding on the Story of Three Kingdoms, this paper revives the research on digital humanities, which seeks to digitize working procedures of sociologists and historians in the field of humanities by using state-of-the-art data science technologies.

An algorithm is developed to extract social networks of stories narrated in two books w.r.t. the Three Kingdoms. Particularly, the advanced NLP model BERT is employed in our character identification work, and a satisfying outcome is obtained. Subsequently, we conduct a series of topological analysis to quantify and characterize the extracted social networks, where we additionally present a quantitative comparison between the two books. Specifically, network topological features, such as small-world, scale-free, and centrality of specific characters, are measured. The results reveal that the social network is more entangled in the narrative of the *Romance* than that of the *Records*, especially, more protagonist-oriented. Moreover, this provides a quantitative reference for the macro (e.g., structural features of a story) and micro levels (e.g., the influence or sentiment of a specific character), and the extent of the grandness vividness of a story can be expressed scientifically.

This work can help both researchers and non-expert readers gain an insight into the story of the Three Kingdoms and the procedure of its digital analysis. Moreover, numerous involved sub-works can be refined in the future. First, the definition of interactions between characters is coarse-grained. Second, a mere five-slice dynamic network is built in this project, and hopefully, a large-scale dynamic network, which can incorporate hundreds even thousands of slices, can be obtained if the story is subdivided in fine granularity, for instance, year-to-year or day-to-day.

6 ACKNOWLEDGMENTS

REFERENCES

[1] Jeff Alstott, Pietro Panzarasa, Mikail Rubinov, Edward T Bullmore, and Petra E Vértes. 2014. A unifying framework for measuring weighted rich clubs. *Scientific reports* 4 (2014), 7258.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.. In *Lrec*, Vol. 10. 2200–2204.

[3] Rahul C Basole and Marcus A Bellamy. 2014. Visual analysis of supply network risks: Insights from the electronics industry. *Decision Support Systems* 67 (2014), 109–120.

[4] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. 2009. Gephi: an open source software for exploring and manipulating networks. *Icwsm* 8, 2009 (2009), 361–362.

[5] Marcus A Bellamy and Rahul C Basole. 2013. Network analysis of supply chain systems: A systematic review and future research. *Systems Engineering* 16, 2 (2013), 235–249.

[6] Francis Bloch, Matthew O Jackson, and Pietro Tebaldi. 2019. Centrality measures in networks. *Available at SSRN 2749124* (2019).

[7] Oriol Borrega, Mariona Taulé, and M AntøâĂŹnia Martı. 2007. What do we mean when we speak about Named Entities. In *Proceedings of Corpus Linguistics*.

[8] Charles Henry Brewitt-Taylor and Timothy Richard. 1931. *Romance of the Three Kingdom.*

[9] Ronald S Burt. 2009. *Structural holes: The social structure of competition.* Harvard university press.

[10] Alessio Cardillo, Jesús Gómez-Gardenes, Massimiliano Zanin, Miguel Romance, David Papo, Francisco Del Pozo, and Stefano Boccaletti. 2013. Emergence of network features from multiplexity. *Scientific reports* 3, 1 (2013), 1–6.

[11] Anirban Chakraborti and Marco Patriarca. 2009. Variational principle for the Pareto power law. *Physical review letters* 103, 22 (2009), 228701.

[12] Tapan Chowdhury, Samya Muhuri, Susanta Chakraborty, and Sabitri Nanda Chakraborty. 2019. Analysis of adapted films and stories based on social network. *IEEE Transactions on Computational Social Systems* 6, 5 (2019), 858–869.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Shoshana R Dobrow, Dawn E Chandler, Wendy M Murphy, and Kathy E Kram. 2012. A review of developmental networks: Incorporating a mutuality perspective. *Journal of Management* 38, 1 (2012), 210–242.

[15] David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. 138–147.

[16] Francis J Flynn, Ray E Reagans, and Lucia Guillory. 2010. Do you two know each other? Transitivity, homophily, and the need for (network) closure. *Journal of personality and social psychology* 99, 5 (2010), 855.

[17] Roberto Franzosi. 2010. *Quantitative narrative analysis.* Number 162. Sage.

[18] Nipendra Kayastha, Dusit Niyato, Ping Wang, and Ekram Hossain. 2011. Applications, architectures, and protocol design issues for mobile social networks: A survey. *Proc. IEEE* 99, 12 (2011), 2130–2158.

[19] Matthew G Kirschenbaum. 2016. What is digital humanities and whatâĂŹs it doing in English departments? In *Defining Digital Humanities*. Routledge, 211–220.

[20] Lluís Màrquez and Horacio Rodríguez. 1998. Part-of-speech tagging using decision trees. In *European Conference on Machine Learning*. Springer, 25–36.

[21] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces* 35, 5 (2013), 482–489.

[22] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[24] Michel Mitri. 2020. Story Analysis Using Natural Language Processing and Interactive Dashboards. *Journal of Computer Information Systems* (2020), 1–11.

[25] Zachary P Neal. 2017. How small is it? Comparing indices of small worldliness. *Network Science* 5, 1 (2017), 30–44.

[26] Evelien Otte and Ronald Rousseau. 2002. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science* 28, 6 (2002), 441–453.

[27] Partha Pakray, Arunagshu Pal, Goutam Majumder, and Alexander Gelbukh. 2015. Resource Building and Parts-of-Speech (POS) Tagging for the Mizo Language. *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)* (2015), 3–7.

[28] Irene M Pepperberg. 1999. Rethinking syntax: A commentary on E. KakoâĂŹs âĂIJElements of syntax in the systems of three language-trained animalsâĂİ. *Animal Learning & Behavior* 27, 1 (1999), 15–17.

[29] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[30] Maria Polo, Umberto Dello Iacono, Giuseppe Fiorentino, and Anna Pierri. 2019. A social network analysis approach to a digital interactive storytelling in mathematics. *Journal of e-Learning and Knowledge Society* 15, 3 (2019), 239–250.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI* (2018).

[32] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[33] Eleonora Rosati. 2014. Google Books' Library Project is fair use. *Journal of Intellectual Property Law & Practice* 9, 2 (2014), 104–106.

[34] Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050* (2003).

[35] Janez Strehovec. 2016. *Text as Ride: Electronic Literature and New Media Art*. Center for Literary Computing.

[36] Hai-Long Trieu, Duc-Vu Tran, Ashwin Ittoo, and Le-Minh Nguyen. 2019. Leveraging Additional Resources for Improving Statistical Machine Translation on Asian Low-Resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18, 3 (2019), 1–22.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[38] Michaël C Waumans, Thibaut Nicodème, and Hugues Bersini. 2015. Topology analysis of social networks extracted from literature. *PloS one* 10, 6 (2015), e0126470.

[39] Sarah E Worth. 2008. Storytelling and narrative knowing: An examination of the epistemic benefits of well-told stories. *Journal of Aesthetic Education* 42, 3 (2008), 42–56.

[40] Heiga Zen and Haşim Sak. 2015. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4470–4474.

[41] Jun Zhang, Hai Zhao, Jiu-qiang Xu, and Jin-fa Wang. 2014. Small-world and Scale-free Features in Harry Potter. (2014).

[42] Shi Zhou and Raúl J Mondragón. 2004. The rich-club phenomenon in the Internet topology. *IEEE Communications Letters* 8, 3 (2004), 180–182.