IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 25, 2020, accepted September 10, 2020, date of publication September 15, 2020,
date of current version September 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024210

Identifying Key People in Chinese Literary Works Using e-Core Decomposition

WEIFENG PAN[1], XINXIN XU[1], HUA MING[2], (Member, IEEE), PING GONG[3], BO JIANG[1], CHUNLAI CHAI[1], AND BAILIN YANG[1]

[1]School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
[2]School of Engineering and Computer Science, Oakland University, Rochester, MI 48309, USA
[3]College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350007, China

Corresponding author: Weifeng Pan (wfpan@zjgsu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1400602 and Grant 2016YFB0800400, in part by the National Natural Science Foundation of China under Grant 61572371 and Grant 61702378, in part by the Natural Science Foundation of Fujian Province under Grant 2018J01781, in part by the Key Research and Development Program Project of Zhejiang Province under Grant 2019C01004, and in part by the Commonweal Project of Science and Technology Department of Zhejiang Province under Grant LGF19F020007.

ABSTRACT The plots of many literary works are very complex, which hinders the readers' comprehension of these literary works. Thus, to help readers' comprehension of complex literary works, tools should be proposed to support their comprehension by presenting the most important information to readers. In the case of literary works, the most important information may be the most important people, also called the key people. Key people play an important role in promoting the development of the plot of literary works. Thus, identifying key people helps to simplify readers' comprehension of literary works. The traditional way to comprehend literary works mainly depends on intensive reading, and no previous work has been done to explore the problem of key people identification in literary works. In this paper, we define the concept of key people and propose an approach, IPWC, to Identify key People in Chinese literary Works using *e*-Core decomposition. First, it uses the Weighted People co-occurrence Network (WPN) to represent people and their coupling relationships, and the frequency of the relationship. Second, an *e*-core decomposition is proposed and applied to decompose the WPN into shells and obtain the coreness of each people. Finally, all the people are sorted in a descending order according to their coreness, and the top-ranked people are the key people identified by our approach. The empirical experiments performed on a famous Chinese literary work, *The History of the Three Kingdoms*, show that WPN has *small world* and *scale-free* features. Furthermore, our approach is feasible and more effective than other compared approaches. Our approach can be used to build an automatic tool that can identify the key people for readers to aid their comprehension of a Chinese literary work.

INDEX TERMS Literary work comprehension, social network, culture analysis, key people, e-core decomposition.

I. INTRODUCTION

Literary works reflect the social life of people in a specific time period in a history. Through the comprehension of literary works, readers can roughly understand the politics, economy and culture of the time period described in a literary work [1]. Generally, literary works mainly contain two parts (i.e., people and plots). People usually closely interact with each other to promote the development of the plots [2].

The traditional way to comprehend a specific literary work mainly depends on the intensive reading of the literary work. Indeed, the manner that a reader uses to build up his/her understanding of a specific literary work may vary greatly, but we do realize that building up knowledge of a literary work is a daunting task, especially for large complex literary works. Just think of how difficult it can be to comprehend a literary work containing more than fifty people and hundreds of pages. In fact, the plots of many literary works are very complex; readers may face many challenges in comprehending these complex literary works.

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Hua Chen.

What make the comprehension of literary works so difficult are the increasing complexity of plots and the lack of appropriate tools to support the comprehension process. Generally, a complex literary work usually contains a large number of people, which interact with each other to promote the development of plots. The complexity enclosed in people and their coupling relationships hinders the readers' comprehension of a literary work. Harrington, a leading researcher in Management Science, believes that *if you cannot measure something, you cannot understand it*. Thus, to promote the comprehension of literary works, it is necessary to propose tools to reasonably describe and effectively quantify people and their coupling relationships in literary works.

Literary works usually contain a lot of information. The goal of this work is to build tools to analyze a literary work and filter out unimportant information such that the important information can be presented to the readers. In the case of literary works, the most important information may be the most important people, which can also be called the key people. To the best of our knowledge, this work is the first work on key people identification. In this work, the key people are described as:

> *The key people are described as the people that promote the development of the main plot of a literary work, and have a main contribution to the formation of the social structure of that time period. Usually, the main plot of a literary work is promoted by very few key people. They are tightly coupled with a large number of other people in the literary work to implement their roles in that time period.*

From this definition, we can observe that key people are depicted by the property that they are tightly coupled with other people. Thus, we can build our approach to identify key people by detecting those tight couplings. The specific angle of our approach is to focus on the coupling information gathered from the literary works. Thus, how to gather the coupling relationship between people is the key step of our approach.

Nowadays, a large number of literary works have corresponding digital versions, which provide a great opportunity to analyze literary works using techniques from other fields [5]. It also provides opportunities for us to gather people and their coupling relationships automatically. In recent years, some researchers introduced theories and techniques in the field of network science to analyze literary works. They used a graph model, People Relationship Network (PRN), to represent people and their relationships, with nodes being people and edges being their relationships. Then they introduced the parameters in complex networks to analyze the structural characteristic of PRN [6], and found that the PRN built from literary works has some complex network features such as *small world* and *scale-free* [4], [6]–[12]. Generally, scale-free networks have a significant feature, i.e., nodes have different levels of importance according to their different positions in the network [13], [14]. Key (or important) nodes play an important role in the complex networks, and are of great significance to the control of the transmission of rumors (or diseases), the allocation of resources, and the formulation of market planning [14]. In literary works, key people usually play an important role in promoting the development of the main plot. If readers ignore these key people, then they may have difficulties in understanding the development of the main plot of a literary work. Thus, when understanding the people and their relationships in the literary works, it is a practical way to choose to start the comprehension process with the key people. But how do readers identify key people of a literary work? It is a tough problem especially when it comes to the understanding of a brand new literary work that readers have never read before.

To the best of our knowledge, there is no previous work on the identification of key people in literary work. In view of this, in this paper, we propose an approach, IPWC (**I**dentifying key **P**eople in Chinese literary **W**orks using *e*-**C**ore decomposition), to identify the key people in Chinese literary works. First, it uses the **W**eighted **P**eople co-occurrence **N**etwork (WPN) to represent people and their relationships in Chinese literary works. Second, an *e*-core decomposition technique is proposed and applied to decompose the WPN into layers, and then the coreness of each people in the WPN is obtained. Finally, all the people are sorted in a descending order according to their coreness. The top-ranked people are the key people identified by our approach. To evaluate our approach, IPWC is applied to identify the key people in a famous Chinese literary work, *The History of the Three Kingdoms* [15], and the feasibility and effectiveness of our approach are illustrated when compared with six other related approaches. The empirical results show that the WPN we built in this work has some complex network features such as *small world* and *scale-free*. We also find that our approach is feasible and more effective than other compared approaches. The key people identified by our approach are largely consistent with humans' judgments.

The main contributions of this paper are as follows:

- We propose the problem of key people identification in literary works. We define the concept of key people, and highlight the importance of key class identification in helping readers' comprehension of complex literary works.
- We propose an approach to identify key people in Chinese literary works by a network representation of people and their coupling relationships, and an *e*-core decomposition technique.
- By extensive experiments on a famous Chinese literary work, we demonstrate that our approach is superior to many other approaches in identifying key people.
- We provide a complete replication package via the link https://github.com/wfpan/IPWC_IEEE_Access.

The package includes the WPN network we built from literary works, and the scripts software to compute the coreness of each people.

The rest of the paper is organized as follows. Section II contains the related work on analyzing people and their coupling relationships in the literary works. Section III describes our IPWC approach in detail, mainly including the extraction of people and their coupling relationships, the definition of WPN, and the *e*-core decomposition technique. Section IV presents the empirical validation of our approach on a Chinese literary work, *The History of the Three Kingdoms*. Section V discusses the threats to validity, and Section VI concludes the paper and discusses future work.

II. RALATED WORK

There has been some research work on the analysis of people and their relationships in literary works. People and their relationships are usually represented as networks, and then statistical indicators or complex network parameters are applied to analyze their structural characteristics [6]–[12], and many shared structural properties have been found. Some representative work is described as follows.

Elson *et al.* [6] built a dialogue network of people for sixty English novels, with nodes being people and un-weighted edges being the dialogue relationships between people. They analyzed many structural characteristics of the dialogue network, such as *scale*, *cliques*, and *average degree*, and also revealed the correlation between these characteristics and the novel's background and culture.

Liu and Yin [7] built a people co-occurrence network for four novels of Fitzgerald, with nodes being people and un-weighted edges being the co-occurrence relationship between people in sentences. They analyzed the *degree*, *degree distribution*, *shortest path length*, *network diameter*, *characteristic path length*, *clustering coefficient* and other structural features of the network, and found many complex network properties such as *small world* and *scale-free*.

Prado *et al.* [8] built a multi-layer time-varying network between people in *Alice's Adventures in Wonderland* and *La Chanson de Roland*, with nodes being people and un-weighted edges being the face-to-face interactions between people. They analyzed the *vitality index* and *Freeman index* of the people.

Zhao *et al.* [9] built the People Relationship Network (PRN) in *The Romance of the Three Kingdoms*, with nodes being people and un-weighted edges being the co-occurrence relationships between people in a chapter or in the whole literary work. They analyzed many structural features of the PRN such as *degree distribution*, *centrality*, and *cohesive subgroups*, and found the PRN shares some structural properties as other complex networks such as *small world*, *scale-free*, and *community structures*.

Wang *et al.* [10] applied the co-word analysis to construct a co-occurrence network of people in the literary work, *The Romance of the Three Kingdoms*, and analyzed the relationships between people quantitatively by using clustering analysis, strategic coordinate analysis, and core-periphery analysis, and found some characteristics, e.g., *some people often co-exist*, and *the relationship between people of Shu Kingdom is closer than that of Wei and Wu kingdoms*.

Lin *et al.* [11] built the people relationship network in the literary work, *A Dream of Red Mansion*, with nodes being people and un-weighted edges being the neighboring relationships between people in the whole literary work. They analyzed the community map of the key people such as *the twelve beauties of Jinling* and *Baoyu Jia* in the first eighty chapters and the last forty chapters. The results showed that the community structures of the key people of the two parts are consistent.

Zhang *et al.* [12] built the people relationship network of fourteen martial arts novels of Yong Jin, with the nodes being the people, edges being the co-occurrence relationships of people, and weights on the edges being the number of co-occurrence of the people in a same paragraph. They proposed a new metric, *character intimacy index*, to analyze the love-mode between the people.

The above described work has made great achievements in the analysis of literary works and also laid a foundation for the work proposed in this paper. But the following problems in the existing research work need to be further explored:

- The people relationship networks constructed in the existing work are un-weighted, only considering whether there is a relationship between the people, and ignoring the strength of the relationship, which cannot accurately describe the relationship between people, especially their coupling strength.
- The existing research work mainly focuses on the analysis of the structural characteristics of people relationship networks by applying the theories and technologies in the field of network science, and no previous work has been done on the identification of key people in the literary works.

Thus, the problem to be solved in this paper can be defined as: how to identify the key people in the Chinese literary works by analyzing their digital versions, so as to provide support for comprehension of the literary works.

III. THE PROPOSED APPROACH

The text of Chinese literary works contains people and their relationships, which can be naturally represented by a people relationship network to formally reflect the internal structure of literary works. Such a *network* representation provides us with an opportunity to apply the theories and techniques in network science to identify the key people in Chinese literary works by the analysis and measurement of the structural information enclosed in the network. Fig. 1 shows the framework of IPWC, which mainly consists of four steps (i.e., steps ① to ④), and it also shows the relationship between these steps. The following sections discuss the main steps of IPWC in detail.

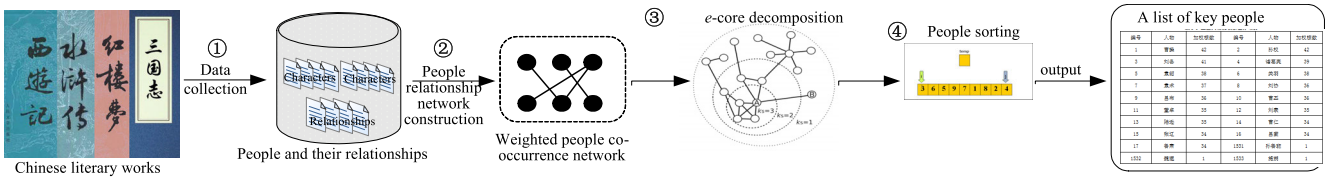


FIGURE 1. The framework of IPWC.

A. DATA COLLECTION

As mentioned above, we take Chinese literary works as the research objects, and apply IPWC to identify the key people. For the convenience for computer processing, in this work, we take the digital version of the subject Chinese work as the input of IPWC. Note that digital version is a corresponding concept with regard to the printed version, which means its content can be easily processed by some software programs. In order to extract people and their relationships, IPWC needs to perform word segmentation, part-of-speech tagging, and anaphora resolution on the digital versions of the corresponding Chinese literary works [9]. These key steps are described as follows:

- Chinese word segmentation [16] divides a sequence of Chinese people into independent words. IPWC treats the Chinese and English periods, question marks, and exclamation marks as the separator of sentences, and performs word segmentation processing sentence-by-sentence. In the results, the words extracted from each sentence occupy one separate line.
- People in literary works are often in the form of nouns, names, and pronouns. Thus, in order to facilitate the extraction of people, IPWC needs to perform part-of-speech tagging on the results obtained by the segmentation [17]. After that, IPWC can identify the people in each sentence by using the nouns (labeled as n), names (labeled as nr), and pronouns (labeled as r) in the text.
- Anaphora is a common phenomenon in Chinese literary works, which often manifests that a people may be denoted in many ways. The emergence of anaphora makes the sentence be concise and coherent and have clear themes. A demonstrative pronoun is often referred to as a referent, which is used to point to an expression. This expression is also called an antecedent. The main task of anaphora resolution is to determine the antecedent object corresponding to the anaphora. In order to identify the people in the text and the relationships between them, IPWC needs to clarify the orientation of the pronouns. Thus, after the word segmentation and part-of-speech tagging, IPWC needs further anaphora resolution of the results [18].

In this work, IPWC applies the Stanford CoreNLP [19] to implement the sentence segmentation, word segmentation, part-of-speech tagging, named entity recognition, and anaphora resolution. Stanford CoreNLP is an open-source software system developed by the Natural Language Processing Group of Stanford University. It is one of the few software systems that open source the anaphora resolution component. After word segmentation, part-of-speech tagging, and anaphora resolution, IPWC can obtain people in the literary works. The relationships between the people are established according to their co-occurrence in the sentences, i.e., if two people appear in at least one same sentence, then it is considered that the two people have a co-occurrence relationship. Note that, although we focus on Chinese literary works, the approach proposed in this work can be extended to English literary works. The main difference is how to divide a sentence into independent words. Interested readers can refer to [19] for details.

B. PEOPLE RELATIONSHIP NETWORK MODEL

Formally abstracting the structural information enclosed in a literary work is a prerequisite for mining the knowledge therein. IPWC uses network models to represent people and their co-occurrence relationships. In this section, we first give the definition of the people relationship network model used in this paper.

Definition 1 (Weighted People co-occurrence Network):

The Weighted People co-occurrence Network (WPN) can be formalized as WPN$= (N, E)$, where N is the set of nodes denoting all the people in a specific Chinese literary work, and $E = \{(c_i, c_j)|c_i \in N, c_j \in N\}$ is the set of undirected edges indicating the people nodes at both ends of the edge have a co-occurrence relationship, i.e., if people c_i and people c_j appear in a same sentence, then there is an undirected edge in WPN between the nodes denoting the two people. Each edge will be assigned with a weight to signify the number of co-occurrence times of the two people nodes in the whole literary work. In fact, the weight is equal to the sum of the co-occurrences of the two people in all sentences in the whole literary work. The connection relationship between people nodes in WPN can be described by the adjacency matrix ψ, with its elements being

$$\psi_{ij} = \begin{cases} n, & \{c_i, c_j\} \in E \\ 0, & others, \end{cases} \tag{1}$$

where ψ is an $|N| \times |N|$ matrix. If $\psi_{ij} = n > 0$, then $\{c_i, c_j\} \in E$, otherwise $\{c_i, c_j\} \notin E$. n represents the weight on the edge $\{c_i, c_j\} \in E$.

Note that WPN and other types of people relationship networks all do not consider the coupling direction between people. It does not mean the direction conveys no meaning. But for the problem of key people identification, we believe the edge direction has less influence. As mentioned in

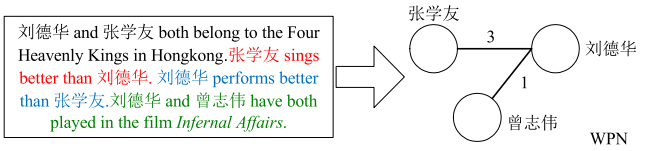


FIGURE 2. Illustration of the process to build WPN from a text.

Definition 1, our WPN only considers the co-occurrence relationships between two people. Co-occurrence relationships only depict the two people (their names) co-existing in at least one same sentence. Thus, such a relationship may not have direction. However, for other types of networks, e.g. people dialogue network, we believe the edge direction cannot be neglected since it depicts the direction of information flow.

Fig. 2 gives a simple example to illustrate the process to build a WPN from a specific text. The left part of Fig. 2 is the text, and the right part is its corresponding WPN. It can be seen from Fig. 2 that the text contains four sentences (identified by different colors), which involve three people (i.e., '刘德华',[1] '张学友',[2] and '曾志伟'[3]). Thus, there are three nodes in the corresponding WPN, being '刘德华', '张学友', and '曾志伟'. Furthermore, '刘德华' and '张学友' co-existed in the first, second, and third sentences. Thus, there is an undirected edge between the nodes denoting the two people, with the weight being 3. '刘德华' and '曾志伟' co-existed only in the fourth sentence. Thus, there is an undirected edge between the nodes denoting the two people, with the weight being 1. '张学友' and '曾志伟' did not co-exist in any sentence. Thus, there is no edge between the nodes denoting the two people.

C. E-CORE DECOMPOSITION

In this paper, WPN is used to represent people and their relationships. It is a weighted network, which is significantly different from the un-weighted people relationship networks widely used in the existing research work. Thus, as an effective metric to quantify the importance of people in a literary work, it should take into consideration such a difference in people relationship networks. Specifically, in this paper, our metric to quantify the importance of people considers the following three properties:

- The different coupling strength between different pairs of people: WPN uses different edge weights to signify the coupling strength among people. However, the un-weighted people relationship networks do not differentiate the coupling strength, and treat all the coupling strength as same. Obviously, the weight will have a great influence on the importance of people, and cannot be ignored in importance computation.
- Different numbers of neighbors that different people nodes have in the WPN: different people nodes may have different numbers of neighbors in the WPN, which will influence the importance computation.
- Different weight distributions between one people and its neighbors: any non-isolated nodes in the WPN have a set of neighbors; different nodes may have different sets of neighbors, with different weight sets on the edges. Such a difference in the weight distributions may also affect the importance of each node in the WPN.

In this work, we propose a new technique, which is called e-core decomposition (e-core for short) to quantify the importance of people in a specific WPN. E-core is an extension of the weighted k-core decomposition (k-core$^{\mathrm{w}}$ for short) [20] and the classic k-core decomposition (k-core for short) [21] for un-weighted networks. Different from k-core$^{\mathrm{w}}$ and k-core, e-core measures the importance of nodes based on the entropy degree of the nodes in a specific weighted network. The entropy degree considers the influence of edge weight, node degree (number of neighbors) and weight distribution on the node importance. In contrast, k-core$^{\mathrm{w}}$ only considers the influence of edge weight and node degree; k-core only considers the influence of node degree. Thus, e-core improves k-core and k-core$^{\mathrm{w}}$. In e-core, the entropy degree of node i, $e(i)$, can be defined as

$$\begin{aligned} e(i) &= \sqrt{k(i)(1-\sum_{x\in n(i)}(p_{ix}\ln p_{ix}))\sum_{y\in n(i)} w_{ix}} \\ p_{ix} &= \frac{w_{ix}}{\sum_{y\in n(i)} w_{iy}} \end{aligned} \tag{2}$$

where $k(i)$ is the degree of node i, i.e., the number of nodes connected to node i, w_{ix} is the weight on the edge between nodes i and x, and $n(i)$ denotes the neighbor set of node i. In formula (2), $\sum_{y\in n(i)} w_{ix}$ is the weighted degree of node i, $-\sum_{x\in n(i)}(p_{ix}\ln p_{ix})$ is the sum of the entropy of the relative weight of the edge between nodes i and x, and $k(i)$ is the degree of node i. Thus, in essence, formula (2) uses $\sum_{y\in n(i)} w_{ix}$ to signify the difference in the coupling strength, uses $-\sum_{x\in n(i)}(p_{ix}\ln p_{ix})$ to signify the difference in the weight distribution, and uses $k(i)$ to signify the difference in the number of neighbors. Thus, our entropy degree considers all the three properties that we listed above.

The entropy degree $e(i)$ returned by formula (2) is usually not an integer. In order to keep the form of $e(i)$ consistent with the degree in the k-core$^{\mathrm{w}}$ and k-core, e-core discretizes $e(i)$ as an integer. Thus, the final value of $e(i)$ is

$$e(i) = \begin{cases} \lfloor e(i) \rfloor, & e(i) - \lfloor e(i) \rfloor < 0.5 \\ \lceil e(i) \rceil, & others, . \end{cases} \tag{3}$$

For the convenience of expression, this paper proposes the theory of e-core by imitating the theory of k-core [21]. Assuming that $G = (N, E)$ is a weighted undirected network,

[1] '刘德华' are Chinese Characters, denoting the Chinese name of Andy Lau, an actor in the Four Heavenly Kings in Hongkong.

[2] '张学友' are Chinese Characters, denoting the Chinese name of Jacky Cheung, an actor in the Four Heavenly Kings in Hongkong.

[3] '曾志伟' are Chinese Characters, denoting the name of a famous actor in Hongkong.

including $|N|$ nodes and $|E|$ undirected weighted edges. The related concepts are defined as follows:

Definition 2 (e-Core of the Weighted Networks):

The e-core of the weighted network refers to the remaining subgraph after repeatedly removing node i with $e(i) \leq e$ and the edge connected to node i. Furthermore, this subgraph is the largest subgraph with such a feature.

Definition 3 (Coreness of the Node in Weighted Networks):

In a weighted network, a node with coreness e indicates that the node exists in the e-core of the weighted network, but does not exist in the $(e+1)$-core.

Definition 4: (e-Shell of the Weighted Networks):

In a weighted network, a subgraph composed of nodes with coreness e and the edges between them is called the e-shell of the weighted network.

e-core applies the pruning process similar to k-core^{w} and k-core to obtain the e-core of the weighted network: recursively remove all nodes whose entropy degree is less than e until the entropy degree of all nodes in the remaining network is at least e. This paper borrows some idea from the efficient k-core decomposition algorithm proposed in [20] and proposes an e-core decomposition algorithm for WPNs (see Algorithm 1). In Algorithm 1, eDegree[n] and eCore[n] represent the entropy degree and coreness of node n, respectively. neighbors(n) returns the neighbors of node n. weight[u] returns the weight of the edge between nodes n and u.

Algorithm 1 e-Core Decomposition Algorithm

Input: WPN$=$ (N, E) represented by neighbor list and edge weight list;
Output: all nodes and their corresponding corenesses.
1. Calculate the entropy degree of all nodes according to formula (3);
2. Sort all nodes in an ascending order according to their entropy degree;
3. for each $n \in N$ do // N is the sorted node set
4. eCore[n]: $=$ eDegree[n];
5. for each u neighbors(n) do
6. if eDegree[u] $>$ eDegree[n] then
7. eDegree[u]: $=$eDegree[u]–weight[u];
8. Re-sort the nodes in an ascending order according to their entropy degree;
9. end
10. end
11.Output all nodes and their corresponding corenesses.

Step 1 can be achieved by traversing the edges with the time complexity being O($|E|$). Step 2 can be achieved by using a quick sort algorithm with the time complexity being O($|N|\log_2|N|$). The key step from steps 3 to 10 is step 8, which can use binary search to find the sorting position of u, and insert it, with the time complexity being O($\log_2|N|$). Thus, the time complexity of step 3 to step 10 is O($|N|\log_2|N|$), and the time complexity of step 11 is O ($|N|$). Therefore, the total time complexity of Algorithm 1 is O($|N|\log_2|N| + |E|$), where $|E|$ is the number of edges in WPN and $|N|$ is the number of nodes in WPN.

Note that the main procedure of Algorithm 1 is similar to the algorithm proposed in [20]. The only difference is, in this work, we compute the entropy degree of each node while [20] computes the traditional node degree.

Fig. 3 illustrates the process of using e-core to divide a simple WPN into the e-core structure. The WPN only contains 3 nodes and 2 edges (see the leftmost part of Fig. 3). First, we calculate the entropy degree of all nodes (see the bottom part of the leftmost part of Fig. 3). Since the entropy degrees of all nodes are $\geq$1, we can directly obtain the 1-core without removing any nodes and edges (see the middle part of Fig. 3). Then, we recalculate the entropy degree of the nodes in the 1-core (see the bottom part of the middle part of Fig. 3). By removing the nodes with entropy degree $<$2 (i.e., '袁术') and the edges between '袁术' and '鲁肃', we can get the 2-core (see the rightmost part of Fig. 3). Next, we recalculate the entropy degree of the remaining nodes in the 2-core (see the bottom part of the rightmost part of Fig. 3). After removing all nodes with entropy degree $<$3 (i.e., '周瑜' and '鲁肃') and the edges between '周瑜' and '鲁肃', there are no nodes and edges left in the 2-core. Thus, e-core terminates. '袁术' exists in 1-core, but is deleted in 2-core, so the coreness of '袁术' is equal to 1. By a similar way, we can obtain that the corenesses of '周瑜' and '鲁肃' are both 2.

D. PEOPLE SORTING

IPWC regards the top-ranked people according to their coreness as the key people. Thus, after obtaining the coreness of each people, IPWC will sort these people in a descending order. However, the coreness of some people may be same. For example, as is shown in Fig. 3, the coreness of '周瑜' and '鲁肃' are 2. How should the IPWC sort the people with a same value of coreness?

IPWC uses *bucket sorting* [22] to sort the people. Specifically, it first groups the people into each *bucket* according to their coreness, i.e., people in the same bucket have a same coreness. Then, for the people in the same bucket, it sorts them in a descending order according to their entropy degree before discretization in the whole WPN. Finally, it outputs the people bucket-by-bucket in a descending order according to the priority of the buckets (priority is the coreness of the people in it) until the number of the outputted people reaches $\lfloor p \times |N| \rfloor$, where $p \in (0, 1]$ is a filtering value, and $|N|$ represents the number of people in the WPN. IPWC uses *bucket sorting* with the aim to avoid the situation where two people cannot be sorted (i.e., two people have a same value of coreness, making the relative importance of these people cannot be distinguished). However, this situation may still exist. Thus, once this *unsortable* situation is encountered, IPWC determines their relative positions randomly.

We use the example shown in Fig. 3 to illustrate the sorting process of IPWC (see Fig. 4). First, '周瑜' and '鲁肃' are grouped into the *bucket* #2 (2 is the coreness of people in this bucket), and '袁术' is grouped into the *bucket* #1 (1 is

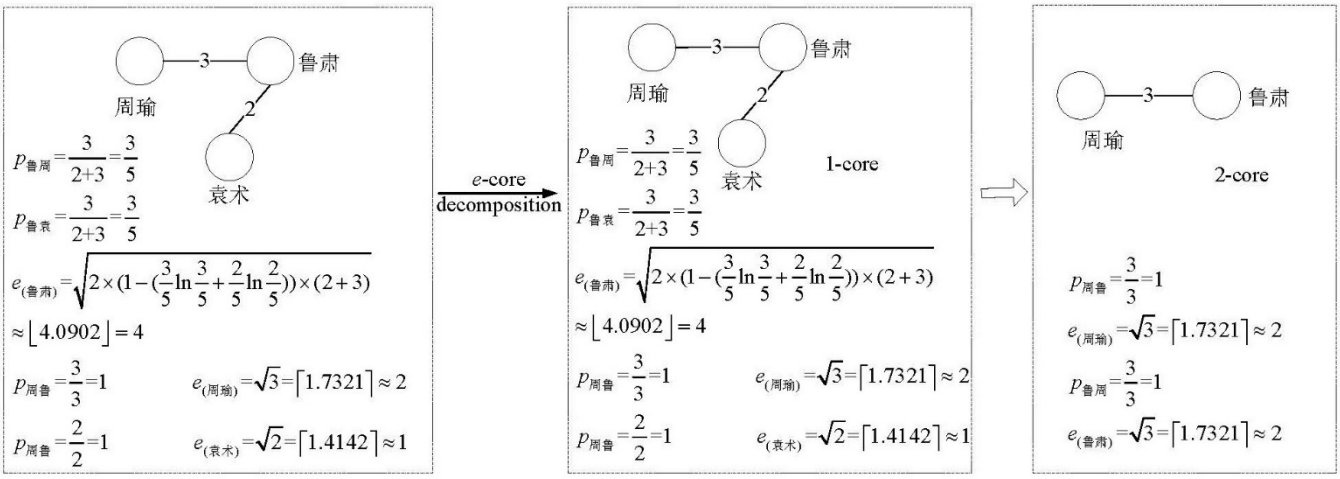


FIGURE 3. Illustration of the *e*-core decomposition when applied to a simple WPN.

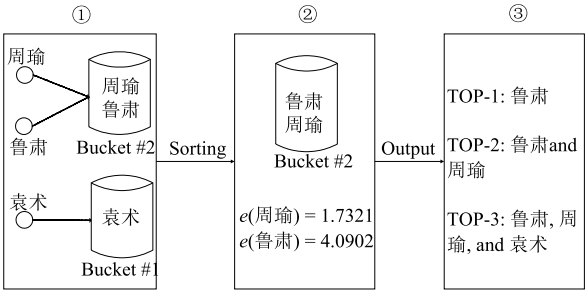


FIGURE 4. Illustration of the process to sort people.

the coreness of people in this bucket). Second, the nodes in *bucket* #2 are sorted in a descending order with regard to their entropy degrees before discretization in the whole WPN (i.e., '周瑜' is 1.7321, and '鲁肃' is 4.0902). Thus, the order is: '鲁肃', '周瑜'. Finally, we output the people in each *bucket* in a descending order according to the priority of the buckets. Thus, if we output the top-1 people, then the result is '鲁肃'. If we output the top-2 people, then the result is '鲁肃' and '周瑜'. If we output the top-3 people, then the result is '鲁肃, '周瑜' and '袁术'.

IV. EMPIRICAL EVALUATION

In order to validate the feasibility and effectiveness of our IPWC approach in identifying key people in Chinese literary works, we performed empirical experiments on a famous Chinese literary work, *The History of the Three Kingdoms* [15].

Our experiments were carried out on a ThinkPad Laptop with a Windows 7 64-bit operating system (Ultimate), Intel (R) Core (TM) i7-5600U@2.6 GHz CPU, and 8G RAM.

In the following subsections, we describe in detail the objects of study, research questions that we focused on, and the results and analysis.

A. OBJECTS OF STUDY

As mentioned in the related work, researchers have carried out some research work on famous Chinese literary works such as *Dream of the Red Mansion*, *Water Margin*, and *Romance of the Three Kingdoms*. They mainly focused on analyzing the structural properties of the people relationship network, and found some features shared by complex networks, e.g., *small world* and *scale-free*. However, to the best of our knowledge, there is no previous research work on identifying key people in Chinese literary works.

In this paper, we used the famous Chinese literary work, *The History of the Three Kingdoms* as our object of study. The rationale is twofold:

- The existing work mainly uses novels as the objects of study. However, novels are usually fictitious and cannot accurately reflect the real life or history of a time period. *The History of the Three Kingdoms*, as orthodox historical factual material, has higher historical credibility. Thus, the study of *The History of the Three Kingdoms* has higher historical research value than the study of fictitious novels.
- We use Chinese literary works as our research subjects rather than English literary work. The main reason is that we cannot recruit enough volunteers to identify the true key people in a specific English literary work. The true key people are indispensable for empirical validation of our approach.
- Most Chinese readers are very familiar with *The History of the Three Kingdoms*. Thus, using *The History of the Three Kingdoms* as our research object lays a basis for the human validation of the results provided by our approach and other approaches we used for comparisons.

Note that though, in this work, we only validated the feasibility and effectiveness of our IPWC approach on one Chinese literary work. In fact, IPWC can be used on any Chinese literary works with digital versions. The main reason why we performed experiments only on one Chinese literary work is that it is a hard and time-consuming task for us to recruit enough volunteers to identify the key people in a

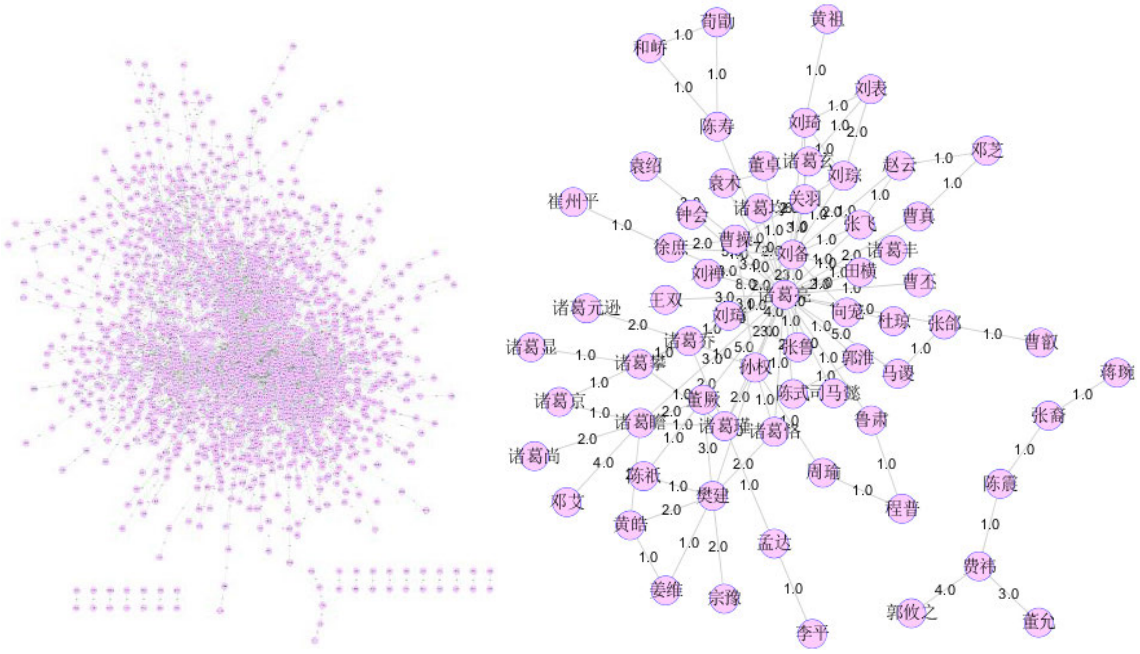


FIGURE 5. The WPNs built by our approach. The left part is built from the whole *The History of the Three Kingdoms*, while the left part is built from the *35th Chapter of The History of the Three Kingdoms, i.e., The book of Shu - Biography of Liang Zhuge.*

specific literary work. Thus, it is sometimes impossible for us to perform experiments on a large data set containing large numbers of literary works.

We follow the main steps shown in Fig. 1 to perform word segmentation, part-of-speech tagging, and anaphora resolution on the digital version of *The History of the Three Kingdoms*, to extract people and their relationships. Our final data set contains 1,529 people and 3,700 edges. Fig. 5 (the left part) shows the WPN that we built from the whole *The History of the Three Kingdoms*. However, the scale of the diagram is too large. We cannot show the details of the WPN clearly. To illustrate the WPN clearly, we built a small WPN for the 35th Chapter of *The History of the Three Kingdoms, i.e., The book of Shu - Biography of Liang Zhuge* (see the right part of Fig. 5).

B. RESEARCH QUESTIONS

We performed experiments to address the following three research questions (RQ):

- RQ1: Does the WPN built from *The History of the Three Kingdoms* have some complex network features such as *small world* and *scale-free*? The existing research work analyzed the structural features of the people relationship network built from the literary works, and found many structural features shared by complex networks such as *small world* and *scale-free*. In this paper, we first check whether the WPN we built from *The History of the Three Kingdoms* also has these structural properties. If WPN is a *scale-free* network, it means that most of the people are connected with only a few people, and there are a small number of people who have contact with a large number of other people and become key people.
- RQ2: Are the key people identified by IPWC meaningful from a historical perspective or the viewpoint of readers? As a feasible approach to identify key people, the key people it identified should be comparable to the historical fact or the results returned by readers through intensive reading of the literary works.
- RQ3: Does IPWC perform better than other similar approaches in identifying key people? There are many centrality metrics in the network science that can be used to measure the importance of nodes in a specific network. We want to know whether the *e*-core decomposition used in IPWC is better than these centrality metrics.

C. RESULTS AND ANALYSIS

In this section, we analyzed the obtained experimental results with the aim to answer the three research questions raised in Section B.

1) RQ1: DOES THE WPN BUILT FROM THE HISTORY OF THE THREE KINGDOMS HAVE SOME COMPLEX NETWORK FEATURES SUCH AS SMALL WORLD AND SCALE-FREE?

(1) Does the WPN built from *The History of the Three Kingdoms* have the *small world* feature?

Generally, networks with small average path length and large clustering coefficient are called *small world* networks [23]. The average path length and clustering coefficient can be defined as follows:

Definition 5 (Average Path Length [24]):

TABLE 1. Statistical parameters of the WPN built from *The History of the Three Kingdoms*.

Network	$\|N\|$	$\|E\|$	$<k>$	L	C	L_{rnd}	C_{rnd}
WPN	1,529	3,700	4.836	3.983	0.253	4.652	0.003

The shortest path d_{ij} between any pair of nodes i and j is the path with the least number of edges among all paths connecting these two nodes. The average path length of a network $L =< d_{ij} >$ is the average value of d_{ij} over all pairs of nodes in the network.

Definition 6 (Clustering Coefficient [24]):

The clustering coefficient C_i of node i reflects the probability that the neighbors of node i to be themselves neighbors. If l_i is the actual number of edges between neighbors of nodes i, and k_i is the total number of neighbors of node i, then C_i can be calculated as $C_i = 2l_i/k_i(k_i-1)$. The clustering coefficient of the whole network is the average of the clustering coefficients over all nodes, i.e., $C =< C_i >$.

We calculated the average path length (L) and clustering coefficient (C) of the WPN built from *The History of the Three Kingdoms*, and compared them with the average path length (L_{rnd}) and clustering coefficient (C_{rnd}) of the corresponding random networks of the same size. Specifically, the average path length and clustering coefficient of the random network are calculated using $L_{rnd} = ln(|N|)/ln(<k>)$ and $C_{rnd} = <k>/|N|$, respectively. The results are shown in Table 1.

By comparing the average path length and clustering coefficient of the WPN built from *The History of the Three Kingdoms* with that of the random network of the same size, we can find that their average shortest path length is very close to each other (i.e., L is very close to L_{rnd}), but the clustering coefficient of WPN is much larger than that of the corresponding random network (C is 84 times larger than C_{rnd}). This shows that the WPN built from *The History of the Three Kingdoms* has the *small world* feature.

(2) Does the WPN built from *The History of the Three Kingdoms* have the *scale-free* feature?

Degree distribution is often used to check whether a network is *scale-free*. If the degree distribution of the network follows the power-law distribution, then the network is a *scale-free* network [25]. The degree distribution reflects the probability that a randomly selected node has a degree k. Mathematically, it can be defined as $P(K = k) \sim k^{-\alpha}$. Power law can also be checked by the cumulative degree distribution, i.e., $P_{cum}(k) = P(K > k) \sim k^{-\beta}$. A network is *scale-free*. It indicates that the degree of nodes in the network has no obvious characteristic length. Most nodes have only a few connections, and a few nodes have a large number of connections and become hubs.

Fig. 6 shows the cumulative degree distribution of the WPN built from *The History of the Three Kingdoms*. It can be seen that the cumulative degree distribution follows a power-law

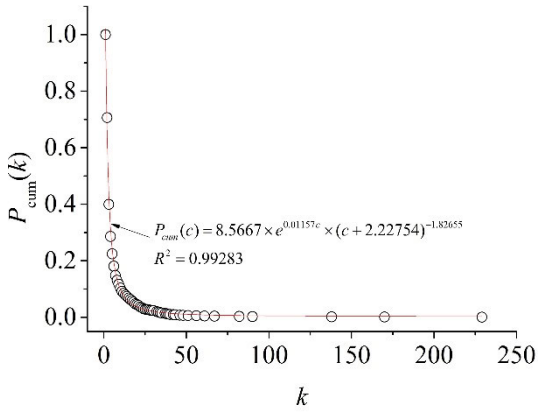


FIGURE 6. Cumulative degree distribution of the WPN built from *The History of the Three Kingdoms*.

distribution with an exponential cutoff, and the measures for goodness of fit (R^2) reaches 0.99283. Thus, the WPN built from *The History of the Three Kingdoms* is a *scale-free* network. It means that most of the people are only connected to a few people, while there are a small number of people who are connected to a large number of other people and become key people. Thus, there are indeed some key people in *The History of the Three Kingdoms*. They promoted the development of the main plots of *The History of the Three Kingdoms*.

Thus, the answer to RQ1 is: the WPN built from *The History of the Three Kingdoms* has *small world* and *scale-free* features, which is consistent with the results of existing work.

2) RQ2: ARE THE KEY PEOPLE IDENTIFIED BY IPWC MEANINGFUL FROM *A* HISTORICAL PERSPECITVE OR THE VIEWPOINT OF READERS?

IPWC applied the *e*-core to decompose the WPN built from *The History of the Three Kingdoms*, so as to obtain the *e*-shells belonging to different levels and the coreness of each people. All the people are sorted according to their corenesses in a descending order, and the top-ranked people are the key people identified by IPWC.

As mentioned above, there is no previous work on the identification of key people in *The History of the Three Kingdoms*. Thus, we cannot obtain a benchmark set of true key people from the existing work. However, as a feasible approach to identify key people, the identified key people should be comparable to the historical fact or the results returned by readers through intensive reading of the literary works.

Table 2 shows the top-6 key people identified by IPWC. In Table 2, the first column shows the ranks of the corresponding people, the second column is the name of the people, the third column is the coreness of the corresponding people, and the fourth column is the entropy degree before discretization obtained from the whole WPN built from *The History of the Three Kingdoms*.

TABLE 2. The top-6 people identified by IPWC.

Ranks	People	Coreness	Entropy degree before discretization
1	曹操[4]	94	1314.24
2	孙权[5]	94	927.20
3	刘备[6]	93	772.48
4	袁绍[7]	78	418.27
5	诸葛亮[8]	78	407.98
6	关羽[9]	77	234.46

It can be seen from Table 2 that IPWC can extract *e*-shells at different levels. The top three people are '曹操', '孙权' and '刘备'. They are household persons in China, and constitute the innermost two cores of the WPN. Shiming Fang, a famous contemporary historian in China, pointed out in his book *People of the Three Kingdoms* that *Three Kingdoms in fact refers to Wei Kingdom founded by '曹操', Shu Kingdom founded by '刘备', and Wu Kingdom founded by '孙权'*. Thus, '曹 操,刘 备' and '孙 权' are the people who established the situation of the three kingdoms and played a decisive role in the development of the history of the three kingdoms. In addition, Zhongtian Yi, a famous scholar in China, also pointed out in his book *Discussion of Three Kingdoms* that *those who can contribute to the reunification of the country and push forward the cause of the unification are the heroes of the times.* The hero ranking 1st is '曹操'. He pacified the north and laid the foundation for the three Kingdoms (i.e., Wei Kingdom, Shu Kingdom, and Wu Kingdom) to return to Jin; '孙权' and '刘备' rank 2nd and 3rd, respectively. '孙权' pacified Jiangdong, and '刘 备' pacified Yizhou. The three persons have made their own contributions to the reunification of the country. Thus, '曹操,孙权' and '刘备' were heroes of that era. It is consistent with the historical facts that IPWC identifies these three persons as the top three people in *The History of the Three Kingdoms*. In addition, it is reasonable that IPWC grouped '曹操' and '孙权' into the same shell (94-shell), and grouped '刘备' into the 93-shell. Shiming Fang believes that '孙权' started his career earlier than '刘备' and existed longer than '刘备'. Therefore, '孙权' had a more far-reaching impact on the history than '刘备'.

'袁绍' and '诸葛亮' constitute the 78-shell. '袁绍' once occupied four states (i.e., Jizhou, Bingzhou, Qingzhou, and Youzhou) of the nine states in the later period of Han Dynasty, and was the most powerful northern hegemon at that time. Although in the battle of Guandu, '袁绍' was defeated by '曹操', from a historical perspective, '袁绍' promoted the development of the Cao-Wei Group and laid a sound foundation for the establishment of the Wei Kingdom. '诸葛亮' is the actual ruler of Shu Kingdom. He has made great achievements in the domestic affairs, foreign affairs, and military affairs. He once captured '孟获'[10] seven times and pacified the southern minority areas. He is one of the most influential persons in the later period of the three kingdoms. Shiming Fang commented in his book, *People of the Three Kingdoms,* that *'刘备' cooperated with '诸葛亮' from Jingzhou where they met with each other, and the history of Shu Kingdom was largely created by '诸葛亮'*. Zhongtian Yi also pointed out in his book, *Discussion of Three Kingdoms,* that *'诸葛亮' was the ranking 1st hero of Shu Kingdom*. Thus, it is in line with the historical facts that IPWC identifies '诸葛亮' as the ranking 4th person in *The History of the Three Kingdoms*.

'关羽' constitutes the 77-shell. '关羽' is a true god of war. Since worshipping with '刘备, he has been loyal to '刘备', and won glory in the field of battle. He helped '刘备' achieve his dominance in the three kingdoms. Thus, from a historical perspective, it is also reasonable that IPWC ranks '关羽' at the 6th position.

Thus, the answer to RQ2 is: IPWC is a feasible approach to identify key people. The key people it identified are comparable to the results returned by readers through intensive reading of the literary works. The key people identified by IPWC are reasonable from a historical perspective or the viewpoint of readers.

3) RQ3: DOES IPWC PERFORM BETTER THAN OTHER SIMILAR APPROACHES IN IDENTIFYING KEY PEOPLE?

As mentioned above, no previous work has been done on the identification of key people in the Chinese literary works. IPWC uses WPN to represent people and their relationships in the literary works, and applies the *e*-core to identify key people in the WPN according to the coreness of each people. In fact, coreness is a metric to quantify the centrality of nodes in a network. As we all know, there are many centrality metrics in the field of complex networks. These centrality metrics can also be used to identify key people in WPN. In this section, we compared *e*-core with other centrality metrics to check if the *e*-core used in IPWC is better than them.

(1) Baseline approaches

e-core will be compared with other centrality metrics in the field of complex networks. Thus, we first give the definition of these centrality metrics.

Definition 7 (Degree Centrality [24]):

Degree centrality is the most direct metric to quantify the centrality of nodes in complex networks. Generally, the larger

[4]'曹操' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

[5]'孙权' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

[6]'刘备' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

[7]'袁绍' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

[8]'诸葛亮' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

[9]'关羽' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

[10]'孟获' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms.*

the degree of a node is, the higher its degree centrality will be. Thus, degree centrality of a node reflects its importance in a specific network. The degree centrality of node k, $C_D(k)$, is defined as

$$C_D(k) = \deg(k), \tag{4}$$

Definition 8 (Coreness [26]):

The k-core of an un-weighted network refers to the remaining subgraph after repeatedly removing node i with degree $\leq k$ and the edges connected to node i. This subgraph is the largest one with such a feature. If node i exists in the k-core of the un-weighted network, but does not exist in the $(k+1)$-core of the un-weighted network, then the coreness of node i is k. The biggest difference between the coreness of un-weighted networks and the coreness of the weighted network is the former uses the traditional degree of nodes while the latter uses the weighted degree of nodes.

Definition 9 (Betweenness Centrality [24]):

The betweenness centrality of node v, $C_B(v)$, is defined as the ratio of the number of shortest paths through node v to the number of shortest paths in the whole network. Betweenness centrality has been widely used in many different complex networks such as biological networks, social networks, and transportation networks. Formally, $C_B(v)$ can be defined as

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{5}$$

where σ_{st} denotes the number of shortest paths from nodes s to t, and $\sigma_{st}(v)$ denotes the number of shortest paths from nodes s to t that node v lies on.

Definition 10 (Closeness Centrality [26]):

The closeness centrality of a node quantifies how close a node is to all other nodes in a specific network. The closer a node is to the center, the closer it is to other nodes. Generally, the closeness centrality of node i, $C_c(i)$, can be defined as

$$C_c(i) = \frac{n-1}{\sum_{i \neq j} d_{ij}}, \tag{6}$$

Definition 11 (PageRank Value [26]):

The PageRank algorithm evaluates the importance of a web page by assigning a *PR* value (PageRank value) to each page. In an un-weighted and un-directed network, the *PR* value of page i, $PR(i)$, can be defined as

$$PR(i) = \frac{1-d}{n} + d \sum_{j \in s(i)} \frac{PR(j)}{|s(j)|}, \tag{7}$$

where n represents the number of nodes in the network, d is the damping factor (the default value is 0.85), $s(i)$ is the neighbors of web page i, and $|s(j)|$ returns the number of elements in $s(j)$. When applying the PageRank algorithm to WPN built from *The History of the Three Kingdoms*, we treated the people as web pages. Through the iterative calculation of formula (7), we can finally obtain the *PR* value for each people. In our previous work, we successfully applied PageRank algorithm to identify the important classes and packages in a software system for software comprehension [27].

		Actual label	
		Key people	Non-key people
Predicted label	Key people	TP	FP
	Non-key people	FN	TN

FIGURE 7. The confusion matrix for key people identification.

(2) Metrics

Key people identification in its nature is a classification problem: the people in a specific literary work are classified into *key* or *non-key* people. Thus, in this paper, we introduced three metrics widely used in classification problems, i.e., *precision*, *recall*, and *F1*. Furthermore, we also introduced the *average ranking score* [28] widely used in the recommendation systems to evaluate the accuracy of each approaches in identifying the key people in the WPN.

For the key people identification, its confusion matrix is shown in Fig. 7, where

- *TP* refers to the number of true key people that are also identified as key people.
- *FP* refers to the number of true non-key people that are identified as key people.
- *TN* refers to the number of true non-key people that are also identified as non-key people.
- *FN* refers to the number of true key people that are identified as non-key people.

Based on the above confusion matrix, *precision*, *recall*, and *F1* can be defined as

$$precision = \frac{TP}{TP + FP}, \tag{8}$$

$$recall = \frac{TP}{TP + FN}, \tag{9}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}, \tag{10}$$

For a particular approach m, the *average ranking score* of the key people c, RS_m^c, can be defined as

$$RS_m^c = \frac{l_m^c}{n}, \tag{11}$$

where n denotes the total number of people in the WPN built from *The History of the Three Kingdoms*, and l_m^c is the ranking position of the key people c in the people list returned by approach m. The average ranking score of the approach m is defined as the average ranking score of all key people. Formally, the *average ranking score* of approach m, RS_m, can be defined as

$$RS_m = \frac{1}{|KS|} \sum_{c \in KS} RS_m^c, \tag{12}$$

where KS is the set of key people identified by the approach m, and $|KS|$ is the number of key people. RS_m is actually the

arithmetic average of the ranking scores of all key people. For a specific approach, the smaller the average ranking score is, the more likely the approach is to rank the key people in the front of people list; otherwise, it indicates that the approach ranks the key people at the back of people list. Thus, the smaller the average ranking score is, the better the approach is in identifying key people.

(3) Results and analysis

This paper focuses on the evaluation of each approach in identifying the top-10 key people in the Chinese literary work, *The History of the Three Kingdoms*. However, as mentioned above, no previous work has been done on the identification of the key people in *The History of the Three Kingdoms*. Thus, there is no benchmark set of true key people in *The History of the Three Kingdoms*. Thus, in order to perform the comparative study, it is necessary to identify the top-10 true key people in *The History of the Three Kingdoms*.

We used a questionnaire survey to collect the top-10 true key people. In the questionnaire, each participant needs to answer two questions: 1) Are you familiar with *The History of the Three Kingdoms*? 2) Please list the ten most influential people in *The History of the Three Kingdoms* in a descending order (i.e., from the most important ones to the least important ones). For the first question, the participant needs to choose an answer from *yes* or *no*. For the second question, the participant needs to write the names of the ten most important people. Questionnaires were sent via email, and the answers were returned via email. Participants are students from the School of Humanities of Zhejiang Gongshang University (ZJGSU), borrowers of the three kingdoms-related books in the library of ZJGSU, and authors who published papers on three kingdoms between 2010 and 2018 in the CNKI database. There were a total of 1500 participants. We collected their emails from the educational administration system of ZJGSU, the library lending system of ZJGSU, and the author information of the papers. The questionnaire survey was conducted from August 1, 2018 to December 31, 2018. We sent out total 1500 emails, and 735 responses were received (49% response rate). After excluding 313 responses (the participants chose *no* for the first question), we finally obtained 422 valid responses. We calculated the frequency of all people in the responses manually, and finally obtained the top-10 people with the largest frequencies in the responses. Note that, in the valid responses, the ranking positions of the key people are not the same. Thus, in the experiments, we only considered the emergence of the key people and ignored their relative ranking positions. The top-10 key people are '曹操', '孙权', '刘备', '诸葛亮', '司马懿', '周瑜', '孙策', '陆逊', '袁绍', and '贾栩'.

For each people in the WPN, all approaches will return a value to signify its importance. Thus, all the people in the WPN can be ranked in a descending order according to the returned importance by a specific approach. After we obtain such a ranked list of people, we can use a fixed threshold x to organize all the people into two different groups, i.e., the top-ranked x people in the list are key people, and the bottom (T-x) people are non-key people. Here we use T to denote the total number of people in a specific literary work. In the case of *The History of the Three Kingdoms*, as mentioned above, $T = 1{,}529$. In practice, x can vary with the requirements of readers according to their time pressure. To evaluate the feasibility and effectiveness of different approaches in identifying key people, we vary x from 5 to 50 with a step of 5, and compute the *precision*, *recall*, *F1* and RS_{m}. Then we use these values to conduct a comprehensive comparison between different approaches. Generally, if an approach has a larger value of *precision*, *recall*, and *F1* under the same threshold, then it is better in identifying key people in the literary works. If two approaches have a same value of *precision*, *recall*, and *F1* under the same threshold, then the approach with a smaller value of RS_{m} is better.

TABLE 3. Average ranking of the seven approaches.

Approaches	*precision*	*recall*	*F1*
degree	3.50	3.50	3.50
coreness$^{\mathrm{u}}$	5.65	5.65	5.65
Betweenness	4.60	4.60	4.60
closeness	4.60	4.60	4.60
PR	3.95	3.95	3.95
coreness$^{\mathrm{w}}$	3.15	3.15	3.15
coreness$^{\mathrm{e}}$	2.55	2.55	2.55

Fig. 8 shows the *precision*, *recall*, and *F1* of different approaches under different thresholds, where coreness$^{\mathrm{u}}$, coreness$^{\mathrm{w}}$ and coreness$^{\mathrm{e}}$ denote the coreness obtained by k-core decomposition, weighted k-core decomposition and e-core decomposition, respectively. From Fig. 8, we can observe that coreness$^{\mathrm{e}}$ is superior to six other approaches in most of the threshold settings, only with several exceptions. Specifically, when threshold $x = 15$, betweenness seems better than coreness$^{\mathrm{e}}$, and when threshold $x = 35$, degree and coreness$^{\mathrm{w}}$ seem better than coreness$^{\mathrm{e}}$. But in the top-50 people, only coreness$^{\mathrm{e}}$ can find all the top-10 people. From this perspective, our coreness$^{\mathrm{e}}$ performs best.

As mentioned above, our approach does not perform best under all the threshold settings. In order to evaluate the overall performance of our approach in the whole set of threshold settings, we conducted the average ranking of the Friedman test [29]. Table 3 shows the average ranking (the smaller, the better) of the seven approaches on the 10 different threshold settings with regard to *precision*, *recall*, and *F1*, respectively. As shown in Table 3, the seven approaches can be roughly sorted by average ranking into the following order: coreness$^{\mathrm{e}}$, coreness$^{\mathrm{w}}$, degree, *PR*, and betweenness (or closeness). It means in the whole set of threshold settings, coreness$^{\mathrm{e}}$ performs best, and betweenness (or closeness) performs worst.

Table 4 lists the true key people (the leftmost column), the ranking positions of the key people returned by each

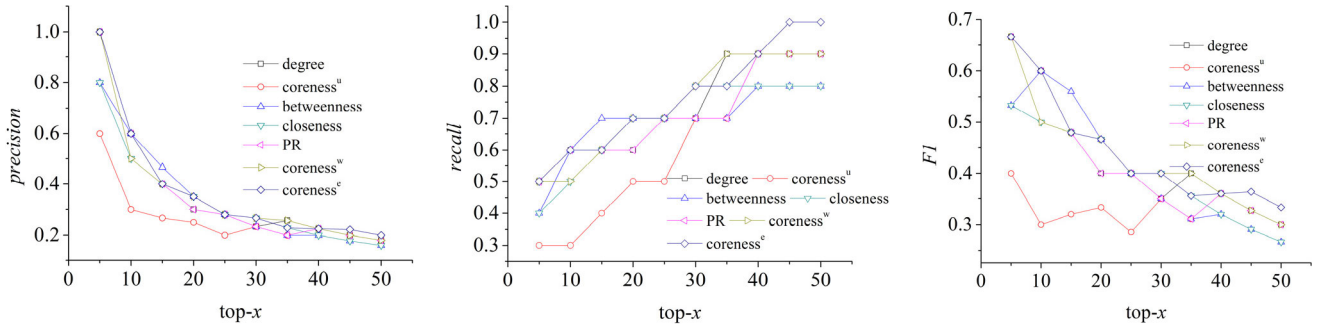


FIGURE 8. Comparison of *precision*, *recall*, and *F1* of different approaches under different thresholds (*x* denotes the threshold).

TABLE 4. The Ranking positions of key people returned by each approach.

	degree	coreness^u	betweenness	closeness	*PR*	coreness^w	coreness^e
曹操	1	1	1	1	1	1	1
孙权	2	12	2	2	2	2	2
刘备	3	2	3	3	3	3	3
诸葛亮	4	16	4	4	4	4	5
司马懿[11]	34	28	36	66	40	52	45
周瑜[12]	35	29	57	18	37	20	19
孙策[13]	23	39	14	13	24	28	26
陆逊[14]	9	37	7	7	7	11	10
袁绍	5	3	8	29	5	5	4
贾栩[15]	58	65	53	70	54	31	37
RS_m	174/1529	232/1529	185/1529	213/1529	177/1529	157/1529	152/1529

approach, and the RS_m of each approach. It can be seen from Table 4 that these approaches can be roughly sorted by the average ranking score into the following order: coreness^e, coreness^w, degree, *PR*, betweenness, closeness, coreness^u. RS_m of the coreness^e is smallest, indicating that IPWC performs better than other approaches in identifying key people in the Chinese literary work, *The History of the Three Kingdoms*.

V. THREATS TO VALIDITY

There are some factors that might influence the validity of our study. In this section, we discuss these threats to validity, including internal validity and external validity.

A. THREATS TO INTERNAL VALIDITY

The first threat to the internal validity of our study is related to the WPN that we used to describe people and their co-occurrence relationships in a target literary work. WPN is an undirected network which does not consider the direction of relationship. Worse still, WPN only considers the presence of co-occurrence relationship between two people, but cannot distinguish the nature of the relationship, i.e., our approach cannot tell what kind of relationship the two people have. In short, it seems that WPN cannot precisely describe the rich information between a pair of people. However, for the problem of key people identification, we believe ignoring the direction of edges and the nature of relationship has less influence. Generally, a people with a lot of in-neighbors and out-neighbors might indicate it is a key people. Similarly, a people with a lot of friends and enemies also can indicate it is a key people. But in other types of networks, e.g. people dialogue network, it seems the edge direction cannot be neglected since it depicts the direction of information flow. Also in the work on unrevealing the relationship between relationship types and the formation of the social structure in the literary work, the nature of the relationship between people cannot be ignored. In fact, it is very hard tasks for us to identify two people are friends or enemies by using nature language processing techniques.

Another threat to the internal validity of our study is related to the fact that many literary works have already provided a list of people involved in the literary works. It seems that, in such a scenario, the benefit of our approach is very limited. But we believe, even in such a scenario, our approach can still help the readers. Although these literary works provide a list of people involved in the literary works, they do not rank these people such that readers cannot focus on the most important ones. Our approach provided a ranked list of people, so it can help people focus on the most important people. Furthermore, our approach can also help readers comprehend the relationship between people by following the edges between people in the WPN.

The third threat to the internal validity of our study is related to the baseline approaches that we used to validate our approach. As is shown in the Section II, many studies have made great achievements in the analysis of literary works and also laid a foundation for the work proposed in this paper. However, the existing research work mainly focuses on the analysis of the structural characteristics (e.g., *scale-free* and *small-world*) of people relationship networks by applying the theories and technologies in the field of network science; their focus is not the identification of key people in the literary works. Thus, we cannot compare the performance of our

[11] '司马懿' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms*.
[12] '周瑜' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms*.
[13] '孙策' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms*.
[14] '陆逊' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms*.
[15] '贾栩' are Chinese Characters, denoting the name of a character in *The History of the Three Kingdoms*.

approach with these approaches in the related work. Although we can compare the difference between our WPN with the dialogue network, people co-occurrence network, and multi-layer time-varying network in the related work, it is not the focus of this paper. Our focus in this work is to explore the possibility of identifying key people in a target literary work from a network perspective rather than charactering the WPN. Thus, in the experiments, we did not select the approaches mentioned in Section II as our baseline approaches. To validate our approach, we compared the performance of our approach with the centrality metrics widely used in the field of complex networks. To the best of our knowledge, this work is the first work on key people identification. Indeed, it is preliminary. However, we believe that our preliminary investigation is necessary before future work on this topic can be accomplished.

B. THREATS TO EXTERNAL VALIDITY

A threat to the external validity of our study is related to the literary work that we used to validate our approach. Our experiments are performed on one Chinese literary work. In fact, IPWC can be used on any Chinese literary works with digital versions. The main reason why we performed experiments only on one Chinese literary work is that it is a hard and time-consuming task for us to recruit enough volunteers to identify the key people in a specific literary work. Thus, it is sometimes impossible for us to perform experiments on a large data set containing large numbers of literary works. As we only used one Chinese literary work in the experiments, there is a possibility that the used literary work may not be representative enough. Thus, our study lacks the ability to be generalized to other non-Chinese literary works. To mitigate this threat, we should replicate our study in much more literary works in the future work.

VI. CONCLUSION

This paper proposed an approach to identify the key people in Chinese literary works using *e*-core decomposition. The identified key people can be used to simplify readers' comprehension of literary works. Specifically, when understanding a literary work, we can first focus on the key people, and then follow the edges to key people to understand other people, so as to understand the people and their relationships in the whole literary work. Our approach is based on a network representation (weighted people relationship network) of people and their relationships in literary works. We applied the *e*-core decomposition to identify the key people in *The History of the Three Kingdoms*, and compared it with six other approaches to illustrate the feasibility and effectiveness of our approach. Our approach can be used to build an automatic tool, which can identify the key people for readers to start the comprehension process of a Chinese literary work.

To the best of our knowledge, our current work is the first work on key people identification. Thus, it is preliminary. Specifically, WCN as other types of people relationship networks in the existing work, do not consider the coupling direction between people. For the problem of key people identification, it seems that ignoring the edge direction has less influence. But in other types of networks, e.g. people dialogue network, it seems the edge direction cannot be neglected since it depicts the direction of information flow. Thus, in the future work, we should explore the influence of edge direction on the effectiveness of a specific approach on key people identification. Worse still, in this work, we only validated the feasibility and effectiveness of our IPWC approach on one Chinese literary work; the samples are too few. Thus, in the future, we will try to revalidate our approach in a large set of literary works. In the future, we will also detect the community structures of people relationship networks in the literary works, and reveal the correlation between community structures and the politics, economy and culture of the time period that the literary works described. We believe that our preliminary investigation is necessary before future work on this topic can be accomplished.

ACKNOWLEDGMENT

The authors would like to thank all the students who participated in their study and all the reviewers for their positive and valuable comments and suggestions regarding their manuscript. They would also like to thank T. Wang and Y. Hu who participated in the collection of the initial experimental data.

REFERENCES

[1] K. S. McCarthy and S. R. Goldman, "Constructing interpretive inferences about literary text: The role of domain-specific knowledge," *Learn. Instruct.*, vol. 60, pp. 245–251, Apr. 2019.
[2] Y. H. Ke, S. W. Yu, Z. F. Sui, and J. H. Song, "Research on corpus annotation method based on collective intelligence," *J. Chin. Inf. Process.*, vol. 31, no. 4, pp. 108–131, Aug. 2017.
[3] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *J. Netw. Comput. Appl.*, vol. 132, pp. 86–103, Apr. 2019.
[4] T. Stanisz, J. Kwapień, and S. Drożdż, "Linguistic data mining with complex networks: A stylometric-oriented approach," *Inf. Sci.*, vol. 482, pp. 301–320, May 2019.
[5] J. Baetens, "Conversations on cognitive cultural studies: Literature, language, and aesthetics," *Leonardo*, vol. 48, no. 1, pp. 93–94, Feb. 2015.
[6] D. K. Elson, N. Dames, and K. R. McKeown, "Extracting social networks from literary fiction," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Uppsala, Sweden, 2010, pp. 138–147.
[7] H. Y. Liu and X. H. Yi, "The 'Small World' in literary works?An experimental analysis of characters relation networks in Fitzgerald's novels," *Statist. Inf. Forum.*, vol. 30, no. 12, pp. 102–107, Dec. 2015.
[8] S. D. Prado, S. R. Dahmen, A. L. C. Bazzan, P. M. Carron, and R. Kenna, "Temporal network analysis of literary texts," *Adv. Complex Syst.*, vol. 19, no. 03, May 2016, Art. no. 1650005.
[9] J. S. Zhao, L. Zhang, and Q. M. Zhu, "Extracting and analyzing social networks from Chinese literary," *J. Chin. Inf. Process.*, vol. 31, no. 2, pp. 99–116, Feb. 2017.
[10] Y. B. Wang, J. S. Yu, and C. Y. Zhao, "Research on application of co-word analysis on relationships of characters in the Romance of the Three Kingdoms," *Inf. Res.*, vol. 7, pp. 52–56, Nov. 2017.
[11] F. Lin, G. P. Zhao, N. Lin, and Y. N. Wu, "Analysis of the social network structure of the text in A Dream of Red Mansions," *J. Shijiazhuang Tiedao Univ. (Social Sci. Ed.)*, vol. 21, no. 1, pp. 58–63, Jan. 2018.
[12] X. Zhang, X. Liang, Z. Y. Li, S. S. Zhang, and X. L. Zhao, "Identification and analysis of love relationships of protagonists in Jin Yong's fictions," *J. Chin. Inf. Process.*, vol. 33, no. 4, pp. 109–119, Apr. 2019.
[13] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[14] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Phys.*, vol. 6, no. 11, pp. 888–893, Aug. 2010.
[15] S. Chen, *The History of the Three Kingdoms*. Beijing, China: China Federation of Literary and Art Circles Publishing House, 2016, pp. 1–660.
[16] W. J. Zhang, H. M. Zhang, L. E. Yang, and E. D. Xun, "Multi-grained Chinese word segmentation with Lattice-LSTM," *J. Chin. Inf. Process.*, vol. 33, no. 1, pp. 18–24, 2019.
[17] H. Liu, M. T. Liu, and Y. J. Zhang, "Improved character-based chinese dependency parsing based on stack-tree LSTM," *J. Chin. Inf. Process.*, vol. 33, no. 1, pp. 10–17, Jan. 2019.
[18] F. Kong, H. L. Wang, and G. D. Zhou, "Survey on Chinese discourse understanding," *J. Softw.*, vol. 30, no. 7, pp. 2052–2072, 2019.
[19] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, Baltimore, MD, USA, 2014, pp. 55–60.
[20] A. Garas, F. Schweitzer, and S. Havlin, "A k-shell decomposition method for weighted networks," *New J. Phys.*, vol. 14, no. 8, pp. 83030–83043, Aug. 2012.
[21] H. Li, H. Zhao, J. Q. Xu, B. Li, P. Li, and J. L. Wang, "Research on hierarchy of large-scale software macro-topology base on k-core," *Acta Electronica Sinica*, vol. 38, no. 11, pp. 2635–2643, Nov. 2010.
[22] E. Corwin and A. Logar, "Sorting in linear time-variations on the bucket sort," *J. Comput. Sci. Colleges*, vol. 20, no. 1, pp. 197–202, Oct. 2004.
[23] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998.
[24] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, "Characterization of complex networks: A survey of measurements," *Adv. Phys.*, vol. 56, no. 1, pp. 167–242, Jan. 2007.
[25] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
[26] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Phys. Rep.*, vol. 650, pp. 1–63, Sep. 2016.
[27] W. F. Pan, H. Ming, K. Carl Chang, Z. J. Yang, and D.-K. Kim, "ElementRank: Ranking java software classes and packages using a multilayer complex network-based approach," *IEEE Trans. Softw. Eng.*, early access, Oct. 8, 2019, doi: 10.1109/TSE.2019.2946357.
[28] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 4, pp. 1–7, Oct. 2007.
[29] W. Pan and C. Chai, "Structure-aware mashup service clustering for cloud-based Internet of things using genetic algorithm based clustering algorithm," *Future Gener. Comput. Syst.*, vol. 87, pp. 267–277, Oct. 2018.

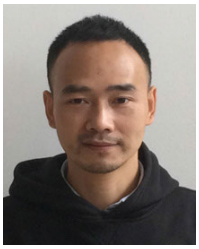
WEIFENG PAN received the Ph.D. degree from the School of Computer, Wuhan University, China, in 2011. He has been a Visiting Scholar with Western Michigan University. He is currently an Associate Professor and the M.S. Supervisor with the School of Computer Science and Information Engineering, Zhejiang Gongshang University. He has published more than 50 articles in international journals, such as the IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, *Future Generation Computer Systems*, and *Cluster Computing*. His current research interests include software engineering, service computing, complex networks, and intelligent computation. He is also a member of the China Computer Federation (CCF) and the CCF Service Computing Association.

XINXIN XU is currently pursuing the M.S. degree with the School of Computer Science and Information Engineering, Zhejiang Gongshang University. Her research interests include software engineering and complex networks.

HUA MING (Member, IEEE) received the Ph.D. degree in computer science from Iowa State University, in 2012. He is currently an Associate Professor with the School of Engineering and Computer Science, Oakland University, Rochester, MI, USA. While his research works primarily focus on the area of software engineering and software intensive systems, his research practice allows him to deeply incorporate computer science theory and novel programming language techniques into discovering, analyzing, and understanding emerging software services. He has published his research work in various reputable journals and conference proceedings. He is a member of ACM.

PING GONG received the Ph.D. degree from the School of Computer, Wuhan University, China, in 2009. He has been a Visiting Scholar with Ulm University. He is currently an Associate Professor with the College of Mathematics and Informatics, Fujian Normal University. He has published more than 20 articles in international journals and conferences. His current research interests include business process management and service oriented computing. He is also a member of the China Computer Federation (CCF).

BO JIANG received the Ph.D. degree from the School of Computer, Zhejiang University, China. She is currently a Professor and the M.S. Supervisor with the School of Computer Science and Information Engineering, Zhejiang Gongshang University. She has published more than 30 articles in international journals and conferences. Her current research interests include service computing and complex networks. She is also a member of the China Computer Federation (CCF) and the CCF Service Computing Association.



CHUNLAI CHAI received the M.S. degree from the School of Computer, Hohai University, China. He is currently an Associate Professor and the M.S. Supervisor with the School of Computer Science and Information Engineering, Zhejiang Gongshang University. He has published more than ten articles in international journals. His current research interests include software engineering and intelligent computation.

BAILIN YANG received the Ph.D. degree from the School of Computer, Zhejiang University, China. He is currently a Professor and the M.S. Supervisor with the School of Computer Science and Information Engineering, Zhejiang Gongshang University. He has published more than 60 articles in international journals and conferences. His current research interests include graphic computing, mobile graphics, and knowledge graph. He is also a member of China Computer Federation (CCF).