Hindawi
Evidence-Based Complementary and Alternative Medicine
Volume 2020, Article ID 8810016, 10 pages
https://doi.org/10.1155/2020/8810016

Research Article

Topological Analysis of the Language Networks of Ancient Traditional Chinese Medicine Books

Qunsheng Zou,[1,2] **Yinyan Wang**,[1,2] **Zixin Shu**,[1,2] **Kuo Yang**,[1,2] **Jingjing Wang**,[1,2] **Kezhi Lu**,[1,2] **Qiang Zhu**,[1,2] **Baoyan Liu**,[3] **Runshun Zhang**,[4] **and Xuezhong Zhou**[1,2]

[1]*Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*
[2]*Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*
[3]*Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China*
[4]*Guang'anmen Hospital, China Academy of Chinese Medicine, Beijing 100700, China*

Correspondence should be addressed to Xuezhong Zhou; xzzhou@bjtu.edu.cn

Received 13 September 2020; Accepted 24 November 2020; Published 11 December 2020

Academic Editor: Zhaohui Liang

This study aims to explore the topological regularities of the character network of ancient traditional Chinese medicine (TCM) book. We applied the 2-gram model to construct language networks from ancient TCM books. Each text of the book was separated into sentences and a TCM book was generated as a directed network, in which nodes represent Chinese characters and links represent the sequential associations between Chinese characters in the sentences (the occurrence of identical sequential associations is considered as the weight of this link). We first calculated node degrees, average path lengths, and clustering coefficients of the book networks and explored the basic topological correlations between them. Then, we compared the similarity of network nodes to assess the specificity of TCM concepts in the network. In order to explore the relationship between TCM concepts, we screened TCM concepts and clustered them. Finally, we selected the binary groups whose weights are greater than 10 in *Inner Canon of Huangdi* (ICH, 黄帝内经) and *Treatise on Cold Pathogenic Disease* (TCPD, 伤寒论), hoping to find the core differences of these two ancient TCM books through them. We found that the degree distributions of ancient TCM book networks are consistent with power law distribution. Moreover, the average path lengths of book networks are much smaller than random networks of the same scale; clustering coefficients are higher, which means that ancient book networks have small-world patterns. In addition, the similar TCM concepts are displayed and linked closely, according to the results of cosine similarity comparison and clustering. Furthermore, the core words of *Inner Canon of Huangdi* and *Treatise on Cold Pathogenic Diseases* have essential differences, which might indicate the significant differences of language and conceptual patterns between theoretical and clinical books. This study adopts language network approach to investigate the basic conceptual characteristics of ancient TCM book networks, which proposes a useful method to identify the underlying conceptual meanings of particular concepts conceived in TCM theories and clinical operations.

1. Introduction

As a traditional medicine with medical theories and concepts mainly established thousands of years ago, TCM has abundance of high-value ancient books written or printed in the form of Chinese classical binding before 1912, which conceive important TCM theories and concepts and the clinical principle for disease diagnosis and treatment [1, 2]. Although many TCM antecessors performed significant theoretical investigations to digest the knowledge employed in these books, which has promoted the advances of contemporary clinical practical solutions for the managing of various complicated diseases in real-world clinical settings [3], it is particularly important to investigate the language characteristics of ancient TCM books, which will help understand the theoretical knowledge exactly expressed in those texts [4]. However, little research was conducted to understand the language regularities of the key TCM concepts (e.g., Yin, Yang, and Qi) in these ancient books using computational linguistics and complex network approaches [5].

Complex network has been developed as a mainstream approach for investigating the regularities in the fields with complex phenomena, such as social science, biological science, and linguistics [6–8]. For this approach, network or graph consisting of nodes and links is the form to represent the structures of the related systems. Due to the complex organization and interactions between various medical entities, complex network approaches have been used for exploring the rules of associations between herbs, symptoms, syndromes, and human meridians [9–12]. However, rare work was conducted on the analysis of language regularities of TCM books by complex network approaches [13].

In this paper, firstly, we constructed directed networks from the full texts of ancient TCM books. Then, we analyzed the statistical characteristics of the networks, identified the centrality patterns of core TCM concepts, and explored the similarities and differences between different ancient books. In addition, we demonstrated what the diversity of the concepts, such as "Qi (气)", "Yin (阴)", "Yang (阳)", "Xie (泻)", and "Li (痢)", in Chinese medicine, and what special conceptual meanings they would have.

2. Materials and Methods

2.1. Dataset of 80 Ancient TCM Books. The data we used was derived from texts of 80 ancient Chinese medicine books (Table 1 shows a typical collection of 30 books), with an emphasis on the analysis of the books of *Inner Canon of Huangdi* (ICH) and *Treatise on Cold Pathogenic Disease* (TCPD). For example, ICH contains 189,984 characters and TCPD contains 43,331 characters. Data cleaning was performed to remove the characters in ancient texts except for Chinese and periods and separated the whole text into sentences. For example, ICH and TCPD, after cleaning the data of these two ancient books, we obtained 6,237 sentences and 1,366 sentences, respectively.

Table 1: The power exponents of TCM language networks.

Category	Book	Power exponent
Medical classics	*ICH*	1.0169
	CMP	1.1703
	YJYZ	0.9864
Basic theory	*HSZZJ*	1.2469
	YXQY	1.1514
	WXDY	1.1367
Typhoid	*TCPD*	1.2000
	SGC	1.2169
	ZQZSHL	1.0271
Diagnostic methods	*CBZN*	1.2360
	YDXY	1.1020
	ZJSY	1.1778
Acupuncture and massage	*A-B-CAM*	1.0533
	ZJZN	1.3123
	ZJZSJ	1.1251
Materia medica	*EMM*	1.1205
	CNCMM	1.1880
	SNCMM	1.2647
Medical formulary	*VPE*	**1.3246**
	BZYBY	1.1889
	JYF	1.2502
Clinical examination	*BQHB*	1.2099
	SBOC	1.0794
	CDP	1.0983
Health preserving	*BPZNP*	1.1610
	BPZWP	1.1988
	YSML	1.3086
Comprehensive work	*GJMYHC*	1.0098
	MCB-A-M	**0.9499**
	EM	0.9586

Note. ICH: Inner Canon of Huangdi; *CMP: Classic on 81 Medical Problems* (黄帝八十一难经); *YJYZ: Yijing Yuanzhi* (医经原旨); *HSZZJ: Huashi Zhongzang Jing* (华氏中藏经); *YXQY: Yixue Qiyuan* (医学启源); *WXDY: Wuxing Dayi* (五行大义); *TCPD: Treatise on Cold Pathogenic Diseases*; *SGC: Synopsis of Golden Chamber* (金匮要略); *ZQZSHL: Zhangqingzi Shanghan Lun* (张卿子伤寒论); *CBZN: Chabing Zhinan* (察病指南); *YDXY: Yideng Xuyan* (医灯续焰); *ZJSY: Zhenjia Shuyao* (诊家枢要); *A-B CAM: A-B Classic of Acupuncture and Moxibustion* (针灸甲乙经); *ZJZN: Zhenjing Zhinan* (针经指南); *ZJZSJ: Zhenjiu Zisheng Jing* (针灸资生经); *EMM: Essentials of Matea Medica* (本草备要); *CNCMM: Collective Notes to the Canon of Materia Medica* (本草经集注); *SNCMM: Shennong's Classic of Materia Medica* (神农本草经); *VPE: Valuable Prescriptions for Emergency* (备用千金要方); *BZYBY: Buzhi Yi Biyao* (不知医必要); *JYF: Jiyan Fang* (集验方); *BQHB: Bian Que Heart Book* (扁鹊心书); *SBOC: Secret Book of Orchid Chamber* (兰室秘藏); *CDP: Confucians' Duties to Parents* (儒门事亲); *BPZNP: Baopuzi Neipian* (抱朴子内篇); *BPZWP: Baopuzi Waipian* (抱朴子外篇); *YSML: Yangsheng Milu* (养生秘录); *GJMYHC: Gujin Mingyi Huicui* (古今名医汇粹); *MCB A-M: Medical Complete Book, Ancient and Modern* (古今医统大全); and *EM: Elementary Medicine* (医学入门).

2.2. Language Network Construction Using 2-Gram Model. In the field of computational linguistics, n-gram is a widely used method to model natural language; particularly, an n-gram is a contiguous sequence of n items from a given sequence of text [14, 15]. Here, we used 2-gram model to obtain the sequential links of characters in ancient TCM books. Given a sentence, we would generate a directed path with characters as nodes and the sequential associations between them as links. When all the sentences of a given book were processed, we would obtain a weighted directed language network, in which the number of identical sequential associations is considered as the weight of the link. For example, the sentence "阴阳者, 天地之道也" in ICH can be processed to a directed path (Figure 1(a)) [16]. We have built the language networks for all the 80 ancient TCM books. In particular, the network of ICH contained 2,367 nodes and 35,502 directed links (see Figure 1(b)).

2.3. Basic Network Characteristics. The number of links connected to each node in TCM book networks, that is, the degree of the node [17]. We counted the degree of each node in the network and figured up the number of nodes with the same degree and attempted to find out whether the degree distributions of ancient book networks are consistent with the power law [18]. By calculating the average path length [19] and clustering coefficients [20] of networks, we judged whether these networks possess the small-world property. Thus,

$$l_G = \frac{1}{n(n-1)} \cdot \sum_{i=j} d\left(v_i, v_j\right), \qquad (1)$$

where l_G is the average path length of graph G, n is the number of nodes, and $d\left(v_i, v_j\right)$ denotes the shortest distance between v_i

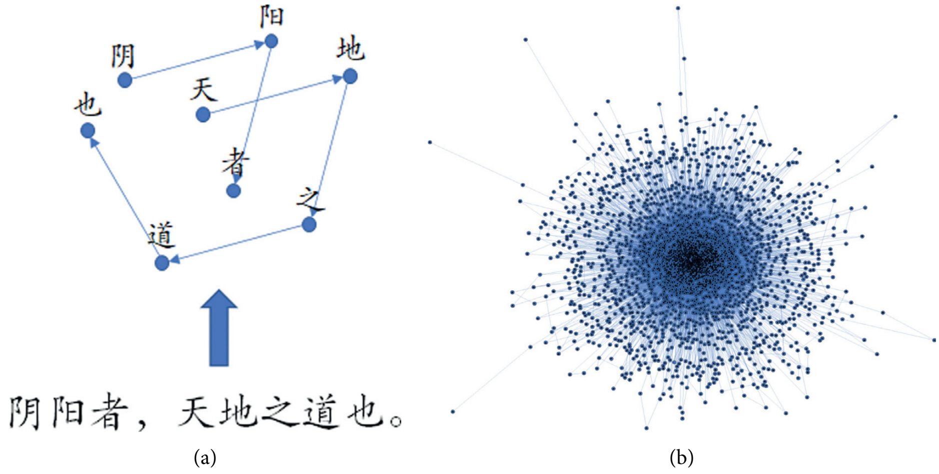


Figure 1: (a) An example of ancient TCM book network. (b) The language network of ICH.

and v_j. When v_j cannot be reached from v_i, $d(v_i, v_j) = 0$. Moreover, clustering coefficients are acquired by

$$C_i = \frac{\left|\left(e_{jk}\colon v_j, v_k \in N_i, e_{jk} \in E\right)\right|}{d_i\left(d_i - 1\right)},$$
$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i. \qquad (2)$$

Consider C_i as the local clustering coefficient of node i, where d_i is the degree of node i, N_i is a set of nodes which immediately connected with node i, E is defined as a set of edges in graph G, and e_{jk} is the edge of nodes j and k. Then, all C_i were summed and averaged to get the average clustering coefficient $\overline{C}$.

2.4. Homogeneity of the Centrality of Similar TCM Concepts in the Language Network. In TCM theories, there are hundreds of similar basic concepts (often in the form of a single character word), such as “Yin and Yang (阴阳)”, “the five elements (五行)”, and “the five internal organs (五脏)”, which are essential for TCM theories and clinical solutions [21]. We supposed that this kind of essentiality or importance of the concepts could be captured by vector representation of nodes in the directed language networks. In addition, we assumed that for those similar concepts, they would finally have similar centralities; that is, the similar concepts would display the same degree of centrality homogeneity compared with the random sets of concepts. The vector representation of each node in book network is calculated by the Node2Vec framework which learns low-dimensional representations for nodes in a graph [22]. To investigate the homogeneity phenomenon of these similar concepts, we proposed the following methods to differentiate between similar concepts and their random controls:

$$\text{sim} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \times \sqrt{\sum_{i=1}^{n} (b_i)^2}}, \qquad (3)$$

where A and B are the low-dimensional representation of nodes, a_i and b_i are their components, and $i = 1, 2, \ldots, n, n$ is the dimension of vectors.

Determine S as a set of similar concepts and R as the set of another nodes in network. The similarity of concepts in groups will be defined by $\text{sim}_{i,j}$ and similarity between in-group concepts and out-of-group concepts will be defined by $\text{sim}_{i,k}$, where $i, j \in S$ and $k \in S$. If our hypothesis is correct, then $\text{sim}_{i,j}$ is generally greater than $\text{sim}_{i,k}$.

2.5. t-Test on Similarity Sequences. Through the above method, we gained the difference between similarity of each basic concept and random concept. However, the current results were only relative to a single basic concept, which did not indicate that similar concepts are homogeneous to some extent. There is a way to solve this problem, named the Student *t*-test, which is often used to assess whether the means of two classes are statistically different from each other by calculating a ratio between the difference of two class means and the variability of the two classes [23, 24]. In this way, we can verify the homogeneity of basic concepts indirectly through the results of *t*-test. Whereupon, we combined results of each basic concept into similarity sequence M and results of random concepts into similarity sequence N. After performing *t*-test on these two sequences, if P value is less than 0.05, we believe these basic concepts are similar to each other and consistent with homogeneity.

2.6. Identifying the Concept Clusters. It is well recognized that the complex networks like language network often hold a kind of community structures with some subnetworks involving dense links while sparse links outside those subnetworks. These subnetworks, which are considered as network clusters or communities, would deliver domain meaningful knowledge for further investigation. To detect the concept clusters or communities in the TCM language network, we applied the Fast Unfolding Algorithm (FUA) which was a well-known community detection method based on modularity [25] to detect the communities of a given network by

$$Q = \frac{1}{2m} \sum_{ij} \left[W_{ij} - \frac{k_i k_j}{2m} \right] \delta\left(C_i, C_j\right). \tag{4}$$

Think of Q as the modularity of the entire network, where $m = (1/2)\sum_{ij} W_{ij}$ represents the sum of the weights of all the edges in the network, W_{ij} is the weight between node i and node j, $k_i = \sum_j W_{ij}$ is the sum of the weights of the edges which connected to node i, and C_i indicates the community which node i is assigned to. The value of $\delta(C_i, C_j)$ is 0 or 1; when $\delta(C_i, C_j) = 1$, it means node i and node j are in the same community; otherwise, node i and node j are not in the same community. Then, iteratively making the modularity reach the maximum value, the final clustering results were obtained.

3. Results

3.1. Basic Characteristics of TCM Language Network. It is observed that the degree distribution of ancient Chinese medicine nodes is consistent with the power law distribution [26, 27] (Figures 2(a) and 2(b)), which means that although most characters were rarely used together with other characters; there are some "Hub" characters, such as "Qi", "Yin", and "Yang", connecting to a various number of characters in the sentences. We listed the basic network features of the 30 typical TCM books which are divided into 10 categories. It can be found that the power exponents of books are close to 1.0; the biggest one is 1.3246 of the *Valuable Prescriptions for Emergency*, and the smallest one is 0.9499 of the *Medical Complete Book, Ancient and Modern*. The node degree distributions of these ancient books follow

$$p(k) \propto k^{-\gamma}, \tag{5}$$

where k is the degree of the node, $p(k)$ is the ratio of the number of nodes with a degree of k to the total number of nodes, and γ is the power exponent which floats above and below 1.

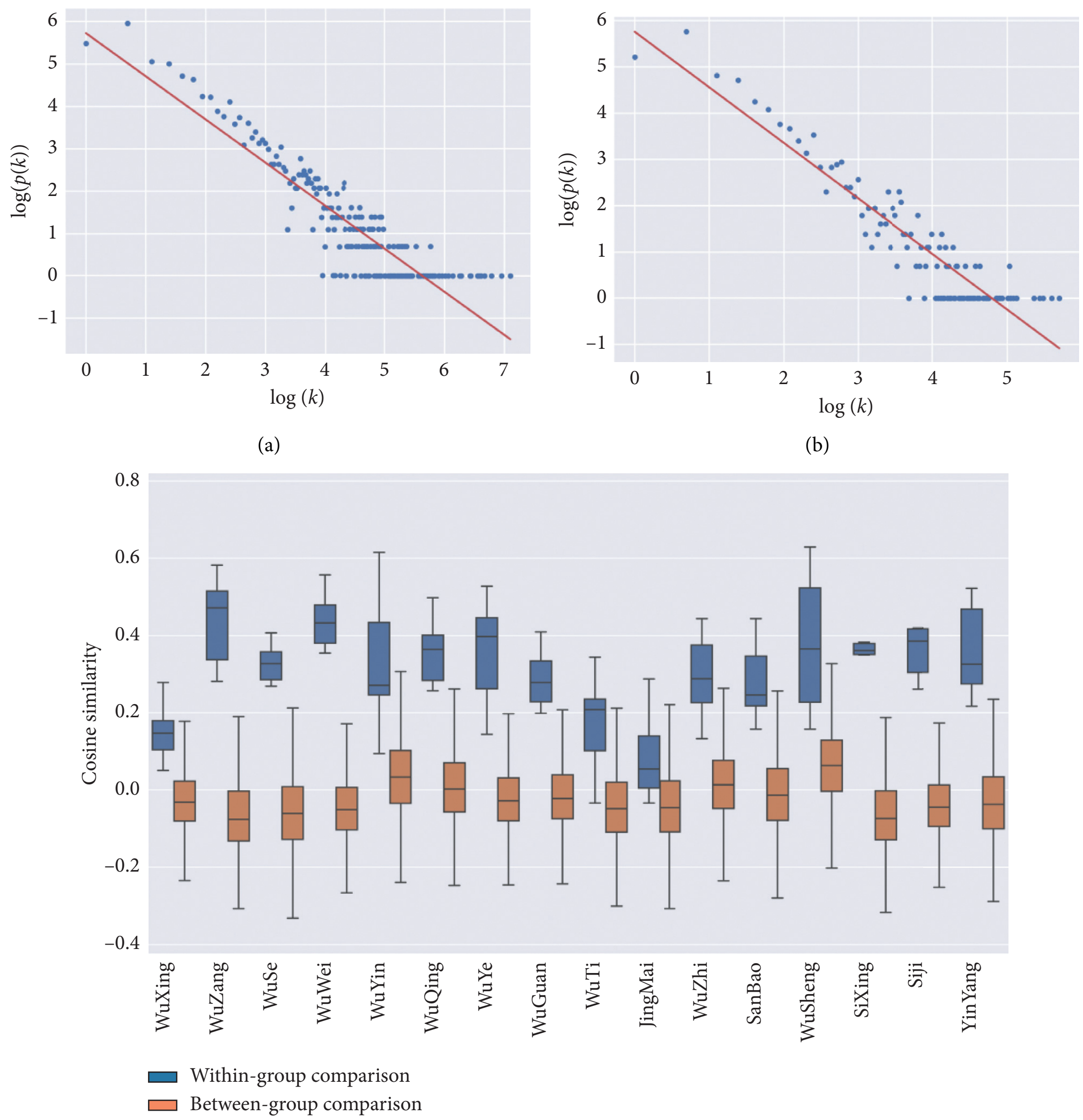
In addition, the average path lengths of these networks are around 3, in which the largest one is 3.819 and the smallest one is 2.727. The clustering coefficients are distributed between 0.1 and 0.3 (Table 2). Comparing these ancient books with random networks of the same scale, it is found that their average path lengths are smaller than random networks and the clustering coefficients are larger than random networks. It means that TCM language networks conform to the small-world pattern [7].

3.2. Topological Homogeneity of TCM Basic Concept Groups. To validate the power of complex network approach to differentiate the semantic groups of basic TCM concepts from the language network, we calculated the cosine similarity of each node vector of 16 basic TCM concept groups (Table 3). We assumed that the basic TCM concept groups, such as these concepts of five elements, would have similar values for the centrality measures, which would reflect their similar semantic importance in the language network from the topological measures. The results showed that most of the basic TCM concept groups in basic theoretical books (e.g., ICH) are more similar to each other than those of random controls (Table 4), which indicated that these basic TCM concept groups display a kind of linking homogeneity reflecting their close category semantic similarities. For example, the five elements concept category includes Mu, Huo, Tu, Jin, and Shui as closely related members. We found that, in the ICH book (Figure 2(c)), the cosine similarity of these five elements ranges from 0.0499 to 0.2786 with a rather high value of very narrow variance (mean: 0.1498 + std: 0.0669). This demonstrated the central role of the concepts in the five elements category for TCM and the categorical homogeneity of these five concepts. In addition, these category similar concepts could be identified by community detection methods due to their similar connection patterns in the context of network. For example, in ICH network, using FUA (see methods), we could identify the concept groups as same communities, such as "Yin and Yang", "the five elements", "the five notes", and "the five colors" from the whole network (Figure 3).

However, the results were different for those clinical books (e.g., TCPD). The cosine similarity of the basic TCM concept groups did not tend to show homogeneous patterns. This might be due to the differently focused subjects of these books. For example, TCPD is mainly focusing on the manifestations of six types of syndromes and their regularities of herb treatment.

3.3. Diversity of TCM Language Networks. To further investigate the distinct topological patterns involved in different TCM language networks, we screened the links whose weights are >10 in ICH and TCPD and regarded the related nodes (Chinese characters) as key concepts in these two books (Figure 4). It is illustrated that the key concepts in ICH mainly include the basic theoretical characters in TCM, such as "Yin/Yang", "the five elements" and its associated concepts, quantifiers, emotions, and pulse (Figure 4(b)). In contrast, although several basic theoretical concepts, such as "Yin/Yang", are still included in TCPD as the key concepts, most of the others are related to herb prescriptions and symptoms (Figure 4(b)). These results indicated the distinct category of knowledge delivered in these two books. It is well known that ICH ensembles the basic theories of TCM, while TCPD is recognized as a representative clinical book focusing on disease manifestations, pathologies, and their corresponding herb prescriptions.

3.4. Exploring the Specific Semantic Intensions of Core TCM Concepts. To identify the specific meaning of a given concept, we would like to see what exactly words or phrases it occurred. The TCM language network could give help to this investigation. It is well known that some basic concepts, such as "Qi" and "Yin and Yang", are of great significance to TCM; however, the connotations of these concepts are rather complicated [28–30]. We constructed an integrated language network with various character



(a)

(b)

(c)

Figure 2: The degree patterns of TCM language networks. (a) Degree distribution of ICH. (b) Degree distribution of TCPD. (c) Cosine similarity of with-group comparison and between-group comparison.

triples derived from 30 ancient TCM books, which finally consists of 6118 nodes and 381467 links. Here, we extracted the 1-order neighborhood of a given node and took four concepts, namely, “Qi”, “Yin/Yang”, “Xie”, and “Li” for demonstrations (Figure 5). It is interesting that for the basic concept of “Qi”, there are about 1864 characters (nodes) directly connecting to this node, in which the characters, such as “Yang”, “Xue”, “Yin”, “Yuan”, “Zheng”, and “Jing” together with “Qi” represent the main types of “Qi” recognized in TCM theories. The other connecting characters to “Qi” obtained the various manifestations and pathologies, such as “XiaQi”, “QiNi” and “QiXu”; “XieQi”, “HanQi”. Although these concepts are usually adopted by professional TCM practitioners, our network results could grasp and demonstrate the global connecting characters for TCM researchers. Similarly, we obtained 1736 characters related to “Yin and Yang”, which could form different “Yin and Yang”-related basic concepts, such as meridian-related concepts (e.g., “TaiYin”, “TaiYang”, “YangMai”, and “YinMai”), syndrome-related concepts (e.g., “YinXu” and “YangXu”). For the character “Xie”, our network approach clearly showed two types of semantics involved. One type of concept is related to different manifestations, such as “XieXie”, “TuXie”, “ShuiXie”, and “FengXie”. Another type of concept is related to the principles for prescriptions including “XieXin” and “XieHuo”. However, the concepts related to “Li” are only associated with disorders or diseases, such as “XueLi”, “NueLi”, “GanLi”, “LiChang”, and “LenLi” [1]. The rigorous evaluation of these related concepts would help with the precise understanding of the

Table 2: Basic topological characteristics of TCM ancient books.

Book	Average path length (APL)	APL random	P value	Average clustering coefficients (ACC)	ACC random	P value
ICH	3.036	2.973	$<2.23E-308$	0.249	0.0077	$<2.23E-308$
CMP	3.590	4.012	$<2.23E-308$	0.123	0.0092	$7.7E-71$
YJYZ	2.912	2.957	$<2.23E-308$	0.297	0.0077	$<2.23E-308$
HSZZJ	3.674	3.219	$<2.23E-308$	0.097	0.0078	$9.44E-129$
YXQY	3.434	3.292	$<2.23E-308$	0.140	0.0076	$3.7E-164$
WXDY	3.329	2.967	$<2.23E-308$	0.154	0.0076	$1.9E-298$
TCPD	3.584	3.367	$<2.23E-308$	0.119	0.0077	$5.06E-120$
SGC	3.548	3.340	$<2.23E-308$	0.107	0.0067	$3.34E-127$
ZQZSHL	3.199	3.164	$2.77E-167$	0.202	0.0074	$6.57E-276$
CBZN	3.702	3.509	$<2.23E-308$	0.098	0.0073	$4.76E-91$
YDXY	3.103	2.847	$<2.23E-308$	0.188	0.0075	$<2.23E-308$
ZJSY	3.505	3.712	$<2.23E-308$	0.135	0.0082	$2.03E-89$
A-B CAM	3.096	3.060	$<2.23E-308$	0.232	0.0074	$<2.23E-308$
ZJZN	3.749	3.472	$<2.23E-308$	0.085	0.0069	$7.3E-74$
ZJZSJ	3.275	3.062	$<2.23E-308$	0.158	0.0077	$5.62E-275$
EMM	3.227	2.884	$<2.23E-308$	0.154	0.0074	$<2.23E-308$
CNCMM	3.327	2.929	$<2.23E-308$	0.144	0.0074	$<2.23E-308$
SNCMM	3.622	2.935	$<2.23E-308$	0.112	0.0076	$4.21E-209$
VPE	**3.819**	3.259	$<2.23E-308$	**0.078**	0.0075	$3.66E-99$
BZYBY	3.406	3.064	$<2.23E-308$	0.127	0.0075	$5.41E-231$
JYF	3.521	3.086	$<2.23E-308$	0.106	0.0075	$3.49E-180$
BQHB	3.488	3.009	$<2.23E-308$	0.116	0.0076	$2.49E-197$
SBOC	3.302	3.259	$<2.23E-308$	0.158	0.0075	$3.06E-206$
CDP	3.143	2.875	$<2.23E-308$	0.191	0.0076	$<2.23E-308$
BPZNP	3.185	2.831	$<2.23E-308$	0.168	0.0073	$<2.23E-308$
BPZWP	3.167	2.877	$<2.23E-308$	0.159	0.0074	$<2.23E-308$
YSML	3.814	3.461	$<2.23E-308$	0.086	0.0077	$2.05E-75$
GJMYHC	3.043	2.932	$<2.23E-308$	0.244	0.0072	$<2.23E-308$
MCB A-M	**2.727**	2.767	$<2.23E-308$	**0.319**	0.0075	$<2.23E-308$
EM	2.843	2.795	$<2.23E-308$	0.270	0.0075	$<2.23E-308$

t-test was used to compare the real and random measures.

Table 3: The mean and standard deviation of cosine similarity of each node of 16 basic TCM concept groups in ICH.

Concept groups	Mean	Standard deviation
WuXing (五行)	0.1498	0.0669
WuZang (五脏)	0.4378	0.0996
WuSe (五色)	0.3407	0.0712
WuWei (五味)	0.4363	0.0633
WuYin (五音)	0.3351	0.1487
WuQing (五情)	0.3581	0.0756
WuYe (五液)	0.3587	0.1144
WuGuan (五官)	0.2855	0.0681
WuTi (五体)	0.1692	0.1104
WuZhi (五志)	0.3144	0.1496
WuSheng (五声)	0.3703	0.1674
SiXing (四性)	0.3496	0.0767
SiJi (四季)	0.3963	0.1243
YinYang (阴阳)	0.3606	0.1178
SanBao (三宝)	0.2824	0.0969
JingMai (经脉)	0.0904	0.1222

Table 4: The difference between basic TCM concept groups and random controls in ancient books (by t-test).

Book	P value
ICH	$1.08E-12^{c}$
CMP	0.064492
YJYZ	$6.28E-23^{c}$
HSZZJ	0.418429
YXQY	0.361999
WXDY	$3.58E-23^{c}$
TCPD	0.856763
SGC	0.807354
ZQZSHL	0.182582
CBZN	0.650913
YDXY	0.239566
ZJSY	0.677128
A-B-CAM	$3.93E-06^{c}$
ZJZN	0.061295
ZJZSJ	0.018853^{a}
EMM	0.086299
CNCMM	0.575891
SNCMM	0.059049
VPE	$5.95E-06^{c}$
BZYBY	0.012146^{a}
JYF	0.027935^{a}
BQHB	0.000277^{c}
SBOC	0.499433
CDP	0.281959
BPZNP	$4.97E-30^{c}$
BPZWP	$1.10E-27^{c}$
YSML	0.740321
GJMYHC	$3.84E-06^{c}$
MCB-A-M	0.008839^{b}
EM	0.572089

Note. P value < 0.05 means that most of the basic TCM concept groups in this book are more similar to each other than random controls. [a]P value < 0.05, [b]P value < 0.01, and [c]P value < 0.001.

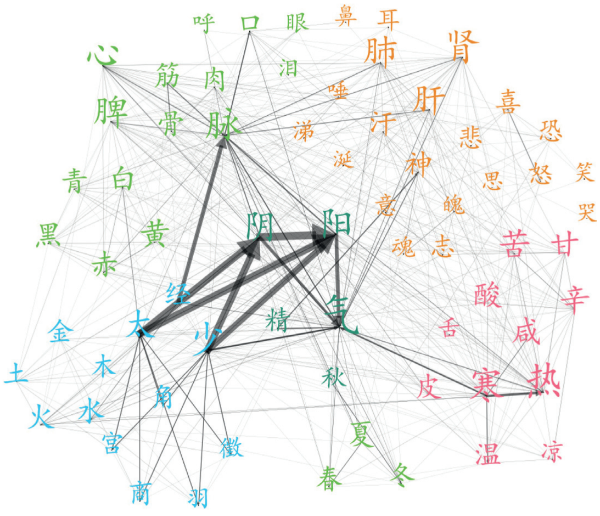


Figure 3: TCM concepts clusters of *Inner Canon of Huangdi.*

manifestations related to "Li" and improve distilling the high-value disease or prescription knowledge from TCM ancient literatures.

4. Discussion

Medical concepts constitute the basic knowledge framework of TCM theories devoted to the clinical observation of the complicated manifestations and their understanding of the underlying pathologies from TCM perspectives. Therefore, the development of TCM terminologies even with international translations is an important task in TCM field [31–33]. However, as most TCM concepts derived from ancient textbooks, it is difficult for contemporary practitioners to definitely grasp the whole meanings and connotations in the framework of TCM theories, in which the semantic diversity of a specific TCM concept is one of the key issues. Language network proposes an efficient approach to investigate the semantic properties of concepts of words in large-scale text corpora [34]. The application of complex network in linguistics has made it possible for us to adopt real network analysis tools in ancient TCM book studies. Unfortunately, the current researchers of TCM are mostly those with medical science background, who usually concern themselves with clinical medicine. It lacks some approaches, which focus on ancient TCM books' concepts, are not only helpful to the research on basic theory of TCM but also helpful to nonprofessionals understanding the basic concepts.

Ancient TCM books are carriers of Chinese medicine knowledge and have great significance for the entire Chinese civilization [35]. In this paper, we analyzed TCM books in the form of network and explored some characteristics of ancient language networks. First of all, the node degree distributions, average path lengths, and clustering coefficients of the networks showed that TCM character-language networks follow a kind of scale-free and small-world networks. Secondly, we analyzed the basic concepts in ancient TCM book networks and found that these concepts play special roles in language networks. Furthermore, we extracted key TCM concepts of each book and found that the key concepts in different categories of ancient books have obvious differences. Finally, we drew a conclusion that Chinese medicine concepts such as "Qi" have rich medical connotations in ancient books.

There are several limitations in our manuscript. Firstly, we only constructed dozens of language networks, which might influence the extensions of the obtained results to more general context. In addition, the character-based 2-gram modeling also limits the investigation capability of the language network for semantic issues. Secondly, although most TCM basic concepts could be grasped by single character (e.g., Qi), there exist many key concepts, such as those of acupuncture points, herbs, and disorders, which would necessarily be represented by words or phrases to further explore their semantic regularities. Furthermore, it is notable that network approach is adept in investigating the global patterns of a given domain, which could be combined with other data analysis methods (e.g., association rules) to generate more specific results to deliver TCM meaningful knowledge.

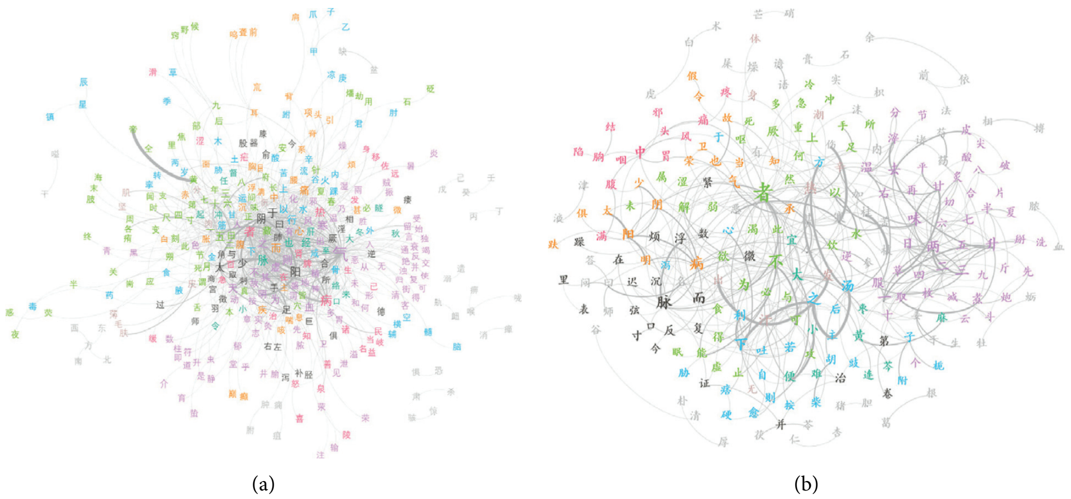


Figure 4: Two cases of TCM language networks. (a) The key TCM concepts of *Inner Canon of Huangdi*. (b) The key TCM concepts of *Treatise on Cold Pathogenic Diseases.*

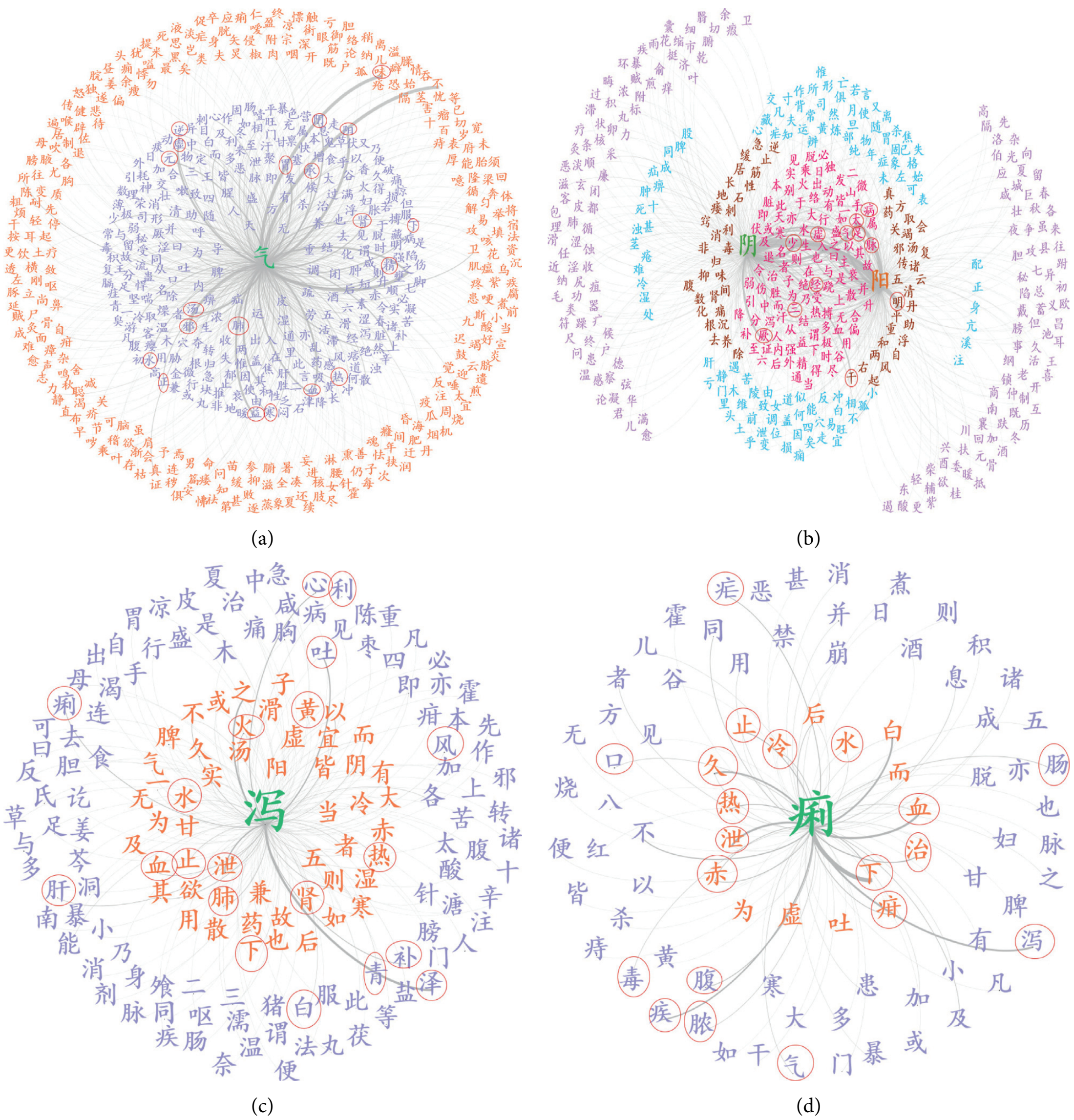


Figure 5: (a) The network centered on "Qi." (b) The network centered on "Yin" and "Yang." (c) The network centered on "Xie." (d) The network centered on "Li."

5. Conclusion

In summary, we found that the degree distribution of ancient TCM book networks is consistent with power law distribution and small-world patterns. In addition, similar concepts in ancient books are displayed and linked closely. Moreover, we realized that there are essential differences in language and conceptual patterns between theoretical and clinical books. To sum up, the exploration of ancient TCM books provides an effective method to identify the underlying conceptual meanings of particular concepts conceived in TCM theories and clinical operations.

Data Availability

Data used in this paper are found at https://gitee.com/zouqunsheng/ancient-tcm-books.git.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National Key Research and Development Program, China (2017YFC1703506, 2017YFC1703502, 2017YFC1703505, and 2020YFC0841600), the Special Programs of Traditional Chinese Medicine China (JDZX2015168, JDZX2015171, and JDZX2015170), and the Fundamental Research Funds for the Central Universities, China (no. 2018JBZ006).

References

[1] P. Jiang, Y. Li, Y. Zhou et al., "Reflections on the collation of ancient books of traditional Chinese medicine," *Chinese Medicine Modern Distance Education of China*, vol. 16, no. 6, pp. 62-63, 2018.

[2] F. Meng, W. Shang, S. Li et al., "Classification systems of ancient books on traditional Chinese medicine and their evolution," *Chinese Journal of Medical Library and Information Science*, vol. 24, no. 9, pp. 62–66, 2015.

[3] T. Yu, J. Li, Q. Yu et al., "Knowledge graph for TCM health preservation: design, construction, and applications," *Artificial Intelligence in Medicine*, vol. 77, pp. 48–52, 2017.

[4] J. Xie and C. Jia, "Types and functions of metaphorical language in internal classic of huang di," *Acta Chinese Medicine and Pharmacology*, vol. 39, no. 1, pp. 1–4, 2011.

[5] X. Zhou, Y. Peng, and B. Liu, "Text mining for traditional Chinese medical knowledge discovery: a survey," *Journal of Biomedical Informatics*, vol. 43, no. 4, pp. 650–660, 2010.

[6] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[7] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, p. 268, 2001.

[8] H. Liu, "Linguistic Complex Networks: a new approach to language exploration," *Grundlagenstudien Aus Kybernetik Geisteswissenschaft (grkg/Humankybernetik)*, vol. 52, pp. 151–170, 2011.

[9] S. Li, B. Zhang, D. Jiang et al., "Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae," *BMC Bioinformatics*, vol. 11, no. 11, p. S6, 2010.

[10] C. van Borkulo, L. Boschloo, D. Borsboom, B. W. J. H. Penninx, L. J. Waldorp, and R. A. Schoevers, "Association of symptom network structure with the course of depression," *JAMA Psychiatry*, vol. 72, no. 12, pp. 1219–1226, 2015.

[11] R. Goekoop and J. G. Goekoop, "A network view on psychiatric disorders: network clusters of symptoms as elementary syndromes of psychopathology," *PLoS One*, vol. 9, no. 11, Article ID e112734, 2014.

[12] Y. Bai, J. Wang, J.-P Wu et al., "Review of evidence suggesting that the fascia network could be the anatomical basis for acupoints and meridians in the human body," *Evidence-based Complementary and Alternative Medicine*, p. 2011, 2011.

[13] H. Wan, M.-F. Moens, W. Luyten et al., "Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, p. 356, 2016.

[14] P. F. Brown, P. V. Desouza, R. L. Mercer et al., "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[15] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," 1994.

[16] J. H. Martin and D. Jurafsky, "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition," *Pearson/Prentice Hall*, vol. 23, 2009.

[17] M. Brede, "Networks—an introduction," *Artificial Life*, vol. 18, no. 2, pp. 241-242, 2010.

[18] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[19] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[20] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[21] Q. Jane, "Traditional medicine: a culture in the balance," *Nature*, vol. 448, no. 7150, p. 126, 2007.

[22] A. Grover and J. Leskovec, "node2vec: scalable feature learning for networks," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 4, pp. 855–864, 2016.

[23] J. F. Box, "Guinness, Gosset, Fisher, and small samples," *Statistical Science*, vol. 2, no. 1, pp. 45–52, 1987.

[24] J. L. Devore and R. Peck, "Statistics: the exploration and analysis of data," *Duxbury Press*, vol. 4, 2005.

[25] V. D. Blondel, J. L. Guillaume, R. Lambiotte et al., "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics*, vol. 2008, no. 10, pp. 155–168, 2008.

[26] C. M. Bernat and R. V. Solé, "Universality of Zipf's law," *Journal of Statistical Mechanics*, vol. 82, no. 1, Article ID 011102, 2010.

[27] B. Albert-László, "Scale-free networks: a decade and beyond," *Science*, vol. 325, no. 5939, pp. 412-413, 2009.

[28] X. Dong, "The meaning evolution of "material force" and the theory in traditional Chinese medicine," *Journal of Liaoning Medical College: Social Science Edition*, vol. 14, no. 4, pp. 63–66, 2016.

[29] M. Yan and Y. Wang, "Objectization of conceptual and state of affairs of spatiotemporal sequence by qi's reference in traditional Chinese medicine," *Journal of Beijing University of Traditional Chinese Medicine*, vol. 41, no. 9, pp. 717–725, 2018.

[30] Z. Wang, "Study on the essence of Yin and Yang," *Traditional Chinese Medicine and Related Issues*, vol. 41, 2014.

[31] Z. Li, "Standardizing English translation of traditional Chinese medical terminology:an analysis of the Concepts,Principles and methods concerned," *Chinese Translators Journal*, vol. 29, no. 4, pp. 63–70, 2008.

[32] J. Zhu, "Standardization of terms of traditional Chinese medicine (TCM) and modernization and internationalization of TCM," *China Journal of Traditional Chinese Medicine and Pharmacy*, vol. 1, pp. 6–8, 2006.

[33] Z. Xie, "Discussion on English translation of traditional Chinese medicine terms," *Chinese Journal of Integrated Traditional and Western Medicine*, vol. 9, pp. 706–709, 2000.

[34] C. Biemann, S. Roos, and K. Weihe, "Quantifying semantics using complex network analysis," *Proceedings of COLING*, vol. 2012, pp. 263–278, 2012.

[35] Y. Fu, G. Liu, B. Li et al., "Digital research on ancient books of traditional Chinese medicine," *Chinese Journal of Information on Traditional Chinese Medicine*, vol. 6, pp. 563-564, 2004.