Complex Network Analysis of Classical Chinese Literature: Insights from the Confucian Canon of Scripta Sinica Corpus

Yichi Zhang u7748799

May, 2024

Abstract

The study of language as a complex system has attracted a lot of attention recently due to the quick development and widespread use of NLP and LLM in real-world. Network science offers a strong framework for detecting the structure of linguistic systems. Chinese, one of the most significant and unique languages in the world, has been studied by numerous academics. Nonetheless, while being the ancestor of Modern Chinese, Classical Chinese has received little attention, most likely because of corpus limitations. In this project, we investigate the topological characteristics of Classical Chinese literature using complex network analysis, concentrating on Scripta Sinica corpus. Through the construction of two distinct network types from the corpus, a Character Co-occurrence Network (CCN) and a Character-Sentence Network (CSN), we reveal a number of features that provide insight into the organizing principles that underlie Classical Chinese literature. According to our research, both networks have hierarchical organization, disassortative mixing patterns, small-world characteristics, and scale-free degree distributions, all of which are in line with findings in other language networks. This work not only advances our knowledge of Classical Chinese literature from the standpoint of complex systems, but it also emphasizes the usefulness of network science in revealing relationships and hidden patterns in textual data.

1 Introduction

Complex network studies have made it possible to analyze and comprehend the dynamics and structure of different real-world systems, such as biological, social, and technical networks **?????**. And there is no exception in human language network.

By representing language as a network, where words or characters are nodes and their relations are edges **??**, researchers have found statistical features in language networks such as scale-free degree distributions, small-world effects, and hierarchical structures. These findings have not only deepened our understanding of the complexity of human language but have also provided us with new insights into language learning and processing **?**.

Given that it is the existing language with the longest history and the largest speaker population in the world, Chinese offers a good case for the academic study of language networks. Chinese is an ideal language to study the connection between complex network theory and linguistic principles because of its large vocabulary, special reading and writing system, and rich literary history.

However, a thorough examination of the topological characteristics of Classical Chinese language networks based on sizable corpora is still absent, in spite of some groundbreaking research on Modern Chinese networks **??**.

We investigate the statistical characteristics of Chinese language networks built from the Scripta Sinica corpus, one of the largest digital collections of ancient Chinese writings, in order to close this gap. Two types of networks are constructed: the character co-occurrence network (CCN) and the character-sentence network (CSN). Our goal is to identify the network features of the Classical Chinese language and any possible linguistic or cultural conclusions by combining quantitative investigation and qualitative interpretation.

This project makes two primary contributions. Firstly, we present a thorough examination of the degree distributions, small-world effects, hierarchical structures, disassortativity, and other pertinent aspects of the topological characteristics of Chinese language networks. In addition to confirming some general features of language networks, our results also point to several distinct patterns that might be indicative of the particular qualities of the classical Chinese language. Secondly, we present the possibilities of using network science to the quantitative analysis of classical Chinese literature and culture. We demonstrate how complex network theory may be used to digital humanities research by building and analyzing networks from extensive historical corpora, creating new opportunities for multidisciplinary cooperation.

This is how the remainder of the essay is structured. The Scripta Sinica corpus and the creation of CCN and CSN, as well as the data and methodologies employed in this investigation, are covered in Section 2. The primary findings of our investigation are shown in Section 3, with an emphasis on the statistical characteristics of the two networks. Section 4 focus on the dicsussions of our results, how they relate to previous research and the limitations of the dataset. The paper is finally concluded in Section 5, which also suggests options for future research.

2 Data and Methods

2.1 Scripta Sinica Corpus

The Scripta Sinica corpus, a comprehensive digital collection of ancient Chinese literature, is managed by the Taiwanese Academia Sinica **?**. From the pre-Qin era (before 221 BCE) to the Qing dynasty (1644–1912 CE), it contains an extensive range of historical works, including classical literature, historical accounts, philosophical writings, and scientific treatises. Every work in the corpus has been sorted into a distinct subset.

In this study, we focus on a subset of the Scripta Sinica corpus, specifically the Confucian Canon (Ruzang, 儒藏). This subset, notated as SS-CC, contains more than three hundred classical works of Confucianism, covering annotations of Confucian classics by various literati in all eras. The SS-CC subset represents the canonical works of Classical Chinese literature and provides a suitable dataset for studying the language networks of Classical Chinese.

The text in SS-CC is not consistent in various aspects such as traditional and simplified Chinese,

punctuation, and so on. Before we start building the network, we need to preprocess the text. Traditional characters are converted to simplified characters to maintain consistency, and then all unpunctuated text is punctuated, preprocessed, and segmented using the Jiayan model, an NLP tool developed specifically for processing Classical Chinese.

2.2 Network Construction

We construct two types of language networks from the SS-CC corpus, namely the character co-occurrence network (CCN) and the character-sentence network (CSN). Both networks are represented as undirected and unweighted graphs, with nodes representing Chinese characters and edges indicating their co-occurrence relationships.

2.2.1 Character Co-occurrence Network (CCN)

In CCN, two characters are connected by an edge if they appear adjacent to each other in at least one text in the corpus. Specifically, for each text in SS-CC, we first extract all the unique characters and then create edges between characters that are adjacent in the original text. The edges are then accumulated across all texts to form the final network. Formally, the CCN can be defined as a graph $G_1 = (V_1, E_1)$, where V_1 is the set of unique characters and E_1 is the set of edges representing adjacent co-occurrences. Figure 1 illustrates the construction of CCN using a simple example. Given a

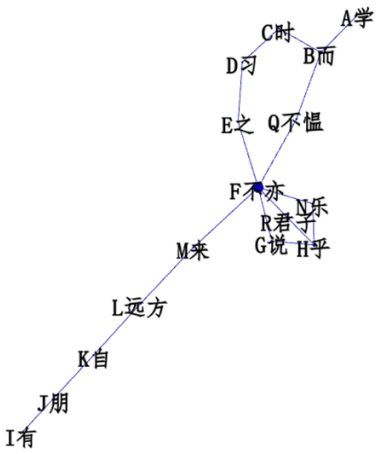

Figure 1: An example of constructing the character co-occurrence network (CCN) from a short text.

short text ”学而时习之,不亦说乎? 有朋自远方来,不亦乐乎? 人不知自而不愠,不亦君子乎?” (Isn’t it a pleasure to study and practice what you have learned? Isn’t it delightful to have friends coming from distant quarters? Isn’t it characteristic of a noble person to be unaffected when others misunderstand them?), we first extract the unique characters 学,而,时,习,之,不亦,说,乎,有,朋,自,远方,来,不亦,乐,乎,人,不知,而,不愠,不亦,君子,乎 as nodes, and then create edges between adjacent characters, such as (学,而), (而,时), (时,习), (习,之), and so on. The resulting subgraph is then merged into the global CCN.

2.2.2 Character-Sentence Network (CSN)

In CSN, two characters are connected by an edge if they co-occur in the same sentence, regardless of their positions. For each sentence in the SS-CC corpus, we create a complete subgraph among all the unique characters in the sentence, and then merge these subgraphs to form the final network. Formally, the CSN can be defined as a graph $G_2 = (V_2, E_2)$, where V_2 is the set of unique characters and E_2 is the set of edges representing sentence-level co-occurrences. Figure 2 shows an example of constructing CSN

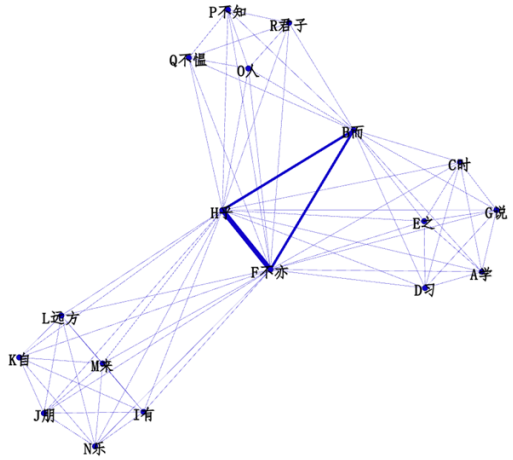

Figure 2: An example of constructing the character-sentence network (CSN) from a set of sentences.

from three sentences: ”学而时习之,不亦说乎?” (Isn’t it a pleasure to study and practice what you have learned?), ”有朋自远方来,不亦乐乎?” (Isn’t it delightful to have friends coming from distant quarters?) and ”人不知自而不愠,不亦君子乎?”(Isn’t it characteristic of a noble person to be unaffected when others misunderstand them?). For each sentence, a complete subgraph is created among its characters. These subgraphs are then merged to form the global CSN.

The basic statistics of the two networks are summarized in Table 1 and the statistics of corresponding random networks are in Table 2. CCN contains 4,253 nodes and 3,119,616 edges, with an average degree of 39.77. CSN is smaller but denser, with 74,837 nodes and 2,766,801 edges, and an average degree of 73.96. Due to arithmetic limitations, we are unable to directly compute the clustering coefficients and average shortest path lengths (When the network is not connected, it is not possible to calculate the average shortest path length. So we calculate the average shortest path length of only the largest connected subgraph of the network), that is why we use sampling to calculate. In the table following, we uniformly set the sampling ratio to 0.01 and the number of samples to 100.

2.3 Network Analysis Methods

To analyze the constructed networks, we employ a range of network measures and techniques commonly used in complex network analysis. These include:

- Node Strength: Although we analyze unweighted networks, we also introduce and analyze the

Table 1: Basic statistics of CCN and CSN

Graph	CCN	CSN
N (Number of nodes)	156879	74837
E (Number of edges)	3119616	2766801
$\langle k \rangle$ (Average degree)	39.77	73.96
C (Clustering coefficient)	0.022	0.046
L (Average shortest path length)	2.41	2.74

Table 2: Basic statistics of control group

Property	CCN	CSN
N (Number of nodes)	156879	74837
E (Number of edges)	3119616	2766801
$\langle k \rangle$ (Average degree)	39.77	73.96
C (Clustering coefficient)	2.83×10^{-5}	1.20×10^{-4}
L (Average shortest path length)	3.66	2.93

concept of node strength. Node strength is related to edge weights and represents the sum of the weights of all edges connected to a node. If we define the weight of an edge between two nodes as the frequency with which they appear adjacent to each other in CCN or simultaneously in a sentence in CSN, then the strength of a node corresponds to its total word frequency. By analyzing node strength, we discover a nontrivial power-law relationship between node strength s and degree k: $s \sim k^{\beta}$. This correlation demonstrates the existence of the "rich get richer" phenomenon in the network, which may account for the power-law degree distribution observed in the unweighted networks.

- Degree distribution: We examine whether the degree distributions of CCN and CSN show scale-free behavior, a characteristic that many real-world networks share. Scale-free networks are characterized by a power-law degree distribution, wherein a small number of nodes, called hubs, have a large number of connections, while the bulk of nodes have relatively few connections. Power-law distributions have a major impact on information transmission and network resilience and are an indication of heterogeneous network architecture.

- Clustering coefficient: And we look at is the networks' clustering coefficient, which indicates how likely it is for nodes to form clusters. We may determine if the networks have a hierarchical structure—that is, node clusters arranged in a multi-level manner—by examining the correlation between the clustering coefficient and node degree. Within language networks, the presence of a hierarchical structure implies the existence of meaningful sub-units.

- Average shortest path length: To evaluate the efficiency of information transfer and the overall connectivity of the networks, we calculate the average shortest path length. The small-world effect, characterized by small average shortest path length relative to the network size, is another feature of many real-world networks. High local clustering and short global distances are characteristics of small-world networks that facilitate effective communication within the network.

- Degree correlations: In addition, we investigate the average nearest neighbor degree and degree-degree correlations in order to better understand the degree correlations in CCN and CSN. The results of this study show whether the networks show assortative or disassortative mixing patterns, which represent the inclination of nodes to form connections with other nodes of correspondingly different degrees. Disassortative mixing says that high-degree nodes prefer to link with low-degree nodes, whereas assortative mixing argues that nodes with high degrees like to connect with other high-degree nodes. Understanding the fundamental structure and evolutionary history of the language networks can be gained from the mixing patterns.

The network analysis is performed using Python and popular network analysis libraries NetworkX. As mentioned earlier, due to computational limitations, we employed a sampling approach when calculating the relationship between the network clustering coefficient and degree, the average shortest path length of the network, and the relationship between the average nearest neighbor degree and degree. This approach allows us to maintain the overall network properties while reducing the computational burden and improving efficiency.

3 Models and Results

This section presents the main findings of our study on the topological properties of CCN and CSN. We focus on the small-world properties, scale-free characteristics, disassortativity and hierarchical structure. These results provide valuable insights into the organizational principles and structural features of Chinese language networks.

3.1 Small-World Properties

Compared to a random network with same scale, if a specific network maintains a similar average shortest path length and has a larger clustering coefficient, it is considered to have small world properties **?**. Therefore, we investigated the clustering coefficient, and average shortest path length of CCN and CSN to assess their small world properties. Compared to an equivalent random network ($C_{\text{rand}} = 2.83 \times 10^{-5}, L_{\text{rand}} = 3.66$), CCN has a significantly higher clustering coefficient $C_1 = 0.022$ and a slightly lower average shortest path length $L_1 = 2.41$. Similarly, CSN has a higher $C_2 = 0.046$ and a lower $L_2 = 2.74$ ($C_{\text{rand}} = 1.20 \times 10^{-4}, L_{\text{rand}} = 2.93$). These findings suggest that Classical Chinese lexical networks have strong small-world effects, making information processing and retrieval easier due to their highly connected local clusters and efficient global communication.

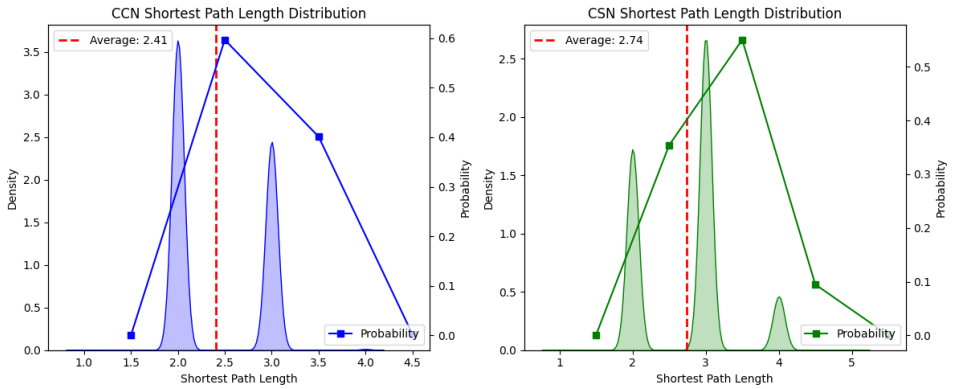

Figure 3: Shortest path length distributions of CCN and CSN. The average shortest path lengths are 2.41 for CCN and 2.74 for CSN.

3.2 Scale-Free Characteristics and Matthew Effect

Many real-world networks exhibit a power-law degree distribution, indicating a scale-free structure **?**:

$$P(k) \sim k^{-\gamma} \tag{1}$$

The degree distribution $P(k)$ represents the probability that a randomly selected node in the network

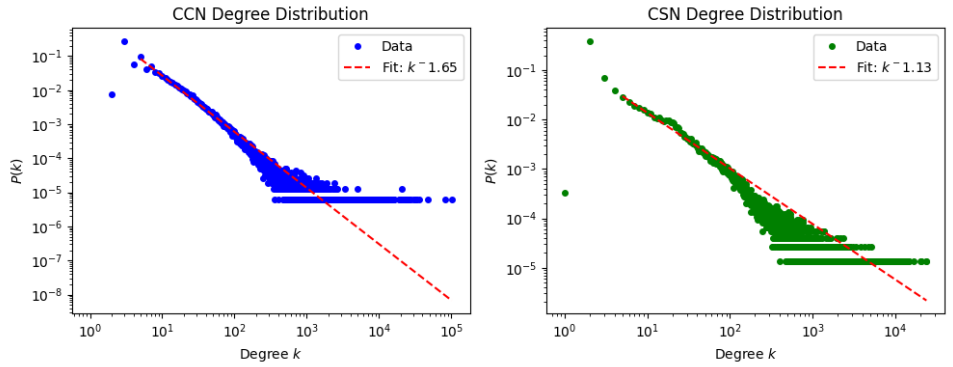

Figure 4: Degree distributions of CCN and CSN on a double logarithmic scale. The power-law exponents are 1.65 for CCN and 1.13 for CSN.

has degree k. Figure 4 shows the degree distributions of CCN and CSN on a double logarithmic scale. Both networks exhibit power-law tails, indicating a scale-free structure. The scale-free nature suggests that a small number of highly connected characters (e.g., function words and common content words, which are usually single characters rather than word combinations in Classical Chinese) play a crucial role in maintaining the connectivity of the Classical Chinese language system. To further investigate the relationship between node importance and connection strength, we analyzed the node strength distribution and its correlation with degree. Figure 5 shows the node strength distributions of CCN and CSN on a double logarithmic scale. Both networks exhibit heavy-tailed distributions, indicating a highly heterogeneous allocation of connection strengths among nodes. Moreover, we observed a super-linear

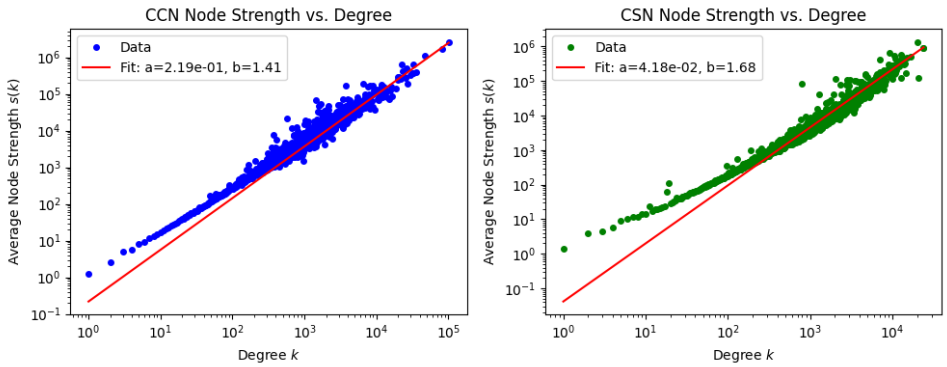

Figure 5: Average node strength $s(k)$ vs. degree k for the CCN and CSN networks on a double logarithmic scale. The data is fitted to a power law $s(k) = ak^b$. For CCN, $a = 2.19e-01$ and $b = 1.41$, indicating a super-linear scaling of node strength with degree. For CSN, $a = 4.18e-02$ and $b = 1.68$, also showing a super-linear scaling but with a higher exponent compared to CCN.

relationship between node strength $s(k)$ and degree k (Figure 5):

$$s(k) \sim k^{\alpha} \tag{2}$$

For CCN, the fitted exponent is $\alpha = 1.41$, while for CSN, $\alpha = 1.68$. This result suggests that high-degree nodes tend to have disproportionately larger strengths compared to low-degree nodes, reflecting the preferential attachment mechanism commonly observed in evolving networks and pointing to a "Matthew effect" or "rich-get-richer" phenomenon **?**.

3.3 Hierarchical Structure and Disassortativity

By observing the power-law relationship between the clustering coefficient $C(k)$ and the node degree k **?**, we can examine whether a complex network possesses a hierarchical structure:

$$C(k) \sim k^{-\beta} \tag{3}$$

The hierarchical structures of CCN and CSN reflect the multi-scale, modular organization of the Chinese language system, as shown in Figure 6, with $\beta = 0.55$ for CCN and $\beta = 0.34$ for CSN. There are tightly connected groups at various linguistic unit levels, from radicals to words and phrases. In addition, we studied the degree correlations in CSN and CCN to understand the assortativity of the networks. Disassortative mixing is the tendency of high-degree nodes to connect with low-degree nodes, measured by a negative degree correlation coefficient r **?**.

Based on the Pearson correlation coefficient, $r_1 = -0.2702$ for CCN and $r_2 = -0.6054$ for CSN. Our findings confirm the disassortative nature of CCN and CSN (Figure 7):

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2} \tag{4}$$

High-degree function words typically associate with low-degree content words, a phenomenon known as disassortativity, which reflects syntactic and semantic constraints in language organization. Moreover,

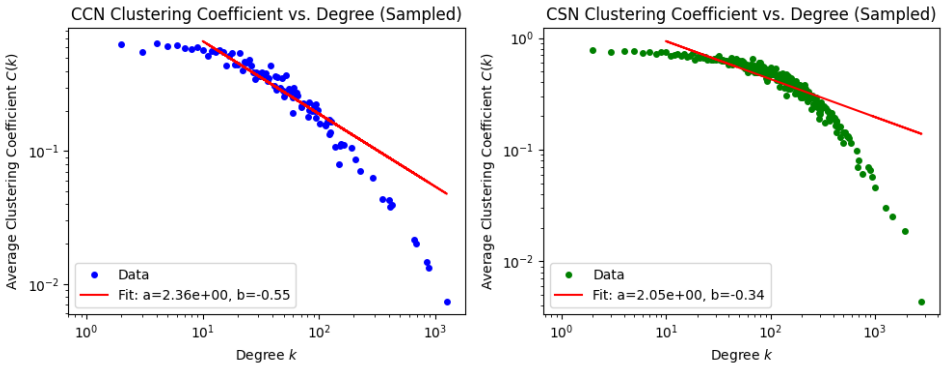

Figure 6: Average clustering coefficient $C(k)$ vs. degree k for the CCN and CSN networks on a double logarithmic scale, using sampled data. The data is fitted to a power law $C(k) = ak^b$. For CCN, $a = 2.36e{+}00$ and $b = -0.55$, indicating a sub-linear decrease of clustering coefficient with degree. For CSN, $a = 2.05e + 00$ and $b = -0.34$, also showing a sub-linear decrease but with a slower decay compared to CCN.

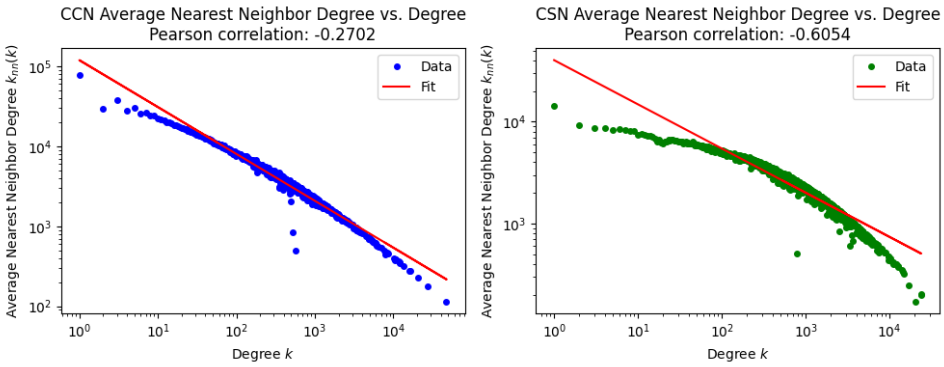

Figure 7: Degree correlations in CCN and CSN. The Pearson correlation coefficients are -0.2702 for CCN and -0.6054 for CSN.

it may enhance the robustness of the language network against errors and attacks **??**. In summary, our study shows that CCN and CSN exhibit a range of complex network characteristics, including hierarchical organization, small-world effect, scale-free structure, and disassortative mixing. These features provide a quantitative description of the Chinese language system and shed new light on its dynamics, structure, and function.

4 Discussion

Our complex network analysis of the character co-occurrence network (CCN) and the character-sentence network (CSN) constructed from the Scripta Sinica corpus provides novel insights into the organizational principles and statistical regularities of the classical Chinese language system. The results reveal several non-trivial topological properties, including the scale-free structure, small-world effect, hierarchical organization, and disassortative mixing, which are consistent with the findings of previous studies on modern Chinese and other language networks **???**. These properties suggest that the Chinese language system, despite its unique features and diachronic evolution, shares some universal characteristics with other human languages, reflecting the common cognitive and communicative constraints in language processing and acquisition.

Comparing our findings with the modern Chinese language networks analyzed in the paper **?**, we observe many similarities. Scale-free degree distributions, small-world features, and disassortative mixing patterns are present in both classical and contemporary Chinese networks. The power-law degree distribution of the scale-free structure suggests the existence of strongly connected hub characters or key words in the language system. The small-world effect, which has short average path lengths and substantial local clustering, points to an effective structure for processing and retrieving information. High-degree nodes typically connect with low-degree nodes in disassortative mixing, which is a reflection of the syntactic and semantic limitations in language use. These shared characteristics show how the Chinese language system is both structurally sound and functionally ideal throughout time.

These can be further understood from a language evolution perspective. Grammar, vocabulary, and writing system simplifications and regularizations occurred throughout the Chinese language's transition from the classical to the modern forms **?**. But the fundamental organizational ideas—like small-world and scale-free properties—have mostly held true, preserving the harmony between stability and adaptability in language processing and communication.

At the same time, we also observe some differences between the classical and modern Chinese networks. The statistical structure of language networks has changed due to the advent of vernacular Chinese, the standardization of character usage, and the influence of other languages. Based on linguistic theories, we have several speculations about the statistical structure of Classical Chinese texts.

First, due to the concise and highly contextualized nature of classical Chinese, where characters are more closely related to each other within and across sentences, Classical Chinese networks would be likely to have higher clustering coefficients and shorter average path lengths than their modern counterparts. Second, syntactic rules and semantic categories in classical Chinese are more rigid compared to the modern vernacular, implying a stricter separation between function and content characters, so that

the disassortative mixing pattern is likely to be more pronounced in the classical Chinese networks.

However, not all inferences can be empirically verified through data analysis.

We can demonstrate the more pronounced disassortative mixing pattern in the classical Chinese networks by comparing the Pearson correlation coefficients: -0.2702 for CCN and -0.6054 for CSN, which are significantly larger in magnitude than -0.0759 for CLN1 and -0.0707 for CLN2. But we cannot obtain the result that the clustering coefficient is higher and the average path length is shorter in the classical Chinese networks.

The indicators that do not align with existing linguistic understanding may have several explanations:

1. Our study is based on a limited corpus of classical Chinese texts and is not comprehensive. Additionally, statistical result (Total number of characters: 36,762,062, Total number of Chinese characters: 27,868,060, Total number of noise characters: 160,150, Total noise ratio: 0.44%) shows a significant amount of noise in the texts, which could be a source of bias.

2. Due to technical limitations, we used sampling methods for calculating multiple indicators. Although these methods can yield relatively reliable results, they are still less complete compared to full-scale calculations.

3. During text preprocessing, we were constrained by the lack of a robust NLP model for classical Chinese. Consequently, we could not create a network that accurately reflects classical Chinese word segmentation habits, leading to inaccuracies in network construction.

5 Conclusion

This study presents a comprehensive complex network analysis of the Classical Chinese language based on the character co-occurrence network (CCN) and the character-sentence network (CSN) constructed from the Scripta Sinica corpus. Our research sheds light on the statistical regularities and organizational principles of the Classical Chinese language system by revealing a variety of non-trivial topological features of the language networks. The following succinctly describes the primary findings of our investigation:

1. The power-law degree distribution of the Classical Chinese language networks indicates a scale-free structure. This suggests that most characters have relatively low degrees, whereas a tiny number of strongly related characters that are essential to the language system exist.

2. The networks display small-world properties, with high local clustering coefficients and short average path lengths. This indicates an efficient organization of the language system for information processing and retrieval, striking a balance between local specialization and global integration.

3. The networks possess a hierarchical modular structure, as evidenced by the power-law relationship between clustering coefficient and degree. This reflects the multi-scale compositional nature of the language system, with characters forming tightly connected clusters at different linguistic levels.

4. The networks show disassortative mixing patterns, where high-degree characters tend to connect with low-degree characters. This property is related to the syntactic and semantic constraints in language organization and may contribute to the robustness and flexibility of the language system.

5. It appears that some organizing principles of human language are universal because the topological characteristics of the Classical Chinese language networks are substantially compatible with those observed in modern Chinese and other language networks. Higher clustering coefficients, shorter path lengths, and more pronounced hierarchical and disassortative structures are some of the major distinctions seen in the Classical Chinese networks, which may be a reflection of the language's distinctive characteristics and evolutionary changes.

These findings offer new insights into the cognitive and communicative mechanisms underpinning language processing and acquisition, as well as the evolutionary and cross-linguistic patterns in human language. They also show how effective complex network theory is at quantitatively characterizing the structure and dynamics of the Chinese language system.

However, it is important to acknowledge the limitations of our study and consider them for future research. First, although extensive, the Scripta Sinica corpus only covers a specific historical period and genre of Classical Chinese literature. The generalizability of our findings to other periods, styles, and registers of the language requires further investigation. Second, our network construction methods, based on character co-occurrence and sentence structure, only capture certain aspects of the language system. Other dimensions of language, such as semantics, pragmatics, morphology, and phonology, are not explicitly represented in the current networks. Third, our analysis focuses primarily on the statistical properties of the language networks, without directly examining the cognitive and social processes underlying language use and evolution. Integrating the network approach with experimental, computational, and historical methods would provide a more comprehensive understanding of the language system.

Despite these limitations, our study opens up new avenues for future research on the complex network analysis of Chinese and other languages. Some promising directions include:

1. Expanding the data sources to include more diverse and representative corpora of Classical and modern Chinese, as well as other languages and dialects, to investigate the universality and specificity of language network properties.

2. Developing more sophisticated network construction methods and analytical tools, such as weighted, directed, and multiplex networks, to capture richer linguistic information and dynamics.

3. To find typological and evolutionary patterns in language networks, conducting methodical comparisons and statistical assessments of network features across various linguistic domains, historical eras, and genres.

4. Combining language acquisition, processing, and change theories and methodologies with network analysis and other linguistic, cognitive, and social theories and methodologies to create a more thorough and explicative framework for comprehending the intricate system of human language.

To conclude, our intricate network analysis of the networks of Classical Chinese languages built from the Scripta Sinica corpus uncovers a wealth of structural features that offer fresh perspectives on the statistical regularities and organizing principles of the language system. Through quantitative characterization and comparison of the structural and functional aspects of human language, the study shows the potential of complex network theory. It also creates new avenues for interdisciplinary research

on the cognitive, communicative, and evolutionary bases of language diversity and complexity.

Acknowledgments

The data for this study were sourced from the Scripta Sinica database(https://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm), supported by the Institute of History and Philology, Academia Sinica(https://www1.ihp.sinica.edu.tw/). We would like to thank @mahavivo(https://github.com/mahavivo/scripta-sinica) on GitHub and "Daizhige(殆知阁)" for their organization and open sourcing of the documents. We also extend our gratitude to @jiaeyan for providing the classical Chinese text processing model, Jiayan(甲言)(https://github.com/jiaeyan/Jiayan).

Code and Data Availability

The code and data used in this study are available on GitHub at https://github.com/WhiteMasky/Complex-Network-Analysis-of-Classical-Chinese-Literature

Ethical Considerations and Broader Impacts

References

Appendices