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

1 Introduction

2 Data and Methods

2.1 Scripta Sinica Corpus

2.2 Network Construction

2.2.1 Character Co-occurrence Network (CCN)

2.2.2 Character-Sentence Network (CSN)

2.3 Network Analysis Methods

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

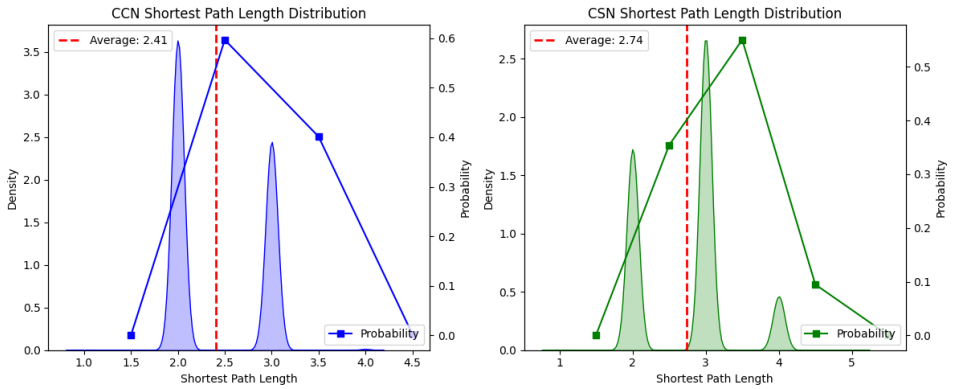

Figure 1: Shortest path length distributions of CCN and CSN. The average shortest path lengths are 2.41 for CCN and 2.74 for CSN.

lexical networks have strong small-world effects, making information processing and retrieval easier due to their highly connected local clusters and efficient global communication.

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

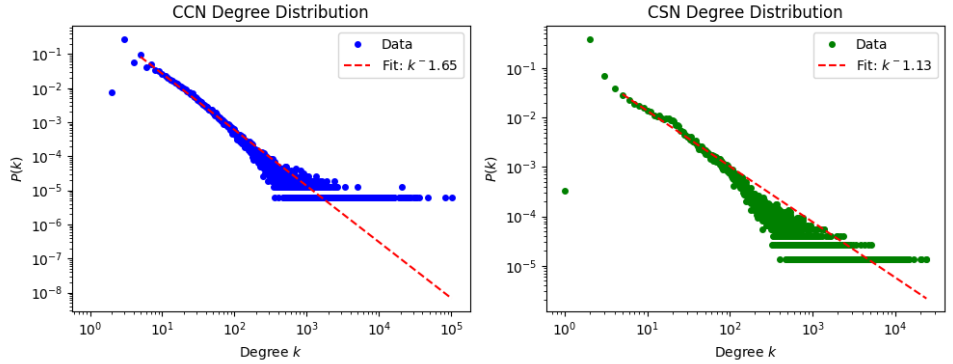

Figure 2: Degree distributions of CCN and CSN on a double logarithmic scale. The power-law exponents are 1.65 for CCN and 1.13 for CSN.

has degree k. Figure 2 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 3 shows the node strength distributions of CCN and CSN

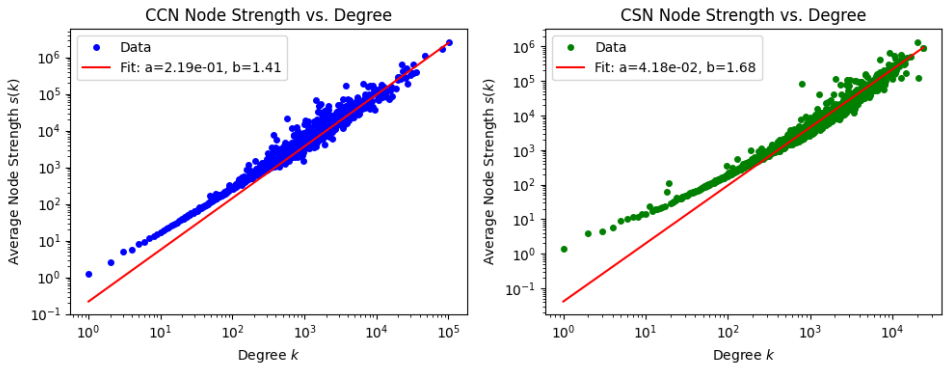

Figure 3: Average node strength $s(k)$ vs. degree k for the CCN and CSN networks on a double logarithmic scale. The data is fitted to a power law $s(k) = ak^b$. For CCN, $a = 2.19e-01$ and $b = 1.41$, indicating a super-linear scaling of node strength with degree. For CSN, $a = 4.18e-02$ and $b = 1.68$, also showing a super-linear scaling but with a higher exponent compared to CCN.

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 relationship between node strength $s(k)$ and degree k (Figure 3):

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

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 4, 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 5):

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

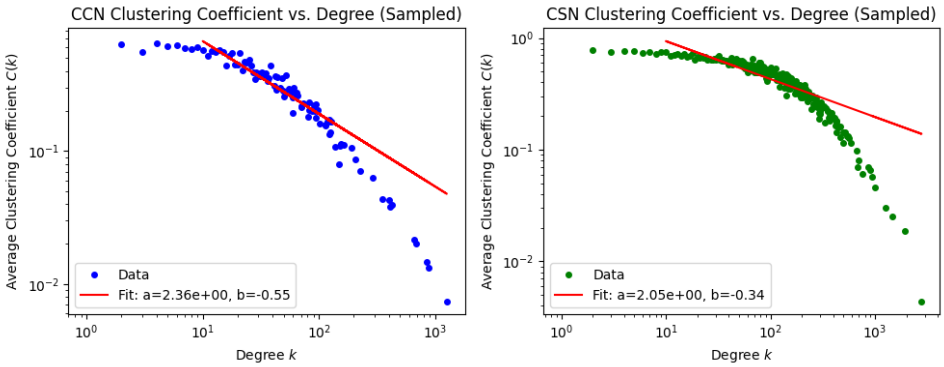

Figure 4: Average clustering coefficient $C(k)$ vs. degree k for the CCN and CSN networks on a double logarithmic scale, using sampled data. The data is fitted to a power law $C(k) = ak^b$. For CCN, $a = 2.36e{+}00$ and $b = -0.55$, indicating a sub-linear decrease of clustering coefficient with degree. For CSN, $a = 2.05e + 00$ and $b = -0.34$, also showing a sub-linear decrease but with a slower decay compared to CCN.

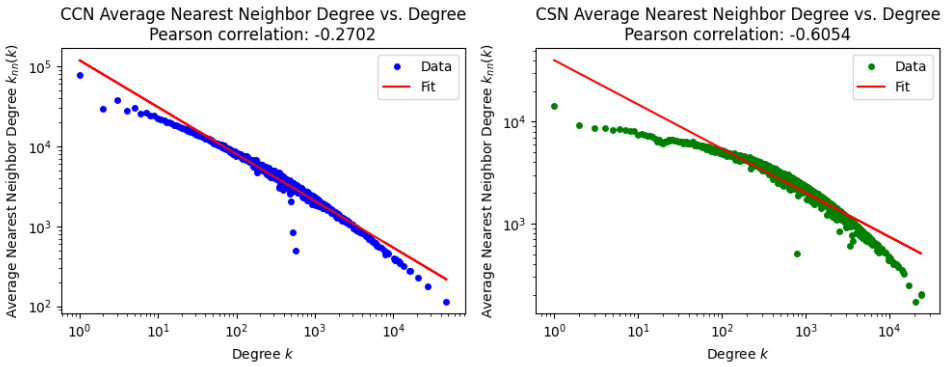

Figure 5: Degree correlations in CCN and CSN. The Pearson correlation coefficients are -0.2702 for CCN and -0.6054 for CSN.

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, 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

5 Conclusion

Acknowledgments

Ethical Considerations and Broader Impacts

References

Appendices

Supplementary Materials

Network Visualization

Statistical Tests

Code and Data Availability