Physica A 450 (2016) 49–56

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Spectral analysis of Chinese language: Co-occurrence networks from four literary genres

Wei Liang [a,*], Guanrong Chen [b]

[a] School of Mathematics and Information Science, Henan Polytechnic University Jiaozuo, Henan, 454000, China
[b] Department of Electronic Engineering, City University of Hong Kong, Hong Kong Special Administrative Region

HIGHLIGHTS

- 408 Chinese networks' spectra and principal eigenvectors are computed.
- The largest eigenvalue depends on N, E, L, and C as scale free, respectively.
- The number of different eigenvalues is $\propto \log N$ for novel, while $\propto N$ for the others.
- A triangle or an "M" shape appears in each of the incorporated networks' spectral densities.
- The principal eigenvector is localized to the node with the largest degree.

ARTICLE INFO

Article history:
Received 26 October 2015
Available online 11 January 2016

Keywords:
Chinese language
Co-occurrence network
Spectral analysis
Adjacency matrix

ABSTRACT

The eigenvalues and eigenvectors of the adjacency matrix of a network contain essential information about its topology. For each of the Chinese language co-occurrence networks constructed from four literary genres, i.e., essay, popular science article, news report, and novel, it is found that the largest eigenvalue depends on the network size N, the number of edges, the average shortest path length, and the clustering coefficient. Moreover, it is found that their node-degree distributions all follow a power-law. The number of different eigenvalues, N_λ, is found numerically to increase in the manner of $N_\lambda \propto \log N$ for novel and $N_\lambda \propto N$ for the other three literary genres. An "M" shape or a triangle-like distribution appears in their spectral densities. The eigenvector corresponding to the largest eigenvalue is mostly localized to a node with the largest degree. For the above observed phenomena, mathematical analysis is provided with interpretation from a linguistic perspective.

1. Introduction

Human languages can be viewed as a complex system emerged from a long-time evolution. The study of language networks has been quite popular in recent years, which provides us with some interesting insight into the nature of languages.

Most of the existing studies on language networks, for example co-occurrence, syntax and semantic frameworks [1–7], focus on the static structural properties, such as degree distribution, shortest path length and clustering coefficient. On the other hand, spectral analysis is a powerful tool capable of revealing the global structural patterns underlying a large-scale complex environment of interacting entities. It refers to the systematic study of eigenvalues and eigenvectors of the adjacency matrix of a network [8]. Spectral methods based on the analysis of the largest eigenvalues and their corresponding

* Corresponding author.
E-mail address: wliang@hpu.edu.cn (W. Liang).

http://dx.doi.org/10.1016/j.physa.2015.12.121
0378-4371/

eigenvectors have proven successful in detecting communities and in characterizing synchronization of scale-free dynamical networks [9,10]. In fact, spectral analysis has been successfully employed in the analysis of technological, biological, and social networks [11–13].

Although spectral properties have not been widely investigated for language networks, there are some reports in the literature. Belkin and Goldsmith analyzed the eigenvectors of the Laplace matrices to obtain a two-dimensional visualization of network models for English and French languages in 2002 [14]. In 2007, Cancho et al. applied spectral methods to community detection in a syntactic dependency network, and found that the spectral methods can cluster words of the same class [15]. In 2009, Mukherjee et al. investigated spectral densities and eigenvalue distributions of phonetic networks [8]. In 2010, Choudhury et al. investigated the spectral densities of word co-occurrence networks for seven different languages, and showed that they have triangle-like shapes [16]. Recently, we investigated the spectral densities of 606 Chinese character and word co-occurrence networks and 404 English word co-occurrence networks, constructed from Chinese and English poems, respectively [17,18]. We found that "M" shape eigenvalue distributions appear in the spectral densities of 1007 networks, while the other 3 spectral densities of the incorporated Chinese character networks exhibit triangle-like shapes, which are similar to that of the BA network [17,18]. Is the "M" shape eigenvalue distribution accidental? Will this interesting phenomenon appear in other Chinese literary genres? Can useful information be drawn from the largest eigenvalues and theirs corresponding eigenvectors of the adjacency matrices in Chinese language networks? The present paper attempts to address these interesting questions.

The rest of the paper is structured as follows. Section 2 introduces some concepts involved in spectral analysis. Section 3 contains mathematical analysis and numerical simulations on the spectra and a special eigenvector: the largest eigenvalue in Section 3.1, the spectral distribution in Section 3.2, the spectral density in Section 3.3, and the principal eigenvector in Section 3.4. Finally, some conclusions are drawn from the above investigations.

2. Basic concepts

First, some basic concepts involved in spectral analysis are briefly reviewed.

For an undirected network consisting of N nodes, its *adjacency matrix* A is defined by $(a_{ij})_{N\times N}$, where $a_{ij} = 1$ if nodes i and j are connected via an edge and $a_{ij} = 0$ otherwise. λ is an *eigenvalue* of A if there is an N-dimensional nonzero vector x such that $Ax = \lambda x$. A is real and symmetric, so all eigenvalues are real and the largest eigenvalue is not degenerate. Any $N \times N$ real symmetric matrix A has N (possibly non-distinct) real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ with N corresponding eigenvectors that are mutually orthogonal [19]. The *spectrum* of a network is the set of all the eigenvalues of its adjacency matrix, denoted by $\{\lambda_1^{[n_1]}, \lambda_2^{[n_2]}, \ldots, \lambda_k^{[n_k]}\}$, where $[n_i]$ is the multiplicity of λ_i. *Spectral density* $\rho(\lambda)$ is defined as the probability that a randomly chosen eigenvalue of the adjacency matrix is λ [12].

3. Main results

A total of 200 Chinese articles of essay, popular science article, news report and novel were collected, 50 for each. In a character (word) co-occurrence network, nodes are characters (words); two characters (words) are connected by an edge if they are adjacent to each other in a sentence. Using this method, character and word co-occurrence networks (denoted by C-network and W-network, respectively) are constructed from each single article. For the four types of articles, C-networks and W-networks were also constructed by incorporating together 50 articles of the same type. For convenience, networks constructed from single and incorporated articles are denoted as S-networks and I-networks, respectively.

In this section, basic parameters especially the largest eigenvalues (λ_1) and their corresponding eigenvectors, the second largest eigenvalues (λ_2), the smallest eigenvalues (λ_N), and the number of different eigenvalues (N_λ) of the above-mentioned networks are computed, thereby the average values of the S-networks and that of the I-networks are obtained as summarized in Table 1, wherein and throughout, PS stands for popular science.

3.1. The largest eigenvalue

It is found that the largest eigenvalues, λ_1, of the S-networks fall into the ranges of 8.06–15.57 (essay), 6.08–13.54 (PS), 4.93–11.78 (news), 8.23–35.02 (novel) for the C-networks, and 6.76–14.61 (essay), 4.77–12.41 (PS), 4.61–10.54 (news), 7.36–30.23 (novel) for the W-networks. By inspecting the data in Table 1, it is found that (1) the λ_1 values of the C-networks are larger than those of the W-networks in each type of literary genres; (2) $\lambda_1 \gg \sqrt{N-1}$ for the incorporated C-networks, while $\lambda_1 \ll \sqrt{N-1}$ for the others, and for a fully-connected and a star network with N nodes, the largest eigenvalue was found to be $N-1$ and $\sqrt{N-1}$, respectively in Ref. [20], but these special networks might exist as subnetworks in our networks herein; (3) novel has the largest λ_1, while for news it is the opposite, which is perhaps due to the fact that novel has the largest E/N; (4) the value of λ_1 is significantly larger than the values of λ_2 and $|\lambda_N|$, which is particularly apparent in the C-networks. In Ref. [21], it is reported that if the first few eigenvalues of a matrix are much larger than the rest of the eigenvalues, then the corresponding graph has a few subgraphs to be repeated for a large number of times in order to obtain the global structure of the graph [21]. Interestingly, this phenomenon is observed across all the networks