Physica A 420 (2015) 284–293

Contents lists available at ScienceDirect

Physica A

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

Co-occurrence network analysis of modern Chinese poems

Wei Liang [a,*], Yanli Wang [b], Yuming Shi [b], Guanrong Chen [c]

[a] School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, Henan 454000, China

[b] Department of Mathematics, Shandong University Jinan, Shandong 250100, China

[c] Department of Electronic Engineering, City University of Hong Kong, Hong Kong Special Administrative Region

HIGHLIGHTS

- 98.5% of networks have scale-free properties.
- 19.8% of networks do not have small-world features, especially the clustering coefficients in 5.6% of networks are zero.
- 61.4% of networks have significant hierarchical structures.
- 98% of networks are disassortative.
- "M" shape distributions appear in the spectral densities of 603 networks, while the other 3 spectral densities are similar with that of the BA network.

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form 16 October 2014
Available online 13 November 2014

Keywords:
Chinese language
Poem
Co-occurrence network
Hierarchical structure
Spectral analysis

ABSTRACT

A total of 606 co-occurrence networks of Chinese characters and words are constructed from rhymes, free verses, and prose poems. It is found that 98.5% of networks have scale-free properties, while 19.8% of networks do not have small-world features, especially the clustering coefficients in 5.6% of networks are zero. In addition, 61.4% of networks have significant hierarchical structures, and 98% of networks are disassortative. For the above observed phenomena, analysis is provided with interpretation from a linguistic perspective.

1. Introduction

Human languages can be viewed as a complex system that emerged from a long-time evolution in communications. The study of linguistic networks has provided us with interesting insights into the nature of languages. Examples include co-occurrence, syntax, semantics, and conception frameworks [1–8]. However, there was not much research from the complex-network approach about poems, although it is one of the most important literary forms of human languages.

Poem originates from labor productivity, religious entertainment, and literary culture. After 1919, some Chinese poets combined the artistic techniques of poems existing in foreign languages and the ancient Chinese literature, to finally create the modern Chinese poems, which have three main styles: rhymes, free verses, and prose poems [9]. Can the writing characteristics of poems be reflected by the statistical parameters of their structural networks? Qifang He, a modern Chinese poet, said "poem contains a wealth of imagination and feelings, whereas its language differs from that of essays on the level of refining and harmony, especially in rhythm". If so, can one make use of the statistical parameters of networks to measure the similarities and differences between poems and other literary genres?

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

http://dx.doi.org/10.1016/j.physa.2014.11.001

It is well known that the spectrum of a network (see definition (10) in Section 2.1) can provide important information about its global structure. However, results on spectral analysis in linguistic networks are rarely found in the literature. In 2002, Belkin et al. studied the spectral properties of the Laplace matrices of some network models for English and French languages [10]. Mukherjee et al. analyzed spectral densities and eigenvalue distributions of phonetic networks [11]. In 2010, Choudhury et al. investigated the spectral densities of co-occurrence networks in English, French, and other languages [12]. Yet, as of today there are no spectral analysis results reported in the literature about Chinese language networks.

In this paper, 100 Chinese articles of rhymes, free verses, and prose poems are selected, respectively, from which 606 character and word co-occurrence networks are constructed. Degree distributions, small-world features, hierarchical organization, disassortative property, and spectral analysis of adjacency matrices of these networks are carefully studied. We found that using larger number of articles for study leads to the nearly the same conclusions, therefore 100 articles were used in this investigation for brevity. By comparing to the results in Ref. [7], we found that different literary genres lead to significant differences in their statistical parameters. We conclude that using statistical parameters of networks to measure the similarities and differences between poems and other literary genres is practical.

2. Basic concepts and co-occurrence construction

We first briefly review some basic concepts of networks, and then introduce the construction of co-occurrence networks.

2.1. Basic concepts

All the networks studied in this paper are undirected and unweighted.

(1) The *degree* of node i, denoted by k_i, is the number of edges adjacent to the node. The average degree of the network is denoted by $\langle k \rangle$.

(2) *Degree distribution* $p(k)$ is defined as the probability that a randomly-picked node in the network has exactly degree k. If $p(k)$ satisfies the power law $p(k) \propto k^{-\gamma}$, where γ is a positive constant, then the network is said to have a *scale-free* property in the sense that this power law is independent of the network size.

(3) Let d_{ij} be the shortest path length that connects two given nodes i and j. The *average shortest path length* of the network is defined as $L = 2\sum_{i>j} d_{ij}/N(N-1)$, where N is the total number of nodes in the network.

(4) The *clustering coefficient* C_i of node i is the probability that any two neighbors of node i are also connected to each other. Specifically, $C_i = 2E_i/k_i(k_i - 1)$, where E_i is the number of the existing edges among the neighbors of node i. The clustering coefficient of the whole network, denoted by C, is the average of C_i.

(5) A network is said to have a *small-world* property if $L \approx L_r$ and $C \gg C_r$, where L_r is the average shortest path length and C_r is the clustering coefficient of the ER random graph [13] with the same numbers of nodes and edges.

(6) *Average nearest-neighbor degree* $k_{nn}(k)$ is the average degree of nearest (namely, adjacent) neighbors for all nodes of degree k [14], i.e.,

$$k_{nn}(k) = \sum_{k'} k' P(k'|k),$$

where $P(k'|k)$ is the probability of nodes with degree k connecting to nodes with degree k'. If $k_{nn}(k) \propto k^{-\mu}$, then μ is called the assortative exponent.

(7) The assortativeness coefficient Γ of a network [15] is defined by

$$\Gamma = \frac{E^{-1}\sum_i j_i k_i - \left[E^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{E^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[E^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2},$$

where j_i, k_i are the degrees of the end nodes of edge i, respectively, and E is the total number of edges in the network. If $\Gamma > 0$, the network is *assortative*; if $\Gamma < 0$, it is *disassortative*.

(8) Let $C(k)$ denote the average clustering coefficient of all nodes with degree k. If

$$C(k) \propto k^{-\beta},$$

where $\beta > 0$, then the network has a *hierarchical organization* [16].

(9) For a graph G with N nodes, its *adjacency matrix* is defined by $(a_{ij})_{N\times N}$, where a_{ij} is 1 if nodes i and j are connected, and is 0 otherwise.

(10) The *spectrum* of network G is the set of all the eigenvalues of its adjacency matrix, denoted by $sp(G) = \{\lambda_1^{[n_1]}, \lambda_2^{[n_2]}, \ldots, \lambda_k^{[n_k]}\}$, where eigenvalues λ_i are listed in decreasing order, and $[n_i]$ is the multiplicity of λ_i.

(11) *Spectral density* $\rho(\lambda)$ is defined as the probability that a randomly chosen eigenvalue of the adjacency matrix is λ [17]. It can be proved that the $\rho(\lambda)$ of an ER random graph converges to a semicircle law (Fig. 1(a)); numerical simulations indicate that the $\rho(\lambda)$ of a BA scale-free network [18] has a triangle-like shape (Fig. 1(b)); while the $\rho(\lambda)$ of a WS small-world network [19] contains some singularities (Fig. 1(c)).