



论 文

# 不同时期汉语散文的字同现网络之研究

梁伟<sup>①②</sup>, 史玉明<sup>①\*</sup>, 谢智刚<sup>③</sup>, 王艳丽<sup>①</sup>

① 山东大学数学学院, 济南 250100

② 河南理工大学数学与信息科学学院, 焦作 454000

③ 香港理工大学电子及资讯工程学系, 香港

\* 通信作者. E-mail: ymshi@sdu.edu.cn

收稿日期: 2010-03-10; 接受日期: 2011-03-31

高等学校博士学科点专项科研基金 (批准号: 20100131110024)、香港理工大学研究基金 (批准号: 1-BBZA)、国家自然科学基金天元基金 (批准号: 11126120) 和河南理工大学博士基金 (批准号: B2011-032) 资助项目

**摘要** 基于汉语 4 个不同时期: 古代汉语、魏晋南北朝的汉语、近代汉语以及现代汉语的散文, 文中建立了字同现网络, 并研究了它们的统计特征. 研究发现 99.6% 的网络具有无标度特性, 95.0% 的网络具有小世界特性. 该研究从复杂网络的角度揭示了汉语语言的发展不是平稳的, 同时还揭示了不同时期中文文章的异同. 在汉语语言学的研究中, 对魏晋南北朝时期的文学是属于古代汉语还是近代汉语一直存在很大的争议. 研究结果表明, 魏晋南北朝时期网络的统计数据与其他时期的统计数据之间的确存在明显的不同, 且魏晋南北朝时期的文学属于近代汉语似乎更合理.

**关键词** 汉语 散文 字同现网络 小世界 无标度 语言学 非齐次网络 统计方法

## 1 引言

语言是人类智慧的结晶, 是一个经过漫长演化而形成的复杂系统<sup>[1]</sup>. 据估计, 目前世界上正在使用的主要语言有 6800 多种, 其中汉语是使用人口最多的语言<sup>[2]</sup>. 汉语网络, 如词同现网络<sup>[3~5]</sup>、句法网络<sup>[6]</sup>、语义网络<sup>[7]</sup>, 已经得到了广泛的研究, 发现它们都具有小世界特性和无标度特性中的一种或两种. 这些网络有一个共同的特点, 就是它们是由大量文章合在一起而建立的. 基于单篇文章我们也可以建立网络. 另外, 汉语中有字、词之分, 因而和词同现网络的构造方式一样, 我们也可以建立字同现网络. 那么单篇文章所建立的字、词同现网络是否仍然具有小世界特性和无标度特性呢? 为了回答这一问题, 我们分别在现代中文散文、小说、科普和新闻 53 篇文章的基础上建立了 114 个字、词同现网络, 研究发现它们都同时具有小世界特性和无标度特性<sup>[8]</sup>. 最近, 为了从复杂网络的角度来研究 4 种类型文章: 散文、小说、科普和新闻之间的异同, 我们选取了现代中文 200 篇, 针对每篇文章建立了一个字同现网络和一个词同现网络. 研究结果表明所有的网络都具有小世界特性, 并且绝大部分网络都具有无标度特性; 散文和科普具有较多的共性<sup>[9]</sup>.

中华民族历史悠久, 文化源远流长. “距今三千多年以前的殷商时期, 我国产生了最初的文字——甲骨文, 开始了有文字记载的历史.”<sup>[10]</sup> “据安阳殷墟中出土的甲骨卜辞证实, 其中甲骨文已有 4500 个左右, 说明至迟在商朝后期, 汉字已经基本成熟.”<sup>[11]</sup> 到了春秋战国时期中国文化得到空前的发展, 出

现了一大批具有独创精神、对后世文化产生深远影响的经典性著作,如《诗经》、《尚书》、《礼》、《易经》、《春秋》、《论语》、《老子》等<sup>[12]</sup>,至此中华文化体系形成。公元前 221 年秦统一中国,文化上实行了“焚书坑儒”的残酷政策,使“学术文化遭受灭顶之灾,加之秦王朝仅有 15 年短命,因此在文学上几无成就可言。”<sup>[13]</sup> 西汉时期,“文化思想上对各家学说采取了宽容并蓄的政策,思想文化比较活跃自由;”“汉武帝期间北击匈奴,打通西域,扩大了文化交流,罢黜百家,独尊儒术,确定了儒学的统治地位。”<sup>[13]</sup> 因而汉朝文学创作的成就非常可观。魏晋南北朝时期,封建国家长期分裂,但却是“文学自觉的时代,文学上取得了巨大成就,呈现出繁荣的景象。”<sup>[10]</sup> 隋唐时期是“中国古代文学发展的繁荣期”,“与汉朝社会的独尊儒术不同,唐朝的统治者在思想上采取开放的态度,儒、道、释三教并行不悖,并乐于接受外来文化的影响;”<sup>[14]</sup> 加之国家统一,南北文化交融,经济发达和科举制度的实施等,使得唐朝文学登上了中国文学的最高峰<sup>[15]</sup>。“宋朝文学创作趋于理性,在内忧外患的环境中,爱国主义文学主题得到弘扬,宋朝城市的繁荣促使词的繁荣兴盛”<sup>[15]</sup>,所以在文学建设和学术思想的发展方面,呈现出承先启后、宏通广博的繁荣景象。元朝是中国历史上第一个由少数民族统治的朝代<sup>[15]</sup>。大统一的局面促进了国内各民族文化的相互交流,同时西方的天文学等陆续传入中国,促进了中西方科学文化的发展。当时文学中最具代表性的文学样式是元曲。明朝初期实行了以八股文为主要科目的科举制度,极大束缚了文人的思想;明朝后期,文学获得了巨大发展,出现了如《三国演义》、《西游记》等经典著作,谱写了中国文学史上辉煌灿烂的一页<sup>[15]</sup>。清朝是我国封建社会最后的一个王朝,在文化方面,通过提倡科举或博学鸿词的考试,促进了文化的发展。清朝文学是中国古代文学的综合鼎盛期并带有终结的性质<sup>[15,16]</sup>。

中华文化经历了漫长的发展变化。那么从复杂网络的角度来看不同时期的汉语之间有何异同? 依据我们研究发现现代汉语所对应的字、词同现网络都具有小世界特性和无标度特性<sup>[8,9]</sup>。那么从春秋战国至清朝的文章所对应的网络是否仍然具有这两大特性呢? 从语言学的角度来看,春秋战国和两汉时期的文学属于古代汉语,唐、宋、元、明、清时期的属于近代汉语,而对魏晋南北朝时期的文学是属于古代汉语还是近代汉语一直存在很大的争议<sup>[17]</sup>。那么这种争议能否在文章所对应的网络中有所体现? 我们能否为解决这一争议提供一点理论依据呢? 然而,到目前为止,在现有文献中我们还没有发现这方面的研究成果。

基于春秋战国、两汉、三国、两晋、南北朝、唐、宋、元、明、清、现代共 11 个时期各 50 篇散文,本文建立了 550 个字同现网络,每个时期的 50 篇散文合在一起建立一个大的字同现网络,共建立了 561 个字同现网络。研究了它们的统计特征: 聚类系数、平均最短路径、度分布、连通性、平均度等。研究发现 99.6% 的网络具有无标度特性, 95.0% 的网络具有小世界特性。另外,我们的研究结果表明,魏晋南北朝时期的统计数据与其他时期的统计数据之间的确存在明显的不同,且魏晋南北朝时期的文学属于近代汉语似乎更合理。

## 2 网络的构造

国家语言文字工作委员会、新闻出版署 1990 年 3 月修订发布的《标点符号用法》中规定: 标点符号是书面语中用来表示停顿、语气以及词语性质和作用的符号,是书面语的有机组成部分。但是在我国古代的典籍中,标点符号很少,断句只能靠人的经验来完成。据考证,中国从先秦时代就已经有标点符号,只不过这些符号没有统一的标准。例如,在战国时代的竹简中就可以见到“[”型符号,通常表示一篇文章的结束。到了汉代发明了“句读”符号,文辞语意已尽之处为“句”,句中语意未完,语言可停

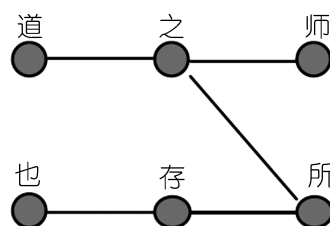


图 1 例句所对应的字同现网络

Figure 1 Character co-occurrence network of the sample sentence

顿之处为“读”。到了宋代,句读有了新的发展,用“。”作为句号,用“,”作为逗号,也就是我们现在使用的句号和逗号。之后,随着时代的进步和文学的需要,标点符号逐步丰富和发展,至今常用的标点符号已达 16 种之多。为了研究不同时期汉语之间的异同,我们所选取的古文是经过处理以后的带有现代常用标点符号的文章。

文言文中往往是用单字来表示白话文中一个词语的意思,加之对古汉语分词比较困难,所以我们在研究不同时期汉语之间的异同时,仅建立了字同现网络。为了研究中国不同时期汉语之间的异同,我们分别在春秋战国(公元前 770~公元前 221)、两汉(公元前 206~220 年)、三国(220 年~280 年)、两晋(280 年~439 年)、南北朝(420 年~589 年)、唐(618 年~907 年)、宋(960 年~1279 年)、元(1279 年~1368 年)、明(1368 年~1644 年)、清(1644 年~1911 年)、现代(1912 年~至今)各选取了 50 篇散文,针对每篇文章建立了一个字同现网络。现代 50 篇散文及其对应字同现网络的统计数据取自文献 [9]。在给定的一篇文章所对应的字同现网络中,节点表示字,如果两个字出现在同一个句子中且相邻,它们之间有一条边相连。为方便起见,所建的网络均是无向、无加权的。例如,考虑下面的句子:

道之所存,师之所存也。

相应的字同现网络见图 1。

一篇文章的长度是指文章中所有字的数目,包括字的重复。因为文章的长度有可能会影响对应网络的统计参数,所以为了避免这一影响,我们所选取的不同时期散文的平均长度几乎相同。各个时期文章长度的分布范围分别为 291~2895 (春秋战国), 354~2555 (两汉), 260~4002 (三国), 320~3031 (两晋), 425~2836 (南北朝), 503~3145 (唐), 536~4068 (宋), 382~3207 (元), 457~4202 (明), 367~3366 (清), 613~2170 (现代); 平均长度分别为 1070 (春秋战国)、1061 (两汉)、1083 (三国)、1056 (两晋)、1021 (南北朝)、992 (唐)、1003 (宋)、938 (元)、1031 (明)、1077 (清)、1161 (现代)。

### 3 实验结果及分析

本节将探索中国不同时期散文之间的异同。我们计算了所有网络及其最大连通子图的统计参数,其平均值见表 1。特别地,为了更加直观地比较各时期的  $L$ 、 $C$ 、 $\gamma$ ,我们在图 2 中画出了其相应的柱状图。

#### 3.1 小世界特性

网络是否具有小世界特性是由它的平均最短路径和聚类系数决定的。因此,首先来介绍这两个基本概念。

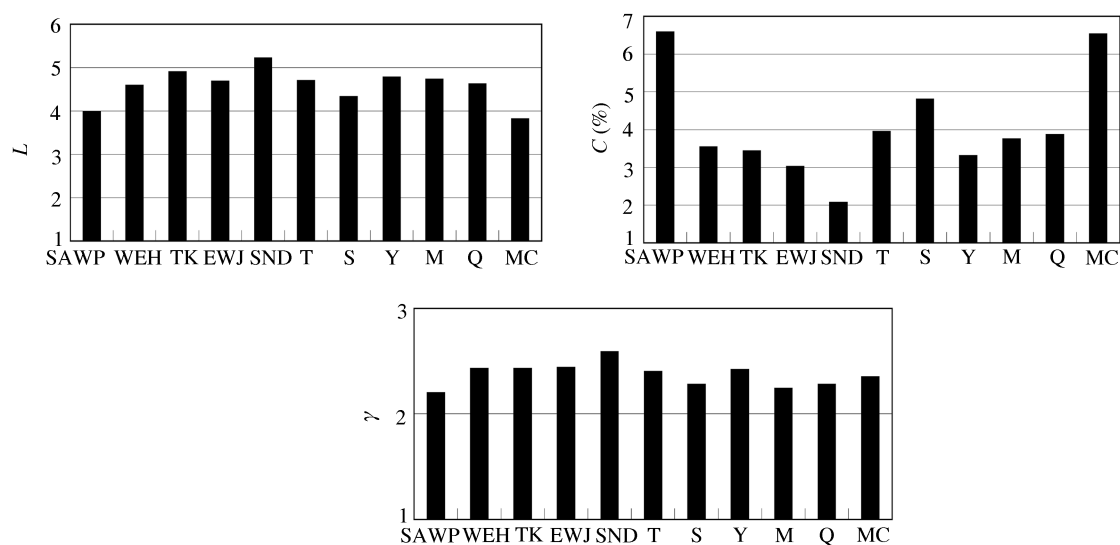
表 1 不同时期散文字同现网络的统计参数的平均值, 其中  $N$  为节点数,  $E$  为边数,  $D$  为直径,  $\langle k \rangle$  为平均度,  $L$ 、 $L_r$  和  $C$ 、 $C_r$  分别表示网络及其相应随机网络的平均最短路径和聚类系数,  $\gamma$  为幂律指数

Table 1 Average values of statistical parameters of networks for essays in different periods of China

| Period | Style | Length | $N$ | $E$ | $D$ | $\langle k \rangle$ | $L/L_r$   | $C/C_r(\%)$ | $\gamma$ | Subnetwork |
|--------|-------|--------|-----|-----|-----|---------------------|-----------|-------------|----------|------------|
| SAWP   | WN    | 1070   | 354 | 658 | 11  | 3.55                | 3.99/4.94 | 6.41/1.16   | 2.20     | 6          |
|        | LC    |        | 342 | 650 |     | 3.62                | 3.99/4.81 | 6.58/1.23   | 2.19     |            |
| WEH    | WN    | 1061   | 457 | 733 | 13  | 3.09                | 4.59/5.69 | 3.42/0.74   | 2.43     | 8          |
|        | LC    |        | 436 | 719 |     | 3.16                | 4.59/5.48 | 3.54/0.80   | 2.41     |            |
| TK     | WN    | 1083   | 458 | 756 | 14  | 3.03                | 4.90/5.86 | 3.32/0.74   | 2.43     | 8          |
|        | LC    |        | 435 | 740 |     | 3.11                | 4.90/5.60 | 3.43/0.82   | 2.41     |            |
| EWJ    | WN    | 1056   | 481 | 758 | 13  | 3.01                | 4.68/5.78 | 2.92/0.68   | 2.44     | 7          |
|        | LC    |        | 462 | 746 |     | 3.07                | 4.68/5.59 | 3.02/0.73   | 2.42     |            |
| SND    | WN    | 1021   | 535 | 758 | 14  | 2.72                | 5.22/6.47 | 2.01/0.57   | 2.59     | 9          |
|        | LC    |        | 506 | 737 |     | 2.79                | 5.22/6.19 | 2.07/0.62   | 2.56     |            |
| T      | WN    | 992    | 464 | 711 | 14  | 3.05                | 4.70/5.77 | 3.76/0.73   | 2.40     | 9          |
|        | LC    |        | 440 | 694 |     | 3.14                | 4.70/5.54 | 3.95/0.81   | 2.38     |            |
| S      | WN    | 1003   | 415 | 695 | 12  | 3.33                | 4.33/5.17 | 4.67/0.90   | 2.28     | 6          |
|        | LC    |        | 400 | 685 |     | 3.40                | 4.33/5.02 | 4.80/0.95   | 2.27     |            |
| Y      | WN    | 938    | 432 | 671 | 14  | 2.97                | 4.78/5.75 | 3.16/0.79   | 2.42     | 9          |
|        | LC    |        | 407 | 655 |     | 3.06                | 4.78/5.48 | 3.31/0.87   | 2.39     |            |
| M      | WN    | 1031   | 439 | 725 | 13  | 3.20                | 4.73/5.36 | 3.62/0.85   | 2.24     | 7          |
|        | LC    |        | 422 | 714 |     | 3.27                | 4.73/5.19 | 3.75/0.91   | 2.22     |            |
| Q      | WN    | 1077   | 424 | 746 | 13  | 3.28                | 4.62/5.28 | 3.73/0.86   | 2.28     | 7          |
|        | LC    |        | 407 | 735 |     | 3.36                | 4.62/5.09 | 3.87/0.93   | 2.26     |            |
| MC     | WN    | 1161   | 426 | 821 | 11  | 3.80                | 3.82/4.55 | 6.47/0.94   | 2.35     | 8          |
|        | LC    |        | 422 | 819 |     | 3.83                | 3.82/4.52 | 6.53/0.96   | 2.35     |            |

若网络中任意两个节点之间至少有一条边相连, 则称该网络是连通的. 假设具有  $N$  个节点的网络是连通、无向、无加权的, 则网络的平均最短路径定义为  $L = 2 \sum_{i>j} d_{ij} / N(N-1)$ , 其中  $d_{ij}$  表示连接节点  $i, j$  之间的最短路径. 节点  $i$  的聚类系数  $C_i$  定义为此节点的任意两个邻居节点之间相连的概率, 即  $C_i = 2E_i / k_i(k_i - 1)$ , 其中  $k_i$  表示节点  $i$  的度,  $E_i$  表示节点  $i$  的邻居节点之间实际存在的边数. 整个网络的聚类系数  $C$  定义为所有  $C_i$  ( $1 \leq i \leq N$ ) 的平均值. 对 ER 随机图<sup>[18]</sup>而言, 设网络的平均度为  $\langle k \rangle$ , 则平均最短路径为  $L_r \approx \ln N / \ln \langle k \rangle$ , 聚类系数为  $C_r \approx \langle k \rangle / (N-1)$ . 若网络的平均最短路径  $L \approx L_r$  且聚类系数  $C \gg C_r$ , 则称该网络具有小世界特性<sup>[19]</sup>. 特别地, 如果网络是不连通的, 则该网络的  $L$  和  $C$  也可以像上述定义方式那样来定义: 当节点  $i$  和  $j$  之间没有任何连边时, 令  $d_{ij} = 0$ , 且当节点  $i$  和网络中的其他节点之间没有任何连边时, 令  $C_i = 0$ .

因为小世界特性是针对连通网络而言的, 所以我们先来研究所建网络的连通性. 经统计发现: 在 550 个网络中 92.9% 的网络是不连通的. 但是在不连通的网络中总是存在一个最大连通子图, 其节点数和边数与整个网络的相差很小, 从而导致整个网络的统计参数与其最大连通子图的统计参数之间相

图 2 不同时期散文字同现网络的  $L, C, \gamma$  的平均值Figure 2 Average values of  $L, C$ , and  $\gamma$ 

差很小 (表 1). 尽管如此, 为了准确, 在研究网络的小世界特性时, 我们仅考虑其最大连通子图的平均最短路径和聚类系数.

研究发现 550 个网络都具有小的平均最短路径  $L$ . 不同时期  $L$  的范围分别为 3.17~6.34 (春秋战国), 3.42~6.41 (两汉), 3.46~7.90 (三国), 3.87~6.47 (两晋), 3.78~8.13 (南北朝), 3.26~6.67 (唐), 3.31~5.84 (宋), 3.60~6.87 (元), 3.26~6.31 (明), 3.37~6.39 (清), 3.43~4.45 (现代). 不同时期  $L$  的平均值见表 1. 因此, 由图 2 知, 在所有的网络中南北朝时期的具有最大的  $L$ , 现代的则最小. 这说明对任意给定的两个节点而言, 在平均意义下连接这两个节点所需要的节点数为南北朝时期的最多, 而现代的最少.

用  $L_r$  表示随机网络的平均最短路径, 则不同时期  $L_r$  的范围分别为 3.47~6.34 (春秋战国), 4.00~6.98 (两汉), 3.98~7.11 (三国), 4.40~7.09 (两晋), 4.00~7.36 (南北朝), 3.76~7.00 (唐), 3.56~6.99 (宋), 4.03~6.80 (元), 3.69~6.22 (明), 3.96~6.50 (清), 3.95~5.45 (现代). 不同时期  $L_r$  的平均值见表 1. 550 个网络均有  $L \approx L_r$  (表 1).

所有的网络都具有非常小的聚类系数  $C$ . 不同时期  $C$  的范围分别为 0.0086~0.1318 (春秋战国), 0.0057~0.0931 (两汉), 0.0038~0.1051 (三国), 0.0043~0.0745 (两晋), 0.0016~0.0998 (南北朝), 0.0049~0.1563 (唐), 0.0033~0.1320 (宋), 0.0089~0.1159 (元), 0.0048~0.1295 (明), 0.0070~0.0982 (清), 0.0295~0.1028 (现代). 在所有的网络中南北朝时期的具有最小的  $C$ , 春秋战国的则最大 (图 2).

不同时期网络所对应的随机网络的  $C_r$  的范围分别为 0.0050~0.0308 (春秋战国), 0.0033~0.0137 (两汉), 0.0034~0.0138 (三国), 0.0043~0.0134 (两晋), 0.0028~0.0105 (南北朝), 0.0033~0.0187 (唐), 0.0048~0.0194 (宋), 0.0037~0.0178 (元), 0.0030~0.0248 (明), 0.0048~0.0158 (清), 0.0060~0.0138 (现代). 不同时期  $C_r$  的平均值见表 1.

经研究发现在我们所构造的网络中, 绝大部分网络的聚类系数都满足  $C \gg C_r$ , 但是却存在 28 个网络其聚类系数为  $C < C_r$ . 这些网络分别出现在从春秋战国到清朝这 10 个时期. 这些网络个数分别为 1, 2, 5, 2, 7, 4, 1, 1, 2, 3, 所对应的文章长度分别为 360, 428, 462, 329, 329, 349, 497, 568, 320, 405,

443, 455, 546, 565, 572, 650, 653, 579, 610, 654, 688, 574, 455, 490, 561, 448, 468, 483. 将各时期 50 篇文章的长度由小到大排序, 这 28 个网络所对应的文章长度排序分别为 4, 5, 7, 4, 4, 8, 12, 17, 1, 3, 2, 4, 8, 9, 10, 18, 19, 3, 5, 10, 15, 7, 6, 2, 8, 5, 6, 9. 这表明这些网络所对应文章的长度远远小于各个时期文章长度的平均值, 且节点数目相对较少. 进一步, 在这 28 个网络中存在一个共同的现象: 聚类系数为零的节点非常多, 也就是说网络中存在很多的星形子网络. 因此, 在这 28 个网络中, 导致  $C < C_r$  的原因可能与文章长度相对较短有关.

综上所述, 所构造的网络除去 28 个不具有小世界特性之外, 其余网络均具有小世界特性. 但是现代汉语的散文、小说、科普和新闻的字同现网络均具有小世界特性<sup>[9]</sup>. 另外, 由大量文章合在一起而构成的中文、英文、斯洛伐克语的词同现网络也都具有小世界特性<sup>[4,5,20,21]</sup>.

春秋战国和宋代都是文化较繁荣昌盛的时期. 而现代散文与这两个时期的散文所对应的网络的统计参数比较接近 (表 1), 因而我们认为现代文学可能正处于蓬勃发展的阶段.

南北朝是中国历史上为数不多的南北分裂时期之一. 这种南北分裂、南北对峙的局面使得社会动荡不安, 出现了一种新的文学题材 — 民歌. 南朝民歌清丽缠绵, 更多地反映了人民真挚纯洁的爱情生活. 北朝民歌粗犷豪放, 广泛地反映了北方动乱不安的社会现实和人民的生活习俗. 13 世纪初, 蒙古统治者经过半个多世纪的征服战争, 建立了元朝. 元朝民族压迫和阶级压迫非常突出, 蒙古统治者对汉人实行高压政策, 在这种尖锐的社会矛盾下一种可以针砭时事的文体 — 元曲应运而生, 成为我国戏剧史上的一座里程碑. 因而在南北朝和元朝时期, 新文化的融入和新文体的出现极有可能是使得该时期网络具有较大的  $L$  和较小的  $C$  (图 2) 的主要原因. 从南北朝时期到宋朝和由元朝至现代的网络具有越来越小的  $L$  和越来越大的  $C$ , 则可能与文体越来越趋于成熟有关.

### 3.2 幂律分布

网络的度分布  $p(k)$  是刻画网络统计性质的另一个重要参数, 定义为网络中随机选取一个节点其度恰好为  $k$  的概率. 若  $p(k)$  满足幂律分布:  $p(k) \propto k^{-\gamma}$ , 其中  $\gamma$  为正常数, 则称该网络具有无标度特性<sup>[22]</sup>.

我们首先分析 550 个网络的度分布情况. 对这些网络而言, 99.6% 的网络服从幂律分布, 显示了不同时期网络的无标度特性. 在服从幂律分布的网络中, 幂律指数  $\gamma$  的范围分别为 1.88~2.69 (春秋战国), 1.81~3.39 (两汉), 1.70~3.56 (三国), 1.89~2.97 (两晋), 2.03~3.56 (南北朝), 1.97~3.18 (唐), 1.77~2.96 (宋), 1.78~3.32 (元), 1.79~2.71 (明), 1.86~2.84 (清), 2.02~2.98 (现代). 不同时期  $\gamma$  的平均值见表 1. 由图 2 知, 南北朝时期的网络具有最大的幂律指数, 春秋战国时期的则最小. 图 3 中给出了春秋战国、两晋和清各一篇散文所对应的字同现网络的度分布图, 其中三篇散文的长度分别为 511、1209、1818.

众所周知, BA 无标度网络的幂律指数约为 3<sup>[22]</sup>. 在我们的网络中  $\gamma$  范围为 1.7~3.6. 从而我们的网络除了具有 BA 无标度网络中的优先连接机制以外还可能存在其他连接机制.

550 个网络中不具有无标度特性的网络仅有两个, 图 4 给出了它们的度分布图. 这两个网络分别属于两汉、三国时期, 节点数分别为 233、207; 所对应的文章长度分别为 354、282. 由表 1 中的数据知, 这两个网络所对应的文章长度都相对较短. 因此, 网络不具有无标度特性可能与文章长度偏短有关.

在所建的网络的度分布中存在“胖尾”现象. 同样, 在其他语言网络<sup>[5,6,8,9,23,24]</sup>的度分布中也存在这种现象. 此现象说明了在一个度分布为幂律形式的无标度网络中, 绝大部分的节点的度相对很低, 但存在少量的度相对很高的节点<sup>[22]</sup>. 经研究发现, 在春秋战国至清代的散文字同现网络中度相对较

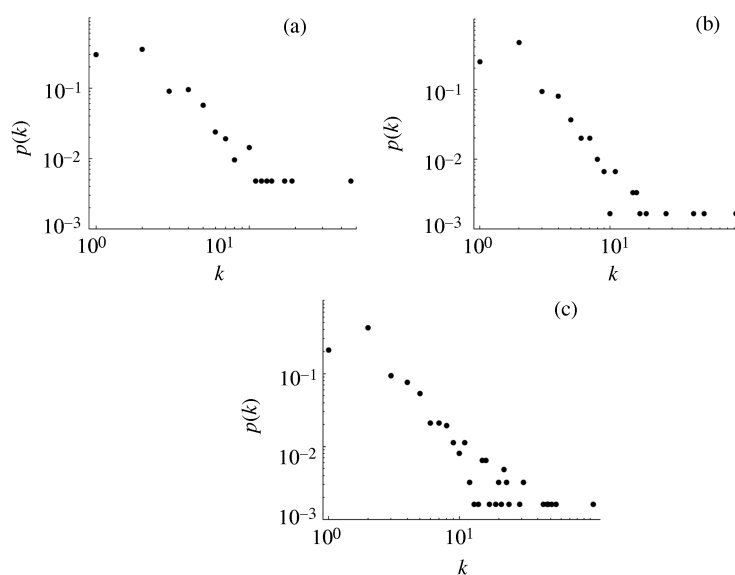


图 3 在双对数坐标下字同现网络的度分布图

**Figure 3** Degree distributions of character co-occurrence networks in the log-log scale. (a)  $\gamma = 2.32$  for SAWP; (b)  $\gamma = 2.64$  for EWJ; (c)  $\gamma = 2.16$  for Q

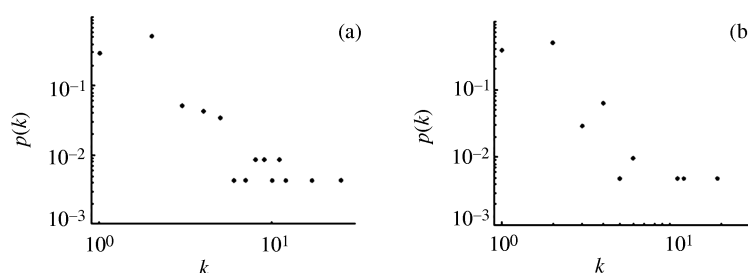


图 4 在双对数坐标下不具有幂律特性的网络的度分布图

**Figure 4** Degree distributions of networks, which are unlike power law. (a) WEH; (b) TK

大的节点几乎相同, 为“之”、“而”、“以”及与文章主题相关的字. 古文中“之”一般表示领属关系, 可译为“的”, 如“道之所存, 师之所存也.”(《师说》) 或者代替人或事物, 如“人非生而知之者, 孰能无惑?”(《师说》)“而”一般表示递进关系, 可译为“并且”、“而且”, 如“君子博学而日参省乎己.”(《劝学》) 或者表示转折关系, 可译为“但是”、“却”, 如“青, 取之于蓝, 而青于蓝.”(《劝学》)“以”表示动作所使用的工具, 可译为“用”、“凭借”, 如“愿以十五城请易璧.”(《廉颇蔺相如列传》) 或者表示动作、行为产生的原因, 可译为“因为”, 如“时操军兼以饥疫, 死者大半.”(《赤壁之战》) 如上所述, 因为“之”、“而”、“以”在古文中为一词多义, 所以在文中出现的频率较高, 因而导致在字同现网络中它们具有较大的度. 而在现代散文字同现网络中度相对较大的节点为“的”、“了”、“是”<sup>[9]</sup>. 什么原因导致古文和现代文所对应的字同现网络中度较大的节点发生变化了呢? 众所周知, 清朝之前的文章绝大部分为文言文, 到 1919 年五四运动以后, 白话文才取代了文言文, 成为写作的主流, 使文言文慢慢地退出了历史的舞台. 因此, 文言文向白话文的转变是导致古文和现代文的字同现网络中度相对较大的节点发生变化的主要原因.

表 2 各时期 50 篇散文合一后所对应字同现网络的统计参数

Table 2 Statistical parameters of networks from the concatenated articles in different periods of China

| Period | Style | Length | $N$  | $E$   | $D$ | $\langle k \rangle$ | $L/L_r$   | $C/C_r(\%)$ | $\gamma$ |
|--------|-------|--------|------|-------|-----|---------------------|-----------|-------------|----------|
| SAWP   | WN    | 53479  | 3045 | 21998 | 9   | 14.45               | 2.93/3.00 | 24.62/0.47  | 1.48     |
|        | LC    |        | 3030 | 21990 |     | 14.51               | 2.93/3.00 | 24.74/0.48  | 1.48     |
| WEH    | WN    | 53109  | 3805 | 27298 | 12  | 14.35               | 3.04/3.10 | 17.93/0.38  | 1.47     |
|        | LC    |        | 3772 | 27283 |     | 14.47               | 3.04/3.09 | 18.08/0.38  | 1.47     |
| TK     | WN    | 54148  | 3511 | 27745 | 9   | 15.80               | 2.94/2.96 | 18.43/0.45  | 1.45     |
|        | LC    |        | 3503 | 27743 |     | 15.84               | 2.94/2.96 | 18.47/0.45  | 1.45     |
| EWJ    | WN    | 52788  | 3388 | 28285 | 9   | 16.70               | 2.92/2.89 | 17.18/0.49  | 1.43     |
|        | LC    |        | 3378 | 28284 |     | 16.75               | 2.92/2.88 | 17.23/0.50  | 1.43     |
| SND    | WN    | 51037  | 3822 | 30030 | 10  | 15.71               | 2.97/3.00 | 14.34/0.41  | 1.43     |
|        | LC    |        | 3818 | 30029 |     | 15.73               | 2.97/2.99 | 14.32/0.41  | 1.43     |
| T      | WN    | 49582  | 3608 | 27436 | 9   | 15.21               | 3.00/3.01 | 15.76/0.42  | 1.42     |
|        | LC    |        | 3597 | 27433 |     | 15.25               | 3.00/3.01 | 15.81/0.42  | 1.42     |
| S      | WN    | 50135  | 3332 | 24830 | 8   | 14.90               | 2.98/3.00 | 18.73/0.45  | 1.48     |
|        | LC    |        | 3317 | 24826 |     | 14.97               | 2.98/3.00 | 18.81/0.45  | 1.48     |
| Y      | WN    | 46885  | 3809 | 26868 | 9   | 14.11               | 3.07/3.12 | 15.33/0.37  | 1.48     |
|        | LC    |        | 3797 | 26867 |     | 14.15               | 3.07/3.11 | 15.38/0.37  | 1.48     |
| M      | WN    | 51539  | 3693 | 28438 | 9   | 15.40               | 3.01/3.00 | 17.34/0.42  | 1.46     |
|        | LC    |        | 3683 | 28434 |     | 15.44               | 3.01/3.00 | 17.39/0.42  | 1.46     |
| Q      | WN    | 53855  | 3466 | 28061 | 8   | 16.19               | 2.94/2.93 | 18.97/0.47  | 1.47     |
|        | LC    |        | 3451 | 28054 |     | 16.26               | 2.94/2.92 | 19.05/0.47  | 1.47     |
| MC     | WN    | 65548  | 3015 | 25388 | 8   | 16.84               | 2.75/2.84 | 25.70/0.56  | 1.41     |
|        | LC    |        | 3006 | 25387 |     | 16.89               | 2.75/2.83 | 25.77/0.56  | 1.41     |

#### 4 各时期文章合一后所对应的网络

我们将各时期 50 篇散文合在一起建立了 11 个字同现网络, 其统计参数见表 2.

统计结果表明 11 个网络均具有小世界特性和无标度特性. 由表 2 中的数据知, 各时期网络的统计参数之间比较接近; 现代网络具有最大的  $C$  和最小的  $L, \gamma$ . 进一步由  $L, C$  可以看出, 现代散文和春秋战国时期的散文比较接近; 这与在单篇文章所建立的网络中所得的结论类似.

尽管合一后现代散文的长度最长, 但是节点数目却最少 (表 2). 因而与古文相比现代文中常用字的数目可能有所减少. 然而, 据统计在现代 50 篇散文合一后文章所对应的词同现网络中, 节点数却高达 8222 个<sup>[9]</sup>. 众所周知, 文言文与白话文相比有一个很大的不同: 文言文中往往是用单字来表示白话文中一个词语的意思, 而在现代文中一个字却可以在多个词语中出现. 这导致了现代文所对应的字同现网络中节点较少但在词同现网络中节点数目却很多. 因此, 现代文中常用字虽然相对较少但在用词方面却很丰富.



表 3 古代汉语、魏晋南北朝时期的汉语、近代汉语、现代汉语中散文字同现网络的统计参数的平均值

**Table 3** Average values of statistical parameters for networks constructed from the articles in the ancient Chinese language; Wei, Jin, and Southern-Northern Dynasties; the recent Chinese language; and the modern Chinese language

| Period     | Style | Length | $N$ | $E$ | $D$ | $\langle k \rangle$ | $L/L_r$   | $C/C_r(\%)$ | $\gamma$ | Subnetwork |
|------------|-------|--------|-----|-----|-----|---------------------|-----------|-------------|----------|------------|
| Ancient    | WN    | 1065   | 406 | 696 | 12  | 3.32                | 4.29/5.32 | 4.92/0.95   | 2.32     | 7          |
|            | LC    |        | 389 | 685 |     | 3.39                | 4.29/5.15 | 5.06/1.02   | 2.30     |            |
| WeiJin SND | WN    | 1053   | 491 | 757 | 14  | 2.92                | 4.93/6.04 | 2.75/0.66   | 2.49     | 8          |
|            | LC    |        | 468 | 741 |     | 2.99                | 4.93/5.79 | 2.84/0.72   | 2.46     |            |
| Recent     | WN    | 1008   | 435 | 710 | 13  | 3.17                | 4.63/5.47 | 3.79/0.83   | 2.33     | 8          |
|            | LC    |        | 415 | 697 |     | 3.25                | 4.62/5.26 | 3.94/0.89   | 2.30     |            |
| Modern     | WN    | 1161   | 426 | 821 | 11  | 3.80                | 3.82/4.55 | 6.47/0.94   | 2.35     | 8          |
|            | LC    |        | 422 | 819 |     | 3.83                | 3.82/4.52 | 6.53/0.96   | 2.35     |            |

## 5 不同时期散文的比较

本节将根据表 1 和文献 [9] 中的数据从复杂网络的角度来分析古代汉语、魏晋南北朝时期的汉语、近代汉语和现代汉语之间的异同.

在语言学中对汉语有如下分类: 春秋战国和两汉时期的文学属于古代汉语, 唐、宋、元、明、清时期的属于近代汉语, 中华民国至今的属于现代汉语. 但是, 对魏晋南北朝时期的文学是属于古代汉语还是近代汉语一直存在很大的争议. 这一争议能否在文章所对应的网络中得到体现呢? 为了回答这一问题, 我们给出了古代汉语、魏晋南北朝的汉语、近代汉语以及现代汉语中各 100、150、250、50 篇文章对应网络统计参数的平均值, 见表 3.

由表 3 中的数据知, 魏晋南北朝时期的网络具有最大的  $L$ 、 $\gamma$  和最小的  $\langle k \rangle$ 、 $C$ ; 现代汉语则具有最大的  $\langle k \rangle$ 、 $C$  和最小的  $L$ ; 魏晋南北朝时期的统计数据与古代汉语和近代汉语的统计数据之间的确存在明显的不同. 从而, 汉语语言学中对魏晋南北朝时期汉语的归属性的争议的确可以在文章所对应的网络中得到体现. 进一步, 由表 3 中的数据知, 魏晋南北朝时期的汉语与近代汉语所对应网络的统计参数更为接近. 因而, 从复杂网络的角度来看, 魏晋南北朝时期的散文属于近代汉语似乎更合理. 更进一步, 我们可以推测魏晋南北朝时期的文学有可能属于近代汉语.

吕叔湘在 1988 年发表过这样的观点: “从三国到唐末, 这 700 年该怎么划分? 这个时期的口语肯定跟秦汉以前有很大差别, 但是由于书面语的保守性, 口语成分只能在这里那里露个一鳞半爪, ……从古典书面语的立场说, 这些都是 ‘俗语’, 也就是说, 都可以算是近代汉语的 ‘露头’”<sup>[25]</sup>. 高福生也认为魏晋南北朝时期的俗语词确实是近代汉语的 “露头”<sup>[26]</sup>. 因此, 本文所得到的魏晋南北朝时期的文学属于近代汉语这一结论与上述观点一致.

在中国文学史上, 魏晋南北朝是一个酝酿着新变的时期. 这种新变可以概括为以下 3 点: 文学进入自觉的阶段, 文学创作趋于个性化; 玄学的兴起和佛教的传入为文学创作带来新的因素; 语言形式美的发现及其在文学上的运用. 就文体的发展看来: 一种诗化的散文即骈文的兴盛, 成为这个时期重要的文学现象, 中国文学增添了一种新的、抒情性很强的、可以充分发挥汉语语言形式美的文体; 在汉代盛极一时的大赋, 演变而为抒情小赋, 并因骈文的兴盛而增加了骈俪的成分, 骈文、骈赋在梁陈两代进入高峰. 魏晋南北朝时期文学表现领域的扩大和表现技巧的丰富, 促进了文章体裁的变化与扩展, 这

为唐朝人们迎来了一个诗的辉煌时代奠定了基础. 如果没有魏晋南北朝时期的酝酿, 就没有唐朝文学的全面繁荣. 因此, 在本文中得到的魏晋南北朝时期的文学属于近代汉语更为合理. 该结论具有两个方面的学术意义. 首先, 这可以印证中国古代文学研究认为魏晋南北朝属于“文学自觉”时代的观点, 语言统计显示出的特异性为其提供了一个有利的证据. 其次, 魏晋南北朝时期出现的骈文、骈赋对近代汉语文学的影响是深刻的, 唐朝文学特别是散文和诗的繁荣可以由此得到很好的解释.

## 6 结论

基于春秋战国、两汉、三国、两晋、南北朝、唐、宋、元、明、清、现代共 11 个时期各 50 篇散文, 本文建立了 561 个字同现网络, 并研究了它们的统计特征. 研究发现 99.6% 的网络具有无标度特性, 95.0% 的网络具有小世界特性. 在单篇文章建立的字同现网络中, 从平均意义上看, 关于平均最短路径  $L$ : 南北朝时期最大, 现代最小, 春秋战国时期次小; 关于聚类系数  $C$ : 南北朝时期最小, 春秋战国时期最大, 现代次大. 这说明春秋战国时期的文学和现代文学比南北朝时期的文学发展成熟, 语言表述简练. 另外, 南北朝和元朝的文章所对应的网络具有较大的  $L$  和较小的  $C$ , 可能与该时期时局动荡不安、新文化的融入和新文体的出现有关; 从南北朝时期到宋朝和由元朝至现代的网络具有越来越小的  $L$  和越来越大的  $C$ , 可能与社会相对越来越稳定而且文体越来越趋于成熟有关. 各时期 50 篇散文合一后所对应的字同现网络的统计数据说明, 尽管现代文的常用字数相对较少但是在用词方面却很丰富, 反映出文言文与白话文的不同. 因此, 本文所建立的语言网络较好地反映了不同时期的文化发展变化. 统计结果表明汉语语言不是平稳地发展, 它是受各个时期的军事、政治和经济的影响.

到目前为止, 我们在以往的文学研究中还没有发现以上结论. 这说明从复杂网络的角度来研究文学, 可以为文学研究提供新方法和新观点.

本文的研究为魏晋南北朝时期的文学的归属争议提供了大量详实的统计数据. 在语言学的研究中, 魏晋南北朝时期的文学是属于古代汉语还是近代汉语一直存在很大的争议. 本文的统计数据显示魏晋南北朝时期的网络特征与古代汉语和现代汉语网络特征的确存在明显的不同. 相对而言, 魏晋南北朝时期的网络特征与近代汉语的网络特征更接近. 因此, 魏晋南北朝时期的文学属于近代汉语似乎更合理. 此结论与吕叔湘和高福生的观点基本一致<sup>[25,26]</sup>. 这种一致性也恰恰说明从复杂网络的角度来研究语言或者文学发展变化是有价值的.

另外, 本文的工作为我们进一步研究汉语的演化和建立汉语演化模型提供了必要的统计数据.

**致谢** 感谢华东师范大学中国文字研究与应用中心的臧克和教授和山东大学文科学报刘运兴编审提供的汉语史分期上的分歧问题和有益讨论. 感谢审稿人所提出的有意义的建议.

## 参考文献

- 1 Steels L. Language as a complex adaptive system. In: Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2000
- 2 Grimes B F. Ethnologue: Languages of the World. 14th ed. Dallas: Summer Institute of Linguistics, 2000
- 3 Wei L X, Li Y, Kang S Y, et al. Organizational structure and scale-free properties of Chinese words networks. Chinese Sci Bull, 2005, 50: 1575–1579 [韦洛霞, 李勇, 康世勇, 等. 汉语词组网的组织结构与无标度特性. 科学通报, 2005, 50: 1575–1579]

- 4 Liu Z, Sun M. Chinese word co-occurrence network: its small world effect and scale-free property. *J Chinese Inf Process*, 2007, 6: 52–58
- 5 Zhou S, Hu G, Zhang Z, et al. An empirical study of Chinese language networks. *Phys A*, 2008, 387: 3039–3047
- 6 Liu H. The complexity of Chinese syntactic dependency networks. *Phys A*, 2008, 387: 3048–3058
- 7 Tang L, Zhang Y, Fu X. Structures of semantic networks: how do we learn semantic knowledge. *J Southeast Univ*, 2006, 22: 413–417
- 8 Shi Y, Liang W, Liu J, et al. Structural equivalence between co-occurrences of characters and words in Chinese language. In: *International Symposium on Nonlinear Theory and its Applied Budapest*, 2008. 94–97
- 9 Liang W, Shi Y, Tse C K, et al. Comparison of co-occurrence networks of the Chinese and English languages. *Phys A*, 2009, 388: 4901–4909
- 10 Zhang F Y, Zhao J. *History of Ancient Chinese Literature (I)*. Tianjin: Nankai University Press, 2003 [张峰屹, 赵季. 中国古代文学发展史 (上). 天津: 南开大学出版社, 2003]
- 11 Ma J G, Huang J. *Ancient Chinese Literature*. Changsha: Hunan Art Publishing House, 2006 [马积高, 黄钧. 中国古代文学史. 长沙: 湖南文艺出版社, 2006]
- 12 Gao Y. Ancient Chinese system and ancient Chinese cultural type. *J Xinjiang Univ (Social Sci)*, 2003, 31: 84–90 [高玉. 古代汉语体系与中国古代文化类型. 新疆大学学报 (社会科学版), 2003, 31: 84–90]
- 13 Zhou J Z. *Ancient Chinese Literature*. Nanjing: Nanjing University Press, 2003 [周建忠. 中国古代文学史. 南京: 南京大学出版社, 2003]
- 14 Zhang Y. *History of Ancient Chinese Literature (II)*. Tianjin: Nankai University Press, 2003 [张毅. 中国古代文学发展史 (中). 天津: 南开大学出版社, 2003]
- 15 School of Chinese Language and Culture, Beijing Normal University. *Ancient Chinese Literature*. Beijing: Beijing Normal University Press, 2008 [北京师范大学文学院. 中国古代文学史. 北京: 北京师范大学出版社, 2008]
- 16 Ning J Y, Li R S. *History of Ancient Chinese Literature (III)*. Tianjin: Nankai University Press, 2003 [宁稼雨, 李瑞山. 中国古代文学发展史 (下). 天津: 南开大学出版社, 2003]
- 17 Zhang Y. A summary of upper limit of the recent Chinese language. *J Henan Univ (Social Sci)*, 1995, 35: 52–55 [张玉萍. 近代汉语上限问题讨论综述. 河南大学学报 (社会科学版), 1995, 35(4): 52–55]
- 18 Erdős P, Rényi A. On the evolution of random graphs. *Publ Math Inst Hung Acad Sci*, 1960, 5: 17–60
- 19 Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ network. *Nature*, 1998, 393: 440–442
- 20 Cancho R Ferrer I, Solé R V. The small world of human language. *Proc R Soc Lond B*, 2001, 268: 2261–2265
- 21 Markošová M, Náther P. Language as a small world network. In: *Proceedings of the 6th International Conference on Hybrid Intelligent Systems*. Washington: IEEE Computer Society, 2006. 37
- 22 Barabási A L, Albert R. Emergence of scaling in random network. *Science*, 1999, 286: 509–512
- 23 Yu Y, Wang Z, Gao W, et al. Chinese language processing with complex network theory. In: *International Conference on Computer Science and Software Engineering*. Washington: IEEE Computer Society, 2008. 710–713
- 24 Sigman M, Cecchi G. Global organization of the wordnet lexicon. *Proc Natl Acad Sci*, 2002, 99: 1742–1747
- 25 Jiang L. *Interpretation of Words in Novel in Wei, Jin, and Southern-Northern Dynasties*. Beijing: Language and Literature Press, 1988 [江蓝生. 魏晋南北朝小说词语汇释. 北京: 语文出版社, 1988]
- 26 Gao F. Notes of “Interpretation of words in novel in Wei, Jin, and Southern-Northern Dynasties”. *J Jiangxi Norm Univ*, 1995, 24: 98–100 [高福生. 读《魏晋南北朝小说词语汇释》札记. 江西师范大学学报, 1995, 24: 98–100]

## Study on co-occurrence character networks from Chinese essays in different periods

LIANG Wei<sup>1,2</sup>, SHI YuMing<sup>1,\*</sup>, TSE Chi K<sup>3</sup> & WANG YanLi<sup>1</sup>

1 *Department of Mathematics, Shandong University Jinan, Shandong 250100, China;*

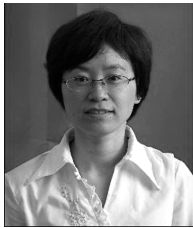
2 *School of Mathematics and Information Science, Henan Polytechnic University Jiaozuo, Henan 454000, China;*

3 Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China

\*E-mail: ymshi@sdu.edu.cn

**Abstract** Co-occurrence networks of Chinese characters are constructed from collections of essays in different periods of China: the ancient Chinese language, the Chinese language in Wei, Jin, and Southern-Northern Dynasties, the recent Chinese language, and the modern Chinese language, and their statistical parameters are studied. It has been found that 99.6% networks have the scale-free feature and 95.0% networks have the small-world effect. This study reveals some commonalities and differences among articles in different periods of China from a complex network perspective. There has been a controversial question as to whether the literatures in Wei, Jin, and Southern-Northern Dynasties should belong to the ancient Chinese language or the recent Chinese language in the linguistic study. Our work shows that the statistical parameters of networks in Wei, Jin, and Southern-Northern Dynasties are clearly different from those of networks in the other periods of China, and it seems more reasonable that the literatures in Wei, Jin, and Southern-Northern Dynasties belong to the recent Chinese language.

**Keywords** chinese language, essay, co-occurrence character network, small-world, scale-free, linguistics, heterogeneous networks, statistical method



**LIANG Wei** was born in 1979. She received the Ph.D. degree in mathematics from Shandong University, Jinan in 2010. Currently, she is a Lecturer of School of Mathematics and Information Science at Henan Polytechnic University. Her research interests include chaos theory, complex networks and their applications.



**SHI YuMing** was born in 1963. She received the Ph.D. degree in mathematics from Shandong University, Jinan in 2000. Currently, she is a Professor of School of Mathematics at Shandong University. Her research interests include spectral theory of Hamiltonian systems, chaos theory, complex networks and their applications. Dr. Shi is now associate editors of Abstract and Applied Analysis and International

Journal of Bifurcation and Chaos.



**TSE Chi K.** was born in 1965. He received the BEng (Hons) and Ph.D. degrees in electrical engineering from the University of Melbourne, Australia, in 1987 and 1991, respectively. Currently, he is a Chair Professor and Head of Department of Electronic and Information Engineering at Hong Kong Polytechnic University. His research interests include nonlinear systems, power electronics, and complex networks applications.

Dr. Tse is a Fellow of IEEE, Fellow of IE Australia, Editor-in-Chief of IEEE CAS Magazine, Editor-in-Chief of IEEE CASS Newsletter, and panel member of Hong Kong Research Grants Council. He holds the Chang Jiang Scholars Chair Professor at Huazhong University of Science and Technology, Wuhan.