BRILL

BULLETIN OF CHINESE LINGUISTICS 9 (2016) 218-241

BULLETIN of CHINESE LINGUISTICS
brill.com/bcl

Using Network Models to Analyze Old Chinese Rhyme Data

Johann-Mattis List
CNRS-CRLAO-INALCO-EHESS
mattis.list@lingpy.org

Abstract

The evidence one can draw from the rhyming behavior of Old Chinese words plays a crucial role for the reconstruction of Old Chinese, and is particularly relevant to recent proposals. Some of these proposals are no longer solely based on the intuition of scholars but also substantiated by statistical arguments that help to assess the probability by which a given set of rhyming instances can be assigned to an established rhyme group. So far, however, quantitative methods were only used to confirm given hypotheses regarding rhyme groups in Old Chinese, and no exploratory analyses that would create hypotheses regarding rhyme groups in a corpus were carried out. This paper presents a new method that models rhyme data as weighted undirected networks. By representing rhyme words as nodes in a network and the frequency of rhymes in a given corpus as links between nodes, rhyme groups can be inferred with help of standard algorithms originally designed for social network analysis. This is illustrated through the construction of a rhyme network from the *Book of Odes* and comparing the automatically inferred rhyme groups with rhyme groups proposed in the literature. Apart from revealing interesting general properties of rhyme networks in Chinese historical phonology, the analysis provides strong evidence for a coda **-r* in Old Chinese. The results of the analysis and the rhyme network of the *Book of Odes* can be inspected in form of an interactive online application or directly downloaded.

Keywords

rhyme network – *Book of Odes* – Old Chinese phonology – Old Chinese reconstruction methodology

1 Introduction

1.1 *Rhyme Analysis in Old Chinese Reconstruction*

The analysis of rhyme patterns is one of the core methods for the reconstruction of Old Chinese phonology. It emerged when scholars of the Suí 隋 (581–618) and Táng 唐 (618–907) dynasties realized that old poems, especially those in the *Book of Odes* (Shījīng 詩經 ca. 1050–600 BCE), were full of inconsistencies regarding the rhyming of words. While the first reaction was to attribute inconsistencies to a different, less strict attitude towards rhyming practiced by the ancestors (as advocated by Lù Démíng 陸德明, 550–630), or to a habit of the elders to switch the pronunciation in certain words in order to make them rhyme (a practice called *xiéyīn* 諧音 'sound harmonization', Baxter 1992:153). Later scholars from the Míng 明 (1368–1644) and Qīng 清 dynasties (1644–1911) realized that the inconsistencies in the rhyme patterns reflect the effects of language change (Baxter 1992:153–157). Table 1 illustrates how a poem that

 | DOI 10.1163/2405478X-00902004

TABLE 1 *Strange rhymes in the Book of Odes. The table shows the third stanza of Ode 28 (邶風 · 燕燕) with the English translation taken from Karlgren (1950). RW shows the rhyme words, which are also highlighted by drawing boxes around the characters in the Chinese text. Patterns and Middle Chinese readings (MCH) follow Baxter (1992), and Old Chinese rhymes are given in the Old Chinese Baxter-Sagart system (Baxter and Sagart 2014). When comparing Middle Chinese readings with the rhymes reconstructed for Old Chinese, it becomes obvious that from the Middle Chinese perspective, the rhyme pattern B rhymes imperfectly, including words ending in -om and words ending in -im, while the Old Chinese readings all show -ə m.*

Chinese Text	Translation	RW	Patterns	MCH	OCBS-Rhyme
燕燕於飛	The swallows go flying	*fēi* 飛	A	*pjɨj	*-ər
下上其音	falling and rising are their voices;	*yīn* 音	B	*ʔim	*-əm
之子於歸	This young lady goes to her new home,	*guī* 歸	A	*kjwɨj	*-əj
遠送於南	far I accompany her to the south.	*nán* 南	B	*nom	*-əm
瞻望弗及	I gaze after her, can no longer see her,	[*jí* 及]	–	[**gip*]	[**-əp*]
實勞我心	truly it grieves my heart	*xīn* 心	B	*sim	*-əm

rhymed regularly when it was originally written lost certain rhyme patterns in the course of history. Already before scholars had clearly recognized that the strange rhymes in the *Odes* where due to the effects of language change, they began to assemble systematic collections of the obvious problematic rhyme patterns. Scholars like Wú Yù 吳棫 (1100–1154) started from the idea of sound harmonization and established collections of Chinese characters that belonged to distinct rhyme groups in contemporary rhyme books like the *Guǎngyùn* 廣韻 (published in 1008) but seemed to rhyme freely in the *Book of Odes* (Hé 2006[1985]:163). Once established as a field of philological investigation, the rhyme categories (*yùnbù* 韻部) became more and more refined. Wú Yù only identified nine different rhyme categories (as reflected in his *Yùnbǔ* 韻補 'rhyme addendum'), later scholars identified more than 30 different categories (Baxter 1992:141–150).

Assuming that rhyming was originally rather consistent, with rhyme words being mostly identical in the pronunciation of nucleus and coda, the analysis of rhyme words makes it not only possible to establish rhyme categories but also to interpret them further phonetically or phonologically. The classical approach for rhyme analysis, which is called *sīguàn shéngqiān fǎ* 絲貫繩牽法 'link-and-bind method' (Gěng 2004),[1] or *yùnjiǎo xìlián fǎ* 韻腳系聯法 'rhyme linking method' (Lǔ 2009), consists of roughly two steps: In a first step, groups of Old Chinese words, mostly represented by one Chinese character and identified to rhyme with each other in a given text are collected. In a further step, these groups are compared with each other. If identical words are found in different groups, those groups can be combined to form larger groups. This procedure is then repeated until categories of rhymes can be identified that ideally do not show any more transitions among each other. This approach is essentially similar to the 'linking method' *xìlián fǎ* 系聯法 see Liú 2006:56–67), first proposed in Chén Lǐ's 陳禮 (1818–1882) *Qièyùnkǎo* 切韻考 (1848), by which characters used in *fǎnqiè* 反切 readings in rhyme books[2] are clustered into groups of supposedly common pronunciations for initials and rhymes. In both approaches, similarities in pronunciation are indirectly inferred by spinning a web of direct links between characters.

1 According to Liú (2006) the term was originally coined by Luó Chángpèi 羅常培 (1899–1958).

2 See Branner (2000) for a detailed description of the *fǎnqiè* method in traditional Chinese phonology.

As a concrete example with data taken from Gěng (2004), a couple of rhyme words extracted from the *Book of Odes* that are traditionally all assigned to the classical *zhī* 之 group are given in Table 2 along with their modern Pīnyīn reading. Figure 1 shows the same data, but re-arranges the words in such a way that all co-occurring words in each of the stanzas are placed in the same column. From this arrangement, it can be easily seen that all rhyme groups are connected with each other, either directly, by sharing one or more rhyme words, or indirectly, by sharing one or more rhyme words with the same group. Based on this information, one can conclude that all of the words in Table 2 should be assigned to the same rhyme category.

TABLE 2 *A collection of rhymes in the Book of Odes that are traditionally assigned to the zhī 之 group*

Number	Ode	Rhyme Groups								
27.3.A	邶風·綠衣	sī 絲	chí 治	yóu 訧						
30.2.A	邶風·終風	mái 霾	lái 來	lái 來	sī 思					
33.3.A	邶風·雄雉	sī 思	lái 來							
39.1.A	邶風·泉水	qī 淇	sī 思	jí 姬	móu 謀					
54.4.B	鄘風·載馳	yóu 尤	sī 思	zhī 之						
58.1.A	衛風·氓	chī 蚩	sī 絲	sī 絲	móu 謀	qí 淇	qīu 丘	qī 期	méi 媒	qī 期
58.6.B	衛風·氓	sī 思	zāi 哉							
59.1.A	衛風·竹竿	qí 淇	sī 思	zhī 之						
66.1.A	王風·君子於役	qī 期	zāi 哉	shí 塒	lái 來	sī 思				
130.1.A	秦風·終南	méi 梅	qiú 裘	zāi 哉						
204.4.A	小雅·四月	méi 梅	yóu 尤							
227.2.A	小雅·黍苗	niú 牛	zāi 哉							

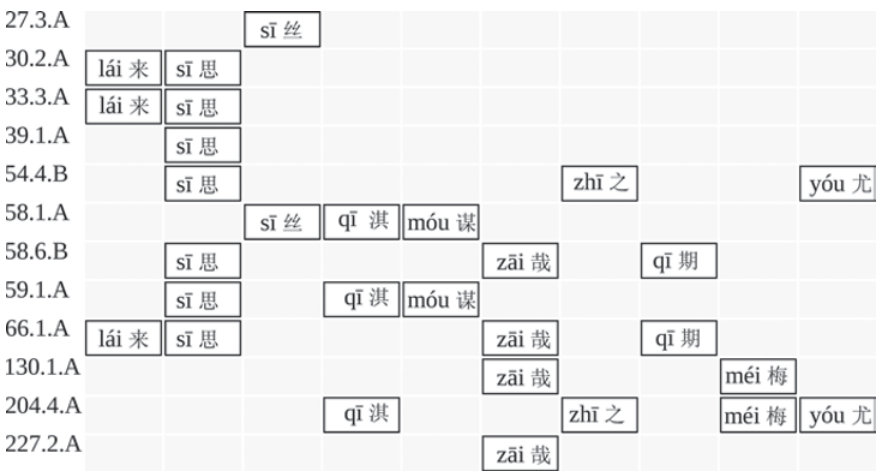


FIGURE 1 *Rhyme groups as a connected component. The figures show all rhyme words given in Table 2 that occur in at least two different stanzas, thereby re-ordering the rhyme words in such a way that identical words are placed in the same column. When comparing these co-occurrences, one can easily see that all groups are connected with each other, either directly or via other groups.*

1.2 *Problems of the Classical Rhyme Analysis*

One striking disadvantage of classical rhyme analysis is its resolution power. If one mechanically follows the idea of connected components in a web of rhyme words, rhyme categories will inevitably tend to become very large, and the number of distinct categories themselves will tend to become very small, since the distinction of rhyme groups requires 'the absence of rhymes of certain types in the corpus' (Baxter 1992:140). But it is not only the absence of evidence that favors lumping over splitting, but the overestimation of the significance of the evidence itself. Since the method does not strictly weight how many co-occurrences across different stanzas are needed to conclude that two words truly rhyme, it is particularly vulnerable for wrongly identified rhyme patterns and irregular rhymes. This is the reason why classical rhyme pattern analysis usually yielded low numbers of categories with large inventories of words (compare the overview in Baxter 1992:150–171). Comparing the number of 20 to 30 rhyme categories on average with the 193 distinct rhymes of the Qièyùn 切韻 (601), or the 206 rhymes in the later Guǎngyùn 廣韻 (1007–1008), which represent almost 100 rhyme categories when ignoring the tones,[3] this would point to a drastic amount of splits in the change from Old Chinese to Middle Chinese. While this does not need to be impossible *per se*, it is not the most realistic assumption regarding the changes from Old to Middle Chinese, and—as Baxter's (1992) refined analysis of the rhymes of Old Chinese has shown—even not a necessary one.

As an example for the vulnerability of the classical method regarding wrongly identified rhyme patterns, Table 3 contrasts the rhyme pattern analyses of the early Qīng scholar Gù Yánwǔ 顧炎武 (1613–1682, quoted after Gěng 2004:15), Wáng (1980), and Baxter (1992). As can be seen from the table, Gù Yánwǔ identifies *léi* 靁 'thunder', sī 斯 'this', *zǐ* 子 'son', and *zāi* 哉 *particle* as words that rhyme with each other in the text, while Wáng marks only 子 and 哉 as rhyming, and Baxter none of the four words. While Baxter's refutation of 子 and 哉 as rhyme words can surely be debated, given that the coda OCH *-ə can be found to rhyme with the coda *-əʔ in other poems of the *Odes*, the separation of 靁 and 斯 from each other and from 子 is crucial for the determination of rhyme categories in general. If one follows the rhyme schema proposed by Gù Yánwǔ, one is forced to either assume that the codas of all words sounded alike and discard the reconstruction proposed by Baxter and Sagart (2014), or one has to assume that the poem consists of irregular rhymes. Since Gù Yánwǔ assumed the former, his analysis inevitably lumped many rhyme words together that were only later shown to belong to different categories (compare the tabular overview in Baxter 1992:156).

TABLE 3 *Comparing differences in rhyme identification for Ode 19.1*《周南·殷其雷》

Text	RW	MCH	Gù Yánwǔ		Wáng (1980)		Baxter (1992)		OCBS-R
殷其靁	*léi* 靁	*lwoj	靁	A, 之部	靁	-	靁	-	*-uj
在南山之陽	*yáng* 陽	*yang	陽	B, 陽部	陽	A, *jiang*, 陽部	陽	A, **ljang*	*-aŋ
何斯違斯	*sī* 斯	*sje	斯	A, 之部	斯	-	斯	-	*-e
莫敢或遑	*huáng* 遑	*hwang	遑	B, 陽部	遑	A, *huang*, 陽部	遑	A, **wang*	*-aŋ
振振君子	*zǐ* 子	*tsiX	子	A, 之部	子	B, *tziə*, 之部	子	-	*-əʔ
歸哉歸哉	*zāi* 哉	*tsoj	哉	A, 之部	哉	B, *tzə*, 之部	哉	-	*-ə

3 When ignoring the tones, the *Guǎngyùn* contains 61 rhymes plus 34 additional rhymes with entering tone (*rùshēng* 入聲), which ended in a plosive coda and was thus segmentally distinct from the other three tone categories.

A further problem of the classical rhyme analysis derives from the interpretation of results. Even if different scholars identify the same words as rhyming in a given poem, their reconstructions may still differ largely, depending on the degree of regularity they assume for the given rhyme patterns. It is well known that rhyming may follow cultural patterns that may diverge from the basic rule claiming that words that rhyme should end in the same sounds. Apart from these regular rhyming habits by which phonetic similarity is broadened (compare, for example, German, where the final [–ɔi] rhymes regularly with [–ai], as in [mai] *Mai* 'May' which regularly rhymes with [nɔi] *neu* 'new', as in *Alles neu, macht der Mai* (the month May makes everything new), spontaneous irregular rhyming may also occur and can be frequently observed in poetry, as it is, for example, frequently met in modern Hip Hop,[4] but also in folk songs, nursery rhymes, and popular music (Zwicky 1976).

An example for the problems of interpretation resulting from irregular rhymes is given in Table 4, where the interpretations of Ode 43.2 by Jiāng Yǒugào 江有誥 (?–1851, quoted after Cáo 2010), Wáng (1980), and Baxter (1992) are contrasted. While Jiāng Yǒugào and Baxter assign *xǐ* 洗 'sprinkle' and *měi* 浼 'flowing water' to the same category, thus distinguishing them from *tiǎn* 殄 'destroy', which consequently forms an irregular rhyme, Wáng however, reconstructs the same reading for all three characters. Furthermore, according to the reconstruction by Baxter and Sagart (2014), *xǐ* and *měi* are reconstructed with the coda *-r*, which is phonetically closer to the coda *-n* in *tiǎn*, thus providing an additional explanation why the irregularity would have been tolerated. Wáng does not give a reason why he reconstructs the same nucleus and nasal coda for all words, although 洗 and 浼 lack nasal endings in Middle Chinese (see the MCH readings in Table 4), but his reconstruction is consistent with the phonetics of the two characters, *xī* 西 and *miǎn* 免 that Wáng also places into his *wén* 文 category (Wáng 1980:23, see also Cáo 2010). As can be seen from this example, it is not only the judgment regarding the rhyme patterns which has a direct impact on the rhyme analysis, but also the interpretation of a given rhyme pattern in terms of the specific pronunciation of the words.

The examples point to a further obvious problem underlying the classical rhyme analysis: its underlying circularity. On the one hand, the rhyme structure of a given poem can only be detected when having initial hints regarding the pronunciation of the words in the poem. On the other hand, the rhyme analysis itself is carried out in order to determine the pronunciation of words in a first instance.[5] Similar to the

TABLE 4 *Comparing differences in rhyme identification for Ode 43.2*《邶風·新臺》

Text	RW	MCH	Jiāng Yǒugào		Wáng (1980)		Baxter (1992)		OCBS-R
新臺有洒	*xǐ* 洒	*srjeX	洒	元部	洒	A, *syən, 文部	洒	A, *s.ij?	*-ərʔ
河水浼浼	*měi* 浼	*mwojX	浼	元部	浼	A, *miən, 文部	浼	A, *mɨjʔ	*-ərʔ
燕婉之求	*qiú* 求	*gjuw	求	-	求	-	求	-	*-u
籧篨不殄	*tiǎn* 殄	*denX	殄	文部	殄	A, *dyən, 文部	殄	A, *dɨnʔ	*-ənʔ

4 Compare Eminem's *Lose yourself* (2002), where *music* [-ik] rhymes with *own it* [-it] in the refrain 'Lose yourself in the music, you own it, you never never let it go, [...]'.

5 As Lǐ and Mài (2008) have illustrated, the very fact that a given collection of poems actually rhymes with high probability, is amenable to statistical investigations. So far, however, no account is known to me, in which an algorithm would be able to automatically identify those words that rhyme in a corpus of Chinese poetry. Although it may even be feasible to employ machine learning approaches for this task, this would go beyond the scope of this paper, and future research needs to show whether automatic approaches could indeed decrease circularity while increasing objectivity in our approaches to Old Chinese reconstruction.

problem of identifying cognate words by identifying sound correspondences and the identification of sound correspondences by identifying cognate words in the classical comparative method in historical linguistics (List 2014:57f), one can circumvent the problem of circularity by applying an iterative procedure in which one starts from a given hypothesis, tests its consequences, and refines the original hypothesis based on the test of the consequences, and this was probably also how the method of rhyme analysis was applied by all scholars in the past. What is important in this context is to note that due to the very nature of rhyme analysis, and the obstacles we encounter when trying to identify which words rhyme in a given poem and whether they rhyme regularly, we should never forget that all analyses of rhyme patterns, be it the one by Gù Yánwǔ, Wáng (1980), or Baxter (1992), are *hypotheses* that may eventually be refined by future analysis. Furthermore, all proposals for the reconstruction of Old Chinese phonology directly based on rhyme analysis should ideally be tested against the rhyme patterns of the *Book of Odes*.

1.3 *Enhanced Approaches to Rhyme Analysis*

In 1992, Baxter proposed a radically new approach for the analysis of rhyme patterns, based on probabilistic arguments. The key idea was to use the evidence drawn from the co-occurrence of rhyme words in order to test hypotheses regarding the grouping of rhyme words into rhyme categories. The basic way to model co-occurrences was a statistical test that checked how likely it was that a given grouping had occurred by chance. As a result of this method, Baxter proposed 52 rhyme groups for the 31 rhyme groups that had been identified by the classical approach to rhyme analysis. One of the core features of this system was the six vowel hypothesis which stated that Old Chinese had only six different vowels, which Baxter originally reconstructed as **a*, **e*, **i*, **o*, **u*, **ɨ*. Interestingly, this hypothesis was independently proposed by other scholars at about the same time (Starostin 1989, Zhèngzhāng 2003)[6] thus providing independent evidence for the validity of this hypothesis.

At around the same time, Zhū (1989) proposed similar statistical tests to distinguish whether words that rhyme in a text form coherent groups,[7] applying these methods to poems from the Northern Sòng 北宋 dynasty (960–1126). Zhū's methods were further applied to different datasets in order to investigate rhyme behavior, rhyme practice, and major rhyme groups in different varieties and historical stages of Chinese, including early Cantonese (Cheng 2004), contemporary Cantonese (Cheng 2009), and Suí 隋 dynasty (581–618) Chinese (Mài 1999).

One obvious disadvantage of Baxter's and Zhū's approach is that they can only be used to test given hypotheses on the data, not to derive hypotheses from the data. In order to apply the tests, one needs to have a hypothesis, and it is not possible to use the approaches to generate potentially fruitful hypotheses. This means that, in the end, the scholars are still left to the painstaking work of inspecting a multitude of rhyme patterns in a corpus of poems in order to come up with valid hypotheses that could then be tested. Given the necessarily iterative character of rhyme analysis, it would be desirable to take some shortcut to derive initial hypotheses regarding the data without being forced to inspect poem after poem, and stanza after stanza.

In many branches of science, it is nowadays common to employ methods of *exploratory data analysis* prior to deriving any initial hypotheses (Morrison 2011:51–56). Exploratory data analysis does not involve any significance testing, but rather seeks to assist scientists in analyzing their data before they come up with a 'real', full-fledged conclusion. In this respect, exploratory data analysis lacks the elegance of statistical models. Its advantage is, however, is that it offers quick access to large datasets and assists

6 Zhèngzhāng's ideas were previously published in separate papers starting from the late 1980s.

7 See Cheng (2009:36–42) for an overview on a range of similar methods and further developments by different authors.

scientists in developing initial hypotheses regarding their data. The following approach can be seen as an example of exploratory data analysis, insofar as it does not involve rigid statistical testing of hypotheses. Instead, it simply tries to harness common computational approaches to network analysis for the evaluation of Chinese rhyme data. Nevertheless, as shall be shown in the application of the approach in Section 4, the weighted network perspective on Chinese rhyme data does not only offer multiple ways to explore large *corpora* of Chinese poetry, but it can also be used to test and enhance given reconstruction proposals.

2 Network Modeling of Rhyme Patterns

2.1 *Networks*

From a computational perspective, networks are a specific data structure that consists of nodes and edges. In a network (or graph), nodes (or vertices) represent objects, and edges (or links) represent relations between the objects. The information regarding the nodes and the edges in a network can be further enriched by tagging, weighting, or labeling the nodes and by labeling, weighting, or directing the edges. Network representations are used for a multitude of different tasks in science in general as well as in linguistics specifically. In social sciences, networks can be used to model social interaction (Ahn et al. 2010). In molecular biology, *gene similarity networks*, in which genes represent the nodes in a weighted undirected network and edges between the nodes reflect the degree of sequential similarity between the genes, can be used to infer deep homologies, lateral gene transfer events, and instances of gene fusion (Alvarez-Ponce et al. 2013). In historical semantics, cross-linguistic polysemic patterns can be modeled with the help of weighted undirected networks in which nodes represent concepts and edges the

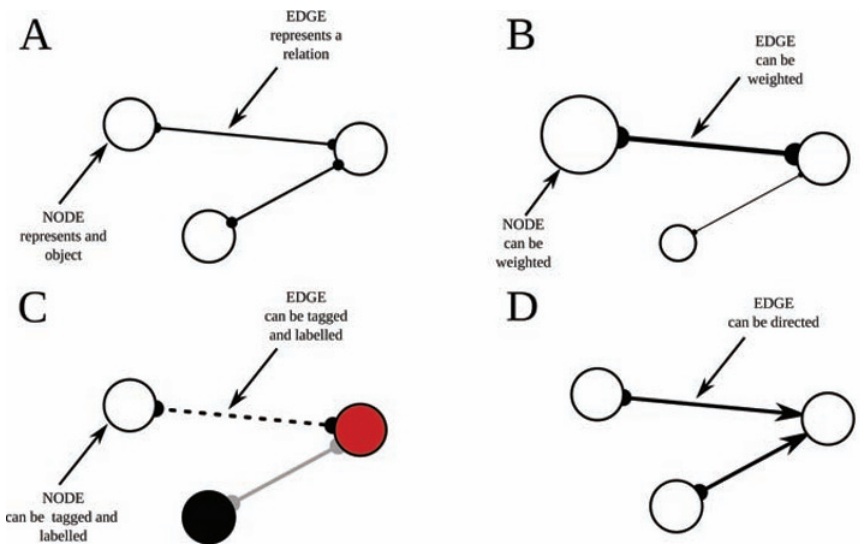


FIGURE 2 *Basic properties of networks. A shows a simple undirected and unweighted network. B shows a weighted network in which the node size represents the node weights and the edge width represents the edge weights. C shows a network in which node and edge labels are represented through the use of different colors for the nodes and different colors and stroke styles for the edges. D shows a directed network.*

frequency of cross-linguistically attested polysemies (List et al. 2013 and 2014). In automatic approaches to sequence comparison, sequence similarity networks, in which nodes represent words and weighted edges represent the phonetic similarity between words have been successfully applied to identify cognate morphemes in Sino-Tibetan language data (List et al. 2016). Figure 2 illustrates some basic properties of networks, such as weights, labels, and directions.

2.2 *Networks of Rhyme Patterns*

The key idea behind classical rhyme analysis is that although two given words may not rhyme in a given collection of poems, we can still find evidence that the words sounded alike by looking at the other words with which they rhyme. In this sense, the classical rhyme linking analysis, as well as the above-mentioned linking analysis applied to *fǎnqiè* readings, are true network approaches. They both infer similarities among research objects from their *connections* to other research objects.

In a simple network model of rhyme patterns, every word that rhymes in a given set of poems can be represented as a node of a network. Edges between the nodes are drawn whenever two words rhyme in a stanza. Edges can further be weighted by counting the frequency of stanzas in which a given pair of words rhymes. The advantage of such a representation is that one exhaustively represents all the evidence that can be drawn from a set of poems, including potentially irregular or spurious rhymes. This increases the objectivity, since it automatically reduces the bias when analyzing the network. As mentioned earlier in footnote 5, the disadvantage is that the data structure cannot be simply reproduced automatically, since it would be difficult to design an algorithm that is capable of inferring both pronunciations and rhyme patterns at the same time, especially since the data is probably too small for the application of machine learning approaches. For this reason, the construction of rhyme networks requires that all rhyme words in a set of poems are identified manually. This holds especially for Chinese collections of poems, like the *Book of Odes*, since the Chinese characters, when taken in isolation, do not give us any hint regarding their pronunciation. Figure 3 illustrates how a rhyme network can be reconstructed from a set of Chinese poems taken from the *Book of Odes*.

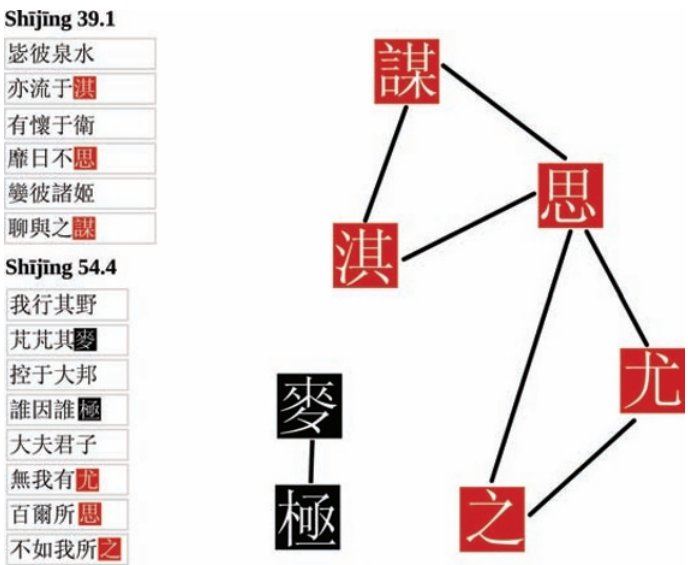


FIGURE 3
Reconstruction a rhyme network from the Book of Odes*. The figure shows two stanzas along with the identified rhymes, illustrated by coloring the background of the Chinese characters. The word* sī 思 *'think' occurs as rhyme word in both Ode 39.1 and 54.4 in the example. As a result, the groups of three rhyming words in the two stanzas can be linked with each other, and form a large cluster of rhyme words.*

3 Constructing a Network for the *Book of Odes*

3.1 *Material*

3.1.1 Data Preparation

The Shījīng version which was used employed in this study was originally taken from the Project Gutenberg (http://www.gutenberg.org/ebooks/23873). This was a purely pragmatic choice, since no other digital sources known to the author were available in full text along with a free license by the time of the preparation of the data. Since the Gutenberg version of the *Book of Odes* is not free of errors, lacking some rare characters and diverging from other editions of the Shījīng, it was only taken as a starting point, and the rhyme words in all 305 poems were thoroughly compared with the version in the Shísān Jīng Zhùshù 十三經註疏 (digitally available via the Chinese Text Project, http://ctext.org, and the CHANT project, http://chant.org). The digital version of the *Book of Odes* was manually annotated by adding the rhyme patterns presented in Baxter (1992). All rhyme words (including potential rhyme words) were further annotated by adding the reconstructions of the Old Chinese character readings by Baxter and Sagart (Baxter and Sagart 2014),[8] along with Middle Chinese readings, following the system of Baxter (Baxter 1992), and reconstructions by Pān Wùyún 潘悟雲 (see Pān 2000), as given in the Thesaurus Linguae Sericae (Harbsmeier and Jiang 2009, http://tls.uni-hd.de). It was not in all cases possible to provide full information for the characters. In some cases, the Middle Chinese readings or the reconstructions by Pān Wùyún were missing in TLS, and in some cases, the Old Chinese reconstruction by Baxter and Sagart (2014) was not available. There are several reasons why the respective data is missing. In some cases, the characters were not included in the available reconstructions, like, for example, the character *jiē* 罝 'net for catching rabbits' (Ode 7.1-3), which is missing in Baxter and Sagart (2014), or *jǔ* 筥 'round basked' (Odes 15.2, 222.1, and 291.1), which could not be found in the reconstructions of Pān Wùyún in the TLS. In other cases, the digital version of the Shījīng, or the resources on Old Chinese phonology might differ in specific character variants. Since the recent reconstructions of Old Chinese involve a multitude of different data sources as well as a very good knowledge of how to weight the different pieces of evidence against each other, no attempts were undertaken to fill these gaps for this study. Since the amount of missing data is rather small (around 5% for OCBS reconstructions), future research might quickly fill the gaps. This should, however, not be done without double-checking the cases with the parents of the different Old Chinese reconstruction systems.

The Shījīng is traditionally divided into four parts which were compiled in different centuries. It comprises a total of 305 different poems, which contain as many as 1,142 stanzas. A stanza can be further divided into multiple *verses*, which roughly constitute a self-contained unit of thought, and a verse can be subdivided into multiple *sections*, which constitute a unit that may potentially contain a rhyme word.[9] According to our data, the Shījīng contains 3,518 verses, and 7,285 sections. The raw data which was compiled for the study is represented in tabular form in which rows represent rhyme words and columns contain the valid information concerning these rhyme words, such as their poem, stanza, rhyme, and

8 The author is deeply indebted to Laurent Sagart and William Baxter for sharing the data on character readings as well as digital versions of the Shījīng. Their character readings are now also officially available in digital form and can downloaded at http://ocbaxtersagart.lsait.lsa.umich.edu/.

9 The term *section* is probably not the best choice for this unit, yet it was taken due to a lack of alternatives for the moment. William Baxter calls the smallest unit of text blocks that usually contains the rhyme words a "line". This may be confusing for the reader, since a "verse" in the notion adopted here usually occupies a *line of a text*. For this reason, the term section was chosen to denote those units in the poems of the Shījīng which potentially end in a rhyme word.

section number, rhyme pattern information, and phonological information regarding the rhyme words. Note that this corpus does not only contain the rhyme words that were identified as such by Baxter (1992), but also *potential* rhyme words. These were automatically determined by taking the last character of each section in the corpus. If the last character turned out to be one of the frequently occurring functional characters, like *zhī* 之 *grammatical marker*, *zhǐ* 止 'to stop', or *yě* 也 *final particle, yǐ* 矣 *id.*, the character *preceding* these words was tagged as potential rhyme word. The raw data is given in the Supplementary Material accompanying this paper. Figure 4 shows how the first stanza of Poem 4 of the *Book of Odes* is represented in tabular form.

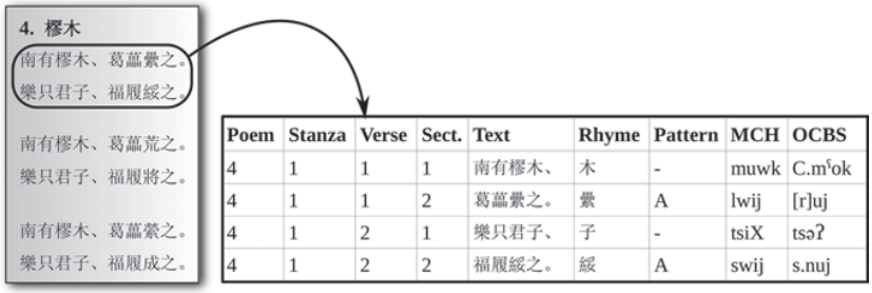
3.1.2 Interactive Shījīng Browser

In order to make it more convenient for the readers to investigate the data underlying this paper in full detail, an interactive web-based application was created. This freely available Shījīng Browser[10] lists all potential rhyme words in tabular form along with additional information including the *pīnyīn* transliteration, the Middle Chinese reading, the reconstruction by Baxter and Sagart (2014), the reading by Pān Wùyún, the GSR index (Karlgren 1957), and the number of poem, stanza, and section. With help of interactive search fields, the data can quickly be filtered, enabling the users to search for specific poems, for specific characters, or for specific readings. When clicking on the "Poem" field in the application, a window pops up and shows the whole poem, in which all rhyme words are highlighted. In certain cases, where potential alternative rhymes were identified, this is marked in an additional column.

3.2 *Methods*

3.2.1 Preliminary Thoughts

When analyzing rhyme patterns from which to reconstruct a network, one needs to be careful to not overstate our hypothesis regarding the closeness of rhyme words. For example, a common motif in the *Book of Odes* is the repetition of certain lines throughout all stanzas of a poem. This way of structuring poems through the repetition of certain parts across stanzas is quite common in the *Book of Odes*, and



Poem	Stanza	Verse	Sect.	Text	Rhyme	Pattern	MCH	OCBS
4	1	1	1	南有樛木、	木	-	muwk	C.mˤok
4	1	1	2	葛藟纍之。	纍	A	lwij	[r]uj
4	1	2	1	樂只君子、	子	-	tsiX	tsəʔ
4	1	2	2	福履綏之。	綏	A	swij	s.nuj

FIGURE 4 *Structured representation of the data in the* Book of Odes*. On the left, the Poem 4,* 《國風・樛木》*, is shown as it is represented in the digital text. On the right, the first stanza of the poem is represented in tabular form. The numbers for poem, stanza, verse, and section are all indicated in separate columns. The column "Rhyme" shows all potential rhyme words, and the column "Pattern" the rhyme words identified by Baxter (1992). MCH and OCBS show Middle and Old Chinese readings, the former following Baxter (1992), and the latter following Baxter and Sagart (2014).*

10 The browser is accessible at http://digling.org/shijing/.

lines are not only repeated across multiple stanzas inside one poem, but may even re-occur in different poems. The bias that may be introduced by repetitions should not be underestimated. Of the 7285 lines in the Shījīng corpus, only 6068 are unique. An extreme example is the line *xīn zhī yōu yǐ* 心之憂矣 'the sorrow of my heart' that occurs in 11 different poems (26, 27, 63, 109, 150, 183, 192, 197, 207, 233, 264), and as many as 26 different stanzas. While the frequency of this line is less problematic for the Shījīng network reconstruction, since the line barely rhymes in any of the poems in which it occurs, we can also find many recurring line pairs, most of them within the same poem, but also more than 40 pairs which rhyme across two and more poems. One example is the line pair *yè bǐ nán mǔ* 饁彼南畝 'bringing food to the fields in the south' and *tián jùn zhì xǐ* 田畯至喜 'causing the supervisor of the field to be happy'. This pair reoccurs as many as three times (Odes 154.1, 212.4, and 211.3).

It seems to be obvious that these cases need to be handled in some way, since they probably reflect formulaic language, and one can never be sure whether the rhyme pattern reflects the natural rhyme intuition of the poets or was just an allusion to earlier or common texts. How to treat concretely these cases, however, is not a simple question. One could either count repeating lines only once, or one could ignore them completely. Both approaches have advantages and disadvantages. When counting the lines once, one may still capture an unwanted signal, since they could reflect fossilized or dialectal speech. On the other hand, one could expect that repeated lines in a poem are specifically "pure" in their rhyming, similar to refrains in modern songs, and poets probably spent more time on the creation of repeating lines than on the creation of the rest of the lines in a poem. For this reason, it seems better to not completely ignore repeating lines, but to decrease their influence on the whole network by counting them only once.

Two further problems one should keep in mind are (a) identical rhyme words occurring in the same stanza, and (b) the overall size of the rhyme groups in which the rhyme words occur. Identical rhyme words occurring in the same stanza should, of course, be counted only once, since they would otherwise suggest a closer affinity between the words that occur multiple times and the rest of the words with which they rhyme. As an example, consider Ode 17.1, in which *lù* 露 'dew' occurs in the first and the third line, rhyming with *yè* 夜 'night' in the second line. If we did not normalize by the repeated occurrence of identical rhyme words in the same stanza, we would count two instances in which 露 rhymes with 夜, although, in fact, the poet probably did not decide twice that the two words rhyme nicely with each other. As to the size of rhyme groups, we also need to be careful of over-counting our evidence, as the number of links drawn between characters exponentially increases, the more rhyme words are found in a rhyme group. If there are only two words in a rhyme group, only one link will be added to the network; if there are three rhyme words, three links will be added. In the instance of four words, six links are added, and for five, ten will be added, etc. The problem is that rhyme words in large rhyme groups will seem to be perfectly integrated in the whole rhyme network, since the they are all interlinked with each other, even if they occur only once. This makes large rhyme groups very vulnerable to irregular rhymes. In order to cope with this problem, it is important to normalize the data, and to reduce the weight one adds to a link in the network in proportion to the size of the rhyme group. A simple way to do so is to divide each co-occurrence of two rhyme words by the size of the group of rhymes with which they rhyme. This results in Formula

$$W_{AB} = \frac{1}{G_{AB} - 1}, \tag{1}$$

where w_{AB} is the weight added to the overall score of character concurrence of characters A and *B*, and G_{AB} is the size of the rhyme group in which *A* and *B* co-occur. In this normalization, the highest value for

w_{AB} is exactly 1 for the minimal group size of 2, and the *weighted degree* of each node (the sum of the weight of all links of a node) is equal to its occurrence as a rhyme word in the whole corpus.

Apart from weighting the links in our network of rhyme connections, one can also weight the nodes. Here one could either weight the nodes by counting how often they occur in rhyme groups throughout the Shījīng corpus, or count how often they occur in potential rhyme positions in all sections of the corpus. The latter value may be more interesting, since it reflects the more general distribution of potential rhyme words in the corpus of ancient Chinese poems. If a given rhyme word, for example, occurs very often in a potential rhyme position, but rarely rhymes with other words, the few instances when they do rhymes should be treated with a certain suspicion. Moreover, as demonstrated later in this current study, there are algorithms for network partitioning that include information regarding the weight of the nodes in a network.

3.2.2 Rhyme Network Construction

Based on the preliminary thoughts discussed in the previous section, the rhyme network was reconstructed as follows:

1. All characters occurring in the Shījīng corpus in a rhyme group, according to the annotation by Baxter, were represented as nodes in the network.
2. Links between two characters were drawn whenever they occurred in a group of rhyme words.
3. Edge weights between two characters *jiǎ* 甲 and *bǐng* 丙 were derived from the number of times both characters occurred in the same rhyme group in different stanzas of the Shījīng.
4. Node weights were derived from the number of times each of the characters occurred in a potential rhyme position in the Shījīng corpus.

The edge weights were further normalized as follows:

1. The co-occurrence of two characters 甲 and 丙 in a given rhyme group was only counted once per stanza, no matter how often 甲 and 丙 occurred in the rhyme group.
2. All co-occurring sections were stored in memory and only counted the first time they appeared.
3. The concrete values for the co-occurrence of two characters in a rhyme group were normalized by applying Formula (1).

3.2.3 Implementation

The code was implemented in Python. A couple of third party libraries are required to run the code, including LingPy (http://lingpy.org, List and Forkel 2016), which was used for data representation and handling, and Networkx (Hagberg 2009, http://networkx.org) which was used for common network operations and network representation, and igraph (http://igraph.org, Csárdi and Nepusz 2006), a high-level C-library which that offers state-of-the-art methods for network analysis along with a Python interface. The source code along with the input data and further instructions on how to replicate the network reconstruction, as well as further analyses described below, is provided along with the Supplementary Material accompanying this paper.[11]

11 The source code along with further instructions can be downloaded from this link: https://zenodo.org/badge/latestdoi/43676744.

4 Analyzing the Rhyme Network

4.1 *General Network Properties*

The entire network reconstructed from the Shījīng rhyme corpus is shown in two flavors in Figure 5. It comprises 1845 nodes and 5266 links between the nodes. Given that the theoretical number of links between the nodes would be

$$\frac{n^2 - n}{2} = \frac{1845^2 - 1845}{2} = 1701090, \tag{2}$$

the density of the network, that is, the proportion of attested links compared to possible links, is very small, amounting to less than one percent.[12] It is the more interesting that the network is almost entirely connected, with one very large connected component of 1,539 characters, comprising 83% of all nodes of the network, and one smaller connected component of 67 characters, comprising 4% of all nodes of the network, and several smaller ones. In total, the network consists of 90 connected components, most of them consisting of less than five characters. Some statistics regarding the size of the clusters (connected components), the total number of clusters of one given size, and the coverage with respect to the complete network are given in Table 5.

The fact that rhyme networks are almost completely connected comes as a bit of a surprise. Given the number of rhyme categories identified by the classical rhyme analysis, one might expect a clearer separation of at least the categories that seemed to be rather obvious to the classical scholars. Yet when taking

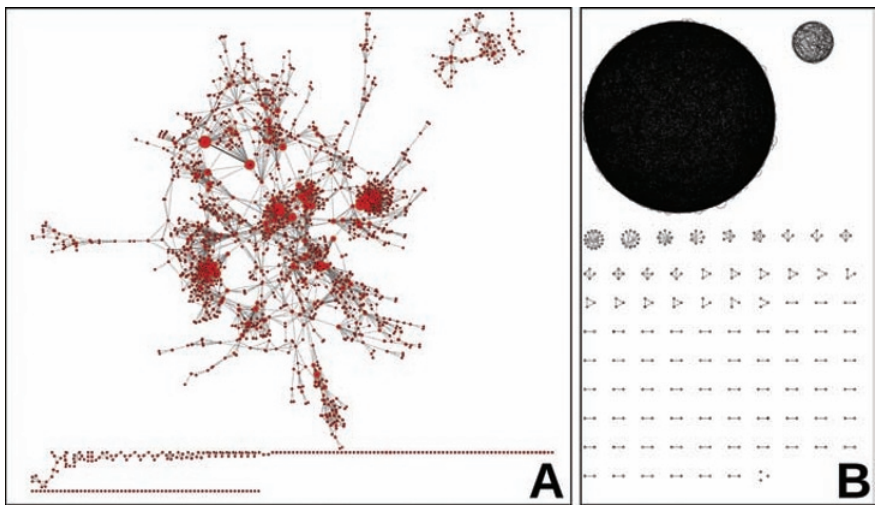


FIGURE 5 *The full network reconstructed from the Shījīng in two different views. A is based on a force-directed layout, and B shows the connected components of the network. The graphic was created with Cytoscape (Smoot et al. 2011: http://cytoscape.org). The size of nodes is proportional to the node weight, as measured by the number of occurrences in the corpus. The width of edges is proportional to the edge weight, measured as the normalized number of co-occurrences of characters in different rhyme groups across different stanzas.*

12 Exactly 0.3% when one divides the number of 5266 attested edges by the number of 1701090 potential edges.

TABLE 5 *General statistics regarding the network structure. The table shows the results of a connected component analysis of the network, contrasting the cluster size, the number of components with a given cluster size, the total number of characters with a certain cluster size, and the coverage compared to the number of characters in the whole network. As can be seen from the numbers, 83% of all characters occur in one big connected component.*

Cluster Size	1	2	3	4	5	7	8	12	14	67	1539
Number of Components	3	59	13	7	2	1	1	1	1	1	1
Total Nr. of Characters	3	118	39	28	10	7	8	12	14	67	1539
Coverage	0.00	0.06	0.01	0.02	0.01	0.00	0.00	0.01	0.01	0.04	0.83

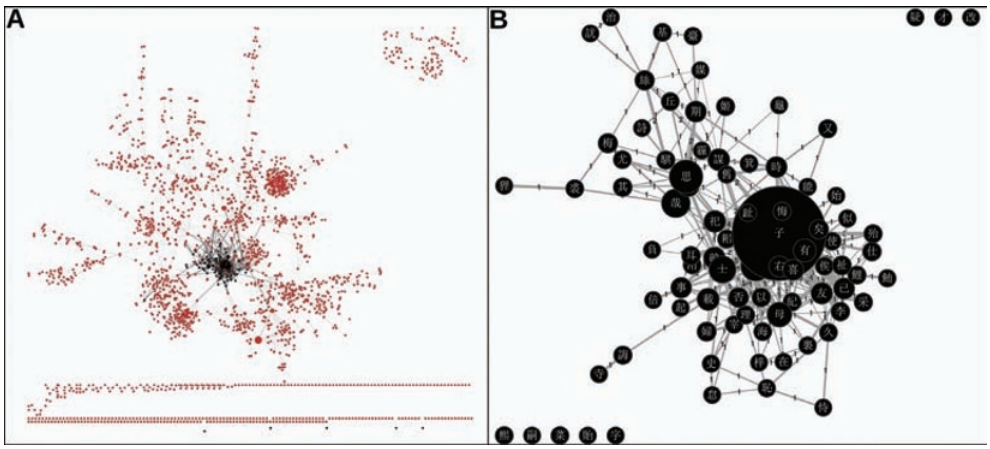


FIGURE 6 *The location of the* zhī 之 *category in the network (A), and the connections between the members of the* zhī 之 *category (B). Rhyme words identified as belonging to the* zhī 之 *category (*-ə according to the reconstruction by Baxter and Sagart 2014), are shown in black.*

a closer look at the large connected component in the network, one can easily see that this component is itself structured, consisting of several groups which are more densely connected among themselves than with other groups. The *zhī* 之 rhyme group for example, which was briefly mentioned above (see the characters in Table 2), occupies the center of the main component of the network, as shown in Figure 6A. When looking at this group in isolation, we can further see that most of the rhyme words form a connected component, with only a few outliers that do not occur in the biggest component of the network. This demonstrates that the structures postulated by the scholars can definitely be found in the network. However, in order to test and investigate these structures further, one needs to turn to more specific methods of network analysis.

4.2 *Communities and the *-r-Coda*

In network approaches, natural groups of similar objects are called *communities*. More specifically, the term *community*, which obviously stems from social network analysis, denotes groups of nodes in a network among which the number of connections is large, while the number of connections to the "outer world" is much smaller (Newman 2004). One should, of course, always be careful making analogies between different domains. The analogy between social communities and rhyme groups, however, turns

out to be quite fruitful. As mentioned above, classical rhyme analysis tries to determine groups of similar objects in the network of rhyme connections by searching for *connected components*. Connected components, however, are very vulnerable to irregular rhymes, as demonstrated in Section 4.1, where it was shown that the biggest connected component of the network reconstructed from our Shījīng corpus contains as many as 83% of all nodes in the network. The analogy with communities, however, allows us to refine the strong claim of connected components. One no longer requires that all rhyme words connected in the network form a cluster of words with a similar pronunciation, but instead it can be proposed that the identified clusters should have more common edges among themselves than with other nodes outside the cluster. The task of identifying rhyme categories in a network of rhyme patterns can thus be modeled as a community detection task.

We still need to be careful regarding the analogy between communities and rhyme categories. Rhyme categories are usually thought to represent clusters of words with similar finals, and similarities are defined as commonalities between the nucleus and the coda of the words. When this is the case, however, it is by no means necessary that words with similar pronunciation in their finals *actually* rhyme in any collection of texts. However, it may well be possible that distinct communities discovered in this network still have the same pronunciation. The reasons for this are manifold. It could be pure coincidence that connections are not made, but the semantics of rhyme words might also play a role, forcing groups of words from incompatible semantic or pragmatic domains to never rhyme with each other.[13]

One could think of similar social communities in different geographical locations here: the communities of football fans from Manchester and London are beyond doubt tightly connected in Manchester and London, but due to geographic distance and typical rivalries among football fans, it may be hard to find a football fan from Manchester who befriends a football fan from London in social networks. In order to find the missing link between similar but separated communities, further data is required. The similarity between the Manchester and the London community of football fans could, for example, be shown by comparing the itineraries of the representatives, which may show that both regularly attend the same football stadiums. For our rhyme networks, we could include information regarding the phonetics of the characters, which can give us additional hints regarding their phonetic similarity. This investigation would, however, go beyond the scope of this paper, and it is left for future research to pursue and test the fruitfulness of networks constructed from mixed data types.

Even by eyeballing the big rhyme network shown in Figure 5, we can see that the big connected component has an inherent community structure. With help of algorithms for community detection, we can make this structure become transparent. For this purpose, an *Infomap* community detection analysis was carried out on the rhyme data. *Infomap* is (Rosvall and Bergstrom 2008) is a fast community detection algorithm with a very good performance. It handles weighted nodes and weighted edges, and uses random walks through the network in order to find the best partition of the network into communities.

The *Infomap* analysis splits the network into 345 distinct communities of varying size. This number is much higher than the 59 basic rhyme categories proposed by Baxter and Sagart (2014). This is, however, not surprising. Firstly, these 59 basic categories do not include information on *shǎng* 上 tone (usually reconstructed as a glottal stop coda *-*ʔ*), and especially the *shǎng* tone had a considerable impact on rhyme behavior. Secondly, the network originally consisted of 90 connected components that community detection algorithms will automatically keep separate, since there is no evidence to connect the

13 It is probably easier to substantiate this claim by showing that semantically or pragmatically connected rhyming words rhyme more frequently. It also may be quite possible that one could find two different clusters of words with similar pronunciation in collections of poetry that never effectively rhyme with each other due to their semantics or pragmatics.

groups further. In contrast, the largest component of the network is clustered into 90 components by *Infomap*, which comes quite close to the proposed number of 59 rhymes in Baxter and Sagart's reconstruction when keeping in mind that the *shǎng* tone often rhymes in separate groups. This is further illustrated in Figure 7, which shows how the six largest communities detected by *Infomap* are distributed over the network. In the figure, the category labels for these six categories were determined by taking the most frequently occurring rhyme in the reconstruction by Baxter and Sagartas representative of the whole group. As can be seen, from the figure, *Infomap* proposes a group labeled **-əʔ*, thus underlining the importance of the glottal stop coda for the rhyming behavior of old Chinese words.

The results of this community detection analysis are available as an interactive web-based application.[14] In this application, the users can search for certain communities by their unique identifier, or by filtering by character, or by one or more of the codas (with *shǎng* tone coda being counted separately) following the reconstruction system by Baxter and Sagart. The application displays all different codas

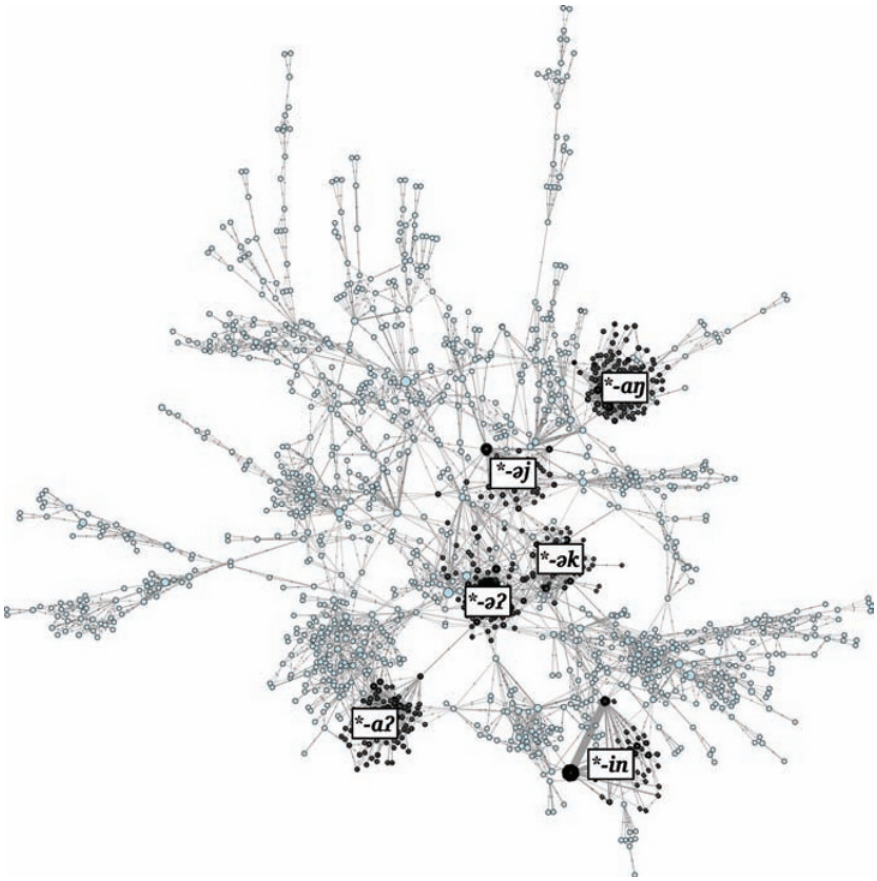


FIGURE 7 *The distribution of the six largest communities inferred by the* Infomap *algorithm over the network. The labels, following the reconstruction by Baxter and Sagart (2014) are determined in a majority-rules fashion by taking the most frequently occurring coda per community as the representative value of the whole set of characters.*

14 Available online at http://digling.org/shijing/infomap.html.

accounted for in a given *Infomap* community. It further lists the size of the community and enables the viewing of all characters constituting a community on one click, giving further information on Middle and Old Chinese character readings as well as the occurrence of the respective characters in potential rhyme positions in the Shījīng.

What is particularly surprising is the high- resolution power of the *Infomap* analysis. As an example, consider the communities numbered 2 and 10, which were given the labels **-əʔ* and **-ə*, respectively. While the frequent co-occurrences of words ending in **-əʔ* and words ening in **-ə* in the *Book of Odes*, make it difficult to draw the border between both communities by eyeballing the visualization of the network in Figure 8, the communities are highly consistent with respect to the Baxter and Sagart's reconstruction system. Of the 74 words assigned to community 2, those directly reconstructed as **-əʔ* are 59 in number, six words are reconstructed as **-ə*, and nine words are assigned further different values. Of the 39 members of community 10, those reconstructed as *-ə, total 30 while nine are reconstructed to other values, with none of the values being **-əʔ*.[15] Some readers might think that these scores are rather low,

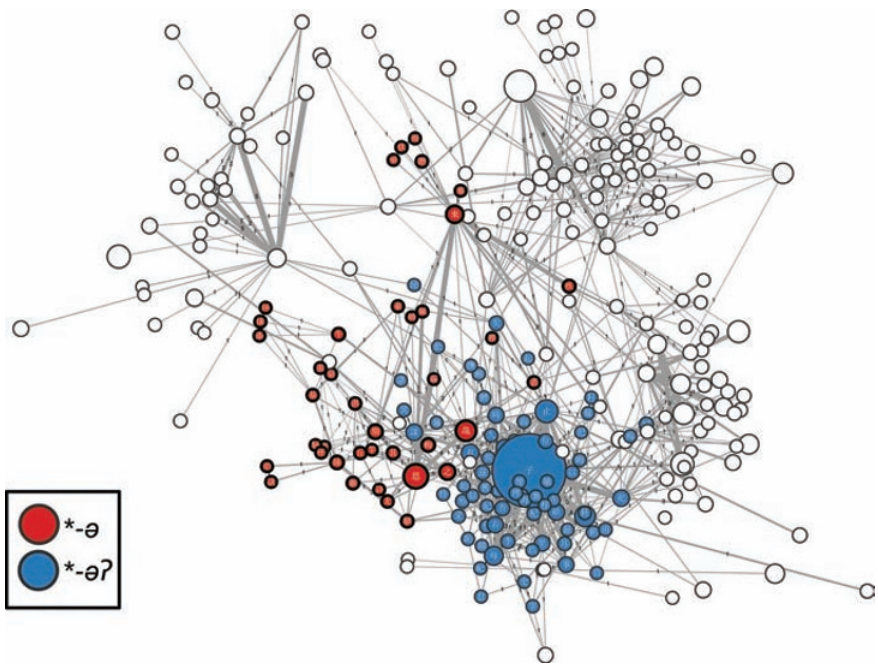


FIGURE 8 *Resolution power of the* Infomap *analysis, exemplified by comparing the communities 2 (*-əʔ) and 10 (*-ə). The frequent co-occurrences of words ending in *-əʔ and words ending in *-ə in the* Book of Odes *makes it difficult to draw the border connection between both communities by eyeballing the visualization of the network, the communities are highly consistent with respect to Baxter and Sagart's reconstruction system as reflected by the fact that only four words out of 74 assigned to community 2 are reconstructed as *-ə (with a total of nine words being assigned different values than *-əʔ).*

15 Due the missing data for Old Chinese reconstructions mentioned in Section 3, values for 16 missing values in community 2 and 11 missing values in community 10 were determined by re-checking the character readings manually with Baxter (1992).

given that only 80% of identical characters for community 2 and 77% for community 10 were reached. One needs to keep in mind, however, that the algorithm is solely based on the analysis of the rhyme patterns, and no additional evidence, be it Middle Chinese readings, the phonetics of the characters, or Sino-Xenic readings, was used. Furthermore, as it was mentioned above, the realistic limits of our analogy between communities and rhyme categories need to taken into consideration. Not all words that show a similar rhyme behavior are necessarily also similar or identical in the pronunciation of their finals. The uniformity of the evidence sets the limits for community detection approaches applied to rhyme networks, but it remains a very useful starting point for both exploratory rhyme analysis, and for the testing of specific hypotheses.

How can specific hypotheses be tested or refined with help of the *Infomap* cluster analysis? As an example, let us have a closer look at those nodes in the network for which Baxter and Sagart reconstruct the rhymes **-ar,* **-an*, and **-aj*. In the Baxter-Sagart system of Old Chinese reconstruction, the rhyme analysis of Baxter (1992) was left largely unchanged, with exception of the additional coda **-r*, which is proposed to account for certain rhyme connections between the codas **-n* and **-j*, occurring in a couple of stanzas in the *Book of Odes*. This hypothesis goes originally back to a proposal by Starostin (1989), and has been constantly gaining acceptance among researchers (Hill 2014, this volume). Since the **-an* and **-aj* are the most frequently occurring finals with the coda **-n* and **-j*, they seem to be a good starting point to test the hypothesis that an additional coda **-r* should be proposed.

The test proposed in this paper is fairly simple. Figure 9 illustrates the basic idea in more detail. In a first step, a subnetwork is reconstructed from all nodes which rhyme in **-ar*, **-aj*, and **-an*. In order to make it easier to inspect the visual representation of this subnetwork, different colors are used to label the nodes. When looking at the force-directed visualization in Figure 9A, it is possible to recognize three larger groups in the network. As the color labels show, however, these groups do not seem to be really homogeneous and we find **-ar* rhymes in both the **-aj* group on the left and what could be an **-an* group on the right. When referencing the inferred *Infomap* communities in Figure 9B, this picture changes slightly, and the communities suggest an increased homogeneity of the groups. Some communities, however, remain mixed, as the one highlighted in Figure 9B (community 19 in the web-application) which is also shown in larger resolution in Figure 9C. When investigating this community closer, however, it turns out that the mixed status is due to an artifact of the data. One of the specific features of the reconstruction system by Baxter and Sagart is that they are very explicit about *uncertainties* of their reconstruction. In cases where evidence is not found to be sufficient to decide for one value only, they propose a tentative reconstruction value but put it in square brackets, thus making clear that they are not completely sure about the validity of the claim. In the case of codas ending in **-n*, this means that they could likewise be interpreted as codas ending in **-r*, and the spelling **-[n]* in the reconstruction system serves as a mere placeholder to distinguish the coda from other rhyme groups (such as *rù*-tone codas ending in **-p*, **-t*, **-k*, for example). The original annotation in the Baxter and Sagart reconstruction, allowing for both the final **-an* and **-ar*, is reflected in Figure 9D, where an additional label color for uncertainties has been introduced. These are displayed as **-a[n]* in the reconstruction by Baxter and Sagart, but here being more readily interpreted as **-an* or **-ar*. It turns out that all the nodes that were labeled as **-an* rhymes in Figure 9C are indeed cases of uncertainty, and the whole cluster seems to represent a true **-ar* rhyme. The *Infomap* analysis not only justifies the uncertainty displayed in the Baxer and Sagart reconstruction, it also provides us with new suggestions regarding the reconstruction of the finals of this cluster. This example shows that we can actively use the *Infomap* community analysis to review, test, and improve given reconstruction systems.

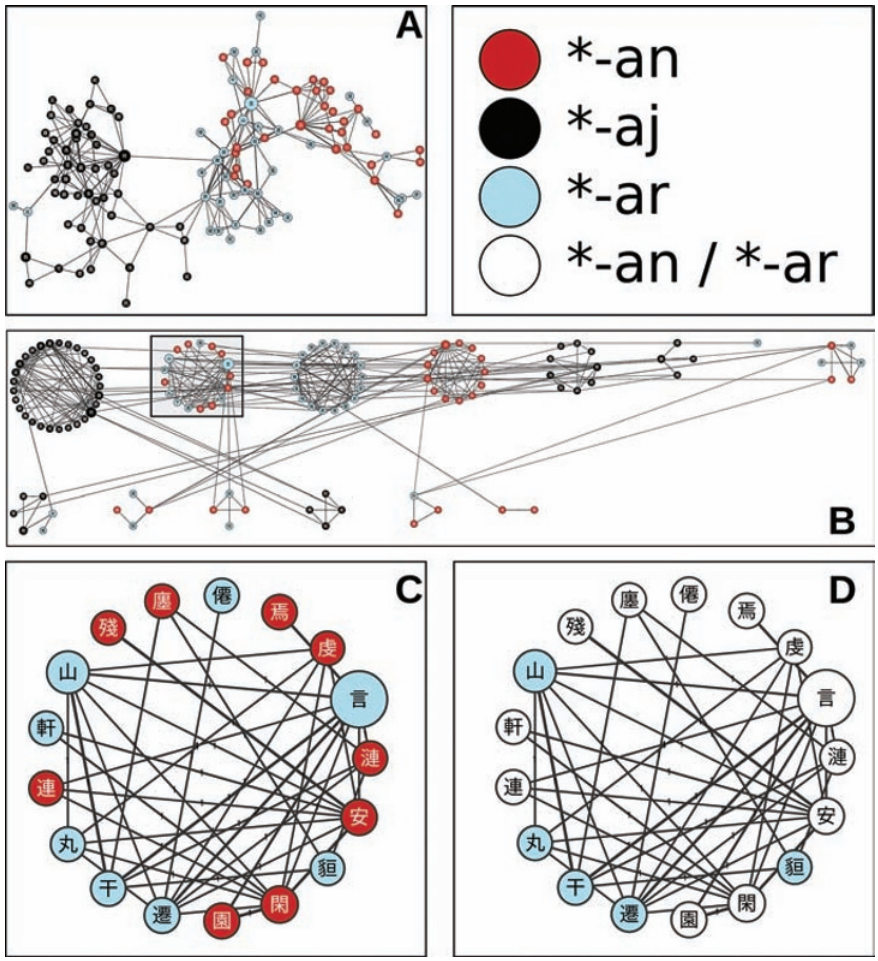


FIGURE 9 *Testing the *-r-rhyme hypothesis of Old Chinese. The data shows rhyme networks reconstructed from all rhymes that were given the coda *-an, *-aj, or *-ar in the reconstruction by Baxter and Sagart. A gives a force-directed view of the network while B shows the* Infomap *clusters. C and D show two different views of the* Infomap *clusters. In C, rhyme reconstructions follow the suggestions of Baxter and Sagart's reconstruction, but ignore the marked uncertainties. In D the uncertainties are noted by adding a specific color for cases marked as *-a[n] in the reconstruction (reflecting *-an or *-ar rhymes) and thus reveals that the cluster originally looked like a mixed cluster of rhymes ending in *-ar and *-an is now better interpreted as a pure *-ar cluster.*

5 Conclusion and Outlook

This study presented a new way to approach the problem of rhyme analysis for the reconstruction of Old Chinese phonology. It was shown how the classical rhyme analysis, based on rhyme judgments applied

to the *Book of Odes* can be used to construct a weighted network of rhyme words that was then further investigated through the standard approaches to network analysis. The experiments presented in the paper reveal interesting facts regarding the general topology of the Shījīng network. They illustrate that there is strong evidence for the six vowel hypothesis and support the proposal of an **-r* coda in Old Chinese that was first proposed by Starostin and now also employed by Baxter and Sagart.

Since the approach is still strictly experimental, no complete revision of current problems is now presented. Instead, the study attempted to demonstrate how strict network models of rhyme data can be used to test hypotheses in Old Chinese phonology and improve certain reconstruction proposals. Several improvements need to be made in the future. The data needs to be enhanced, ideally incorporating additional analyses of rhyme patterns in the *Book of Odes*, similar to the one presented by Wang. Potentially many other studies should be added, although it is difficult to digitize the rhyme judgments in cases where the data is not presented transparently. In instances when the versions of the Shījīng differ, these should be annotated in order to allow to run competing analyses. The meta-data also needs to be refined and completed, including the different available reconstructions of Old Chinese phonology. It would also be beneficial to incorporate alternative perspectives on the available data, especially the phonetic series. Ideally, all of the recent reconstruction proposals for Old Chinese should be digitally available as a series of rhyme judgments along with the proposed reconstruction for the rhyme words.

This analysis presents two interactive applications that are supposed to ease the work of experts who often are not satisfied by just seeing the grand picture, but also wish to zoom in to reconstruct better a unique history of each word. It is important to find ways to incorporate this incredibly valuable knowledge into the big data perspective, thus reconciling automatic and manual approaches to linguistic reconstruction. No matter whether one is accustomed to automatic approaches or not, it seems indispensable that historical linguists generally enhance the way they present their ideas to others, especially in those cases involving larger datasets. In an ideal world, all the different ideas regarding the reconstruction of Old Chinese would be presented in a form that is both human- and machine-readable, thus enabling computational scientists to run large-scale analyses, while at the same time saving the experts invaluable time by providing quick access to the opinions of their colleagues.

Network analyses are very common in many branches of science. It is surprising in this context that, despite the fact that classical rhyme analysis, the linking analysis of *fǎnqiè* readings, as well as the analysis of phonetic series are inherently network-like, no computational network approaches have been carried out so far. Hopefully, this study represents starting point, and many more analyses of other aspects of Chinese historical linguistics that are amenable for network modeling will follow. Here, it is less important to disprove the great work that has been done by classical Chinese linguists and current experts. Instead it would be desirable to ease the painstaking work that scholars such as Chén Lǐ 陳禮 Duàn Yùcái 段玉裁 (1735–1815), and Jiāng Yǒugào 江有誥 have started, and make their methods applicable for other epochs of Chinese language history. Enhanced, computer-assisted approaches to the analysis of rhyme patterns, *fǎnqiè* readings, and phonetic series promises not only an investigation of the oldest stages of Chinese phonology, but also a means to trace their development across times and places.

Acknowledgements

I am very thankful to Laurent Sagart, William Baxter, Guillaume Jacques, and Nathan Hill, as well as the anonymous reviewers, who all commented on earlier versions of this manuscript and the network

methods and helped me with their inspiring critics. I am also very thankful to Wolfgang Behr, who gave me many hints for additional references and turned my awareness to the large amount of literature on rhyming in general, which I could only marginally touch on in this paper. Last not least, I am very indebted to my biological collaborators from the team AIRE at UPMC Paris, especially Eric Bapteste and Philippe Lopez, who introduced me to the basics of graph theory and the fruitfulness of network approaches in tackling various problems in the sciences and the humanities.

Supplementary Material

The supplementary material accompanying this paper includes two interactive applications and a repository of source code along with additional data that can be used to replicate all the analyses described in this paper.

The interactive Shījng browser is accessible via the following link: http://digling.org/shijing/

The interactive display of the Infomap communities is accesible via this link: http://digling.org/shijing/infomap.html

The full repository is accesible via this link: https://zenodo.org/badge/latestdoi/43676744

Funding Information

This research was supported by the DFG grant 261553824 (http://gepris.dfg.de/gepris/projekt/261553824).

References

Ahn, Yeol, James Bagrow, and Sune Lehmann. 2010. Link communities reveal multiscale complexity in networks. *Nature* 466.7307:761–764.

Alvarez-Ponce, David, Philippe Lopez, Eric Bapteste, and James O. McInerney. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. In *Proceedings of the National Academy of Sciences* 110.17:1594–1603.

Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. Berlin: Walter de Gruyter.

Baxter, William H. and Laurent Sagart. 2014. *Old Chinese. A New Reconstruction*. Oxford: Oxford University Press.

Branner, David P. 2000. The Suí-Táng tradition of Fǎnqiè phonology. In *History of the language sciences*. eds. S. Auroux, E. Koerner, E. H. Niederehe, and K. Versteegh, 36–46. Berlin, New York: Walter de Gruyter.

Cáo, Qiáng 曹強. 2010.〈江有誥《詩經韻讀》與王力《詩經韻讀》再比較〉[Recomparison of the rhyme analysis of the Book of Songs of Jiang Yougao and Wang Li].《延安大學學報(社會科學版)》[Journal of Yanan University (Social Science)] 32.3:106–109.

Cheng, Siu Kei 鄭绍基. 2004.〈從數理統計方法看《粵謳》押韻系統與近代粵語音系〉[The rhyming system of Yue Ou and the sound system of Early Modern Cantonese. From the perspective of mathematical statistics]. In *Proceedings of the 4th Postgraduate Research Forum on Linguistics,* 184–192.

Cheng, Siu Kei 鄭绍基. 2009. *An Optimality Theoretical Account of Contemporary Cantonese Rhyming Based on Inferential Statistics*. Hong Kong: The Hong Kong University of Science and Technology dissertation.

Csárdi, Gábor, and Tamás Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*. 1695:1–9. http://igraph.org.

Gěng, Zhènshēng 耿振生. 2004.《20世紀漢語音韻學方法論》[20th century's methods in traditional Chinese phonology]. 北京: 北京大學出版社.

Hagberg, A. 2009. NetworkX. High productivity software for complex networks. http://networkx.github.io/.

Harbsmeier, Christoph, Jiang, S. 2009. TLS – Thesaurus Linguae Sericae. A historical and comparative encyclopedia of Chinese conceptual schemes. http://tls.uni-hd.de/.

Hé, Jiǔyíng 何九盈. 2006[1985].《中國古代語言學史》[History of ancient Chinese linguistics]. 北京: 北京大學出版社.

Hill, Nathan W. 2014. Cognates of Old Chinese *-n, *-r, and *-j in Tibetan and Burmese. *Cahiers de Linguistique Asie Orientale* 43:91–109.

Karlgren, Bernhard. 1950. *The Book of Odes. Chinese Text, Transcription and Translation*. Stockholm: Museum of Far Eastern Antiquities.

Karlgren, Bernhard. 1957. Grammata serica recensa. *Bulletin of the Museum of Far Eastern Antiquities* 29:1–332.

Lǐ, Shūxián 李書嫻, Mài Yún 麥耘. 2008.〈證《詩經》押韻〉 [Proof that the Book of Odes rhymes].《中國語文》56.4:371–384.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis, and Robert Forkel. 2016. LingPy. A Python library for historical linguistics. Version 2.5. http://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

List, Johann-Mattis, Anselm Terhalle, and Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*, 347–353.

List, Johann-Mattis, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics*, Vol. 2: *Short Papers*, 599–605.

List, Johann-Mattis, Thomas Mayer, Anselm Terhalle, and Matthias Urban. 2014. CLICS: Database of Cross-Linguistic Colexifications. Marburg: Forschungszentrum Deutscher Sprachatlas. http://clics.lingpy.org.

Liú, Xiǎonán 劉曉南. 2006.《漢語音韻研究教程》[Reader in traditional Chinese phonology]. 北京: 北京大學出版社.

Lopez, Philippe, Johann-Mattis List, and Eric Bapteste. 2013. A preliminary case for exploratory networks in biology and linguistics: the phonetic network of Chinese words as a case-study. In *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, eds. Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, 181–196. Stuttgart: Franz Steiner Verlag.

Lǚ, Shèngnán 呂勝男. 2009.〈古韻研究方法論發微. 兼論今文《尚書》用韻研究〉 [A brief study of the methodology of the study of ancient rhyme. And Concurrently on the study of the rhyme of "Jinwen Shangshu"].《南陽師範學院學報(社會科學版)》[Journal of Nanyang Normal University (SocialSciences)] 8.2:57–61.

Mài, Yún 麥耘. 1999.〈隋代押韻材料的數理分析〉 [Quantitative analysis of Suí dynasty rhyme material].《語言研究》37.2:112–128.

Morrison, David. 2011. *An Introduction to Phylogenetic Networks*. Uppsala: RJR Productions.

Newman, M. 2004. Analysis of weighted networks. *Physical Review E* 70.5:056131.

Pān, Wùyún 潘悟雲. 2000.《漢語歷史音韻學》[Chinese historical phonology]. 上海: 上海教育出版社.

Rosvall, Martin, and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105.4:1118–1123.

Smoot, M., K. Ono, J. Ruscheinski, PL. Wang, and T. Ideker. 2011. Cytoscape 2.8. New features for data integration and network visualization. *Bioinformatics* 27.3:431–432.

Starostin, Sergej A., Старостин, Сергей А. 1989. *Реконструкция древенекитайсой фонологической системы* [Reconstruction of the phonological system of Old Chinese]. Москва: Наука.

Wáng, Lì 王力. 1980.《詩經韻讀》[Rhyme readings in the Book of Odes]. 上海: 上海古籍出版社.

Zhèng, Línxiào 鄭林嘯. 2004. 〈音韻學中統計法的比較〉[Comparing statistical approaches in traditional Chinese phonology]. *Studies in Language and Linguistics* 24.3:18–22.

Zhèngzhāng, Shàngfāng 鄭張尚芳. 2003.《上古音系》[Old Chinese phonology]. 上海: 上海教育出版社.

Zhū, Xiǎonóng 朱曉農. 1989.《北宋中原韻轍考.一項數理統計研究》[Investigation of the Zhōngyuán rhymes in the Northern Sòng dynasty. A statistical analysis]. 北京: 语文出版社.

Zwicky, Arnold. 1976. Well, this Rock and Roll Has Got to Stop. Junior's Head is Hard as a Rock. In *Papers from the Twelfth Regional Meeting of the Chicago Linguistic Society*, 676–697.

用網絡模型來分析古代漢語的韻母數據

游函
法國國家科學研究院
mattis.list@lingpy.org

提要

古代漢語的詞語所反映的韻為對上古音系的構擬，特別是對於最近的一些上古漢語構擬系統，異常重要。其中有一些構擬系統不再僅僅靠於學者的直覺，而且還用統計參數證實來評估分韻和派韻的概率。然而，迄今為止，定量方法僅用於確認關於上古韻部的假設，並且沒有進行探索性數據分析來創建初步分韻假設。本文提出了一種將韻母數據模型為加權無向網絡的新方法。此方法將韻母模型為網絡中的頂點，將某個語料庫的合韻率模型為聯頂點的邊緣，用社會網絡分析的標準算法來推斷語料庫所反映的韻母。為了更具體的說明此方法，本文用“詩經”來構建韻母網絡，而且比較自動與學者所推斷的上古韻部。除了揭示古代漢語韻網的一些有趣特點，“詩經”韻網分析了支持上古漢語韻尾* -r的新證據。“詩經”韻網和韻網分析的結果可以用交際在線應用來訪問而下載。

關鍵詞

韻母網絡、詩經、上古音系、上古漢語構擬法