DHQ: Digital Humanities Quarterly
2019
Volume 13 Number 4

Reading the *Quan Tang shi*: Literary History, Topic Modeling, Divergence Measures

Peter Broadwell <broadwell_at_stanford_dot_edu>, Center for Interdisciplinary Research, Stanford University
Jack W. Chen <jwc8v_at_virginia_dot_edu>, University of Virginia
David Shepard <dave_at_humnet_dot_ucla_dot_edu>, Scholarly Innovation Lab, UCLA

Abstract

The present paper addresses the problem of literary history as a problem of data comprehensiveness and selection, seeking not to resolve the impossibility of literary historical narrative, but to reframe it through a computational perspective. Our focus is on the *Quan Tang shi* 全唐詩 (Complete Tang poetry), the massive comprehensive anthology of Tang poetry that was produced at the height of the Qing dynasty (1644–1912). The sheer quantity of Tang poetry preserved in the *Quan Tang shi* (over 50,000 poems and poem fragments) exceeds the human-scale perspectives of close reading. To make sense of the corpus as a whole, we will show how two related forms of distant reading — topic modeling and divergence measures — allow us to reframe and rethink these literary historical questions and provide a new perspective on what it means to read Tang poetry.

The writing of literary history is perhaps an impossible endeavor. Even beyond the broader problem of justifying what counts as literature, we are faced with the question of how a historical narrative framework can effectively and comprehensively represent a given literary tradition. In many ways, this is a data modeling problem, one that turns on the question of reducing the vast arrays of literary data into a comprehensible narrative. [1] That is, even if one lowers one's sights to consider only a particular genre within a specific literary tradition — say, poetry written during the Tang dynasty (618–907) — there is still too much writing to account for and to fit into a coherent narrative. The problem is further complicated by the lack of a clear standard for what should be selected and what omitted, and how the selected portion could justifiably represent the whole of the data. As a result, literary histories tend to recycle the same small sample of super-canonical works and authors, unless the literary history is specifically aimed at broadening the canon or focused on a previously marginalized tradition. 1

The present paper addresses the problem of literary history precisely as a problem of data comprehensiveness and selection, seeking not to solve the impossibility of literary historical narrative, but to reframe it through a computational perspective. We take as our critical object the *Quan Tang shi* 全唐詩 (Complete Tang poetry; hereafter *QTS*), the massive comprehensive anthology of Tang poetry that was commissioned by imperial decree in 1705, at the height of the Qing dynasty (1644–1912). Because the poetry of the Tang had become the standard for poetic composition over the intervening dynasties, the compilation of the *QTS* thus may be said to represent the culmination of some 800 years of literary canonization. This work is *the* major source for extant Tang poetry, and moreover, constitutes a literary historical perspective on the development of Tang poetry in its complex and multidimensional classificatory scheme (more on this below). However, the sheer quantity of Tang poetry that is preserved in the *QTS* — over 50,000 poems and poem fragments — simply exceeds the human-scale perspectives of close reading. Yet without reading the entirety of the *QTS*, there can be no comprehensive understanding of what Tang poetry is, let alone a literary historical account of Tang poetry that can make use of the entire range of evidence and itself remain intelligible. To this end, what is required is access to a different scale of reading, one that has been characterized as "distant reading" [Moretti 2000a] [Underwood 2017], but might be more accurately thought of as machinic reading or as computational criticism. 2

Two points should be noted. First, on the question of distant reading versus close reading, we are mindful of Andrew Piper's caution against the oft-brandished either/or dichotomy between the two scalar approaches. He writes, 3

> I want us to see how impossible it is not to move between these poles when trying to construct literary arguments that operate at a certain level of scale (although when this shift occurs remains unclear). In particular, I want us to see the necessary integration of qualitative and quantitative reasoning, which, as I will try to show, has a fundamentally circular and therefore hermeneutic nature. [Piper 2015, 69]

We fully agree with the necessity of a both/and approach that allows for recursively circular processes of insights between scales of reading, though the question remains of what it means to mediate between machinic and human understandings of text and the equally important question of how a machinic vision of reading might transform how we as humans understand hermeneutics. Second, there has been relatively little work on distant reading methodologies with regard to East Asian corpora, not only because of technical issues such as character encoding and tokenization, but also because of the higher degree of linguistic difficulty in mastering East Asian languages and textual traditions. This said, in recent years there has been work on topic modeling and classical Chinese text corpora [Nichols et al. 2018], focusing on particular issues in early thought. Mention should also be made of an endeavor to apply topic modeling to a very large classical Chinese corpus, though its results are more general and preliminary, and there are some basic issues with document selection and preparation [Allen et al. 2019]. There have also been some studies of East Asian literary history using computational methods that take up questions of genre and style as related to problems of scalar reading [Vierthaler 2016][Long So 2016], and along similar lines more specifically for Tang poetry [Liu et al. 2018]. 4

This essay begins with a discussion of literary history and how it has framed problems of data selection and completeness, before turning to an overview of the *QTS*. To make sense of the corpus as a whole, we will then show how two related forms of distant reading — topic modeling and divergence measures — allow us both to reframe and rethink these literary historical questions and to access a new perspective on the practice of literary history. 5

I. What Do We Do When We Write Literary History?

Literary history, as a subfield within literary scholarship, has occupied an ambivalent place, one that has seen a decline in terms of institutional or disciplinary training, even while new literary histories are still being written. The purpose of literary history can be defined in various ways, including the emergence and development of a genre or genres, the rise of a national literature, or the situation of literary writing within particular historical and cultural contexts. Yet, however literary history is defined, it is complicated by a number of factors, among which are the restrictions that both the individual critic and the narrative form impose on what authors and texts can be included, and what it means to include certain authors or texts and not others, not to mention the broader problems of working with historical context. David Perkins, in his now–seminal study, *Is Literary History Possible?*, writes: 6

> Whatever else they have also hoped to accomplish, all literary historians have sought to represent the past and to explain it. To represent it is to tell how it was and to explain it is to state why — why literary works acquired the character they have and why the literary series evolved as it did. … Of course representation and explanation can never be complete, as literary historians and theorists have always recognized. Even if a historian knew all the relevant facts and answers, he could not crowd them into a book. The only complete literary history would be the past itself, but this would not be a history, because it would not be interpretive and explanatory.

Perkins goes on to ask "how much incompleteness is acceptable" and to note how the fact that "a literary history must be written from a point of view" — the limited perspective of the critic—inflects the explanatory argument of the narrative [Perkins 1992, 13]. These are problems that underlie all narrative historiography, since the nature of narrative historical argument involves simplification by means of representationality. That is, certain data are selected to serve as representative of the whole, and such choices are always informed by value judgments and critical interventions.

To speak of data selection entails a need for evidentiary standards by which the selection is made, but how standards are determined is not always clear in literary historical writing. Data are not agnostic, because what gets chosen as data always already implies value judgments [Drucker 2011] [Rosenberg 2013]. Decades earlier, René Wellek and Austin Warren had already noted this exact point: 7

> There are simply no data in literary history which are completely neutral “facts”. Value judgements are implied in the very choice of materials: in the simple preliminary distinction between books and literature, in the mere allocation of space to this or that author. Even if we grant that there are facts comparatively neutral, facts such as dates, titles, biographical events, we merely grant the possibility of compiling the annals of literature. But any question a little more advanced, even a question of textual criticism or of sources and influences, requires constant acts of judgement. [Wellek and Warren 1956, 40]

Neither Wellek and Warren nor Perkins are arguing against historiographic simplification, since literary histories, much like maps, provide us with signposts by which to navigate the practically boundless territories of literary data. Indeed, one might go even further: We rely on literary history to tell us what to read, to tell us what counts as literature — in short, to demarcate the very boundaries of the vast and otherwise amorphous realm of textuality. The map is not and should not be the territory if the map is to be of any use. That said, there is still a problem of *how* one makes the initial selections. To allocate space in the historical narrative to a particular author means that other authors must be neglected, or simply reduced to passing mention; at some point the linearity of the narrative will require that data be excluded or discussed in aggregate; and this refracts the data in ways that can be obscured through the graceful telling of literary history. [2]

While one reduction of data takes place at the point of the historical narrative’s composition, another reduction takes place over the history of readership. Margaret Cohen speaks of the “Great Unread”, the unquantified mass of writings that have been lost through readerly neglect [Cohen 1999, 23]. This same idea is expressed rather more colorfully by Franco Moretti, who speaks of the “slaughter of literature” and “the butchers–readers: who read novel A (but not B, C, D, E, F, G, H…) and so keep A ‘alive’ into the following one, and so on until eventually A becomes canonized” [Moretti 2000b, 209]. Moretti’s model has its flaws, both because it assumes “the shape of the past from that of the present”, effectively rediscovering the canon it claims to question [Bode 2017, 90], and because it treats each hypothetical novel as a stable and discrete entity. The fact that Novels B, C, or D are not chosen does not necessarily mean that their traits go untransmitted to later generations. Novel A, whose traits are passed on, may also have traits similar to Novels B, C, and D, etc., and indeed, most novels within a particular genre that are produced around the same time are likely to be more similar to one another than not. If Moretti’s model does not quite hold, we may at least say that readership plays some kind of role in the selection process of literary history, a process that is largely unsupervised and distributed, in the sense that there is no single intelligence that dictates which literary works will be canonized and which will not. 8

Yet, although Moretti views the composition of literary history as largely an afterthought, a rubber stamp of approval as he says, with the real work done by the decentered agency of readership, the writing of literary history is never simply reflective of reading histories, just as literary historiography is never the sole constitutive authority of what is read. The literary historian stands at the end of a lengthy process of selection, charged with representing the history of the successfully transmitted genetic material of literary data, but she also provides a critical intervention into the history of readership, one that potentially renegotiates the terms of readerly histories. That is to say, literary historiography both articulates unarticulated processes of canonization and consciously makes further selections of material both from within the readerly canon and from without. 9

However the immense accumulations of literary data are pruned and managed before they reach us, we might consider such reductions of data to be blessings, since to paraphrase Ann Blair, there is simply too much to read [Blair 2011]. Nevertheless, with the advent of massive text digitization projects and the development of quantitative, rather than qualitative, reading methodologies, we might begin to explore what a literary history of a whole corpus (defined in terms of genre, period, author, or even on greater scales, such as tradition) would look like. We are mindful that qualitative perspectives will not be supplanted by quantitative methods, and that quantitative perspectives will not simply serve to 10

confirm (or disprove) qualitative judgments. At the same time, the impact of computational methodologies has not yet been fully assessed; how one reads across multiple scales is not yet integrated into literary and cultural studies. This essay is a first step toward articulating a set of epistemological methodologies for a computational understanding of literary history.

II. What Is the *QTS*?

11 To explore this problem of literary history, we will take as our object of analysis the *Quan Tang shi*, a comprehensive anthology of (almost) all Tang dynasty (618–907) poetry [QTS 1960] that was compiled between 1705 and 1707 under the nominal editorship of the prominent official and imperial confidant Cao Yin 曹寅 (1658–1712) and a team of scholars led by Peng Dingqiu 彭定求 (1645–1719) [Spence 1966, 156–65]. [3] Despite oft-voiced complaints about textual errors, misattributions, and duplicate or omitted poems, this is an extraordinary work of scholarship that has preserved almost all poetry composed between the seventh and tenth centuries that might otherwise have been lost. And because the *QTS* corpus is a historically defined and comprehensive collection of extant Tang verse, the flaws inherent to Matthew L. Jockers's macroanalytic approach to English literature, namely that his corpus is, despite his grander claims, "incomplete, interrupted, haphazard," are not relevant here [Jockers 2013, 172] [Bode 2017, 83].

12 The structure of the *QTS* is itself remarkable, representing an ingenious attempt during its time to manage a vast array of literary data and allow for efficient retrieval of documents within a complex system. It consists of nine hundred chapters (*juan* 卷), with 49,403 poems (*shou* 首), as well as 1,055 poem fragments and couplets (*ju* 句) [Tong 1998, 17]. This work was based on two earlier anthologies of Tang poetry: first, the *Tangyin tongqian* 唐音統籤 (Assembled documents of Tang tones), which was compiled by the Ming scholar Hu Zhenheng 胡震亨 (1569–ca. 1644) [Hu 2003]; and an earlier project, also entitled *Quan Tang shi*, that was begun by the Ming-Qing transition poet Qian Qianyi 錢謙益 (1582–1664) and continued by Ji Zhenyi 季振宜 (b. 1630), now referred to as the *Quan Tang shi gaoben* 全唐詩稿本 (Draft version of the complete Tang poems) [Qian and Ji 1979]. The *QTS* is organized for the most part in a chronological fashion with certain significant exceptions. The contents of the work are identified as follows:

Chapters	Contents
1–9	Poems by emperors, empresses and imperial members
10–16	Poems relating to rituals and sacrifices
17–29	*Music Bureau* poetry
30–731	Poems by individual poets of the Tang dynasty
732–733	Poems by dynastic villains and rebels
734–766	Poems by individual poets of the Five Dynasties Period
767–784	Poems by poets with partial biographical information
785–787	Poems without authorial attribution
788–794	Linked verse poems (*lianju* 聯句)
795	Incomplete poems and lines by poets not listed above
796	Incomplete poems and lines without authorial attribution
797–805	Poems by women
806–851	Poems by Buddhist figures
852–859	Poems by Daoist figures
860–862	Poems by (male) immortals (*xian* 仙)
863	Poems by female immortals (*nüxian* 女仙)
864	Poems by spirits (*shen* 神)
865–866	Poems by ghosts (*gui* 鬼)
867	Poems by anomalies (*guai* 怪)
868	Poems sent in dreams (*meng* 夢)
869–872	Jest and insult poems (*xienüe* 諧謔)
873	Poems inscribed on walls (*tiyu* 提語) and judgments (*pan* 判)
874	Songs (*ge* 歌) sung by groups or local communities
875	Prophetic verse (*chenji* 讖記)
876	Sayings in verse form (*yu* 語)
877	Proverbs and enigmatic verse (*yanmi* 諺謎)
878	Prescient popular ditties (*yao* 謠)
879	Drinking songs (*jiuling* 酒令)
880	Divination songs (*zhanci* 占辭)
881	The *Mengqiu* 蒙求 primer by Li Han 李瀚
882–888	Poems left out of previous sections (*buyi* 補遺)
889–900	Song-lyrics (*ci* 詞)

Table 1.

The first nine chapters are devoted to poems by members of the Tang imperial household, while chapters 10 to 29 cover ritual-related poetic texts and *yuefu* 樂府 poetry (poetry that relates to the Han dynasty Music Bureau or were composed under titles related to this office) [Allen 1992] [Owen 2006b, 301–307]. However, the bulk of the anthology's contents (chs. 30–731) follow an approximate chronological order of the lives of the individual poets included within. The anthology breaks from historical chronology to include poems by dynastic villains such as the eunuch Gao Lishi 高力士 (684–762) and the rebel Huang Chao 黃巢 (d. 884) (chs. 732–33) before the historical chronology resumes with poets of the Five Dynasties (chs. 734–66). However, following this, one finds poets with increasingly less biographical information, beginning with poets for whom full names and rank or geographic association are known, but who have no verified dates, then taking up poets who lack names but have some kind of other association that can be used to refer to them (i.e., "Traveler at Mount Li" 麗山遊人), and ending with poems that no longer have attributed authors (chs. 767–

87). This is followed by linked verse (poems composed collaboratively, with each poet contributing alternating lines or couplets), incomplete poems (lines, couplets, stanzas) by poets without otherwise extant complete poems, and incomplete poems without any attribution (chs. 788–96).

14 Excepting the sections for poems written by imperial household members and dynastic villains, the anthology is, up to this point, organized according to a literary historical logic, one that seeks to order the poems by chronological precedence according to the authors' biographical information, and then places those poems that lack such information afterwards in descending order of how much information is available. Pauline Yu, writing on Tang poetic anthologies, has called attention to the essential relationship between anthologies and literary history. She points out how poems in traditional Chinese anthologies "were generally arranged in chronological order, if known, within an individual poet's corpus, and then from author to author as well." As Yu goes on to note, this practice was distinct from Japanese anthologies, "whose topical, associative, or thematically sequential organizational schemes suggest principles that are distinctly non-historical" [Yu 1990, 170] (also see [Yu 1994]). This latter form of anthology is based on the *leishu* 類書 (categorized writings), which were most commonly organized by a hierarchical system of topics, from the cosmic (heaven and earth) to the miniscule (insects) [Drège 2007]. *Leishu* are comparable to encyclopedias insofar as they were intended as comprehensive repositories of textual information, though there are significant differences from the Western concept of the encyclopedia [Tian 2017].

15 We see a different organizational logic beginning with chapter 797, where the *QTS* departs from a literary historical mode and follows something that is much closer to the topical categories associated with Japanese-style anthologies. This next section (chs. 797–805) is concerned with female poets, from "famous beauties" (*mingyuan* 名媛) and professional female performers (*jinü* 妓女) to well-known named authors such as Xue Tao 薛濤 (d. 832), Yu Xuanji 魚玄機 (ca. 844–68), and Li Ye 李冶 (d. 784). Following this are sections on Buddhist and Daoist poets (chs. 806–59), and then sections on immortals (*xian* 仙), spirits (*shen* 神), ghosts (*gui* 鬼), and anomalies (*guai* 怪) (chs. 860–66). The last chapters of the anthology are truly miscellaneous, including examples of divination verse (*zhanci* 占辭), proverbs and enigmatic verse (*yanmi* 諺謎), prescient popular ditties (*yao* 謠), poems from dreams (*meng* 夢), and poems inscribed on walls (*tiyu* 提語). Wherever possible, these chapters begin with figures related to the imperial house (emperors, palace women) and then proceed in chronological order.

16 The organizational logic of the *QTS* is worth a brief comment as it combines two earlier organizational schemes for literary anthologies: the categorical scheme associated with the encyclopedic *leishu* and the comprehensive chronological scheme that became prominent with Ming dynasty literary anthologies. Both forms of organization express a critical intelligence that orders documents in a way that makes theoretically possible the search and retrieval of particular documents. Yet because the *QTS* is constructed on a larger scale than what an individual reader can productively manage, the anthology functions in many ways more like an archive or a database than a work intended for personal perusal.

17 One might contrast the *QTS* to other smaller Tang poetic anthologies that were clearly meant for individual readers, such as the slightly later Qing anthology, *Tang shi sanbaishou* 唐詩三百首 (The three hundred poems of the Tang). Compiled by Sun Zhu 孫誅 (1722–78), this extremely popular selection of approximately three hundred poems has served as an elementary poetry primer for generations through to the present day [Qian 2000, 364]. To frame this within broader issues of literary history, while the *QTS* has served as the imperially sanctioned and culturally canonical edition of Tang poetry since its compilation, it was the *Tang shi sanbaishou* (and similar anthologies) that provided convenient and familiar access and thus actually shaped the idea of Tang poetry for most readers. In this way, the *Tang shi sanbaishou* served as a primary training dataset for literary historical knowledge, curating the popular experience of Tang poetry throughout society, and thus constructing standards of taste and readerly expectation for what a Tang poem should be.

18 While literary histories of the Tang are not usually based on elementary anthologies like the *Tang shi sanbaishou*, they are nevertheless shaped by other cultural datasets, whether these are more scholarly anthologies, critical studies, or university curricula. If this is the case, then what might a history of Tang poetry look like if the dataset were comprehensive, or at least more robust? That is to say, how might *reading* the *QTS* alter our view of Tang poetic

history? To be sure, both the act of reading the *QTS* and any critical representation of this knowledge would certainly depart from the usual narrative forms associated with conventional literary history. Indeed, while the *QTS* has often been mined as a resource for histories of Tang poetry (in English, see Owen 1977, Owen 1981, Owen 2006a; in Chinese, see Xu 1994, Yang 1996), it is rarely considered in its own right. After all, how many poets and poems, representing the literary accumulation of nearly three hundred years, can be fruitfully discussed within a single narrative framework? And even if such a narrative framework were possible, what examples of poetry should be selected and what omitted? How should the oeuvre of an individual poet be represented? How does the unit of analysis, whether it be the individual poet, a poetic style, or a literary theme, shape and inflect the literary historical account? If we are to understand Tang poetry not in terms of the limited dataset of authors and texts, but in terms of the whole of transmitted literary data, then we will need to take a macroscopic perspective, one that initially forgoes the granularity of close reading for something that takes place at a more stratospheric level.

III. On Topic Modeling as Literary Methodology

19 In an age in which massive text digitization projects have been carried out by research centers, universities, government agencies, and individuals, literary historical claims need not rely solely upon microscale reading. The problem of close reading is that it is only possible at the scale of the individual text, which then necessitates an exemplary approach to the selection and analysis of the critic's dataset [Hickman and McIntyre 2012][North 2013]. To avail oneself of distant reading methods, however, is to adopt a somewhat different understanding of what it is to read a text, one that does not rely upon the anecdotal, limited nature of individual data-gathering. Here it is the machine that transforms the interpretative process, providing the scalar shift and quantitative capacity that makes reading large corpora possible. Computational analysis is not necessary for the individual poem, but when confronted by thousands of poems (or more), then the computer makes it possible to extract meaningful information on a macroscopic level from the aggregated texts. N. Katherine Hayles has written on this point, noting that we are entering an age of "machine reading", which is to say, "computer algorithms used to analyze patterns in large textual corpora where size makes human reading of the entirety impossible" [Hayles 2010, 72]. And Stephen Ramsay writes:

> It is one thing to notice patterns of vocabulary, variations in line length, or images of darkness and light; it is another thing to employ a machine that can unerringly discover every instance of such features across a massive corpus of literary texts and then present those features in a visual format entirely foreign to the original organization in which these features appear. Or rather, it is the same thing at a different scale and with expanded powers of observation. It is in such results that the critic seeks not facts, but patterns. And from pattern the critic may move to the grander rhetorical formations that constitute critical reading. [Ramsay 2011, 17]

It is important to note that Ramsay also argues (elsewhere in his work) that computer-aided reading, which he calls "algorithmic criticism", is not a radical break from traditional modes of literary reading, since the human critic's acts of interpretation can also be said to depend on rule-based processes — which is to say, upon algorithms — even though the critic's manipulations and deformations of text may not be as explicit or programmatic as those of the computer. And to return to Hayles' point, the major difference that can be realized through the computer's algorithmic reading is simply that of the object of study, which is no longer confined to the single text or a succession of single texts, but a macroscalar aggregation of texts.

20 More recent work such as Andrew Piper's *Enumerations* follows essentially the same guiding principles: Piper uses statistical methods to approximate the judgements of a human reader at a macroscopic scale. Piper arrives at some impressive conclusions, including that characters within novels are more different from each other than they are from characters in other novels, and that heroes in twentieth-century science fiction are just as introspective as the heroines of nineteenth-century women's fiction. While critics' chosen statistical methods have become more sophisticated, the basic approach of using statistics to approximate the human reader's critical judgments at scale persists.

21 Topic modeling, as a form of text mining, is one method of reading a large corpus, and while it is becoming increasingly familiar to practitioners of literary analysis, we will provide a quick summary here, one that is intended to set the terms

for the following section. There are various introductions and overviews for those seeking more detailed summaries [Meeks 2012]; [Underwood 2012]; [Riddell 2012]). In brief, topic modeling deploys a set of probabilistic algorithms to identify latent semantic patterns within a set of documents. This is how David Blei, co-author of the most widely used probabilistic model — *latent Dirichlet allocation* or LDA (see Blei, Ng, and Jordan 2003) — describes it:

> …the goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure — the topics, per-document topic distributions, and the per-document per-word topic assignments — is *hidden structure*. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as "reversing" the generative process — what is the hidden structure that likely generated the observed collection?

[Blei 2012, 79]

22 There are a number of terms and concepts that need to be explained. First of all, there is the issue of *latency*, which Blei refers to as the "hidden structure" of the corpus. Topic modeling posits that there is a latent semantic structure beneath the observable corpus, and that there is a difference between what we can read in the everyday sense of the word and what only the computer can discern. This hidden structure is what topic modeling seeks to recover as the corpus's *generative model*, or the matrix of thematically significant word clusters that make up the corpus as it presently exists. These thematic word clusters are the *topics* that the algorithm seeks to model, composed of words that have a high rate of co-occurrence within documents (understood here as any significant unit of text). The topics are probabilistically significant distributions of words that comprise the corpus's documents, and each of the documents is understood as a mixture of topics. Document word order is not important, as the topic model treats the corpus simply as a "bag of words," drawing out each word and assigning it to a topic based on the frequency of its co-occurrence with other frequently co-occurring words.

23 How can this model claim to reflect the process of composing literature, let alone literary history? The answer is that it does not, but it approximates it. As readers, we may assume the poetic process for a poet — or a group of poets — as something like: first, a poet sets out to express an idea, which is composed of one or more themes, like "love" and "beauty". Next, the poet chooses concrete images to express these themes and ideas, such as "love" and "roses". From this, the poet puts the ideas into words, e.g., "My love is like a red, red rose". A human reader, reading the poem, works in reverse: she encounters the poet's actual words and transforms them into a set of images and thoughts, ultimately seeking to get a sense of "what the poet is actually writing about" — in other words, the themes he is trying to express. The specific layers described above perhaps do not matter so much: there may be more to the processes of composition or reading. Our point is that what we experience as the richness of human reading are the different levels of experience that a poem inspires in a reader: visualized images, emotional experiences, resonances with other words, historical context, and the like. The processes of producing or consuming literature is a process of either expressing or discovering these layers.

24 LDA follows a similar process: it starts from the poet's word choice and assumes that some latent variables (images or themes) drive the poet's word choices (the observed variables). It models the latent variables from the observed variables. While LDA is an approximate model, with far less sophistication than a human reader, it scales better. LDA approximates what a human being does with less sophistication: it has only two layers of meaning (observed and latent variables) versus the reader's potentially many interpretive layers. However, it can work at a much larger scale, and with much greater accuracy because it is not subject to the failures of human memory. A human critic using LDA works off of this approximation: she assumes that the latent variables match some of the things that a human reader will discover using their own interpretive apparatus, and interprets LDA's topics as patterns of imagery, themes, or ideas. She must negotiate between the rich meanings that a reader can extract from the few texts that a human reader can read, and the far less rich latent patterns that the LDA model can discover.

25 From this negotiation, we are able to derive a different model for doing literary history, one that scales to a larger set of texts. If one conceives of literary history as an epistemological process of discovery, one that seeks to understand the

parameters for how the literary past is constructed, then this perspective has the potential to reconceptualize our received evidentiary standards. What follows may not resemble literary historical work, and indeed, may read more like laboratory reports than critical analysis, but this is necessary: we are proposing that a macroanalytic literary history will require different methodologies, ones that negotiate between human and machine visions of textuality. To highlight the applications of such a computationally enhanced methodology to poetic studies and perhaps more broadly, we illustrate how judicious examination of a topic model of the *QTS* calls out poems whose authors and genres have not received the blessing of concerted attention from scholars, yet upon close inspection present exemplary literary and historical insights. Moreover, the identification of these poems and their revelatory topical similarities (and divergences) follows from what is arguably the key analytical contribution of LDA topic modeling: its ability to infer semantically alloyed relationships between sets of words that may appear together in the same document infrequently or even not at all, and to do this at a corpus-wide scale that is inaccessible to the minds of even the most well-read scholars. The discussion also emphasizes the importance of keeping the human "in the loop" throughout such computer-aided analyses, as computational measures of significance do not always correspond to humanistic notions of meaning.

IV. The *QTS* and Topic Modeling

On the most basic level: the topic modeling program that we used is *MALLET (MAchine Learning for LanguagE Toolkit)* [McCallum 2002]. For the *QTS* corpus, we excluded all poetic fragments (generally orphan couplets and lines) from the topic model, as the shortness of these documents statistically skews the output, though we retained sections of significant length, such as Chapter 881, which is comprised solely of the poetic primer *Mengqiu*, as it was feasible to divide these into poem-length subsections. We did not use a stop-word list, as many classical Chinese grammatical particles have other, non-particle meanings depending on the context. We specified that *MALLET* should generate 150 topics, which was the number of topics (following an iterative trial-and-error process) that provided the most informational complexity while minimizing interpretive perplexity.[4] From this, we received three reports: a topic-keys report (TKR), a topic-phrase report (TPR), and a document-topics report (DTR) — all of which will be discussed in turn. While these reports may be familiar to those who have experience with topic modeling, discussing these reports in detail is important for our argument, as we will demonstrate. 26

The first output file, the topic-keys report, is a non-ranked list of topics that shows the twenty strongest correlated words in each topic in ranked order (note that these are only the top twenty correlated words for a given topic, as this is the default number in *MALLET*). Here are the first eleven topics from the TKR: 27

1. 馬 車 騎 行 塵 出 長 鞭 門 走 白 嘶 金 蹄 道 鞍 青 黃 馳 驅
2. 酒 醉 杯 飲 一 客 酌 醒 勸 滿 尊 傾 對 笑 壺 歡 酣 送 倒 且
3. 不 人 能 知 誰 生 自 有 無 來 豈 得 在 與 作 但 令 可 古 解
4. 相 年 家 見 弟 同 兄 來 逢 長 少 還 自 許 多 喜 說 親 作 時
5. 江 海 迢 潮 帆 吳 越 楚 孤 滄 客 去 洲 波 歸 遠 湖 浪 遞 南
6. 寂 遙 寥 空 落 寞 朝 風 獨 見 中 思 林 月 清 在 寒 逍 想 招
7. 夜 月 明 燈 曉 殘 漏 星 聲 更 照 暗 寒 宿 燭 火 半 光 露 鐘
8. 我 不 為 此 言 爾 者 何 有 亦 如 生 人 與 得 苦 無 所 願 令
9. 水 東 流 西 山 日 去 雲 空 落 歸 中 路 陵 白 復 見 長 暮 盡
10. 金 玉 珠 銀 紫 錦 重 寶 黃 光 刀 環 裁 佩 衣 雙 盤 珊 龍 垂
11. 蕭 風 雨 秋 暮 條 寒 雲 吹 起 獨 颯 晚 上 日 樹 葉 空 向 索

The first column is the topic ID number, assigned by *MALLET* without rank significance. The second is the set of the words assigned to the topics in ranked order of correlation strength.

As the topics simply consist of ranked lists of words, to understand what the topics signify requires that labels be added to the topic-keys report. These labels are necessarily acts of interpretation, though useful ones for translating the machine perspective on the corpus into human terms. Here again are the first eleven topics, now with our provisional labels: 28

1. 馬 車 騎 行 塵 出 長 鞭 門 走 白 嘶 金 蹄 道 鞍 青 黃 馳 驅 horses, traveling
2. 酒 醉 杯 飲 一 客 酌 醒 勸 滿 尊 傾 對 笑 壺 歡 酣 送 倒 且 drinking ale
3. 不 人 能 知 誰 生 自 有 無 來 豈 得 在 與 作 但 令 可 古 解 particles, common verbs
4. 相 年 家 見 弟 同 兄 來 逢 長 少 還 自 許 多 喜 說 親 作 時 families, home
5. 江 海 迢 潮 帆 吳 越 楚 孤 滄 客 去 洲 波 歸 遠 湖 浪 遞 南 southern waterscapes
6. 寂 遙 寥 空 落 寞 朝 風 獨 見 中 思 林 月 清 在 寒 逍 想 招 remoteness and longing
7. 夜 月 明 燈 曉 殘 漏 星 聲 更 照 暗 寒 宿 燭 火 半 光 露 鐘 night: moon and lamp
8. 我 不 為 此 言 爾 者 何 有 亦 如 生 人 與 得 苦 無 所 願 令 particles, pronouns
9. 水 東 流 西 山 日 去 雲 空 落 歸 中 路 陵 白 復 見 長 暮 盡 landscape: traveler
10. 金 玉 珠 銀 紫 錦 重 寶 黃 光 刀 環 裁 佩 衣 雙 盤 珊 龍 垂 precious materials
11. 蕭 風 雨 秋 暮 條 寒 雲 吹 起 獨 颯 晚 上 日 樹 葉 空 向 索 autumn evening scene

Taking a closer look at Topic 1, to which we have attached the label *traveling with horses*, we may further gloss the listed terms as follows: 馬 (horse), 車 (carriage or wagon), 騎 (rider or to ride), 行 (to travel), 塵 (dust), 出 (to depart, set forth), 長 (long), 鞭 (whip), 門 (gate), 走 (to run), 白 (white), 嘶 (to neigh), 金 (gold, metal), 蹄 (hoof), 道 (road), 鞍 (saddle), 青 (green-blue), 黃 (yellow, brown), 馳 (to gallop, to rush), and 驅 (to drive forward, to urge forth). All of these words are either thematically related to horses and traveling or illustrate what the linguist Shuanfan Huang calls syntactic contiguity, such as the three color terms [Huang 2013, 55–80]. 29

More difficult to label is a result such as Topic 4, which consists of many commonly used words, including the particle 相 (mutually, or an adverbial particle replacing a direct object), 年 (year), 家 (family, home), 見 (to see, to meet), 弟 (younger brother), etc., which might relate to a number of poetic themes involving home and family or perhaps something broader. There are a number of such topics, where the meanings of the terms are clear but how they relate semantically remains more or less ambiguous, inviting the human interpreter to fill in the gaps and to imagine possible poems that might be constructed with these terms. Here, the problem becomes one of projecting meaning where meaning is indeterminate, or perhaps, not determined in ways that accord with our (human) cultural categories. 30

Turning back to the *MALLET* output files, the same results are available in a more detailed list with weights for each word and for frequently occurring phrases. This is provided in the second output file, the topic-phrase report. This report first begins with a listing of frequently occurring phrases (combinations of tokens) in the topic and then provides a ranked list of the first nineteen words in the topic, followed by a ranked list of the 21 most common phrases. As the TPR is an even lengthier report than the TKR, here is the section just for Topic 0, reformatted as two separate tables. The first table lists each word in ranked order with word weights and token counts: 31

Word	Word Weight	Count
馬	0.16849291492169124	2937
車	0.049968447019677585	871
騎	0.03568355229189375	622
行	0.031897194653204064	556
塵	0.03172508748780908	553
出	0.030577706385175835	533
長	0.027594515518329414	481
鞭	0.02702082496701279	471
門	0.025299753313062934	441
走	0.020595490792266653	359
白	0.018358097642131834	320
嘶	0.016694395043313638	291
金	0.016120704491997016	281
蹄	0.014973323389363778	261
道	0.014629109058573805	255
鞍	0.01359646606620389	237
青	0.013424358900808904	234
黃	0.012047501577649016	210
馳	0.011990132522517355	209
驅	0.01187539441225403	207

Table 2.

The second table lists frequently occurring phrases, with phrase weights and counts:

Phrase	Phrase Weight	Count
車 馬	0.0615748963883955	104
馬 蹄	0.04262877442273535	72
走 馬	0.04085257548845471	69
馬 嘶	0.04026050917702783	68
驄 馬	0.03197158081705151	54
騎 馬	0.028419182948490232	48
鞍 馬	0.02427471876850207	41
駟 馬	0.020130254588513915	34
匹 馬	0.018354055654233273	31
駿 馬	0.017761989342806393	30
駑 駘	0.017169923031379514	29
躞 蹀	0.013025458851391355	22
匆 匆	0.012433392539964476	21
驅 馬	0.011841326228537596	20
騏 驥	0.010657193605683837	18
羸 馬	0.010065127294256957	17
車 騎	0.009473060982830076	16
馬 鞭	0.0076968620485494375	13
騏 驎	0.0076968620485494375	13
白 馬	0.0076968620485494375	13
嘶 馬	0.007104795737122558	12

Table 3.

As can be seen, the TPR provides a clearer breakdown of the strength of each word's correlation to the topic. Each word is assigned a value that denotes its topic weight, or the number of times that the word appears within documents identified with this particular topic. For example, the first–ranked character for Topic 0 is 馬 (horse), which is represented in the TPR as having a weighting of 0.1685 (rounded up) and a count of 2937, which means that it represents 16.85% of the 17,431 word occurrences assigned to topic 0 in the corpus (this count is also provided in the TPR). In other words, in the entire corpus, 馬 appears more than 2937 times, but for 2937 of the instances in which it occurs, it is assigned to topic 0. Most words will have a weighting value of <0.001 in the TPR for a specific topic, reflecting the default prior probability that any word in the corpus vocabulary will appear in a topic at some infinitesimal level. Words with this value in the topic–phrase report are considered to have negligible weight within the modeled topic but are nevertheless still represented. For some topics there may be a more even distribution of strongly correlated words, while for others there may be a marked differential between the first word and the next–strongest words. 33

The final *MALLET* output file is the document–topics report, which shows the strength of each topic within every document. This is an extremely long report, consisting of the percentage breakdowns of each of the 150 topics for every poem in the *QTS*. Rather than present the entire report, we have excerpted part of the report for Poem 30_1 in the corpus, which is to say, the first poem from Chapter 30, titled "On Han Gaozu" 詠漢高祖. This is not an arbitrary selection, as this is the point in the anthology where it begins to be organized by individual poets in roughly chronological order, taking on a literary historical logic. Moreover, this poem was composed by the late Sui-early Tang figure Wang Gui 王珪 (570–639), who served under Tang Emperor Taizong 唐太宗 (r. 626–49) and is not particularly remembered for his poetic ability. As such, this poem represents an example of the "Great Unread", a literary work that should have been forgotten except for its inclusion in the *QTS*. 34

Before turning to the report, it might be useful to be able to read the text of the poem, which we have translated in full as follows: 35

漢祖起豐沛， Han Gaozu arose from Feng Town in Pei County,[5]
乘運以躍鱗。 Riding fortune's course with high-leaping scales.[6]
手奮三尺劍， In his hand he brandished a three-foot sword,
西滅無道秦。 And to the west he destroyed the lawless Qin.
十月五星聚， In the tenth month, the five planets clustered,[7]
七年四海賓。 In seven years, all in the four seas were his guests.[8]
高抗威宇宙， Lofty and unyielding, he awed the universe,
貴有天下人。 Exalted, he took charge of all the world's people.
憶昔與項王， I think back on how he accompanied Xiang Yu,[9]
契闊時未伸。 Now long separated, that age cannot be reached.
鴻門既薄蝕， Both at the Hongmen Feast he was treated poorly,
滎陽亦蒙塵。 And at Xingyang, where he fled, covered in dust. [10]
蟣虱生介冑， Lice are born in armor and helmets,
將卒多苦辛。 Generals and soldiers endure much suffering.
爪牙驅信越， As his claws and teeth: Han Xin and Peng Yue forged ahead,
腹心謀張陳。 As his belly and heart: Zhang Liang and Chen Ping laid plots.[11]
赫赫西楚國， How glorious was Western Chu's kingdom,
化爲丘與榛。 Transformed now into mounds and underbrush.[12]

This is a fairly standard example of a "poem on history" (*yongshi shi* 詠史詩), one that focuses on the founding emperor of the Han dynasty. One might trace out the progression of the poem from its broad-stroke portrait of the ruler's rise to power to its focus on his war with his great rival Xiang Yu and the eventual ruin of Xiang Yu's Western Chu. The poem evokes a lost age of heroes, of glorious victories, fabled escapes, and tragic falls, bookending the triumphant founding of the Han dynasty with the barren fate of the Western Chu. 36

However, a different approach to analyzing this poem can be found in *MALLET*'s document-topics report. We have quoted only a portion of the lengthy report, stopping at the point where the percentage falls below 1%: 37

"1969 30_1, 30 0.10952380952380952 16 0.05952380952380953 139 0.05238095238095238 7 0.03809523809523809 135 0.03095238095238095 110 0.03095238095238095 32 0.03095238095238095 15 0.03095238095238095 49 0.02380952380952381 42 0.02380952380952381 125 0.016666666666666666 124 0.016666666666666666 122 0.016666666666666666 115 0.016666666666666666 109 0.016666666666666666 98 0.016666666666666666 92 0.016666666666666666 87 0.016666666666666666 83 0.016666666666666666 66 0.016666666666666666 36 0.016666666666666666 " 38

Following the two poem IDs ("1969", meaning the 1969th poem of the corpus; and "30_1", referring to the document's chapter number and its poem number within the chapter), *MALLET* provides two pieces of information: the topic ID number and the percentage of that topic within the poem. Included in the whole of the report are all 150 topics in descending order of strength (again, *MALLET* assumes that all topics are represented in each document, even if in statistically insignificant percentages). We can see that, for this document, it is Topic 30 that ranks most highly, constituting approximately 10.95% of the poem. Only the top twenty-one topics (topics with values >1%) are actually meaningful in this case, as most topics belong to the long tail of statistical insignificance. Here are the top topics (values are rounded up) in descending order of significance and with our preliminary labels for the topics: 39

TopicID	Percentage	Topic (first 20 terms in rank order)	Topic Label
30	10.952%	天 大 皇 四 帝 海 萬 太 功 夷 方 王 三 命 元 聖 乾 業 宗 武	empire and sovereign power
16	5.952%	不 如 食 生 足 骨 口 死 飢 肉 眼 齒 腹 為 土 耳 力 飲 可 破	mortality, eating and drinking
139	5.238%	塵 人 春 新 身 親 勤 鄰 殷 頻 津 為 巾 輪 濱 四 辰 辛 貧 巡	tsyen 真 *rhyme category*
7	3.81%	我 不 為 此 言 爾 者 何 有 亦 如 生 人 與 得 苦 無 所 願 令	pronouns and particles
135	3.095%	將 軍 兵 戰 旗 馬 旌 功 劍 弓 戎 營 騎 鼓 戈 羽 角 虜 射 箭	military, armies
110	3.095%	未 自 慚 知 豈 雖 薄 終 非 已 顧 愧 負 辭 寧 猶 難 甘 心 敢	shame, regret
32	3.095%	十 三 年 二 五 四 六 七 一 八 百 千 九 月 今 歲 前 載 老 第	numbers, calendrical time
15	3.095%	王 漢 秦 國 宮 吳 陵 帝 楚 武 臺 陽 長 子 安 亡 苑 作 蘇 梁	Han and pre-Han kingdoms
49	2.381%	南 北 東 西 山 風 望 國 城 日 向 斗 河 闕 起 海 長 直 復 從	territorial space, cardinal directions
42	2.381%	庭 中 滿 洞 月 山 風 樹 裏 長 高 入 空 雲 水 煙 起 日 上 陽	courtyard / landscape scene?
125	1.667%	門 深 開 日 閉 高 戶 滿 入 客 出 堂 朱 外 巷 院 掩 牆 庭 靜	gates, halls, and built spaces
124	1.667%	所 言 子 豈 道 志 異 為 良 非 懷 自 徒 昔 貞 乃 可 賢 常 義	particles / will and righteousness?
122	1.667%	魚 水 釣 下 垂 鳥 池 上 時 小 鱗 竿 驚 得 網 有 無 欲 避 坐	fisherman, fishing
115	1.667%	劍 士 氣 生 平 感 侯 壯 橫 長 縱 交 雄 將 子 英 安 報 燕 節	heroes, men of valor
109	1.667%	江 湘 南 楚 水 客 舟 遠 雲 孤 秋 山 瀟 雁 陽 浦 湖 去 渚 夢	southern (Chu) riverscapes
98	1.667%	神 禮 德 惟 肅 樂 靈 降 薦 既 誠 載 以 明 昭 永 斯 福 陳 備	ritual and ceremony
92	1.667%	龍 雷 騰 神 虎 如 驚 蛟 若 蛇 電 橫 當 鼓 倒 大 氣 怪 勢 天	dragons, tigers, images of divine power
87	1.667%	然 化 物 無 浩 天 心 異 造 有 形 方 道 可 忽 中 變 至 窮 言	divine creation and transformation
83	1.667%	古 草 荒 人 空 在 野 木 平 跡 遺 地 舊 原 猶 城 餘 有 無 蕪	desolate wilderness
66	1.667%	遊 留 休 頭 秋 侯 州 收 憂 求 愁 不 丘 浮 酬 裘 子 諸 牛 羞	ghou 尤 *rhyme category*
36	1.667%	天 下 太 地 子 中 一 上 海 出 道 入 平 黃 九 四 守 白 大 成	empire and territory

Table 4.

From this, we might say that “On Han Gaozu” is a document constructed out of assorted topics that might be labeled “empire and sovereign power”, “mortality, eating and drinking”, “tsyen **真** *rhyme category*”,[13] “pronouns and particles”, “military, armies”, “shame, regret”, “numbers, calendrical time”, “Han and pre-Han kingdoms”, “territorial space, cardinal directions”, and the ambiguous “courtyard / landscape scene”. Some of these are topics with a strong semantic identity (for example, both “empire” topics), though some are topics that represent syntactic and grammatical functions (“pronouns and particles”) and formal characteristics of the poem (“tsyen **真** *rhyme category*”). We also find the

presence of topics that are not easy to label, such as the unclear Topic 42 (“courtyard / landscape scene?”), and topics that are inexplicably present, such as Topic 66 (“ghou 尤 *rhyme category*”), which is a second rhyme category topic not otherwise evidenced in the poem. It is important to keep in mind that our interpretive perplexity is the result of differences between how humans and how *MALLET* reads the corpus, and that *MALLET*’s identification of topics for documents are not based on semantic reasoning but on the distribution of vocabularies. At times human and machine readings might converge, leading to deeper insight into the corpus, but at other times, we are reminded of our methodological — and ontological — differences.

V. Using Topics to Think through Literary History

41

Up until his point, the application of topic modeling to the *QTS* has largely been to reframe the question of representing and accessing the poetic corpus, and where we have provided an analytical perspective, it has focused on understanding what a topic model represents in relation to the poetic documents. Here, we propose that topic modeling may help us to think through the problem of selection criteria in constructing a literary historical account, enabling one to navigate the vast unread of Tang poetry, selecting poems that do not reflect the individual critic’s own biases and preferences, or those preferences predetermined by anthological training sets and other unknowable selection processes over the course of history. The resulting literary historical narrative would be semi-supervised, more strongly dependent upon human reading and interpretation at certain points in the process, and more strongly dependent upon machine reading at others. This would not be the utopian fantasy of posthuman reading, where machine objectivity replaces human subjectivity, but a shifting balance between a machine-assisted perspective and a traditional “analog” perspective. Part of this approach, however, is a serious attempt to engage the machine’s perspective, to try to understand how the machine sees the collection of documents, and to translate this into human terms. To this end, we rely on the Jensen-Shannon divergence measure [Lin 1991], which quantifies the difference between any (finite) number of probability distributions.[14] In our case, these distributions are the probabilities of topics constituting documents, as in *MALLET*’s document-topics report described above. Although this measure is formulated in terms of “divergence”, it allows us to explore document similarity, much like vector similarity measures, and to perform unsupervised information retrieval within a large textual corpus.[15]

42

Although the supervised selection of “On Han Gaozu” (Poem 30_1) as a starting point may have been somewhat arbitrary, chosen because of its position as the first poem in the main section of the *QTS* (the chronological-by-author section that runs from Chapter 30 to 731), its literary obscurity makes it a useful example of Tang poetry’s Great Unread. We could imagine, however, that we have arrived at this poem after navigating a topic-based index of the corpus and now seek to find topically similar poems. Using “On Han Gaozu” as our target document, the top five most similar documents according to the Jensen-Shannon divergences of their topic distributions are identified as follows:

Document ID	Poem Title and Author	English Translation	Similarity Score
1969 (30_1)	《詠漢高祖》王珪	“On Han Gaozu” by Wang Gui	1.0
573 (13_68)	《享太廟樂章。象德舞》 段文昌	“Hymns for the Offering to the Ancestral Temple: Dance for Manifesting Virtue” by Duan Wenchang	0.8822478125653975
38884 (767_29)	《苻堅投棰》 孫元晏	“Fu Jian Casts His Whip” by Sun Yuan’an	0.8787439416072281
2751 (52_30)	《過函谷關》 宋之問	“Visiting Hangu Pass” by Song Zhiwen	0.877820744700559
37314 (729_66)	《二廢帝》 周曇	“On the Two Deposed Emperors” by Zhou Tan	0.8722429641594189
2009 (31_26)	《梁郊祀樂章。慶休》 ?	“Hymns for the Liang Suburban Sacrifice: Jubilation” by unknown author	0.872006203906182

Table 5.

The table is mostly self-explanatory, but it is worth quickly pointing out that because “On Han Gaozu” is the benchmark for the similarity measure scores, it has a score of 1.0, meaning that it is identical with itself, and that the more similar a document is to “On Han Gaozu”, the higher its fractional similarity score. 43

One striking aspect of the divergence measure is how “On Han Gaozu”, traditionally understood as a poem on history, correlates strongly to poems originating from ritual and sacrificial occasions. The top document retrieved in this manner is Poem 13_68, “Hymns for the Offering to the Ancestral Temple: Dance for Manifesting Virtue” 享太廟樂章：象德舞, by the mid-/late Tang poet and official Duan Wenchang 段文昌 (773–835): 44

> 肅肅清廟 How magnificent and solemn, the ancestral temple,
> 登顯至德 Radiating forth its perfected virtue!
> 澤周八荒 Your grace covered the eight expanses,
> 兵定四極 Your armies settled the four directions.
> 生物咸遂 Living creatures all followed suit,
> 群盜滅息 All thieving ceased and came to an end.
> 明聖欽承 The Enlightened Sage reverently has inherited this,
> 子孫千億 Transmitting it to descendants for countless generations.
> [QTS 1960, 13.133]

This hymn belongs to the family of ritual songs categorized as “Lyrics for the Suburban Altars and Ancestral Temple” 郊廟歌辭, and according to the *Jiu Tang shu* 舊唐書 (Old history of the Tang dynasty), it was performed at the tomb of Tang Emperor Xianzong 唐憲宗 (r. 805–20) [Liu 1986, 31.1140] (for background on this form, see Kevin Jensen 2012). The hymn uses the archaizing tetrasyllabic form, rather than the pentasyllabic (or heptasyllabic) form used throughout most Tang verse; this is common in other poems composed for ritual or formal court occasions.[16] The lyrics themselves are fairly straightforward and do not contain much of aesthetic interest, mainly serving to translate the ritual moment into linguistic expression. 45

On the face of it, Duan Wenchang’s poem would seem to share little in common with Wang Gui’s poem, given the differences in poetic subgenre and occasion. However, the DTR (document-topics report) for “Hymns for the Offering to the Ancestral Temple: Dance for Manifesting Virtue” reveals a similar topical mixture: 46

TopicID	Percentage	Topic (first 20 terms in rank order)	Topic Label
30	10.163%	天 大 皇 四 帝 海 萬 太 功 夷 方 王 三 命 元 聖 乾 業 宗 武	empire and sovereign power
87	5.285%	然 化 物 無 浩 天 心 異 造 有 形 方 道 可 忽 中 變 至 窮 言	divine creation and transformation
98	4.065%	神 禮 德 惟 肅 樂 靈 降 薦 既 誠 載 以 明 昭 永 斯 福 陳 備	ritual and ceremony
32	2.846%	十 三 年 二 五 四 六 七 一 八 百 千 九 月 今 歲 前 載 老 第	numbers, calendrical time
149	1.626%	忘 隱 清 心 閒 興 林 靜 勝 自 外 幽 景 愛 機 塵 坐 情 對 境	reclusion
131	1.626%	朝 恩 主 重 詔 拜 承 明 從 臣 榮 紫 賜 門 闕 寵 官 命 列 禁	imperial court
126	1.626%	多 過 更 深 遠 近 宜 入 偏 地 好 和 處 隨 經 移 數 重 客 高	visiting someone
113	1.626%	惡 死 禍 者 敢 罪 危 狼 殺 受 非 防 虎 力 反 利 失 亂 命 傷	disaster
111	1.626%	應 出 隨 朝 從 還 臨 先 行 分 迎 暫 逐 近 節 遠 待 將 旌 方	welcoming / serving?
109	1.626%	江 湘 南 楚 水 客 舟 遠 雲 孤 秋 山 瀟 雁 陽 浦 湖 去 渚 夢	southern (Chu) riverscapes
108	1.626%	復 念 歲 已 懷 所 窮 忽 歎 何 良 憂 未 當 豈 夕 役 終 思 歡	remembering, mourning
83	1.626%	古 草 荒 人 空 在 野 木 平 跡 遺 地 舊 原 猶 城 餘 有 無 蕪	desolate wilderness
72	1.626%	山 石 松 巖 溪 泉 林 幽 雲 鳥 谷 蘿 深 澗 隱 青 徑 野 下 竹	mountain scene
40	1.626%	天 人 地 上 不 生 下 此 長 一 道 得 日 為 間 何 高 意 擾 白	Heaven and Earth
36	1.626%	天 下 太 地 子 中 一 上 海 出 道 入 平 黃 九 四 守 白 大 成	empire and territory
15	1.626%	王 漢 秦 國 宮 吳 陵 帝 楚 武 臺 陽 長 子 安 亡 苑 作 蘇 梁	Han and pre-Han kingdoms

Table 6.

47 The fact that “On Han Gaozu” and “Hymns for the Offering to the Ancestral Temple: Dance for Manifesting Virtue” share Topic 30 (“empire and sovereign power”) as their highest ranking topic is a large factor in the high similarity score between the two documents. However, equally important is how the two documents rank Topics 32 (“numbers, calendrical time”), 98 (“ritual and ceremony”), 87 (“divine creation and transformation”), 109 (“southern [Chu] riverscapes”), 15 (“Han and pre-Han kingdoms”), 83 (“desolate wilderness”), and 36 (“empire and territory”). There are clear thematic similarities here, given that both poems treat the foundings of empire, though no traditional literary history would correlate these two poems, as they would be classified differently based on their metrical forms and subgenres: one a pentasyllabic poem on history and the other a tetrasyllabic ritual poem.

48 Moreover, we can see how a comparison of document topic rankings may help us to understand what similarity and divergence means on a cultural and semantic level. If “On Han Gaozu” and “Hymns for the Offering to the Ancestral Temple” both rank Topic 30 (“empire and sovereign power”) first in terms of topic percentage, they diverge in their second highest ranked topic, which is Topic 16 (“mortality, eating and drinking”) for “On Han Gaozu” and Topic 87 (“divine creation and transformation”) for “Hymns for the Offering to the Ancestral Temple”. This suggests that where a poem on history diverges from a ritual poem that also treats historical events is in the combination of secondary vocabularies. The topic we have labeled “mortality, eating and drinking” evokes subjective experience in a way that connects the lament on historical memory to the individual poet, and this would not be fitting for a ritual poem that is

articulated at the higher register of “divine creation and transformation”.

Finally, related to this, we can also address the problem of topics that *MALLET* identifies in a particular document but that are not clearly evident to the human reader. For example, consider Topic 122 (“fishermen, fishing”), which ranks highly in “On Han Gaozu” and the presence of Topic 149 (“reclusion”), which ranks highly in “Hymns for the Offering to the Ancestral Temple”. The two documents diverge on these topics (among others, of course), but what is interesting is that neither topic is particularly evident in the documents in which they are highly ranked. The question becomes, why does *MALLET* consider these “ghost topics” important for their respective documents when their vocabularies are not clearly invoked by their documents? We hypothesize that ghost topics are identified by the machine reader because they are contiguous to topics that are clearly evidenced in the documents, and that such identifications speak to latent semantic networks to which these documents can be connected.[17] That is to say, if “On Han Gaozu” says nothing about fishermen or fishing (a common trope for the wisdom of humble folk), it nevertheless is comprised of a number of vocabulary sets that are adjacent to the topic of fishermen in some way, from the “southern (Chu) riverscapes” of Topic 122 to the “heroes, men of valor” of Topic 115, implying that “fishermen, fishing” should be located within the semantic networks of these probabilistic distributions. Likewise, there is nothing about “reclusion” in “Hymns for the Offering to the Ancestral Temple”, though the highly ranked cluster of topics that emphasize imperial court ritual, such as Topic 87 (“divine creation and transformation”), Topic 98 (“ritual and ceremony”), and Topic 131 (“imperial court”), create a space for the recluse, who is often figured in diametric opposition to the courtier, as the obverse face of the same sociopolitical coin. 49

To conclude: if we think about poetry not in terms of subgenres or standard metrical forms, but rather in terms of constitutive vocabularies, it becomes clear that there is a shared set of terms that crosses different poetic modes, creating hidden textual communities that would not usually be noticed when reading along traditional throughlines. The detection of hidden or unremarked communities of poems is not in itself literary history, but it lays a new evidentiary foundation for how literary history might be imagined. Instead of narrating the development of poetic style or genre based upon the chronological progression of authors, one might take a radically agnostic thematic approach, one that treats literary compositions as “bags of words” and seeks out the commonalities of vocabulary across these bags, using topics to dictate the selection of documents for argument and analysis. Such an approach might not answer the impossible question of map and territory — of data completeness and selectivity — that haunts narrative historical representation, but it would allow pathways into the true unconscious of literary history, the Great Unread that contemporary databases of digitized corpora are beginning to resurrect and make legible. 50

Here, we reaffirm that human reading on its own cannot make sense of this quantity of data, and machinic reading lacks the capacity to translate patterns into meaningful arguments. It is only a conjoining and interweaving of machinic and human perspectives that make it possible to read a collection on the scale of the *QTS*; it is only through reading the collection as a whole that one may arrive at a comprehensive model of its contents; and it is by engaging with a comprehensive model that literary history may find its standards for the necessary reductions of data that make its narratives possible. 51

Acknowledgements

We would like to thank Timothy R. Tangherlini for his advice throughout, as well as Evan Nicoll-Johnson, Yunshuang Zhang, and Ruichuan Wu for their support during the data preparation phase of the project. We are also grateful to the former UCLA Center for Digital Humanities (now HumTech) for providing a meeting space and resources. 52

Notes

[1] See Rosenthal 2017 for a discussion of the relationship between narrative and data, though the problem of how the narrative form impacts the data model is not raised. On the topic of data modeling, see Flanders and Jannidis 2016, though, conversely, there is no consideration of narrative as a data model.

[2] It is worth noting that Wellek and Warren’s first edition of *Theory of Literature* was published in 1948, the same year that Claude Shannon published his “A Mathematical Theory of Information”, the essay that is often cited as ushering in the Information Age. Wellek and Warren, of

course, do not make reference to the kinds of engineering problems with which Shannon was concerned, but they are conscious of how the vast possible quantities of literary facts impact the writing of literary history. Some work has been done on intersection between the history of information science and the history of literary studies, though there is more that can be done [Geoghegan 2011] [Liu 2010, 153–200].

[3] Unsurprisingly, not all Tang poems are collected in the *QTS*. For an anthology of uncollected Tang poems, see Chen 1992. Poetry recovered from Dunhuang would not have been included in the *QTS* because the Dunhuang Library Cave was not discovered until 1900 [Ren 2014].

[4] Note that the problem of optimizing the ideal number of topics for LDA is still unresolved. This is discussed, with relevant technical sources, in Tangherlini and Leonard 2013, 731.

[5] Han Gaozu 漢高祖 (r. 202–195 BCE), born Liu Bang 劉邦 (256–195 BCE), was the founding emperor of the Han dynasty (202 BCE–220 CE).

[6] This same image is used in the first poem in the “Gufeng” 古風 (“Ancient Airs”) series by Li Bo 李白: “The gathered talents entrust themselves to [the Sage’s] peerless brilliance, / Riding fortune’s course, all are leaping scales” 群才屬休明，乘運共躍鱗. See Qu and Zhu 1980, 2.91; and *QTS* 1960, 161.1670.

[7] This is an auspicious celestial event, the appearance of the first five planets, each identified with one of the Five Phase elements (water, metal, earth, fire, and wood), all in one region of the sky.

[8] That is to say, the people of the world all become Han Gaozu’s subjects.

[9] Xiang Yu 項羽 (232–202 BCE) was a nobleman of Chu who led the rebellion against the Qin dynasty (221–206 BCE) and triumphed, in 207 BCE, with the support of Liu Bang. Xiang Yu then founded the Western Chu, declaring himself king, and warred with Liu Bang over the empire. Liu Bang would eventually triumph in the battle at Gaixia, after which Xiang Yu committe suicide.

[10] This couplet refers to Xiang Yu’s plot to kill Liu Bang, the future Han Gaozu, at a feast at Hongmen and later at Xingyang. At Hongmen, Xiang Yu’s cousin Xiang Zhuang 項莊 was to perform a sword dance and stab Gaozu, but Gaozu was protected by Xiang Chan 項纏, uncle of Xiang Yu and Xiang Zhuang, who joined the dance and blocked each attack. Afterwards, Gaozu’s forces were surrounded at Xingyang (in modern-day Henan), and so he dressed up women in armor, sending them out to surrender, while he himself fled with a handful of men [Sima 1993, 30–33, 38–40].

[11] Han Xin 韓信 (d. 196) and Peng Yue 彭越 (d. 196 BCE) were both military commanders under Liu Bang. Zhang Liang 張良 (d. 186 BCE) and Chen Ping 陳平 (d. 178) were strategists who served under Liu Bang.

[12] Western Chu (206–202 BCE) was the name of the kingdom founded by Xiang Yu after the conquest of Qin.

[13] Classical Chinese rhymes were codified as rhyme-books (sometimes spelled as “rimebooks”) in the medieval period, with rhyming words categorized under header-words that represented the entire rhyme category. We have used David Prager Branner’s reconstructed medieval Chinese transcription in representing the rhyme category header-words [Branner 1999] and made use of his online rhyme database Yintong 音通, which can be accessed here: http://yintong.info/yintong/public/index.php.

[14] This measure takes its name from Jensen’s inequality (devised by Johan L. W. V. Jensen; see Jensen 1906, and Needham 1993) and the Shannon entropy (formulated by Claude E Shannon; see Shannon 1948, and Shannon and Weaver 1949).

[15] There are other ways to calculate document similarity, including vector similarity measures such as cosine similarity (cf. Salton and McGill 1983, 201–4).

[16] Tetrasyllabic verse is generally associated with the *Classic of Poetry* (*Shijing* 詩經), the oldest extant collection of poetry in China and one of the Confucian Classics. Although poets continued to compose in tetrasyllabic form throughout the Han dynasty and the Period of Division (220–589 CE), by the Tang dynasty, most poets would compose in pentasyllabic or heptasyllabic meter, reserving the tetrasyllabic form mainly for ritual genres, archaizing modes, and certain formal occasions [Raft 2007].

[17] This also provides a justification for overshooting the value of n in the initial assignment of the number of topics for *MALLET*. If one sets the number of topics so high that *MALLET* begins finding topics that are not noticeably present in specific documents, these extra topics may still be meaningful as semantic ghosts of the topics that are more evident in the corpus. Lowering the value of n most likely would eventually banish these ghosts.

Works Cited

Allen 1992 Allen, Joseph R. *In the Voice of Others: Chinese Music Bureau Poetry*. Ann Arbor: Center for Chinese Studies, The University of Michigan (1992).

Allen et al. 2019 Allen, Colin, Hongliang Luo, Jaimie Murdock, Jianghuai Pu, Xiaohong Wang, Yanjie Zhai, and Kun Zhao. 2017. "Topic Modeling the Hàn diǎn Ancient Classics". *Journal of Cultural Analytics* (2017). http://doi.org/10.22148/16.016. Accessed September 5, 2019.

Blair 2011 Blair, Ann. *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven: Yale University Press (2011).

Blei 2012 Blei, David. "Probabilistic Topic Models". *Communications of the ACM* 55.4 (2012): 77–84.

Blei et al. 2003 Blei, David, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3 (2003): 993–1022.

Bode 2017 Bode, Katherine. "The Equivalence of *Close* and *Distant* Reading, or, Toward a New Object for Data-Rich Literary History". *Modern Language Notes* 78.1 (2017): 77–106.

Branner 1999 Branner, David Prager. "A Neutral Transcription System for Teaching Medieval Chinese". *T'ang Studies* 17 (1999): 1–169.

Branner and Weng 2004 Branner, David Prager and Yi Weng. *Yintong: Chinese Phonological Database* 音通：聲韻學數據庫. 2004-9.http://yintong.info/yintong/public/index.php.

Chen 1992 Chen, Shangjun 陳尚君, ed. *Quan Tang shi bubian* 全唐詩補編 (Supplement to the *Complete Tang Poems*). 3 vols. Beijing: Zhonghua shuju (1992).

Cohen 1999 Cohen, Margaret. *The Sentimental Education of the Novel*. Princeton: Princeton University Press (1999).

Drucker 2011 Drucker, Johanna. "Humanities Approaches to Graphical Display". *DHQ: Digital Humanities Quarterly* 5.1 (2011). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html. Accessed September 10, 2019.

Drège 2007 Drège, Jean-Pierre. "Des ouvrages classés par catégories: les encyclopédies chinoises". In *Qu'était-ce qu'écrire une encyclopédie en Chine? / What Did It Mean to Write an Encyclopedia in China?*, edited by Florence Establet-Bretelle and Karine Chemla, 19–38. *Extrême-Orient, Extrême-Occident*, hors série (2007).

Flanders and Jannidis 2016 Flanders, Julia and Fotis Jannidis. "Data Modeling". In *A New Companion to the Digital Humanities*, edited by Susan Scriebman, Ray Siemens, and John Unsworth, 229–37. Chichester, UK: John Wiley & Sons (2016).

Geoghegan 2011 Geoghegan, Bernard Dionysius. "From Information Theory to French Theory: Jakobson, Lévi-Strauss, and the Cybernetic Apparatus". *Critical Inquiry* 38 (2011): 96–126.

Hayles 2010 Hayles, N. Katherine. "How We Read: Close, Hyper, Machine". *ADE Bulletin* 150 (2010): 62–79.

Hickman and McIntyre 2012 Hickman, Miranda B. and John D. McIntyre, eds. *Rereading the New Criticism*. Columbus: Ohio University Press (2012).

Hu 2003 Hu, Zhenheng 胡震亨 (1569–ca. 1644), *Tangyin tongqian* 唐音統籤 (Assembled documents of Tang tones). 9 vols. Shanghai: Shanghai guji chubanshe (2003).

Huang 2013 Huang, Shuanfan Huang. *Chinese Grammar at Work*. Amsterdam: John Benjamins Publishing Company (2013).

Jensen 1906 Jensen, J. L. W. V. "Sur les fonctions convexes et les inégalités entres les valeurs moyennes". *Acta Mathematica* 30 (1906): 175–93.

Jensen 2012 Jensen, Kevin A. "Wei-Jin Sacrificial Ballets: Reform versus Conservation". Ph.D. diss., University of Washington (2012).

Jockers 2013 Jockers, Matthew L. *Macroanalysis: Digital Methods & Literary History*. Urbana: University of Illinois Press (2013).

Lin 1991 Lin, Jianhua. "Divergence Measures Based on the Shannon Entropy". *IEEE Transactions on Information Theory* 37.1 (1991): 145–51.

Liu 1986 Liu, Xu 劉煦 (887–947), *Jiu Tang shu* 舊唐書 (Old history of the Tang). 16 vols. Beijing: Zhonghua shuju (1986).

Liu 2010 Liu, Lydia. *The Freudian Robot: Digital Media and the Future of the Unconscious*. Chicago: University of Chicago Press (2010).

Liu et al. 2018 Liu, Chao-Lin, Thomas J. Mazanec, and Jeffrey R. Tharsen.. “Exploring Chinese Poetry with Digital Assistance: Examples from Linguistic, Literary, and Historical Viewpoints”. *Journal of Chinese Literature and Culture* 5.2 (2018): 276–321.

Long So 2016 Long, Hoyt and Richard Jean So. “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning”. *Critical Inquiry* 42 (2016): 235–67.

McCallum 2002 McCallum, Andrew Kachites. *MALLET: A Machine Learning for Language Toolkit*. 2002. Available at: http://mallet.cs.umass.edu/index.php.

Meeks 2012 Meeks, Elijah and Scott B. Weingart, eds. Special issue, “The Digital Humanities Contribution to Topic Modeling”, *Journal of Digital Humanities* 2.1 (2012). Available at: http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/. Accessed January 9, 2018.

Moretti 2000a Moretti, Franco. “Conjectures on World Literature”. *New Left Review* 1 (2000a): 54–68.

Moretti 2000b — . “The Slaughterhouse of Literature”. *MLQ: Modern Language Quarterly* 61, no. 1 (2000b): 207–27.

Needham 1993 Needham, Tristan. “A Visual Explanation of Jensen’s Inequality”. *The American Mathematical Quarterly* 100.8 (1993): 768–71.

Nichols et al. 2018 Nichols, Ryan, Edward Slingerland, Kristoffer Nielbo, and Uffe Bergeton. “Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-Learning Approach”. *The Journal of Asian Studies* 77.1 (2018): 19–57. Available at: https://doi.org/10.1017/S0021911817000973. Accessed September 5, 2019.

North 2013 North, Joseph. “What’s *New Critical* about *Close Reading*? I.A. Richards and His New Critical Reception”. *New Literary History* 44, no. 1 (2013): 141–57.

Owen 1977 Owen, Stephen. *The Poetry of the Early T’ang*. New Haven: Yale University (1977).

Owen 1981 — . *The Great Age of Chinese Poetry: The High T’ang*. New Haven: Yale University (1981).

Owen 2006a — . *The Late Tang: Chinese Poetry of the Mid-Ninth Century (827–860)*. Cambridge: Harvard University Asia Center (2006a).

Owen 2006b — . *The Making of Early Chinese Classical Poetry*. Cambridge: Harvard University Asia Center (2006b).

Perkins 1992 Perkins, David. *Is Literary History Possible?* Baltimore: The Johns Hopkins University Press (1992).

Piper 2015 Piper, Andrew. “Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel”. *New Literary History* 46 (2015): 63–98.

Piper 2018 — . *Enumerations: Data and Literary Study*. Chicago: University of Chicago Press (2018).

QTS 1960 *Quan Tang shi* 全唐詩 (Complete Tang poems). 25 vols. Beijing: Zhonghua shuju (1960).

Qian 2000 Qian, Zhonglian 錢仲聯, gen. ed., et al. *Zhongguo wenxue dacidian* 中國文學大辭典 (Great dictionary of Chinese literature). 2 vols. Shanghai: Shanghai guji chubanshe (2000).

Qian and Ji 1979 Qian, Qianyi 錢謙益 (1582–1664) and Ji Zhenyi 季振宜 (b. 1630), comps. *Quan Tang shi gaoben* 全唐詩稿本 (Draft edition of the complete Tang poems). Edited by Qu Wanli 屈萬里 and Liu Zhaoyou 劉兆祐. 71 vols. Taipei: Lianjing chuban shiye gongsi (1979).

Qu and Jincheng 1980 Qu, Tuiyuan 瞿蛻園 and Zhu Jincheng 朱金城, eds. *Li Bai ji jiaozhu* 李白集校注 (Annotated edition of Li Bai’s literary collection). 4 vols. Shanghai: Shanghai guji chubanshe (1980).

Raft 2007 Raft, Zebulon David. “Four-syllable Verse in Medieval China”. Ph.D. diss., Harvard University (2007).

Ramsay 2011 Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press (2011).

Ren 2014 Ren, Zhongmin 任中敏, ed. *Dunhuang geci zongpian* 敦煌歌辭總編 (Complete edition of Dunhuang songs and lyrics). 3 vols. Nanjing: Fenghuang chubanshe (2014).

Riddell 2012 Riddell, Allen B. “A Simple Topic Model (Mixture of Unigrams)”. July 22, 2012. Available at: https://www.ariddell.org/simple-topic-model.html. Accessed January 9, 2018.

Rosenberg 2013 Rosenberg, Daniel. “Data before the Fact”. In *“Raw Data” Is an Oxymoron*, edited by Lisa Gitelman, 15–40. Cambridge: MIT Press (2013).

Rosenthal 2017 Rosenthal, Jesse. “Introduction: Narrative against Data”. *Genre* 50.1 (2017): 1–18.

Salton and McGill 1983 Salton, Gerard and Michael J. McGill. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company (1983).

Shannon 1948 Shannon, Claude E. “A Mathematical Theory of Communication”. *Bell Systems Technical Journal* 27 (July, October 1948): 379–423, 623–56.

Shannon and Weaver 1949 Shannon, Claude E. and Warren Weaver. *The Mathematical Theory of Communication* Urbana: The University of Illinois Press (1949).

Sima 1993 Sima Qian. *Records of the Grand Historian: Han Dynasty I*. Trans. Burton Watson. Rev. ed. Hong Kong and New York: The Research Centre for Translation, Chinese University of Hong Kong and Columbia University Press (1993).

Spence 1966 Spence, Jonathan D. *Ts’ao Yin and the K’ang-hsi Emperor*. New Haven: Yale University Press (1966).

Tangherlini and Leonard 2013 Tangherlini, Timothy R. and Peter Leonard. “Trawling the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research”. *Poetics* 41: 725–49.

Tian 2017 Tian, Xiaofei. “Literary Learning: Encyclopedias and Epitomes”. In *Oxford Handbook of Classical Chinese Literature (1000 BCE–900 CE)*, edited by Wiebke Denecke, Wai-Yee Li, and Xiaofei Tian, 132–46. New York: Oxford University Press (2017).

Tong 1998 Tong Peiji 佟培基. “Jin sanbai nian *Quan Tang shi* de zhengli yu yanjiu” 近三百年《全唐詩》的整理與研究 (Corrections and research on the *Complete Tang poems* over the last three hundred years), *Wenxian* 文獻 (Textual evidence) 59.3 (1998): 17–28.

Underwood 2012 Underwood, Ted. “What Kind of ‘Topics’ Does Topic Modeling Actually Produce?” April 1, 2012. Available at: https://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/. Accessed January 9, 2018.

Underwood 2017 — . “A Genealogy of Distant Reading”. *DHQ: Digital Humanities Quarterly* 11.2 (2017). Available at: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html. Accessed September 11, 2019.

Vierthaler 2016 Vierthaler, Paul. “Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature”. *Journal of Cultural Analytics* (2016). DOI: 10.31235/osf.io/t9b7v. Accessed August 20, 2019.

Wellek and Warren 1956 Wellek, René and Austin Warren. *Theory of Literature*. 3rd ed. New York: Harcourt, Brace & World (1956).

Xu 1994 Xu, Zong 許總. *Tang shi shi* 唐詩史 (A history of Tang poetry). Nanjing: Jiangsu jiaoyu chubanshe (1994).

Yang 1996 Yang, Shiming 楊世明. *Tang shi shi* 唐詩史 (A history of Tang poetry). Chongqing: Chongqing chubanshe (1996).

Yu 1990 Yu, Pauline. “Poems in Their Place: Collections and Canons in Early Chinese Literature”. *Harvard Journal of Asiatic Studies* 50.1 (1990): 163–96.

Yu 1994 — . “The Chinese Poetic Canon and Its Boundaries”. In *Boundaries in China*, edited by John Hay, 105–23. London: Reaktion Books (1994).