Quantitative patterns of stylistic influence in the evolution of literature

James M. Hughes[a], Nicholas J. Foti[a], David C. Krakauer[b,c], and Daniel N. Rockmore[a,b,d,e,1]

[a]Department of Computer Science, Dartmouth College, Hanover, NH 03755; [b]Santa Fe Institute, Santa Fe, NM 87501; [c]Wisconsin Institute for Discovery, University of Wisconsin, Madision, WI 53715; [d]Department of Mathematics, Dartmouth College, Hanover, NH 03755; and [e]Neukom Institute for Computational Science, Dartmouth College, Hanover, NH 03755

Edited by* Michael S. Gazzaniga, University of California, Santa Barbara, Santa Barbara, CA, and approved March 13, 2012 (received for review September 21, 2011)

Literature is a form of expression whose temporal structure, both in content and style, provides a historical record of the evolution of culture. In this work we take on a quantitative analysis of literary style and conduct the first large-scale temporal stylometric study of literature by using the vast holdings in the Project Gutenberg Digital Library corpus. We find temporal stylistic localization among authors through the analysis of the similarity structure in feature vectors derived from content-free word usage, nonhomogeneous decay rates of stylistic influence, and an accelerating rate of decay of influence among modern authors. Within a given time period we also find evidence for stylistic coherence with a given literary topic, such that writers in different fields adopt different literary styles. This study gives quantitative support to the notion of a literary "style of a time" with a strong trend toward increasingly contemporaneous stylistic influence.

cultural evolution | stylometry | culture | complexity | big data

Written works, or literature, provide one of the great bodies of cultural artifacts. The analysis of literature typically involves the aggregation of information on several levels, ranging from words to sentences and even larger scale properties of temporal narratives such as structure, plot, and the use of irony and metaphor (1–3). Quantitative methods have long been applied to literature, most notably in the analysis of style, which can be traced back to a comment by the mathematician Augustus de Morgan regarding the attribution of the Pauline epistles (4) and the late nineteenth century work of the historian of philosophy Wincenty Lutasłowski, who brought basic statistical ideas of word usage to the problem of dating the dialogues of Plato (5). It was Lutasłowski who coined the word "stylometry" to describe such an approach to investigating questions of literary style. Since then, a wide range of statistical techniques have been developed toward this end (6), generally with the goal of settling questions of author attribution (see, e.g., refs. 6–11). Stylometric studies have also been pursued in the study of visual art (12, 13) and music [both in composition (14–16) and performance (17)], and are part of a growing body of work in the quantitative analysis of cultural artifacts (18).

In this paper we report our findings from the first large-scale stylometric analysis of literature. The goal of this work is not author attribution—for the authorship of all the works is well known—but is instead to articulate, in a quantitative fashion, large-scale temporal trends in literary (i.e., writing) style. This type of study has been, until now, impossible to undertake, but the advent of mass digitization has created dramatic new opportunities for scholarly studies in literature as well as in other disciplines (19). Our literature sample is obtained from the Project Gutenberg Digital Library (http://www.gutenberg.org/wiki/Gutenberg:About). Project Gutenberg consists of more than 30,000 public domain texts, music, audiobooks, etc., is freely available online, and is among the digital archives that have become, over the past 60 yr, crucial components of the preservation of cultural artifacts (18).

In scope, our work is related to, but quite different from, recent studies in the dating of literary works (20), the analysis of the coarse-grained structure of literary history (and the evolution of genre) (21), and most notably, a recent analysis of Google Books (22), wherein the temporal trends in content-word usage were articulated. Content words also form the basis of the various topic model analyses that have been applied to large corpuses of science texts (see, e.g., ref. 23).

In contrast, the work presented here focuses on the usage of content-free words as the basis of the first large-scale study of the similarity structure of literary style. Content-free words are the "syntactic glue" of a language: They are words that carry little meaning on their own but form the bridge between words that convey meaning. Their joint frequency of usage is known to provide a useful stylistic fingerprint for authorship (8, 11), and thus suggests a method of comparing author styles. When we consider content-free word frequencies from a large number of authors and works over a long period of time, we can ask questions related to temporal trends in similarity. The primary results of our analysis are that time provides the most coherent means of clustering work and a trend of diminishing stylistic influence as we move forward in time. Such a finding is consistent with a simple evolutionary model for stylistic influence, which assumes that imitation attends preferentially to contemporary authors. In addition, we uncover quantitative support of the previously purely anecdotal notion of a literary "style of a time." Taken together, these two findings suggest the utility and perhaps the creation of a new field of stylometric analysis in culturomics.

Materials and Methods

In our experiments, we studied a subset of the authors in the Project Gutenberg database composed of those who wrote after the year 1550, had at least five works in English in the Project Gutenberg collection, and for whom we had birth and death date information. This left us with 537 authors. For each author, we created a representative feature vector by aggregating the content-free word frequencies for each individual work by that author. In total, we analyzed 7,733 works.

In our experiments, we used a list of 307 content-free words that included prepositions, articles, conjunctions, "to be" verbs, and some common nouns and pronouns (see Table S1 for a complete list). We did not attempt to semantically disambiguate between occurrences of homographs in situ (e.g., when using "to" as a preposition or as an indicator of an infinitive verb). Doing so would require a sophisticated grammatical model, and it was not our aim to model this particular aspect of word usage. We believe that ignoring these distinctions is not likely to greatly affect our results, because words that account for the greatest differences in usage frequency among

Author contributions: J.M.H., N.J.F., D.C.K., and D.N.R. designed research; J.M.H., N.J.F., D.C.K., and D.N.R. performed research; J.M.H., N.J.F., D.C.K., and D.N.R. contributed new reagents/analytic tools; J.M.H., N.J.F., D.C.K., and D.N.R. analyzed data; and J.M.H., N.J.F., D.C.K., and D.N.R. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: rockmore@cs.dartmouth.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1115407109/-/DCSupplemental.

authors are not likely to be homographs. After counting the occurrences of each content-free word for each work by a particular author, we aggregated these counts and normalized them so that the components summed to 1 (L^1-norm). We then took each of these normalized vectors to be the feature vector for the corresponding author. We compared authors by computing the symmetrized Kullback–Leibler divergence between their feature vectors, given by

$$D_{\mathrm{KL}}(P_i, P_j) = \frac{1}{2}\sum_{\omega\in\Omega}\left(P_i(\omega)\log\frac{P_i(\omega)}{P_j(\omega)}\right) + \left(P_j(\omega)\log\frac{P_j(\omega)}{P_i(\omega)}\right), \quad [1]$$

where Ω is the set of content-free words and $P_i(\omega)$ is the corresponding unitized feature vector for author i. Using Eq. **1** to define the distance d_{ij} between authors i and j, we derived a similarity matrix with elements $S_{ij} = \exp(-d_{ij}/\sigma)$, with σ equal to 0.5. The value of σ was chosen to spread the values of S_{ij} out so that they occupied as a whole most of the unit interval [0, 1].

A more explicit means of understanding the connection between stylistic similarity and temporal distance is to consider how the average similarity between two authors changes as the distance between authors in time increases. Specifically, we consider the sets of similarities $S(t) = \{S_{ij}\ \text{s.t.}\ |y_i - y_j| \le t\}$, for several values of t, ranging from 2 yr to 389 yr (the maximum temporal distance between any two authors). For each value of t, we take the average of the values in $S(t)$,

$$S_{\mathrm{avg}}(t) = \frac{1}{|S(t)|}\sum_{s\in S(t)} s.$$

In order to get at a more localized notion we also considered a windowed version

$$S_W(t) = \mathrm{mean}\{S_{ij}\ \text{all}\ i, j\ \text{s.t.}\ (t-3) \le |y_i - y_j| \le (t+3)\}$$

average only among those that authors fall within 3 yr in either direction of the current temporal distance t. Thus, for example, $S_{\mathrm{avg}}(100)$ should be read as the average similarity to all authors separated by 100 yr and less, whereas $S_W(100)$ should be read as average similarity to all authors between 97 and 103 yr apart.

In order to understand the relationships between author styles, we considered the similarities between author feature vectors that were statistically significant according to the local distribution of similarity for each individual author. Specifically, we identified significantly large similarities by using the empirical distribution of similarity values for a given author. In our representation, we view each author i as possessing a distribution of 536 similarity values that describes author i's stylistic relationship to all other authors. We compute the $1-\alpha$ quantile of this distribution and consider all authors whose similarity value exceeds this quantile to have statistically significant stylistic similarity to author i. In our experiments, we chose $\alpha = 0.002$, though the results for different values of α were qualitatively similar. Similar methods have been proposed for the detection of meaningful links in highly connected, complex networks (24, 25).

When considering these statistically significant stylistic connections, temporal structure quickly reveals itself. Authors tend to have important connections to other authors from roughly the same time period. We computed a temporal disparity metric d_i for each author i as the median temporal distance from author i to each of his or her neighbors j, i.e.,

$$d_i = \mathrm{median}\{|y_i - y_j|, \quad \text{for all}\ j \neq i\},$$

where y_k is the so-called representative year of author k, equal to the average of the author's birth and death years.

Results

We first view the extent to which the local structure of the important stylistic connections between authors is composed of temporally localized information by examining the distribution of the temporal disparity statistic across all authors. This is the distribution over all authors of the distance in time of those (other) authors whose style is significantly similar (see *Materials and Methods* for details). Fig. 1 shows this distribution, and it is clearly heavily right-skewed, indicating that authors tended to have statistically significant connections to other authors close to them in time. On average, authors were approximately 24 yr apart from their neighbors. Indeed, over 85% of authors had an associated temporal disparity of less than 37 yr, remarkable considering the means of representing the working period of each author as well as the fact that the similarity metric does not explicitly take time into account.

In order to assess the significance of this result, we performed 500 simulations to measure the average temporal disparity on a set of authors with the same set of similarity values as in our dataset but with author years permuted randomly. We observed that the true average temporal disparity was smaller than any average disparity observed in the simulations, indicating that observing an average temporal disparity as small as the one we observed is a highly significant event. This analysis represents quantifiable support for the anecdotal claims of a literary "style of a time."

Our next set of results are the primary discovery of the paper. Herein we consider the temporal nature of similarity. In Fig. 2 we see that as the temporal distance between authors increases in size, the average similarity between authors tends to decrease, until it converges to the overall average similarity but with one important exception. Just as authors tended to have important stylistic connections to other authors close to them in time (but not necessarily to immediate contemporaries), so does the trend of similarity increase as the time window size increases, before decreasing precipitously. The envelope around the similarity curve represents the ± 2 standard deviation bounds on the true average similarity value for the corresponding time window. These were estimated via bootstrap resampling (26) of the set of values $S(t)$ 500 times. The flat, dashed red line in Fig. 2 plots the global average value. The error envelope around the similarity curve allows us to estimate whether the average similarity value we observed was significantly different from the overall mean. For all values of t less than roughly 310 yr, the average similarity value was significantly different from the overall mean, but at the point $t \approx 310$, the overall average falls within the similarity curve's error bounds. If we take similarity as a proxy for influence (and of course recall that influence can only be exerted from authors who are earlier in time), then essentially, this means that at a certain point in the past, the influence of temporally distant authors is indistinguishable from what we would expect at random.

Figs. 1 and 2 show that the distribution of similarity between writing styles clearly varies (i.e., is not uniformly distributed) as a function of temporal distance between authors. In our next study we look more closely at this variation. We split the authors into

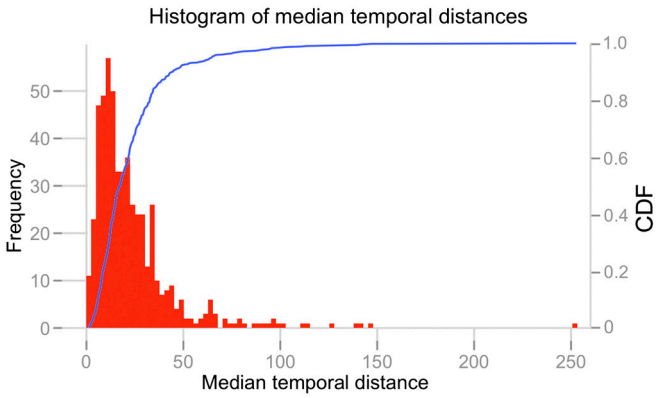

Fig. 1. The distribution of temporal disparity indicates a significant amount of temporal localization, because most authors have important connections with other authors that are close to them in time. More than 85 percent of authors had a temporal disparity of less than 35 yr, and the overall average temporal disparity was approximately 23 yr.

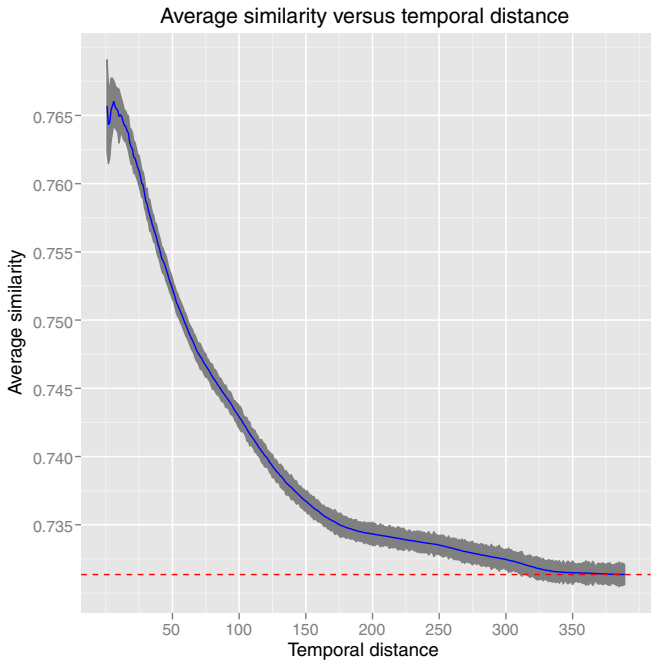

Fig. 2. Average similarity between authors as a function of temporal distance between them. Clearly, as the distance between authors increases, the similarity between them tends to decrease. The flat dashed red line marks the global average.

two time periods of equal length, "early" and "late." For the early authors (those who wrote between 1550 and 1783), the average similarity as a function of temporal distance does not deviate significantly from the overall average, suggesting that authors during this time period influence each other roughly equally, regardless of how far apart in time they are.

However, for the "modern" authors (those who wrote between 1784 and 1952), the average similarity curve was high for shorter temporal distances and decreased rapidly toward the mean, much like the overall trend shown in Fig. 2. In order to examine the extent to which there is a shift in the way authors were influenced based on when they wrote, we split the modern authors into quartiles defined over the range of most densely populated years (see Fig. 3) by partitioning them according to their representative author years. The four partitions consist of the years 1784–1829, 1825–1870, 1866–1911, and 1907–1952. There is a small amount of overlap (5 yr) between these groups in order to mitigate edge effects.

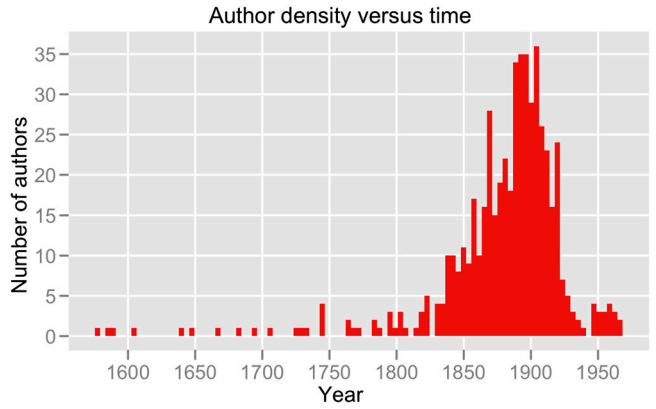

Fig. 3. Density of authors in our dataset as a function of time. The vertical axis indicates how many authors fell into the corresponding time window.

Fig. 4 *A* and *B* display $S_{\text{avg}}(t)$ and $S_W(t)$ for each of these groups separately, indicating that they possess remarkably different patterns of similarity as a function of temporal distance.

In Fig. 4*A* we plot $S_{\text{avg}}(t)$ for all authors in the corresponding time window. For the early modern period (1784–1870), the similarity functions do not differ significantly from the average (indicated by the dashed red line), which suggests that authors during that period tended to draw influence from other authors uniformly as a function of temporal distance. The same pattern is observed for the windowed analysis in the same time period (see Fig. 4*B*). Thus over this period there is no significant evidence for stylistic localization in time.

For the late modern quartiles (1866–1911 and 1907–1952) the pattern is very different. In the period 1866–1911, authors are significantly more similar to members of their own age cohort, and interestingly this similarity decreases toward the average, for the cumulative analysis. In other words, above 30 yr apart, authors are not significantly more like any set of authors chosen at random, regardless of how far apart they are in time. When we consider the windowed analysis, we see more structure. Above 20 yr apart, authors actually tend to be less like each other than the average. This suggests that there is a significant decline in the similarity between authors who are widely separated in time. The "repulsive" effect of temporal distance is consistent within this period, and similarity between authors decreases throughout the years in question.

In the later period, 1907–1952, this pattern repeats but with stronger effects. Contemporaneous authors are most similar and average similarity decays to the within-group average with increasing temporal separation. The rate of decay is now nonlinear and scales quadratically in time ($s \approx t^{-2}$), which suggests that authors tend to be influenced by their contemporaries more strongly than during 1866–1911. In other words, the amount of temporal distance until contributions of authors becomes indistinguishable from the average is smaller than in the earlier period spanning 1866–1911. In the later period, average similarity was indistinguishable from the mean after approximately 23 yr, whereas in the earlier period, it was not until almost 30 yr that influence became random. The windowed analysis, as before, exhibits greater structure. Average similarity $S_W(t)$ is no longer monotonic in time but has a minimum at 25 yr, returning to the average at the maximum calculated separation of 40 yr. This suggests that for the modern period the pattern observed over the complete dataset is reversed. Whereas when we consider the average similarity over the complete dataset, the most similar authors are around 24 yr apart, in the late modern quartile, authors separated by 25 yr are maximally different. These findings are robust under changes in sampling (see Fig. S1).

Discussion

It is a remarkable fact that vectors of content-free words—subject-independent textual features of a book—allow us to cluster authors in time and by narrative theme, and that content-free word frequencies are fairly faithfully transmitted among authors of a similar period, even when imitation at this level of textual resolution seems to be out of the question. As we move into the present, this imitation becomes increasingly localized to our contemporaries.

We propose that for the earliest periods in our dataset, and the early modern period, the number of published works remained relatively low. This allowed authors to have sufficient time to sample (read) very broadly from the full range of historically published works. Common phrasing, and norms of syntax and grammar, remain relatively unchanged for long periods of time. This generates decay rates in similarity as a function of temporal distance that are not significantly different from the average, because authors are influenced by models distributed uniformly in time. However, for more recent authors, the number of possible

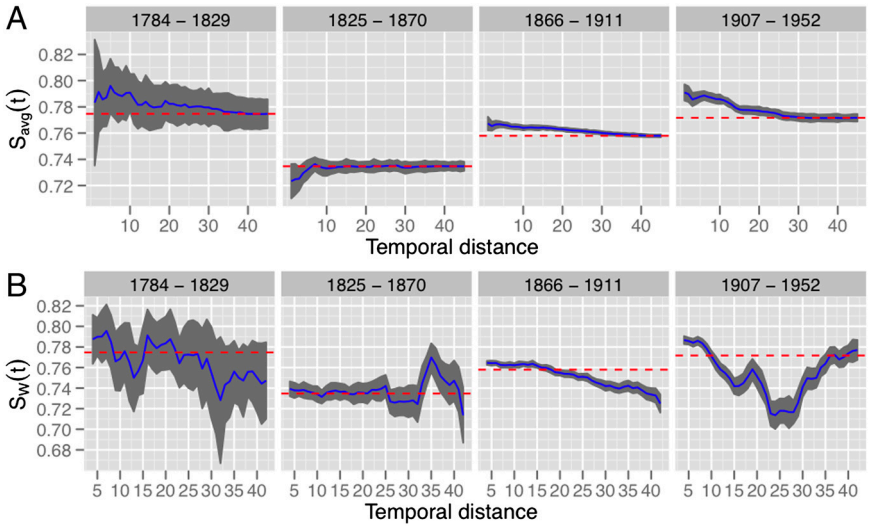

Fig. 4. (*A*) Average similarity between authors as function of the temporal distance between them, for four groups: authors who wrote between 1784–1829, 1825–1870, 1866–1911, 1907–1952, using $S_{avg}(t)$. Note that similarity changes differently as a function of temporal distance, depending on which era is considered. This suggests different "regimes" of stylistic influence. (*B*) Average similarity between authors as function of the temporal distance between them for the same four groups as in *A*, except that these curves depict the windowed analysis $S_W(t)$. Note the vastly different trends in influence between the two earlier and two later subdivisions. The flat dashed line in each interval marks the overall average in each time period.

choices of books to read has increased dramatically, and with a finite amount of time, a subset of these works must be chosen, leading to rather heterogeneous reading patterns and a greater overall diversity of authored works. The pattern accelerates in the later modern period, with even more authors to choose from and selection dominated by contemporaneous authors. This suggests a simple evolutionary model for patterns of influence (see *SI Text*).

The negative influence of authors from a preceding generation in the period 1907–1952 could be explained by the Modernist movement. Modernist authors, who are contained within this time period, display a radical shift in style as they reject their immediate stylistic predecessors yet remain a part of a dominant movement that included many of their contemporaries. The contemporary influence of writing programs and their often close readings of contemporary works and feedback (sometimes called "reflexive modernism") has also been suggested to contribute to this effect (27). The overall pattern that we find is that the stylistic influence of the past is diminishing at an increasing rate, which suggests that style itself is evolving at an accelerating pace.

The patterns of influence are a first discovery from the corpus. Implicit in this is a temporal clustering of similarity and quantitative support for the qualitative suggestions of a notion of a "style of a time." It is also worth noting that the implicit temporal clustering of similarity is not an exclusively temporal phenomenon. Fig. S2 shows a network representation of the authors in which a preliminary investigation reveals evidence of thematic clustering as well. Examples include interesting groupings of English poets and playwrights, military leaders, and a collection of important naturalists, social thinkers, and historians. This is suggestive and supportive of the hypothesis that word frequencies are not only typical of a given time but also of a field of inquiry. Historians and naturalists do not only write about different topics, they write about them differently. Taken together with the patterns of decay in influence this suggests that whereas authors of the 18th and 19th centuries continued to be influenced by previous centuries, authors of the late 20th century are strongly influenced by authors from their own decade. The so-called "anxiety of influence" (28), whereby authors are understood in terms of their response to canonical precursors, is becoming an "anxiety of impotence," in which the past exerts a diminishing stylistic influence on the present. These results are consistent with many complex, scaling phenomena such as those found in urban and technological systems, where there has been an accelerating rate of change into the present. This is a rather intriguing pattern of short-term cultural evolution that is different from the constant rates of change reported for names and pottery (29) or the reduced rates of lexical substitution of frequently used words over thousands of years (30). Further analysis will elucidate not only the transmission mechanisms generating temporally localized styles but additional stylistic factors that help differentiate the style of one author from that of another.

ACKNOWLEDGMENTS. D.C.K.'s contribution to this project/publication was made possible through the support of a grant from the John Templeton Foundation.

1. Aristotle (1997) *Poetics* (Penguin Classics, London).
2. Auerbach E (2003) *Mimesis: The Representation of Reality in Western Literature* (Princeton Univ Press, Princeton, NJ).
3. Booth Wayne C (1983) *The Rhetoric of Fiction* (Univ of Chicago Press, Chicago).
4. de Morgan SE (1882) *Memoir of Augustus de Morgan, by his wife Sophia Elisabeth de Morgan, with Selection of his Letters* (Longmans, Green, London).
5. Lutostowski W (1897) *The Origin and Growth of Plato's Logic* (Longmans, Green, London), Reprint: *The Origin and Growth of Plato's Logic*, Georg Olms, Hildesheim, 1983.
6. Holmes DI, Kardos J (2003) Who was the author? An introduction to stylometry. *Chance* 16(2):5–8.
7. Williams DS (1992) *Stylometric Authorship Studies in Flavius Josephus and Related Literature* (Edwin Mellen Press, Lewiston, NY).
8. Juola P (2006) Authorship attribution. *FTIR* 1:233–334.
9. Mosteller F, Wallace D (1964) *Inference and Disputed Authorship: The Case of the Federalist Papers* (Addison-Wesley, Reading, MA).
10. Williams CB (1975) Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika* 62:207–212.
11. Binongo JNG (2003) Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance* 16(2):9–17.
12. Taylor RP, Micolich AP, Jonas D (1999) Fractal analysis of Pollock's drip paintings. *Nature* 399:422.
13. Hughes JM, Graham DJ, Rockmore DN (2010) Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proc Natl Acad Sci USA* 107:1279–1283.
14. Manaris B, et al. (2005) Zipf's law, music classification, and aesthetics. *Comput Music J* 29(1):55–69.
15. Huron D (1991) The ramp archetype: A study of musical dynamics in 14 piano composers. *Psychol Music* 19(1):33–45.
16. Casey M, Rhodes C, Slaney M (2008) Analysis of minimum distances in high-dimensional music spaces. *IEEE Trans Audio Speech Lang Processing* 16(5):1015–1028.
17. Sapp C (2008) Hybrid numeric/rank similarity metrics for musical performances. *Proceedings of ISMIR* 99:501–506.
18. Hockey S (2004) The history of humanities computing. *A Companion to Digital Humanities*, eds S Schreibman, R Siemens, and J Unsworth (Blackwell, Oxford, UK), pp 3–19.

19. Evans JA, Foster JG (2011) Metaknowledge. *Science* 331:721–725.
20. Stamou C (2009) *Dating Victorians* (VDM, Saarbrücken, Germany).
21. Moretti F (2005) *Graphs, Maps, Trees* (Verso, London and New York) Afterword by Alberto Piazza.
22. Michel J-B, et al., and The Google Books Team (2010) Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182.
23. Lafferty JD, Blei DM (2007) A correlated topic model of science. *Ann Stat* 1(1):17–35.
24. Serrano M, Boguna M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci USA* 106:6483–6488.
25. Foti NJ, Hughes JM, Rockmore DN (2011) Nonparametric sparsification of complex multiscale networks. *PLoS One* 6:e16431.
26. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7:1–26.
27. McGurl M (2009) *The Program Era* (Harvard Univ Press, Cambridge, MA).
28. Bloom H (1997) *The Anxiety of Influence: A Theory of Poetry* (Oxford Univ Press, New York).
29. Benley RA, Hahn MW, Shennan Sj (2004) Random drift and culture change. *Proc R Soc Lond B* 271:1443–1450.
30. Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449:717–720.