The Computational Case against Computational Literary Studies

Nan Z. Da

1

This essay works at the empirical level to isolate a series of technical problems, logical fallacies, and conceptual flaws in an increasingly popular subfield in literary studies variously known as cultural analytics, literary data mining, quantitative formalism, literary text mining, computational textual analysis, computational criticism, algorithmic literary studies, social computing for literary studies, and computational literary studies (the phrase I use here). In a nutshell the problem with computational literary analysis as it stands is that what is robust is obvious (in the empirical sense) and what is not obvious is not robust, a situation not easily overcome given the nature of literary data and the nature of statistical inquiry. There is a fundamental mismatch between the statistical tools that are used and the objects to which they are applied.

Digital humanities (DH), a field of study which can encompass subjects as diverse as histories of media and early computational practices, the digitization of texts for open access, digital inscription and mediation, and computational linguistics and lexicology, and technical papers on data mining, is not the object of my critique. Rather, I am addressing specifically the project of running computer programs on large (or usually not so large) corpora of literary texts to yield quantitative results which are then mapped, graphed, and tested for statistical significance and used

With thanks to my research assistants Tong Ding for his help with all of the data work and Emily Howard for assisting me with data input. Thanks also to Bill McDonald for his insights. Metadata, scripts, and instructions for replication are available at www.github.com/nan-da. The appendix is available online.

Critical Inquiry 45 (Spring 2019)

to make arguments about literature or literary history or to devise new tools for studying form, style, content, and context. Another suitable definition of computational literary studies (CLS) is the statistical representation of patterns discovered in text mining fitted to currently existing knowledge about literature, literary history, and textual production to close what Andrew Piper, in his manifesto "There Will Be Numbers," calls "the evidence gap."[1] CLS claims that literary critics will no longer make unsupported claims about whole periods of literary history using just a few texts or ignore large swaths of literary production—CLS (says Piper) can show us new things and keep us honest by giving us a way to back up claims with empirical evidence, or by using said evidence to challenge various conventional wisdoms about literary history (such as claims about style, genre, periodization, and so on).

Literary scholars have few ways to check CLS work, sometimes owing to problems with access.[2] There are also disciplinary circumstances that have made criticisms against CLS hard to mount, such as the mainstreaming of network literary sociology and the semantic reduction of the meaning of *form* and *formalism* to trackable units and a study of the patterns made by trackable things. CLS has also adopted an approach to critical contribution characterized by modesty, supplementarity, or incrementality, reframing setbacks as a need to modify methodology and generate more testing. So, while Piper comments "there have by now been so many polemics written for and against the use of data to study literature, culture, media and history that to offer one more rationale seems perilously unnecessary," he goes on to say, "What is needed, for sure, is more research—more research into why exactly, why right now, the computational study of culture *is* necessary."[3] CLS claims to produce exploratory tools that, even if wrong, are intrinsically valuable because exploration is intrinsically valuable. Misclassifications become objects of interest, imprecisions become theory, outliers turn into aesthetic and philosophical explorations, and all merit more funding and more publications. This kind of strategic incremental-

1. Andrew Piper, "There Will be Numbers," *Journal of Cultural Analytics*, 23 May 2016, culturalanalytics.org/2016/05/there-will-be-numbers/

2. CLS authors are right to argue that critiques of their work ought to address more than one or two articles but the process of requesting complete, runnable codes and quantitative results (tables, output data, matrices, measurements, and others) took me nearly two years. Authors and editors either never replied to my emails, weren't able or willing to provide complete or runnable scripts and data, or gave them piecemeal only with repeated requests.

3. Piper, "There Will Be Numbers."

Nan Z. Da teaches literature at the University of Notre Dame.

ism has made some of the most vocal critics temper their argument—after all, who would not want to appear reasonable, forward-looking, open-minded?

There are critiques of CLS in place—notably Timothy Brennan's "The Digital Humanities Bust" and Danielle Allington, Sarah Brouillette, and David Golumbia's "Neoliberal Tools (and Archives): A Political History of the Digital Humanities."[4] Political and philosophical critiques of DH have made significant contributions to our understanding of the institutional and ideological underpinnings of the subfield, but they either take CLS at its word that it does what it claims to do or they overlook CLS's argumentative arbitrariness. It is in fact true that data mining text labs are given institutional resources disproportionate to what they offer and how little computing power (excepting large-scale digitization efforts) their work actually requires. It only took one laptop to recreate almost all of the works here, and a single smart phone could have supplied the computing power, which begs the question of why we need "labs" or the exorbitant funding that CLS has garnered.[5] Nevertheless, because of the way it approaches textual analysis, CLS can use similar data-mining methods to back up very different positions and has already made a case for itself as something that offers new ways to catch inequalities and "read" corpora left out by the canon for reasons of access or judgments of aesthetics and value.

This essay does not argue that "numbers are neoliberal, unethical, inevitably assert objectivity, aim to eliminate all close reading from literary study, fail to represent time, and lead to loss of 'cultural authority' " or that

4. See Timothy Brennan, "The Digital-Humanities Bust," *The Chronicle of Higher Education*, 15 Oct. 2017, www.chronicle.com/article/The-Digital-Humanities-Bust/241424, and Danielle Allington, Sarah Brouillette, and David Golumbia, "Neoliberal Tools (and Archives): A Political History of Digital Humanities," *Los Angeles Review of Books*, 1 May 2016, lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/. Other critiques include Maurizio Ascari, "The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres," *Genre* 41 (Spring 2014): 1–19, and Daisy Hildyard's incisive description of the crudeness of interpretation in Jodie Archer and Matthew L. Jockers's *The Bestseller Code* (2016), in Daisy Hildyard, "Writing is Heavy Bombing," *Times Literary Supplement*, 8 Feb. 2017, www.the-tls.co.uk/articles/public/writing-is-heavy-bombing/. Kathryn Schulz's objection to counting words rings truest: "The trouble is that Moretti isn't studying a science. Literature is an artificial universe, and the written word, unlike the natural world, can't be counted on to obey a set of laws. Indeed, Moretti often mistakes metaphor for fact" (Kathryn Schulz, "What Is Distant Reading?" *New York Times*, 24 June 2011, www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html).

5. One CLS author has received $4,489,019 just in the past four years even though CLS predominantly uses databases, software, and algorithms that are either 100 percent free or free for anyone with a university affiliation (ECCO, NCCO, Archive.org, Chadwyck-Healy, Stanford CoreNLP, PC-ACE, Gephi, Word2Vec, TensorFlow, WORDHOARD, GraphML, Python & R, MALLET, and Microsoft Excel); see Piper's curriculum vitae at piperlab.mcgill.ca/pdfs/AndrewPiperCV2017.pdf

"numbers inevitably (flatten time/reduce reading to visualization/exclude subjectivity/fill in the blank)."[6] Nor does it make any claims about "the hegemony of data and data science," or the instability of the objectivity of data itself.[7] Others have already done this thoughtfully and eloquently. That human and literary phenomena are irreducible to numbers and that good interpretation and style in literary criticism are as objective as the sciences are personal convictions that do not enter into this critique. We can use nonideological reasoning to see that CLS as it currently exists has very little explanatory power that is not negated by its operations.

I discuss a handful of CLS arguments (chosen for their prominent placement, for their representativeness, and for the willingness of authors to share data and scripts or at least parts of them). Each of the papers I look at suffers from conceptual fallacies from a literary, historical, or cultural-critical perspective, but here I am taking them on their own terms completely—their sampling (too often the only point of contention from outsiders), their testing, their codes, and their truth claims. Drawing on basic statistical principles, I discuss these examples alongside text mining's known uses and applications and situations in which textual quantitative analysis and simplified reconfigurations of information would be useful. I will explain true applications in simple ways that do not do justice to their myriad complexities (mostly due to my own limitations) but that I believe still capture these applications' rightful functions and limitations. Critics in digital humanities have provided accompanying explanations for their methods but generally only to initiate more people into the subfield by making the entry bar seem low or so that their audience can follow along. I believe in reintroducing these methodologies in intuitive and efficient ways so that we can begin to understand the logics that drive them and better evaluate CLS's utility, discerning instances where tools and methods are used suboptimally or for no foreseeable reason. This essay is not an attempt to address all of the errors and oversights that occur in CLS work. Oversights in implementation; lack of robustness, precision, and recall; and less than ideal measurements are endemic to data-mining work. For these reasons, although I go over technical issues, the case against CLS won't rest on technicality, nor can one person take on that much work hunting down incom-

6. Ted Underwood, "It Looks Like You're Writing an Argument against Data in Literary Study . . . ," The Stone and the Shell, 21 Sept. 2017, tedunderwood.com/2017/09/21/it-looks-like-youre-writing-an-argument-against-data-in-literary-study/

7. Piper, "Why are Non-Data Driven Representations of Data-Driven Research in the Humanities So Bad?" .TXTLAB, 17 Sept. 2017, txtlab.org/2017/09/why-are-non-data-driven-representations-of-data-driven-research-in-the-humanities-so-bad/. For Piper, these constitute the "bad" examples of humanist objections to CLS work.

plete data work and debugging broken scripts. A clear explanation of the computational work that CLS actually does is enough to constitute a provocation to the rest of us to appreciate the circumstances under which such errors would be permissible and which not. The nature of my critique is very simple: the papers I study divide into no-result papers—those that haven't statistically shown us anything—and papers that do produce results but that are wrong. I discuss what it is about the nature of the data and the statistical tools that leads to such outcomes.

2

CLS papers are more or less all organized the same way, detecting patterns based in word count (or 1-, 2-, n-grams, a 1-gram defined as anything separated by two spaces) to make one of six arguments: (1) the aboutness of something; (2) how much or how little of something there is in a larger body, (3) the amount of influence that something has on something else; (4) the ability of something to be classified; (5) if genre is consistent or mixed; and (6) how something has changed or stayed the same. It will become clear that all six are basically the same argument, with aboutness, influence, relatedness, connectedness, generic coherence, and change over time all represented by the same things, which are basic measurements and statistical representations of overlapping vocabulary—and not even close to all of the vocabulary, as much culling must take place to have any kind of statistical workability at all. Data sets with high dimensionality are decompressed using various forms of scalar reduction (typically through word vectorization) whose results are plotted in charts, graphs, and maps using statistical software. Finally, this model (a newly derived instrument for measuring literary patterns or distinguishing between them) is tested using in- and subsample testing. The argument itself is usually the description of the results of data mining.[8] Quantitative analysis, in the strictest sense of that concept, is usually absent in this work. Hypothesis testing with the use of statistical tools to try to show causation (or at least idiosyncratic correlation) and the explanation of said causation/correlation through fundamental literary theoretical principles are usually absent as well.

8. This is the format laid out in Sarah Allison et al., "Quantitative Formalism: An Experiment," *n+1* 13 (Winter 2012), nplusonemag.com/issue-13/essays/quantitative-formalism-an-experiment/; hereafter abbreviated "QF." An essay by Alan Liu provides a good overview of the methods that are typically applied and the perceived rationale for their applications, updated to new statistical software for natural language processing or topological rendering; see Alan Liu, "From Reading to Social Computing," MLA Commons, dlsanthology.mla.hcommons.org/from-reading-to-social-computing/

No matter how fancy the statistical transformations, CLS papers make arguments based on the number of times *x* word or gram appears. CLS's processing and visualization of data *are not* interpretations and readings in their own right.[9] To believe that is to mistake basic data work that may or may not lead up to a good interpretation and the interpretive choices that *must be made* in any data work (or have no data work at all) for literary interpretation itself. In CLS data work there are decisions made about which words or punctuations to count and decisions made about how to represent those counts. That is all. The highest number of consecutive words (1-grams) that CLS work has looked at is three (trigrams). Mark Algee-Hewitt looks at the probabilities of bigrams (the likelihood that one word will be followed by another specific word) to calculate corpus "entropy," but this is just another way of saying "two words that appear together" (I will return to this paper later). Jean-Baptiste Michel and others' "Quantitative Analysis of Culture Using Millions of Digitized Books" tracks 5-grams (five 1-grams in a row), but their payoff is for lexicography and for tracking large-scale grammatical shifts, not for literary history or criticism.[10] Roberto Franzosi claims to find "narrative events" using trigram tagging.[11] Though it is already outdated in the field, his is the only case I know of natural-language-process tagging that tries to get beyond basic word frequencies. But there, "narrative events" are just 3-gram length subject+verb+object sequences, and accounting for "time" and "space" amounts to little more than known time markers and geographical locations (extremely hard from a coding perspective, reductive from a literary perspective).[12]

9. CLS has developed literary metaphors for what coding and statistics actually are and involve, turning elementary coding decisions and statistical mechanics into metanarratives about interpretive choice, as in, for instance, Clifford E. Wulfman's claims in "The Plot of the Plot" that raw data processing is like "a semiotic machine" or a "computational play of signification" (Clifford E. Wulfman, "The Plot of the Plot: Graphs and Visualizations," *Journal of Modern Periodical Studies* 5, no. 1 [2014]: 96).

10. See Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Books," *Science*, 14 Jan. 2011, science.sciencemag.org/content/331/6014/176

11. Franzosi, *Quantitative Narrative Analysis* (Los Angeles, 2010), p. 5. Franzosi's PC-ACE (Program for Computer-Assisted Coding of Events) is supposed to be user ready and yet, according to the website, you still have to manually develop a coding scheme with the assistance of PC-ACE on one body of texts to generate semantic triplets on another body of similar texts; see Robert Franzosi, "About PC-ACE," PC-ACE: Program for Computer-Assisted Coding of Events, pc-ace.com/about/.

12. This includes the fifty thousand articles on strikes in partisan publications in Italy or fifty-five years of news on lynchings in Georgia (1875–1930). Franzosi trains an SQL query to find the correct subject+verb+object from language that's not automatically ordered that way but not too far from that ordering; see Franzosi, Gianluca De Fazio, and Stefania Vicari, "Ways of Measuring Agency: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875–1930)," *Sociological Methodology* 42 (Nov. 2012): 42.

Despite claims to the contrary, CLS is not able to look at anything like plot beyond three words. And it's not just a matter of letting a nascent field mature (corpus analysis for literature has been around for half a century) but a matter of their objects being too few and too complex. Suggestions for quantifying literature from experimental early structuralism, such as Claude Lévi-Strauss's attempt to define the structure of myths using the formula $^{f}x(a) : {}^{f}y(b) \cong {}^{f}x(b) : {}^{f}(a-1)(y)$, are not operationalizable at all, as such patterns are too difficult and abstract to code and define far too few texts for machine learning to successfully code even one such appearance in a handful of texts.[13] Therefore all the things that appear in CLS—network analysis, digital mapping, linear and nonlinear regressions, topic modeling, topology, entropy—are just fancier ways of talking about word frequency changes. Breaking down CLS mistakes will clarify why, despite the fact that different semantic and syntactical tagging methods have existed since the 1970s, CLS tends to stick to counting words and is, in an even more limited sense, forced to find many of its significances by tweaking stop words.

3

The CLS papers I studied sort into two categories. The first are papers that present a statistical no-result finding as a finding; the second are papers that draw conclusions from its findings that are wrong.

I start with a paper that presents a no-result as a finding from using a measuring device too weak to capture a known difference, a paper that will also help us see the problem of measuring so-called homology, repetitiveness, or self-similarity through word frequency. Ted Underwood's "The Life Cycle of Genres," which tries to see if genres change over time, models the genre of detective fiction based only on word homogeneity and tests the accuracy of the model by seeing if it can distinguish *B* (post-1941 detective fiction) from *C* (a random motley of works) the same way it can distinguish *A* (pre-1941 detective fiction) from *C*.[14] Underwood compares *A* to *B* and claims that detective fiction is much more coherent over one hundred fifty years than literary scholars have claimed. Underwood wants to argue that genres do not change every generation and that they do not only consolidate in the twentieth century—as others, namely Franco Moretti, have

13. The formula stipulates that the analogy between function x of term a and function y of term b and remains true when the terms are inverted (the function x of a becomes the function x of b) and when the function and term value of one of the two things is inverted. This formula is from Claude Lévi-Strauss, "The Structural Study of Myth," *Journal of American Folklore* 68 (Oct.–Dec. 1955): 442.

14. Underwood, "The Life Cycle of Genres," *Journal of Cultural Analytics*, 23 May 2016, culturalanalytics.org/2016/05/the-life-cycles-of-genres/

argued—but have in fact been more or less consistent from the 1820s to the present. The problem here is that his model does nothing for his objective. Underwood should train his model on pre-1941 detective fiction (*A*) as compared to pre-1941 random stew and post-1941 detective fiction (*B*) as compared to post-1941 random stew, instead of one random stew for both, to rule out the possibility that the difference between *A* and *B* is not broadly descriptive of a larger trend (since all literature might be changed after 1941).[15] All that Underwood has shown in using word frequency homogeneity to differentiate detective fiction from random fiction is that the difference between pre- and post-1941 detective fiction *is not as significant as* its difference from random fiction. This does not mean that the same method can capture the difference between different types of detective fiction. After all, statistics automatically assumes that 95 percent of the time there is no difference and that only 5 percent of the time there is a difference. That is what it means to look for p-value less than 0.05. Think of it this way: if everyone can agree that something is changing—even Underwood concedes that genres evolve—but you have devised one way that concludes that it does not, it does not necessarily mean that you have found something. It just means your instrument of measurement might be too weak—your method might have too little power—to capture this kind of change.

The use of data mining to present naturally occurring statistical significances as results is a problem that can be seen in Matthew Jockers and Gabi Kirilloff's paper, "Understanding Gender and Character Agency in the Nineteenth-Century Novel," which claims that certain verbs are highly correlated with gendered pronouns (*he*/*him*, *she*/*her*) in the dataset.[16] (Gender is a preferred analytic in CLS, most likely because it's one of the few things that can offer a clean second-order classification—into male/female.) The authors use a parser to find accurate pronoun-verb pairs in their data and then build a classifier to predict the correct gender for given verb, claiming an 81 percent accuracy rate (30 percent improvement over pure chance). They find fifty verbs most correlated with males and fifty verbs

15. Whether homogeneity measurements are actually able to capture change or changelessness also depends on the study and whatever it's disproving sharing the same definition of change. Underwood uses Moretti's thesis (genres change) as a null hypothesis, but Moretti does not have the same definition of change as Underwood, so they're working from different premises; Underwood would have to build a model that tests for the same phenomenon of difference as the one Moretti describes using language to actually disprove his claim.

16. See Matthew Jockers and Gabi Kirilloff, "Understanding Gender and Character Agency in the Nineteenth-Century Novel," *Journal of Cultural Analytics*, 1 Dec. 2016, culturalanalytics.org/2016/12/understanding-gender-and-character-agency-in-the-19th-century-novel/; hereafter abbreviated "UG."

most correlated with females, with ten of these each that "the machine found most useful for differentiating between male and female pronouns" ("UG"). Putting aside the errors endemic to dependency parsing and OCR recognition, and the lack of accounting for association by negation (when a person *doesn't* do something), some of their results are obvious; some are not. As the authors themselves admit, this can make for a backward understanding of gender (it is binary; women *weep*, men *take*), but I'll leave this for others to discuss.[17]

First, there were always going to be top five, top ten, top fifty, top one hundred statistically significant pronoun-verb pairs. That is a function of finding all pronoun-verb pairs, ranking them by correlation, and cutting off the ranking where one chooses. In good statistical work, the burden to show difference within naturally occurring differences ("diff in diff") is extremely high. Let us say that you are measuring the overlap of features between two sets of data using a standard 5 percent confidence level; out of n possible shared features, 0.05n will automatically be significant. Data-mine something and you will always find significant associations.[18] Their claim that "there is a strong correlation between character gender and verbs in the nineteenth century" is true by fiat, as by their definition of correlation one could claim that about any group of literature in any century ("UG"). The paper does not perform a bootstrap, which means the literary-historical suggestions that follow this genre classification do not stand. But let's say they did. Just to find the top ten verbs for each gender, a far simpler method—a simple regression of pronoun-verb associations on an almost-identical corpus—regressing male percentage on female percentage of each verb—produces commensurate results. A good-faith replication using a commensurate parser delivers different results.[19] So what is the value added here? Their model has, in-sample, a 22 percent error rate when the true pronoun class was female and 16 percent when the true class was male. The authors account for high error rates by suggesting that gen-

17. Even the authors themselves admit that their findings "[correspond] with general notions of gender propriety . . . [because] verbs connoting emotion and sentiment (such as to cry, to love, to weep, etc.) were more strongly associated with female characters while verbs connoting action and motion were more strongly associated with male characters (to advance, to approach, to ride, etc.)" ("UG").

18. For this reason, practitioners have to apply the Bonferroni Correction to conventional statistical thresholds of significance used for data mining. If you have one sample and you have one test, then finding a t-value over 2 would be significant, but if you have multiple samples—as you do with text mining, which looks at lots of texts (samples)—then t-stats have to become more restrictive; the critical value of t has to go up. Standard statistical tests don't account for this because they tend to focus on one sample.

19. See section 1 of the online appendix.

dering of verbs might be less stable for females—but you cannot turn predictive weakness into an argument unless you can prove that your predictive ambiguity is not due to lack of power in your measurement. To extend their contribution by recasting it as a measurement of the gender rigidity of novelistic genres, Jockers and Kiriloff add that their model had 58 percent, 63 percent, and 67 percent accuracy rates for classifying the correct gender for their six bildungsroman novels, four silver-fork novels, and three historical novels; 80 percent for thirty-three gothic novels; and 100 percent accuracy rates for six industrial novels and two Newgate novels. There is no statistical rigor in this, never mind that we're talking about a very small number of books. Whatever sample size you start with you can always cut it in such a way so that you get something for which there's a 100 percent accuracy rate. By pure chance, there will be variation in accuracy rates; that does not mean there's systematic variation or a true pattern between genre and the model's gender-predicting powers.

Because of the way the data is treated, CLS can make macrohistorical claims that are statistically uninformative. Consider this graph, "A Network of Three Thousand Novels," which depicts similarity based on vocabulary and which Matthew Jockers argues reveals things about over three thousand novels over time (fig. 1).[20] According to Jockers, this network map, in which "books are being pulled together (and pushed apart) based on the similarity of their computed stylistic and thematic distances from each other" is "extraordinary" because it obeys chronology (clustering based on time written) and "chronological alignment reveals that thematic and stylistic change does occur over time. The themes that writers employ and the high-frequency function words they use to build the frameworks of their themes are nearly, but not always, tethered in time."[21] In other words, Jockers is arguing that because there is a separation between light dots and dark dots, because they're not all mixed together, and because the network visualization is itself agnostic on date of publication, he has proven that older works are more similar to one another and newer works are more similar to one another: that they reflect their times. Sampling errors notwithstanding, this network graph represents a tiny percentage of the data. What you learn from this 3 percent is tautological.[22] Jockers cal-

20. See Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana, Ill., 2013), p. 166.

21. Ibid., pp. 164–65.

22. "For computational convenience and for network simplicity," the author reduced the total number of unique observations between the books in his dataset to those within one standard deviation (ibid., p. 163). Deviating from conventional expectations, this cull resulted in only 3 percent of total edges, suggesting either a highly skewed distribution or something going awry in the execution. Because scripts were not shared, this is only speculative.

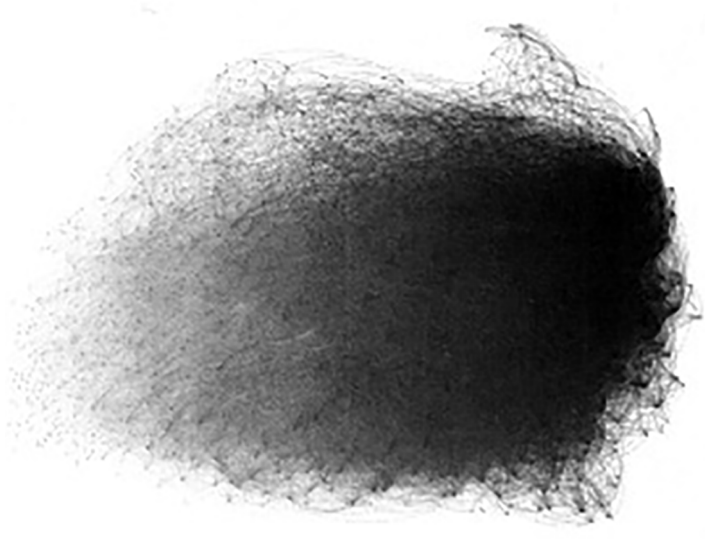

FIGURE 1. "Nineteenth-century Novel Network with Date Shading," from Matthew Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana, Ill., 2013), p. 165.

culates similarity between books (Euclidean distance) based on 578 features—five hundred are topics extracted from LDA topic modeling (more below), and the rest are frequently used words and punctuation. LDA topics and frequently used words tend to cluster in time, so these features already have time correlation built in. If you take a similar dataset (texts over one hundred years) and regress absolute Euclidean distances (based on similarly determined features) on absolute distances in time, you will see super significant positive correlation.[23] This is neither unique nor insightful; you've mechanically guaranteed the capture of a generic time trend—what is discussed over time plus language evolution.

Computational literary criticism is prone to fallacious overclaims or misinterpretations of statistical results because it often places itself in a position of making claims based purely on word frequencies without regard to position, syntax, context, and semantics. Word frequencies and the measurement of their differences over time or between works are asked to do an enormous amount of work, standing in for vastly different things.

Piper's essay "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel" offers a good example of this problem,

23. See section 2 of the online appendix for this correlation.

matching a word-frequency difference to structural difference in an argument that is historically and hermeneutically over-specific. "Novel Devotions" makes two claims: first, that the last three chapters of Augustine's *Confessions* are significantly different than the first ten and significantly different from each other.[24] In other words, things start to feel different after the tenth chapter and continue to feel increasingly more different as the book goes on. Piper attributes this to the experience of conversion that happens in book 10—an experience that he argues makes a real difference in terms of vocabulary output. This, he says, is what makes *Confessions* and books influenced by *Confessions* act on the reader in measurable ways, what makes them "devotional."[25] Second, Piper claims that English and German novels share the same structure as Augustine's *Confessions*; the second half of the text is radically different from the first half of the novel and increasingly different within its parts.[26] The amount of change that happens in word frequency (for each word) between first and second half, and between the chunks within the second half, is measured through cross-half and in-half scores, respectively, which are simply Euclidean measurements of the square root of the sum of the square of differences between terms' frequencies in text 1 versus text 2 (and up to text n). Piper derives an in-half and a cross-half score to capture this word frequency change and uses multidimensional scaling (MDS) to visualize the result, which essentially reduces a twenty-dimensional set of relationships down to two so it can be visualized (fig. 2).[27]

There are multiple things wrong with this study. Anyone who has read *Confessions* knows that the last three chapters differ from the first ten because Augustine turns to discussions of *Genesis* after spending ten chapters on autobiography, and so of course different words will start to show up.

24. See Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," *New Literary History* 46 (Winter 2015): 63–98.

25. See ibid.

26. In his computations, Piper doesn't divide the English and German novel (or Augustine's *Confessions*, for that matter) into first ten parts versus last three parts but rather uses an even split, dividing the works into twenty equal chunks and comparing first ten to next ten. His justification for this adjustment, which is problematic if you care about the moment of conversion wherever it happens in the text as the moment of change, is that he just wants to capture a basic difference in the second half, so there is no need for the specificity of last three and so on. In actuality, Piper had no choice but to do it this way. Technically, it would be inaccurate to derive in-half and cross-half scores from two texts of unequal lengths. For this reason, Piper's in-half and cross-half scores for are also based on, and not a 10/13 versus 11–13/13 split.

27. Piper divides the dataset into twenty sections of equal length (this is a 20×20 table, where each "document" is a section of the book, which was divided into twenty equal parts). This table is generated from the lexical features in the sections—in other words, a document-term matrix. He's asking: given a feature space of words for all documents, what are the similarities between each document compared with every other document?

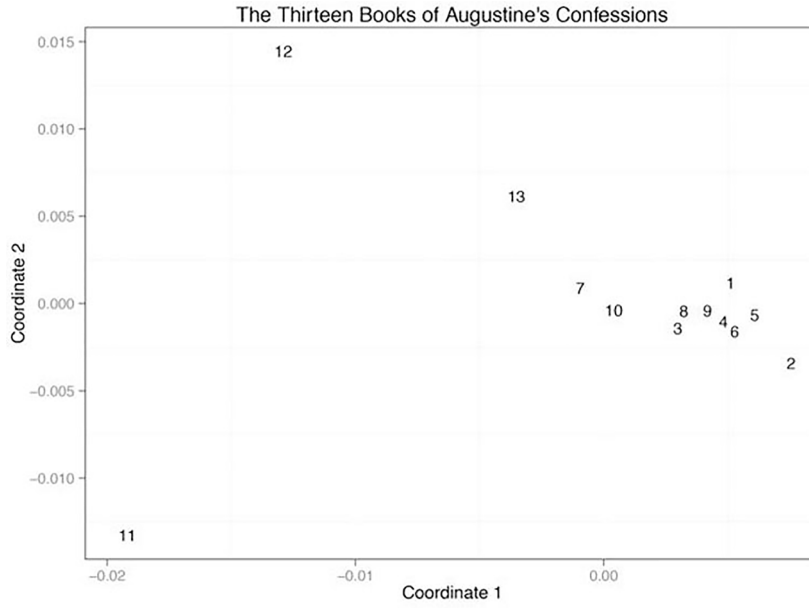

FIGURE 2. "The Thirteen Books of Augustine's Confessions," from Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," *New Literary History* 46 (Winter 2015): p. 72.

This has nothing intrinsically to do with conversion. His in-half and cross-half scores do not necessarily indicate this pattern of change and should not be taken as benchmarks for novels having such a "devotional" structure.[28] More technically: Piper did not stem the Latin text (pare words down to verb and noun roots) even though he stemmed the English and German texts.[29] He was counting conjugated verbs and declined nouns as different words in Latin but as the same words in English. Once the Latin text was stemmed and the distance matrices properly scaled for variables, we get scores that are different from his, and his results no longer obtain. I have recreated Piper's plot with a *stemmed* text, properly scaled (fig. 3). In my rendering, books 1 and 2 are not clustered with the other books in the first half, nor is book 13 as distant from the first half.

It is easy to see the problem with structuralist arguments that are at bottom tied to word frequency: word frequency differences show up in all

28. See section 3 of the online appendix for sample in-half and cross-half scores of other works and notes on scaling.

29. These stemmers are packaged in Python. The only Latin stemmer currently available is the Schinke Stemmer (C code not available in Python); see Martin Porter, "The Schnike Latin stemming algorithm," Snowball, snowball.tartarus.org/otherapps/schinke/intro.html

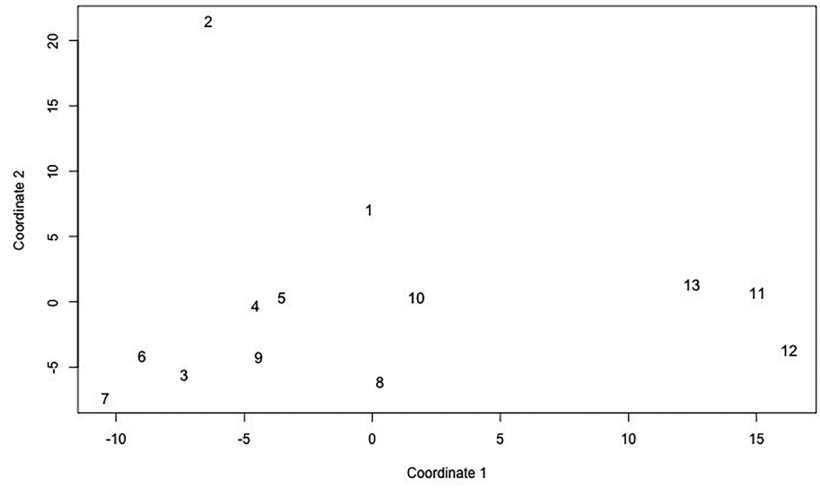

FIGURE 3. My corrected version. Each number here correlates to the same thing as those in figure 2: a one-twentieth chunk of the entire text.

kinds of texts and situations that do not match what you want them to represent. Piper cannot stop the second half of texts being quantitatively different from the first half where he does not want them. To define word frequency changes as *change* itself (and by a conceptual slide, conversion) is both tautological and risky. There is no reason to mystify the process; as more concepts are introduced into a text, more words come with them. An MDS for Exodus, for example, demonstrates this (fig. 4). The Exodus plot shows a spread similar to what Piper finds in Augustine's *Confessions*, with the first half closer together and the second half not only farther than the first but with data points farther from each other. Unless Piper is prepared to argue that the Hebrew Bible also follows Augustine's confessional structure (as he defines it), he has to admit that such a pattern is not limited to *Confessions*. Of course, this may be confusing necessary with sufficient conditions—the fact that Christian conversion narratives exhibit this phenomenon does not mean that nonconversion narratives do not. An effective argument of this sort about religious texts would require further evidence and commentary. In the meantime, a Chinese translation of Augustine's *Confessions*, for example, produces an MDS (using Piper's methods) that does not look at all like his graph for the Latin *Confessions* (fig. 5). Do conversion experiences not carry through translation?[30]

30. I used character 字 frequencies in accordance with Paul Vierthaler's assumption that while Chinese words are irreducible to characters, measuring 1-grams contain capture a sufficient amount of linguistic meaning, and also in accordance with CLS's claim that stemming or

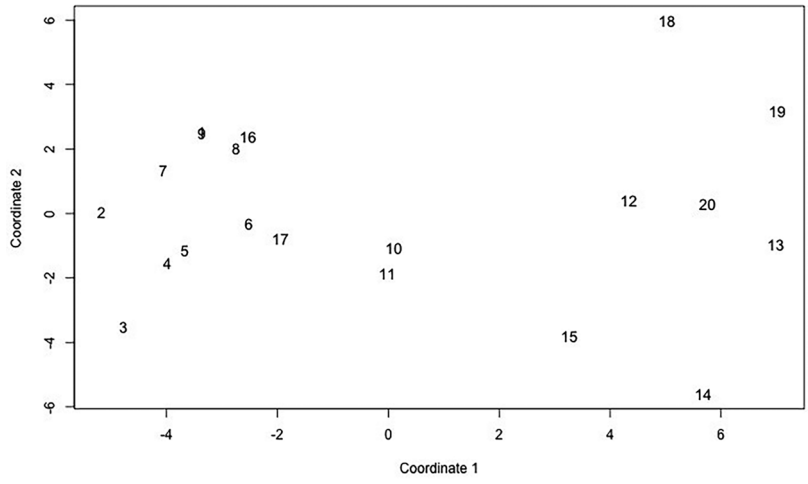

FIGURE 4. MDS of English translation of *Exodus*, each number representing a one-twentieth chunk of the book. The first ten chunks are clustered together and the last ten are farther away and farther apart from one another, just as in Piper's *Confessions* MDS.

Reducing similarities and differences to word frequency differences forces one to produce findings when there might be an underlying explanation that both obviates your claims and obviates the need for your modeling. A ready example of this problem can be found in Paul Vierthaler's work on the difference between different types of Chinese writing.[31] The author claims that two genres of Chinese writing, historical apocrypha (野史) and fiction or narratives (小说), are not as similar as literary historians have assumed. He looks at three very small corpora (14, 126, and 524 texts) and compares their word frequencies (1-gram 字 frequencies) using a hierarchical clustering algorithm (HCA) that makes dendrograms based on "similarity scores" and PCA. Because he split each book into ten-thousand-character chunks and then took the one thousand most-used Chinese characters in that chunk (determined through simple term frequency),[32] each dot on his PCA represents a ten-thousand-character sec-

not stemming does not make very much difference to outcome; see Paul Vierthaler, "Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature," *Journal of Cultural Analytics*, 23 May 2016, culturalanalytics.org/2016/05/fiction-and-history-polarity-and-stylistic-gradience-in-late-imperial-chinese-literature/

31. See ibid.

32. Already, Chinese language readers will challenge this study, as one cannot look at *zi* 字 as contained units of meaning. The author admits that determining which are the *ci* 词 (words) in a Chinese document is difficult since there are insurmountable parsing issues (a *zi* 字 plus another zi 字 often produce another word entirely) and punctuating and sentence-separating 断句 issues (classical Chinese often appears without punctuation and so semantics and grammar

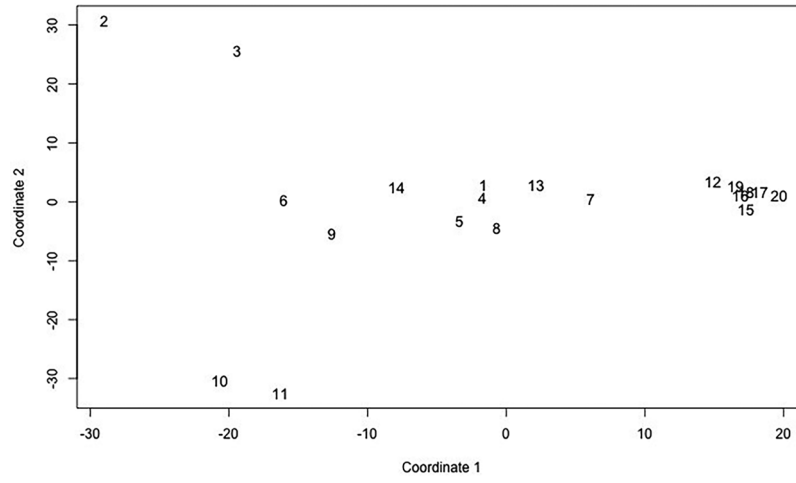

FIGURE 5. Chinese translation of *Confessions*, each number representing a one-twentieth chunk of the book. The first ten chunks are not close together nor are the last ten far apart from one another.

tion, not a whole book (fig. 6). And in comparing the one thousand most common 字 in ten thousand 字 segments, the author has already made data points that will look extremely similar and allowed the PCA to look more robust than it is.[33] In other words, the author has already homogenized the data points and needlessly increased their number. Therefore, the number of data points on the PCAs seem to make a strong case, but in fact the data points from each kind of genre are very close to one another simply because of the way the author prepared his data. More pressingly, Vierthaler uses computational methods to prove to us that Chinese historical apocrypha is in fact closer to official history because of similar formal language use. This assertion is based on common tokens that clearly divide between classical language and vernacular, but he describes a bridge between official history and fiction based on common tokens having to do with theme and plot. This relationship is already known to readers of classical Chinese literature. Historical apocrypha and official histories of the Ming and Qing period were predominantly written by a similar set of literati or functionaries. Apocrypha are differentiated by content, not formal lan-

must be inferred from context) that counting 1-grams (字)s is extremely inaccurate. Still, Vierthaler insists that 1-grams (字) frequencies are still meaningful and predictive; see ibid.

33. For the Chinese language especially, if we extract the one thousand most commonly used characters, we have already secured a degree of similitude among each one-thousand-character chunk.

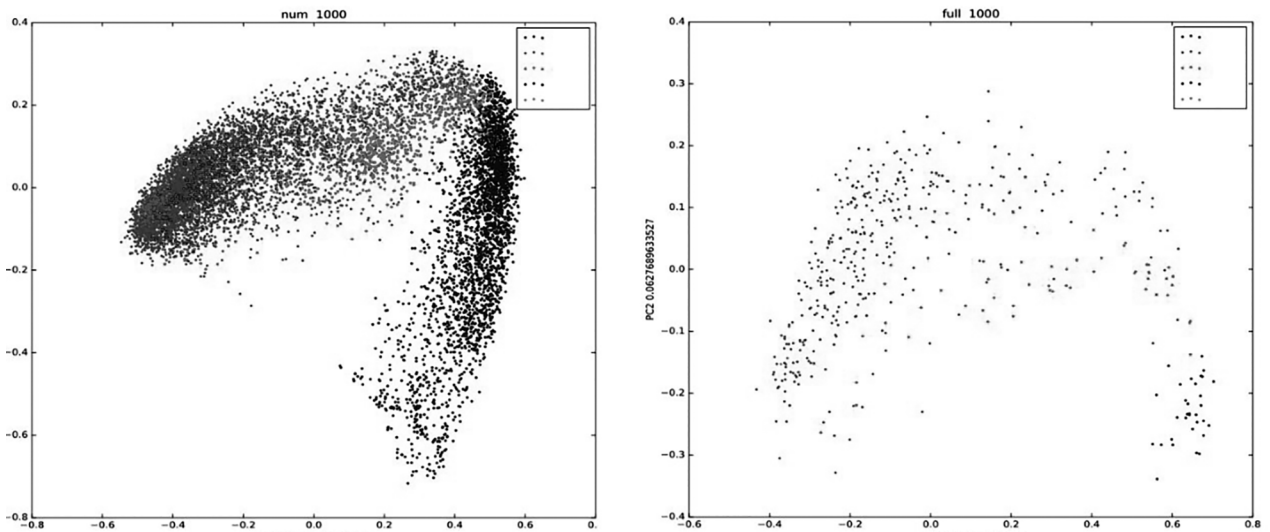

FIGURE 6. The same MDS using the full texts, without splitting into one-thousand-character blocks, still using term frequency, and yielded a similar spread but with much fewer data points. The three apocryphal texts are in gray in the upper right corner.

guage use, while fiction and narratives are primarily written in the vernacular (or an admixture that leans toward vernacular) and contain themes shared with historical apocrypha. As Vierthaler writes, if historical apocrypha and fiction or narrative have been traditionally grouped together, it is because both tend to be collected from hearsay. It is redundant to challenge this classification based on criteria that the classification was never confused about in the first place.

Hoyt Long and Richard Jean So's "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning" sets out to measure formal influence from East to West by building a Naïve-Bayes classifier to find haikus that are not self-identified as haikus—in part offering a classificatory tool and in part tracking English poetry not expressly designated as haikus.[34] They train their classifier on four hundred haikus (in translation and as adaptations) and one thousand nine hundred nonhaiku short poems and then run it on an unclassified combined set. The Bayes rule is a widely used rule that with every new observation updates the probability distribution; the system is "naïve" because the features are supposed to be independent of one another.[35] You do not tell the algorithm the exact criteria by which to make its classificatory decisions; you tell it what to pay attention to and it learns the decision rule based on some basic features,

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

35. The presence of the Bayes formula in CLS work does not mean that a paper has captured a Bayesian phenomenon, nor does a Bayesian phenomenon mean literary complexity. CLS papers use the Bayes rule to refine their classifications, as the Bayes rule works based on regression of conditional expectation. In practical applications the Bayes rule can update probabilities as new conditional expectations come in, allowing us to calibrate actionable versus nonactionable information in real time. In CLS its applications are mostly to incrementally refine its measurements. In the methodology paper "A Bayesian Mixed Effects Model of Literary Character" David Bamman, Underwood, and Noah Smith seek to "learn" probabilistically the different kinds of character types in about fifteen thousand eighteenth- and nineteenth-century English novels. Before, the words that are associated with a character may have been calculated based on their proximity to one another; with their modeling improvements, the words that are associated with a "persona" can now be sorted out from the influences of author, period, and genre continuously with new information. Using a hierarchical Bayesian approach just means that finding a Mr.-Darcy-like character now means finding a group of words whose distribution over the role types chosen, updated with each different input, is closest in distance to Darcy's. This paper's contribution to literary criticism is that it marginally improved correct identification of group of words called "personae" with respect to the confounding factor of authorship, resulting in the discovery of such personae as "king emperor throne general officer guard soldier knight hero" and "beautiful fair fine good kind ill dead living died" (David Bamman, Underwood, and Noah A. Smith, "A Bayesian Mixed Effects Model of Literary Character," *Proceedings* of the 52nd Annual Meeting of the Association for Computational Linguistics [Baltimore, Maryl., 23–25 June 2014], acl2014.org/acl2014/P14-1/pdf/P14-1035.pdf, p. 378).

changing the probability distribution every time a new thing comes in and so getting smarter and better at classifying the next thing. Technically, Long and So use Naïve-Bayes (N-B) to refine their classifier, treating each poem in the test sample as a new observation. But instead of letting the N-B figure out the cutoff syllable count, the authors hard code that decision rule into their script (whether a poem is under nineteen syllables if it's a translation, or under thirty syllables if it's an adaptation).[36] The only other thing that their classifier uses to classify haikus is a simple likelihood of appearance score for individual words (the word *sky* becomes 5.7 times more likely to show up in a nonhaiku, for example). They end up with a model that is overfitted and for which features are learned very quickly. I ran their N-B classifier on English translations of Chinese couplets that are similarly long and filled with similar imagery as well as two hundred English translations of short Chinese and nonhaiku Japanese poems from *Wakan Rōei Shū* (Collection of Japanese and Chinese Poems to Sing) from the tenth century (predating the consolidation of the haiku form by almost seven hundred years). Their classifier heavily misclassified the Chinese and prehaiku poems because of the primitiveness of its criteria;[37] in fact, as the reduction threshold increases (removing features that don't occur often enough to prevent overfitting) the accuracy declines even more. It turns out, in other words, if you define haikus as poems under thirty syllables with words that show up a lot in haikus, you will in effect collapse the diversity of many types of East Asian poems into the haiku form.

The power of a statistical test comes from having meaning and setting up a null/alternative hypothesis that's informative and that explains something with respect to fundamental insights. It is not enough to find a pattern in the data that rejects a poorly chosen null such as "most frequently used words don't change" / "most frequently used words do change." It may be an extremely rigorous test, but it tests the wrong problem. All it accomplishes is the data mining of results. Researchers in the sciences and social sciences are extremely wary of such results. Statistical tools are designed to do certain things and solve specific problems; they have a spe-

36. Let's say that on average three hundred characters constitute roughly fifty-five words and that fifty-five words yield on average one hundred ten syllables. If one of the authors' parameters for defining the English haiku is syllable count (above or below nineteen syllables for translations, thirty syllables for adaptations), that means that most of the data in their control group is eliminated off the bat (being close to one hundred syllables over). To have a more honest approach, they need to cut off the poems in the control group at much lower character counts.

37. See section 4 of the online appendix for misclassifications.

cific utility and should not be used just for the sake of dressing up word counts. This is not at all to argue that literary analysis must have utility—in fact I believe otherwise—but if we are employing tools whose express purpose is functional rather than metaphorical, then we must use them in accordance with their true functions.

The reasons for quantifying narrative text, running algorithms based on word frequencies, and topographically visualizing textual data are not very transferable to our discipline. Typical applications of textual data mining involve a trade-off: speed for accuracy, coverage for nuance. Such methods are efficient for industries, sectors, and disciplines that are dealing with so much textual data at such fast speeds that they cannot possibly (nor would want to) read it all or where one wants to extract from a large data set a relatively simple piece of information that is either actionable or that can be quickly labelled and classified along simple features. Whatever one's feelings are towards deterministic algorithmic approaches to worldly phenomena, text mining is ethically neutral. In legal discovery, large volumes of largely identical legal documents (such as contracts) can be machine read for errant phrasing or diction (including misapplied terms of art) amidst standard terminology and formally repetitive syntactic patterns to quickly identify problematic or intentionally misleading clauses. The information that is extracted is not supposed to be semantically complicated. Investors use text mining to determine if a news report or press release by a company has a negative or positive tone so that a trading decision can be made very quickly. Every second, news is released by companies—annual reports, quarterly reports, stock earnings announcements, and so on—that no one wants to read; nor could anyone possibly have the time to read it all. Simple measures of terms that drive certain measurable changes are what one can and would wish to glean from these modes of inquiry; speed is the most important consideration because the corresponding decision often has to happen within seconds if not nanoseconds. We could theoretically verify each report individually—text mining knows that human reading can catch more nuances, exceptions, ambiguities, and qualifications—but why would we? Your email server uses a machine-learning classifier trained on a mix of all the previous emails marked as spam and nonspam by the user to decide if an incoming document is spam. An email might get sorted into the wrong folder or flagged as important for no good reason, but the classifier works instantaneously and is accurate enough that you wouldn't prefer to do it yourself.

To look for homologies in literature, CLS must eliminate much of high-dimensional data and determine the top drivers of statistically significant

variation. This always involves a significant loss of information; the question is whether that loss of information matters. One popular way to decompress high-dimensional data is factorization, a way of parsimoniously explaining lots of variance in numerical data. Take for example tools like principal component analysis (PCA) or multidimensional scaling (MDS) that were used in Piper and Vierthaler's paper and that are used widely in CLS to capture morphology and represent quantitative findings. PCA performs an orthogonal transformation of data, reducing the total number of aspects in multivariate data without knowing exactly what commonalities and differences to look for in the first place. PCA will sort multivariate data into principle components and provide quantitative descriptions of differences between data entities based on their loading on shared vectors. If you have three hundred thousand metric profiles of multivariate data (for example, patients exhibiting one or more illnesses and their chromosome maps) and wish to know what they might have in common—but not *everything* they have might in common, just three or four things, and again without knowing what those could even be—PCA will help you sort the data by these *principal* components. It does not tell you what to call these categories or what themes they share, descriptively, but tells you what characteristics (different chromosome maps) might be driving a clustering (patients who all have heart disease). In textual analysis, this means that the greatest difference between one article, piece of literature, or book and another would be their loading on a few of the shared vectors—information that is given quantitatively, not descriptively. You wouldn't want to go all the way to all of the vectors because that would simply be a reproduction of the entire data set (where you stop is a professional choice); therefore, it is necessarily a significant reduction of information. It is one thing to statistically identify the shared drivers of a medical illness and another to say that the difference between Immanuel Kant's third critique and G. W. F. Hegel's *Lectures on Aesthetics* can be captured in two or two numbers derived from their overlap on two or three vocabulary lists. There are many different ways of extracting factors and loads of new techniques for odd data sets, but these are atheoretical approaches, meaning, strictly, that you can't use them with the hope that they will work magic for you in producing interpretations that are *intentional*, that have meaning and insight defined with respect to the given field.

Consider this plot by the Stanford Literary Lab (and originally produced by Michael Witmore and Jonathan Hope) that argues that perhaps "narrative genres can be reduced to two basic variables" and that perhaps something besides genre drives the differences among William Shakespeare's

comedies, tragedies, histories, and late plays ("QF") (fig. 7).[38] No one has ever said, though, that consistent word frequency is what distinguishes Shakespeare's comedies from tragedies, tragedies from histories, and so on—and no one would ever say that because such distinctions cannot be captured with word frequencies. Put another way, the only way that this PCA diagram would be interesting is if word frequencies *were* recognized as what actually drove the genre differences. That is, if the first and second principal components precisely identified the tragedy and comedy *factors*. Again, this would be highly unlikely but statistically sound. Hypothetically, researchers could have complied all of the works from each category into one vector so that there are only four data points in the PCA, one for each genre. Then they could go in and look at the vector of word frequencies to see which words are driving the differences. That would actually teach us something, even if it would still be reductive as literary criticism. (In fact, it is good practice to ask users of CLS to show their vectors—it demystifies much of the process and often reveals conceptual weaknesses.) The authors of "Quantitative Formalism" did try to do that, generating multiple PCAs only to find repeatedly that PCA cannot capture generic differences. They then looked at the Docuscope scatterplot to see which component loadings (words) were mostly driving the differences and found mostly stop words; they then presented this phenomenon as a literary-critical argument: "Do you want to write a story where each and every room may be full of surprises? Then locative prepositions, articles and verbs in the past tense are bound to follow" ("QF"). Whether we find this reasoning sound or not, it is not a revelation but rather an attempt to make something meaningful out of a stop word problem.

The hitch of using textual pattern mining for forensic stylometry is that even if you apply pattern recognition techniques that reduce noise and nonlinear interactions between data, the stylistic differences that can be captured for literature tend to be driven by stop words—*if, but, and, the, of*.[39]

38. See Michael Witmore and Jonathan Hope, "Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays," in *Early Modern Tragicomedy*, ed. Subha Mukherji and Raphael Lyne (Rochester, N.Y., 2007), pp. 133–53.

39. The most famous example of CLS forensic stylometry—the use of statistical text mining to differentiate style, genre, and authorship—is the argument by Hugh Craig and Arthur F. Kinney that the late works by Shakespeare were written by Christopher Marlowe, even though Marlowe died more than a decade before Shakespeare's last known work (in accordance with the Marlovian Theory, which argues that Marlowe faked his death in 1593 and continued to write in Shakespeare's name); see *Shakespeare, Computers, and the Mystery of Authorship*, ed. Hugh Craig and Arthur F. Kinney (New York, 2009), and Neal Fox, Omran Ehmoda, and Eugene Charniak, "Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate" (MA thesis, Brown University, Providence, R.I., 2012), cs.brown.edu/research/pubs/theses/masters/2012/ehmoda.pdf

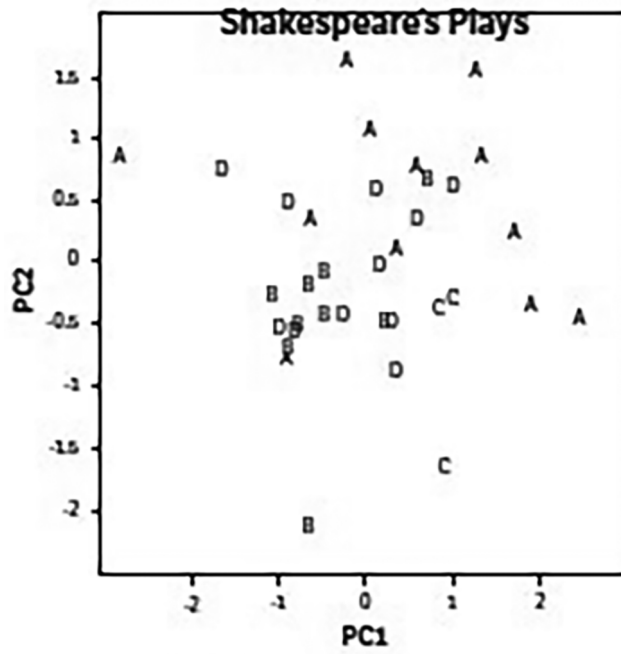

FIGURE 7. Hope and Witmore's PCA of Shakespeare's plays, classed by genre; see Sarah Allison et al., "Quantitative Formalism: an Experiment," *Stanford Literary Lab*, Pamphlet 1, 15 Jan. 2011.

Why is that the case? Mark Algee-Hewitt and Piper tell us that "stop words are usually semantically poor and yet stylistically rich. . . . The best means so far for determining authorship attribution and classifying texts as categorically different."[40] In reality, stylistic differences boiling down to stop words is not surprising at all. To locate a statistical difference of occurrence means having enough things to compare in the first place. If the word *cake* only occurs once in one text and four times in another, there's no way to really compare them, statistically. By the numbers, stop words are the words that texts have most in common with one another, which is why their differentiated patterns of use will yield the readiest statistical differences and why they have to be removed for text mining.

The stop word dilemma—keep them and they produce the only statistical significance you have; remove them and you have no real results—can be seen in Long and So's "Turbulent Flow: A Computational Model of

40. Piper and Mark Algee-Hewitt, "The Werther Effect I: Goethe, Objecthood, and the Handling of Knowledge," in *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ed. Matt Erlin and Lynne Tatlock (Rochester, N.Y., 2014), p. 158.

World Literature," a paper that tries to come up with a predictive algorithm for the literary phenomenon stream of consciousness (SOC). The paper argues that SOC travels amongst countries and that this "diffusion" can be tracked.[41] Long and So compare three hundred twelve-hundred-character SOC passages based on what other scholars have said were SOC passages and repeated for sixty realist novels (since realist novels are often seen as not having or using SOC), to build a classifier testing for thirteen linguistic traits unique to SOC (token-type ratio, onomatopoeia, neologisms, noun-ending sentences). They argue that they can predict a piece of SOC literature with 95 percent accuracy (97 percent accuracy for Japanese literature). Of the thirteen features tested for, authors identify token-type ratio (the number of words divided by number of types of words in a sentence) to be the single most important factor in predicting SOC; this is a concept that critics had already claimed in the 1970s but "never with such precision or on such a scale."[42] When Long and So's classifier is less accurate in dealing with SOC in Japanese literature, the authors call this "turbulent flow"—when formal influence doesn't carry all the way through.

However, their strongest predictor for if something is SOC or realism—token-type ratio—is too sensitive to the nonstandard stop words that the authors chose themselves. If you do not remove the stop words, then the statistical significance flips the other way (realist texts have higher token-type ratios). Removing the stop words flips the equation because the ratio between distinct stop words to total words in passage for SOC is statistically higher. And this is because SOC stop words are similar, while realist stop words are more varied if we are using the stop words that the authors chose themselves (the list of which, even with proper names removed, was three hundred words longer than the standard stop word list).[43] Using this list, realist texts will have significantly more stop words than SOC texts. This explains why the removal of stop words changes the token-type ratio enough to make SOC's token-type ratio statistically higher than realism's. Thus, the only thing the authors needed to do to differentiate SOC texts from realist was to tabulate stop word frequencies—their strongest indicator above any of the four they isolated; their strongest explanatory feature, in other words, is an unnecessary measurement. I reran their code using a standard stop word list, and once we only remove the standard stop words, the difference

41. Long and So, "Turbulent Flow: A Computational Model of World Literature," *Modern Language Quarterly* 77 (Sept. 2016): 345.

42. Ibid., p. 350.

43. See section 5 of the online appendix for their stop words and t-test using standard stop words.

between token-type ratios for realist texts and SOC texts loses statistical significance.

In other sectors and applications, texts with stop words removed can further be categorized—into *economic terms*, *political terms*, *female consumer*, for example. Another level of *simple and accurate-enough* classification has to occur so that categories can be compared rather than an individual word's frequencies—this is what allows for the statistical analysis of words. When CLS tries to do this for literature, using various methods to reduce large corpora of words to sensible groupings, it realizes that after the necessary dimensionality reduction is performed—uncommon words taken out, stop words removed, groups of words vectorized to become single points in space—it's left with only a small portion of what it was originally purporting to study, and these are corralled into groupings so general as to preclude meaningful interpretations.

4

To deal with secondary classification problems, CLS often use topological data analysis (TDA) tools, network analysis tools, and topic modeling tools like latent dirichlet allocation (LDA) and latent semantic analysis (LSA). This represents one of the most questionable uses of statistical tools in CLS. Topic modeling, which treats each text as a distribution over topics and each topic as a distribution over words (thus still dealing with texts as an unordered collection of words), is used to discover topics unsupervised in large bodies of texts. It is extremely sensitive to parametricization, prone to overfitting, and is fairly unstable as an "aboutness" finder for sophisticated texts because you need only tweak small details to discover completely different topics. There is no real way to measure the accuracy of the topics found since LDA's recall depends on having true classes of topics arrived at through human decision making. Its utility is most observable in circumstances where recall and precision really do not matter very much, as with content-based recommendation systems (such as Facebook advertising products to its users).

Without meaningful applications, topic modeling will look like a word-cloud generator for literary criticism. Jockers and David Mimno use LDA to extract themes from the Literary Lab Corpus and find that women writers are twice as likely to focus on female fashion (one word cloud for female fashion) and male writers are much more likely to focus on the topic of enemies (another word cloud of war-related terms).[44] In contrast, Under-

44. See Jockers and David Mimno, "Significant Themes in Nineteenth-Century Literature," *Poetics* 41 (Dec. 2013): 755, 759.

wood argues that topic modeling is only useful for literary studies if it can find "meaningfully ambiguous" clusters instead of "intuitive" ones—"ones that are clearly about war, or seafaring, or trade." But that would mean banking on instances where topic modeling isn't working as it should.[45] The truth is that "meaningfully ambiguous" clusters where unexpected words gather either turn out to have rather mundane explanations or to simply replicate the order of appearance of actual words in the works. In that same article Jockers and Mimno try to extend the uses of topic modeling to find writers who hide political information in religious topics—words clustering around "Convents and Abbeys"—only to find that two texts in the anonymous corpus drive most of the content of the topic of abbeys and convents.[46] This is simply because the phenomenon of displacement—when talking about cats is actually talking about one's mother—is not what topic modeling, which is based on probabilistic models of likely coappearance, is designed for. Underwood discovers topic 22 in women's poetry from 1815–1835, but because it does not intuitively cohere—reading like a poem assembled from the most frequently occurring words in a poetry corpus—interpreting it is nonsensical, which is why interpretation is missing from his presentation of the possibilities of topic modeling.[47]

Topic modeling has also been used in a new genre of academic surveillance in which academics catch each other out by interrogating the things they have been covering. Ethical considerations aside, there is the question of whether such models can even effectively determine areas of study. Underwood and Andrew Goldstone's survey, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us" looks for what scholars have been "talking about" in about thirteen thousand scholarly articles from 1889 to 2016 and finds that many topics are becoming more prevalent (fig. 8).[48] For instance, they find that increase in topic 80—ten words that cluster around the word "power"—is a "trend specific to lit-

45. Underwood, "Topic Modeling Made Just Simple Enough," The Stone and the Shell, 7 Apr. 2012, tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

46. Jockers and Mimno, "Significant Themes in Nineteenth-Century Literature," p. 763.

47. Topic 22 consists of "thy, where, over, still, when, oh, deep, bright, wild, eye, yet, light, tis, whose, brow, each, round, through, many, dark, wave, beneath, twas, around, hour, like, while, away, thine, those page, hath, lone, sky, spirit, song, oft, notes, home, mid, grave, vaine, again, though, far, mountain, shore, soul, ocean, and night" (Underwood, "Topic Modeling Made Just Simple Enough").

48. See Andrew Goldstone and Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45 (Summer 2014): 363. They also find decade-specific increases in the presence of other topics related to theory and politics such as "*new cultural culture theory*," "*social work form own*," "see new media information," and "*reading text reader read*" (pp. 370, 376, 377, 370).

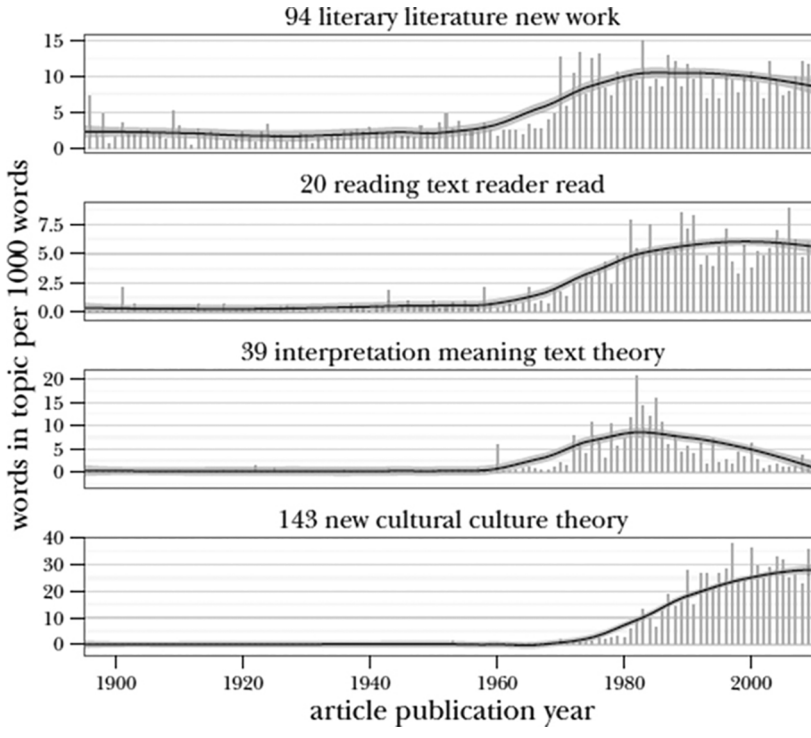

FIGURE 8. Topic-year distributions from Underwood and Goldman's "The Quiet Transformations."

erary study" that peaks in the 1980s.[49] If the authors wanted to nonarbitrarily study the change in topics covered in journal articles over time, they could have saved time by just looking at journal abstracts. Treating all the articles published in a year as a single sample (and not separating their data set into training and test sets) and running an LDA without fitting a prior to a posterior means that the algorithm will tend to form topics based on consecutive years in the corpus. They want to argue that some topics are increasing while others are decreasing, but conducting the topic modeling this way will *mechanically* produce topics that increase and decrease over time.

As a literature of scholarship grows, there will be more literature. The topics (a cooccurrence of words) that are found are driven by more recent scholarship because there's more of it; therefore, back mining the earlier scholarship for the topic will obviously show the topic to have increased over time. The authors find it counterintuitive that topic 80 has risen over

49. Ibid., p. 363.

time while the individual words have not (using Google n-gram) but if topic 80 exists over the whole period but is primarily driven by the latter period of scholarship, then *definitionally* the words in topic 80 do occur but do not co-move in the earlier period.[50] In the presentation of their results the authors eventually perform year-topic scaling for the topics they found, but that doesn't change the fact that they still found those topics in the first place still using the full sample. Ideally, a study either chooses a list of reasonable words to associate with a topic beforehand and looks just for those words in the full sample as a trend, or the study down weighs more recent articles to avoid the clustering effect. If using the full sample to find topics, as Underwood and Goldstone have done, an author cannot make arguments about time-series variation.

When topic modeling is used in a reasonably correct fashion, one can identify interesting and unexpected topics *only when* the other topics that are found (say, forty-seven out of fifty topics) pass the smell test.[51] This is not the case for this study; basic robustness tests also fail. To see how article length might influence topics found, I performed two robustness tests. In a partial double test (which randomly doubles the lengths of 30 percent of documents, all other parameters staying the same, and which should not affect the LDA because it's based on a *bag-of-words* model), all the topics changed. When I randomly removed just 1 percent of the original sample, all the topics changed. This paper also does not pass the test of reproducibility; if the method were effective, someone with comparable training should be able to use the same parameters to get basically the same results without swimming through tailored codes and buried filters. I took their dataset and used a Python LDA script (scaled for each document's length) to find one hundred fifty topics of ten words each, exactly as they did.[52] The

50. Using Google n-gram, which consists of an entirely different sample set, to make claims about topic 80 is just wrong. While the individual words in topic 80 are not rising in Google n-gram over the same time period, they are very much in the original data set.

51. On how to properly conduct LDA given its inherent shortcomings, see David Hall, Daniel Jurafsky, and Christopher D. Manning, "Studying the History of Ideas Using Topic Models," *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hi., 2008), nlp.stanford.edu/pubs/hall-emnlp08.pdf, pp. 363–73, and David M. Blei and John D. Lafferty, "Dynamic Topic Models," *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Penn., 2006), mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf, pp. 113–20.

52. See section 6 of the online appendix for the topics these tests produced versus mine. Because the authors provided scripts with missing functions and poor documentation, I had to rerun their results using a method as close to theirs as possible, using their extremely odd 6,970-word stop word list and their frequency cutoff points (take out words occurring less than two times and not in the top one hundred thousand used words).

topics I generated were not at all the same.[53] This does not mean that one of us did not do due diligence, but it does mean that topic modeling is like a kaleidoscope that turns out something entirely different with the slightest tweaking.

5

These days there is no shortage of fancy statistical tools to aid in machine learning. Computation is relatively easy and cheap; tools exist that allow you to run every pathway, make a decision at every step along the way, and provide many ways to tweak the modeling to identify patterns. In the end, statistics is about identifying a higher-order structure in quantifiable data; if the structure is not there (or is ontologically different) statistical tools cannot magically produce one. For instance, while text mining often uses topology, it is meaningless if it does not retain topology's function, which is a meaningful reduction of complexity to make quicker, more intuitive, noncontingent calculations. In a math problem foundational to graph theory, "The Seven Bridges of Konigsberg," one has to determine if a path exists that crosses only once each of the bridges in a particular configuration of rivers and land masses (fig. 9). You could do it manually, but this becomes too arduous if we are dealing with larger areas with many more intersections, bridges, and odd-shaped land masses, or indeed a whole municipality. Leonhard Euler proposed a scalar reduction in complexity, reformulating each land mass as a node (the blue dots on the third image), and each crossing to another land mass as an edge, producing a graph that only records nodes and edges. This graph is not a formal rearrangement of the map but a fundamental transformation of the information in the map. It no longer matters how the river runs, how big or what shape the islands are, and how they lay with respect to one another (these are local). You can take any area and count the number of land masses and their exit points. If zero or two of the nodes have odd number of edges, then such a walk is possible. If not, then not (so in the original problem the walk is not possible).

A reduction in complexity is necessary in this case because you don't want to have to exhaust all the combinations of routes in order to know the answer for urban planning. Topology, which grew from this problem, relies on a reduction of complexity from actual layout to schematic representation, preserving the relationship between two points under their continuous deformations. Topological maps such as a subway map transform

53. For their list, see Underwood and Goldstone, "List of Stop Words Used in Topic Modeling Journals, Summer 2013," IDEALS, www.ideals.illinois.edu/handle/2142/45709

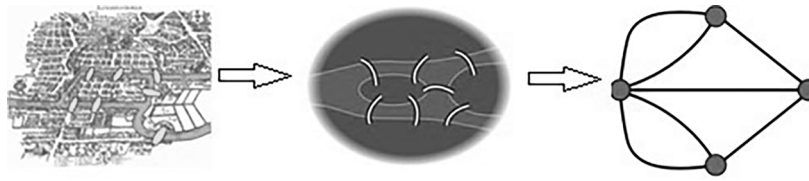

FIGURE 9. "The Seven Bridges of Konigsberg": a topological transformation.

complicated and contingent geospatial information into essential nodes (the map doesn't have to reflect the myriad topographical details in an actual map or even resemble it in any way by being *to scale*—the only things that matter are the points of interchange). These examples illustrate the criterion by which to judge the usefulness of a topological transformation.

CLS understands the topological terms *global* and *local* in ways that are no longer imbued with graphic theoretical meaning—network diagramming and topology function interchangeably in its practices—and tends to reconfigure information for the purpose of visualizing lower-dimensional homologies (similarities based not on the whole texts but on very finite aspects of it).[54] A corpus is mapped on vectors, each one representing a document by encapsulating it using a measure of relative weight of each term. This vector space model generates a set of data points in a non-Euclidean coordinate system that CLS then presents as topological information. Topological modeling is used, for example, to calculate sociability and social interactions in literary landscapes, using an extremely metaphorical interpretation of the topological edge. What the literary sociologist Alan Liu means by "latent social network" or "core circuit" is simply a visualization of connections using a functionally reductive definition of "ties."[55]

CLS network analysis can easily become recommender-system literary sociology, in which consumer and discursive associations are visualized without regard to the tone, context, emphasis, rhetoric, and so on—exactly in the way recommender systems function. In these, word frequency overlaps constitute *spatial connectivity* and a *network* means a simple visu-

54. On the limits of applying computational topology to high-dimensional data, see Herbert Edelsbrunner and John L. Harer, *Computational Topology: An Introduction* (Providence, R.I., 2010), and Hubert Wagner, Pawel Dłotko, and Marian Mrozek, "Computational Topology in Text Mining," in *Computational Topology in Image Context: 4th International Workshop CTIC 2012 Proceedings* (Bertinoro, 28–30 May 2012), pp. 68–78.

55. Liu, "From Reading to Social Computing," *Literary Studies in the Digital Age: An Evolving Anthology*, ed. Kenneth M. Price and Ray Siemens, dlsanthology.commons.mla.org /from-reading-to-social-computing/. Transferring Pierre Bourdieu's notion of a literary economy into literary sociality, Liu argues that we can visualize a piece of literature's "latent social network and that of the characters in its imaginative worlds" using a "core circuit" in which "editors, publishers, translators, booksellers" act as "vital nodes" (ibid.).

alization of a very small number of these connections. Such diagrams are often rendered with "off-the-shelf-social-computing tools and platforms created for other purposes."[56] But these off-the-shelf tools, such as the Facebook Friend Wheel, are useful if you wish to promote socialization or enterprising opportunities by mapping out your networks, which are dynamic and complex not in reference to the *nature* of the connections in question but in reference to their order of magnitude and the amount of topological information embedded therein. Network maps are used to calculate the centrality of nodes based on directional vectors; so Google, for example, knows how to turn up most relevant searches because it calculates the number of nodes (sites) in its network that link to another site and in so doing calculates the relative centrality of a site. A network map cannot be replaced by other forms of data representation. It becomes complex because of size and connections (which increase at a rate of 2^n): seating arrangements for a wedding with five hundred guests—some of whom cannot sit with some others and who all have a descending list of proximity preferences—become that much more complicated if the total number of guests increases to five million. Capturing this kind of complication—or capturing network complexity by studying a network whose degree distribution of nodes to links is neither arbitrary nor regular but obeys some other mathematical law—is not the same as saying that network diagrams of who is talking to whom in Shakespeare can capture the complexity of connections in Shakespeare or character discourse. We are dealing with fundamentally different definitions of complication and complexity.

Network diagramming for low-volume data is not a meaningless activity if it can help us see what we otherwise cannot see, but this payoff is often missing from such visualization. Ed Finn creates a network map of Junot Diaz's *The Wondrous Life of Oscar Wao*'s Amazon page with "book reviews and website recommendations . . . as links" and book "titles as nodes," which is intended to visualize consumer and discursive associations.[57] Using scripts that recursively gather recommendations, Finn maps out the first ten "Customers Who Bought This Also Bought" links and the first ten recommendations for each of these links over several months (from December 2010 to March 2011) in order to produce a network map (fig. 10). In this map, though, where is the network *analysis*? Where are the centrality scores? What are the assortativity measures? The statistical inference?

56. Ibid.

57. Ed Finn, "Revenge of the Nerd: Junot Diaz and the Networks of American Literary Imagination," *Digital Humanities Quarterly* 7, no. 1 (2013), www.digitalhumanities.org/dhq/vol/7/1/000148/000148.html

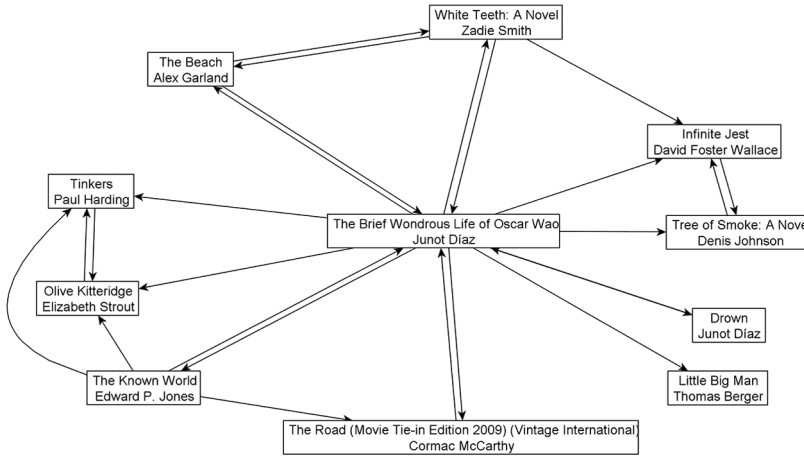

FIGURE 10. "Amazon Recommendations, Diaz, Late December 2010," figure 4 in Ed Finn, "Revenge of the Nerd: Junot Diaz and the Networks of American Literary Imagination," Digital Humanities Quarterly 7, no. 1 (2013), www.digitalhumanities.org/dhq/vol/7/1/000148/000148.html

Properly defining nodes has no operationalizable end here, in contrast to, say, the NSA keeping track of terrorist webs on social media by investigating up to three nodes of connection. For Finn, each mention of another author (regardless of the nature of the mention), whether in Amazon's recommender system or in these sundry reviews, is proof that *Oscar Wao* serves as a "literary gateway from this genre of ethnic literature to a distinct canon of mainstream prize-winners" or proof of "the process of literary reverse colonization, of deliberately contaminating the language of one discourse with the icons of another."[58] These are captivating ideas, but Finn performs no network analysis (he has made *Oscar Wao* the de facto center of his graphs) because these are simply eleven items and their connections to one another. This is not a network map but only a very, very small piece of a network map—one that could easily be represented in a table. There might be an order of magnitude between the first recommended book and the second one, but Amazon does not reveal that information to the consumer. Finn is weighing the referrals equally because he only has partial access to Amazon's item-to-item collective filtering algorithm (full access would mean that Finn will simply duplicate what is already on Amazon).

Topological insight and topologically structured visualization tools for word frequency arguments: these are not the same things. Piper describes

58. Ibid.

his topological renderings as "autochthonic" and "protocologic," a Latourian network of "'quasi-objects,'" a Deleuzian "'relation of the non-relation,'" a Badiouian questioning of "'the aura of the limit,'"[59] "a different kind of thinking about the beyond," a Judd-Morrissey-inspired "radical act of interleaving," something that allows us to "think more about language in agential terms (what it does),"[60] a Foucauldian "'field of regularity,'"[61] and something that "move[s] past the ontology of discourse" but also "allows for a far more nuanced sense of discursive being."[62] These inspired comparisons are hard to reconcile with what his uses of topology actually amount to, and they get in the way of seeing what topological maps actually do. His project with Mark Algee-Hewitt, "The Werther Effect," for example, is a series of topological visualizations that capture the influence of Johann Wolfgang von Goethe's *The Sorrows of Young Werther* (1774) on his later works (and other English and German works after Goethe).[63] "Influence" means tracking ninety-one representative words in *The Sorrows of Young Werther* and their frequencies in x other works, a measurement deemed important because Goethe supposedly wrote differently after this renunciation of *Werther* and because *Werther* is known to have influenced later works, but we don't know how or to what degree. Piper and Algee-Hewitt take the Euclidean distance of word frequency measurements to measure the similarity of lexicon across works and then, in order to visualize their matrix, try and find the best way to collapse the information in the matrix into a picture because this distance matrix is large and the information is not easily grasped. They chose a Voronoi diagram, a very useful and intuitive form of data visualization that allows you to see geometrically how distant a work is from every other work to scale.[64] Topology functions here as an optimal way to visualize a matrix of word frequency differences; it is not a representation of how we read, visually, no matter how it is plied metaphorically. And generating a Voronoi diagram aside (whose application in these types of data situations is not the authors' original contribution), what these distance measurements—which now can be seen all at once—represent is the way that ninety-one words show up (regardless of location, order, context, syntax, speaker, voice, tone, proximity

59. Piper, "Reading's Refrain: From Bibliography to Topology," *English Literary History* 80 (2013), pp. 386, 384, 381, 386.

60. Piper and Algee-Hewitt, "The Werther Effect I," pp. 162, 157.

61. Piper, "Novel Devotions," p. 71.

62. Andrew Piper, "Reading's Refrain," p. 381.

63. Much of this project has been published on McGill University's blog .TXTLAB, but I will only look at the parts that are already published in *English Literary History* as "Reading's Refrain."

64. See section 7 of the online appendix.

to one another) in the rest of Goethe's oeuvre.[65] At the end of the day, the repetitions of those ninety-one words indicate the influence of *Werther* on other texts. In another forum we as literary critics have to decide how much we're invested measuring the precise indices of influence and if the fact that a set of words in *A* also show up frequently in *B* means that *A* influenced *B*; here, it is enough to see that it is this is the same argument that we've seen in every paper: overlapping most-frequently-used vocabulary denotes influence, and when *A* isn't *exactly B*, word for word, *B* has definitionally influenced *A* by degrees.

6

Quantitative visualization is intended to reduce complex data outputs to its essential characteristics. CLS has no ability to capture literature's complexity. Mark Algee-Hewitt wants to go beyond word-frequency counts to measure for literary entropy, or level of redundancy in a piece of work, which would *seem* to be a complexity measure. His contribution to the multiauthor Stanford Lab article "Canon/Archive. Large-scale Dynamics in the Literary Field" is to argue that noncanonical texts are less entropic (more redundant) than canonical texts, using 260 titles from the Chadwyck-Healey corpus as the canon corpus and 949 titles from the same time period as the noncanon corpus. He measures number and probabilities of consecutive pairs of words therein on the reasoning that the more entropic a piece of literature is, the less redundant it is and the more information it contains.[66] Entropy measure sounds sophisticated (and seems analogous to literary complexity), but what it does, actually, is measure the number of distinct pair of words and their distribution in the total number of bigram pairs.[67] It is not a mysterious property but directly tied to variety of words (two thousand, twenty thousand, or two million distinct words make a huge difference) and skewedness of words (whether a couple of words are the ones that are always appearing or if each of the words appears just once). Entropy levels are highest in a situation wherein bigram pairs are diverse but no particular bigrams dominate others, leading to more information in a text which, as Warren Weaver says, "must not be confused with meaning."[68] Even if we agree that more mathematical entropy somehow means more literary novelty or less literary redundancy, as Algee-Hewitt would have it, his calculations are still wrong. Using an Archive corpus of

65. See section 6 of the online appendix for the English and German list.

66. Scripts and metadata for articles published through the Stanford Literary Lab prior to December 2016 are not available.

67. Algee-Hewitt's measurement is an adaptation of the Shannon-Weaver formula.

68. Claude E. Shannon and Warren Weaver, *The Mathematical Theory of Communication* (1949; Urbana, Ill., 1998), p. 8.

356 books (thus closer in size to their Chadwyck-Healy corpus of 260 books), I recalculated entropy for both (scaled entropy scores = 0.796391 and 0.793993, respectively) and found no statistical difference between the two after robustness tests.[69] Algee-Hewitt's larger entropy for the Chadwyck-Healy corpus is driven by the large size of his archive corpus (263 versus 949), which creates a magnitude of difference between the number of distinct bigrams for Chadwyck and for the archive, which causes the archive entropy score to go down. His finding, the basis for a significant portion of "Canon/Archive," is just a scaling oversight.[70]

CLS has not kept pace with corpus linguistics in accounting for things like coreferentiality or sentence processing that care about words as embedded in linguistic structures (local discourse).[71] CLS does use natural language processing (NLP) to tag parts of speech and phonemes, seemingly to move beyond summary statistics to grab words in a more semantically meaningful way, but these efforts are halfhearted, and there are reasons for this beyond the fact that NLP has only recently taken off. Speech tagging is extremely inaccurate for literary texts. Lexical, syntactic, and grammatical ambiguities make it difficult for an algorithm to know whether a word is a participle or a gerund, if an adjective is a noun, or if entire phrases are functioning as a single part of speech. NLP is said to have a 93–95 percent accuracy level, but that depends on what you're using it for and the degree of classification you require (thus, formal evaluation is very difficult). Having 95 percent accuracy for building an online chatbot or for basic translations is very different than 95 percent accuracy at picking out all the parts in a piece of literature. NLP software for narrative parts of speech tagging is also not very user friendly because it requires that one manually annotate a training set.

69. See section 7 of the online appendix.

70. Algee-Hewitt tries to account for the discrepancy of corpus size by using a Kullback-Leibler Divergence measure (KL), but a KL measure is used to see how one probability distribution diverges from the target probability distribution. Insofar as the two corpuses diverge when we looked at which items contributed most to the divergence between the two corpuses, it was the most common 1/10 of intersecting bigrams; see section 7D of the online appendix. This means that one corpus is not more entropic than the other but where they do not 100 percent intersect is mostly in commonly occurring pairs of words.

71. As modeled, for example, in Kerry Ledoux et al., "Coreference and Lexical Repetition: Mechanisms of Discourse Integration," *Memory and Cognition* 35 (June 2007): 801–15. Henry Kučera and W. Nelson Francis's *Computational Analysis of Present-Day American English* (Providence, R.I., 1967) does not seem to be in use. Nor does Michael A. K. Halliday's classification of processes into six types of actors and six types of actions for functional linguistics seem to have been taken up; see Michael A. K. Halliday, *An Introduction to Functional Grammar*, ed. Christian M. I. M. Matthiessen (New York, 2004), p. 81.

You quickly run into a data scarcity and data complexity problem with literature. How many distinct sets of literature are out there—that you would be able and willing to manually annotate—that would be large enough to allow you to accurately run NLP on the rest of the set? And what do you do after you've tagged a text? Suppose someday all literary things (including homonymy, figuration, polysemy, irony, transference) can be accurately tagged—a pretty big supposition. The researcher would still be left with a list of tags and their frequencies, which would have to be heavily reduced in dimensionality to have any extractable statistical meaning. In this case semantics or basic plot are still being ignored (unless we are willing to accept their premise that words statistically co-occurring with others can effectively represent semantics, topicality, or plot). For other fields of study, named entity recognition tasks can be used to provide that second layer of classification, sorting tagged words into predefined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, and so on. But widening out in this way to get workable categories only make sense when you have truly large data sets and when you desire to quickly extract some usable information. Tagging errors and imprecision in NLP do not sufficiently degrade the extraction of information in many other contexts, but they do for literature.

Even when applied to the kinds of text most suited for it—NLP lends itself particularly well to reportage of data that are abundant but similar[72]—Franzosi spent thirty years manually training his tagger on newspaper articles ("10–15 minutes per one-page document for an experienced coder working with fairly complex coding schemes") to confirm simplistic versions of basic historical facts.[73] Martin Paul Eve also tries to get beyond stop word frequencies by turning to NLP to prove that David Mitchell's *Cloud Atlas* is a mishmash of genres.[74] It is an exemplary case because Eve only uses the statistical tools that are needed, explains the relative simplicity of his measurements and gives credit to these measurements as things already available in coding packages instead of presenting them as though he devised them from scratch. Instead of making homology calculations after removing stop words, Eve saw that a much simpler classifier

72. See Franzosi, De Fazio, and Vicari, "Ways of Measuring Agency."

73. Roberto Franzosi, *Quantitative Narrative Analysis* (Los Angeles, 2010), p. 149. These being the "historians' division of the period into 'red years' and 'black years,'" wherein the police were the primary agents of violence against the working class in the red years and fascists the main agents of violence against the working class in the black years, thus confirming historians' designation of the factory occupation movement of September 1920 as the pivot between the two phases (ibid.).

74. See Martin Paul Eve, "Close Reading with Computers: Genre Signals, Parts of Speech, and David Mitchell's *Cloud Atlas*," *SubStance* 46, no. 3 (2017): 76–104.

could be had by the frequency measurements of common stop words (*the*, *a*, *I*, *to*, *of*, and *in*) that can accurately classify all the sections of *Cloud Atlas* save one (with twenty common stop words being able to classify *all* the sections), and he takes the Manhattan distance of z-scores and the clustered dendrograms of the five thousand most frequently used words (or two words) to predict the likelihood that different sections in *Cloud Atlas* were written by the same author. Eve then turns to NLP to show that the Luisa Rey section of *Cloud Atlas* has statistically significant occurrences of the tagged trigram NNP+NNP+VBZ (proper noun singular + proper noun singular + third-person singular present verb). But this turns out to have a completely prosaic explanation. All that Eve has done is to prove that Mitchell's sections are as distinct from one another as they are from some other author using stop words. NLP does not offer any additional insight. To actually explain the tropic substrate of distinctive trigram frequencies, he still had to go in and find the instances of adverb+adjective+noun and distinguish a "'hopelessly uneven gunfight'" from a "'mostly empty wine' glass."[75] Because of UK copyright laws, Eve typed out the novel manually. This is a lot of work to learn with certainty that one chapter pairs more full names of characters with actions than another.

For an even clearer instance, the problem with So and Long's haiku classifier is not its accuracy rate or even its parametricization but its functionality. Of course, the classifier does not have to be 100 percent accurate—one cannot reject it simply by finding examples of misclassification. If for Long and So (1) "translations and adaptation," (2) things calling themselves haikus, and (3) things other people have classified as haikus are indeed the same kinds of things—haikus (whatever their differences)—then whatever the Naïve-Bayes classifier classifies as an English haiku is, by their very definition, an English haiku, as they don't have a rigorous definition to begin with. But are we talking about enough ambiguous cases (or even total number of very short poems) to justify this error? Are we facing a situation where millions of short poems are published but we cannot possibly find the time to read them? The authors themselves do not have a good way to find and scrape all the short poems that exist in the world without knowing in advance where to look, so they have not saved us any time on that account. Couldn't someone trained in poetry just find, read, and classify them?

75. Eve, "Close Reading with Computers," p. 101. Eve acknowledges the limitations to validating his cluster dendrogram analysis and uses bootstrap consensus tree plotting, and he also unveils the hidden driver right away by pointing out that section's tendency to pair full character names with actions ("Rufus Sixsmith leans," "Luisa Rey hears," "Maharaj Aja says," and more) (p. 88).

Supporters of CLS argue that it does not matter that it takes a long time to do something we already know, as the innovation is in a computer being able to do basic reading tasks at all (an argument for artificial intelligence). But it does matter because computation is being used here as an investigative tool that shows you where to look or what to casually opine on, and CLS authors simply pick up influence, change over time, no change over time, generic coherence, or generic difference arguments along the way because they've defined these to be identical with the only kind of data processing they can do in order to use these particular tools and have statistical inference at all. This is not artificial intelligence but humans working in summary statistics.

CLS has also excused its methodological and argumentative flaws by appealing to a trade-off: *Who can possibly read all the literary texts that are out there? Machine reading is not perfect, but it's better than nothing, and it can show us latent patterns that no one reader can see.* Literary critics, especially those studying contemporary literature, tend to look to DH to help them account for literary objects that they feel are exponentially increasing. They naturally assume that computational methods can help them tackle this scale in a faster, more comprehensive, and nonarbitrary manner. As all the above examples prove, that is a misperception. Looking for, obtaining copyrights to, scraping, and drastically paring down "the great unread" into statistically manageable bundles, and testing the power of the model with alternative scenarios, takes nearly as much, if not far more, time and arbitrariness (and with much higher chance of meaninglessness and error) than actually reading them. CLS's methodology and premises are similar to those used in professional sectors (if more primitive), but they are missing economic or mathematical justification for their drastic reduction of literary, literary-historical, and linguistic complexity. In these other sectors where we are truly dealing with large data sets, the purposeful reduction of features like nuance, lexical variance, and grammatical complexity is desirable (for that industry's standards and goals). In literary studies, there is no rationale for such reductionism; in fact, the discipline is about *reducing* reductionism. Even macroanalytical results cannot themselves be the products of reductionist thinking.

With regard to the overabundance argument, it is important to remember that many of the key examples come from corpora or texts that have already been read. CLS is really not dealing with nearly as much data or complexity (of the order that justifies the use of the tools they use) as authors like to think. Basic math also helps here: one million words roughly equals ten novels; one and a half billion represents about fifteen thousand novels, which at one novel a month will only take one thousand people one

year to read. At the end of the day, the overabundance claim is not a legitimate argument in and of itself. In the sciences and social sciences there is also an inestimable volume of texts, data sets, and scenarios that haven't been covered. There are a lot of things about which we don't know and lots of questions we haven't answered. That does not mean that any pattern that can be found in that unknown data, any answer to any previously unasked question, or any question, is automatically worthy of attention. The basic criteria should always be to not confuse what happens mechanically with insight, to not needlessly use statistical tools for far simpler operations, to present inferences that are both statistically sound and argumentatively meaningful, and to make sure that functional operations would not be far faster and more accurate if someone just read the texts.[76] It may be the case that computational textual analysis has a threshold of optimal utility, and literature—in particular, reading literature well—is that cut-off point.

76. These basic criteria are detailed in section 9 of the online appendix.