Macroanalysis

Matthew L. Jockers

Published by University of Illinois Press

Matthew L. Jockers.
Macroanalysis: Digital Methods and Literary History.
Champaign: University of Illinois Press, 2013.
Project MUSE. Web. 7 Jul. 2015. http://muse.jhu.edu/.

9 INFLUENCE

> Every work of art is the result of a complex interrelation of individual features of creative art. The author's role is to use these features and to combine them into a definite artistic product. The elements of which the artwork is created are external to the author and independent of him. The author merely uses them for his work, with a greater or lesser degree of success.
>
> In every period there is a certain number of artistic methods and devices available for creative use. Changing these methods and devices is not a matter of the individual author's volition, but is the result of the evolution of artistic creativity.
>
> —Osip Brik, "Teaching Writers" (1929)

Examining macro patterns in style and theme allows us to contextualize our close readings in ways that have hitherto been impossible or, at the very minimum, impractical. We see, for example, that while Melville may be best remembered for *Moby Dick, Moby Dick* was only the apex text in a longer tradition of whaling- and seafaring-themed fiction, a tradition that stretches back at least to Sir Walter Scott's book *The Pirate* (1821) and through the work of Frederick Marryat.* Along the way, from Scott to Marryat to Melville, other writers touch upon and help build the themes that ultimately find full expression in *Moby Dick.* If we look only at those novels in the corpus containing at least 1 percent of the "Seas and Whaling" topic, we find thirty-six, including books by James Fenimore Cooper, Edward Augustus Kendall, Edgar Allan Poe, Nathaniel Hawthorne, J. H. Ingraham, and thirteen others. We know that Melville was a borrower, and the evidence that he borrowed from Poe's *Narrative of Author Gordon Pym* and from the Reverend Henry Cheever's book *The Whale and His*

* Marryat was a prolific novelist and naval officer who, like Scott, wrote a novel titled *The Pirate* (1836). Marryat also developed the maritime-flag signaling code that bears his name—"Marryat's code."

Captors is fairly well known and easy for close readers of these works to detect (see, for example, Lee 1984; Simon 2005). What is not as clear are the more subtle spheres of influence; consider, for example, an "allusionary" chain. At the opening of *Moby Dick* (1851), Melville cites a line from Hawthorne's *Twice Told Tales* that reads: "I built a cottage for Susan and myself and made a gateway in the form of a Gothic Arch, by setting up a whale's jaw bones." Hawthorne may well have picked this up from Scott, for in Scott's *Pirate* (1821), we are told of a Scottish burgh that had been renovated in a Gothic style with an entrance gate "supported by a sort of arch, constructed out of the jaw-bones of the whale." Ten years after Scott's *Pirate,* the narrator of Anna Marie Hall's *Sketches of Irish Character* (1831) tells of a "mysterious arch, composed of the jaw-bone of a whale" upon which she gazed from her "cottage" where she "kept all her favorite books," including, we are told, books by Walter Scott! Thirteen years after that, Eliza Lanesford Cushing's *Fatal Prediction* (1844) describes a similar jaw-bone arch found outside a fortune-teller's "cottage." Is Hall Hawthorne's literary ancestor, or a more distant relative of Melville, or a descendant of Scott? Are recycled elements and allusions such as these a matter of coincidence or design? Or are they, as Osip Brik suggests, part and parcel of an involuntary creative evolution?*

That such arches existed in reality is a matter of fact; why they become a touchstone in a series of "unrelated" literary works is uncertain. Whatever the reason, the presence of recurring themes and recurring habits of style inevitably leads us to ask the more difficult questions about influence and about whether these are links in a systematic chain or just arbitrary, coincidental anomalies in a disorganized and chaotic world of authorial creativity, intertextuality, and bidirectional dialogics. This kind of big question takes us beyond single books, beyond recurring allusions, and even beyond the macro patterns and trends that have been graphed and charted in previous chapters. At the very least, they demand that we look for some significance in the apparent chaos.

"Evolution" leaps to mind as a possible explanation.† Information and ideas can and do behave in ways that seem evolutionary. Nevertheless, I prefer to avoid the word *evolution* (even though I have just used it several times, and even though it is a favorite trope of the Russian formalists, whose approach I obviously admire): books are not organisms; they do not breed. The metaphor for this process breaks down quickly, and so I do better to insert myself into the safer, though perhaps more complex, tradition of literary "influence" and to simply investigate literary influence on a grander scale than close observation and anecdotal speculation allow. Before abandoning the word *evolution*

* See the epigraph.

† Whereas the dialogic text, in Bakhtin's sense of the word, is in dialogue with works both before and after, allusion and influence, whether intentional or accidental, exist in only one direction. In this sense, *evolution* may be a more appropriate term or analogue.

completely, however, one minor point: evolution is not moving us toward anything in particular. It is only movement and change. There is no end point to evolution. Nor is there any grand objective behind literary change. This is not to say that individual authors have no agency or do not strive to create something better or different or new—they do; they strive. Instead, I wish to suggest that a writer's creativity is tempered and influenced by the past and the present, by literary "parents," and by a larger literary ecosystem. We cannot argue that *Middlemarch* is a great novel, any more than we can argue that *Homo erectus* was a great man; we can only argue about what makes one different or similar to its peers. I do, however, accept that in literature, as in nature, there are survivors, thrivers, outliers, mutations, and there is also that which does not survive. *Middlemarch* and *Homo erectus* may fit into one of these categories, and it may, therefore, be entirely appropriate to call attention to these forms as exemplars. Some forms (of life and of literature) excel and become more common; others shine just briefly. My interest is in tracing where and when these forms emerge and then where and when they die. My interest is in finding the context in which change occurs, for it is only by understanding the larger context that we might then move to address the deeper question of creation, of how and why such forms come into being in the first place.

Attempts to demonstrate literary imitation, intertextuality, and influence have relied almost entirely upon close reading.* It seems very likely that Melville's "Call me Ishmael" is a direct echo of Poe's "My name is Arthur Gordon Pym." And though knowledge of this might add to our understanding of Melville's art (and perhaps also to our appreciation of Poe . . . that is, Melville could not have done it without him), this is not the scale of influence about which I am thinking. To chart influence empirically, we need to go beyond the individual cases and look to the aggregate. Borrowing a whale's jaw or a catchy opening sentence is neither imitation nor influence, not in the full-throated sense that I mean to explore. Melville's echo of Poe likely goes deeper; it may be just one among a hundred similar echoes in a hundred other novels. The existence of such a state would certainly alter our understanding of what it means to be influenced.

Within the field of observational learning, there exists a theory of "information cascades." The landmark essay defining these phenomena was published in 1992; it begins as follows: "An informational cascade occurs when it is optimal for an individual, having observed the actions of those ahead of him, to follow the behavior of the preceding individual without regard to his own information" (Bikhchandani, Hirshleifer, and Welch 1992, 992). Information cascades offer a theory of social behavior that serves as a close, if sometimes imperfect, corol-

* I say "almost" because text searching and tools such as Google's "Popular Passage" finder have made it possible to employ computational search in place of sustained and concentrated reading. For more on Popular Passage, see Schilit and Kolak 2007.

lary for the kind of thematic and stylistic change explored in previous chapters. The authors of this landmark paper write that their model of social behavior "explains not only conformity but also rapid and short-lived fluctuations such as fads, fashions, booms and crashes" (ibid., 994). More important for our purposes, they argue, as stated above, that an information cascade occurs when "an individual . . . follow[s] the behavior of the preceding individual *without regard to his own information*" (ibid.; emphasis added). In other words, once a cascade begins, it tends to continue and to create a situation of mass imitation in which individuals repeatedly *avoid* the road less taken. Were a whole series of writers to begin writing whaling novels after the publication of *Moby Dick,* it could be argued that these subsequent authors were caught up in an information cascade in which their own independent ideas about what to write were trumped by some herd instinct.* Likewise, this same theory might offer some manner of explanation for the upward trend in British and Irish usage of the "confidence" markers seen in figure 7.6. At the same time, the theory tells us that cascades are fragile; the introduction of a disruptive force, a new "signal," can cause the cascade to collapse and move in an entirely new direction. Not everyone would follow Melville's lead; some mutant writer would take some other road, and a new cascade would follow. As a way of modeling literary influence and intertextuality at scale, information cascades provide an attractive theoretical framework. Whether the data in the literary record can be explained by this theory of information exchange is worth exploring.†

For every book in the Literary Lab corpus, I have extracted both stylistic (as in chapter 6) and thematic information (as in chapter 8). These data can then be combined into an aggregated numerical representation or "expression" of the stylistic and thematic content of every book in the corpus. The resulting data matrix is 3,346 by 578.‡ In each row, 578 different feature measurements represent a book's thematic-stylistic expression, or "signal."§

* The contemporary fascination with vampires may be another and more familiar example.

† Consider the genre trends that Moretti describes in *Graphs, Maps, Trees* (2005). Moretti's graphs show us how genres appear, build steam, and then fade out, in what are essentially twenty- to thirty-year cycles. An alternative to the generational hypothesis is that genres represent a type of "information cascade." Such a model could help explain some of the genre-generation discrepancies explored in chapter 6.

‡ Not included here were the uninterruptable topics and the topics that were clearly derived from either book metadata or from bad optical-character-recognition data. For details, see chapter 8.

§ I cannot resist the great temptation to liken these data to the genome. Still, 578 "genes" does not come close to the 20,000–25,000 genes that are estimated to make up human DNA; at best, it is an incomplete, or partial, *literary* genome.

To explore these data, to test the waters and get a sense of how well this amalgamation of features might approximate or represent the book from which it was extracted, I trained a classifier and then tested how well nationality and gender could be predicted from the features. Here I could not test genre prediction because not all of the texts in the full corpus have been coded with genre metadata, nor did it make sense to try to classify by author given the size of the corpus and number of authors in it. Gender and nationality, where the number of classes was limited to two and three, respectively, would be enough. The gender and nationality results were perfectly consistent with what had been observed in the classification tests described in the previous chapters. Surprisingly enough, the combination of stylistic and thematic information neither improved nor worsened classification accuracy. This result may suggest that theme and style are to some extent interdependent: perhaps thematic choices entail stylistic ones. Such a conclusion would be in keeping with, if a slight extension of, the discovery in "Quantitative Formalism: An Experiment" that Gothic novels seem to demand a higher proportion of locative prepositions (Allison et al. 2012). Interesting and inconclusive, this is a fruitful area for further exploration and one that may have implications for scholars working in authorship attribution in particular. This is, however, beyond the scope of the present study.

My objective now is not to classify novels into nationalities or genders but rather to capture for each book a unique book signal and then to look for signs of historical change from one book to the next. Using the "Euclidian" metric, I calculated every book's *distance* from every other book in the corpus.* Assume that we have three books and only two features for each book. Let's calls the three books b1, b2, and b3 and the two features f1 and f2. Table 9.1 shows these data and some "dummy values" for each feature in each book. These points can each be plotted in a two-dimensional space, as in figure 9.1, where books b1 and b2 are closest (least distant) to each other in the lower-right corner. These simple distances can be perceived visually, measured with a standard ruler, or, of course, calculated with a simple equation—really just a version of the familiar Pythagorean equation.

This fairly simple equation, thought of in two dimensions, becomes more complex when thought of in terms of 578 dimensions. The closeness of items in this high-dimensional space can, however, still be calculated. Assume a new data set in which we have just four features (as in table 9.2). Using the Euclidean metric, the distances "d" between books (b1, b2, b3) are calculated as follows:

$$d(b1, b2) = \sqrt{(10-11)^2 + (5-6)^2 + (3-5)^2 + (5-7)^2} = 3.162278$$

$$d(b1, b3) = \sqrt{(10-4)^2 + (5-13)^2 + (3-2)^2 + (5-6)^2} = 10.09950$$

$$d(b2, b3) = \sqrt{(11-4)^2 + (6-13)^2 + (5-2)^2 + (7-6)^2} = 10.39230$$

Table 9.1. Example feature data, version 1

Book	f1	f2
b1	10	5
b2	11	6
b3	4	13

Note: "*f*" is any arbitrary feature.

Table 9.2. Example feature data, version 2

Book	f1	f2	f3	f4
b1	10	5	3	5
b2	11	6	5	7
b3	4	13	2	6

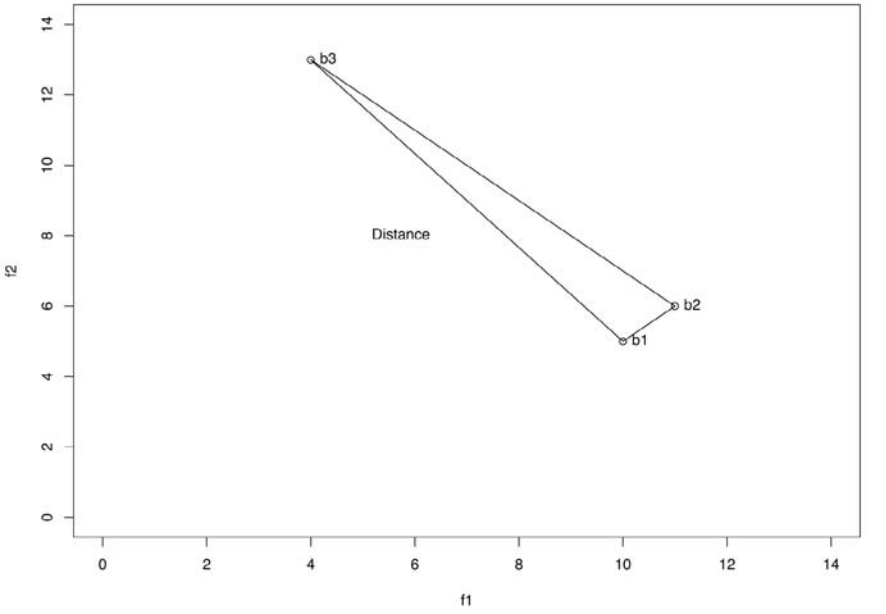

Figure 9.1. Example plotting of distance between books

* In "Textual Analysis," Burrows reports on his own finding that "complete linkages, squared Euclidean distances, and standardized variables yield the most accurate results" when seeking to cluster texts by similarity (2004, 326).

Table 9.3. Euclidean distances from *Pride and Prejudice* based on 578 features

Rank	Author	Title	Distance
0	Austen, Jane	*Pride and Prejudice*	0
1	Austen, Jane	*Sense and Sensibility*	0.042557864
2	Austen, Jane	*Mansfield Park*	0.049754052
3	Austen, Jane	*Emma*	0.050242054
4	Burney, Sarah	*Traits of Nature*	0.056073837
5	Cathcart	*Adelaide: A Story of Modern Life*	0.057314379
6	Waddington, Julia	*Misrepresentation; or, Scenes in Real Life*	0.058382231
7	D'Arblay, Frances	*Cecilia; or Memoirs of an Heiress*	0.058646462
8	Burney, Sarah	*Tales of Dancy*	0.059090054
9	Humdrum	*Domestic Scenes: A Novel*	0.059223492
10	Lister, Thomas	*Herbert Lacy*	0.059397822

In this case, the distance between b1 and b2 (3.162278) is much smaller than the distance between b1 and b3. This indicates that b1 and b2 are more similar to each other in terms of these four features. Using the R statistics package, it is a trivial matter to calculate the distances between every book and every other book in the corpus along all 578 dimensions. The result is a 3,346 × 3,346 distance matrix. Every row of this table represents a single book, as does every column; the values in the individual cells are the calculated distances between them.*

From this "distance matrix," any book in the corpus may be selected, and a ranked list (based on distance) of all of the other books in the corpus can be returned for inspection. The first thing one discovers by going through a few of these lists is that books by the same authors tend to show up at the top of the list. Books by the same author tend to be stylistically and thematically similar. A ranked list of books closest to *Pride and Prejudice,* for example, is shown in table 9.3.

The presence of three other books by Austen at the top of a list, which began with a search for books most similar to *Pride and Prejudice,* confirms much of what we have already learned from the authorship-attribution literature and from what has been reported in previous chapters. Austen wavers only slightly when it comes to her core themes and even less so when it comes to her linguistic signature. Put rather too bluntly, neither Austen's stylistic nor her thematic

* As you can imagine, there is a diagonal of "0" values in the cells where the row from book b1 intersects with a column for the same book, b1. There is zero distance between a book and itself.

range is exceedingly vast. It must be kept mind that these rankings of similarity are relative to the corpus as a whole, relative to all 3,346 books. In other words, *Sense and Sensibility* may not ultimately be the most similar book in the universe to Austen's *Pride and Prejudice,* but it is the most similar of the books in this corpus. As the corpus grows, the values may change to greater or lesser degrees. A similar study can be made of Dickens's *Tale of Two Cities.* As with Austen's novel, *Tale* returns several other books by Dickens (table 9.4), but also shows the strong presence of another author, George Payne Rainsford James. James was a close contemporary of Dickens—born thirteen years earlier—and, like Dickens, was prolific, publishing more than forty novels.

Worth noting too is that in the top-ten results for *Pride and Prejudice,* we find only one male-authored book, and all are books by British authors. Excepting *Life's Masquerade,* which is of unknown authorship, in the *Tale of Two Cites* list, we find all male and British authors. The result is much the same when charting books similar to *Moby Dick.* We observe an all-male list that includes at the top other works by Melville (table 9.5) followed by works of his American contemporaries, including most prominently James Fenimore Cooper and Edgar Allan Poe. The two exceptions are Robert Ballantyne and Robert Louis Stevenson, two Scots. The lesser-known Ballantyne spent six years in Canada in the employ of the Hudson's Bay Company and then wrote works largely based on this experience. Interestingly, the work of Ballantyne's identified here, *The Coral Island,* is known to have been an influence upon Stevenson, who included mention of Ballantyne in the introductory poem that prefaces *Treasure Island*.*

Table 9.4. Euclidean distances from *Tale of Two Cities* based on 578 features

Rank	Author	Title	Distance
0	Dickens, Charles	*A Tale of Two Cities*	0
1	Dickens, Charles	*Master Humphrey's Clock*	0.046820931
2	Dickens, Charles	*Little Dorrit*	0.0472454
3	James, George Payne Rainsford	*The False Heir*	0.048010961
4	James, George Payne Rainsford	*Lord Montagu's Page: A Historical Romance*	0.048851065
5	James, George Payne Rainsford	*The Vicissitudes of a Life: A Novel*	0.050046016
6	Locker, Arthur	*Sir Goodwin's Folly: A Story of the Year 1795*	0.051164434
7	Collins, Wilkie	*After Dark*	0.051527178
8	Dickens, Charles	*Dombey and Son*	0.051644288
9	Dickens, Charles	*Barnaby Rudge*	0.051970992
10	Unknown	*Life's Masquerade: A Novel*	0.052783311

* Were table 9.5 expanded from the top ten to the top eleven, *Treasure Island* would be the eleventh book in the list.

Table 9.5. Euclidean distances from *Moby Dick* based on 578 features

Rank	Author	Title	Distance
0	Melville, Herman	*Moby-Dick; or, The Whale*	0
1	Melville, Herman	*Omoo: A Narrative of Adventures in the South Seas*	0.057784134
2	Melville, Herman	*Mardi and a Voyage Thither*	0.073077699
3	Cooper, James	*The Crater; or, Vulcan's Peak: A Tale of the Pacific*	0.073798396
4	Cooper, James	*The Sea Lions; or, The Lost Sealers*	0.08339971
5	Melville, Herman	*Typee: A Peep at Polynesian Life*	0.101393295
6	Ballantyne, Robert	*The Coral Island: A Tale of the Pacific Ocean*	0.117425226
7	Poe, Edgar Allan	*The Narrative of Arthur Gordon Pym of Nantucket*	0.13125092
8	Stevenson, Robert	*Island Nights' Entertainments*	0.146050418
9	Williams, William	*The Journal of Llewellin Penrose, a Seaman*	0.176751153
10	Payn, James	*A Prince of the Blood*	0.18207467

These tables listing the distances between books take us in the direction of gauging influence, but they are still too small in scale to give us the broad picture of literary history that we are looking for. Having computed the stylistic-thematic distances among all the books in the corpus, it is possible to move even further away from individual data points and into a larger-scale visualization of the entire corpus. For this, network visualization software is well suited.

In terms of literary history and literary *influence,* our corpus is a type of network. Whether consciously influenced by a predecessor or not, every book is in some sense a descendant of, or "connected to," those before it. Its relationship may be familial, that is, a new book by the same author, or it may be parodic, as in *Shamela,* a book meant to be a direct response to some other book. Or the relationship may be indirect and subtler, as when an author unconsciously "borrows" elements from the book(s) of some predecessor(s), or simply pulls from the same shared pool of stylistic and thematic materials. Previous chapters have shown how writers can draw elements from what is available on their stylistic and thematic "buffets": that is, writers may consciously or unconsiously adopt the habits of prose that are typical to their time period, their gender, their nation, or the genre in which they are writing. Like the master craftsman teaching an apprentice by example, so too does each subsequent generation learn from and then evolve beyond the former, while all the while being constrained by the available resources. The artistry—as Brik and other formalists have argued—comes in the assembly of these resources.

To visualize this corpus as a network, then, and to interrogate my hypothesis of literary progression and influence, I converted the distance matrix described above into a long-form table with 11,195,716 rows and three columns. Each row captures a distance relationship between two books: the first cell contains one book, a "source," and the second cell another book, the "target." A third cell

contains the measured distance between the two. I reduced these data by removing all of the records in which the second book was published *before* the first: influence only works in one direction! This reduced the data from around 11 million records to a more manageable 5,548,275 records. The distance measures in this final data set ranged from 0.05946 to 107.44473, with a mean distance of 10.45770.* I then further reduced the data by calculating the standard deviation for the distances from every source book to all of the other books in the corpus.† I then removed those target books that were more than one standard deviation from the source book. This winnowing is done both for computational convenience and for network simplicity. The process has the effect of retaining only those books that are particularly close, or similar, to each other. In the initial distance matrix, every book is connected to every other book; however, at some distance the argument that two works are related or connected breaks down. After this culling, the number of edges or "connections" from one book to another was reduced to 165,770 book-to-book relationships; 5,382,505 weaker connections are ignored. Using custom scripts in R, these data, along with a separate table of metadata for each individual novel, were combined and converted to the "Graph Exchange XML Format" (GEXF). This file was then imported into the open-source network-analysis software package Gephi (Bastian, Hemann, and Jacomy 2009).

Networks (or "graphs," as they are frequently called) are constructed out of two primary elements: nodes and edges. For our purposes, nodes are individual books, and edges are the distances between them. In this data set, the edges are weighted using the distance measure calculated with the Euclidean metric. Nodes with smaller distances are more similar and more closely connected. When plotted, nodes with larger distances will spread out farther in the network diagram. Figure 9.2 offers a simplified example.

Gephi provides a number of layout options and analysis routines for network data. The layout algorithms provide methods for displaying the data, that is, methods for making the intricacies of the network most visible. With a large network such as this, generating and then plotting images that can be displayed, on a standard, book-size, page, are challenging. Despite this challenge, a few

* Some readers may find it useful to think of these distances in terms of a familiar measure of distance such as "inches" or "centimeters." The two closest books in the corpus are 0.05 inches apart, and the two that are farthest from each other are 107.4 inches; on average, books are about 10 inches apart.

† It may be easier to consider this process one book at a time. For *Moby Dick,* for example, I calculate the standard deviation from *Moby Dick* to all of the other books in the corpus that were published after *Moby Dick.* I then keep only those books that have a distance less than one standard deviation above the minimum distance.

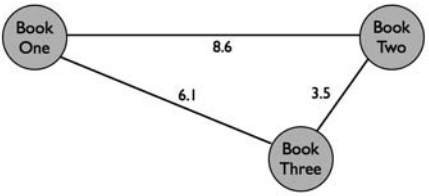

Figure 9.2. Example network graph

useful plots can be shown here. Figure 9.3 shows the entire network laid out using the Force Atlas 2 algorithm. The individual book nodes are the gray-scaled "dots," and the edges are the more visible arching lines.

Gephi provides an option that allows for coloring of the nodes based on the metadata contained in the node records. With the addition of this coloring or gray-scaling, several macrostructures can be made visible in this graph.* In figure 9.3, the book nodes have been colored according to the publication years of the books. The edges that are directed outward from these source nodes have been colored with the same shade of gray. The lighter-gray nodes and edges indicate works from the earlier part of the century; the darker nodes are later. The further back in time we go, the lighter the nodes become. This shading of the nodes by year reveals a clear time signature to the stylistic-thematic data. Beginning in the lighter, western, section of the graph, we move eastward through time. It is critical to bear in mind here that the novels are *not* being clustered in the network based on their publication dates; in fact, dates play no role whatsoever in determining how close the books are to each other or how they are laid out in the network visualization. Books are being pulled together (and pushed apart) based on the similarity of their computed stylistic and thematic distances from each other. The fact that they line up in a chronological manner is incidental, but rather extraordinary.† The 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

* Color versions of figures 9.3 and 9.4 can be found online at http://www.matthewjockers.net/macroanalysisbook/color-versions-of-figures-9–3-and-9–4/.

† I say "rather" because some amount of chronological organization is to be expected given that I have removed the possibility that a book in the future could influence a book in the past. Nevertheless, the possibility exists that a book from 1800 is most stylistically and thematically similar to a series of books published in the 1890s and that this book from 1800 will be situated in the network alongside these more similar works that are published ninety years later. This is, in fact, exactly what is observed for some books in the corpus.

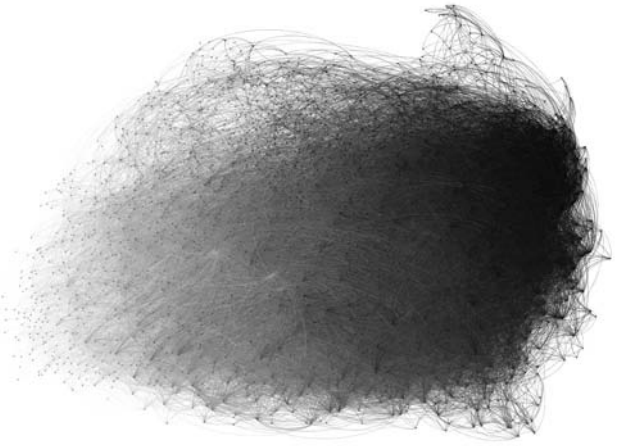

Figure 9.3. Nineteenth-century novel network with date shading

the frameworks for their themes are nearly, but not always, tethered in time. At this macro scale, style and theme are observed to evolve chronologically, and most books and authors in this network cluster into communities with their chronological peers. Not every book and not every author is a slave to his or her epoch; there are a few outliers who buck the trend. Before moving to a discussion of the outliers, however, a few more observations about the macro structures of this network are in order.

Figure 9.4 shows the same network layout reshaded according to author gender.* Male nodes and edges are colored lighter gray, and the female nodes are black. A clear boundary can be seen dividing the network into male and female regions. Works by female authors are more stylistically and thematically similar to each other, and they cluster together in the south and southeast portions of the main network. Males are drawn together in the north.

In both renderings of the network, we can see the presence of outliers: in figure 9.3 there are works from earlier in the century that cluster in portions of the network dominated by works from a later period, and in figure 9.4 there are male authors placed firmly in the more female-dominated regions of the graph, and vice versa for several female authors. Without a large screen and an interactive program in which to view this entire network, many of these individual subtleties and outliers are lost. Three slightly larger outlier "communities," however, are clearly visible at this scale. First, at the lower-right corner of the main network is a community of nodes extending outward and down—where Florida would be if this were a map of the United States. These nodes all belong

* Anonymous authors have been filtered out of the image.

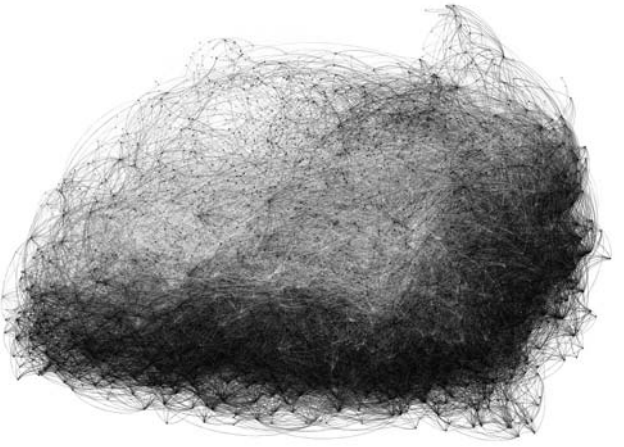

Figure 9.4. Nineteenth-century novel network with gender shading

to books authored by Margaret Oliphant. Remember here that the network has organized itself in this manner independent of any metadata about the books. The graphing software does not know who the authors are. This layout suggests not only that Oliphant's stylistic and thematic "signal" is unique, but that her signal is an extreme within the context of the major subcluster of works dominated by female authors. She is at once connected to the female section of the graph and an isolated peninsula. In other words, Oliphant's signal is unusual both within her gender and to the network as a whole.

Similar to this outlier cluster of works by Oliphant is a similar outcropping of works found farther west and to the north, approximately where Montana would be on a U.S. map. This cluster is made up of six books by Walter Scott alongside works by a series of other Scottish authors, including Robert Louis Stevenson and Henrietta Keddie.* Also present in this "highland" cluster are several works by George P. R. James. James was not Scottish, but he was indebted to Walter Scott. According to James N. MacKenzie (1992), George P. R. James sent his first novel (*Richelieu: A Tale of France*) to Walter Scott for review. It was only after receiving a positive reply from Scott that James found the confidence to send the manuscript off for publication. Scott's approval of James's novel, and

* Keddie, who wrote under the pseudonym Sarah Tytler, was known primarily as a writer of women's fiction. The work of hers that appears in this cluster, however, is atypical of her oeuvre. *Saint Mungs City: A Novel* presents a picture of the Scottish urban world and industrialization. None of her other works appears in this group, a fact that suggests there may be something special or unique about this particular book.

the novel's eventual appearance in this cluster of works similar to Scott's own, is suggestive of an entirely different sort of influence, the influence of endorsement. It is a tantalizing idea, and one that would require a closer reading both of James and of Scott to fully explicate.

The third and final outlier community is found very obviously in the northeast. Unlike the other two, this one is not so easily understood. Indeed, I can offer no unifying thread at all. Table 9.6 provides a listing of the fifteen works that make up this cluster. Aside from all being published within nine years of each other, I find nothing here but noise. Perhaps more knowledgeable scholars will see some link that I have missed.

Beyond these more obvious outlier clusters, there are any number of specific oddities and surprises. What, for example, does it mean that Maria Edgeworth's *Belinda* is mapped to a place in the network that puts her book twenty years ahead of its time? Why are all of Harriet Beecher Stowe's books firmly rooted in the male sections of the network and many of James Payn's in the female half? These are questions that I will not try to answer; they were chosen arbitrarily and are just several among many, and still we have barely sampled the diverse offerings of graph theory and network analysis: Ego networks, for example, can be calculated to explore a single book's sphere of influence. Node-centrality measures can provide a sense of a book's importance to and within the larger network. The Gephi software provides tools for calculating these and many other measures, and through such measures Gephi gives us the power to sift and rank the relative importance of one node versus another. Gephi's PageRank statistic, for example, is based on the algorithms developed by Google founders

Table 9.6. A cluster of books

Author	Title	Publication year
Bates, Emily	*George Vyvian*	1890
Beale, Anne	*Courtleroy*	1887
Davidson, Hugh	*The Green Hills by the Sea*	1887
Deccan, Hilary	*Light in the Offing*	1892
Fitzclarence, Wilhelmina	*Dorinda*	1889
Grant, James	*Colville of the Guards*	1885
Hake, Thomas	*In Letters of Gold*	1886
Harwood, John	*Sir Robert Shirley Bart*	1886
Hayward, Gertrude	*Dulcibel*	1890
Lambert, George	*The Power of Gold*	1886
Linton, Elizabeth	*Through the Long Night*	1889
Spender, Emily	*Until the Day Breaks*	1886
Spender, Lillian	*Mr. Nobody*	1884
Wilkins, William	*The Forbidden Sacrifice*	1893
Woollam, Wilfred	*All for Naught*	1890

Sergey Brin and Lawrence Page. It is designed as a tool for assigning "a numerical weighting to each element of a hyperlinked set of documents, . . . with the purpose of 'measuring' its relative importance within the set" (Wikipedia 2011b). When applied to this nineteenth-century corpus in which the links are measures of stylistic and thematic affinity, the algorithm points us first to Laurence Sterne's *Tristram Shandy,* next to George Gissing's novel *The Whirlpool,* and then to Benjamin Disraeli's *Venetia. Tristram Shandy* is a book frequently lauded as one of the highest achievements of the novel form, and, by all accounts, Gissing was one of the century's most accomplished realists. Disraeli's minor novel *Venetia* is harder to understand. Maybe its presence here is a sign that the method has failed, or perhaps it is a sign that close readers need to reevaluate *Venetia.* In short, these network data are rich—too rich, in fact, to take much further in these pages because they demand that we follow every macroscale observation with a full-circle return to careful, sustained, close reading. This is work for the future. With more than three thousand books, the observations we could make and the questions we might ask about the context in which a title appears are overwhelming. My purpose here has been to describe the landscape and offer a glimpse of the possibilities and a few of the provocations.

At the macro scale, we see evidence of time and gender influences on theme and style. By superimposing these two network snapshots in our minds, we can begin to imagine a larger context in which to read and study nineteenth-century literature. What is clear is that the books we have traditionally studied are not isolated books. The canonical greats are not even outliers; they are books that are similar to other books, similar to the many orphans of literary history that have been long forgotten in a continuum of stylistic and thematic change. Whether these orphans are worth fostering is an entirely different and more complicated question. In terms of their potential influence on those other works that we already know and care about, they are certainly worth our attention, and macroanalysis offers us a way of finding them in the haystack of literary history.