The emotional arcs of stories are dominated by six basic shapes

Andrew J. Reagan,[1] Lewis Mitchell,[2] Dilan Kiley,[1] Christopher M. Danforth,[1] and Peter Sheridan Dodds[1]

[1]*Department of Mathematics & Statistics, Vermont Complex Systems Center, Computational Story Lab, & the Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401*

[2]*School of Mathematical Sciences, The University of Adelaide, SA 5005 Australia*

(Dated: July 8, 2016)

Advances in computing power, natural language processing, and digitization of text now make it possible to study our a culture's evolution through its texts using a "big data" lens. Our ability to communicate relies in part upon a shared emotional experience, with stories often following distinct emotional trajectories, forming patterns that are meaningful to us. Here, by classifying the emotional arcs for a filtered subset of 1,737 stories from Project Gutenberg's fiction collection, we find a set of six core trajectories which form the building blocks of complex narratives. We strengthen our findings by separately applying optimization, linear decomposition, supervised learning, and unsupervised learning. For each of these six core emotional arcs, we examine the closest characteristic stories in publication today and find that particular emotional arcs enjoy greater success, as measured by downloads.

I. INTRODUCTION

The power of stories to transfer information and define our own existence has been shown time and again [1–5]. We are fundamentally driven to find and tell stories, likened to *Pan Narrans* or *Homo Narrativus*. Stories are encoded in art, language, and even in the mathematics of physics: We use equations to represent both simple and complicated functions that describe our observations of the real world. In science, we formalize the ideas that best fit our experience with principles such as Occam's Razor: The simplest story is the one we should trust. We also tend to prefer stories that fit into the molds which are familiar, and reject narratives that do not align with our experience [6].

We seek to better understand stories that are captured and shared in written form, a medium that since inception has radically changed how information flows [7]. Without evolved cues from tone, facial expression, or body language, written stories are forced to capture the entire transfer of experience on a page. A often integral part of a written story is the emotional experience that is evoked in the reader. Here, we use a simple, robust sentiment analysis tool to extract the reader-perceived emotional content of written stories as they unfold on the page.

We objectively test the theories of folkloristics [8, 9], specifically the commonality of core stories within societal boundaries [4, 10]. A major component of folkloristics is the study of society and culture through literary analysis. This is sometimes referred to as *narratology*, which at its core is "a series of events, real or fictional, presented to the reader or the listener", who further define narrative and plot [11]. In our present treatment, we consider the plot as the "backbone" of events that occur in a chronological sequence. We first find an analogous definition in Aristotle's theory of the three act plot structure: A central conflict emerges in act one, followed by two major turning points in acts two and three before concluding with a final resolution. While the plot captures the mechanics of a narrative and the structure encodes their delivery, in the present work we examine the emotional arc that is invoked through the words used. The emotional arc of a story does not give us direct information about the plot or the intended meaning of the story, but rather exists as part of the whole narrative. This distinction between the emotional arc and the plot of a story is one point of misunderstanding in other work [12]. Through the identification of motifs [13], narrative theories [14] allow us to analyze, interpret, describe, and compare stories across cultures and regions of the world [15]. We show that automated extraction of emotional arcs is not only possibly, but can test previous theories and provide new insights with the potential to quantify unobserved trends as the field transitions from data-scarce to data-rich [16, 17].

There have been various hand-coded attempts to enumerate and classify the core types of stories from their plots, including models that generalize broad categories and detailed classification systems. We consider a range of these theories in turn while noting that plot similarities do not necessitate a concordance of emotional arcs.

- Three plots: In his 1959 book, Foster-Harris contends that there are three basic patterns of plot (extending from the one central pattern of conflict): the happy ending, the unhappy ending, and the tragedy [18]. In these three versions, the outcome of the story hinges on the nature and fortune of a central character: virtuous, selfish, or struck by fate, respectively.

- Seven plots: Often espoused as early as elementary school in the United States, we have the notion that plots revolve around the conflict of an individual with either (1) him or herself, (2) nature, (3) another individual, (4) the environment, (5) technology, (6) the supernatural, or (7) a higher power [19].

- Seven plots: Representing over three decades of work, Christopher Booker's *The Seven Basic Plots: Why we tell stories* describes in great detail seven narrative structures: [20]
  - Overcoming the monster (e.g., *Beowulf*).
  - Rags to riches (e.g., *Cinderella*).
  - The quest (e.g., *King Solomons Mines*).
  - Voyage and return (e.g., *The Time Machine*).
  - Comedy (e.g., *A Midsummer Night's Dream*).
  - Tragedy (e.g., *Anna Karenina*).
  - Rebirth (e.g., *Beauty and the Beast*).

  In addition to these seven, Booker contends that the unhappy ending of all but the tragedy are also possible.
- Twenty plots: In *20 Master Plots*, Ronald Tobias proposes plots that include "quest", "underdog", "metamorphosis", "ascension", and "descension" [21].
- Thirty-six plots: In a translation by Lucille Ray, Georges Polti attempts to reconstruct the 36 plots that he posits Gozzi originally enumerated [22]. These are quite specific and include "rivalry of kinsmen", "all sacrificed for passion", both involuntary and voluntary "crimes of love" (with many more on this theme), "pursuit", and "falling prey to cruelty of misfortune".

The rejected master's thesis of Kurt Vonnegut—which he personally considered his greatest contribution—defines the emotional arc of a story on the "Beginning–End" and "Ill Fortune–Great Fortune" axes [23]. Vonnegut finds a remarkable similarity between Cinderella and the origin story of Christianity in the Old Testament (see Fig. S18), leading us to search for all such groupings. In a recorded lecture available on YouTube [24], Vonnegut asserted:

> "There is no reason why the simple shapes of stories can't be fed into computers, they are beautiful shapes."

We proceed as follows. We first introduce our methods in Section II, we then discuss the combined results of each method in Section III, and we present our conclusions in Section IV.

II. METHODS

A. Emotional arc construction

To generate emotional arcs, we analyze the sentiment of 10,000 word windows, which we slide through the text (see Fig. 1). We rate the emotional content of each window using our Hedonometer with the labMT dataset,

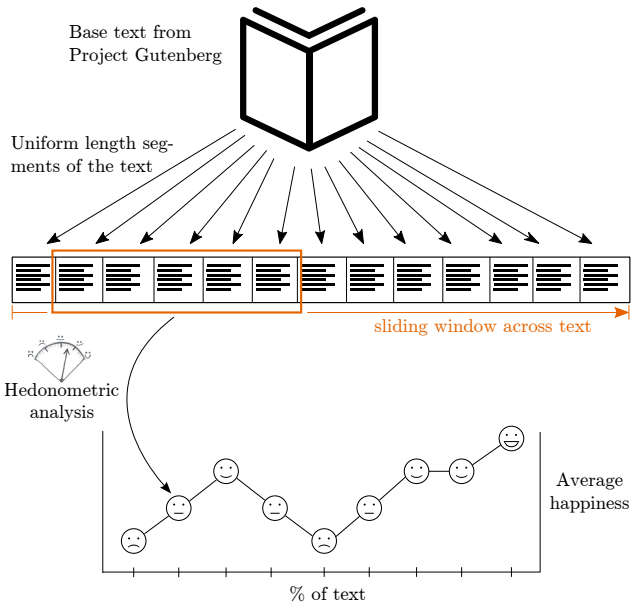


FIG. 1: Schematic of how we compute emotional arcs. The indicated uniform length segments (gap between samples) taken from the text form the sample with fixed window size set at $N_w = 10,000$ words. The segment length is thus $N_s = (N-(N_w+1))/n$ for N the length of the book in words, and n the number of points in the time series. Sliding this fixed size window through the book, we generate n sentiment scores with the Hedonometer, which comprise the emotional arc [27].

chosen for lexical coverage and its ability to generate of meaningful word shift graphs, using 10,000 words to generate meaningful sentiment scores [25, 26]. We emphasize that dictionary-based methods for sentiment analysis can perform worse than random on individual sentences [26], a misunderstanding of similar efforts [12]. In Fig. 2, we show the emotional arc of *Harry Potter and the Deathly Hallows*, the final book in the popular Harry Potter series by J.K. Rowling. While the plot of the book is nested and complicated, the emotional arc associated with each subnarrative is clearly visible. We analyze the emotional arcs corresponding to complete books, and to limit the conflation of multiple core emotional arcs we restrict our analysis to shorter books (by selecting a maximum number of words when building our filter). Further details of the emotional arc construction can be found in Appendix A.

B. Project Gutenberg Corpus

For a suitable corpus we draw on the freely available Project Gutenberg data set [29]. We apply rough filters to the entire collection in an attempt to obtain a set of 1,737 books that represent English works of fiction. Two slices of the data are shown in Fig. S1. We select for only English books, with total words between 10,000 and 200,000, with unique words between 1,000

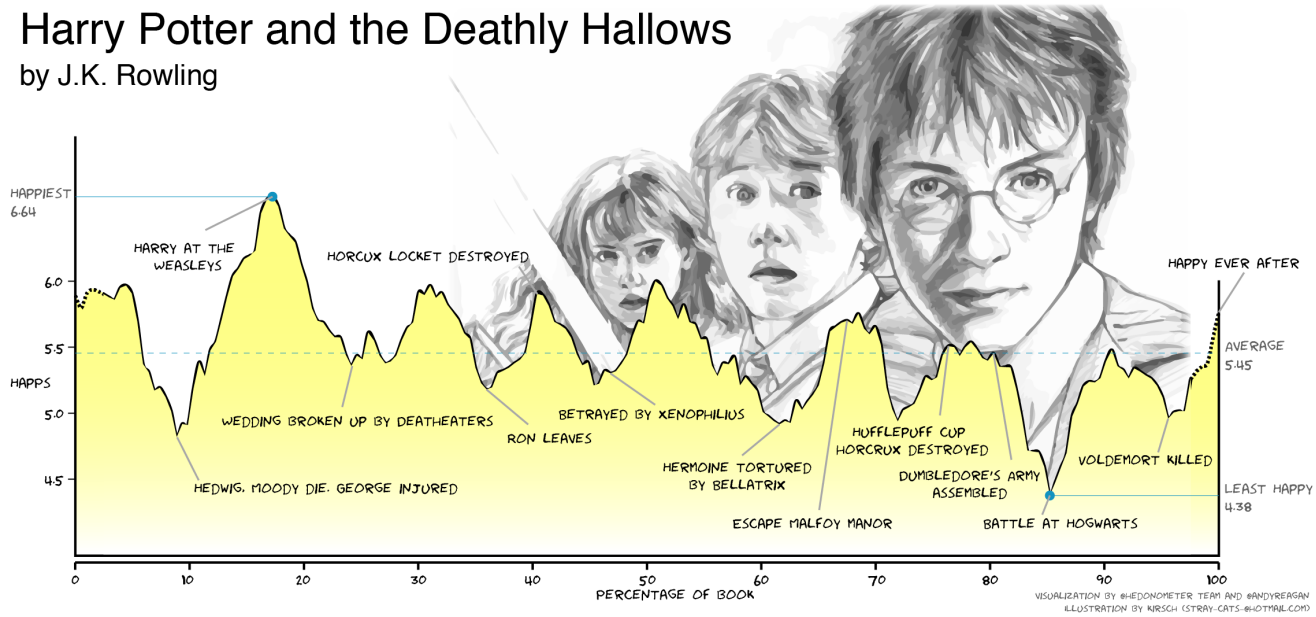


FIG. 2: Annotated emotional arc of *Harry Potter and the Deathly Hallows*, by J.K. Rowling, inspired by the illustration made by Medaris for The Why Files [28]. The entire seven book series can be classified as a "Rags to riches" and "Kill the monster" story, while the many sub plots and connections between them complicate the emotional arc of each individual book. The emotional arc shown here, captures the major highs and lows of the story, and should be familiar to any reader well acquainted with Harry Potter. Our method does not pick up emotional moments discussed briefly, perhaps in one paragraph or sentence (e.g., the first kiss of Harry and Ginny). We provide interactive visualizations of all Project Gutenberg books at `http://hedonometer.org/books/v3/1/` and a selection of classic and popular books at `http://hedonometer.org/books/v1/`.

and 18,000, and with more than 150 downloads from the Project Gutenberg website. Beyond these broad filters, we also remove dictionaries and transcriptions by the Human Genome Project (all three copies of the 24 books). We provide a list of the book ID's which are included for download in the online appendices at `http://compstorylab.org/share/papers/reagan2016b/`.

C. Principal Component Analysis (SVD)

We use the standard linear algebra technique Singular Value Decomposition (SVD) to find a decomposition of stories onto an orthogonal basis of emotional arcs. Starting with the sentiment time series for each book b_i as row i in the matrix A, we apply the SVD to find

$$A = U\Sigma V^T = WV^T, \tag{1}$$

where now U contains the projection of each sentiment time series onto each of the right singular vectors (rows of V^T, eigenvectors of A^TA), which have singular values given along the diagonal of Σ, with $W = U\Sigma$. Different intuitive interpretations of the matrices U, Σ, and V^T are useful in the various domains in which the SVD is applied; here, we focus on right singular vectors as an orthonormal basis for the sentiment time series in the rows of A, which we will refer to as the modes. We combine Σ and U into the single coefficient matrix W for clarity and convenience, such that W now represents the mode coefficients.

In Fig. 3 we show the leading 12 modes in both the weighted (dark) and un-weighted (lighter) representation. In total, the first 12 modes explain 80% and 94% of the variance from the mean centered and raw time series, respectively. The modes are from mean-centered emotional arcs such that the first SVD mode need not extract the average from the labMT scores nor the positivity bias present in language [25]. The coefficients for each mode within a single emotional arc are both positive and negative, so we need to consider both the modes and their negation. We can immediately recognize the familiar shapes of core emotional arcs in the first four modes, and compositions of these emotional arcs in modes 5 and 6. We observe "Rags to riches" (mode 1, positive), "Tragedy" or "Riches to rags" (mode 1, negative), Vonnegut's "Man in a hole" (mode 2, positive), "Icarus" (mode 2, negative), "Cinderella" (mode 3, positive), "Oedipus" (mode 3, negative). We choose to include modes 7–12 only for completeness, as these high frequency modes have little contribution to variance and do not align with core emotional arc archetypes (more below).

We emphasize that by definition of the SVD, the mode coefficients in W can be either positive and negative, such that the modes themselves explain variance with both the positive and negative version. In the right panels of each

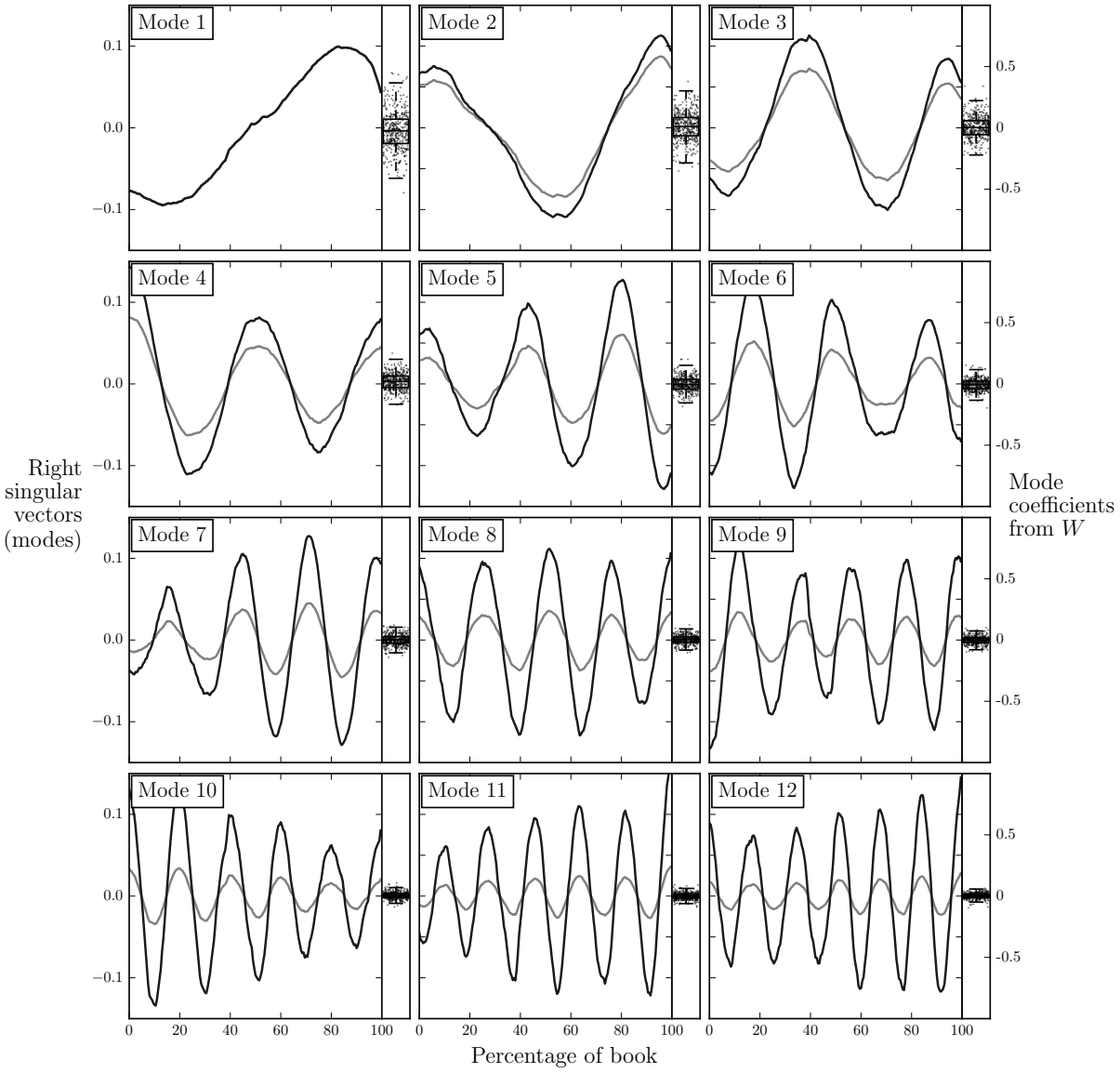


FIG. 3: Top 12 modes from the Singular Value Decomposition of 1,737 Project Gutenberg books. We show in a lighter color modes weighted by their corresponding singular value, where we have scaled the matrix Σ such that the first entry is 1 for comparison (for reference, the largest singular value is 27.3). The mode coefficients normalized for each book are shown in the right panel accompanying each mode, in the range -1 to 1, with the "Tukey" box plot.

mode in Fig. 3 we project the 1,737 stories onto each of first six modes and show the resulting coefficients. While none are far from 0 (as would be expected), mode 1 has a mean slightly above 0 and both modes 3 and 4 have means slightly below 0. To sort the books by their coefficient for each mode, we normalize the coefficients within each book in the rows of W to sum to 1, accounting for books with higher total energy, and these are the coefficients shown in the right panels of each mode in Fig. 3. In Appendix B, we provide supporting, intuitive details of the SVD method, as well as example emotional arc reconstruction using the modes (see Figs. S3–S5). As expected, less than 10 modes are enough to reconstruct the emotional arc to a degree of accuracy visible to the eye.

We show labeled examples of the emotional arcs closest to the top 6 modes in Figs. 4 and S2. We present both the positive and negative modes, and the stories closest to each by sorting on the coefficient for that mode. For the positive stories, we sort in ascending order, and vice versa. Mode 1, which encompasses both the "Rags to riches" and "Tragedy" emotional arcs, captures 30% of the variance of the entire space. We examine the closest stories to both sides of modes 1–3, and direct the reader to Fig. S2 for more details on the higher order modes. The two stories that have the most support from the "Rags to

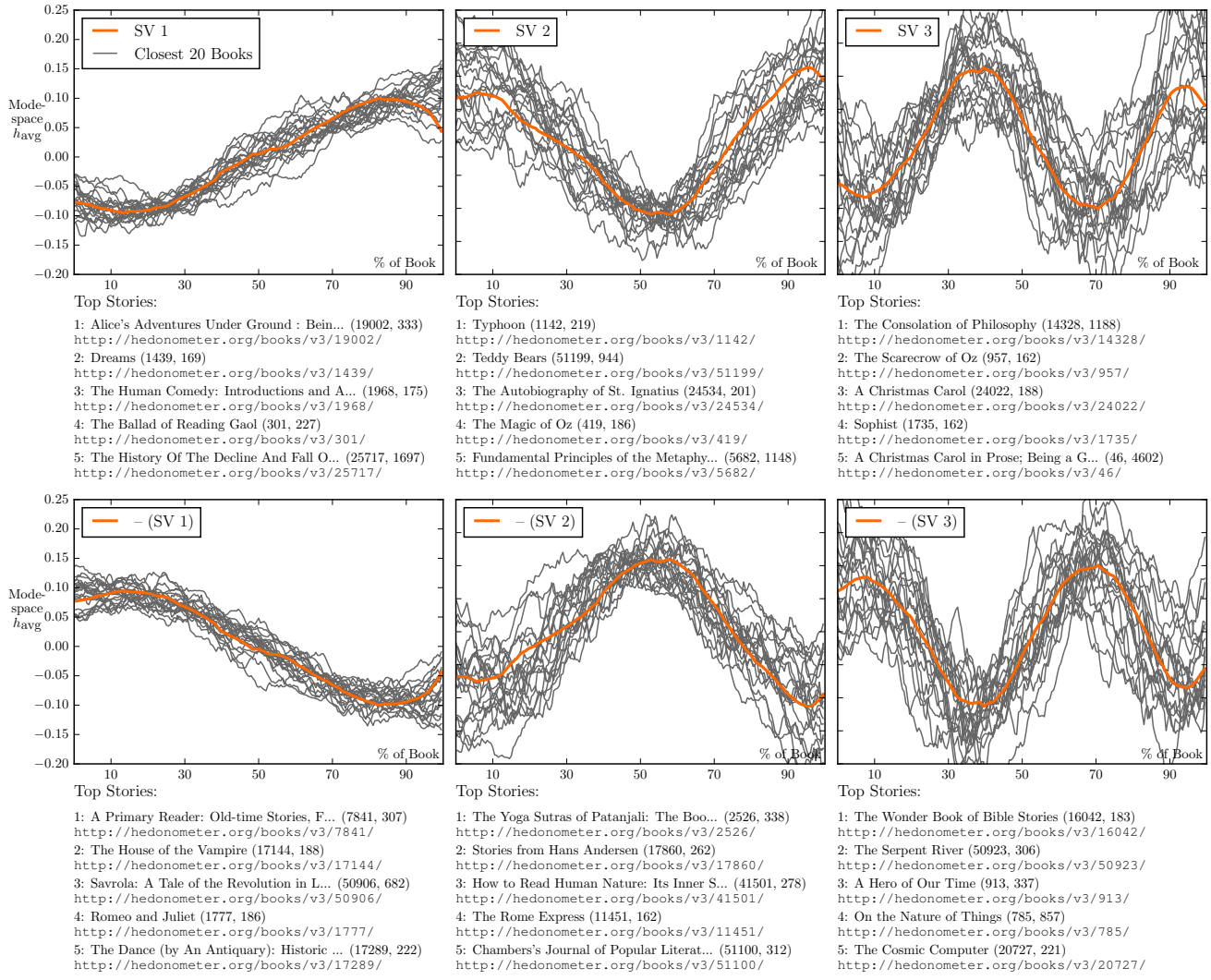


FIG. 4: First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W. In parenthesis for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on `http://hedonometer.org` which enables detailed exploration of the emotional arc for the story.

riches" mode are *Alice's Adventures Under Ground* and *Dreams*. Among the most categorical tragedies we find *A Primary Reader* and *The House of the Vampire*. The top 5 also includes perhaps the most famous tragedy: *Romeo and Juliet* by William Shakespeare. Mode 2 is the "Man in a hole" emotional arc, and we find the stories which most closely follow this path to be *Typhoon* and *Teddy Bears*. The negation of mode 2 most closely resembles the emotional arc of the "Icarus" narrative. For this emotional arc, the most characteristic stories are *The Yoga Sutras of Patanjali* and *Stories from Hans Andersen*. Mode 3 is the "Cinderella" emotional arc, and includes *The Consolation of Philosophy* and *The Scarecrow of Oz*. The negation of Mode 3, which we refer to as "Oedipus", is found most characteristically in *The Wonder Book of Bible Stories*, *The Serpent River*, and *A Hero of Our Time*. We also note that the spread of the stories from their core mode increases strongly for the higher modes.

III. RESULTS

We obtain a collection of 1,737 books that are mostly, but not all, fictional stories by using metadata from Project Gutenberg to construct a rough filter. Using principal component analysis, we find broad support for six emotional arcs:

Mode	Mode Arc	N_m	N_m/N	DL Median ▼	DL Mean ▽	DL Variance	% > Average	Download Distribution
SV 1		267	15.4%	289.0	638.0	2176764	20.6%	
-SV 1		440	25.4%	337.5	633.6	907943	25.0%	
SV 2		219	12.7%	327.0	652.3	1122421	21.9%	
-SV 2		167	9.7%	297.0	540.2	554142	16.8%	
SV 3		104	6.0%	298.0	896.3	7829052	22.1%	
-SV 3		109	6.3%	303.0	803.9	2839614	26.6%	
SV 4		108	6.2%	311.5	823.5	2728083	26.9%	
-SV 4		47	2.7%	286.0	790.6	1637200	19.1%	
SV 5		48	2.8%	280.0	397.1	146597	8.3%	
-SV 5		44	2.5%	280.5	452.0	188580	13.6%	

FIG. 5: Download statistics for stories whose SVD Modes comprise more than 2.5% of books, for N the total number of books and N_m the number corresponding to the particular mode. Modes *SV 3* through *-SV 4* (both polarities of modes 3 and 4) exhibit a higher average number of downloads and more variance than the others. Mode arcs are rows of V^T and the download distribution is show in $\log_{10}$ space from 150 to 30,000 downloads.

- "Rags to riches" (rise).
- "Tragedy", or "Riches to rags" (fall).
- "Man in a hole" (fall–rise).
- "Icarus" (rise–fall).
- "Cinderella" (rise–fall–rise).
- "Oedipus" (fall–rise–fall).

Importantly, we also find these emotional arcs using two other methods: As clusters in a hierarchical clustering using Ward's algorithm and as clusters using unsupervised machine learning.

We again find the first four of these six arcs appearing among the eight most different clusters from a hierarchical clustering (Fig. S9). The clustering method groups stories with a "Man in a hole" emotional arc for a range of different variances, separate from the other arcs, in total these clusters (Panels E, F, and G of Fig. S9) account for 23% of the Gutenberg corpus. The remainder of the stories have emotional arcs that are clustered among the "Icarus" arc, "Rags to riches" arc, and the "Tragedy" arc. A detailed analysis of the results from hierarchical clustering can be found in Appendix C, and this result agrees with other attempts that use only hierarchical clustering [12].

Finally, we apply Kohonen's Self-Organizing Map (SOM) and find these core arcs from unsupervised machine learning on the emotional arcs (Fig. S11 and Appendix D). On the two dimensional component plane, the prescribed network topology, we find three spatially coherent groups. These spatial groups are comprised of stories with core emotional arcs of differing variance. Grouping together the nine most active nodes, we find the "Rags to riches" arc accounts for 19% of the corpus, the "Tragedy" arc with 6%, the "Icarus" arc with 12% across three nodes, and the "Man in a hole" arc with 3% on one node.

There are many possible emotional arcs in the space that we consider. To demonstrate that these specific arcs are uniquely compelling as stories written by and for *homo narrativus*, we compare the emotional arcs of "word salad" versions for each book with randomly permuted word locations. An example of the emotional arc for a single book is shown in Fig. S12, along with 10 word salad versions. We re-run the SVD, hierarchical clustering, and unsupervised machine learning on the Gutenberg Corpus with the word salad version of each book and verify that the emotional arcs of real stories are not simply an artifact. The singular value spectrum from the SVD is flatter, with higher-frequency modes appearing more quickly, and in total representing 45% of the total variance present in real stories (see Figs. S14 and S13). Hierarchical clustering generates less distinct clusters with lower linkage cost for the emotional arcs from word salad books, and the winning node vectors on a self-organizing map lack coherent structure (see Figs. S15 and S17 in Appendix E).

To examine how the emotional trajectory impacts success, in Fig. 5 we examine the downloads for all of the books that are most similar to each SVD mode (for additional modes, see Fig. S19 in Appendix F). We find that the first four modes, which contain the greatest total number of books, are not the most popular. Both polarities of modes 3 and 4 have markedly higher downloads, and somewhat higher variance. The success of the stories underlying these emotional arcs shows that the emotional experience of readers strongly affects how stories are shared. We find the "Cinderella" (SV 3), "Oedipus" (-SV 3), two sequential "Man in a hole" arcs (SV 4), and "Cinderella" with a tragic ending (-SV 4) are the most successful emotional arcs.

IV. CONCLUSION

Using three distinct methods, we have demonstrated that there is strong support for six core emotional arcs. Our methodology brings to bear a cross section of data science tools with a knowledge of the potential issues that

each present. By considering the results of each tool in support of each other we are able to confirm our findings. We have also shown that consideration of the emotional arc for a given story is important for the success of that story.

Our approach could applied in the opposite direction: namely by beginning with the emotional arc and aiding in the automatic generation of compelling stories [30]. The emotional arcs of stories may be useful to aid in constructing arguments [31] and teaching common sense to artificial intelligence systems [32].

Extensions of our analysis that use a more curated selection of full-text fiction can answer more detailed questions about which stories are the most popular throughout time, and across regions [10]. Automatic extraction of character networks would allow a more detailed analysis of plot structure for the Project Gutenberg corpus used here [11, 33, 34]. Bridging the gap between the full text stories [35] and systems that analyze plot sequences will allow such systems to undertake studies of this scale [36]. Place could also be used to consider separate character networks through time, and to help build an analog to Randall Munroe's Movie Narrative Charts [37].

We are producing data at an ever increasing rate, including rich sources of stories written to entertain and share knowledge, from books to television series to news. Of profound scientific interest will be the degree to which we can eventually understand the full landscape of human stories, and data driven approaches will play a crucial role.

PSD and CMD acknowledge support from NSF Big Data Grant #1447634.

[1] T. Pratchett, I. Stewart, and J. Cohen. *The Science of Discworld II: The Globe*. Ebury Press, London, UK, 2003.

[2] J. Campbell. *The Hero with a Thousand Faces*. New World Library, California, third edition, 2008.

[3] J. Gottschall. *The Storytelling Animal: How Stories Make Us Human*. Mariner Books, New York, NY, 2013.

[4] S. Cave. The 4 stories we tell ourselves about death. `http://www.ted.com/talks/stephen_cave_the_4_stories_we_tell_ourselves_about_death`, Jul 2013.

[5] P. S. Dodds. Homo Narrativus and the trouble with fame. Nautilus Magazine, 2013. http://nautil.us/issue/5/fame/homo-narrativus-and-the-trouble-with-fame.

[6] R. S. Nickerson. Confirmation Bias; A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220, 1998.

[7] J. Gleick. *The Information: A History, A Theory, A Flood*. Pantheon, New York, 2011.

[8] V. Propp. *Morphology of the Folktale. 1928*. Texas University Press, Texas, 1968.

[9] M. R. MacDonald. *Storytellers Sourcebook: A Subject, Title, and Motif Index to Folklore Collections for Children*. Gale Group, Michigan, 1982.

[10] S. G. da Silva and J. J. Tehrani. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science*, 3(1), 2016.

[11] S. Min and J. Park. Narrative as a complex network: A study of Victor Hugo's les misérables. In *Proceedings of HCI Korea*, 2016.

[12] M. Jockers. A novel method for detecting plot. `http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/`, June 2014.

[13] A. Dundes. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, pages 195–202, 1997.

[14] S. K. Dolby. *Literary Folkloristics and the Personal Narrative*. Trickster Press, Indiana, 2008.

[15] H.-J. Uther. *The Types of International Folktales. A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson. Part I. Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction (FF Communications, 284)*. Finnish Academy of Science and Letters, Helsinki, Finland, 2011.

[16] M. G. Kirschenbaum. The remaking of reading: Data mining and the digital humanities. In *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Maryland*, 2007.

[17] F. Moretti. *Distant Reading*. Verso, New York, 2013.

[18] W. F. Harris. *The basic patterns of plot*. University of Oklahoma Press, Oklahoma, 1959.

[19] H. P. Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, Massachusetts, 2008.

[20] C. Booker. *The Seven Basic Plots: Why We Tell Stories*. Bloomsbury Academic, New York, 2006.

[21] R. B. Tobias. *20 Master Plots: And How to Build Them*. Writer's Digest Books, Ohio, 1993.

[22] G. Polti. *The Thirty-Six Dramatic Situations*. James Knapp Reeve, Ohio, 1921.

[23] K. Vonnegut. *Palm Sunday*. RosettaBooks LLC, New York, 1981.

[24] K. Vonnegut. Shapes of stories. https://www.youtube.com/watch?v=oP3c1h8v2ZQ, 1995.

[25] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth. Human language reveals a universal positivity bias. *PNAS*, 112(8):2389–2394, 2015.

[26] A. Reagan, B. Tivnan, J. R. Williams, C. M. Danforth, and P. S. Dodds. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. Preprint available at https://arxiv.org/abs/1512.00531, 2015.

[27] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):e26752, 12 2011.

[28] D. J. Tenenbaum, K. Barrett, S. Medaris, and T. Devitt. In 10 languages, happy words beat sad

ones. http://whyfiles.org/2015/in-10-languages-happy-words-beat-sad-ones/, February 2015.
[29] Various. Project Gutenberg.
[30] B. Li, S. Lee-Urban, G. Johnston, and M. Riedl. Story generation with crowdsourced plot graphs. In *AAAI*, 2013.
[31] F. J. Bex and T. J. Bench-Capon. Persuasive stories for multi-agent argumentation. In *AAAI Fall Symposium: Computational Models of Narrative*, volume 10, page 04, 2010.
[32] M. O. Riedl and B. Harrison. Using stories to teach human values to artificial agents. 2015.
[33] X. Bost, V. Labatut, and G. Linarès. Narrative smoothing: dynamic conversational network for the analysis of tv series plots, 2016.
[34] S. D. Prado, S. R. Dahmen, A. L. C. Bazzan, P. M. Carron, and R. Kenna. Temporal network analysis of literary texts, 2016.
[35] A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer, Berlin, Germany, 2012.
[36] P. H. Winston. The strong story hypothesis and the directed perception hypothesis. 2011.
[37] R. Munroe. Movie narrative charts. http://xkcd.com/657/, 11 2009.
[38] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Appendix A: Emotional Arc Construction

To generate emotional arcs, we consider many different approaches with the goal of generating time series that meaningfully reflect the narrative sentiment. In general, we proceed as described in Fig. 1 and consider a method of breaking up the text as having three (interdependent) parameter choices for a sliding window:

1. Length of the desired sample text.
2. Breakpoint between samples.
3. Overlap of each sample.

These methods vary between rating individual words with no overlap to rating the entire text. To make our choice, we consider competing two objectives of time series generation: meaningfulness of sentiment scores and increased temporal resolution of time series. For the most accurate sentiment scores, we can use the entire book. The highest temporal resolution is possible with a sliding window of length 1, generating time series that have potentially as many data points as words in the book.

Since our goal is not only the generation of time series, but the comparison of time series across texts, we consider the additional objective of consistency. We seek time series which are consistent both in the accuracy of the time series, as well as consistent in the length of the resulting time series. Again these goals are orthogonal, and we note that our choice here can be tuned to test the sensitivity.

We normalize the length of emotional arcs for books of different length (while using a fixed window size) by varying the amount that the window needs to move. To make a time series of length l from a book with N words, we fix the sample length at k and set the overlap of samples to

$$(N-k-1)/l$$

words. This guarantees that we have temporal resolution l and sample length k for any $N > k+l$. We do not consider books with $N \leq k+l$ words.

To generate a sentiment score as in Fig. 1, we use a dictionary based approach for transparency and understanding of sentiment. We select the LabMT dictionary for robust performance over many corpora and best coverage of word usage. In particular, we determine a sample T's average happiness using the equation:

$$h_{\text{avg}}(T) = \frac{\sum_{i=1}^{N} h_{\text{avg}}(w_i) \cdot f_i(T)}{\sum_{i=1}^{N} f_i(T)} = \sum_{i=1}^{N} h_{\text{avg}}(w_i) \cdot p_i(T), \tag{A1}$$

where we denote each of the N words in a given dictionary as w_i, word sentiment scores as $h_{\text{avg}}(w_i)$, word frequency as $f_i(T)$, and normalized frequency of w_i in T as $p_i(T) = f_i(T)/\sum_{i=1}^{N} f_i(T)$.

We note here for the general case, and additionally specify with the details of each method, that for each emotional arc we subtract the mean before computing the distance or clustering. To compute the distance between two emotional arcs, we use the city block distance metric:

$$D(b_i, b_j) = l^{-1} \sum_{t=1}^{l} |b_i(t) - b_j(t)|. \tag{A2}$$

Our null set of emotional arc time series is generated by randomly shuffling the words of each book that we consider. Other variations on generating this null set include sampling from a phrase-level parse of the book with a Markov process, using continuous space random walks directly, or shuffling on sentences.

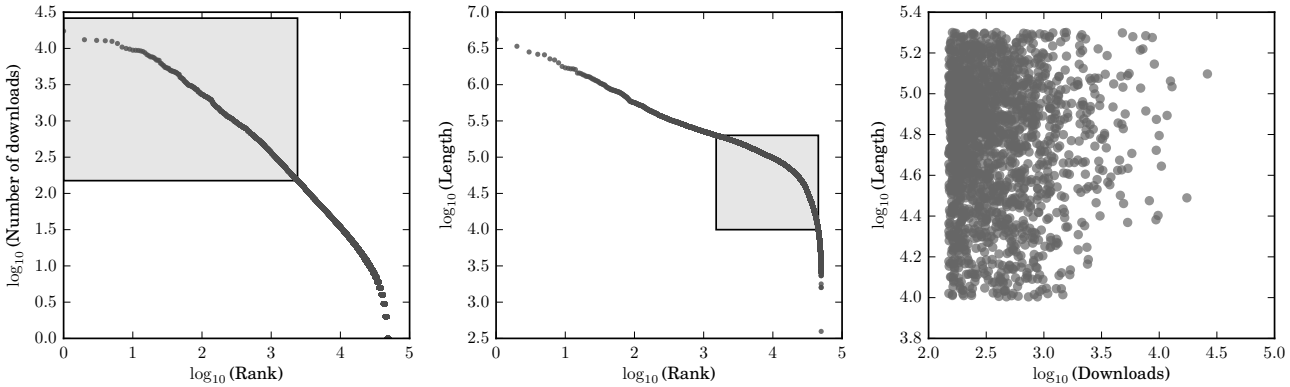


FIG. S1: Rank-frequency distributions of book downloads and length in the Gutenberg corpus: (A) downloads, (B) book length in words, and (C) both downloads and length. We filter by both number of downloads and book length to select for fiction books, with the filters shown as grey boxes in Panels A and B. In Panel C, we plot each of the resulting 1,748 books in download-length space.

Appendix B: Principal Component Analysis (SVD)

In this section we provide (1) more results from the SVD analysis and (2) a more in-depth, intuitive explanation of the method.

First we have consider modes 4–6 and their closest stories in Fig. S2.

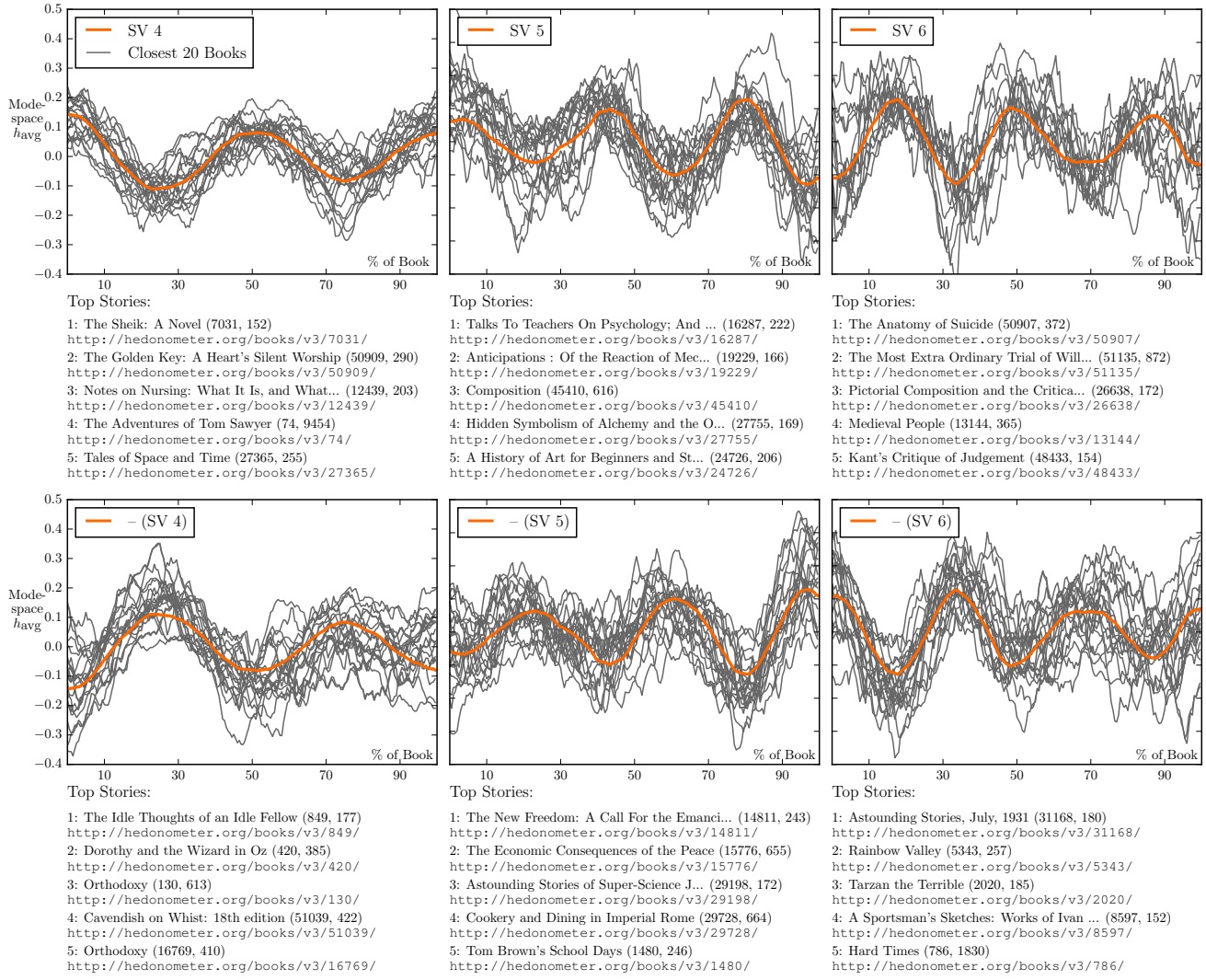


FIG. S2: SVD modes 4–6 (and their negation) with closest stories. First 3 SVD modes and their negation with the closest stories to each. Again, to show the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. Shown in parenthesis for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on `http://hedonometer.org` which enables detailed exploration of the emotional arc for the story.

In an effort to develop a better intuition for the results of the principal component analysis by way of SVD, we plot Eq. 1 along with representations of the matrices in Fig. S3.

Further, we considered in Eq. 1 the mode coefficient in the matrix W, and in Fig S4 we plot the second line of the equation with W:

With A written as $W \cdot V^T$, the coefficients for each mode (row of V^T) for a book i are given as the rows of W. To reconstruct the emotional arc of book i, using mode j from V^T, we simply multiply $W[i,j] \cdot V^T[j,:]$. Shown below in Fig. S5, we built the emotional arc for an example story using only the first mode through the first 12 modes.

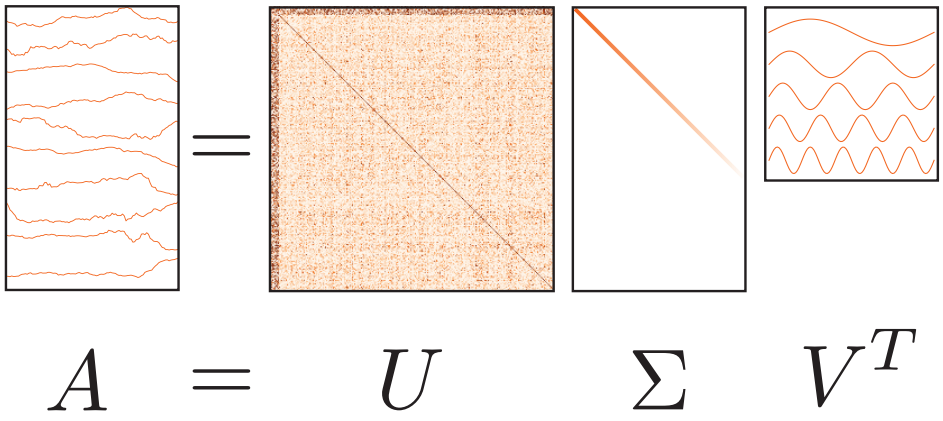


FIG. S3: Schematic of the Singular Value Decomposition applied to emotional arcs of Project Gutenberg books. Shown in A are 10 randomly chosen emotional arcs, in U a "spy" of the matrix, in Σ the decreasing singular values, and in V^T sinusoidal modes. We emphasize that this representation is purely for intuition, as only U is a image of the actual matrix, and A has only 10 of the 1,737 books.

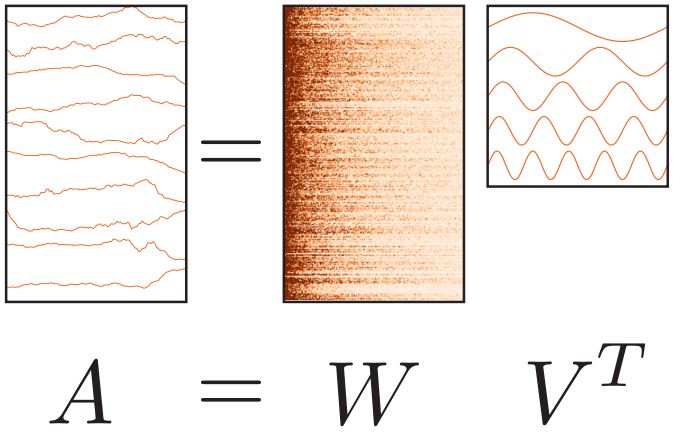


FIG. S4: Schematic of the Singular Value Decomposition applied to emotional arcs of Project Gutenberg books, with $W = U\Sigma$ containing the mode coefficients. Again shown in A are 10 randomly chosen emotional arcs, in W a "spy" of the matrix used in the analysis, and in V^T representative sinusoidal modes.

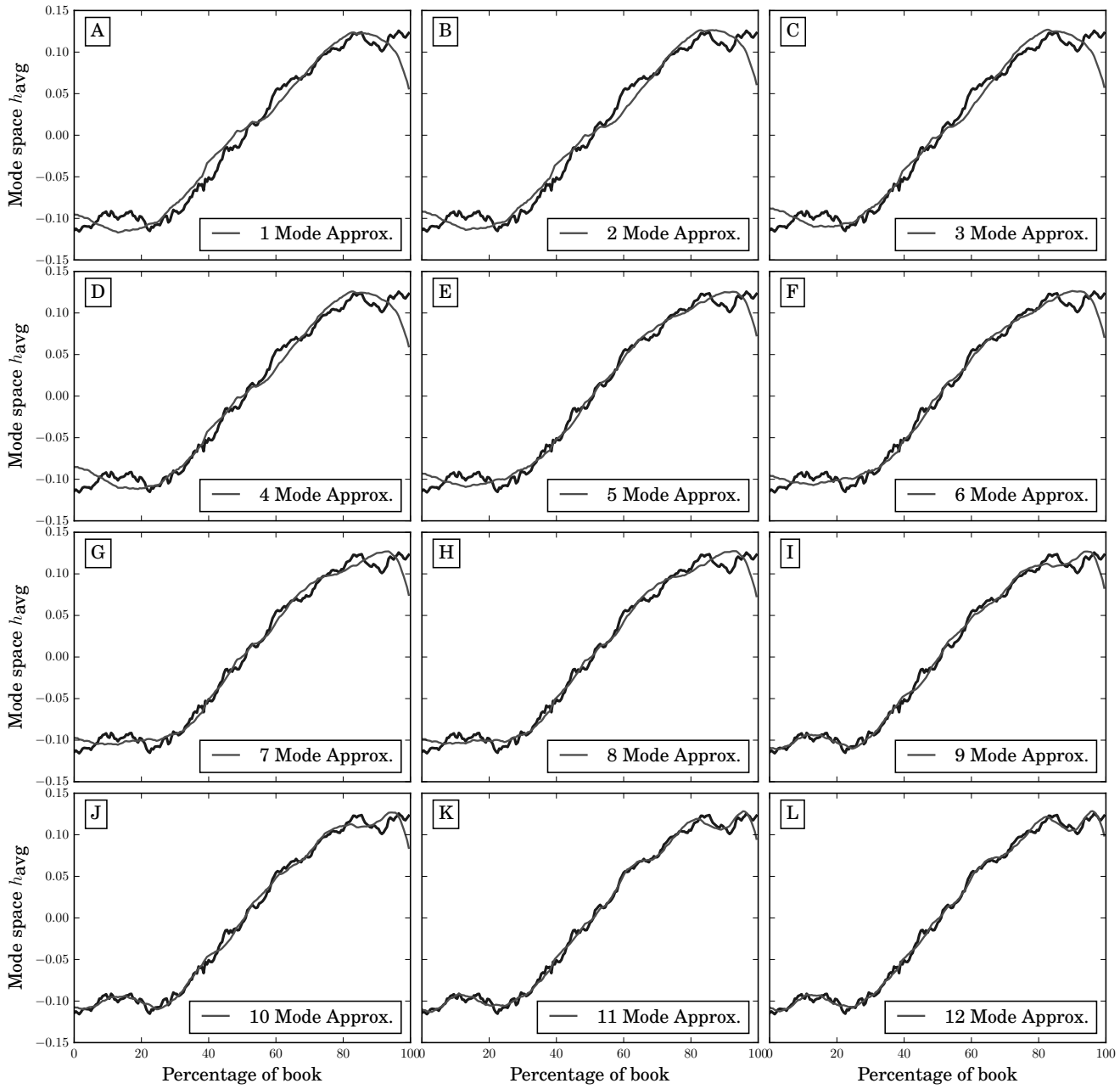


FIG. S5: Reconstruction of the emotional arc from *Alice's Adventures Under Ground*, by Lewis Carroll. The addition of more modes from the SVD more closely reconstructs the detailed emotional arc. This book is well represented by the first mode alone, with only minor corrections from modes 2-11, as we should expect for a book whose emotional arc so closely resembles the "Rags to Riches" arc.

Appendix C: Hierarchical Clustering

We use Ward's method to generate a hierarchical clustering of stories, which proceeds by minimizing variance between clusters of books [38]. We again use the mean-centered books and the distance function given in Eq. A2 to generate the distance matrix.

We show a dendrogram of the 60 clusters with highest linkage cost in Fig. S6. A characteristic book from each cluster is shown on the right by sorting the books within each cluster by the total distance to other books in the cluster (e.g., considering each intra-cluster collection as a fully connected weighted network, we take the most central node), and in parenthesis the number of books in that cluster. In other words, we label each cluster by considering the network centrality of the fully connected cluster with edges weighted by the distance between stories. By cutting the dendrogram in Fig. S6 at various costs we are able to extract clusters of the desired granularity. For the cuts labeled 0, 1, and 2, we show these clusters in Figs. S7, S8, and S9.

The final linkage in the hierarchical clustering combines the two clusters in Fig. S7 of size 1203 (Panel A, Threshold 8000 Cluster 1) and size 548 (Panel B, Threshold 800 Cluster 2). Cluster 1 is much larger, and the main difference between these clusters appears to be their variance. Nevertheless, we are able to see from the cluster average emotional arc (shown in orange) that Cluster 1 most closely follows a "Tragedy with the happy ending" and Cluster 2 follows the "Man in a hole" story. The top stories in Cluster 1 are *Piper in the Woods* and *Butterfly 9*, in the Figure they are shown with their ID in the Project Gutenberg database. The most central stories for Cluster 2 are *Asmodeus; or, The Devil on Two Sticks* and *Ghost Stories of an Antiquary*.

Going through the linkage procedure in reverse, we see in Fig. S8 that Threshold 8000 Cluster 1 is the linkage of Threshold 3850 Cluster 1 and Threshold 3850 Cluster 3. We see that what was a "Tragedy with the happy ending" splits into a tragedy in Cluster 1 (with perhaps a small uptick) and a mostly flat Cluster 3. Threshold 8000 Cluster 2 is the linkage of Threshold 3850 Cluster 2 and Threshold 3850 Cluster 4, Panels B and D in Fig. S8, respectively. Cluster 2 here again shows the most variance, with what mostly closely resembles a "Romance" story shape. Cluster 4, on average, is the "man in the hole" story type.

Finally, in Fig. S9 we see the 8 most different clusters as a collection of a familiar story types. In Panel A, Threshold 2050 Cluster 1 we have the "Romance" story. In Panel C, we have the "Rags to riches" (Cinderella) story. Panels E, F, and G are all "Man in a hole" stories of varying strength. In Panel H we find a large number of stories (360) that fall squarely into the "Tragedy" story type.

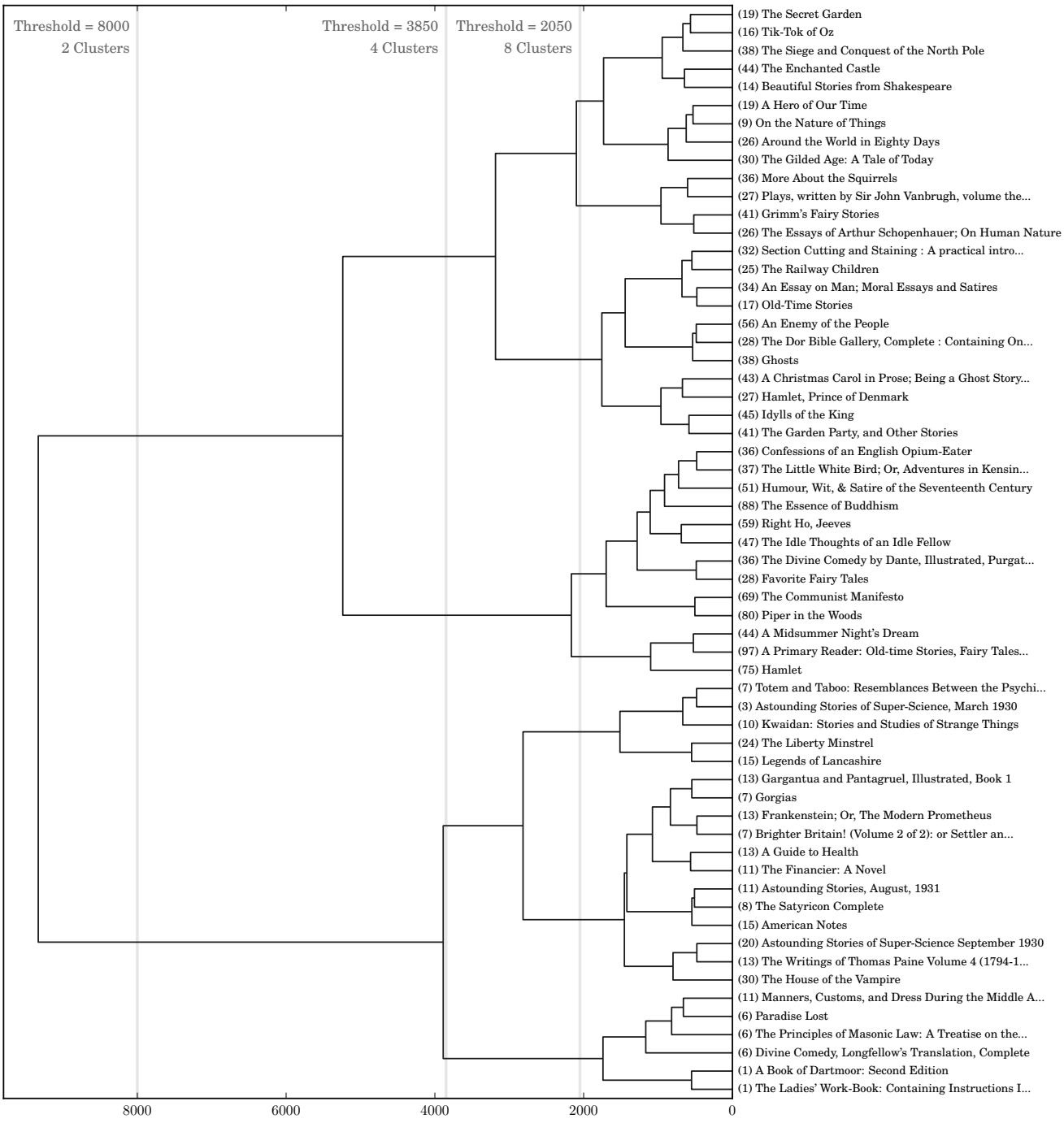


FIG. S6: Ward clustering.

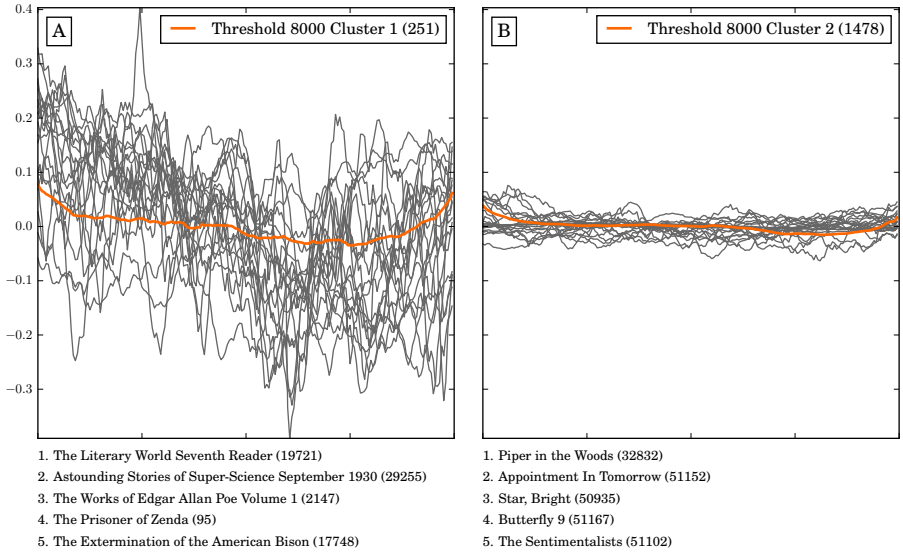
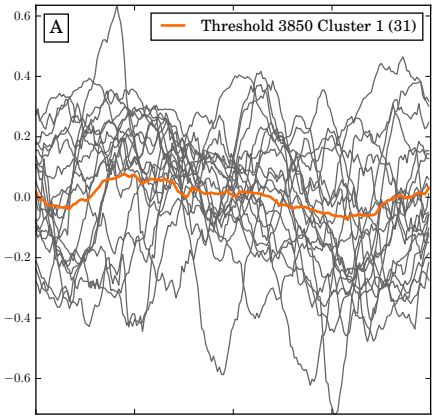
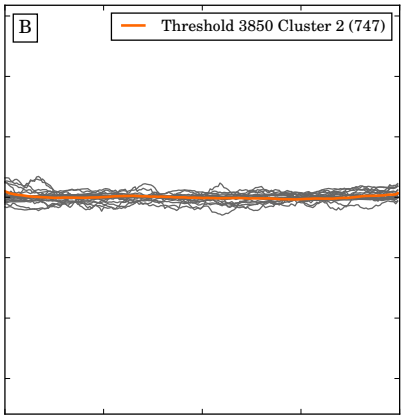


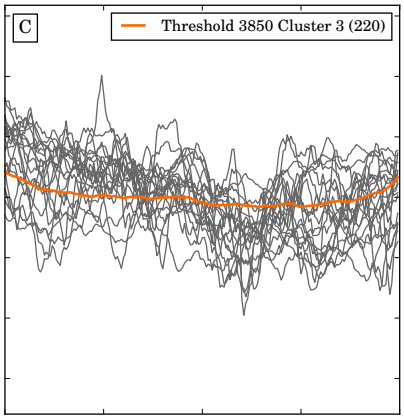
FIG. S7: Ward cluster number 1, showing the largest 2 clusters.



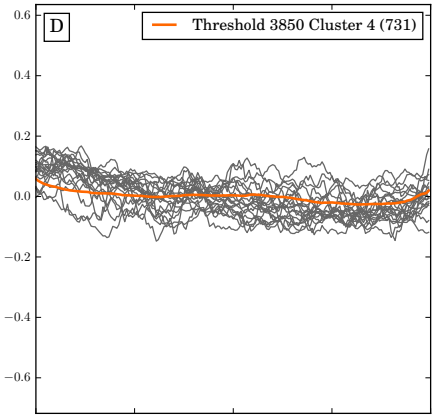
1. A General Introduction to Psychoanalysis (38219)
2. Paradise Lost (20)
3. Second Treatise of Government (7370)
4. Paradise Lost (26)
5. The Divine Comedy by Dante, Illustrated (8800)



1. Piper in the Woods (32832)
2. Butterfly 9 (51167)
3. Appointment In Tomorrow (51152)
4. Transfer Point (51115)
5. Euthyphro (1642)

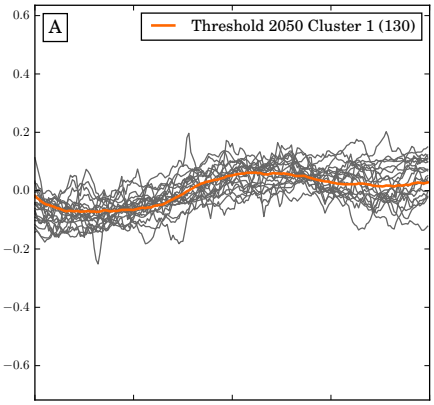


1. The Literary World Seventh Reader (19721)
2. Astounding Stories of Super-Science September 1930 (29255)
3. The Extermination of the American Bison (17748)
4. The Works of Edgar Allan Poe Volume 1 (2147)
5. The Prisoner of Zenda (95)

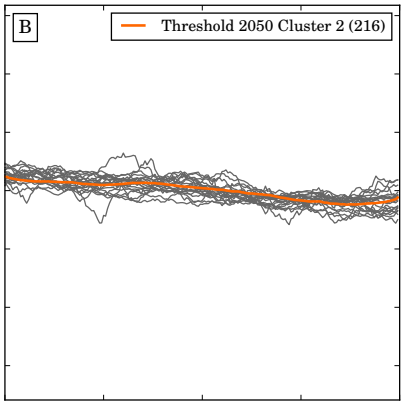


1. An Enemy of the People (2446)
2. Ghosts (8121)
3. Othello (2267)
4. Pierre and Jean (3804)
5. All Things Considered (11505)

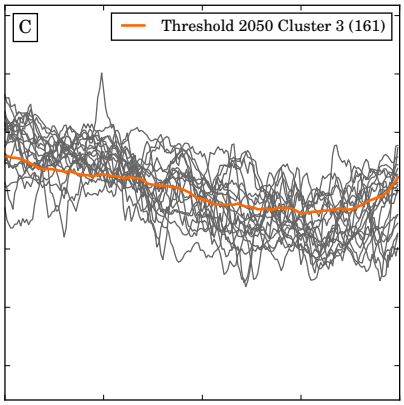
FIG. S8: Ward cluster number 2, showing the 4 largest clusters.



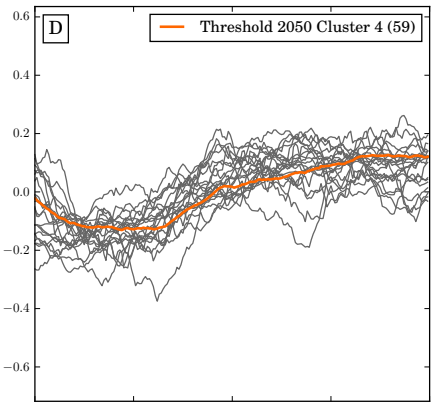
1. Woodcraft and Camping (34607)
2. The Inspector-General (3735)
3. The Governess; Or, The Little Female Academy (1905)
4. The Essays of Arthur Schopenhauer; On Human Nature (10739)
5. The Lost Princess of Oz (959)



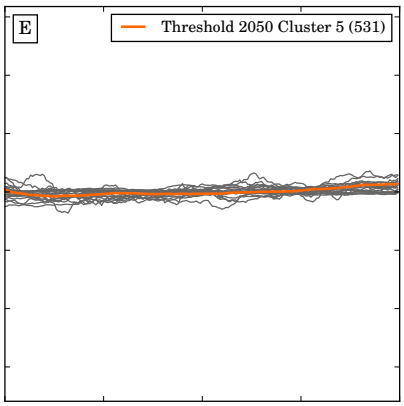
1. The Lifted Veil (2165)
2. Shakespeare's Sonnets (1041)
3. A Primary Reader: Old-time Stories, Fairy Tales... (7841)
4. The Adventures of Buster Bear (22816)
5. The Clouds (2562)



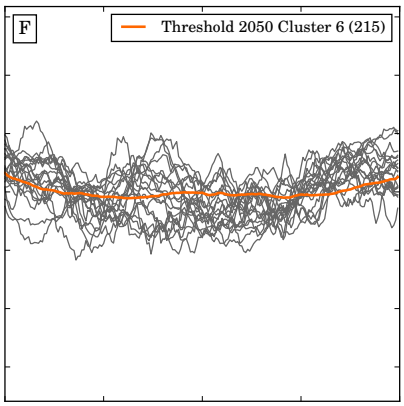
1. The House of the Vampire (17144)
2. Astounding Stories of Super-Science September 1930 (29255)
3. The Prisoner of Zenda (95)
4. The Literary World Seventh Reader (19721)
5. The Works of Edgar Allan Poe Volume 1 (2147)



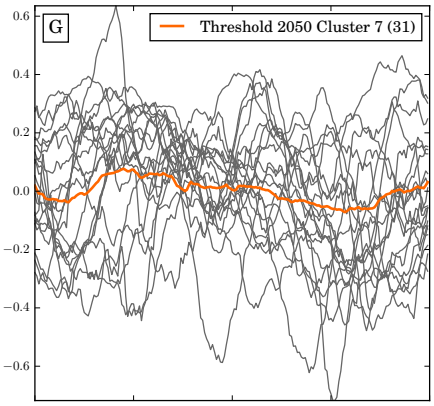
1. The Liberty Minstrel (22089)
2. The Adventures of Pinocchio (500)
3. Incidents in the Life of a Slave Girl, Written ... (11030)
4. Dreams (1439)
5. Legends of Lancashire (51177)



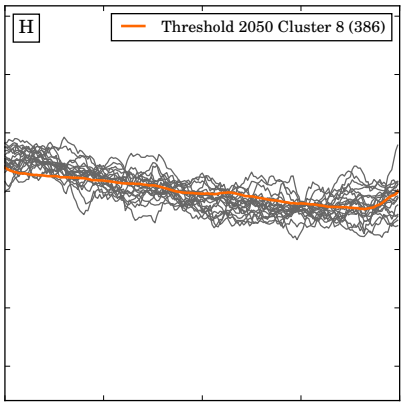
1. The Light Princess (697)
2. Euthyphro (1642)
3. Piper in the Woods (32832)
4. Appointment In Tomorrow (51152)
5. Star, Bright (50935)



1. Asmodeus; or, The Devil on Two Sticks (51145)
2. The Enchanted Castle (3536)
3. Ghost Stories of an Antiquary (8486)
4. Among the Burmans: A Record of Fifteen Years of ... (51080)
5. Bushido, the Soul of Japan (12096)



1. A General Introduction to Psychoanalysis (38219)
2. Paradise Lost (20)
3. Second Treatise of Government (7370)
4. Paradise Lost (26)
5. The Divine Comedy by Dante, Illustrated (8800)



1. Othello (2267)
2. Ghosts (8121)
3. An Enemy of the People (2446)
4. Pierre and Jean (3804)
5. Volpone; Or, The Fox (4039)

FIG. S9: Ward cluster number 3, showing the 8 largest clusters.

Appendix D: Unsupervised Machine Learning

The SOM works by finding the most similar emotional arc in a random collection of arcs. We use a 13x13 SOM (for 169 nodes, roughly 10% of the number of books), connected on a square grid, training according to the original procedure (with winner take all, and scaling functions across both distance and magnitude). We take the neighborhood influence function at iteration i as

$$\mathrm{Nbd}_k(i) = \left[j \in \mathcal{N} \mid d(k,j) < \sqrt{N} \cdot (i+1)^{\alpha} \right] \tag{D1}$$

for a node k in the set of nodes $\mathcal{N}$, with distance function d and total number of nodes N. For result shown here we take $\alpha = -0.15$. We implement the learning adaptation function at training iteration i as

$$f(i) = (i+1)^{\beta}, \tag{D2}$$

again with $\beta = -0.15$

In Fig. S10 we see both the B-Matrix to demonstrate the strength of spatial clustering and a heat-map showing where we find the winning nodes. The A–I labels refer to the individual nodes shown in Fig. S11, and we observe three spatial groups in the both panels of S10: (1) A, C, and D, (2) G and H, and (3) I and F. These spatial clusters reinforce the visible similarity of the winning node arcs. We show the winning node emotional arcs and the arcs of books for which they are the winners in Fig. S11. The legend shows the node ID, numbers the cluster by size, and in parenthesis indicates the size of the cluster on that individual node. In Panels A, C, and D we see varying strengths of the "Rags to Riches" emotional arc. In Panel B, the second largest individual cluster consists of the "Tragedy" stories. In Panels F and I we see the "Boy Meets Girl" story type, with a more pronounced positive start and happier ending in Panel F. In Panels G and H we again see the "Boy Meets Girl" with more variation and a deeper "Man in the Hole" at the end. Each of these top stories are all readily identifiable, yet again demonstrating the universality of these story types.

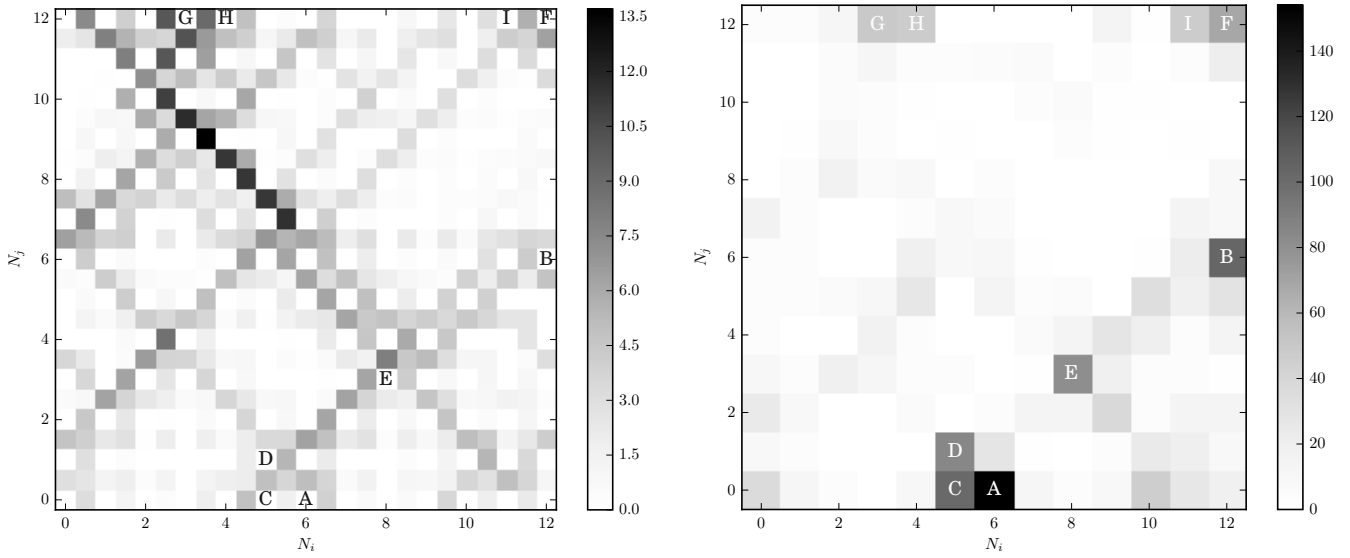


FIG. S10: Left panel: Nodes on the 2D SOM grid are shaded by the number of stories for which they are the winner. Right panel: The B-Matrix shows that there are clear clusters of stories in the 2D space imposed by the SOM network.

The top 9 SOM stories for our set of null emotional arcs are shown in Fig. S17. We observe that, as hypothesized, the shuffled versions of stories lack coherent structure and are not emotional experiences that we can attach to meaningful stories.

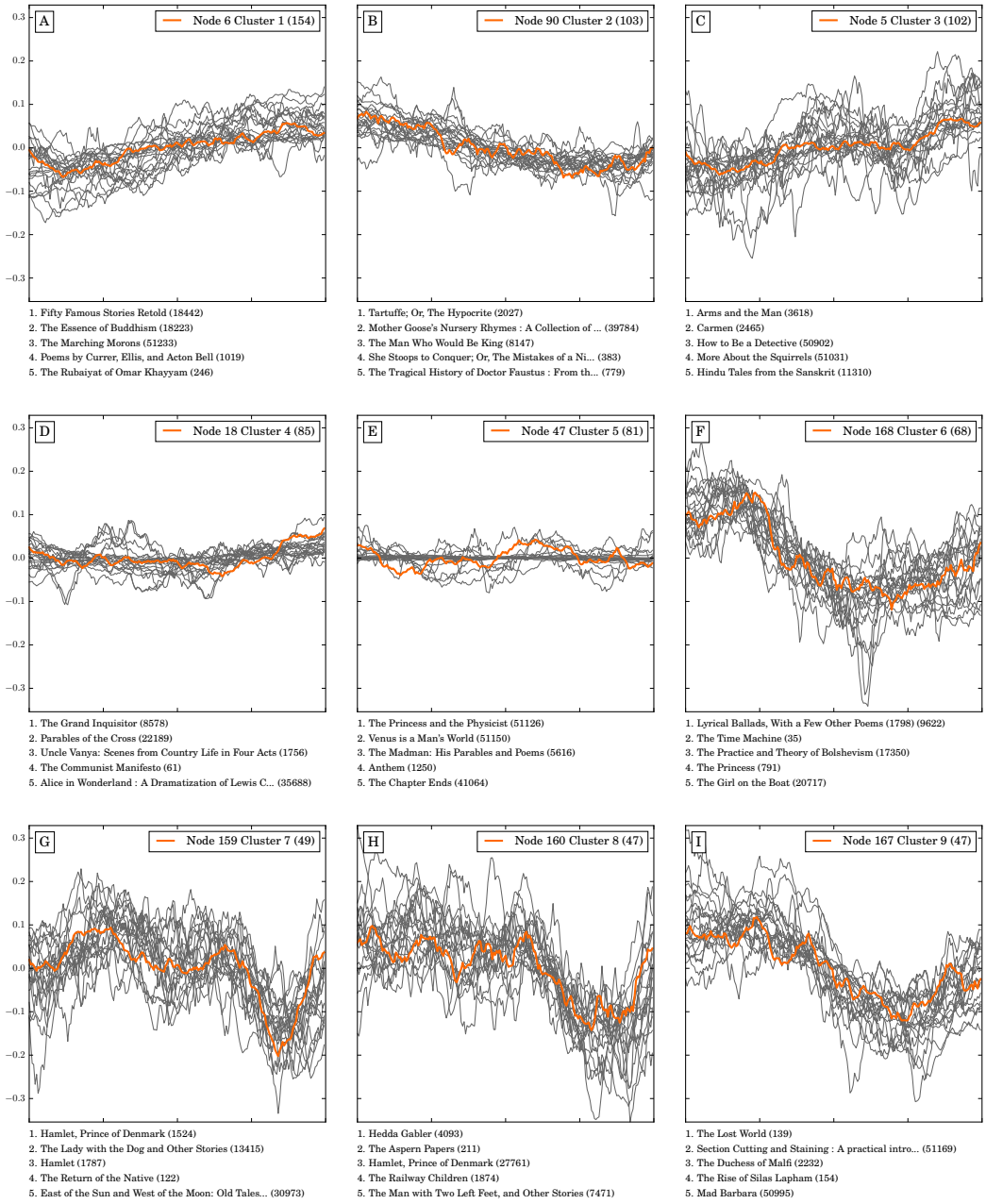


FIG. S11: The vector for each of the top 9 SOM nodes, accompanied with those sentiment time series which are closest to that node. The core stories which we have found with other methods are readily visible.

Appendix E: Word salads (null) comparison

Figs. S12, S13, S14, S15, S16, and S17 showcase the emotional arc analysis applied to word salad books.

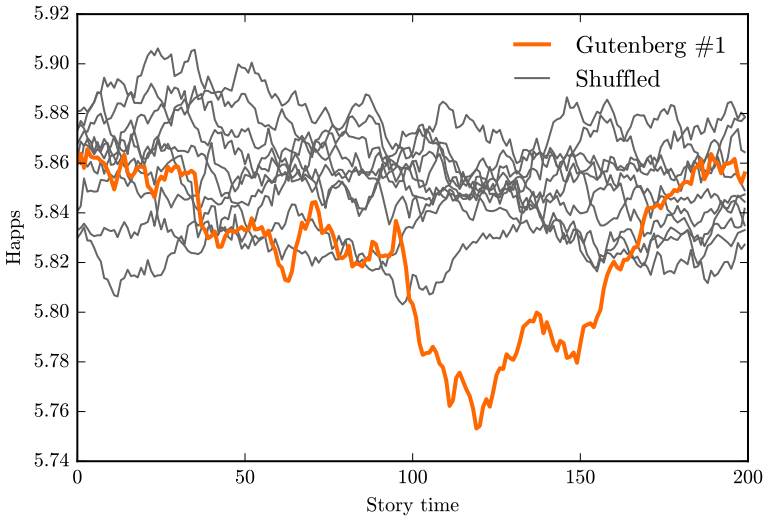


FIG. S12: The first book in Project Gutenberg, *The Declaration of Independence of the United States of America* by Thomas Jefferson, along with word salad versions.

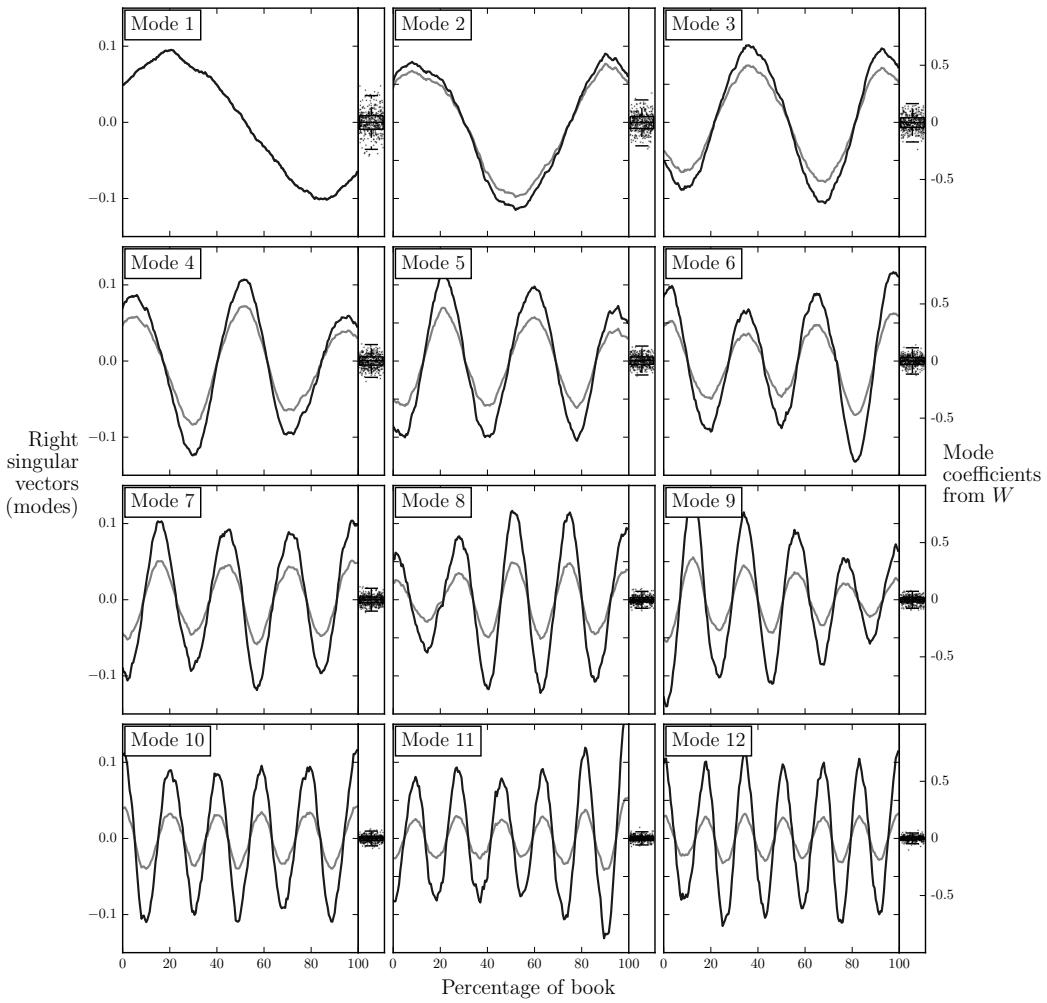


FIG. S13: SVD modes from the emotional arcs of word salad books. We observe higher frequency modes appearing more quickly, and a more even spread of mode coefficients.

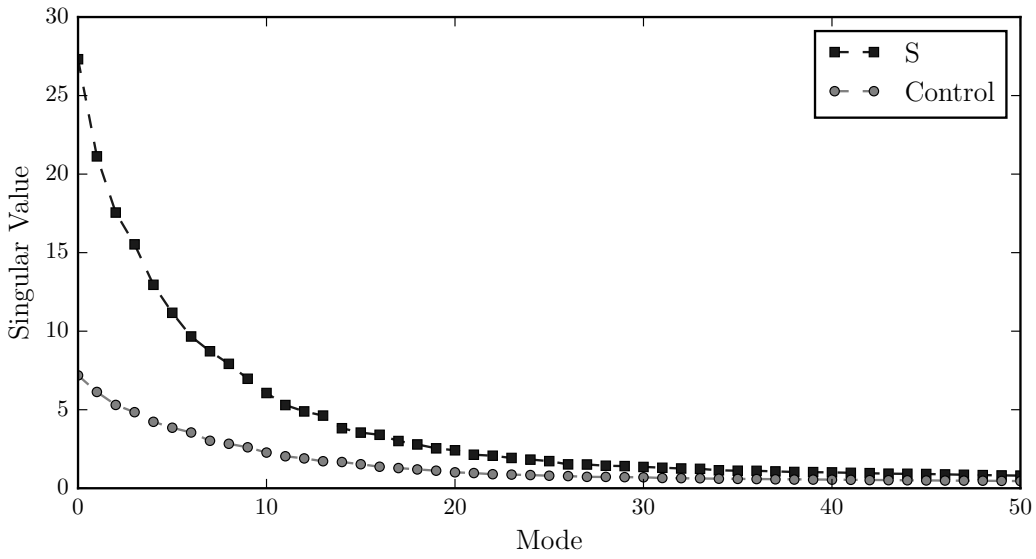


FIG. S14: Comparison of the singular value spectra from the word salad emotional arcs and the emotional arcs of individual Project Gutenberg books. The spectra from the word salad books is muted, indicating both lower total variance explained and less important ordering of the singular vectors.

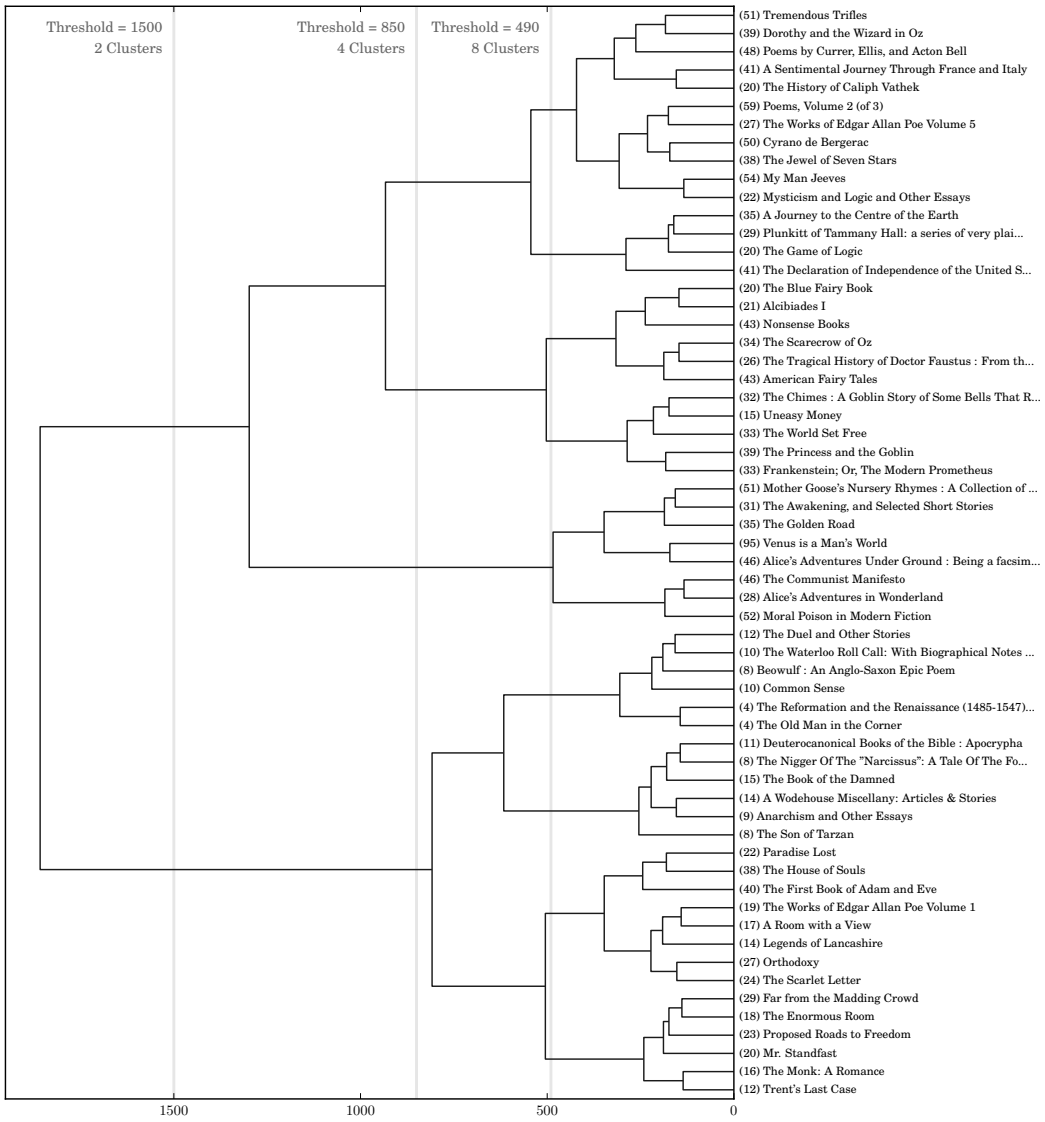


FIG. S15: Dendrogram of clustering using Ward's method on the emotional arcs of word salad books. We observe comparatively low linkage cost for these emotional arcs, indicating the absence of distinct clusters.

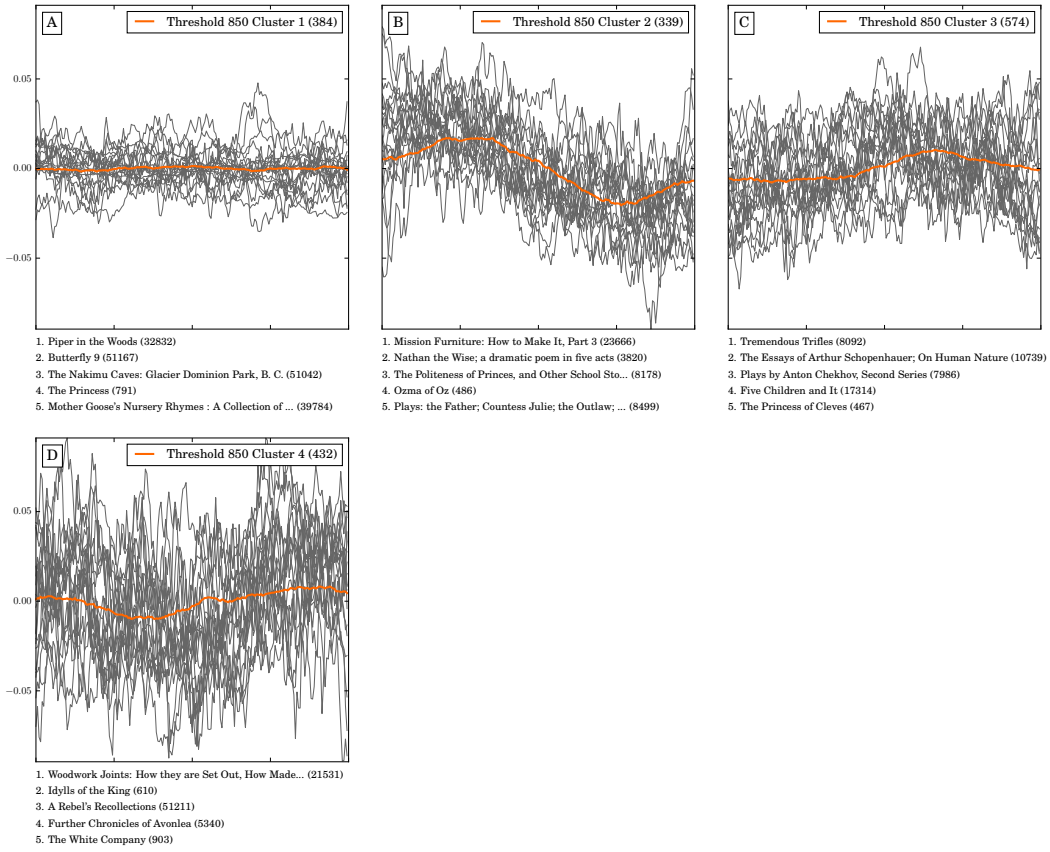


FIG. S16: Four clusters (linkage threshold 850) from the hierarchical clustering of word salad books. We observe that the cluster mean emotional arc and the most central emotional arcs have high variance, without a visible signal.

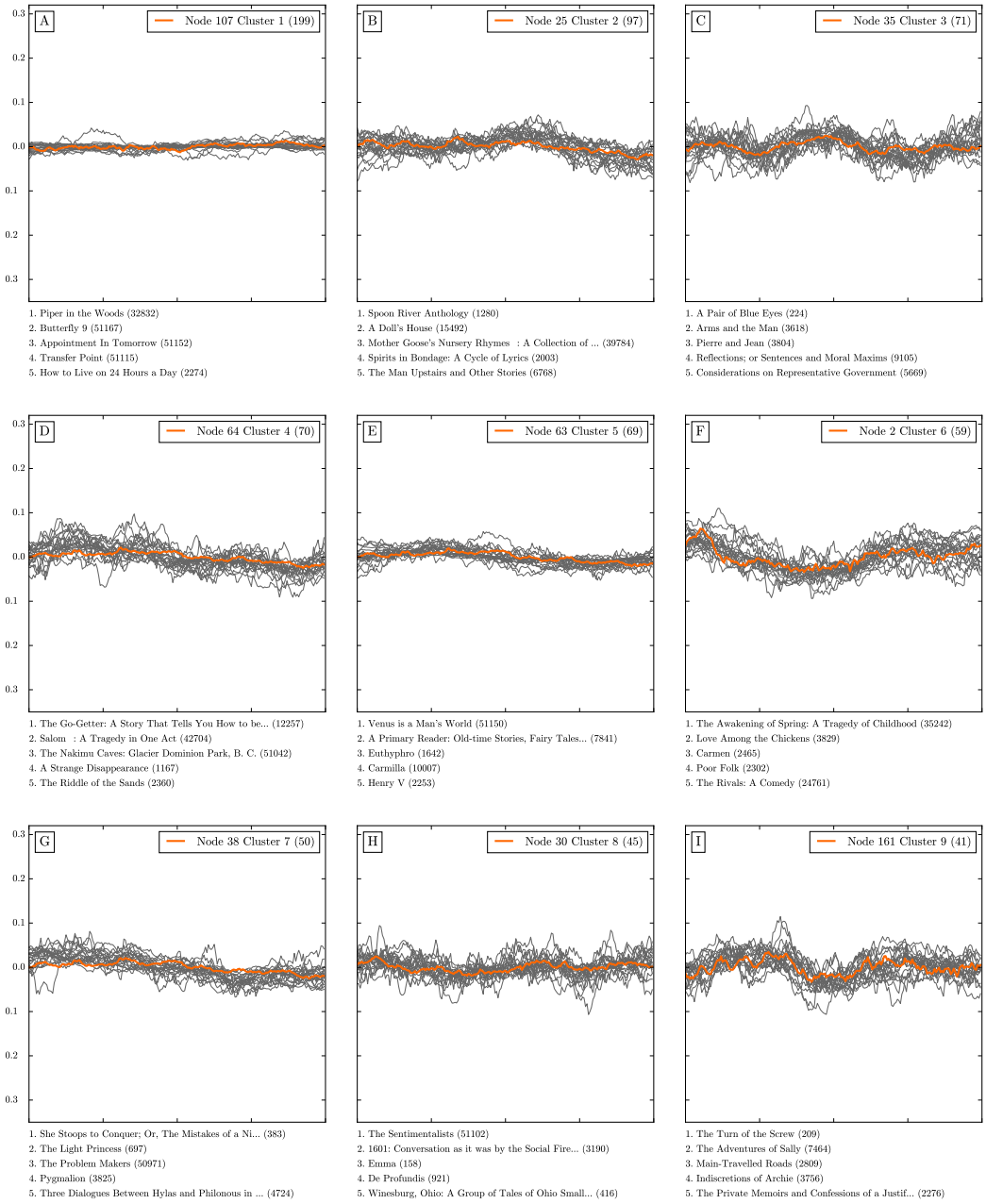


FIG. S17: The vector for each of the top 9 SOM nodes for null emotional arcs, accompanied with those sentiment time series which are closest to that node. We see that the emotional arcs from null arcs show little coherent structure, collapsing heavily onto a single node with muted variance.

Appendix F: Additional Figures

Here we include additional supporting information.

The steps, you see, are all the presents the fairy godmother gave to Cinderella, the ball gown, the slippers, the carriage, and so on. The sudden drop is the stroke of midnight at the ball. Cinderella is in rags again. All the presents have been repossessed. But then the prince finds her and marries her, and she is infinitely happy ever after. She gets all the stuff back, and *then* some. A lot of people think the story is trash, and, on graph paper, it certainly looks like trash.

But then I said to myself, Wait a minute—those steps at the beginning look like the creation myth of virtually every society on earth. And then I saw that the stroke of midnight looked exactly like the unique creation myth in the Old Testament. And then I saw that the rise to bliss at the end was identical with the expectation of redemption as expressed in primitive Christianity.

The tales were identical.

I was thrilled to discover that years ago, and I am just as thrilled today. The apathy of the University of Chicago is repulsive to me.

They can take a flying fuck at the moooooooooooooooon.

FIG. S18: Kurt Vonnegut writes in his autobiography *Palm Sunday* on the similarity of certain story shapes [23]. The exposition of this particular similarity would place both of these stories in our grouping of "Rags to Riches" emotional arcs.

Mode	Mode Arc	N_m	N_m/N	DL Median ▼	DL Mean ▽	DL Variance	% > Average	Download Distribution
SV 1		267	15.4%	289.0	638.0	2176764	20.6%	
-SV 1		440	25.4%	337.5	633.6	907943	25.0%	
SV 2		219	12.7%	327.0	652.3	1122421	21.9%	
-SV 2		167	9.7%	297.0	540.2	554142	16.8%	
SV 3		104	6.0%	298.0	896.3	7829052	22.1%	
-SV 3		109	6.3%	303.0	803.9	2839614	26.6%	
SV 4		108	6.2%	311.5	823.5	2728083	26.9%	
-SV 4		47	2.7%	286.0	790.6	1637200	19.1%	
SV 5		48	2.8%	280.0	397.1	146597	8.3%	
-SV 5		44	2.5%	280.5	452.0	188580	13.6%	
SV 6		15	0.9%	336.0	500.8	337828	13.3%	
-SV 6		43	2.5%	267.0	689.5	929111	25.6%	
SV 7		19	1.1%	336.0	678.7	643264	21.1%	
-SV 7		24	1.4%	286.5	786.5	1597462	25.0%	
SV 8		12	0.7%	279.5	509.0	479830	8.3%	
-SV 8		14	0.8%	275.0	764.9	1067329	28.6%	
-SV 9		10	0.6%	215.0	275.2	13927	0.0%	
SV 10		10	0.6%	410.0	756.4	666121	20.0%	

FIG. S19: Download statistics for SVD Modes with more than 0.5% of books.