Detecting and Evaluating Local Text Reuse in Social Networks

Shaobin Xu[*], **David A. Smith**[*], **Abigail Mullen**[†], and **Ryan Cordell**[‡]
NULab for Texts, Maps, and Networks
College of Computer and Information Science[*], Department of History[†], Department of English[‡]
Northeastern University, Boston, MA

Abstract

Texts propagate among participants in many social networks and provide evidence for network structure. We describe intrinsic and extrinsic evaluations for algorithms that detect clusters of reused passages embedded within longer documents in large collections. We explore applications of these approaches to two case studies: the culture of free reprinting in the nineteenth-century United States and the use of similar language in the public statements of U.S. members of Congress.

1 Introduction

While many studies of social networks use surveys and direct observation to catalogue actors (nodes) and their interactions (edges), we often cannot directly observe network links. Instead, we might observe behavior by network participants that provides indirect evidence for social ties.

One revealing form of shared behavior is the reuse of text by different social actors. Methods to uncover invisible links among sources of text methods would have broad applicability because of the very general nature of the problem—sources of text include websites, newspapers, individuals, corporations, political parties, and so on. Further, discerning those hidden links between sources would provide more effective ways of identifying the provenance and diverse sources of information, and to build predictive models of the diffusion of information.

There are substantial challenges, however, in building a methodology to study text reuse, including: scalable detection of reused passages; identification of appropriate statistical models of text mutation; inference methods for characterizing missing nodes that originate or mediate text transmission; link inference conditioned on textual topics; and the development of testbeds through which predictions of the resulting models might be validated against some broader understanding of the processes of transmission.

In this paper, we sketch relevant features of our two testbed collections (§2) and then describe initial progress on developing algorithms for detecting reused passages embedded within the larger text output of social network nodes (§3). We then describe an intrinsic evaluation of the efficiency of these techniques for scaling up text reuse detection (§4). Finally, we perform an extrinsic evaluation of the network links inferred from text reuse by correlating them with side information about the underlying social networks (§5). A preliminary version of the text reuse detection system was presented for a single, smaller corpus in (Anonymous, 2013), but without the extrinsic or much of the intrinsic evaluation and without data on the underlying networks.

2 Case Studies in Text Reuse

The case studies in this paper, which form the basis for our experimental evaluations below, involve two fairly divergent domains: the informational and literary ecology of the nineteenth-century United States and of twenty-first century U.S. legislators.

2.1 Tracking Viral Texts in 19c Newspapers

In *American Literature and the Culture of Reprinting*, McGill (2003) argues that American literary culture in the nineteenth century was shaped by the widespread practice of reprinting stories and poems, usually without authorial permission or even

Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 50–57,
Baltimore, Maryland USA, 27 June 2014. ©2014 Association for Computational Linguistics

knowledge, in newspapers, magazines, and books. Without substantial copyright enforcement, texts circulated promiscuously through the print market and were often revised by editors during the process. These "viral" texts—be they news stories, short fiction, or poetry—are much more than historical curiosities. The texts that editors chose to pass on are useful barometers of what was exciting or important to readers during the period, and thus offer significant insight into the priorities and concerns of the culture.

Nineteenth-century U.S. newspapers were usually associated with a particular political party, religious denomination, or social cause (e.g., temperance or abolition). Mapping the specific locations and venues in which varied texts circulated would therefore allow us to answer questions about how reprinting and the public sphere in general were affected by geography, communication and transportation networks, and social, political, and religious affinities. These effects should be particularly observable in the period before the Civil War and the rise of wire services that broadcast content at industrial scales (Figure 1).

To study the reprint culture of this period, we crawled the online newspaper archives of the Library of Congress's *Chronicling America* project (`chroniclingamerica.loc.gov`). Since the Chronicling America project aggregates state-level digitization efforts, there are some significant gaps: e.g., there are no newspapers from Massachusetts, which played a not insubstantial role in the literary culture of the period. While we continue to collect data from other sources in order to improve our network analysis, the current dataset remains a useful, and open, testbed for text reuse detection and analysis of overall trends. For the pre-Civil War period, this corpus contains 1.6 billion words from 41,829 issues of 132 newspapers.

Another difficulty with this collection is that it consists of the OCR'd text of newspaper issues without any marking of article breaks, headlines, or other structure. The local alignment methods described in §3 are designed not only to mitigate this problem, but also to deal with partial reprinting. One newspaper issue, for instance, might reprint chapters 4 and 5 of a Thackeray novel while another issue prints only chapter 5.

Since our goal is to detect texts that spread from one venue to another, we are not interested in texts that were reprinted frequently in the same newspaper, or *series*, to use the cataloguing term. This includes material such as mastheads and manifestos and also the large number of advertisements that recur week after week in the same newspaper.

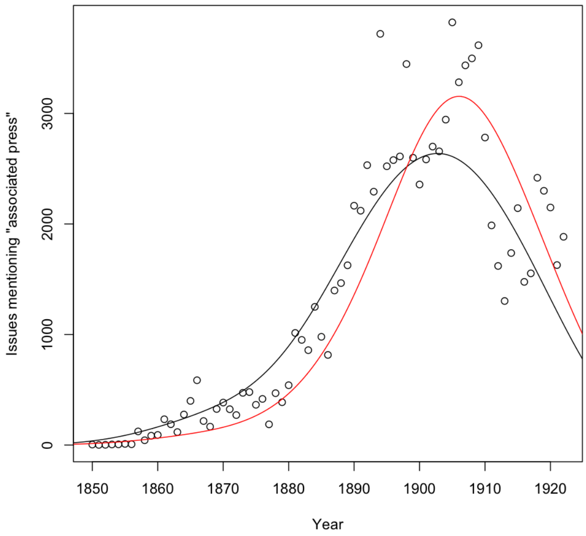

Figure 1: Newspaper issues mentioning "associated press" by year, from the *Chronicling America* corpus. The black regression line fits the raw number of issues; the red line fits counts corrected for the number of times the Associated Press is mentioned in each issue.

2.2 Statements by Members of Congress

Members of the U.S. Congress are of course even more responsive to political debates and incentives than nineteenth-century newspapers. Representatives and senators are also a very well-studied social network. Following Margolin et al. (2013), we analyzed a dataset of more than 400,000 public statements made by members of the 112th Senate and House between January 2011 and August 2012. The statements were downloaded from the Vote Smart Project website (`votesmart.com`). According to Vote Smart, the Members' public statements include any press releases, statements, newspaper articles, interviews, blog entries, newsletters, legislative committee websites, campaign websites and cable news show websites (Meet the Press, This Week, etc.) that contain direct quotes from the Member. Since we are primarily interested in the connections *between* Members, we will, as we see below, want to filter out reuse among different statements by the same member. That information could be interesting for other reasons—for instance, tracking slight changes in the phrasing of talking points or substantive positions.

We supplemented these texts with categorical data chambers and parties and with continuous representations of ideology using the first dimension of the DW-NOMINATE scores (Carroll et al., 2009).

3 Text Reuse Detection

As noted above, we are interested in detecting passages of text reuse (poems or stories; political talking points) that comprise a small fraction of the containing documents (newspaper issues; political speeches). Using the terminology of biological sequence alignment, we are interested in *local alignments* between documents. In text reuse detection research, two primary methods are n-gram shingling and locality-sensitive hashing (LSH) (Henzinger, 2006). The need for local alignments makes LSH less practical without performing a large number of sliding-window matches.

In contrast to work on near-duplicate document detection and to work on "meme tracking" that takes text between quotation marks as the unit of reuse (Leskovec et al., 2009; Suen et al., 2013), here the boundaries of the reused passages are not known. Also in contrast to work on the contemporary news cycle and blogosphere, we are interested both in texts that are reprinted within a few days and after many years. We thus cannot exclude potentially matching documents for being far removed in time. Text reuse that occurs only among documents from the same "source" (run of newspapers; Member of Congress) should be excluded. Similarly, Henzinger (2006) notes that many of the errors in near-duplicate webpage detection arose from false matches among documents from the same website that shared boilerplate navigational elements.

3.1 Efficient N-gram Indexing

The first step is to build for each n-gram feature an inverted index of the documents where it appears. As in other duplicate detection and text reuse applications, we are only interested in the n-grams shared by two or more documents. The index, therefore, does not need to contain entries for the n-grams that occur only once. We use the two-pass space-efficient algorithm described in Huston et al. (2011), which, empirically, is very efficient on large collections. In a first pass, n-grams are hashed into a fixed number of bins. On the second pass, n-grams that hash to bins with one occupant can be discarded; other postings are passed through. Due to hash collisions, there may still be a small number of singleton n-grams that reach this stage. These singletons are filtered out as the index is written.

In building an index of n-grams, an index of (n-1)-grams can also provide a useful filter. No 5-gram, for example, can occur twice unless its constituent 4-grams occur at least twice. We do not use this optimization in our experiments; in practice, n-gram indexing is less expensive than the later steps.

3.2 Extracting and Ranking Candidate Pairs

Once we have an inverted index of the documents that contain each (skip) n-gram, we use it to generate and rank document pairs that are candidates for containing reprinted texts. Each entry, or *posting list*, in the index may be viewed as a set of pairs (d_i, p_i) that record the document identifier and position in that document of that n-gram.

Once we have a posting list of documents containing each distinct n-gram, we output all pairs of documents in each list. We suppress repeated n-grams that appear in different issues of the same newspaper. These repetitions often occur in editorial boilerplate or advertisements, which, while interesting, are outside the scope of this project. We also suppress n-grams that generate more than $\binom{u}{2}$ pairs, where u is a parameter.[1] These frequent n-grams are likely to be common fixed phrases. Filtering terms with high document frequency has led to significant speed increases with small loss in accuracy in other document similarity work (Elsayed et al., 2008). We then sort the list of repeated n-grams by document pair, which allows us to assign a score to each pair based on the number of overlapping n-grams and the distinctiveness of those n-grams. Table 1 shows the parameters for trading off recall and precision at this stage.

3.3 Computing Local Alignments

The initial pass returns a large ranked list of candidate document pairs, but it ignores the order of the n-grams as they occur in each document. We therefore employ local alignment techniques to find compact passages with the highest probability of matching. The goal of this alignment is

[1] The filter is parameterized this way because it is applied after removing document pairs in the same series.

n	n-gram order
w	maximum width of skip n-grams
g	minimum gap of skip n-grams
u	maximum distinct series in the posting list

Table 1: Parameters for text reuse detection

to increase the precision of the detected document pairs while maintaining high recall. Due to the high rate of OCR errors, many n-grams in matching articles will contain slight differences.

Unlike some partial duplicate detection techniques based on global alignment (Yalniz et al., 2011), we cannot expect all or even most of the articles in two newspaper issues, or the text in two books with a shared quotation, to align. Rather, as in some work on biological subsequence alignment (Gusfield, 1997), we are looking for regions of high overlap embedded within sequences that are otherwise unrelated. We therefore employ the Smith-Waterman dynamic programming algorithm with an affine gap penalty. This use of model-based alignment distinguishes this approach for other work, for detecting shorter quotations, that greedily expands areas of n-gram overlap (Kolak and Schilit, 2008; Horton et al., 2010). We do, however, prune the dynamic programming search by forcing the alignment to go through position pairs that contain a matching n-gram from the previous step, as long as the two n-grams are unique in their respective texts. Even the exact Smith-Waterman algorithm, however, is an approximation to the problem we aim to solve. If, for instance, two separate articles from one newspaper issue were reprinted in another newspaper issue in the opposite order—or separated by a long span of unrelated matter—the local alignment algorithm would simply output the better-aligned article pair and ignore the other. Anecdotally, we only observed this phenomenon once in the newspaper collection, where two different parodies of the same poem were reprinted in the same issue. In any case, our approach can easily align different passages in the same document to passages in two other documents.

The dynamic program proceeds as follows. In this paper, two documents would be treated as sequences of text X and Y whose individual characters are indexed as X_i and Y_j. Let $W(X_i, Y_j)$ be the score of aligning character X_i to character Y_j. Higher scores are better. We use a scoring function where only exact character matches get a positive score and any other pair gets a negative score. We also account for additional text appearing on either X or Y. Let W_g be the score, which is negative, of starting a "gap", where one sequence includes text not in the other. Let W_c be the cost for continuing a gap for one more character. This "affine gap" model assigns a lower cost to continuing a gap than to starting one, which has the effect of making the gaps more contiguous. We use an assignment of weights fairly standard in genetic sequences where matching characters score 2, mismatched characters score -1, beginning a gap costs -5, and continuing a gap costs -0.5. We leave for future work the optimization of these weights for the task of capturing shared policy ideas.

As with other dynamic programming algorithms such as Levenshtein distance, the Smith-Waterman algorithm operates by filling in a "chart" of partial results. The chart in this case is a set of cells indexed by the characters in X and Y, and we initialize it as follows:

$$\begin{aligned} H(0,0) &= 0 \\ H(i,0) &= E(i,0) = W_g + i \cdot W_c \\ H(0,j) &= F(0,j) = W_g + j \cdot W_c \end{aligned}$$

The algorithm is then defined by the following recurrence relations:

$$H(i,j) = \max \begin{cases} 0 \\ E(i,j) \\ F(i,j) \\ H(i-1,j-1) + W(X_i, Y_j) \end{cases}$$

$$E(i,j) = \max \begin{cases} E(i,j-1) + W_c \\ H(i,j-1) + W_g + W_c \end{cases}$$

$$F(i,j) = \max \begin{cases} F(i-1,j) + W_c \\ H(i-1,j) + W_g + W_c \end{cases}$$

The main entry in each cell $H(i,j)$ represents the score of the best alignment that terminates at position i and j in each sequence. The intermediate quantities E and F are used for evaluating gaps. Due to taking a max with 0, $H(i,j)$ cannot be negative. This is what allows Smith-Waterman to ignore text before and after the locally aligned substrings of each input.

After completing the chart, we then find the optimum alignment by tracing back from the cell with the highest cumulative value $H(i,j)$ until a cell with a value of 0 is reached. These two cells represent the bounds of the sequence, and the overall SW alignment score reflects the extent to which the characters in the sequences align and the overall length of the sequence.

In our implementation, we include one further speedup: since in a previous step we identified n-grams that are shared between the two documents, we assume that any alignment of those documents must include those n-grams as matches. In some cases, this anchoring of the alignment might lead to suboptimal SW alignment scores.

4 Intrinsic Evaluation

To evaluate the precision and recall of text reuse detection, we create a pseudo-relevant set of document pairs by pooling the results of several runs with different parameter settings. For each document pair found in the union of these runs, we observe the length, in matching characters, of the longest local alignment. (Using matching character length allows us to abstract somewhat from the precise cost matrix.) We can then observe how many aligned passages each method retrieves that are at least 50,000 character matches in length, at least 20,000 character matches in length, and so on. The candidate pairs are sorted by the number of overlapping n-grams; we measure the pseudo-recall at several length cutoffs. For each position in a ranked list of document pairs, we then measure the precision: what proportion of documents retrieved are in fact 50k, 20k, etc., in length? Since we wish to rank documents by the length of the aligned passages they contain, this is a reasonable metric. One summary of these various values is the *average precision*: the mean of the precision at every rank position that contains an actually relevant document pair. One of the few earlier evaluations of local text reuse, by Seo and Croft (2008), compared fingerprinting methods to a trigram baseline. Since their corpus contained short individual news articles, the extent of the reused passages was evaluated qualitatively rather than by alignment.

Figure 2 shows the average precision of different parameter settings on the newspaper collection, ranked by the number of pairs each returns. If the pairwise document step returns a large number of pairs, we will have to perform a large number of more costly Smith-Waterman alignments. On this collection, a good tradeoff between space and speed is achieved by skip bigram features. In the best case, we look at bigrams where there is a gap of at least 95, and not more than 105, words between the first and second terms (n=2 u=100 w=105 g=95).

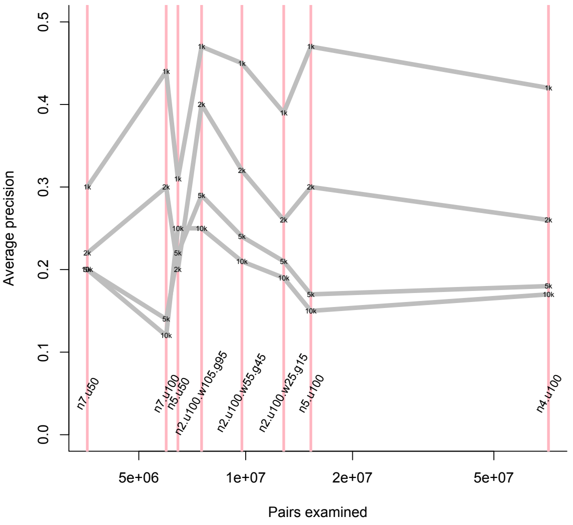

Figure 2: Average precision for aligned passages of different minimum length in characters. Vertical red lines indicate the performance of different parameter settings (see Table 1).

While average precision is a good summary of the quality of the ranked list at any one point, many applications will simply be concerned with the total recall after some fixed amount of processing. Figure 3 also summarizes these recall results by the absolute number of document pairs examined. From these results, it is clear the several good settings perform well at retrieving all reprinted passages of at least 5000 characters. Even using the pseudo-recall metric, however, even the best operating points fail in the end to retrieve about 10% of the reprints detected by some other setting for all documents of at least 1000 characters.

5 Extrinsic Evaluation

While political scientists, historians, and literary scholars will, we hope, find these techniques useful and perform close reading and manual analysis on texts of interest, we would like to validate our results without a costly annotation campaign. In this paper, we explore the correlation of patterns of text reuse with what is already known from other

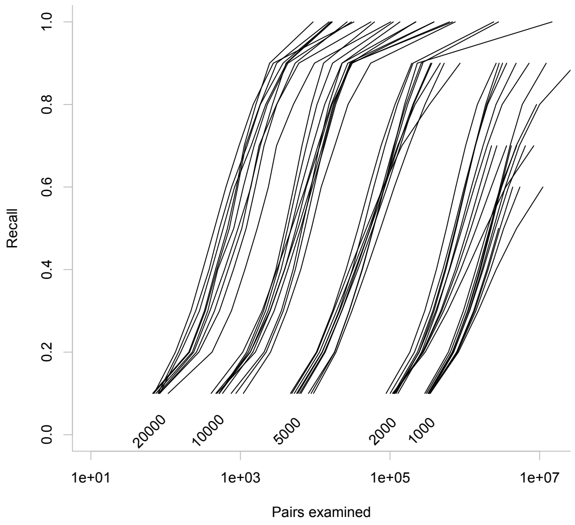

Figure 3: (Pseudo-)Recall for aligned passages of different minimum lengths in characters.

sources about the connections among Members of Congress, newspaper editors, and so on. This idea was inspired by Margolin et al. (2013), who used these techniques to test rhetorical theories of "semantic organizing processes" on the congressional statements corpus.

The approach is quite simple: measure the correlation between some metric of text reuse between actors in a social network and other features of the network links between those actors. The metric of text reuse might be simply the number of exact n-grams shared by the language of two authors (Margolin et al., 2013); alternately, it might be the absolute or relative length of all the aligned passages shared by two authors or the tree distance between them in a phylogenetic reconstruction. To measure the correlation of a text reuse metric with a single network, we can simply use Pearson's correlation; for more networks, we can use multivariate regression. Due to, for instance, autocorrelation among edges arising from a particular node, we cannot proceed as if the weight of each edge in the text reuse network can be compared independently to the weight of the corresponding edges in other networks. We therefore use nonparametric permutation tests using the quadratic assignment procedure (QAP) to resample several networks with the same structure but different labels and weights. The QAP achieves this by reordering the rows and columns of one network's adjacency matrix according to the same permutation. The permuted network then has the same structure—e.g., degree distribution—but should no longer exhibit the same correlations with the other network(s). We can run QAP to generate confidence intervals for both single (Krackhardt, 1987) and multiple correlations (Dekker et al., 2007).

5.1 Congressional Statements

We model the connection between the log magnitude of reused text and the strength of ties among Members according to whether they are in the same chamber and how similar they are on the first dimension of the DW-nominate ideological scale (Carroll et al., 2009). On the left side of Table 2 are shown the results for correlating reused passages of certain minimum lengths (10, 16, 32 words) with these underlying features. On the right are shown the similar results of (Margolin et al., 2013) that simply used the exact size of the n-gram overlap between Members' statements for increasing values of n. The alignment analysis proposed in this paper achieves similar results when passages and n-grams are short. Our analysis, however, achieves higher single and multiple correlations among networks are the passages grow longer. This is unsurprising since the probability of an exact 32-gram match is much smaller than that of a 32-word-long alignment that might contain a few differences. In particular, the much higher coefficients for DW-nominate at longer aligned lengths suggests that ideological influence still dominates over similarities induced by the procedural environment of each congressional chamber.

5.2 Network Connections of 19c Reprints

For the antebellum newspaper corpus, we are also interested in how political affinity correlates with reprinting similar texts. We have also added variables for social causes such as temperance, women's rights, and abolition that—while certainly not orthogonal to political commitments—might sometimes operate independently. In addition, we also added a "shared state" variable to account for shared political and social environments of more limited scope. Figure 4 shows a particularly strong example of a geographic effect: the statement of the radical abolitionist John Brown after being condemned to death for attacking a federal arsenal and attempting to raise a slave rebellion was very unlikely to be published in the

	aligned passages of $\geq n$ words			n-grams of length		
	10	16	32	8	16	32
	First-order Pearson correlations					
DW-nominate	0.26***	0.25***	0.23***	0.26***	0.22***	0.16***
same chamber	0.05*	0.08**	0.13***	-0.05***	0.21***	0.10***
	Regression coefficients					
DW-nominate	0.72***	0.75***	0.74***	1.31***	2.67***	0.36
same chamber	0.15**	0.27***	0.42***	0.20	3.14***	0.81***
R-squared	.069	.070	.073	.068	.073	.010

Table 2: Correlations between log length of aligned text and other author networks in public statements by Members of Congress. $*p < .05, * * p < .01, * * *p < .001$

South.

Using information from the Chronicling America cataloguing and from other newspaper histories, we coded each of the 132 newspapers in the corpus with these political and social affinities. We then counted the number of reprinted passages shared by each pair of newspapers. There is not a deterministic relationship between the number of *pairs* of newspapers sharing an affinity and the number of *reprints* shared by those papers. While our admittedly partial corpus only contains a single pair of avowedly abolitionist papers—a radical position at the time—those two papers shared articles 306 times, compared for instance to the 71 stories shared among the 6 pairs of "nativist" papers.

Table 3 shows that geographic proximity had by far the strongest correlation with (log) reprinting counts. Interestingly, the only political affinity to show as strong a correlation was the Republican party, which in this period had just been organized and, one might suppose, was trying to control its "message". The Republicans were more geographically concentrated in any case, compared to the sectionally more diffuse Democrats. Another counterexample is the Whigs, the party from which the new Republican party drew many of its members, which also has a slight negative effect on reprinting. The only other large coefficients are in the complete model for smaller movements such as nativism and abolition. It is interesting to speculate about whether the speed or faithfulness of reprinting—as opposed to the volume—might be correlated with more of these variables.

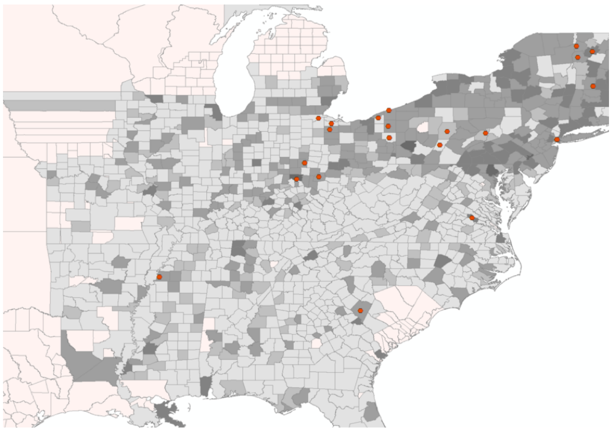

Figure 4: Reprints of John Brown's 1859 speech at his sentencing. Counties are shaded with historical population data, where available. Even taking population differences into account, few newspapers in the South printed the abolitionist's statement.

6 Conclusions

We have presented techniques for detecting reused passages embedded within the larger discourses produced by actors in social networks. Some of this shared content is as brief as partisan talking points or lines of poetry; other reprints can encompass extensive legislative boilerplate or chapters of novels. The longer passages are easier to detect, with prefect pseudo-recall without exhaustive scanning of the corpus. Precision-recall tradeoffs will vary with the density of text reuse and the noise introduced by optical character recognition and other features of data collection. We then showed the feasibility of using network regression to measure the correlations between connections inferred from text reuse and networks derived from outside information.

References

Royce Carroll, Jeff Lewis, James Lo, Nolan McCarty, Keith Poole, and Howard Rosenthal. 2009. Measur-

newspaper	pairs of		regression w/pairs		
affinity	papers	reprints	≥ 1	≥ 10	≥ 100
Republican	1176	134,302	0.74***	0.73*	0.72***
Whig	1176	91,139	-0.35	-0.34	-0.35
Democrat	1081	62,609	-0.08	-0.09	-0.07
same state	672	103,057	1.12***	1.11***	1.13***
anti-secession	435	22,009	-0.58*	-0.58	-0.60
anti-slavery	231	12,742	-0.65	-0.64	-0.60
pro-slavery	120	11,040	-0.35	-0.35	-0.27
Free-State	15	1,194	0.80	0.80	
Constitutional Union	15	1,070	-0.21	-0.21	
pro-secession	15	529	0.11	0.11	
Free Soil	10	1,936	-0.42	-0.42	
Copperhead	10	797	1.53	1.54	
temperance	6	560	0.65		
independent	6	186	-0.22		
nativist	6	71	-1.93*		
women's rights	3	721	1.91		
abolitionist	1	306	3.49**		
Know-Nothing	1	25	1.33		
Mormon	1	3	-1.13		
R-squared	–	–	.065	.063	.062

Table 3: Correlations between shared reprints between 19c newspapers and political and other affinities. While many Whig papers became Republican, they do not completely overlap in our dataset; the identical number of pairs is coincidental.

ing bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Political Analysis*, 17(3).

David Dekker, David Krackhardt, and Tom Snijders. 2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika*, 72(4):563–581.

Tamer Elsayed, Jimmy Lin, and Douglas W. Oard. 2008. Pairwise document similarity in large collections with MapReduce. In *ACL Short Papers*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press.

Monika Henzinger. 2006. Finding near-duplicate web pages: A large-scale evaluation of algorithms. In *SIGIR*.

Russell Horton, Mark Olsen, and Glenn Roe. 2010. Something borrowed: Sequence alignment and the identification of similar passages in large text collections. *Digital Studies / Le champ numérique*, 2(1).

Samuel Huston, Alistair Moffat, and W. Bruce Croft. 2011. Efficient indexing of repeated n-grams. In *WSDM*.

Okan Kolak and Bill N. Schilit. 2008. Generating links by mining quotations. In *Hypertext*.

David Krackhardt. 1987. QAP partialling as a test of spuriousness. *Social Networks*, 9(2):171–186.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD*.

Drew Margolin, Yu-Ru Lin, and David Lazer. 2013. Why so similar?: Identifying semantic organizing processes in large textual corpora. *SSRN*.

Meredith L. McGill. 2003. *American Literature and the Culture of Reprinting, 1834–1853*. U. Penn. Press.

Jangwon Seo and W. Bruce Croft. 2008. Local text reuse detection. In *SIGIR*.

Caroline Suen, Sandy Huang, Chantat Eksombatchai, Rok Sosič, and Jure Leskovec. 2013. NIFTY: A system for large scale information flow tracking and clustering. In *WWW*.

Ismet Zeki Yalniz, Ethem F. Can, and R. Manmatha. 2011. Partial duplicate detection for large book collections. In *CIKM*.