Close Readings of Big Data:

Triangulating Patterns of Textual Reappearance and Attribution

in the *Caledonian Mercury,* 1820-1840

M. H. Beals, Loughborough University

Newspaper digitisation has been hailed as a revolutionary change in how researchers can engage with the periodical press.[1] From immediate global access, to keyword searching, to large-scale text and image analysis, the ever-growing availability of electronic facsimiles, metadata, and machine-readable transcriptions has encouraged scholars to pursue large-scale analyses rather than rely on samplings and soundings from an unwieldy and fragmentary record—to go beyond the case study and attempt the "comprehensive history" of the press that seemed so elusive forty years ago.[2] Yet, after a decade of access to digital newspaper corpora, much of what has been attempted remains fundamentally conservative in approach.[3] In *British Settler Emigration in Print* (2016), Jude Piesse laudably provides URLs to the precise facsimiles she consulted and comments on the search parameters used to obtain her sample. However, her coverage was fragmentary, relying heavily upon select case studies rather than demonstrating general trends, admitting that "[d]igital searches frequently generate thousands of hits, which can be difficult to navigate or to appraise in any detail."[4] She also subtly laments the loss of the immersive offline experience: "Despite the obvious benefits of focused digital searching, it is quite possible that it misses details that research in paper archives would bring to light," the ease of jumping straight to a keyword discouraging a deep contextual understanding of the materials. Online interfaces encourage this type of sampling, with simplified full-text and metadata searches returning a list of "relevant" hits based on often-hidden algorithms, constricting research in ways similar to using a publisher-created newspaper index.[5] Yet, from the beginning, researchers have pushed at the boundaries of these offerings. Through a series of thoughtful search enquires, Dallas Liddle subverted these interface prompts and wrangled a sample of the metadata hidden within the *Times Digital Archive*.[6] Bob Nicholson likewise went to admirable lengths to delineate the work required to effectively sample an externally curated digitised corpus.[7] All of these uses stretched our understanding of the press beyond what was possible a generation ago but a much greater degree of abstraction, or large-scale analysis, is possible. Digitisation offers an opportunity to understand the newspaper press as a multipolar, interactive system—as something other than the sum of its parts.

Although digitisation has impacted many aspects of periodical research, it is perhaps best placed to address the problem of textual reappearance and to understand its role within and between individual publications. Referred to as scissors-and-paste journalism, reprinting, syndication, or simply duplication depending on the nature of the reappearance, the inclusion of a single, recognisable length of text in multiple publications was very common in the nineteenth century.[8] Yet, despite the reappearance of text being oft-observed, much of our evidence for the underlying practices that prompted it is anecdotal: infrequent and sometimes erroneous attributions, memoirs of editors who engaged in the process, and recriminations by victims of the so-called pirate press.[9] The precise scale of the process was largely unknowable before the advent of mass digitisation, as the fallibility of human memory and reading speed prevented a comprehensive view. The modern computer, in contrast, is well-suited to finding these identical blocks of text; rather than be distracted by nuanced interpretation, it needs only to stoically read through billions of lines of meaningless text in the pursuit of a matching pair. Indeed, the more removed the researcher is from the process the more effective the search will be. Using a graphical search interface to obtain a representative sample of even a single text has been proven cumbersome and imprecise.[10] Instead, large-scale interrogations of multiple digital corpora have been required to effectively map wider trends.[11] Yet, just as the serendipity of sampling constrains the wider applicability of a case study, the noise associated with big-data analysis makes applying wider textual trends to specific compositional practices problematic. Duplicates, out of context, tell us little of historical patterns of practice.

In response to these difficulties, this essay explores a middle-scale analysis, one which iterates between the case study and big data to demonstrate how to best leverage digitisation when contextualising both small- and large-scale analyses. It explores the case of the *Caledonian Mercury*—a four-page, thrice-weekly newspaper published in Edinburgh by Thomas Allan—over the course of twenty years, 1820-1840, the heart of what Robert Cowan deemed "the first expansion" of the Scottish newspaper, after the post-Napoleonic boom in provincial titles but before the removal of stamp and advertising duties.[12] This essay first compares a distant reading of the *Caledonian Mercury* against two of its London contemporaries, *The Morning Chronicle* and *The Examiner,* to provide a broad view of how its composition contrasted or aligned with newspapers known to be important sources (*Chronicle*) or curators (*Examiner*) of news content.[13] It then provides a close reading of five "Edinburgh Arrow" issues of the *Mercury* in order to test and contextualise the trends seen in the distant reading.[14] Finally, it offers an additional computational analysis that compares these five issues of the *Mercury* to contemporary issues of other newspapers to determine the proportionality of automatically and manually matched content. Through this iterative process, from large to small and returning to large-scale analysis, this essay demonstrates a digital means for understanding the degree to which the provincial press republished material from other newspapers and made use of implicit and explicit attributions.

A DISTANT READING OF THREE NEWSPAPERS

Unlike close reading, in which a text is read within its original structure, distant reading "aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s).”[15] Through this abstraction, the researcher is at least partially freed from pre-existing biases of focus, aiding them in the discovery of unexpected trends and correlations. Nonetheless, all computational analyses are guided by human-designed parameters that in some way bias the results, if only by limiting those aspects that are to be compared. Here, limits were placed on the category of materials analysed—by genre and corpora—and the unit of analysis—by word count and the temporal distance between units. These limits focused the results on instances of news content—which was quickly disseminated in word-for-word copies—without biasing for or against a specific tone or topic.

In terms of genre and corpora, a desire to focus on news content in the early nineteenth century, alongside practical licencing considerations, led to the use of the *British Library 19th Century Newspapers, Part I* and *Times Digital Archive*.[16] Although far from a complete record of British reportage, access to these machine-readable collections allows a mapping of textual reappearance within newspapers on a previously unachievable scale. Combined, they contain seventeen individual titles for 1820, rising to twenty-seven titles by 1840, distributed regionally and across the political spectrum. Unfortunately, only two other Scottish periodicals are included in this collection, the *Glasgow Herald* and the *Aberdeen Journal*, which may have impacted the number of matches for regional news, discussed in more detail below. Nonetheless, initial analysis of the corpus demonstrated several non-geographically determined clusters of titles with significantly overlapping text, suggesting multiple national networks of news dissemination against which to test initial hypotheses.

The unit of analysis was shaped by the nature of the machine-readable texts. Developed to support full-text searching through a graphical user interface, rather than raw analysis of the text, the dataset first required cleaning. The raw data was transformed from XML into plain text, removing metadata and creating a collection of page-level units of analysis. These files, owing to their creation through optical character recognition, contained a significant number of transcription errors. Rather than attempt to retroactively correct these errors, a highly flexible piece of open-source plagiarism detection software, Copyfind, was used to identify matching texts.[17] Individual pages were compared against each other on a month-by-month and a month-by-succeeding-month basis, producing a list of likely instances of textual reappearance—a match of at least 100 words per page in clusters of at least 20 words each. These manifests were regularised for the typical two-week domestic news-cycle while same-title matches, which were almost exclusively advertisements, were filtered out.[18] These lists were then processed to determine the average word count of a match as well as the percentage of each page and issue that included duplicate text.[19]

The results of these analyses support anecdotal evidence that the *Chronicle*, the *Examiner*, and the *Mercury* had an unusually high level of textual reappearance between them.[20] These newspapers all fall under the general heading of liberal or reform publications—though the *Examiner* held more explicitly radical views than the "moderate constitutional liberalism" of the *Morning Chronicle* and "Whiggishly inclined" *Caledonian Mercury*—so this degree of overlapping text was not unexpected.[21] Despite this, the distribution and nature of textual reappearance varied noticeably between them. Measured by raw word count, the number of words that could be found in an earlier publication, the *Examiner* deviated the most from the corpus average. In the *Chronicle* and the *Mercury*, approximately 55% of matches fell into the range of 100-400 words, evenly distributed, with higher word counts appearing in decreasing frequency. In contrast, over 75% of the *Examiner*'s pages fell into this range. This discrepancy is most readily explained in the larger number and smaller size of its pages. For the *Chronicle* and the *Mercury*, which had similar page sizes and overall word counts, the average length of a matched textual unit—likely representing an article—was 200 words, the same as the corpus average. Throughout the corpus, 1840 saw a marked decrease in the percentage of matches under 300, a more even distribution in the number of matches between 100 and 600, and a significant rise in the number of matches over 1,000—a trend that merits specific consideration when undertaking a close reading of these issues. Of the three publications, the *Mercury* most closely mirrored corpus averages across the period.

These word counts are given greater context by contrasting the percentage of each issue that the computational analysis identified as duplicate material. Again, the results from the *Examiner* vary significantly from the *Chronicle*, the *Mercury,* and the wider corpus. In the latter three, over two-thirds of the issues had fewer than 6% of their content identified as duplicated text, with the strongest clustering around 1-3%. The *Examiner,* conversely, clustered around 10%, with roughly 15% of its issues registering over 15% of their text as duplicate material. As a self-styled "weekly review," these higher percentages are to be expected, even if they appear anomalous within this particular corpus. Taken as a whole, the computational analysis suggests that the occurrence of duplicate material was similar for titles of similar periodicity, scaling logically between weekly and daily publications, and that the preponderance of pages from daily newspapers within the corpus biased the average percentage toward that found in the most frequently issued publications. This highlights the value of pre-processing collections into sub-corpora, ideally over multiple iterations with differing criteria, in order to isolate meaningful trends. Despite these inconsistencies, an average of 3-4% of each issue of the *Mercury* was computationally identified as duplicate material, making it largely representative of the corpus's more frequent publications; it is with this baseline figure that we shall compare our close readings to determine the general accuracy of computation matching.

A CLOSE READING OF THE *CALEDONIAN MERCURY*

A close reading of an issue can focus on the text, images, marginalia, or other manuscript marks; it may also include an examination of the physical layout and the materiality of the item. In this study, close reading entails the dissection of the issue, the careful examination of the text to disambiguate, categorise and number individual textual units by type, topic, geography, source, and word count. This type of newspaper anatomy has several precedents. In *The Press in Australia*, Henry Mayer dissected "twelve random copies" of the *Age* and Melbourne *Herald* for 1855, 1875, 1900, and 1925, cataloguing the average percentage of physical space given to different content types as well as the coverage of various broad topics. From these, he argued that such samples were always in danger of producing misleading trends, with layout and composition fluctuating dramatically in response to changes in "news-values and [the] accessibility of news." [22] In contrast to this broader survey, Charles E. Clark and Charles Wetherell's study of the contents of the *Pennsylvania Gazette* from 1728 to 1765 offers a detailed discussion of a single title.[23] In it, the authors measure the column inches dedicated to different content types and topics, as well as the geographical distribution of stories (by place of action) and their average delay in publication (from time of action). Of particular relevance to this study is their discussion of sources: other publications, correspondence, and oral communication, with the first providing at least 66.4% of the *Gazette*'s news content during the period they studied.[24] Although they had to rely upon explicit attributions and deduction, the degree to which they were able to infer the interconnectivity of colonial American and European newspapers is impressive. By using similar categories, but substituting word count (and, in some sense, compositor effort) for the measurement of physical space, it is hoped that the present study will complement previous close readings and provide new perspectives on how to conceive of the composition of a newspaper. As for the choice of the *Mercury* itself, the wider applicability of this dissection can be seen through the anatomy of a Georgian newspaper put forth by Francis Williams:

> The normal format was a single sheet of 24 1/2 inches by 18 3/4 inches folded once to produce a four-page paper in folio 12 1/4 inches by 18 3/4 inches. Each page was made up in four columns printed solid with the minimum of headings. The news offered consisted normally of summaries of Parliamentary debates, foreign intelligence copied from Continental papers, Court intelligence, reprints of the *London Gazette*, brief reports of decisions in the law and police courts and a certain amount of commercial intelligence. In addition as general readership grew there might be a medley of gossip paragraphs about those in the public eye, a column of jokes and epigrams on social follies, notices of new plays, some verse and a 'Letter to the Printer.'[25]

Based largely on the London dailies, this summation is equally representative of the *Mercury* between 1820 and 1840. It, too, was composed for four pages, printed on a single sheet, though with more columns, growing from five in 1820 to seven in 1840. Its content can, likewise, be placed under the same general headings, though these were perhaps more constricted in any given issue owing to its less frequent publication. Although any delineation is contentious, the *Mercury* contained four general categories of content and their appearance on certain pages was consistent across the period:

	Page One	Page Two	Page Three	Page Four
Advertisements and Notices	127 Items 28903 Words	0 Items 0 Words	25 Items 5000 Words	7 Items 875 Words
Commentary and Miscellany	0 Items 0 Words	17 Items 2654 Words	35 Items 8872 Words	27 Items 7097 Words
Numerical Content	4 Items 908 Words	22 Items 3283 Words	25 Items 2863 Words	54 Items 11234 Words
News	10 Items 6393 Words	88 Items 31233 Words	144 Items 24200 Words	56 Items 18446 Words

Table 1: The combined item and word counts, by page, of the four main types of content printed by the Caledonian Mercury in the sample issues.

The front pages were principally composed of advertisements, stock prices and other formalized notices. Across the period, only 7% of the individual items could be categorised as news content and these appeared only in the 1835 and 1840 issues. Although advertisements were never wholly confined to this page, their place here represented a standard feature of the *Mercury* throughout the period. Page four, the other outer page, was more varied. Of the 144 individual items, 19% were commentaries or miscellany, 39% were news items from local, regional, national, and international sources, and the remaining 42% were advertisements, price lists or other numerical content. As the period progressed, the relative percentage of news and miscellany increased with numerical and advertisement notices remaining largely confined to the final two columns despite an increase in the number of columns per page. The mechanics of printing a four-page, single-sheet newspaper allowed the setting and printing of the outer pages of the issue first, leaving the inner sections for more recent reportage.[26] This idea is plausible for the *Mercury* before 1835, with the inner and outer pages presenting distinct content, but is incompatible with the later issues, where parliamentary reports frequently ran across the first and second pages and information derived from recent London newspapers appeared on the back page. Even in the earlier issues, the chronology of reportage across all four pages, and the insertion dates of advertisements on the first, suggests that the inner and outer pages of each issue were set and printed at roughly the same time.

The inner pages consistently contained most of the news content. Page two allocated over 60% of the items and 85% of the word count to news content, nearly all of which referred to the royal family or Parliament. The remainder came from the *London Gazette*, as Williams noted, discussing recent appointments, bankruptcies, and stock prices. Similarly, 63% of the items appearing on page three were categorised as news, though with a more even split of metropolitan (32%), Scottish (36%) and international (11%) coverage. The remainder of items were local advertisements or sponsored content (11%), numerical content such as price lists (11%), and commentaries or miscellany (15%). Between 1820 and 1840, the *Mercury* saw an increase in parliamentary coverage and commentary, and its concentration on the second page, with a commensurate decrease in inner-page advertisements—only a handful appeared in the 1830s and none in 1840. Soundings from other seasons suggest that despite cyclical fluctuations in shipping, politics, and commerce, the distribution of content was relatively consistent year-on-year, making it suitable for computational analysis.

The relationship between content and placement is immediately clear but a finer cataloguing also reveals the way the *Mercury*'s editors reused and labelled existing materials. Cataloguing began with the disambiguation and numbering of individual texts (hereafter referred to as snippets) by page number, column number (left-to-right), and snippet number (top-to-bottom). These snippets were then coded for five characteristics: the type, the topic, the location of the action discussed, the source of the material, and the word count as determined through optical character recognition. Type was separated into five fixed categories—advertisements and notices, commentaries, miscellany, news, and numerical content—while topic was left open ended. Place of action, derived from a close reading of the snippet, was only entered if explicitly mentioned and left at the resolution of that reference—

leading to some locations being listed as a country and others as a province or city. The source of the material was recorded as either the specific title, an ambiguous title such as "Paris papers," the name of the correspondent, or the location at the resolution given. In this period, the *Mercury* had very few decorative elements to indicate different sections or textual units and I have relied upon my judgement regarding shifts in topic, as well as typographical and semantic clues, to determine where one snippet concluded and another began. Once the snippets were fully disambiguated and coded, they were re-visualised to provide a simplified facsimile of each page, depicting the placement and length of each snippet, and shaded to represent different source types, allowing for a visual comparison of different pages within and across issues. The results provide a basis for understanding the standard practices of curation—the use of previously published material— and attribution in the *Mercury*.

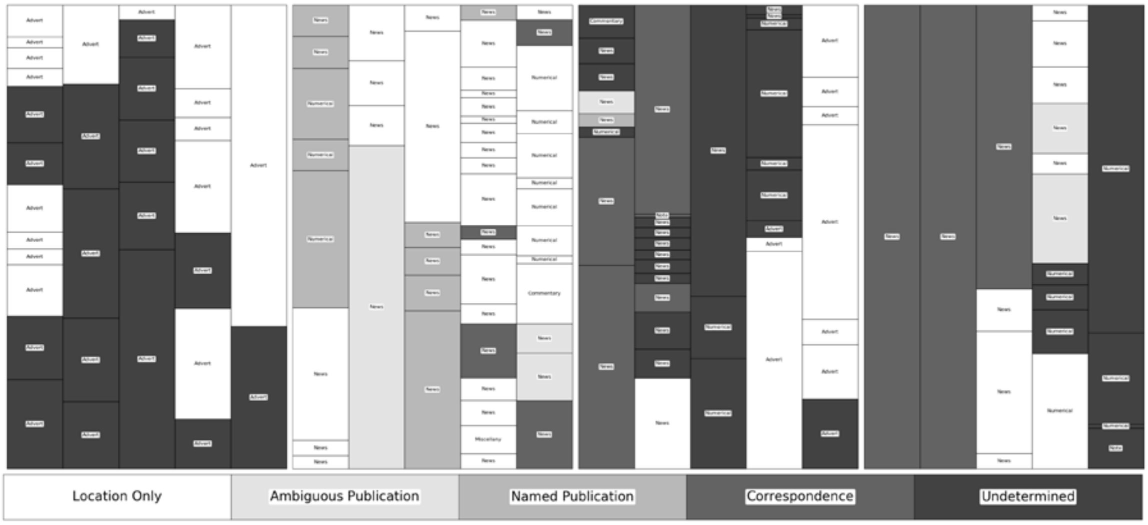


FIGURE 1: RE-VISUALISATION OF THE CALEDONIAN MERCURY FOR JUNE 15, 1820, SHADED BY SOURCE TYPE. HIGH-RESOLUTION AND FULL COLOUR IMAGES AND ASSOCIATED DATA AVAILABLE AT HTTP://DX.DOI.ORG/10.6084/M9.FIGSHARE.5998502.

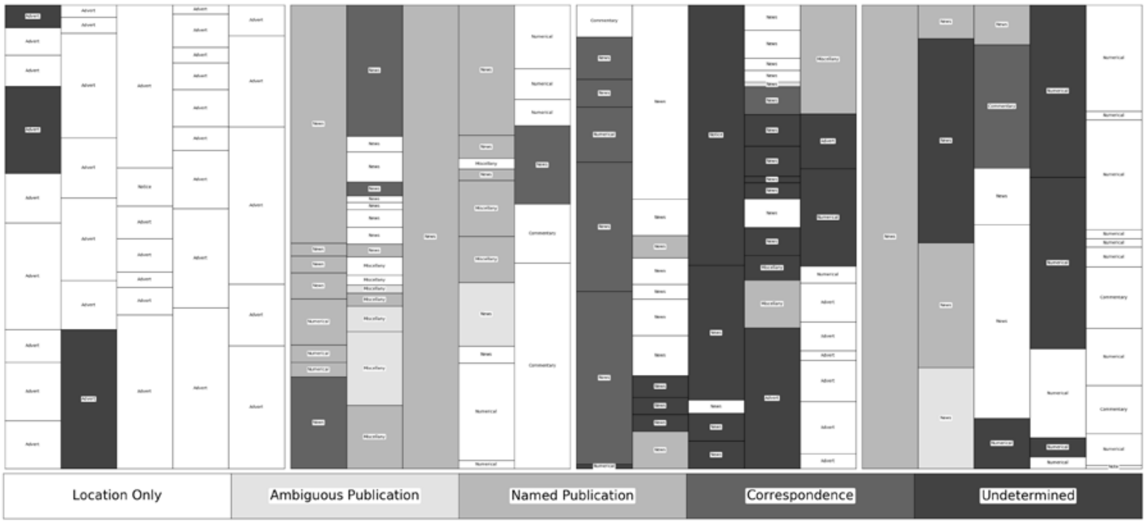


Figure 2: Re-visualisation of the Caledonian Mercury for June 16, 1825, shaded by source type. High-resolution and full colour images and associated data available at http://dx.doi.org/10.6084/m9.figshare.5998496.

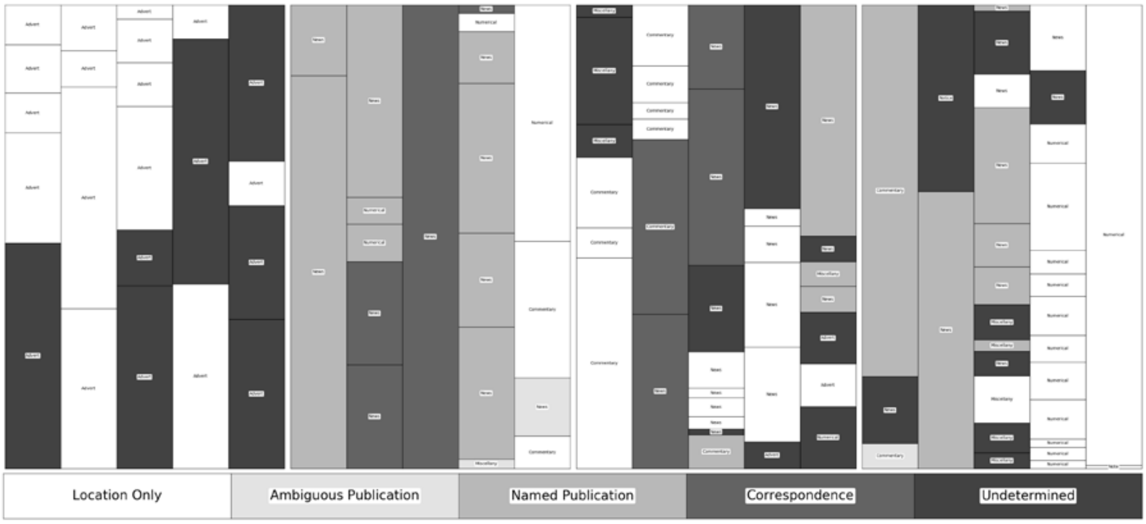


Figure 3: Re-visualisation of the Caledonian Mercury for June 14, 1830, shaded by source type. High-resolution and full colour images and associated data available at http://dx.doi.org/10.6084/m9.figshare.5998493.

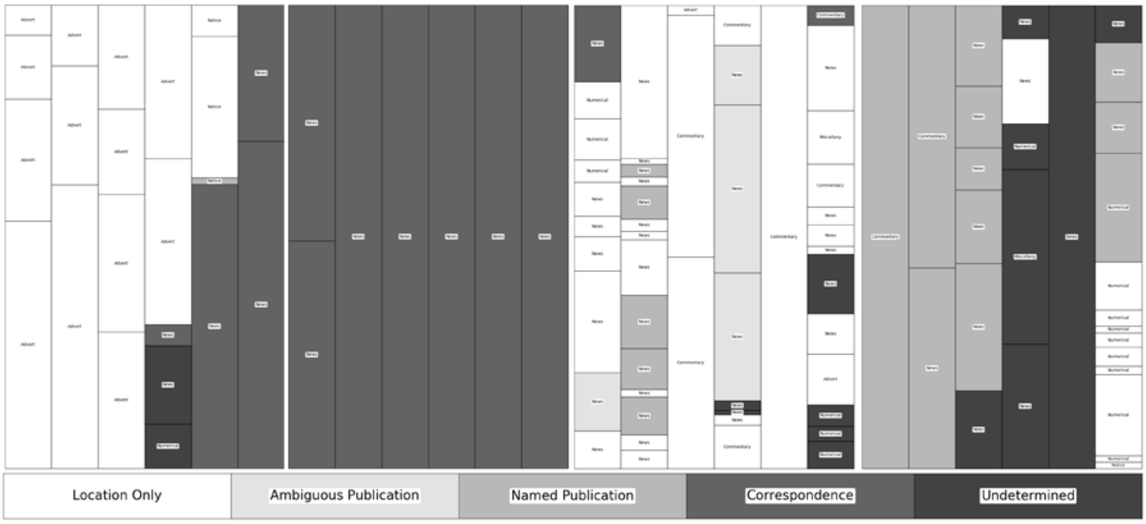


Figure 4: Re-visualisation of the Caledonian Mercury for June 15, 1835, shaded by source type. High-resolution and full colour images and associated data available at http://dx.doi.org/10.6084/m9.figshare.5998454.

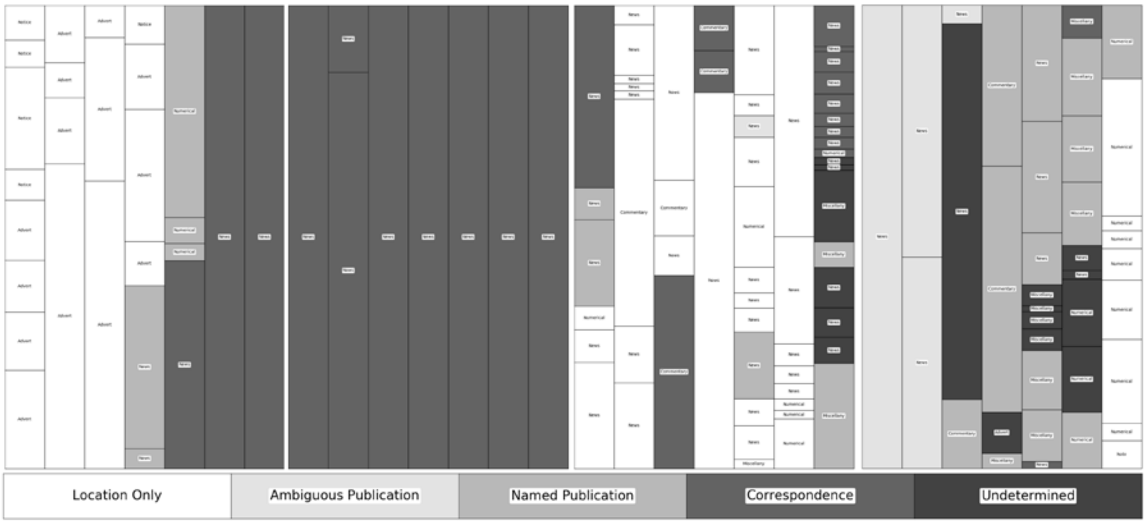


Figure 5: Re-visualisation of the Caledonian Mercury for June 15, 1840, shaded by source type. High-resolution and full colour images and associated data available at http://dx.doi.org/10.6084/m9.figshare.6011597.

Direct comparison across these years has two specific caveats. First, although Allan remained the owner of the *Mercury*, he was not involved in the daily running of it, leaving its management to the editor, who changed twice during the period: in 1827, James Browne was appointed, after which contemporaries commented on an increase in the amount of news that the *Mercury* was able to report ahead of its local competitors; in 1838, the editorship fell to "a knot of young Whig lawyers, suckling politicians and expectant commissioners, who, gratuitously, it is said, furnish the requisite 'leaders.'"[27] These changes mark three distinct eras in the sample period. Second, the 1840 issue was atypical in content owing to coverage of the attempted murder of Queen Victoria, which saturated the *Mercury* and other newspapers throughout the country, skewing the relative percentage of duplicate material from London. Yet, even with these caveats, clear patterns are visible, providing new insights into the general composition of the newspaper.

First, providing an attribution of some description was the standard practice, with over 80% of non-advertising material containing some indication of its source. However, the spatial placement of these attributions varied without correlation to topic or type. They sometimes appeared as explicit lead-ins, tag lines, or mentions of the source somewhere in the text; other times they were implied through a heading preceding multiple items, particularly in the case of local or regional news. The purpose of attribution is implied by relative proportions of attribution types. 60% of attributions were to geographical locations, through datelines or other textual references, rather than to specific publications or individuals. Moreover, nearly all items lacking an attribution provided a specific British location of action, implying that the material was obtained from that location. This suggests that the main purpose of attribution was to signal the physical rather than intellectual source of the information, to foreground the likelihood of its accuracy without the need for evoking the reputation of individual reporters or institutions.[28] Its use also suggests that early nineteenth-century readers were expected to have or to develop a sense of distances and travel times, as well as the importance of certain locations in the political and economic landscape, in order to properly weigh conflicting, anonymous accounts.[29] Local and regional material presents a distinct case. The *Mercury* drew local content from a significant hinterland, including Edinburgh, the port of Leith, and rural Midlothian, as well as from Perthshire, Fife and the eastern Borders, to which Edinburgh had long-standing economic connections. News from these regions were generally attributed implicitly by reference to the place of action or through the suggestion of direct communication with the newspaper's editor, being placed beneath a sectional masthead reading *Caledonian Mercury*. The likelihood that these stories were obtained through oral transmission and personal correspondence makes disentangling direct reportage from unattributed duplicate material seemingly impossible for local items.

When the *Mercury* did indicate that duplicate material came from another publication, it did so with either the title (77%) or the location of the publication (23%). In both cases, London was the most prevalent source, with 50% of titles and 37% of location-based attributions. The remaining attributions were divided between Scottish (19%), provincial British and Irish (11%), and international or colonial (11%) titles as well as a smaller number (9%) of monthlies and separate publications, including books and pamphlets. Across the five issues, few publications were mentioned more than once, the notable exception being the *London Gazette* to which nearly 5% of all non-advertising snippets were attributed, most of them appearing under a heading naming that publication. Distinguishing by category, approximately 25% of news items were attributed to a publication, whereas only 14% of numerical content and notices were—all of which came from the *London Gazette* and *Lloyd's List*. By far the most likely to be attributed to an earlier publication were commentaries, miscellany, and human-interest stories, which may have been considered creative or analytical works and therefore garnered greater legal protections or moral rights, a notion requiring further research. The prevalence of attributing miscellany material may also help account for the relatively low percentage of attribution in 1820, which included only one item of society news and two short commentaries on Parliament and the royal family, attributed geographically or not at all.

If we include only snippets that are in some way attributed to another publication, we find that on average 21% of each issue, as determined by OCR word count, was duplicate material, a significantly higher proportion than the corpus average of 3-4% discovered with Copyfind. The average wordcount of these snippets was 380, at the high end of our expectations, with only a single piece over 1,000—suggesting that the increase in matches over 1,000 words at the end of the period was owing to the use of multiple snippets from a single source rather than the reuse of a single long snippet.[30] The 1840 issue, despite a change in editorship and the extended coverage of the attempted regicide, contained the same proportion of explicitly duplicated material; only the 1820 issue contains a lower proportion (11%), explained by the relatively high proportion of private correspondence used. As for the correlation between Browne's editorship and an increase in early reports, this sample failed to capture any significant decrease in the amount of previously published text, although the space provided to Parliamentary correspondence did increase.

Finally, the paucity of Scottish titles within the corpus may also help account for the disparity between the manual and automatic identification of duplicate material. Across the period, fourteen different Scottish newspapers were explicitly attributed, including the *Aberdeen Journal* and *Glasgow Herald*, which were included in the corpus. Of the others, only the *Perth Courier* has been fully digitised for this period, as part of the *British Newspaper Archive*. Most of the others have been digitised but only for the period after 1840, while the *Galloway Register, Glasgow Chronicle,* and *Glasgow Courier* remain un-digitised. However, the items attributed to Scottish newspapers tended to be short in comparison with overseas or English content, representing only 11.1% of publication-attributed snippets and 2.3% of the entire sample, and many fell below the minimum word count set for automatic identification. The exclusion of Scottish titles from the corpus is, therefore, unlikely to have a significant negative effect on the overall computational analysis.

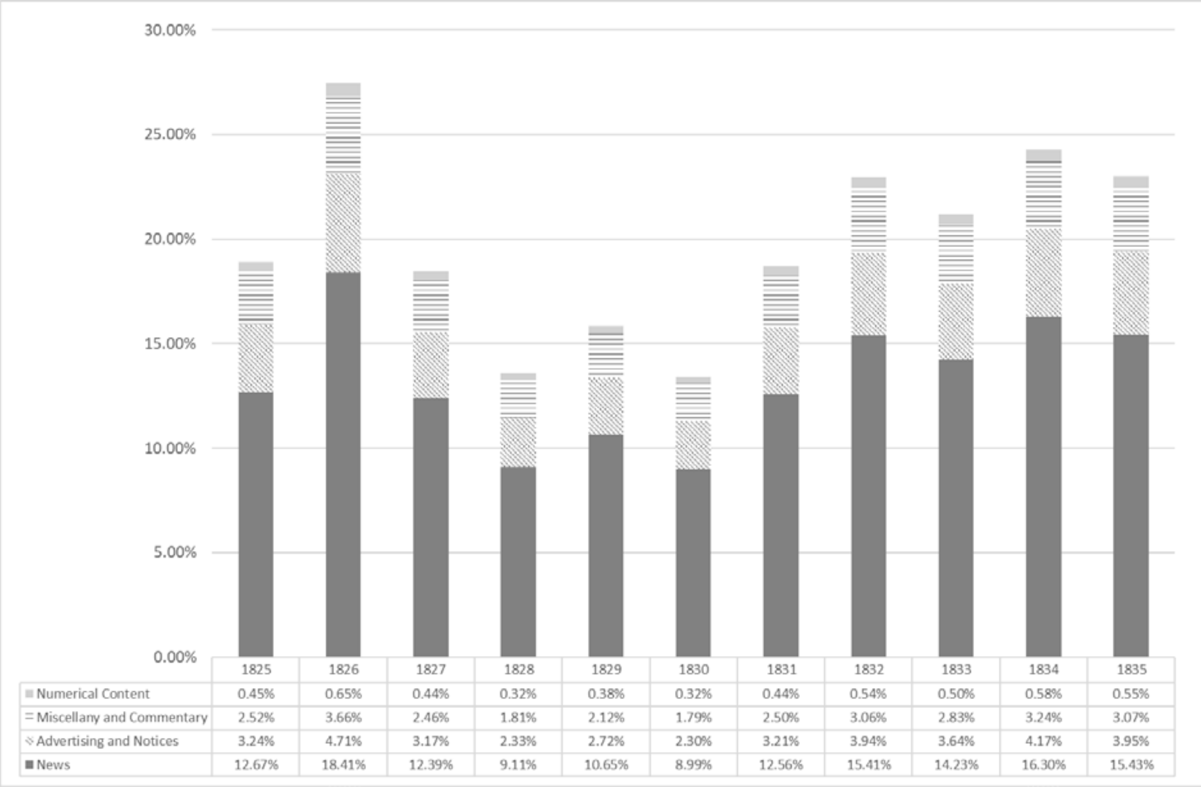
Having examined the sample in detail, we are left with the following conclusions. First, the automatic identification of duplicate material appears to have consistently returned only 15-20% of attributed, and ostensibly duplicated, material. Second, over 80% of non-advertising items were in some way attributed, with roughly 25% giving the name of the source publication. Of the latter, numerical items were relatively less likely and non-news items relatively more likely to be attributed than the average. Third, change over time and editorial control appears to have minimally impacted the percentage of attributed material. Finally, the geographical composition of the corpus does not appear to have significantly biased the likelihood of computational identification.

EXTRAPOLATING FROM CLOSE AND DISTANT READINGS

When the sample issues of the *Mercury* are computationally compared with the wider corpus of digitised newspaper content—all pages from the months of May, June, and July of the selected years—ninety-four instances of duplicate text were found. Of these, twelve (12.7%) were false positives, pointing to similar but non-identical parliamentary transcriptions, and one was an outright false positive. From the opposite perspective, 10% of the items catalogued in the sampled issues were shown to have also appeared in other newspapers within the corpus. Of these, only 47% were matched to publications that were dated at least two days before the *Mercury*; the remainder were printed either simultaneously or subsequently and therefore cannot be inferred as a possible source of the material. Thus, within these specific issues, roughly 5% of the items and 3.8% of the word count were plausibly matched to a source within the corpus, similar to the initial 3-4% estimate but far below that calculated by close reading. Moreover, unless one generously matches vague attributions such as "London Papers" to specific metropolitan titles, no source was confirmed by both explicit attribution and computational matching. For example, a small number of pieces attributed to "London" or without attribution were found in previous issues of the

Times and the *Chronicle* but none of the texts explicitly attributed to the *Chronicle* or the *Times* were found computationally owing to their short length.[31] This suggests that some items were popular enough to appear in multiple publications before appearing in the *Mercury,* even if their actual originator was not included in the corpus. It also introduces the intriguing possibility of a correlation between some popularising characteristic within the text and a purposeful ambiguity or occlusion of its source, similar to patterns found in urban legends. Regardless, the rapid duplication of texts within competitive markets, particularly London, the short length of certain texts, and the possibility of parallel dissemination pathways makes identifying specific lineages solely through computational matching extremely hazardous.

In terms of category, 70% of computational matches were to news items, with 15% catalogued as miscellany, 12% as advertisements, and 3% as numerical content. Within these news matches, two thirds of those indicating a previous printing had been given a publication attribution by the *Mercury*, whereas two thirds of those indicating only simultaneous or subsequent matches were attributed geographically. This suggests that at least one common source of news is missing from the corpus, and, more interestingly, that material from such sources was less likely to be explicitly attributed. The results also prompt speculation on the placement of the *Mercury* within the wider network. Over half of the news articles appeared simultaneously or subsequently with the corpus. If the corpus provides a representative sample, this implies that the *Mercury* existed near the centre of the dissemination timeline, receiving non-local news earlier than many other provincial publications, even if there is little computational evidence that it acted as a direct source of that material to others. As for evidence of biases within the corpus, only one locally authored story was computationally identified elsewhere despite frequent references to the *Mercury* in other Scottish publications.[32] This may indicate that the *Mercury* was only used as a source by other Scottish newspapers, which were not included in the predominantly English corpus.



	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
■ Numerical Content	0.45%	0.65%	0.44%	0.32%	0.38%	0.32%	0.44%	0.54%	0.50%	0.58%	0.55%
= Miscellany and Commentary	2.52%	3.66%	2.46%	1.81%	2.12%	1.79%	2.50%	3.06%	2.83%	3.24%	3.07%
▨ Advertising and Notices	3.24%	4.71%	3.17%	2.33%	2.72%	2.30%	3.21%	3.94%	3.64%	4.17%	3.95%
■ News	12.67%	18.41%	12.39%	9.11%	10.65%	8.99%	12.56%	15.41%	14.23%	16.30%	15.43%

Figure 6: Likely percentage of the OCR word count of the average issue of the Caledonian Mercury for a given year to be duplicate material, shaded by content type. Data available at http://dx.doi.org/10.6084/m9.figshare.6011630.

Scaling these results to a wider analysis of all *Mercury* issues requires caution but offers tantalising results.[33] In 1820 and 1840, Copyfind successfully identified 15% of duplicate items; conversely, in 1825, 1830, and 1835 it identified only 6%. When comparing word counts, the discrepancy between these dates is even greater, 17% and 44% compared with 10%. This, along with other differences in the 1820 and 1840 issues, makes it prudent to begin with a smaller range, 1825-1835. Based on the snippet catalogues, at least 21% of these issues should be composed of duplicate text. Scaling the computationally matched word count for each year by our sample-derived proportions, we find that the reality varies but roughly aligns to our sample percentage. Duplicated news content likely averaged 13% of an issue, with advertising and miscellany at roughly 3% each; numerical content, difficult to represent through OCR transcriptions, would likely account for less than 1% of the computationally-derived word count. Longitudinally, 1827 shows a dip followed by a slow rise, coinciding with the appointment of Browne as editor, suggesting his reputation for early reports may be merited.

CONCLUSIONS

As with all iterative processes, additional sampling will further refine these scaling factors, providing ever-more accurate representations of the *Mercury* and its content—understanding seasonal variations and the effect of key events being important next steps. Yet, it is already clear that iterative distant reading can provide new insights into the limits and considerations one should address during subsequent close readings and that iterative close reading is crucial for the improvement of large-scale queries.

At the start of this study, distant reading signalled the relative value of studying the *Mercury* over its London contemporaries, owing to its general adherence to corpus-wide trends, and highlighted the need to quantify the distribution of short, medium, and long snippets, particularly in 1840, in order to interpret the results of computational matching. Careful sampling and targeted close reading then clarified the distribution of computational matches among different snippet types and lengths, allowing new hypotheses about the attractiveness of certain traits and the specific limits of the current matching processes, increasing the value of a general reckoning of the scale of textual reappearance. It also allowed comparisons between two independently created lists of duplicate texts, leveraging the strengths of both interpretative (human) and literal (computational) identification to build the most complete catalogue possible. Moreover, correlating attribution types with the number of previous and subsequent appearances within the corpus hinted at both the completeness of the collection and the relationship between attribution and the virality of certain text types.

Most importantly, the iterative development and testing of scaling formulae allow us to understand the wider applicability of our soundings and samplings in new, more nuanced ways. The ambiguous and shifting definition of a newspaper in the nineteenth century often makes generalisation a counterproductive pursuit. Instead, a better understanding of correlative and interacting factors at the small scale, alongside the footprints of these at the large, may help us understand the press as something other than sum of its parts.

BIBLIOGRAPHY

Beals, M. H. “The Role of the Sydney Gazette in the Creation of Australia in the Scottish Public Sphere.” In *Historical Networks in the Book Trade*, ed. Catherine Feely and John Hinks, 145-166, Basingstoke: Routledge, 2016:

Beals, M. H. “Scissors and Paste: The Georgian Reprints, 1800–1837,” *Journal of Open Humanities Data* 3 (2017): 1-8.

Beals, M. H. “Stuck in the Middle: Developing Research Workflows for a Multi-scale Text Analysis.” *Journal of Victorian Culture* 22, no.2 (2017): 224-231.

Brownlees, Nicholas. “‘Newes also came by Letters’: Functions and Features of Epistolary News in English News Publications of the Seventeenth Century.” In *News Networks in Early Modern Europe*, ed. Joad Raymond and Noah Moxham, 349-410. Leiden: Brill, 2016.

“Caledonian Mercury, The; (1720 - 1859).” In *The Waterloo Directory of Scottish Newspapers and Periodicals: 1800-1900*, s. v. Accessed December 1, 2017, http://scottish.victorianperiodicals.com.

Clark, Charles E., and Charles Wetherell. “The Measure of Maturity: The Pennsylvania Gazette, 1728-1765.” *William and Mary Quarterly* 46, no. 2 (1989): 279-303.

Conboy, Martin, Joad Raymond, Kevin Williams and Michelle Tusan "Roundtable Discussion of Martin Conboy's Journalism: A Critical History, London: Sage, 2004. (x + 246 pp., ISBN 0761940995, $115 (hbk); 0761941002, $38.95 (pbk)." *Media History* 12, no. 3 (2006): 329-51.

Cowan, Robert M. W. *The Newspaper in Scotland: A Study of Its First Expansion, 1816-1860*. Glasgow: G. Outram & Company, 1946.

Curran, James, and Jean Seaton. *Power Without Responsibility: Press, Broadcasting and the Internet in Britain*. London: Routledge, 2009.

Curwen, Henry. *A History of Booksellers: The Old and the New*. Cambridge: Cambridge University Press, 2010.

Dicken-Garcia, Hazel. *Journalistic Standards in Nineteenth-Century America*. Madison, Wis.: The University of Wisconsin Press, 1989.

"Examiner, The; (1808 - 1881)." In *The Waterloo Directory of English Newspapers and Periodicals: 1800-1900*, s. v. Accessed December 1, 2017, http://english.victorianperiodicals.com.

Freely, Catherine. "'Scissors-and-Paste' Journalism." In *Dictionary of Nineteenth Century Journalism in Great Britain and Northern Ireland*, ed. Laurel Brake and Marysa Demoor, 561. London: Academia Press, 2009.

Feely, Catherine. "'What say you to free trade in literature?' The Thief and the Politics of Piracy in the 1830s." *Journal of Victorian Culture* 19, no. 4 (2014): 497-506.

Garvey, Ellen Gruber. *Writing with Scissors: American Scrapbooks from the Civil War to the Harlem Renaissance*. Oxford: Oxford University Press, 2012.

Gooding, Paul. *Historic Newspapers in the Digital Age: "Search All About It!"* Basingstoke: Palgrave, 2017.

Greengrass, Mark, Thierry Rentet and Stephane Gal, “The Hinterland of the Newsletter: Handling Information in Space and Time.” In *News Networks in Early Modern Europe,* ed. Joad Raymond and Noah Moxham, 616-40. Leiden: Brill, 2016.

Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema and Gerik Scheuermann. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.” *Eurographics Conference on Visualization* (2015): E1-21.

Katrina Navickas and Adam Crymble “From Chartist Newspaper to Digital Map of Grass-roots Meetings, 1841–44: Documenting Workflows,” *Journal of Victorian Culture* 22, no.2 (2017): 232-247.

Krippendorff, Klaus. *Content Analysis: An Introduction to Its Methodology*. London: Sage, 2004.

Liddle, Dallas. “Reflections on 20,000 Victorian Newspapers: ‘Distant Reading’ *The Times* using *The Times Digital Archive*.” *Journal of Victorian Culture* 17, no. 2 (2012): 230-237.

Matheson, Donald. "The Birth of News Discourse: Changes in News Language in British Newspapers, 1880-1930." *Media Culture & Society* 22, no. 5 (2000): 557-573.

Mayer, Henry. *The Press in Australia*. Melbourne: Landsowne Press, 1964.

McMinn, W. G. "A Newspaper Index Report." *Journal of the Royal Australian Historical Society* 53 (1968): 69-71.

“Morning Chronicle and London Advertiser, The (1770 - 1865).” In *The Waterloo Directory of English Newspapers and Periodicals: 1800-1900,* s. v. Accessed December 1, 2017, http://english.victorianperiodicals.com.

Mussell, James. *The Nineteenth-Century Press in the Digital Age*. Basingstoke: Palgrave, 2012.

"The Newspaper Press of Scotland." *Fraser's Magazine for Town and Country* 17 (1838): 559-570.

Nicholson, Bob. "Counting Culture; or, How to Read Victorian Newspapers from a Distance." *Journal of Victorian Culture* 17, no. 2 (2012): 238-246.

Nicholson, Bob. "The Digital Turn" *Media History* 19, no. 1 (2013): 59-73

Nicholson, Bob. "'You Kick the Bucket; We Do the Rest!': Jokes and the Culture of Reprinting in the Transatlantic Press," *Journal of Victorian Culture* 17, no. 3 (2012): 273-86.

Paul, Sir James Balfour. *The History of the Royal Company of Archers: The Queen's Body-guard for Scotland*. Edinburgh: William Blackwood and Sons, 1875.

Piesse, Jude. *British Settler Emigration in Print*. Oxford: Oxford University Press, 2016.

Pigeon, Stephan. "Steal it, Change it, Print it: Transatlantic Scissors-and-Paste Journalism in the Ladies' Treasury, 1857–1895." *Journal of Victorian Culture* 22, no. 1 (2017): 24-39.

Rantanen, Terhi. "The New Sense of Place in 19th-Century News." *Media, Culture & Society* 25 (2003): 435-449.

"Scotch Newspaper Press." *The Westminster Review* 12 (1830): 82-85.

Seville, Catherine. *Literary Copyright Reform in Early Victorian England: The Framing of the 1842 Copyright Act*. Cambridge: Cambridge University Press, 1999.

Shattock, Joanne, and Michael Wolf, eds., *The Victorian Periodical Press: Samplings and Soundings*. Leicester: Leicester University Press, 1982.

Smith, David A., Ryan Cordell, and Abby Mulle. "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers." *American Literary History* 27, no. 3 (September 1, 2015): E1–15.

[1] Mussell, *Press in the Digital Age*, 1; Nicholson, "Digital Turn," 61; Gooding, *Historic Newspapers*, 172. The author would like to express her gratitude to Will Slauter, Paul Fyfe and the participants of the "Copying and Copyright in Nineteenth-Century Newspapers and Periodicals" workshop held at Université Paris Diderot in March 2017 for their thoughtful comments and suggestions on an earlier draught of this article. She would also like to thank Geraint Palmer for his assistance in the development of the visualisation software employed in this study.

[2] Shattock and Wolff, *Victorian Periodical Press*, xvi.

[3] Conboy, Raymond, Williams and Tusan "Roundtable Discussion," 340. A straightforward comparison of historical and modern newspaper methods of sampling can be seen in Dicken-Garcia, *Journalistic Standards*, 66-7 and Curran and Seaton, *Power Without Responsibility*, 43, 52. For a discussion of robust newspaper sampling, see Krippendorff, *Content Analysis*, 111-21.

[4] Piesse, *British Settler Emigration*, 16.

[5] McMinn, "A Newspaper Index Report," 70.

[6] Liddle, "Reflections," 235.

[7] Nicholson, "Counting Culture," 243-4.

[8] Freely, "'Scissors-and-Paste' Journalism," 561.

[9] Pigeon, "Steal it," 27; Garvey, *Writing with Scissors*, 238; Feely, "The Thief," 503.

[10] Beals, "Sydney Gazette," 152-54; Nicholson, "'You Kick the Bucket," 277.

[11] Smith, Cordell, and Mullen, "Computational Methods."

[12] Cowan, *Newspaper in Scotland*, 166; Seville, *Literary Copyright Reform*, 139.

[13] Data used in this study can be obtained from http://www.github.com/mhbeals/sap_reprints. A detailed discussion of the limits used can be found in Beals, "Scissors and Paste."

[14] Each year, the *Caledonian Mercury* included a short notice on the annual archery contest and the awarding of the silver "Edinburgh Arrow." As the event was constitutionally set for the "Second Monday of June," it acts as a marker of cyclical similarity between the issues. The results were consistently reported in the issue nearest June 15 apart from 1820, when the archery contest was postponed to the "next fair Monday" and therefore not reported until July 6, 1820. For consistency sake, June 15, 1820 was used rather than the issue recording the contest, making the issues sampled June 15, 1820, June 16, 1825, June 14, 1830, June 15, 1835, and June 15, 1840. Paul, *Royal Company of Archers*, 315.

[15] Jänicke, Franzini, Cheema and Scheuermann, "On Close and Distant Reading," 2.

[16] For publisher details on these datasets, see http://gale.cengage.co.uk/british-library-newspapers/19th-century-british-library-newspapers-part-i.aspx and http://gale.cengage.co.uk/times.aspx.

[17] For a discussion of correcting OCR transcriptions in this database, see Navickas and Crymble "From Chartist Newspaper," 239.

[18] While the removal of same-title matches reduced the number of advertisements in the final report, advertising material appearing in multiple publications, namely those for national lotteries or patent medicines, were still captured by the automatic matching process.

[19] For a fuller discussion of the rationale behind the methodology employed, see M. H. Beals, "Stuck in the Middle".

[20] The most likely pairing in the corpus was the *Chronicle* with the *Times*, followed by the *Chronicle* and the *Examiner.* The number of instances between the *Chronicle* and the

Mercury was essentially tied for third with several other provincial English pairings, but represents the most common pairing for the *Mercury*, as it had for the *Examiner*. The number of instances of textual reappearance between these three titles rose from a hundred per month in the 1820s to over three hundred by 1840. The increase over this period mirrored the increased number of pages per issue and therefore does not necessarily suggest closer linkages between the publications but rather a proportional increase in their overlapping content.

[21] “Morning Chronicle”; “Examiner”; “Caledonian Mercury”; "Scotch Newspaper Press," 84.

[22] Mayer, *Press in Australia*, 11-14.

[23] Clark and Wetherell, “Measure of Maturity,” 279-303.

[24] Ibid., 295.

[25] Williams, *Dangerous Estate*, 50-1.

[26] Huntzicker, *Popular Press*, 102.

[27] Curwen, *History of Booksellers*, 145-6; "Scotch Newspaper Press," 84; “The Newspaper Press of Scotland,” 563.

[28] Brownlees, “Epistolary News,” 408; Rantanen, "Sense of Place,” 437-41.

[29] Greengrass, Rentet and Gal, “Hinterland of the Newsletter,” 617; Matheson, "Birth of News Discourse,” 566.

[30] “The Murderous Attempt on The Queen,” *The Caledonian Mercury*, (June 15, 1840): 4.

[31] "We copy the above from the Observer", *Caledonian Mercury* (June 15, 1820): 2; "A correspondent of the Times", *Caledonian Mercury* (June 16, 1825): 2; "The Deputies", *Caledonian Mercury* (June 15, 1835): 3; "SPAIN.", *Caledonian Mercury* (June 15, 1835): 4.

[32] Beals, “Sydney Gazette,” 153.

[33] The multiplication factor used was determined by comparing the word count for the highest computational match of each page with the word count of text manually categorised as duplicate material, either through individual Copyfind reports or through explicit attributions in the text. This provided an average multiplication factor of 478.9%. Weighted proportionally, every 1,000 words of duplicate text reported by Copyfind in its maximum match likely represents 3,666 words of news, 737 words of commentary or miscellany, 414 words of advertisements, and 297 words of numerical content.