*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 5865–5878
Marseille, 20-25 June 2022


# Identifying Copied Fragments in an 18th Century Dutch Chronicle

**Eleanor L. T. Smith[1], Lianne Wilhelmus[2], Erika Kuijpers[2], Alie Lassche[3], Roser Morante[2]**
[1] Faculty of Arts, University of Antwerp
Eleanor.Smith@uantwerpen.be
[3] Faculty of Humanities, Leiden University
a.w.lassche@hum.leidenuniv.nl
[2] Faculty of Humanities, Vrije Universiteit Amsterdam
{erika.kuijpers, r.morantevallejo}@vu.nl; l.e.wilhelmus@student.vu.nl

**Abstract**

We apply computational stylometric techniques to an 18th century Dutch chronicle to determine which fragments of the manuscript represent the author's own original work and which show signs of external source use through either direct copying or paraphrasing. Through stylometric methods the majority of text fragments in the chronicle can be correctly labelled as either the author's own words, direct copies from sources or paraphrasing. Our results show that clustering text fragments based on stylometric measures is an effective methodology for authorship verification of this document; however, this approach is less effective when personal writing style is masked by author independent styles or when applied to paraphrased text.

**Keywords:** authorship verification, stylometry, chronicles, Dutch

## 1. Introduction

Recent studies have expanded our understanding of the production, circulation and marketing of news media in Early Modern Europe (c. 1500-1800) (Blair, 2010; Pettegree, 2014; Blair et al., 2021). Far less is known about the *reception* of news and information. How and when did the wider public change their media consumption in this period? An interesting source of information about reading habits and media use are local chronicles. These are hand-written narratives produced by in majority middle class authors, who recorded events and phenomena they considered important and memorable (local politics, upheavals, climate, prices, crime, deaths) (Pollmann, 2016). To do so, authors made use of a wide array of sources and frequently copied older chronicles and history books, excerpts from official documents, local announcements, by-laws and broadsheets, as well as of newspapers and periodicals that became increasingly popular in the course of the eighteenth century. Some chroniclers clearly mention where they got their information, for instance from a newspaper, in church, at the local market or through official government publications. More often authors did not explicitly quote their sources. There were no conventions concerning referencing, and copy rights still had to be conceived of.

This paper explores whether computational techniques of authorship verification can be used to detect copied fragments in chronicles, and thus contribute to our understanding of media use through the practice of copying by early modern chroniclers. More specifically, we apply computational stylometric techniques to an 18th century Dutch chronicle to determine which fragments of the manuscript represent the author's own original work and which show signs of external source use through either direct copying or paraphrasing.

The chronicle is part of a recently digitized corpus[1] that contains c. 320 early modern Dutch language chronicles written between 1500 and 1850 in what are now the Netherlands and Belgium. These manuscripts have been digitized with the help of volunteers and automatic handwritten text recognition.

For this research we selected a chronicle of the small town of Purmerend,[2] written by wine seller Albert Pietersz Louwen (1722-1798) in the period 1766-1791. This period was marked by the rise of a varied media landscape including periodicals and opinionated magazines and heated public debates leading to the revolutionary era. The chronicle consists of five volumes divided in three parts, each containing different types of information. The first part covers the history of the town of Purmerend since its Medieval origins, while the second part describes contemporary events during the period 1766-1780. Part three, covering the revolutionary years from 1775 on-wards, is strongly politically oriented, as Albert Louwen became invested as a 'patriot' in the Batavian Revolution (1794-98). Reasons for selecting Louwen's chronicle to perform authorship verification experiments on, were *(i)* its large size, *(ii)* his regular references to both historical and contemporary sources, and *(iii)* his frequent use of inverted commas, which have been hypothesised to indicate copied or paraphrased sections.

Our main **research question** is to what extent are computational authorship verification techniques useful to differentiate between text by the author in his own words and text copied from other sources, in an eighteenth century Dutch chronicle. To answer this question, we need to determine how reliable the author's quotation marks

[1] https://chroniclingnovelty.com/
[2] Purmerend lies 30 km North of Amsterdam.

are to differentiate between copied and non copied text, and whether it is useful to compare fragments by the author with fragments by external sources.
This work is innovative in that no previous study has applied computational stylometry techniques for authorship verification to a historical Dutch chronicle that is characterized by various practices of digesting and reproducing news and information. The following sections first provide both historical and computational related work. Following this, the full methodology of the project and data used, as well as the experimental set up are described, before the results are outlined and discussed. The final section of the paper provides concluding statements relating the findings to the research questions listed above.
The data and code of the research presented in this paper can be found at `https://github.com/chroniclingnovelty/stylometry-lrec22`.

## 2. Related Work

### 2.1. Historical Work

Medieval studies into literary and historical texts have a long tradition in authorship attribution (Besamusca et al., 2016). Medieval authors often remained anonymous while their manuscripts were frequently copied manually for distribution. While analogue codicological and heuristic approaches remain important to identify medieval authors, it comes as no surprise that medievalists are also pioneers in the development of digital methods for the automatic comparison of various copies, authorship verification and the detection of plagiarism in historical texts (Besamusca et al., 2016; Birnbaum et al., 2017; Jänicke and Joseph Wrisley, 2017). The authors of 16th to 18th century chronicles have much in common with their medieval predecessors in that they frequently copy other texts. The origin and transmission of narratives, source use and the circulation of manuscripts is far less studied for the period after the introduction of the printing press. Some studies on German (Rau, 2002; Schmidt, 1958; Mauer, 2001), English (Waddell, 2018; Waddell, 2020) and Dutch (Blaak, 2009) chronicles pay attention to the practice of chronicling and the sources used by their authors, but in all these case studies a qualitative approach is used. A computational method to distinguish the author's own voice from the texts he copies or paraphrases, would allow us to analyse and compare a larger set of chronicles in the future and map changes in source use over time.

### 2.2. Computational Stylometry

Stylometry can be defined as the statistical analysis of variations in literary style between one writer or genre and another. Stylometry relies on the theory of the human stylome. This theory can be summarised as an idiolect in writing (Daelemans, 2012), in that each individual is seen to have a unique combination of traits in their writing style which can be used to identify them, much like genetic traits in the human genome. This is translated into the field of computational stylometry, as the practice of attributing texts to individual authors or specific author traits using models that recognise writing style (Daelemans, 2012).

The work of Kestemont et al. (2013) is a key case study in applying computational stylometric clustering methods to historical text samples. The authors do not answer an authorship verification question, but focus on authorial collaboration and the synergy hypothesis, which states that texts resulting from collaboration between individuals can display a style markedly different from that of any one of the collaborating authors.

Another key work for this paper is Dalen-Oskam (2014). This case study applies computational stylometry to an 18th century Dutch novel written jointly by two authors. The aim of the study was to distinguish how the two authors divided the task of writing the novel. Future work on authorship tasks using historical data should follow the example of Kestemont et al. (2013) and Dalen-Oskam (2014) in the methodical approach to clustering. Establishing an evaluation of the method by using non-disputed texts before adding the questioned data into the experiments provides a strong foundation for hypothesising the provenience of questioned fragments. By considering the historical linguistic component of the data, Kestemont et al. (2013) provide an excellent example of how feature selection when using historical texts cannot be based only on the current state of the art features for a given task. The contextual knowledge of strong spelling variation rules out the use of character level n-grams for their case. Finally both Kestemont et al. (2013) and Dalen-Oskam (2014) show how hypotheses from qualitative analysis and the results of quantitative clustering experiments can work together to form a strong evidence base for authorship attribution.

Kestemont et al. (2013) and Dalen-Oskam (2014), along with many others, utilise the computational tool Stylo (Eder et al., 2016) to run the experiments and present the results of their studies. This package is built specifically for stylometric experiments, with questions of authorship in mind, making it an ideal candidate for use in this project as its design takes into account the specifics of this highly specialised task. The fact that both the feature extraction process and the stylometric experiments can be run using this tool ensures that results are highly reproducible, with users able to re-run identical experiments as long as they have the same data and the Stylo settings used in the original experiment are reported. One issue with using Stylo for visualisation is that it is unable to show the results of large data sets, with some figures becoming unreadable with as few as 50 text samples. In order to generate readable results for larger data sets, like that of this study, the output of Stylo must be represented in a different way. In this study we used the network visualisation and analysis software Gephi (Bastian et al., ).

| Text Type | Abbreviation | # Fragments | # Tokens |
|---|---|---|---|
| Chronicle Section 1 | *chr-1* | 516 | 157,200 |
| Chronicle Section 2 | *chr-2* | 257 | 171,230 |
| Other texts by Author | *author* | 18 | 4,296 |
| Historical Sources | *source-hist* | 37 | 88,742 |
| Contemporary Sources | *source-contemp* | 21 | 29,622 |
| ALL | - | 849 | 451,090 |

Table 1: Number of fragments and topics of each text type.

## 3. Methods

Authorship attribution is based on the comparison of text fragments. One of the main methodological decisions when setting up these experiments is how to divide the chronicle in fragments. Ideally, one should be able to make the split without relaying on typographical characteristics of a specific chronicle. However, in order to test whether our methods are valid for this task, we have assumed predefined fragments, which we have obtained based on the use of inverted commas by the author throughout the chronicle. Based in the modern use of inverted commas to denote quotation, our hypothesis is that the chronicler uses inverted commas to signal that information from an external source is being used. Following this, we would expect fragments of text found within inverted commas in the original manuscript to be copied or paraphrased text and text fragments outside of the inverted commas to show the original words of the chronicler. We rely on this 'inverted commas hypothesis' to split the chronicle into fragments.
In future work we will extend our experiments to finding copied text through stylometric variation within a chronicle without relying on non-generalizable information. A possible method would be to subdivide a chronicle in random fragments and evaluate the clustering manually based on expert (historical) knowledge.

### 3.1. Data

We use five text types for our experiments: two types belong to Louwen's chronicle, two types are taken from the published sources that Louwen mentions in the chronicle, and one type is made up of other writings by Louwen. The total number of fragments per text type is given in Table 1.
Fragments from the chronicle text were generated by splitting chronicle sections based on the inverted comma hypothesis. The sources texts were collated by the historians in our team, who collected digital copies from the sources (books, newspapers, governmental publications) mentioned by Louwen in the chronicle[3] and selected short passages at random for use in the experiments. The additional author texts were transcribed by the historians from the introductory sections of two other non-chronicle manuscripts by Louwen.

- **Chronicle section 1 (*chr-1*):** Text fragments from the first two volumes of the chronicle. This section focuses on the history of Purmerend and the surrounding area. Fragments of text which are found between inverted commas in the original manuscript are referred to as 'copy' fragments and fragments of text which are outside of inverted commas in the original manuscript are referred to as 'no-copy' fragments.

- **Chronicle section 2 (*chr-2*):** Text fragments from volumes 3, 4, and 5 of the chronicle, which focus on describing the contemporary social and political situation of Purmerend in the time of Louwen. Fragments are also classified into 'copy' or 'no-copy' depending on the inverted commas.

- **Historical Sources (*source-hist*):** Fragments of text from accessible sources explicitly mentioned by Louwen in *chr-1*. This set is used as external data which is compared to the possible copied and paraphrased text in *chr-1* to determine if the style of the chronicle fragments is similar to the sources Louwen lists.

- **Contemporary Sources (*source-contemp*):** Fragments of text from contemporary publications mentioned by Louwen in the content of *chr-2*. This is used as external data which is compared to the possible copied and paraphrased text in *chr-2* to determine if the style of the chronicle fragments is similar to the sources Louwen lists.

- **Other texts by author (*author*):** Samples of Louwen's writing taken from two other non-chronicle manuscripts. This is used as external data which is compared to the possible non-copied text in both types of chronicle data (*chr-1* and *chr-2*) to determine if the style of the chronicle fragments is similar to the examples of Louwen's writing style from outside of the chronicle manuscript.

Five data sets are used in the experiments of this project. They are constructed from the text types listed in Table 1. Table 2 shows how the text types are grouped to create these five data sets.

### 3.2. Experimental Set Up

We design a formalised 5 step method for our experiments:

[3]The full list is available through the Github repository.

| Data Set | Chronicle Data | Author Data | Source Data |
|---|---|---|---|
| *External Data 1* | - | *author* | *source-hist* |
| *Dataset1* | *chr-1* | *author* | *source-hist* |
| *Outliers 1* | *chr-1 outlier fragments* | *author* | *source-hist* |
| *External Data 2* | - | *author* | *source-contemp* |
| *Dataset2* | *chr-2* | *author* | *source-contemp* |

Table 2: Groupings of text types used in the clustering experiments.

1. **Stylo:** First the preprocessed data is loaded into the R package Stylo Eder et al. (2016),[4] which is a tool meant for the high level quantitative analysis of writing style using stylometry. Stylo provides a host of settings which can be used to generate features and run clustering and classification stylometric experiments. The 'network' function is used to generate an edge table using the selected feature set and parameters chosen through parameter tuning experiments. This edge table contains pairwise distances for all combinations of texts in the loaded corpus. They are computed using statistics from a bootstrap consensus tree which runs the analysis 10 times for each experiment, increasing the most frequent word setting by 10, from 10 to 100, each time. As such, these distances describe how similar each pair of texts are to each other, on average across 10 conditions, based on the features used in the experiment. This information is used to cluster the texts.

2. **Gephi OpenOrd:** The second step in the method is to load the Stylo generated edge tables into the network visualisation tool Gephi (Bastian et al., ) and visualise the network using the algorithm OpenOrd (Martin et al., 2011). Gephi is used in the case of this project, as the corpus is too large to be visualised in an interpretable way through Stylo. It was decided to specifically use the OpenOrd algorithm due to its emphasis on both local and global structure, combining the power of a multi-level force-directed approach with an edge-cutting strategy which focuses on encouraging node clustering. This is important in the current project, as we are interested in analysing the relationship between both the clusters that form and also the texts included in each cluster. This meant that the algorithm used needed to balance the importance of structure both locally between individual nodes in the network and also globally on a larger scale between clusters.

3. **Gephi Modularity:** Next, one of Gephi's built in analysis features, the modularity function, is run over the network. This function is designed to measure how strongly the network can be divided into clusters. The approach used by the modularity function to determine clusters is the Louvain method (Blondel et al., 2008), this is a heuristic method which optimizes based on modularity score. Modularity is defined as a scalar score between -1 and 1 given to a potential partition based on the density of links inside the clusters as compared to links between different clusters. The closer the score for a potential partition is to 1 the better the balance between closely related nodes inside the cluster and loosely related nodes in neighbouring clusters. The aim being, that all nodes in a single cluster should be more related to each other than to any other nodes in the network.

4. **Python Statistics:** Following this, the output of running the modularity function on the network in Gephi is exported and Python scripts are used to compute the make up of each cluster and the distribution of texts of each data class. The results of these statistical calculations and the aggregated list of all texts included in each cluster are then examined to determine the results of the experiment.

5. **Comparison to Qualitative Results:** The final step is to both analyse the quantitative results generated in the fourth step and compare them with previous experiments, as well as performing a qualitative analysis from a historic perspective. This involves comparing outliers found through the computational process with fragments of the chronicle flagged as suspicious by volunteers and researchers during the transcription and annotation process. In this case, 'suspicious' is used to denote sections of text which the annotators and researchers felt may not follow the inverted comma hypothesis, e.g. a section inside of inverted commas which appears to be in the authors voice when subjected to close reading. In this phase we compare the judgement of the stylometric approach with the judgement of expert readers of the text. When both agree, this is a clear indication of the nature of the fragment. When the quantitative results and the historians disagree on the nature of a text fragment, further analysis by both parties is needed to understand what causes the disagreement.

## 4. Experiments

### 4.1. Stylo Parameter Tuning Experiments

Before the clustering experiments could be completed, we needed to determine the parameter settings in Stylo that would be used for generating the edge tables. We

[4]https://github.com/computationalstylistics/stylo

conducted experiments using a *development corpus* of 50 randomly selected chronicle data fragments, 25 copy and 25 no-copy. As the *development corpus* was generated using an earlier version of the chronicle preprocessing script with different splitting, the fragments used in these parameter tuning experiments do not exist in the final corpus. Therefore, there is no danger of over-fitting to the chronicle data.

All experiments used Burrow's delta (Burrows, 2002) as the distance measure to ensure comparability. Burrow's delta was chosen as the distance measure for the project due to its wide use in the field and its design specifically for authorship questions. We experimented with four different parameters: normal sampling, random sampling, culling, and character n-grams. The final settings for Stylo, shown in table 3, were decided based on multiple rounds of experimentation with the development corpus.

| Stylo Setting | Decision |
|---|---|
| Normal Sampling | None |
| Random Sampling | None |
| Culling | 20% |
| Character n-grams | Character 4-grams |

Table 3: Final Stylo parameter settings used to generate edge tables.

When we experimented with different levels of representation, we found that **character 4-grams** performed the best for the fragments in the development corpus. The decision to represent the features of the documents on a character level was taken as this provided better clustering results, compared to word level representations in our experiments. There is also strong support for this in previous work in the field. The decision to set n to 4 was taken as the the results for character 4-grams showed a higher level of clustering, with fewer documents left unclustered than when character n-grams of different lengths were used.

We decided to use some level of **culling** in our final parameter settings, as all of the culling experiments yielded improvements in results in comparison to no culling. The specific choice of using 20% culling was made after multiple experiments with different levels of culling showed that 20% provided the best results on the *development corpus*.

The choice not to use any **sampling method** was based on the fact that normal sampling was not an option for the project due to the over-representation of 'no-copy' fragments generated with this approach, and that random sampling experiments showed little improvement in results. Further experiments were conducted using character 4-grams with 20% culling and random sampling but when compared to the same settings without any sampling the results remained constant. As such it was chosen not to include any sampling in the parameter settings of the Stylo experiments.

The final decision to impose a **lower length limit** of 100 word level tokens on the fragments was to ensure that there was an adequate number of features per document to determine that any unclustered document was not excluded from clustering purely due to its size.

## 4.2. Clustering Experiments

| Experiments | Data |
|---|---|
| External Data Experiment 1 | *External data 1* |
| External Data Experiment 2 | *External data 2* |
| Chronicle Data Experiment 1 | *Dataset1* |
| Chronicle Data Outlier Experiment | *Outliers 1* |
| Chronicle Data Experiment 2 | *Dataset2* |

Table 4: Experiments carried out.

The clustering experiments in this study follow the method described in Section 3 and are listed in Table 4.

**External Data Experiments**. For these experiments we use *External data 1* and *External data 2*, so the data sets that contain texts not belonging to the chronicle. These experiments are carried out in order to test that the method applied is able to separate texts of known authors without the noise of the chronicle data.

**Chronicle Data Experiments**. For these experiments we use the data sets that contain both texts from the chronicles and from external texts. The aim of these experiments is to find copy and no-copy fragments. The copy fragments should cluster with texts not written by the author of the chronicle.

**Chronicle Data Outlier Experiment**. After Chronicle Data Experiment 1, there were clear outliers in the results. These outliers are defined as chronicle fragments which do not cluster with the type of data that would be expected given their label (copy or no-copy). An example of this would be a no-copy chronicle fragment from *chr-1* inside of a cluster made up of mostly *source-contemp* and *chr-1* copy fragments. Only chronicle text fragments were classified as outliers. For this experiment we use the dataset that contains the chronicle outliers and external data. Then we perform a three way comparison between: the clustering from Chronicle Data Experiment 1, the clustering from Chronicle Data Outlier Experiment, and the result of the qualitative analysis from historians. We aim at determining if the outlier chronicle fragments continue to cluster in an incorrect way when the extra noise from the rest of the chronicle data is removed. Comparing the Chronicle Data Experiment 1 clustering results with the Chronicle Data Outlier Experiment clustering results provides extra evidence to reinforce the labelling of outlier fragments. Additionally comparing the two clustering results with the opinions of the historians allows for a hypothesis generated outside of the computational method to be taken into account. No outlier experiment is conducted for the *Dataset2* results, because the Chronicle Data Experiment 2 clustering did not provide clear enough outliers.

## 5. Results

Several findings can be pointed out from the results of our experiments provided in this section.

Regarding the clustering obtained from the **external data experiments**, we found that the clusters are correclty build, fully separatiing*author* and *source* fragments. This indicates that the methodology is efficient when no texts from the chronicle are used. [5]

**Chronicle Data Experiment 1** resulted in a strong classification for 84.5% of the chronicle fragments, leaving 15.5% as outliers. A chronicle fragment is defined as strongly classified when it is placed into a cluster which contains only one type of external data (author or source). Some clusters which do not contain any external data show patterns of either copy or no-copy fragments, but this is not deemed a strong classification as external data fragments are not present. Figure 1 shows the network visualisation broken down in Table 5. In yellow we show the classes that obtain a higher percentage for each cluster.

| Cluster ID | Cluster Size | copy % | *source-hist %* | no-copy % | *author %* |
|---|---|---|---|---|---|
| **0** | 65 | 90.77 | 0 | 9.23 | 0 |
| **1** | 16 | 12.50 | 0 | 87.50 | 0 |
| **2** | 41 | 31.70 | 0 | 65.85 | 2.44 |
| **3** | 12 | 25 | 0 | 75 | 0 |
| **4** | 78 | 34.61 | 37.17 | 15.38 | 12.82 |
| **5** | 65 | 83.07 | 12.30 | 4.62 | 0 |
| **6** | 17 | 94.11 | 0 | 5.88 | 0 |
| **7** | 20 | 95 | 0 | 5 | 0 |
| **8** | 28 | 89.28 | 0 | 7.14 | 3.57 |
| **9** | 49 | 12.24 | 0 | 87.76 | 0 |
| **10** | 125 | 28 | 0 | 67.20 | 67.2 |
| **11** | 55 | 100 | 0 | 0 | 0 |

Table 5: The distribution of the fragments for Gephi modularity on the network in Chronicle Data Experiment 1.

Figure 1: Results of running Gephi OpenOrd for Chronicle Data Experiment 1. Green = no-copy; pink = copy; orange = *source-hist*; blue = *author*.

The **Chronicle Data Outlier Experiment** resulted in 83.75% of the outlier fragments from Chronicle Data Experiment 1 being strongly classified. The method was able to provide strong classification for 97.5% overall of the chronicle fragments in *chr-1* across the two experiments. Figure 2 shows the network visualisation for the Chronicle Data Outlier Experiment broken down in Table 6.[6]

| Cluster ID | Cluster Size | copy % | *source-hist %* | no-copy % | *author %* |
|---|---|---|---|---|---|
| **0** | 25 | 4 | 92 | 4 | 0 |
| **1** | 4 | 50 | 0 | 50 | 0 |
| **2** | 32 | 75 | 3.13 | 18.75 | 3.13 |
| **3** | 13 | 61.54 | 30.77 | 7.69 | 0 |
| **4** | 17 | 11.76 | 47.06 | 29.41 | 11.76 |
| **5** | 17 | 64.71 | 0 | 35.29 | 0 |
| **6** | 8 | 0 | 0 | 12.5 | 87.5 |
| **7** | 7 | 100 | 0 | 0 | 0 |
| **8** | 11 | 9.09 | 0 | 18.18 | 72.72 |

Table 6: The distribution of the fragments for Gephi modularity on the network for Chronicle Data Outlier Experiment.

Figure 2: Results of running Gephi OpenOrd for the Chronicle Data Outlier Experiment. Green = no-copy; pink = copy; orange = *source-hist*; blue = *author*.

**Chronicle Data Experiment 2** provided a strong classification for 63.8% of the chronicle fragments in *chr-2*, a lower percentage than for the data in Chronicle Data Experiment 1. Figure 3 shows the network visualisation broken down in Table 7. The results of clustering the *Dataset2* fragments made it difficult to find outliers, as fewer clusters showed strong trends. No outlier experiments were carried out for this data set, due to the high proportion of mixed clusters and lack of clear outlying fragments in Chronicle data experiment 2. We will investigate this further in future work.

| Cluster ID | Cluster Size | copy % | *source-contemp %* | no-copy % | *author %* |
|---|---|---|---|---|---|
| **0** | 49 | 85.71 | 8.16 | 4.1 | 2.04 |
| **1** | 28 | 50 | 21.43 | 3.6 | 25 |
| **2** | 28 | 53.57 | 10.71 | 28.57 | 7.14 |
| **3** | 21 | 38.1 | 38.1 | 23.8 | 0 |
| **4** | 9 | 100 | 0 | 0 | 0 |
| **5** | 16 | 56.25 | 0 | 12.5 | 31.25 |
| **6** | 67 | 34.33 | 0 | 61.2 | 4.47 |
| **7** | 31 | 77.42 | 0 | 22.58 | 0 |
| **8** | 47 | 27.75 | 0 | 72.34 | 0 |

Table 7: The distribution of the chronicle data and corresponding external fragments for Gephi modularity on the network of Chronicle Data Experiment 2.

[5] See appendix B for tables and visualisations of external data experiment 1 and external data experiment 2.

[6] See appendix A for full details of the Chronicle Data Outlier Experiment comparative results.

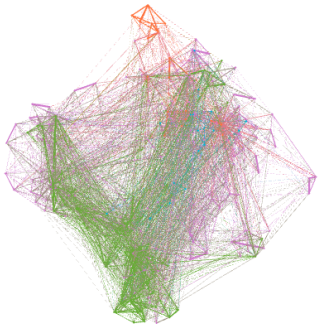

Figure 3: Results of running Gephi OpenOrd on *Dataset2* for Chronicle Data Experiment 2. Green = no-copy; pink = copy; orange = *source-contemp*; blue = *author*.

## 6. Discussion

The results obtained, which are summarized in Table 8 lead to several discussion items that are presented below.

| Experiment | % Strongly classified |
|---|---|
| Chronicle Data Experiment 1 | 84.5% |
| Chronicle Data Outlier Experiment | 83.75% |
| All *chr-1* Experiments (combined) | 97.5% |
| Chronicle Data Experiment 2 | 63.8% |

Table 8: Summary of the clustering experiments results, showing the % of chronicle fragments given a strong classification.

Chronicle Data Experiment 1 shows that the method is able to assign the correct label to 436 of the 516 chronicle fragments in *chr-1*. The results of this experiment rely only on the separation of chronicle fragments into copy and no-copy based on the presence of inverted commas. As such, the presence of inverted commas appears to be a strong indication of authorship in this section of the chronicle.

For the 15.5% of outlying fragments from Chronicle Data Experiment 1, the majority (83.75%) can be assigned to a label with relative certainty by comparing the results of Chronicle Data Outlier Experiment clustering and the opinion of the historians with the original Chronicle Data Experiment 1 clustering result.

The qualitative analysis of the clusters and their outliers performed by the historians resulted in a few observations. First of all, a certain writing style or word use could be a reason for clustering. Other clusters showed that fragments of a certain topic clustered together. Fragments about for example finances (and thus containing a lot of numbers) often ended up in the same cluster, and so did fragments on law, and religion. In the case of copied fragments, the period of the sources from which they were copied could also serve as an important characteristic. Publications from the seventeenth century clearly differ from eighteenth century sources. Finally, the qualitative analysis showed that Louwen made inconsistent use of the inverted commas. Some fragments were marked as copy because there were in inverted commas, but the use of personal pronouns made a strong case for such fragments in fact being the author's original. The opposite could also be the case: some fragments were clearly copied, but no inverted commas were added. Louwen rarely added reference to a source, indicating that copying or paraphrasing was very probable.

For many of the outlier fragments it was found that the clustering in both experiments lined up to give a clear verdict. Where this was not the case, the historians' opinion on the authorship of the text based on qualitative analysis was able to give an explanation for this. This shows that the usefulness of the qualitative analysis should not be understated in the analysis of the quantitative results. The second analysis in Chronicle Data Outlier Experiment and the following three way comparison leaves only 2.5% of *chr-1* with no clear label. These fragments need further analysis.

Chronicle Data Experiment 2 gives less clear results. In this experiment, the method is shown to hold, as defined clusters still appear. But the mixed nature of several of the clusters makes labelling outliers much less clear than in the case of Chronicle Data Experiment 1. Through qualitative analysis we found that cluster 3 (see Table 7) may show a strong limitation of the current method. The outlying fragments in this cluster do not appear to show copied text, but in fact a contemporary author independent style that could be qualified as 'Tale Kanaäns'. This style is impregnated with biblical vocabulary and expressions deriving from the Dutch translation of the Bible authorised by the States General of the Dutch Republic in 1637. The no-copy fragments included in this cluster show Louwen discussing sermons and other religious affairs using this solemn style, but clearly without referring to any sources. These fragments in a biblical style unsurprisingly cluster with bible passages included in *source-contemp*. This, in the context of the current method, flags them as copied text from a biblical source, which is incorrect. We hypothesize that the use of such an author independent style masks the personal style of Louwen and provides a false negative in stylometric methods of authorship verification.

The mixed nature of many clusters of Chronicle Data Experiment 2 makes labelling them much more difficult. No outlier experiments were run on this second data set yet. The single round of experimentation (Chronicle Data Experiment 2) provides explainable clusters for 63.8% of the chronicle fragments in *chr-2*.

The current method is able to make authorship judgements about the content of *chr-1* to a further extent than *chr-2*. Taking the differences in structure and content of the chronicles' volumes into account (as discussed in Section 1), it can be hypothesised that this is partly due to the difference in function of the two sections of the chronicle. In the first section, Louwen provides an overview of the history of the local area. Louwen's knowledge of these events comes from history books,

as he was not alive when they took place. This writing follows a relatively simple format with copied text from sources, some paraphrasing to tell stories, and Louwen's own voice used in his reflections. The second section of the chronicle follows a more complex structure, with Louwen covering contemporary life and political events in the Netherlands, interweaving his own experiences with quotes from news media and publications by political institutions and authorities. In this section the distinction between what Louwen copies from external sources and his own words and experiences becomes more blurred. The fact that Louwen is currently experiencing the events covered in the contemporary sources he uses may also lead him to paraphrase more in *chr-2* than in *chr-1*, inserting his own knowledge, personal experiences and political biases into the information he includes. This hypothesis is supported by the larger portion of mixed clusters for this data set shown in Table 7, although further qualitative analysis is needed to determine if this explains the misclassifications.

Lastly, how can this interdisciplinary research contribute to further historical research on media use in chronicles? The chronicle from Purmerend was selected because of the author's use of inverted commas, which made it possible to split the chronicle into fragments with hypothesised stable authorship. In the complete corpus only few chronicles have this characteristic, which means that it will be necessary to find another method to create fragments. Nonetheless, this particular research that combines historical and computational linguistics and stylometry has proven to be highly informative. Many of the outliers in the stylometric classifications could be explained through qualitative knowledge of the historical context and source material. Analysis of the outliers was very insightful for the historians regarding the writing, quoting behaviour and media use of the chronicler. For future work it is recommended to opt for this interdisciplinary approach.

In relation to the existing body of work outlined in Section 2, these findings agree with the observations of (Kestemont et al., 2013) that collaboration between historians and computational linguists is key to answering questions of textual provenance for historical works. It was also found that as in both Kestemont et al. (2013) and Dalen-Oskam (2014), our methodology was successful in separating texts of known origin before the chronicle fragments were included in the analysis. The Stylo parameter tuning experiments show similar findings to those in (Kestemont et al., 2013), further adding that character level features were beneficial in the context of this study.

## 7. Conclusions

In this paper we presented a computational method for authorship verification using stylometric techniques and applied it to a 18th century Dutch chronicle, the chronicle of Purmerend written in the period 1766-1791 by Albert Pietersz Louwen. Several clustering experiments have been performed with the Stylo tool, combining 5 text types, which include texts from the chronicle, other texts from the author and texts from sources that the author mentions.

Based on the findings outlined in Section 6 we put forward several conclusions. In relation to the usefulness of computational authorship verification techniques to differentiate text originally by the author from text copied from other sources, the results show that stylometric methods can provide an authorship hypothesis for the majority of fragments. As such, the method can be seen as effective in providing evidence of authorship in this case. We established three classes of fragments, own words, direct copies, and paraphrasing. For *chr-1*, all three classes can be identified> Most paraphrase fragments are found through Chronicle Data Outlier Experiment. The fact that paraphrasing in *chr-1* is often marked by contrasting the labels produced by Chronicle Data Experiment 1 and Chronicle Data Outlier Experiment, means that for *chr-2* it is not possible to find paraphrased fragments with certainty, since we did not perform outlier experiments.

As for the question of how helpful the use of inverted commas is to differentiate between copied and non copied text, the results of Chronicle Data Experiment 1 show that 84.5% of the fragments from *chr-1* can be labelled purely based on the presence of inverted commas. For *chr-2* this drops to 63.8%. These results provide evidence that for the majority of cases the inverted commas in this chronicle indicate use of a source, either as copied or paraphrased text.

Finally, in relation to how useful it is to compare fragments by the author with fragments by external sources, we conclude that it is useful in order to test the methods applied, since it was possible to automatically cluster the *author* and *source* texts when no chronicle data was present in the experiment and the results held when texts from the chronicle were added. Additionally the outcome of the Chronicle Data Outlier experiment shows how valuable the collaboration with historians has been within this research. We conclude that, combining the quantitative methods with the results of qualitative analysis is key to understanding and interpreting clustering results of historical texts.

Future work in this project should focus on the in depth analysis of the results of Chronicle Data Experiment 2 in order to test the hypothesis that the multiple mixed clusters found in this experiment show some form of paraphrasing. Further work should also aim to uncover patterns in the inconsistent use of inverted commas to decipher what Louwen signals when using inverted commas outside of copying and paraphrasing. The 2.5% of cases not clearly labelled in *chr-1* provide a starting point for this analysis. Finally, the long term goal is to apply the methodology to the full corpus of chronicles. This requires the development of a sustainable method to generate text fragments so that the fragmentation does not rely on the use of inverted commas by chroniclers.

## Acknowledgements

Research for this paper was funded through the NWO VC project 'Chronicling Novelty. New knowledge in the Netherlands, 1500-1850' and by the Network Institute of the VU Amsterdam, via the project "The Cycle of News in Chronicles from Eighteenth Century Holland: A Stylometric Approach". We are grateful to the three anonymous reviewers who provided useful comments and to our colleagues of the Chronicling Novelty project.

## 8. Bibliographical References

Bastian, M., Heymann, S., and Jacomy, M. ). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 1, pages 361–362.

Besamusca, B., Griffith, G., Meyer, M., and Morcos, H. (2016). Author attributions in medieval text collections: An exploration. 76(1):89–122. Publisher: Brill Rodopi.

Birnbaum, D. J., Bonde, S., and Kestemont, M. (2017). The digital middle ages: An introduction. 92:S1–S38. Publisher: The University of Chicago Press.

Blaak, J. (2009). *Literacy in Everyday Life. Reading and Writing in Early Modern Dutch Diaries*. Egodocuments and History Series; v. 2. Brill, Leiden ; Boston.

Ann Blair, et al., editors. (2021). *Information: A Historical Companion*. Princeton University Press, Princeton.

Blair, A. (2010). *Too Much to Know: Managing Scholarly Information before the Modern Age*. Yale University Press, New Haven London.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct.

Burrows, J. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computingt*, 17(3):267–287.

Daelemans, W. (2012). Tsd conference presentation: Computational stylometry. `https://www.youtube.com/watch?v=xxLH--VAClQ`, September.

Dalen-Oskam, K. H. v. (2014). Epistolary voices: The case of elisabeth wolff and agatha deken. *LLC: the journal of digital scholarship in the humanities*, 29(3):443–451. Publisher: Oxford University Press.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1):107–121.

Jänicke, S. and Joseph Wrisley, D. (2017). Visualizing mouvance: Toward a visual analysis of variant medieval text traditions. 32:106–123.

Kestemont, M., Moens, S., and Deploige, J. (2013). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities*, 30(2):199–224, 10.

Martin, S., Brown, W., Klavans, R., and Boyack, K. (2011). Openord: An open-source toolbox for large graph layout. *Proc SPIE*, 7868:786806, 01.

Mauer, B. (2001). *"Gemain Geschrey" und "teglich Reden": Georg Kölderer - ein Augsburger Chronist des konfessionellen Zeitalters*. Number 29 in Veröffentlichungen der Schwäbischen Forschungsgemeinschaft Reihe 1, Studien zur Geschichte des bayerischen Schwaben. Wißner, Augsburg.

Pettegree, A. (2014). *The Invention of News: How the World Came to Know about Itself*. Yale University Press, New Haven ; London, England.

Pollmann, J. (2016). Archiving the Present and Chronicling for the Future in Early Modern Europe. *Past & Present*, 230(suppl 11):231–252.

Rau, S. (2002). *Geschichte Und Konfession: Städtische Geschichtsschreibung Und Erinnerungskultur Im Zeitalter von Reformation Und Konfessionalisierung in Bremen, Breslau, Hamburg Und Köln*. Number Bd. 9 in Hamburger Veröffentlichungen Zur Geschichte Mittel- Und Osteuropas. Dölling und Galitz, Hamburg.

Schmidt, H. (1958). *Die deutschen städtechroniken als spiegel des bürgerlichen selbstverständnisses im spätmittelalter.* Vandenhoeck & Ruprecht.

Waddell, B. (2018). Writing history from below: Chronicling and record-keeping in early modern england. *History Workshop Journal*, 85:239–264.

Waddell, B. (2020). "verses of my owne making": Literacy, work, and social identity in early modern england. *Journal of Social History*, 54(1):161–184.

## A. *Dataset1* Outlier Comparison Table

| Fragment ID | *Dataset1* Cluster ID (proposed cluster type) | Outlier Only Cluster ID (proposed cluster type) | Historian Opinion | Verdict |
|---|---|---|---|---|
| no_copy_A13997_14029 | 0 (NE, copied text) | 5 (NE, 65%K) | copy | probably copied |
| no_copy_A13687_13717 | 0 (NE, copied text) | 5 (NE, 65%K) | copy | probably copied |
| no_copy_A13391_13423 | 0 (NE, copied text) | 5 (NE, 65%K) | copy | probably copied |
| no_copy_A15539_15584 | 0 (NE, copied text) | 5 (NE, 65%K) | copy | probably copied |
| no_copy_A14525_14593 | 0 (NE, copied text) | 5 (NE, 65%K) | copy | probably copied |
| copy_238_330 | 1 (NE, author) | 4 (weak copied text) | weak author | probably paraphrase |
| copy_A_5263_5278 | 1 (NE, author) | 2 (unclear) | weak author | probably paraphrase |
| copy_979_999 | 2 (author) | 3 (copied text) | author | Needs looking at (poss para) |
| copy_A_14429_14447 | 2 (author) | 7 (NE, 100%K) | unclear | author |
| copy_A_2372_2386 | 2 (author) | 1 (NE, 50/50 K/NK) | unclear | author |
| copy_A_2389_2416 | 2 (author) | 1 (NE, 50/50 K/NK) | unclear | author |
| copy_A_2637_2658 | 2 (author) | 2 (unclear) | unclear | probably paraphrase |
| copy_A_17260_17291 | 2 (author) | 7 (NE, 100%K) | unclear | author |
| copy_A_17225_17257 | 2 (author) | 7 (NE, 100%K) | unclear | author |
| copy_A_17192_17222 | 2 (author) | 7 (NE, 100%K) | unclear | author |
| copy_A_12964_12994 | 2 (author) | 0 (copied text) | unclear | Needs looking at (poss para) |
| copy_A_17537_17571 | 2 (author) | 7 (NE, 100%K) | unclear | author |
| copy_A_14302_14340 | 2 (author) | 5 (NE, 65%K) | unclear | Possibly author (or para) |
| copy_9269_9295 | 3 (NE, author) | 5 (NE, 65%K) | author | Probably author |
| copy_9167_9193 | 3 (NE, author) | 3 (copied text) | author (recounting a service) | Paraphrase |
| no_copy_A5552_5572 | 4 (weak copied text) | 2 (unclear) | copy/paraphrase | Paraphrase |
| no_copy_90_237 | 4 (weak copied text) | 4 (weak copied text) | copy/paraphrase | Paraphrase |
| no_copy_700_725 | 4 (weak copied text) | 4 (weak copied text) | copy/paraphrase | Paraphrase |
| no_copy_559_594 | 4 (weak copied text) | 4 (weak copied text) | copy/paraphrase | Paraphrase |
| no_copy_357_428 | 4 (weak copied text) | 4 (weak copied text) | copy/paraphrase | Paraphrase |
| no_copy_330_351 | 4 (weak copied text) | 4 (weak copied text) | copy/paraphrase | Paraphrase |
| no_copy_9874_9905 | 4 (weak copied text) | 8 (author) | copy/paraphrase | paraphrase |
| no_copy_A611_632 | 4 (weak copied text) | 2 (unclear) | copy/paraphrase | paraphrase |
| no_copy_474_493 | 4 (weak copied text) | 8 (author) | copy/paraphrase | Possible author (or para) |
| no_copy_9772_9794 | 4 (weak copied text) | 2 (unclear) | copy/paraphrase | paraphrase |
| no_copy_9528_9544 | 4 (weak copied text) | 3 (copied text) | copy/paraphrase | Possible copy (or para) |

| Fragment ID | *Dataset1* Cluster ID (proposed cluster type) | Outlier Only Cluster ID (proposed cluster type) | Historian Opinion | Verdict |
|---|---|---|---|---|
| no_copy_A4784_4806 | 4 (weak copied text) | 6 (author) | copy/paraphrase | Possible author (or para) |
| no_copy_A8770_8804 | 5 (copied text) | 0 (copied text) | copy | copy |
| no_copy_A1182_1209 | 5 (copied text) | 1 (NE, 50/50 K/NK) | copy | Probably copy |
| no_copy_A1121_1170 | 5 (copied text) | 1 (NE, 50/50 K/NK) | copy | Probably copy |
| no_copy_A10953_10977 | 6 (NE, copied text) | 2 (unclear) | copy | Probably paraphrase |
| no_copy_A16883_16917 | 7 (NE, copied text) | 5 (NE, 65%K) | copy | Probably copy |
| no_copy_A15891_15995 | 8 (copied text) | 2 (unclear) | unclear | Probably paraphrase |
| no_copy_A20899_20916 | 8 (copied text) | 2 (unclear) | unclear | Probably paraphrase |
| copy_A_13718_13737 | 9 (NE, author) | 5 (NE, 65%K) | unclear | Probably author |
| copy_A_776_795 | 9 (NE, author) | 2 (unclear) | unclear | paraphrase |
| copy_A_15222_15237 | 9 (NE, author) | 2 (unclear) | unclear | paraphrase |
| copy_A_15041_15057 | 9 (NE, author) | 2 (unclear) | unclear | paraphrase |
| copy_A_10879_10893 | 9 (NE, author) | 2 (unclear) | unclear | paraphrase |
| copy_A_11032_11046 | 9 (NE, author) | 5 (NE, 65%K) | unclear | Probably author |
| copy_A_14393_14414 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_13424_13453 | 10 (author) | 5 (NE, 65%K) | Weak author | Probably author |
| copy_A_17992_18020 | 10 (author) | 7 (NE, 100%K) | Weak author | author |
| copy_A_19264_19298 | 10 (author) | 5 (NE, 65% K) | Weak author | Probably author |
| copy_3767_3815 | 10 (author) | 4 (weak copied text) | Weak author | paraphrase |
| copy_7119_7146 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_2962_2992 | 10 (author) | 5 (NE, 65%K) | Weak author | Probably author |
| copy_A_12775_12797 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_10156_10185 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_11055_11074 | 10 (author) | 5 (NE, 65%K) | Weak author | Probably author |
| copy_A_15447_15472 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_4832_4868 | 10 (author) | 3 (copied text) | Weak author | Needs looking into (poss para) |
| copy_A_12556_12577 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_18516_18543 | 10 (author) | 7 (NE, 100%K) | Weak author | author |
| copy_3276_3298 | 10 (author) | 3 (copied text) | Weak author | Needs looking into (poss para) |
| copy_6517_6545 | 10 (author) | 3 (copied text) | Weak author | Needs looking into (poss para) |
| copy_2336_2366 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_8171_8198 | 10 (author) | 3 (copied text) | Weak author | Needs looking into (poss para) |
| copy_2304_2331 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_5298_5321 | 10 (author) | 2 (unclear) | Weak author | paraphrase |

| **Fragment ID** | ***Dataset1* Cluster ID (proposed cluster type)** | **Outlier Only Cluster ID (proposed cluster type)** | **Historian Opinion** | **Verdict** |
|---|---|---|---|---|
| copy_917_944 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_10903_10922 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_836_860 | 10 (author) | 3 (copied text) | Weak author | Needs looking into (poss para) |
| copy_864_885 | 10 (author) | 5 (NE, 65%K) | Weak author | Probably author |
| copy_A_18023_18051 | 10 (author) | 5 (NE, 65%K) | Weak author | Probably author |
| copy_8662_8677 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_6644_6663 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_3505_3540 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_808_827 | 10 (author) | 3 (copied text) | Weak author | Needs looking into (poss para) |
| copy_A_2254_2285 | 10 (author) | 5 (NE, 65%K) | Weak author | Probably author |
| copy_A_15422_15441 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_3562_3605 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_15482_15506 | 10 (author) | 2 (unclear) | Weak author | paraphrase |
| copy_A_15158_15174 | 10 (author) | 8 (author) | Weak author | author |
| copy_A_15382_15403 | 10 (author) | 2 (unclear) | Weak author | Possible paraphrase |

Table 9: This table shows the three way comparison between the clustering of outlier fragments in **chronicle data experiment 1**, the clustering of those same fragments in **chronicle data outlier experiment** and the opinion of the historians. 'NE' signifies no external data in the cluster.

## B. Results of the External Data Experiments

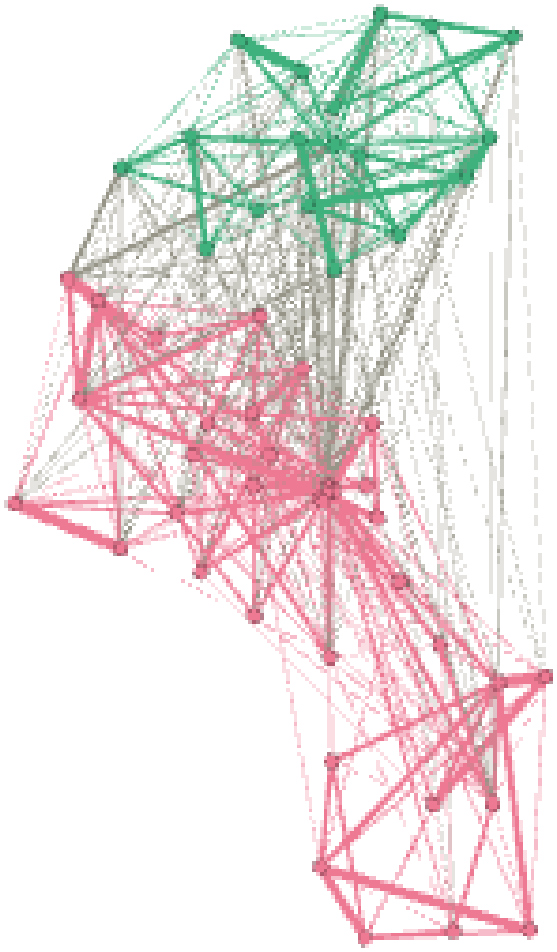

Figure 4: Results of running Gephi OpenOrd for **external data experiment 1**. Red = *source-hist*; green = *author*

| **Cluster ID** | **Cluster Size** | ***source-hist %*** | ***author %*** |
|---|---|---|---|
| **0** | 8 | 100 | 0 |
| **1** | 14 | 100 | 0 |
| **2** | 10 | 0 | 100 |
| **3** | 11 | 100 | 0 |
| **4** | 4 | 100 | 0 |
| **5** | 8 | 0 | 100 |

Table 10: The distribution of fragments for Gephi modularity for **external data experiment 1**.

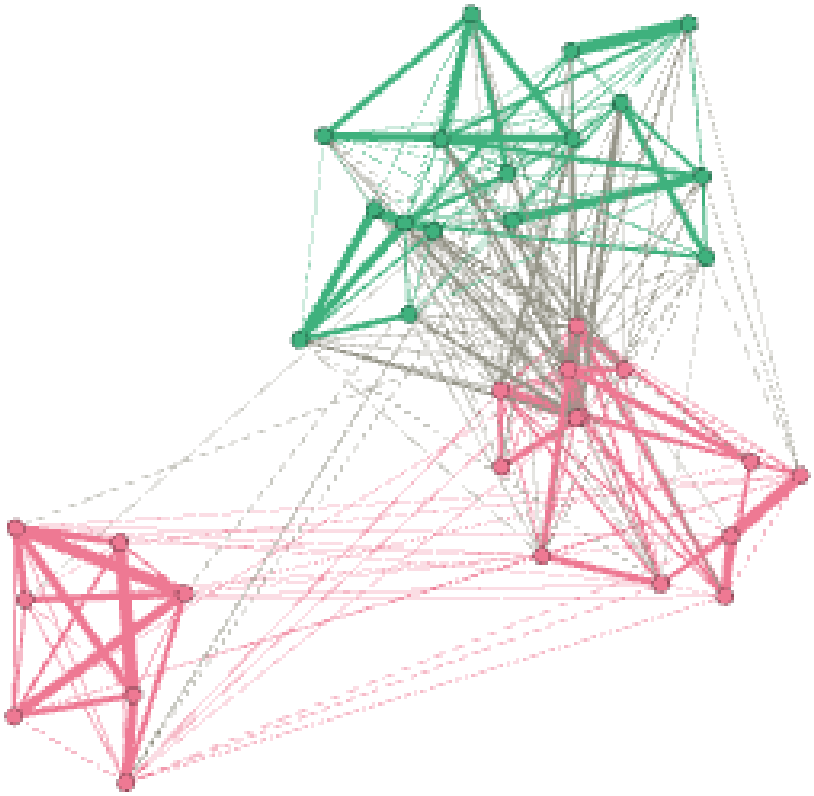

Figure 5: Results of running Gephi OpenOrd for **external data experiment 2**. Red = *source-contemp*; Green = *author*

| **Cluster ID** | **Cluster Size** | ***source-contemp %*** | ***author %*** |
|---|---|---|---|
| **0** | 4 | 100 | 0 |
| **1** | 10 | 100 | 0 |
| **2** | 7 | 100 | 0 |
| **3** | 2 | 0 | 100 |
| **4** | 7 | 0 | 100 |
| **5** | 9 | 0 | 100 |

Table 11: The distribution of fragments for Gephi modularity on the network for **external data experiment 2**.