Exploring the Potential of Bootstrap Consensus Networks for Large-scale Authorship Attribution in Luxdorph's Freedom of the Press Writings

Florian Meier, Birger Larsen, and Frederik Stjernfelt

Science, Policy and Information Studies
Department of Communication and Psychology
Aalborg University, Copenhagen, Denmark
{fmeier,birger,stjern}@hum.aau.dk

Abstract. Authorship attribution (AA) is concerned with the task of finding out about the true authorship of a disputed text based on a set of documents of known authorship. In this paper, we investigate the potential of Bootstrap Consensus Networks (BCN) – a novel approach to generate visualizations in stylometry by mapping similarities of authorial style between texts into the form of a network – for large-scale authorship attribution tasks. We apply this method to the freedom of the press writings (*Trykkefrihedsskrifter*), a corpus of pamphlets published and collected in Denmark at the end of the 18th century. By conducting multiple experiments, we find that the size of the constructed networks depends heavily on the type of variables and distance measures used. Furthermore, we find that, although a small set of unknown authorship problems can be solved, in general, the precision of the BCN method is too low to apply it in a large-scale AA scenario.

Keywords: Authorship Attribution, Stylometry, Bootstrap Consensus Network (BCN).

1 Introduction

In this work, we investigate the feasibility of using Bootstrap Consensus Networks (BCN) for solving large-scale authorship attribution (AA) problems in the so-called *Trykkefrihedsskrifter*, the danish freedom of the press writings.

AA, a sub-field of quantitative text analysis and stylometry, is concerned with the task of finding out about the true authorship of a disputed text based on a set of documents of known authorship. A fundamental assumption of AA is that "individuals have idiosyncratic and largely unconscious habits of language use, leading to stylistic similarities between texts written by the same person" [9, p.ii4]. Quantitative techniques for finding stylistic similarities have been used in literary studies since the 1960s, with Mosteller and Wallace being among the first ones trying to solve the problem of unknown authorship in the Federalist Papers [23]. Over the years, literary AA has seen many approaches to capture stylistic similarities by e.g. looking at the relative frequencies of function words, parts of speech, the degrees of vocabulary richness, or syntactic complexity, and use them in a classification task or to determine the distance between documents in a clustering scenario or vector-space. [4,16,12].

The BCN method used in this work was originally introduced as a visualization technique that tries to overcome the reliability issues of visual methods used in stylometry, especially dendrograms, that also allows for getting a distant reading perspective [22] on a text corpus by visualizing textual relations in a network [6]. However, due to the idea of linking the nearest stylistic neighbour to every text in the bootstrap network generation process, the method seems like a natural fit for AA and finding the true authors of pamphlets.

The data-set we are working with is a collection of pamphlets published in Denmark during the press freedom period between 1770-1773. The rise of the freedom of the press was based on the abolition of censorship on September 14, 1770, announced by JF Struensee, a German physician, who rose to power as de facto regent of the kingdom of Denmark and Norway, and led to an unprecedented publication of ideas in the form of pamphlets discussing topics like the church, economy, societal conditions or patriotism [15]. This period forever changed the conditions for opinion formation and exchange of views in Denmark [15]. Thus, the collection is of high value at present times as questions about what the freedom of the press ideas meant in contemporary and later times are important to be elucidated at a time when we are also trying to understand the premises for how both fact and fake news are spread and exchanged in today's public. However, in order to perform an exhaustive analysis of all pamphlets determining their true authorship is necessary. Unfortunately, about 50% of the collection is of unknown authorship. However, this makes the collection a reasonable use case for our experiments.

This work aims not at solving all problems of unknown authorship in the freedom of the press writings. However, it can be considered to be an introduction into the BCN method and a feasibility study on the applicability of BCNs for AA tasks. The contributions of our paper can be summarized as follows:

- We present general lessons learned from applying the BCN method to a corpus of danish pamphlets published in the 18th century.
- We give insight into the use of the R package *Stylo* and which parameter the function call takes. This is difficult to interpret from the reference manual.
- We investigate what effect the choice of features and distance measures has on BCN characteristics e.g. in terms of the number of edges/links produced.
- We conduct experiments to find the BCN with the highest attribution quality i.e. precision and number of possible attributions by proposing a precision measure that is based on the type of relations (edges) created in the network.

2 Related Work: Authorship Attribution

AA is based on the extraction of authorial information from texts, to determine if, given a set of candidate authors and a sample of texts written by these authors, a text of unknown or disputed authorship can be attributed to one of these authors or not [26].

AA problems are widespread among many disciplines reaching from digital text forensics which is concerned with e.g finding out about the originator of a blackmail message as evidence in court of law [12] over applications in information retrieval where indexing authorial style can help with retrieving answers to writing-style related

queries [26] to applications in the digital humanities where AA techniques are not only used to solve questions of true authorship [16] but also to obtain a broader picture of relations between different literary pieces from a distant-reading perspective [22].

Problems within the field of AA can be distinguished into problems of (1) closed-set attribution (2) open-set attribution and (3) authorship verification [25]. Koppel and Scheer refer to the latter as the fundamental question of authorship analysis: given two documents are they written by the same author [20]? Authorship verification is the essential problem as all authorship attribution tasks can be transferred to separate authorship verification problems [20]. Methods to solve the problem of AA can be classified in various ways [20,25,11]. Koppel and Winter divide automated AA techniques into two main types: (1) machine-learning methods, in which known writings of the candidate authors are used as training documents to construct a classifier and (2) similarity-based methods, which use clustering approaches or some form of dimensionality reduction (PCA) of feature vectors. Potha and Stamatatos distinguish methods that follow (1) the instance-based paradigm, which looks at each document separately and (2) the profile-based paradigm in which multiple documents of one author get combined to an author-profile [25]. Furthermore, they distinguish (1) intrinsic and (2) extrinsic verification models. The former only uses documents from the corpus, in the latter scenario additional documents from external sources get added to create a multi-class classification problem [25].

There are multiple survey papers that summarize machine-learning methods for AA [17,19,29]. However, recently many new and advanced methods, especially in the area of digital forensics, have been proposed [11]. These approaches range from applying Latent Dirichlet Allocation (LDA) [13] over deep learning architectures [3] to the use of compression models [10]. In a recent study Halavni, Winter and Graner compare 12 existing AV approaches, concluding that although some of them reach good performance in AV tasks, they are not reliable enough for real forensic cases [11]. To the best of our knowledge, none of these methods have ever been applied in a literary context, which leaves a potential for future work.

In the digital humanities the use of similarity-based methods – i.e. feature vectors (token, n-gram, tf-idf, delta) in combination with some kind of distance or similarity measure – are widely spread [4,8,18,27]. One reason for their popularity is the fact that those methods are not black-box approaches but their results are traceable and often *speak for themselves* through meaningful visualizations [8]. However, these methods also have drawbacks, which will be outlined in the next chapter.

3 Method: Boostrap Consensus Network

The BCN method is a similarity-based method, following the instance-based paradigm applying an intrinsic verification model. Its development is a reaction to overcome the reliability issues inherent in some AA techniques popular among literary scholars, especially the visualization of clustering results in form of dendrograms [6]. As the final shape of a dendrogram depends on many factors like (1) the distance measure, (2) the linkage algorithm and (3) the number (most frequent words, MFW) and type of vari-

ables (i.e. words, n-grams) a change in one of the three factors, can result in a change of clusters and thus differing visualizations.

BCNs solve two problems of visualization in stylometry. First, the problem of unstable results is by-passed by using consensus techniques. The idea is to combine numerous dendrograms into a single consensus plot [6]. Second, when a corpus reaches a certain size it becomes impossible to visualize it in form of a dendrogram as those get cluttered and illegible [6]. A BCN avoids this problem by mapping textual similarities into the form of a network. Texts are considered as nodes, their relation i.e. authorial or topical similarity represented by the edge that links those nodes (Figure 3 shows an example of a BCN network).

Two algorithms are involved in the creation of a BCN:

- *Linkage computation*: The first algorithm performs the distance computation between each pair of texts in the corpus-based on a distance measure and the type of variables and furthermore ranks the text samples from most to least similar. A link between one text and its closet neighbour gets created, however `X` more runner-ups can be considered. This way researchers can not only extract the authorial signal by looking at the edges with the highest weight but by looking at runner-ups also let "hidden layers of subtle inter-textual correlations" [6, p.58] emerge.
- *Consensus creation*: The second algorithm adds up all connections produced in multiple runs and captures the average behaviour of a corpus for all given frequency strata (i.e. different values of most frequent works e.g. 100, 200 ... 1000). The resulting consensus network is characterised by the robust pattern that emerged across a set of all generated snapshots.

BCNs can be produced in an undirected or directed manner. In case a directed version is created, the weights of mutual relations are simply summed up. In a directed network, connections are kept as independent edges with weight values of their own. Moreover, every single node has at least `1+X` outgoing links whereby `X` represents the number of runner-ups. In our experiments, we only consider the nearest neighbours. The idea is that nearest neighbours only extracts the strongest pattern – in our case the authorial style – filtering out weaker textual similarities and reducing the number of false positives in an authorship identification task.

Although BCNs are a fairly new approach to visualization in stylometry, some research projects have already applied the method in their analysis [28,24,1,7]. The closest to our work is a study by Al-Yahya, that investigates the effect of different distance measures on clustering and BCN visualizations by using Arabic texts in a genre detection task [1]. She concludes that BCN-based genre clustering aligns well with human-defined genres although the clustering result only reaches an accuracy of 63% [1].

In our experiments we use the R implementation of the BCN method, which is part of `Stylo`, an R package that provides a number of functions, supplemented by a GUI, to perform various analyses in the field of computational stylistics e.g. authorship attribution [8]. Section 5.4 presents details on how to set the function parameters.

4 Data-set: The Danish Freedom of Press Writings

The data-set used in the experiments is a digitized and machine-readable version of Bolle Willum Luxdorph's collection of Freedom of the Press Writings also referred to as the 'Trykkefrihedsskrifter' [14,15]. Digitization and Optical Character Recognition (OCR) were performed by the Royal Library of Denmark as part of the Carlsberg funded project *Trykkefriheden og en ny offentlighedsʼ tilblivelse*.

In 1770, Johan Friedrich Struensee, a German physician, took advantage of King Christian VII inability to reign due to mental health issues and rose to power as de-facto regent of the kingdom of Denmark-Norway. One of Struensee's first acts was the first official declaration of the freedom of the press in 1770, which was "followed by a flood of periodicals and newspapers" [15, p.146]. The period between September 1770 and October 1773 – when the control of the press was reasserted by the court – was characterized by an explosion of printed debate in the form of pamphlets, that Bolle Willum Luxdorph collected [21]. In total, Luxdorph's collection contains around 900 pamphlets[1] bound in 45 volumes, which tend to be discussed as two series of 20 and 25 volumes, that cover the period before and after 17 January 1772, which marks the date of Struensee's fall by force [21]. While the topics discussed in series one span from the order of state and church, economic policy, the development of Danish and Norwegian national identity over the role of women in eighteenth-century politics to the moral danger of the lottery, the second series mainly contains songs and poems about Struensee's evil acts [21]. The Luxdorph collection is extremely valuable as it is a major source of research for how the premises for knowledge dissemination changed and public opinion formed. However, gaining knowledge about "when and why printers hid their authors' names" [21, p.71] and finding out about how multifaceted authorship among the pamphlets is is still a desideratum and needs further detailed attention [21].

4.1 Challenges for Authorship Attribution

Although the pamphlets were written and published within a short time-span, the corpus possesses multiple challenges that make identifying pure authorial signal difficult and solving AA tasks particularly hard. In the following, we list a number of challenges that lead to the exclusion of several books for the present study.

- The corpus does not only contain texts in Danish but also other languages like French, German and English. To limit the influence of language differences these books needed to be filtered out.
- The corpus contains publications of different forms e.g. prose, poems in verses, dramas, plays, simple lists of commodities or handwritten notes. To avoid a strong influence of different literary forms or genres, we filtered out books that did not represent pure prosaic text.
- The OCR result is pretty poor. Although we did not filter out books with many OCR errors, the poor quality of OCR needs to be considered when choosing features for AA.

[1] Depending on which bibliography one looks at the number of volumes, books and single pamphlets can vary slightly. Horstbøll explains the differences in more detail in [15].

In total, we excluded 129 of 854 books which resulted in 725 books to be used in the experiments. The books vary strongly in length with the shortest book containing only 117 tokens. The longest pamphlet, however, is 700 times longer and contains 81416 tokens. On average pamphlets are 4706 tokens long (median=2879). The strong difference in length is another point that makes AA challenging, as certain distance measures – presented in section 5.3 – are sensitive towards extreme word frequencies.

4.2 Authorship in the Freedom of the Press Writings

Henrik Horstbøll published a detailed bibliography of the freedom of press writings in which he performs an initial attribution of authors to pamphlets [14]. This bibliography is the gold standard for our AA investigations. However, one has to acknowledge, that these attributions have been created manually through inspection of pamphlet title pages and secondary material, which means that they can be erroneous and might affect our ground truth data.

In total, Horstbøll identifies 166 different authors. The number of authors shows a heavy long-tailed distribution with 116 authors being attributed to only a single book each. On average an author is attributed to 2.22 books (median=1). The author with the most books in the corpus is Martin Brun, who authored 55 books, followed by J.C. Bie (16) and J.L. Bynch (16). However, among the 725 books about half are of unknown authorship (360;49.66%). This underlines the large-scale dimension and challenge of AA in this collection.

5 Experiments & Experimental Setup

In our experiments, we combine different textual features, variations of punctuation removal and distance measures to build different variations of BCNs.

5.1 Data-set Pre-processing

Due to the large number of OCR errors some pre-processing steps were undertaken. In detail we performed the following normalization and cleaning steps:

- We resolved multiple following white-spaces to a single space.
- We resolved multiple following punctuation marks of the same type to a single character of that type.
- We performed lower-casing on all tokens.

After the cleaning process, we started the feature engineering phase.

5.2 Feature Engineering

Individual writing style, and thus the power to distinguish between different authors, is efficiently and effectively well approximated by textual features [12]. In a survey of different authorship attribution methods, Stamatatos distinguishes between character, lexical, syntactic, semantic and application-specific feature types [29].

Halvani, Winter and Pflug suggest an even finer-grained distinction at the character level by introducing nine feature categories among others the in authorship analysis studies widely used and popular category character n-grams [12].

In our experiment we consider: character n-grams (character shingles) and lexical/-token[2] n-grams. In a bag-of-words context token n-grams are a very standard way of representing text. We intentionally choose character n-grams as a feature category due to the reasonable success their applications have in other AA studies [29,5]. Moreover, character n-grams are impressively robust when one deals with a noisy corpus e.g. one with a high number of misspelt characters, or bad OCR [5,30]. As mentioned earlier, poor OCR is a problem that we are also facing.

For character shingles, we consider a length from one to five characters (1-grams to 5-grams). For token n-grams, we considered values between 1-grams (i.e. single tokens) to 3-grams. Besides varying numbers of `N`, we also varied the experimental condition for removing/not removing punctuation. Table 1 shows that when using character shingles we experimented with the option of not removing (`FALSE`) or removing (`TRUE`) punctuation marks.

5.3 Distance Measures

To investigate the differences in frequency patterns among the individual books we need statistical procedures i.e. distance measures. To get an idea of how and whether the used distance measure has an influence on how BCNs get built, we varied the condition between the classic Euclidean distance [8] and Burrow's delta [4,2,9].

Euclidean distance The Euclidean distance is the most basic, straight-line distance measure of points in a possibly multidimensional vector-space and can also be interpreted as a baseline measure. Eder, Rybicki and Kestmont advise against using Euclidean distance in stylometric analysis if one is working with raw word frequencies [8]. In our case, however, we use normalized word frequencies relative to text length. The Euclidean distance between two documents A and B can be calculated in the following way:

$$\delta_{(AB)} = \sqrt{\sum_{i=1}^{n} \left| (A_i)^2 - (B_i)^2 \right|}$$

Where `n`=*the number of most frequent words* and A_i, B_i=*the frequency of a given feature i in text A and B respectively* [8].

Delta distance In the seminal paper introducing the Delta distance, Burrows defines the measure as "the mean of the absolute difference between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in the target text" [4, p.271]. The Delta distance uses z-transformed word frequencies,

[2] In this work a token is defined as a string separated by whitespace at the beginning and at the end of that string.

which means that over the whole corpus the mean for each word is 0 and has a standard deviation of 1 [2]. This results in less influence of top-scoring words, nevertheless, the measure is dependent on the number of texts analyzed and on a balance between theses texts [8]. Following Argamon, the delta distance between two documents `A` and `B` can be calculated as[2]:

$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - B_i}{\sigma_i} \right|$$

The notation is the same as for the Euclidean distance in formula 5.3 with the additional σ_i=*the standard deviation of a given feature* [2].

5.4 The `stylo.network()`-function

To generate a BCN from word frequencies of a corpus a user can call the `stylo.network()`-function from the Stylo package [8]. When calling the function we made the following changes to its parameters:

- We set `linked.neighbors=1` to extract the strongest pattern i.e. the authorial signal thus only links/edges between the closest neighbours get considered and weaker textual similarities get filtered out [6].
- We set `network.type='undirected'` to generate an undirected network of pamphlets. A link/edge between two books(nodes) can be interpreted as the two books being stylistically related and from the same author.
- The parameters `frequencies` and `distance.measure` refer to the feature category and the distance measure used and were set to one of the options discussed above.

The two parameters `mfw.min` and `mfw.max`, which determine the minimum and maximum number of frequent tokens/n-grams to start and end the bootstrap procedure with, were left at their default values of `mfw.min=100` and `mfw.max=1000` respectively. The following is an example of a function call:

```
stylo.network(gui=FALSE,
    frequencies=character_shingles2FALSE,
        linked.neighbors=1,
            distance.measure='delta',
                network.type='undirected')
```

5.5 Evaluation Measure

In total, we compare 26 runs resulting in 26 BCNs, which are listed in Table 1. To be able to judge the quality of a BCN for AA we propose the following classification of edges i.e. relations between two pamphlets. The arrows can be interpreted as edges in an undirected network.

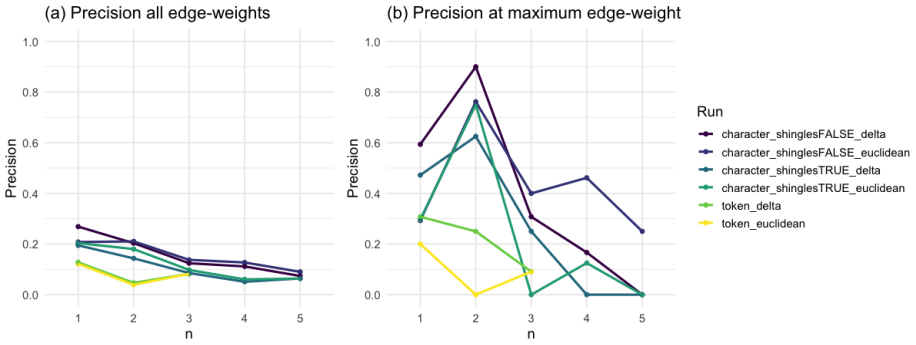

Fig. 1. Overview of the precision values for all feature combinations at different levels of `N` considering (a) all edges and only (b) edges with maximum edge-weight.

- **False attribution,** $X \longleftrightarrow Y$**:** A link between two books written by different authors X and Y
- **Correct attribution,** $X \longleftrightarrow X$**:** A link between two books written by the same author X
- **Possible attribution,** $X \longleftrightarrow ?$**:** A link between a book of which author X is known and a book of unknown authorship (?).
- **Both unknown,** $? \longleftrightarrow ?$**:** A link between two books both of unknown authorship.

To measure the success of a BCN for AA we calculate its *Precision* in the following way:

$$Precision = \frac{\#correct\ attributions}{\#correct\ attributions + \#false\ attributions} \tag{1}$$

The possibility to measure the quality of a BCN is an additional advantage over clustering-based methods, where one would need to estimate an additional threshold value for deciding whether the distance or similarity between two texts hints towards matching authorship or not.

6 Results & Findings

The aim of the experiments is to learn, which combination of input-features used in building a BCN yields to the highest precision. We present the results and findings of our experiments in a four-stage process. First, we describe overall statistics of the built BCNs and how different experiment parameters influence factors like the number established edges. Second, we look at the overall precision that can be achieved if all edge weight groups (i.e. the strength of links between nodes) are considered. Third, we look at the maximum precision that can be reached if we only consider links between books with a maximum edge-weight value. Finally, we show that there is a certain trade-off between the precision that can be achieved and the number of possible attributions.

Table 1 gives an overview of all runs and shows the overall statistics for the different attribution scenarios considering all weight values. The column `Total Links`

represents the total number of edges that get created during the BCN process. The table shows that with a rising number of N more links between books get created. In general, character shingles tend to produce a higher number of links compared to tokens. The maximum number of links gets created at run ID=13 (*character-shingles-5-FALSE-euclidean*) which results in a network with 3793 edges. One can see that the number of links created is mostly dependent on the textual feature. The distance measure seems not to cause any strong variations.

By looking at the precision values in Table 1, which are also visualised in Figure 1(a), we observe that when considering all weight groups the number of correctly linked books and thus the overall precision is extremely poor The best precision lies at P=0.27 and is reached with run ID=17 (*character-shingles-5-FALSE-euclidean*; followed by run ID=4 and run ID=6, which both lie at P=0.21 and use a combination of character shingles and euclidean distance. Intuitively it is much harder to observe correct attributions with a rising number of links. Figure 1(a) shows a clear trend that a rising number of N results in a decreasing number of correctly linked nodes, which supports the intuition that a high number of established links leads to a lower precision.

When only considering edges with maximum weight a different picture emerges. Figure 1(b) visualises all runs considering only the links with maximum edge-weight. It can be observed that in this scenario the precision rises significantly. Figure 1(b) shows that at run ID=19 (*character-shingles-2-FALSE-delta*) a correct attribution rate of P=0.90 is possible. In that context it seems that having punctuation characters not removed contributes a lot to precision as run ID=20 is only able to achieve a precision of P=0.63.

At least two other approaches (run ID=6 and run ID=7) are able to reach a correct attribution ratio of around P$\sim$0.75 and thus link 3 out of 4 nodes correctly. One can

Table 1. Summary of all runs showing the number of different relations created in each network and the achieved precision both in an all edges context and when only considering edges with a maximum edge-weight.

Run ID	Feature	N	Punctuation Removed	Measure	Both Unknown	False Attribution	Correct Attribution	Possible Attribution	Nodes	Total Links	Precision All edge-weights	Precision at Max edge-weight
1	token	1	-	euclidean	265	267	37	489	725	1058	0.12	0.20
2	token	2	-	euclidean	799	491	20	1288	725	2598	0.04	0.00
3	token	3	-	euclidean	318	67	6	492	725	883	0.08	0.10
4	character shingles	1	FALSE	euclidean	169	160	42	251	725	622	0.21	0.29
5	character shingles	1	TRUE	euclidean	159	164	42	258	725	623	0.20	0.30
6	character shingles	2	FALSE	euclidean	308	330	88	489	725	1215	0.21	0.76
7	character shingles	2	TRUE	euclidean	314	378	83	549	725	1324	0.18	0.75
8	character shingles	3	FALSE	euclidean	535	608	97	963	725	2203	0.14	0.40
9	character shingles	3	TRUE	euclidean	764	942	102	1482	725	3290	0.10	0.00
10	character shingles	4	FALSE	euclidean	652	721	105	1269	725	2747	0.13	0.46
11	character shingles	4	TRUE	euclidean	675	1026	66	1593	725	3360	0.06	0.13
12	character shingles	5	FALSE	euclidean	939	1029	102	1723	725	3793	0.09	0.25
13	character shingles	5	TRUE	euclidean	569	499	34	1039	725	2141	0.06	0.00
14	token	1	-	delta	260	259	38	493	725	1050	0.13	0.31
15	token	2	-	delta	783	471	23	1278	725	2555	0.05	0.25
16	token	3	-	delta	318	67	6	492	725	883	0.08	0.10
17	character shingles	1	FALSE	delta	151	166	61	276	725	654	0.27	0.59
18	character shingles	1	TRUE	delta	146	186	45	273	725	650	0.19	0.47
19	character_shingles	2	FALSE	delta	337	415	106	594	725	1452	0.20	0.90
20	character shingles	2	TRUE	delta	313	454	76	643	725	1486	0.14	0.63
21	character shingles	3	FALSE	delta	555	734	104	1206	725	2599	0.12	0.31
22	character shingles	3	TRUE	delta	764	963	90	1491	725	3308	0.09	0.25
23	character shingles	4	FALSE	delta	695	836	105	1393	725	3029	0.11	0.17
24	character shingles	4	TRUE	delta	631	1025	55	1560	725	3271	0.05	0.00
25	character shingles	5	FALSE	delta	886	940	75	1654	725	3555	0.07	0.00
26	character shingles	5	TRUE	delta	576	497	34	1024	725	2131	0.06	0.00

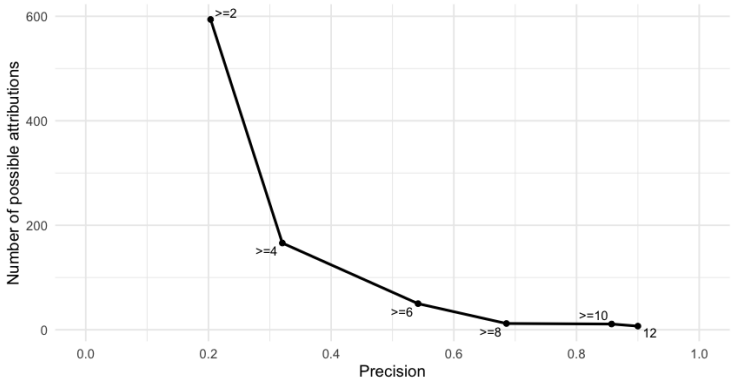

Fig. 2. Relationship between the number of possible attributions and precision for different edge-weight values.

also observe that at a value of `N=3` precision drops heavily to values below `P=0.50` for all feature combinations. The rise in precision can be explained by the low number of links that get considered. At run `ID=19` only 19 edges have a maximum edge-weight, whereby 7 possible attributions exist. Considering the large number of books with unknown authorship a higher value of possible attributions is necessary for the BCNs approach to be applicable in this large-scale AA scenario. Thus the question is: Can we find an optimal edge-weight value at which precision is acceptable, yet the number of possible attributions rises?

To investigate this question, we filtered the BCN created in run `ID=19` at different edge-weight values. Figure 2 visualises the relation between precision and the number of possible attributions at different edge-weight thresholds. For that run, the average edge-weight lies at 3 (median=2) while at maximum edge-weight (max=12) a precision of `P=0.90` can be achieved, as mentioned above.

Although choosing edges with an edge-weight of 2 and higher would result in 594 possible attributions, the precision in that scenario is far too low (`P=0.20`) to be considered in an AA task. It is obvious that at 594 attributions multiple false positives (incorrect attributions) get created. Nevertheless, an acceptable precision of `P=0.87` can be achieved when considering links between nodes with `edge-weight>=10`. However only 11 possible attributions are possible. Of these 11 books with previously unknown authorship, three can now be attributed to Martin Brun, three to J.L. Bynch and one book each to other authors.

To sum up, even in a high edge-weight scenario, at which we try to maximise the number of possible attributions at a reasonable level of precision, too few attributions are possible.

Figure 3 visualizes the BCN for run `ID=19` with pamphlets with known authorship highlighted in green and pamphlets with unknown authors highlighted in red. The overall layout of the network doesn't allow for profound conclusions from a distant reading perspective. This is partially the case because we homogenized the data-set, filtering

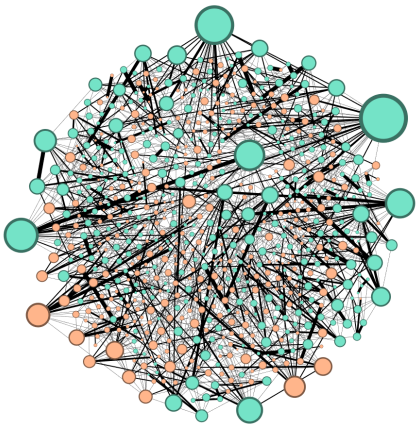

Fig. 3. Visualization of the BCN for run `ID=19`. Node size represents degree i.e. the frequency of how often a pamphlet was chosen as nearest neighbour by the algorithm. Edge-weight is visualized by line size. Pamphlets with known authors are highlighted in green.

out pamphlets not following a specific form of prose or pamphlets in other languages. Nevertheless, by sizing nodes according to their degree, we can see that some pamphlets seem to be very central i.e. being the nearest neighbour to many other pamphlets. The average degree of the network is 4, which means that on average a pamphlet was picked four times to be the nearest neighbour of another pamphlet. The highest degree is 41 which is a pamphlet written by Rasmus Fleischer. Rasmus Fleischer is the author of three pamphlets and compared to others not very prominent in the collection. The reason why it might be a prominent node/pamphlet might originate in the length of the document. This pamphlet is 27 661 tokens long and thus among the longest pamphlets in the collection. This offers many possibilities of stylistic similarities to other texts.

7 Discussion and Conclusion

The BCN method is a visualization technique that tries to overcome the reliability issues of visual methods used in stylometry [6]. However, due to the idea of linking the nearest stylistic neighbour to every text in the bootstrap network generation process, the method seems like a natural fit for investigating AA problems. In this paper, we explored the potential of the BCN method for large-scale AA by applying it to a corpus of 18th-century danish pamphlets.

The results of our experiments show that the best BCN would only link 27% of pamphlets correctly. This way the BCN method – when considering all edge-weights – is simply not precise enough to be applied in a large-scale AA task. When only considering edges with maximum weights, we learn that bi-gram character shingles without punctuation removal (run `ID=19`) lead to the best correct attribution rate. Although a precision of `P=0.90` can be achieved in that scenario only 7 attributions would be possible. Thus we investigated whether we can find an optimal edge-weight value at which

both precision and the number of possible attributions are maximised. At run `ID=19` and `edge-weight>=10` 11 attributions with a precision of `P=0.87` are possible. Although precise enough to be trusted as accurate attribution still only 3% (11 of 360) of anonymous pamphlets would be solved, which again allows the conclusion that the method is not applicable in a large-scale context.

The BCN approach solves some reliability problems that one-snapshot-at-a-time visualizations like classic dendrograms have. Nevertheless, some arbitrary decisions when applying the method can not be avoided [6]. Our study shows that the construction of BCNs is still very sensitive to the distance measure and type of variables used. In our experiments, the number of total links to construct the network varies greatly between 622 and 3793 (see Table 1). This is yet problematic as in our experiments we only used the most similar books (nearest neighbours) not considering any runner-ups. Adding runner-ups would lead to even more links between texts, which consequently means that signals from other layers but authorial style influence the network creation. Although one might not be interested in performing pure authorship analysis, but only in producing a visualization that allows for drawing a conclusion from a distant-reading perspective, researchers should be interested in a BCN that is characterised by as few false-positives as possible.

We have to note, however, that we are working with a real-life case and reasons for why we see a rather poor precision in the attribution scenario can be manifold. On the one hand, it might be that the BCN method is inadequate for the task. On the other hand, it can also be that the data-set we are dealing with is a very challenging one. In section 4.1 we already presented some rationales on why this might be the case. After all, it is a difficult task to reduce a text to its authorial style in the form of feature vectors. As Eder puts it: ”text is a multilayered phenomenon in which particular layers [e.g. authorship, topic or genre] are correlated” [6, p.53]. This might be especially pronounced in our case and make the task even more challenging.

In the future, further experiments applying the BCN method to other corpora should be performed to get a better idea of its applicability in AA tasks. Furthermore, more advanced approaches performing well in the area of digital forensics can be applied in the context of the freedom of the press writings to solve the problem of anonymous authorship within that data-set.

References

1. Al-Yahya, M.: Stylometric analysis of classical arabic texts for genre detection. Electronic library **36**(5), 842–855 (2018).
2. Argamon, S.: Interpreting burrows's delta: Geometric and probabilistic foundations. Digital Scholarship in the Humanities **23**(2), 131–147 (2008).
3. Bagnall, D.: Author identification using multi-headed recurrent neural networks. Proc. of PAN Workshop ECIR (2015).
4. Burrows, J.: Delta: a measure of stylistic difference and a guide to likely authorship. Digital Schoolarship in the Humanities **17**(3), 267–287 (2002).
5. Eder, M.: Mind your corpus: systematic errors in authorship attribution. Digital Scholarship in the Humanities **28**(4), 603–614 (2013).
6. Eder, M.: Visualization in stylometry: Cluster analysis using networks. Digital Scholarship in the Humanities **32**(1), 50–64 (2015).

7. Eder, M.: Elena ferrante: A virtual author. In: Tuzzi, A., Cortelazzo, M.A. (eds.) Drawing Elena Ferrantes Profile, pp. 31–45. Padova University Press, Padova (2018).
8. Eder, M., R.J., Kestemont, M.: Stylometry with r: a package for computational text analysis. R Journal **8**(1), 107–121 (2016).
9. Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielstrm, S., Schch, C., Vitt, T.: Understanding and explaining delta measures for authorship attribution. Digital Scholarship in the Humanities **32**(2), ii4–ii16 (2017).
10. Halvani, O., Winter, C., Graner, L.: On the usefulness of compression models for authorship verification. In: Proc. of ARES'17. pp. 54:1–54:10. ACM (2017).
11. Halvani, O., Winter, C., Graner, L.: Assessing the applicability of authorship verification methods. In: Proc. of ARES'19. pp. 38:1–38:10. ACM, New York, NY, USA (2019).
12. Halvani, O., Winter, C., Pflug, A.: Authorship verification for different languages, genres and topics. Digital Investigation **16**, S33–S43 (2016).
13. Hernandez-Castaneda, A., Calvo, H.: Author verification using a semantic space model. Computacion y Sistemas **21**, 167–179 (2017).
14. Horstbøll, H.: Luxdorphs samling af trykkefrihedens skrifter 1770-1773. Fund og Forskning **44**, 397–440 (2005).
15. Horstbøll, H.: The politics of publishing : Freedom of the press in denmark 1770-1773. In: Ihalainen, P. (ed.) Scandinavia in the age of revolution : Nordic political cultures, 1740-1820, pp. 145–156. Ashgate (2011).
16. Jockers, M.L., Witten, D.M.: A comparative study of machine learning methods for authorship attribution. Digital Scholarship in the Humanities **25**(2), 215–223 (2010).
17. Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval **1**(3), 233–334 (2008).
18. Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., Daelemans, W.: Authenticating the writings of julius caesar. Expert Systems with Applications **63**, 86 – 96 (2016).
19. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. JASIST **60**(1), 9–26 (2009).
20. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. JA-SIST **65**(1), 178–187 (2014).
21. Laursen, J.C.: Luxdorph's press freedom writings: Before the fall of struensee in early 1770s denmark-norway. The European Legacy **7**(1), 61–77 (2002).
22. Moretti, F.: Graphs, maps, trees : abstract models for a literary history. Verso, London (2005)
23. Mosteller, F., Wallace, D.L.: Inference in an authorship problem. Journal of the American Statistical Association **58**(302), 275–309 (1963).
24. O'Sullivan, J., Bazarnik, K., Eder, M., Rybicki, J.: Measuring joycean influences on Flann OBrien. Digital Studies / Le Champ Numrique **8**(1) (2018).
25. Potha, N., Stamatatos, E.: A profile-based method for authorship verification. In: Likas, A., Blekas, K., Kalles, D. (eds.) Artificial Intelligence: Methods and Applications. pp. 313–326. Springer, Cham (2014).
26. Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J.M., Köhler, J., Lötzsch, W., Müller, F., Müller, M.E., Paßmann, R., Reinke, B., Rettenmeier, L., Rometsch, T., Sommer, T., Träger, M., Wilhelm, S., Stein, B., Stamatatos, E., Hagen, M.: Who wrote the web? revisiting influential author identification research applicable to information retrieval. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. pp. 393–407. Springer, Cham (2016).
27. Rebora, S., Herrmann, J.B., Lauer, G., Salgaro, M.: Robert Musil, a war journal, and stylometry: Tackling the issue of short texts in authorship attribution. Digital Scholarship in the Humanities **34**(3), 582–605 (2018).
28. Rybicki, J., Hoover, D., Kestemont, M.: Collaborative authorship: Conrad, ford and rolling delta. Digital Scholarship in the Humanities **29**(3), 422–431 (2014).

29. Stamatatos, E.: A survey of modern authorship attribution methods. JASIST **60**(3), 538–556 (2009).
30. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy **21**(2) (2013).