The Classical Quarterly (2021) 71.2 819–848
doi:10.1017/S0009838821000987

SOME CONSIDERATIONS ON THE ATTRIBUTION OF THE 'NEW APULEIUS'*

ABSTRACT

The 'New Apuleius' is a set of Latin summaries of Plato's works first published in 2016 by Justin Stover, who attributed it to Apuleius. The present article attempts to assess two key aspects of Stover's argument, viz. his reconstruction of the manuscript transmission of the new text and his use of computer-assisted stylometric techniques. The authors suggest that both strands of his argument are inconclusive. First, it is argued that the transposition of gatherings in the archetype of the Apuleian philosophica *as envisaged by Stover is highly unrealistic. Second, replications of Stover's stylometric experiments show that their results are highly dependent on the particular algorithm settings and on the composition of the corpus. It is further shown that Stover's choice of highly specialized stylometric techniques is suboptimal, because popular generalist methods for statistical data analysis are demonstrably more successful in correctly identifying authors of Latin text fragments and do not support the case for Apuleius' authorship of the new text. The authors conclude that there are no solid grounds to conclude that the 'New Apuleius' was indeed written by Apuleius.*

Keywords: 'New Apuleius'; Apuleius; attribution; transmission; computer-assisted stylometry; Burrow's Delta

I. INTRODUCTION

In 2016, Justin Stover produced an *editio princeps* of a new Latin text, which he has identified as a work by Apuleius, in particular as the third book of the *De dogmate Platonis*.[1] While some aspects of Stover's argument have already received ample attention in the ensuing discussion of this attribution, with some scholars accepting the attribution and some rejecting it,[2] the problems connected with two contentions central for his reasoning have not received proper treatment.

* The authors of the article thank B. Kayachev, C.M. Lucarini and the anonymous *CQ* readers for their help and advice.

[1] J.A. Stover, *A New Work by Apuleius: The Lost Third Book of the* De Platone (Oxford, 2016). We shall call this text the 'New Apuleius', since both titles proposed by Stover (*De Platone et eius dogmate liber tertius* and *De Platonis pluribus libris compendiosa expositio*) seem unacceptable to us: the first assumes Stover's identification of the text, and this identification, as we shall show, is extremely controversial; against Stover's risky hypothesis that the second title, preserved in a manuscript that does not contain the new text, originally referred to the 'New Apuleius', see G. Magnaldi, 'Review of J.A. Stover, *A New Work by Apuleius: The Lost Third Book of the* De Platone (Oxford, 2016)', *Exemplaria Classica* 21 (2017), 367–76, at 368.

[2] Thoroughly sympathetic towards Stover's conclusions are H. Tarrant, 'Review of J.A. Stover, *A New Work by Apuleius: The Lost Third Book of the* De Platone (Oxford, 2016)', *JHPh* 55 (2017), 158–9 and J.G. Rheins, 'The arrangement of the Platonic corpus in the newly published *Compendiosa expositio* attributed to Apuleius of Madaura', *Phronesis* 62 (2017), 377–91; reluctant to connect the text with the *De dogmate Platonis* are C. Hoenig, 'Review of J.A. Stover, *A New*

The first is his views on the transmission scenario.[3] Although Stover places most of this section after the rest of his argument, the fact that the 'New Apuleius' is preserved in manuscripts of Apuleius' *philosophica* has clearly been the initial stimulus to attempt attributing it to Apuleius,[4] and the particular scenario reconstructed by Stover is actually the main grounds for his identification of the text as *De dogmate Platonis* Book 3. While Stover's reconstruction of the stemma of Apuleius' *philosophica*, implying that the manuscripts preserving the 'New Apuleius' could have derived it directly from the archetype of the whole tradition,[5] has been discussed by specialists in the transmission of Apuleius' philosophical corpus,[6] the very scenario of transpositions that occurred in this tradition has not so far become an object of special attention.

The second, even more important, hitherto undiscussed part of Stover's argument is his use of stylometry, in particular in the section entitled 'Computational analysis'.[7] In it Stover briefly presents the results of several stylometric experiments and refers the readers to two papers on computational approaches to the attribution of the 'New Apuleius', which he co-authored,[8] for additional argumentation and detail. Generally greeted with uncritical admiration by reviewers, including even those who are sceptical about Stover's attribution,[9] in fact his computational methodology has not received any detailed evaluation whatsoever.

Work by Apuleius: The Lost Third Book of the De Platone (Oxford, 2016)', *CPh* 113 (2018), 227–32 with C. Hoenig, *Plato's* Timaeus *and the Latin Tradition* (Cambridge, 2018), 102–3 n. 1 and J.M. Dillon, 'Review of C. Moreschini, *Apuleius and the Metamorphoses of Platonism* (Turnhout, 2015)', *International Journal of the Platonic Tradition* 12 (2018), 190–2, at 192; rather sceptical about Apuleius' authorship but convinced that the text is connected with Middle Platonism are G. Hays, 'Keeping things Platonic', *Times Literary Supplement* 5903 (2016), 29 (https://www.the-tls.co.uk/articles/keeping-things-platonic/) with G. Hays, 'Notes on the "New Apuleius"', *CQ* 68 (2018), 246–56, at 246, 248, C. Moreschini, 'Review of J.A. Stover, *A New Work by Apuleius: The Lost Third Book of the* De Platone (Oxford, 2016)', *BMCRev* 2017.03.31 (http://bmcr.brynmawr.edu/2017/2017-03-31.html), M. Bonazzi, 'Plato systematized: doing philosophy in the imperial schools', *OSAPh* 53 (2017), 215–36, C.M. Lucarini, 'Über das dem Apuleius zu Unrecht zugeschriebene vaticanische mittelplatonische Fragment', *ZPE* 211 (2019), 64–9; decidedly sceptical about Stover's conclusions are C.P. Jones, 'Review of J.A. Stover, *A New Work by Apuleius: The Lost Third Book of the* De Platone (Oxford, 2016)', *Sehepunkte* 17 (2017), Nr. 10 (http://www.sehepunkte.de/2017/10/28809.html) and Magnaldi (n. 1). The aspects covered in the discussions are style (both general impressions and particular features), philosophical contextualization of the text, Stover's stemma of Apuleius' *philosophica*, parallels with Apuleius, and the reasons for identification with *De dogmate Platonis* Book 3.

[3] Stover (n. 1), 12–18, 45–59.

[4] Stover (n. 1), ix: 'It became increasingly clear to me that … [the text's] collocation with the *philosophica* in the Vatican manuscript was no happenstance'; cf. Tarrant (n. 2), 158.

[5] See also J.A. Stover, 'Apuleius and the Codex Reginensis', *Exemplaria Classica* 19 (2015), 131–54.

[6] Jones (n. 2), Moreschini (n. 2), Magnaldi (n. 1).

[7] Stover (n. 1), 34–44.

[8] J.A. Stover, Y. Winter, M. Koppel and M. Kestemont, 'Computational authorship verification method attributes a new work to a major 2nd-century African author', *Journal of the Association for Information Science and Technology* 67 (2016), 239–42 and J. Stover and M. Kestemont, 'Reassessing the Apuleian corpus: a computational approach to authenticity', *CQ* 66 (2017), 645–72.

[9] Cf. e.g. Tarrant (n. 2), 158: 'a compelling case for Apuleian authorship supported by a proven and up-to-date kind of computer-assisted vocabulary analysis'; Bonazzi (n. 2), 217 n. 9: 'this section is truly remarkable'; Rheins (n. 2), 379 n. 3: 'Perhaps the most impressive line of evidence is the cutting-edge stylometry, which not only gives strong evidence that the author is Apuleius, but also that this text belongs to the *De Platone.*' Moreschini (n. 2) chooses instead to ignore the section without examining it: 'Stover ci scusi se non abbiamo compreso, a causa della nostra età, la "Computational Analysis"'.

The purpose of the present paper is to offer an examination of the reliability of these aspects of Stover's argument. First, in section II, we show that Stover's transmission scenario rests on highly unrealistic assumptions. Section III then presents an overview of the stylometric approaches used by Stover. We show that his experiments can be replicated but their outcome is highly dependent on the composition of the reference corpus. Section IV contrasts Stover's methodology, which relies on the use of novel stylometric methods (Bootstrap Consensus Trees, Impostor Method), with popular generalist approaches for analysing complex data: the UMAP algorithm for dimensionality reduction and gradient-boosting-machine and random-forest-based binary classifiers. We show that these methods do a much better job of attributing textual fragments to correct authors both in Stover's reference corpus and in our extended corpora, and that they do not support Apuleius' authorship of the text. Section V concludes.

II. TRANSMISSION SCENARIO

The main source for the text is MS Vat. Reg. lat. 1572 (thirteenth century). Before the 'New Apuleius', it contains the Pseudo-Apuleian *Asclepius*, Chalcidius' translation of Plato's *Timaeus*, as well as Apuleius' *De deo Socratis*, *De dogmate Platonis* and *De mundo*.[10] The 'New Apuleius' comes last at fols. 77r–86r, without any *incipit* or *explicit* (the beginning and the end of the 'New Apuleius' are lost), and is separated from the *De mundo* by the start of a new line only.

A few lines are also preserved in a thirteenth- or fourteenth-century manuscript in Venice, Marc. lat. VI.81 (3036).[11] Here again the text follows the standard group *De dogmate Platonis* + *De mundo* on fol. 130v and is attached directly to the end of the *De mundo*. After a couple of lines corresponding to the beginning of the 'New Apuleius' in Vat. Reg. lat. 1572, the text ends abruptly with the words *Explicit Apuleius de dogmate Platonis liber tertius*.[12] The incipit of the *De mundo* and the running titles of the manuscript show that this title refers to both the *De mundo* and the fragment of the 'New Apuleius' regarded as a single text.[13] Before the *De dogmate Platonis*, the manuscript contains Cicero, *De finibus*, *Timaeus*, *Lucullus*, *De diuinatione* and *De fato*, Henry Aristippus' translation of Plato's *Phaedo* and Apuleius' *De deo*

[10] Stover (n. 1), 3. A digitized version of the manuscript is now available online: https://digi.vatlib.it/view/MSS_Reg.lat.1572. For the sake of simplicity we shall refer to the *De dogmate Platonis* and the *De mundo* as works by Apuleius, although the question of authorship of both texts is debated: see, for example, S.J. Harrison, *Apuleius: A Latin Sophist* (Oxford, 2000), 174–80. Stover and Kestemont (n. 8), 670 claim that computational methods have allowed them to make the 'unlösbares Echtheitsproblem' (J. Redfors, *Echtheitskritische Untersuchung der apuleischen Schriften De Platone und De mundo* [Lund, 1960], 115) of the *De dogmate Platonis* and the *De mundo* 'a *lösbares* (and even *gelöstes*) *Echtheitsproblem*', but see below, sections III–IV, for the reliability of the computational approaches used by Stover.

[11] Although referred to by Stover as Marc. lat. VI.31 (3036); Moreschini (n. 2) notes that Marc. lat. VI.81 (3036) is how the manuscript is referred to in R. Klibansky and F. Regen, *Die Handschriften der philosophischen Werke des Apuleius: Ein Beitrag zur Überlieferungsgeschichte* (Göttingen, 1993), 120, and the correction is confirmed by the handwritten catalogue of Latin manuscripts in the Biblioteca Marciana now available online, http://cataloghistorici.bdi.sbn.it/file_viewer.php?IDCAT=244&IDGRP=2440008&LEVEL=1&PADRE=2440006&PROV=INT, at fol. 32v and 38r: Marc. lat. VI.31 contains 'Apollinaris Offredi Cremonensis, Quaestiones in Libros Posteriorum Analyticorum Aristotelis', not Apuleius, and has an alternative number 3016, not 3036.

[12] Stover (n. 1), 6.

[13] Stover (n. 1), 48.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

Socratis; after the 'New Apuleius' fragment, it has the *Asclepius*, Gundissalinus' *De unitate et uno* and Robert Grosseteste's translation of the *Testamentum XII patriarcharum* (some of the texts are given in excerpts).[14]

According to Stover, in the archetype of the tradition of Apuleius' *philosophica* (ω), the 'New Apuleius' initially constituted the third book of the *De dogmate Platonis* and consequently occupied the position between *De dogmate Platonis* Book 2 and the *De mundo*. The following sequence of events then unfolded:

1. ω suffered physical damage that led to the transposition of the 'New Apuleius' (which concomitantly lost its beginning and end) to a wrong place after the *De mundo*. The resulting state of ω is called ω_1 in Stover's book.
2. φ—the ancestral manuscript of Vat. Reg. lat. 1572 (R) and Marc. lat. VI.81 (Z)—was copied from ω.
3. ω lost the 'New Apuleius' completely, and this state of the manuscript (referred to by Stover as ω_2) served as an archetype for the rest of the tradition.

The *incipit* of *De dogmate Platonis* Book 3, however, was preserved in the archetype between *De dogmate Platonis* Book 2 and the *De mundo*, for, as Stover argues, it was normal for late antique manuscripts to have 'elaborate *subscriptiones* on the final page of a book, with the text of the following book beginning on the next page, and these *subscriptiones* would include the incipit of the successive book'.[15] If the boundary between books coincided with the boundary between quaternions (or gatherings), the *incipit* could have been left there after the gathering(s) containing the 'New Apuleius' had been transposed. This scenario, according to Stover, accounts for the fact that the *De mundo* is named *De dogmate Platonis* Book 3 in some of the medieval manuscripts, including R and Z, for after such a transposition the *incipit* of *De dogmate Platonis* Book 3 would have appeared to refer to the *De mundo*.

We leave the assessment of Stover's stemma to specialists in the tradition of the *philosophica*,[16] but a few words need to be said about the scenario itself. To get to the position directly after the *De mundo*, the 'New Apuleius' must have begun exactly at the beginning of a gathering and must have ended exactly at the end of the same or, more probably, another gathering; besides, the *De mundo* also must have ended exactly at the end of a gathering.[17] However, such multiple correspondences between boundaries of books and gatherings are very difficult to account for: beginning a new book with a new gathering on purpose in a situation of parallel creation of different gatherings[18] would be a very uneconomical strategy, potentially leading to waste of parchment; and a coincidence of this kind does not seem to be a very likely one.

[14] Klibansky and Regen (n. 11), 120–1.

[15] Stover (n. 1), 49.

[16] Magnaldi (n. 1) actually allows for the possibility that MSS R and Z (she follows Stover in considering the latter an apograph of R) form an independent branch of the tradition, insisting, though, that MS Bruxell. 10054–6 (B) was copied from ω earlier than φ, since R often agrees in error with δ (the third branch) against B. However, at least some of these readings of Rδ could perhaps also be explained by contamination.

[17] Stover (n. 1), 51–2 makes clear that he imagines the transposition of the 'New Apuleius' as a transposition of its gathering(s); the loss of the beginning and end of the text must also have been connected with the damaged state of the external folia of this gathering or group of gatherings on his hypothesis.

[18] See J. Vezin, 'La répartition du travail dans les "scriptoria" carolingiens', *JS* (1973), 212–27; M.D. Reeve, '*Eliminatio codicum descriptorum*: a methodological problem', in J.N. Grant (ed.), *Editing Greek and Latin Texts* (New York, 1989), 1–35, at 28.

As for the identification of the *De mundo* with *De dogmate Platonis* Book 3 in some of the manuscripts, it does not necessarily derive from an *incipit* that survived in the tradition after the work it referred to had been lost. A book (or a poem) following a text consisting of numerous books (poems) was obviously liable to being treated as an additional book (poem) of the same text, as can be seen from the examples of Tacitus' *Historiae*, whose books are numbered in Laur. plut. 68.2 and its descendants as *Annales* 17–21,[19] or Nemesianus' *Eclogues*, numbered in most manuscripts as *Eclogues* 8–11 by Calpurnius.[20]

To sum up, we probably cannot positively say that the transposition scenario proposed by Stover is impossible, but the evidence certainly does not make his interpretation necessary: on the contrary, it appears to make it rather improbable. Consequently, as far as transmission is concerned, there appears to be no reason to believe that the 'New Apuleius' originally formed part of the corpus of Apuleius' *philosophica*.

III. STOVER'S USE OF STYLOMETRY

In this section, we will provide an overview of the stylometric methods used by Stover and his co-authors with a particular focus on Bootstrap Consensus Trees. We will show that, even though the results they obtained can be replicated, these results are unstable and heavily dependent on the composition of the reference corpus.

Preliminary computations

Before reporting the results of computer-assisted stylometric experiments, Stover resorts to a simpler stylometric argument showing that the words *alioquin* and *enimuero* and the conjunction *nec non* are used by both Apuleius and the author of the 'New Apuleius' with frequencies that are much higher than the average in the corpus of Classical Latin prose.[21] In table 1 we contrast the frequencies of these words in the 'New Apuleius' and in Apuleius, on the one hand, and in Classical Latin prose in general, on the other hand.[22]

[19] Although there existed a view that this represented the book-numbering of an ancient complete edition of Tacitus' historical works (see, for example, O. Seeck, 'Der Anfang von Tacitus *Historien*', *RhM* 56 [1901], 227–32, at 227), it appears much more likely that the *Annales* originally consisted of eighteen books (see R. Syme, *Tacitus* [Oxford, 1958], 686–7) and consequently this manuscript book-numbering does not derive from an ancient source. Neither can it derive from an original *incipit* of Book 17 preserved at the end of Book 16 without the book itself, since the end of Book 16 is also not extant. Probably the numbering of the books of the *Annales* was simply continued to the following books by some scribe.

[20] See M.D. Reeve, 'Calpurnius and Nemesianus', in L.D. Reynolds (ed.), *Texts and Transmission: A Survey of the Latin Classics* (Oxford, 1983), 37–8, at 37. Note also that the *explicit* to Porphyrio's commentary on Horace's *Epodes* calls them *carminum liber v*.

[21] Stover (n. 1), 36–8. In absolute numbers, though, the 'New Apuleius' only has two instances of *enimuero*.

[22] For Apuleius, we consider maximum frequencies of each word in a single book instead of the average frequencies of corresponding words in the whole of the Apuleian corpus. This and the following tables show these maxima in the 'Apuleius (max.)' columns. The average will in every case be much lower. The reason for this decision is that frequencies of individual function words in different works by Apuleius can vary significantly. For instance, the frequency of the word *igitur* in the *Apologia* is 0.3%, which is extremely high, while in the *Metamorphoses* Book 8 Apuleius does not use this word at all; cf. e.g. H. Jordan, *Kritische Beiträge zur Geschichte der lateinischen Sprache* (Berlin, 1879),

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

Table 1: The frequencies of *alioquin*, *enimuero* and *nec non*

	'New Apuleius'	Apuleius (max.)	CL prose (average)
alioquin	0.19%	0.12%	0.01%
enimuero	0.04%	0.22%	0.002%
nec non	0.13%	0.09%	0.002%

Stover's statement appears to be corroborated by this table. However, the problem with this argument is that only three words were chosen in support of what appears to be a convincing picture. A different choice gives a different outcome. For instance, if we take the words *deinde*, *autem* and *quia*, which are not so extremely rare (this insures us against relying on values comparable to accidental fluctuations in the data) and at the same time are also clearly used by the author of the 'New Apuleius' with an unusually high frequency, we will have table 2.

Here it is rather the difference between the usage of the 'New Apuleius' and Apuleius (including the *De dogmate Platonis*) that comes to the fore. The 'New Apuleius' (a comparatively short text, 4,742 words long) uses *deinde* 35 times, while in the whole of the Apuleian corpus (*c*.100,000 words) we only find it 14 times. *quia* is used 20 times in the 'New Apuleius' and 14 times in the whole of the Apuleian corpus. To judge from the extremely high frequency of *autem*, a much closer parallel to the 'New Apuleius' than Apuleius would be, for example, John Scottus Eriugena (0.9% in the first book of his translation of the *Areopagitica*); the frequency of *quia* is comparable in some late antique and medieval texts, mainly didactic or technical (0.32% in Martianus Capella 9, 0.37% in *Digesta* 2 and in Eriugena, *Areopagitica* 1, 0.61% in Hugh of Saint Victor, *Didascalicon* 1). We were unable to find a parallel for such a high frequency of *deinde*.[23]

Burrows's Delta and Bootstrap Consensus Trees

Stover's main tool for automated stylometric analysis are Bootstrap Consensus Trees, a text-oriented hierarchical clustering algorithm based on averaging results obtained using different subsets of the original data (relative frequencies of words in textual fragments).

To quantify differences between texts Stover and his co-authors use Burrows's Delta,[24] a widely adopted method of measuring stylistic differences between texts

325, H. Becker, *Studia Apuleiana* (Berlin, 1879), 7–53, E. Löfstedt, *Syntactica: Studien und Beiträge zur historischen Syntax des Latein* (Lund, 1956), 1.334–5 and, in general, E. Norden, *Die antike Kunstprosa: Vom VI. Jahrhundert v. Chr. bis in die Zeit der Renaissance* (Leipzig and Berlin, 1915–18), 2.603, B. Axelson, 'Akzentuierender Klauselrhythmus bei Apuleius. Bemerkungen zu den Schriften "De Platone" und "De mundo"', in id., *Kleine Schriften zur lateinischen Philologie* (Stockholm, 1987), 233–45, at 234. To allow for this variation, we compare the exceptionally high frequencies found in the 'New Apuleius' with maximum frequencies found in Apuleius. We do not include the *Asclepius* and the Περὶ ἑρμηνείας into our count as probably non-Apuleian: see, for example, Harrison (n. 10), 11–13. For Classical Latin prose, we give the average figure for all the Classical Latin prose texts included in the PHI5 corpus.

[23] In Stover and Kestemont (n. 8), 665–70, some other observations on the frequencies of particular words in the 'New Apuleius' are given, but they are even less conclusive than those discussed above.

[24] J. Burrows, '"Delta": a measure of stylistic difference and a guide to likely authorship', *Literary and Linguistic Computing* 17 (2002), 267–87.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

Table 2: The frequencies of *deinde*, *autem* and *quia*

	'New Apuleius'	Apuleius (max.)	CL prose (average)
deinde	0.73%	0.07%	0.08%
autem	1.12%	0.28%	0.19%
quia	0.42%	0.03%	0.10%

from a predefined corpus based on relative frequencies of words in those texts. In order to compute pairwise differences between texts in a corpus, a table is constructed with rows for texts and columns for diagnostic words. First, cells are filled with relative frequencies of words in the texts (how many times a given word is found divided by the length of the text). Then the values in the cells are *z*-scaled: from each cell the mean for its column is subtracted and the result is divided by the standard deviation for the column. This dampens the effect of rare and unusual words actively used by some authors but not by others, as those will have higher standard deviations. The resulting values show how far each author departs from the general trend in her frequency of use of particular words. The Delta for a pair of texts is then equal to the sum of the absolute values of pairwise differences between scaled relative frequencies of all diagnostic words.

A further measure, which ensures the robustness of the procedure, is to select words for the analysis manually: contemporary stylometry often relies on the premise that the best way to ascertain the identity of the author of a given text is to measure the use of function words in it. It is assumed that authors are less likely to be able to control the differential rates of conjunctions, prepositions and other non-content-related words than to keep track of which nouns or verbs they employ.[25] In some of Stover's and our experiments reported below, a manual selection of function words was performed. In others, most frequent words (MFWs) were extracted from the texts and the only kind of filtering applied was the removal of personal pronouns, whose distribution may be coupled with a text's genre.[26]

The result of the application of Burrows's Delta to a corpus is a distance matrix: a table of pairwise differences between texts. This matrix can be used to perform various kinds of analyses, including clustering, that is, dividing texts into several groups or constructing a tree-like graph of their relationships (so-called hierarchical clustering). Clustering can be used both to validate a stylometric method (texts by the same author should be reliably grouped together) and to corroborate or disprove hypotheses about authorship (if we assume that a text was written by some author, we expect samples from this text to end up in the same subtree or cluster as this author's other works).

It is known, however, that hierarchical-clustering trees can be sensitive to small fluctuations in the data. One way to make results more robust is to use different clusterization algorithms and compare their outputs. Another approach is to repeatedly

[25] D.I. Holmes, 'The evolution of stylometry in humanities scholarship', *Literary and Linguistic Computing* 13 (1998), 111–17.

[26] Under both approaches, it is possible to 'cull' the words by excluding those that are found in less than 50, 60, 70, etc. per cent of texts. Given the small size of the text samples used for Stover's and our experiments, it was impractical to use culling, because most of the words were missing from a large proportion of the samples.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

apply the same algorithm to different subsets of the data. The latter method is called *bootstrapping* in the statistical literature.

Bootstrap Consensus Trees (BCTs) used by Stover are built by averaging over hierarchical-clustering trees based on distance matrices computed by applying Burrows's Delta to different subsets of MFWs or manually selected function words. For example, when using MFWs, an initial 'frequency band' of top 50 or 100 words is selected, which is then gradually widened by some amount (usually 50 or 100) until a predefined limit (3,000 MFWs in Stover's experiments) is reached. A distance matrix is computed using Burrows's Delta, and a clustering tree based on this matrix is produced for each intermediate frequency range. Finally, the results are averaged by choosing such clusters of nodes as were contained in half of the trees or more. When manually selecting diagnostic words for analysis, their total number is usually lower (250 in Stover's experiments), the increment is smaller (Stover uses the increment of 1) and frequency bands are therefore more similar to each other.

The concept of Bootstrap Consensus Trees is widely used in bioinformatics.[27] It should be noted, however, that Stover's BCTs are rather idiosyncratic and that they do not adhere to the principles of bootstrapping in the strict sense. Classical bootstrap assumes that 'character vectors' used for repeated experiments (i) have the same size as the full dataset and (ii) are randomly sampled with replacement, which means, in our case, that frequencies of some words may be used several times.[28] Randomization makes sure that no particular subset of the data is overrepresented, and all replication trees have an unbiased view of the original dataset. The rationale for Stover's BCT approach, where the top-frequency bands reappear in the subsequent wider samples,[29] is that we want frequent function words to provide the foundation for the classification, but also do not want to completely discard information contained in the frequencies of content words. Stover used the open-source statistical library *stylo*[30] to build BCTs, and we also used it for our replication and validation experiments.

Replication

In order to ascertain that Stover's results are reliable and that we reproduce his method correctly, we first replicated Stover's experiment using the same corpus, which was published online,[31] and the same algorithm settings as reported in Stover's works. It must be pointed out, however, that the experiment reported in the book[32] was actually performed at least three times with slight modifications.

The first iteration, reported in the 2016 brief communication[33] and then referenced in the monograph,[34] was based on a slightly different corpus: compared to the corpus used in the book version, it lacked Cicero's *Tusculanae disputationes* but included Cicero's

[27] J. Felsenstein, 'Confidence limits on phylogenies: an approach using the bootstrap', *Evolution* 39 (1985), 783–91.

[28] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap* (Boca Raton, 1993).

[29] There is another version of the BCT procedure, where frequency bands are non-overlapping. It was not used in the experiments reported below.

[30] https://sites.google.com/site/computationalstylistics/stylo; cf. M. Eder, J. Rybicki, M. Kestemont, 'Stylometry with R: a package for computational text analysis', *R Journal* 8 (2016), 107–21.

[31] https://github.com/mikekestemont/Apuleius/tree/master/Texts/BCT.

[32] Stover (n. 1), 38–42.

[33] Stover, Winter, Koppel and Kestemont (n. 8), 240 fig. 1.

[34] Stover (n. 1), 38 n. 23.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

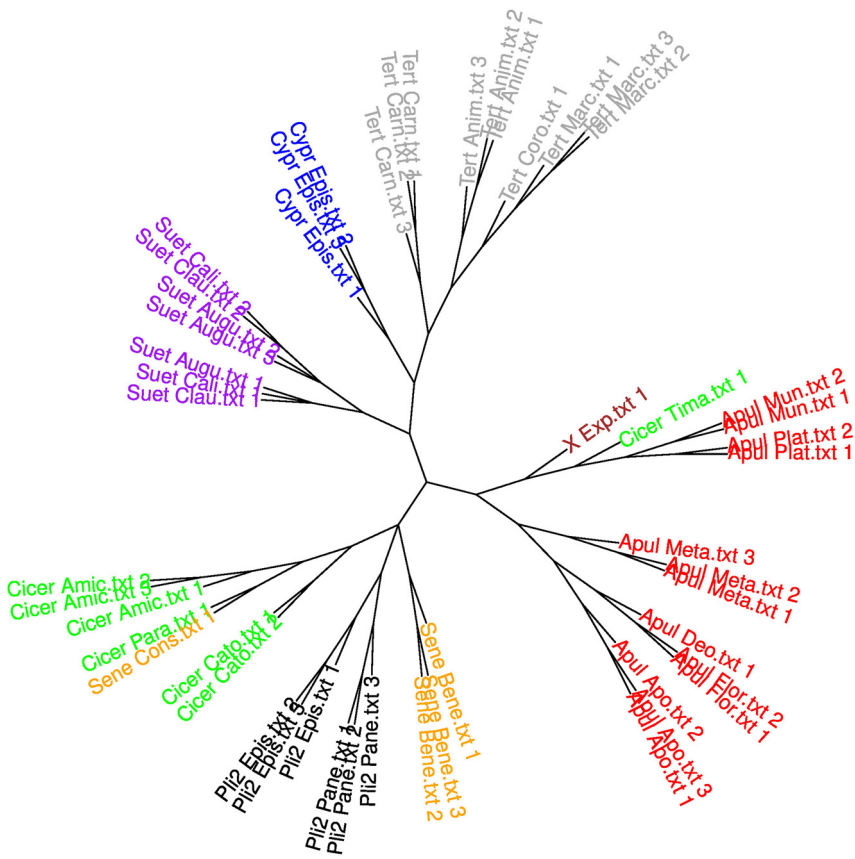

FIG. 1: Replication of the experiment reported by Stover and his co-authors, fig. 1 (in this replication we follow Stover in designating the 'New Apuleius' as *X_Exp*). (This figure is in colour in the online version of our article.)

Timaeus and *Paradoxa Stoicorum* and Tertullian's *Aduersus Marcionem*. A supplement to the paper describing the method in more detail was published online.[35]

According to the supplementary materials and to the description of fig. 1 in the paper, the longer texts were truncated to the first 9,000 words 'to maximise comparability'[36] and then split into two slices of 4,500 words each; texts shorter than 9,000 words are said to have been truncated to the first 4,500 words. However, the tree makes it clear that this description does not correspond to the actual experiment, since the longer texts are divided into *three* pieces instead of two, while shorter texts are sometimes represented by a single piece and sometimes divided into two parts, obviously depending on whether the text in question is longer than 6,000 words. We conjecture that sample lengths of 3,000 words were eventually chosen instead of 4,500 words, without reflecting this change in the

[35] https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fasi.23460&file=asi23460-sup-0001-si.doc.

[36] Supplementary materials (see note 35 above), page 1.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

descriptions. Texts between 6,000 and 9,000 words were truncated to 6,000 words and then divided into two pieces, while texts shorter than 6,000 words were truncated to 3,000 words. This reconstruction is confirmed by the fact that we were able to exactly replicate the results of the experiment using these parameter settings (cf. fig. 1 above).

The automated classification reported by Stover and his co-authors was largely the same as in the book. However, a fragment of the *Timaeus* ended up in the same subtree as the 'New Apuleius', the *De mundo* and the *De dogmate Platonis*, while Seneca's *De constantia sapientis* intruded into the subtree containing other works by Cicero. Consequently, the results of the experiment looked much less definitive and showed that the method is capable of attribution errors. The algorithm settings were also slightly different: instead of hand-picking non-content-related words, as in the book version, the authors simply based the calculation on 3,000 most frequent words (starting from 50) with personal pronouns removed.

The second iteration, eventually published in the 2017 paper,[37] but also referenced in the monograph,[38] also uses 3,000 most frequent words (starting from 100 this time) without manual selection except for personal-pronoun removal. The samples are now officially 3,000 words long. The corpus, however, is markedly different: there are no works by Cicero at all.[39] The results are more similar to those reported in the book: works by Apuleius form a separate subtree, which also includes the 'New Apuleius' but no texts by a different author.

The third iteration is that reported in the monograph.[40] Cicero's works were reintroduced, with the exception of the ill-behaved *Timaeus* and with the *Paradoxa Stoicorum* replaced by the *Tusculanae disputationes*; the set of Tertullian's texts was also reduced. The samples are now 4,522 words long (equal to the length of the shortest text, which is the *De deo Socratis*), and in most cases longer texts are not truncated;[41] an exception, however, was made for the *Metamorphoses*, which was truncated to the length of three samples, 'because in many experiments this text would otherwise form a separate branch in the bootstrap tree'.[42] Raw most frequent words were replaced with 'a list of 250 grammatical function words' with personal pronouns and 'topic-specific words (nouns such as *deus*, verbs such as *dicere*)' removed,[43] with the smallest frequency-bandwidth of 50 and an increment of 1.

This variability in the actual implementation of the analysis begs the question of the freedom that the stylometric method gives to the researcher. By altering the corpus and manually selecting a restricted set of 'grammatical function words', Stover was able to obtain the Consensus Tree that was even able to separate different stylistic strands in Pliny the Younger's writings and thus was a purportedly reliable indicator of the

[37] Stover and Kestemont (n. 8), 669 fig. 17.

[38] Stover (n. 1), 38 n. 23.

[39] Cf. Stover and Kestemont (n. 8), 655: 'We left out Cicero … in order not to overload the analysis and subsequent visualisation with different authors and texts.' What Stover and Kestemont (n. 8), 652 refer to as Tertullian, *Ad Marcianum* is obviously Tertullian, *Aduersus Marcionem*.

[40] Stover (n. 1), 40 fig. 2.

[41] The rationale for this decision was perhaps to make the comparison of Plin. *Ep.* 10 with Plin. *Ep.* 1–9 possible: Stover (n. 1), 41 uses the fact that slices corresponding to Plin. *Ep.* 10 stand in the resulting tree a little apart from the rest of Pliny's *Letters* to underline 'the sensitivity of the experiment', since '[t]he style of Book X, the correspondence with Trajan … is markedly different than that of the other books and was added to the corpus later; it also includes the responses from Trajan'.

[42] Stover (n. 1), 41 n. 30. '[T]he numerical overweight of a single, long text in the corpus' and the different genre of the *Metamorphoses* are given as possible explanations for this fact.

[43] Stover (n. 1), 39.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

authorship of the 'New Apuleius'. We do not know, however, if this combination of algorithm settings still produces results suggesting that Cicero's *Timaeus* was written by the same author. The results presented in the 2016 brief communication and the absence of Cicero's *Timaeus* from the subsequent experiments indicate that there may be content-related signals in the data that the algorithm, *pace* Stover, is unable to tease apart.

We chose to replicate the first of the three iterations, because the detailed (although not entirely accurate; cf. above) description of its methodology in the supplementary materials allowed us to reproduce it exactly. The results of our replication are reported in fig. 1, and they match the results reported by Stover and his co-authors.

In order to check if these results could be improved upon without any changes in the corpus, we created a BCT based on the list of 150 highly frequent, manually filtered, predominantly function words with pronouns removed (starting with the 50 most frequent words and going up to 150 by increments of 1; the same increment and starting range were used by Stover in his monograph).[44]

The results are reported in fig. 2 below. They show that there is a trade-off: basing the classification on the selected function words makes the method more robust (there are no mistaken attributions in the tree), but lowers its insightfulness (the works by Apuleius are split into two unrelated groups, one of which is further split into two subgroups, with the divisions largely corresponding to genre boundaries: the contested group *De dogmate Platonis* + *De mundo* vs the *Metamorphoses* in one subtree vs the rest of Apuleius' texts in another).[45] This arguably more performant approach does not treat the 'New Apuleius' as a text by Apuleius.

If we further make the experiment more conformant to what was reported in Stover's monograph by using the sample length of 4,500 words (and sacrificing the *Timaeus* and the *Paradoxa Stoicorum*, which are too short), we again will see the 'New Apuleius' clustered with the *De dogmate Platonis* and the *De mundo* (fig. 3 below). Cyprian's *Epistles*, however, are in this case also placed within Apuleius' subtree.

Dependence on the corpus

Stover's experiments probably do show that there is some affinity between the 'New Apuleius' and Apuleius (and in particular between the 'New Apuleius' and the *De dogmate Platonis* together with the *De mundo*) in the way in which they use frequent

[44] We used the following list (we were unable to find even 150 function words with high enough frequencies and had to include a couple of very frequent non-specific content words at the end of the list; Stover's list was not published but for the first ten words reported in Stover [n. 1], 39 n. 27): *et*, *in*, *non*, *ut*, *ad*, *cum*, *ab*, *sed*, *ex*, *si*, *de*, *etiam*, *enim*, *aut*, *ac*, *nec*, *per*, *atque*, *nam*, *uel*, *ne*, *quidem*, *autem*, *tamen*, *neque*, *uero*, *ita*, *iam*, *quoque*, *nihil*, *pro*, *modo*, *quia*, *quasi*, *inter*, *nisi*, *tunc*, *post*, *sic*, *igitur*, *tam*, *qua*, *ante*, *an*, *nunc*, *apud*, *magis*, *sine*, *ergo*, *at*, *deinde*, *ubi*, *dum*, *semper*, *minus*, *unde*, *contra*, *maxime*, *itaque*, *sicut*, *satis*, *denique*, *ob*, *simul*, *uti*, *sub*, *saepe*, *quamquam*, *numquam*, *ideo*, *propter*, *siue*, *quippe*, *prius*, *adhuc*, *quoniam*, *usque*, *inde*, *bene*, *sane*, *mox*, *item*, *super*, *quin*, *adeo*, *quamuis*, *cur*, *tamquam*, *postea*, *praeterea*, *potius*, *statim*, *uelut*, *postquam*, *supra*, *ceterum*, *certe*, *omnino*, *licet*, *forte*, *o*, *circa*, *rursus*, *tandem*, *diu*, *praeter*, *umquam*, *tot*, *ibi*, *hinc*, *haud*, *necesse*, *melius*, *paene*, *fere*, *namque*, *amplius*, *uix*, *scilicet*, *quum*, *iterum*, *aliquando*, *aduersus*, *seu*, *parum*, *plerumque*, *interim*, *prope*, *plus*, *intra*, *partim*, *olim*, *iuxta*, *ultra*, *male*, *quare*, *aliter*, *dolorem*, *fortasse*, *malis*, *primis*, *studio*, *agere*, *immo*, *quanto*, *domine*, *eiusdem*, *opera*, *oportet*, *publicam*.

[45] The same threefold division of the Apuleian corpus is often discernible in Stover and Kestemont's PCA experiments: see Stover and Kestemont (n. 8), 658–60, 663, 671.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

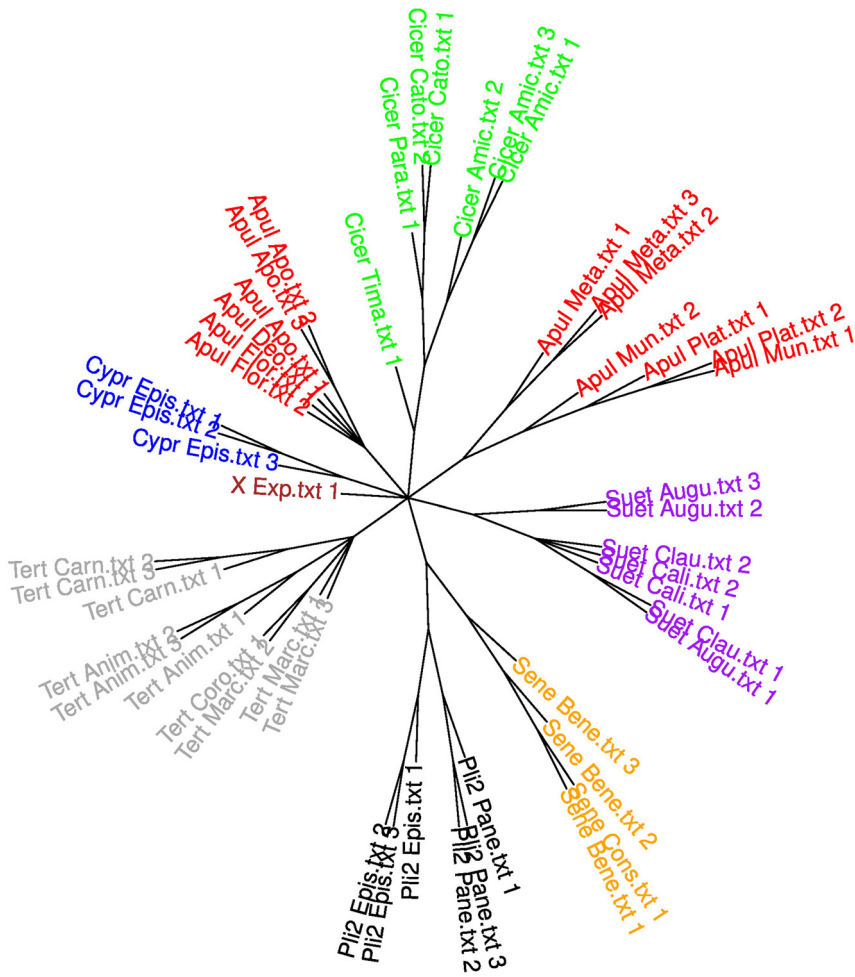

FIG. 2: Replication of the experiment reported by Stover and his co-authors with 150 manually selected function words used as a basis for classification. (This figure is in colour in the online version of our article.)

and function words. However, attributing the new text to Apuleius is not the only way to explain this affinity.

Most importantly, it is not clear whether the experiments picked up features shared specifically by the 'New Apuleius' and Apuleius alone or features shared by these texts together with other texts by different authors. The intrusion of Cicero's *Timaeus* into the same subtree in one of the experiments and of Cyprian's letters in the other seem to confirm the latter interpretation. Moreover, it is notable that in his experiments Stover uses only texts that are earlier or slightly later than the date he ascribes to the 'New Apuleius' (the second century A.D.).[46] As a result, only those texts are used that are

[46] Basing experiments on a preconceived dating is of course to be noted as a lapse on Stover's part. The date is deduced from considerations on the general cultural and philosophical context of the 'New

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

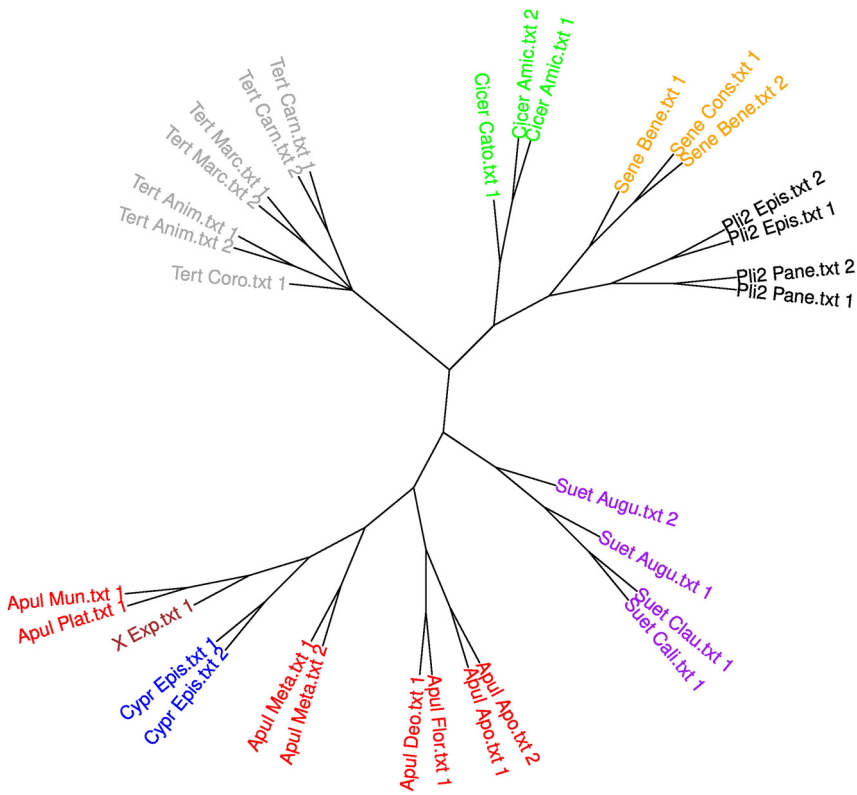

FIG. 3: Replication of the experiment reported by Stover and his co-authors with 150 manually selected function words and 4,500-word samples. (This figure is in colour in the online version of our article.)

patently dissimilar in their style from both Apuleius and the 'New Apuleius'. In other words, what Stover's experiments actually show might be that the 'New Apuleius' looks Apuleian when compared to Cicero, Pliny the Younger, Seneca, Suetonius and the early Christian writers, and not that it is objectively Apuleian.

More precisely, our working hypothesis is that the inclusion of the 'New Apuleius' in the same subtree as Apuleius in Stover's experiments reflects the fact that these texts display a set of features shared by a certain type of works but absent from the background texts used in Stover's experiment.

In order to investigate this possibility, we added to the corpus used in our replication of Stover's experiment (see the subsection on 'Replication' above) several texts that

Apuleius' (Stover [n. 1], 11–12, 23), but see Moreschini (n. 2) for other possible interpretations of the same data, and even Stover (n. 1), 23 confesses that the text 'could conceivably date from after 225; but if so, its contents, aims, and methods would have been decidedly retrograde'. Besides, no reasons are given for rejecting Raymond Klibansky's view that the text is a translation of a lost Greek work (R. Klibansky, *The Continuity of the Platonic Tradition during the Middle Ages* [Munich, 1981], 6–7, Klibansky and Regen [n. 11], 5); if we assume it is, the Latin version can in principle belong to a date much later than the second century A.D.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

share certain features of genre with the 'New Apuleius' and with some of the authentic Apuleian texts (we have chosen technical and didactic texts), represent various periods different from that to which Stover limits his choice of comparative material, and are already known to parallel in certain cases the unusually high frequencies of particular words shown by the 'New Apuleius' (see the subsection on 'Preliminary computations' above). The following texts have been added:

- Digesta[47]
- Martianus Capella
- Eriugena, *Areopagitica* (ninth century A.D.)
- Hugh of Saint Victor, *Didascalicon* (twelfth century A.D.)

In every case except Martianus Capella we used the first 9,000 words of the text adapted according to Stover's practice (that is, we removed the non-alphabetic characters, replaced all *v*s with *u*s and made all the words lowercase; in the case of the *Digesta*, we also deleted source references). In the case of Martianus Capella, using the first 9,000 words would mean working only with the narrative part of the text, while our intention was to consider the 'behaviour' of both the narrative and the technical parts of this text, if possible separately. For this reason we have taken the last 9,000 words of the text (that is, Book 9 and a tiny scrap of Book 8) and placed the fragment of Book 8 (technical in content) after the text of Book 9 (containing a long narrative introduction followed by the technical main body), so that the first piece of Martianus Capella used in our experiment would contain the narrative beginning of Book 9 and exemplify Martianus' narrative style, while the other two pieces would represent his technical style.

Using this corpus, we reran the algorithm using 3,000-word samples and the 3,000 most frequent words (starting from 50) with personal pronouns removed. The results are shown in fig. 4.

Although the results cannot be said to correspond exactly to our initial hypothesis, it is notable that, on the one hand, the 'New Apuleius' again constitutes a separate branch and, on the other hand, the group *De dogmate Platonis* + *De mundo* + Cicero's *Timaeus* is now detached from the rest of the Apuleian corpus and united with other technical/didactic texts, viz. Hugh of Saint Victor and the technical parts of Martianus Capella. As for the narrative piece of Martianus Capella, it landed in the same branch as Apuleius' *Metamorphoses*, which Martianus obviously imitates in his narrative.[48] Thus, in the way in which the samples by Martianus Capella are grouped, features of genre and imitated style clearly appear to prevail over authorial features, contrary to

[47] It may be pointed out that the *Digesta* is not a text composed by a single author and therefore is not suitable as the reference point. However, in our experiments the fragments of the *Digesta* usually cluster together and are not scattered randomly over different clusters. Investigating the reasons for this behaviour of the *Digesta* in stylometric experiments is a promising avenue for future work, but whatever these reasons are (whether the dominance of particular authors in the *Digesta*, a certain stylistic unification introduced by the sixth-century editors of the compendium, or the inability of the algorithms to distinguish an authorial identity from a generic identity), comparing this collective identity of Roman jurists to the supposed authorial identity of the author of both the authentic Apuleian corpus and the 'New Apuleius' appears not inappropriate for testing the validity of Stover's results.

[48] See, for example, Norden (n. 22), 2.624–5, W.H. Stahl, R. Johnson and E.L. Burge, *Martianus Capella and the Seven Liberal Arts* (New York and London, 1971), 1.30–2, R.H.F. Carver, *The Protean Ass: The* Metamorphoses *of Apuleius from Antiquity to the Renaissance* (Oxford, 2007), 36–9.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

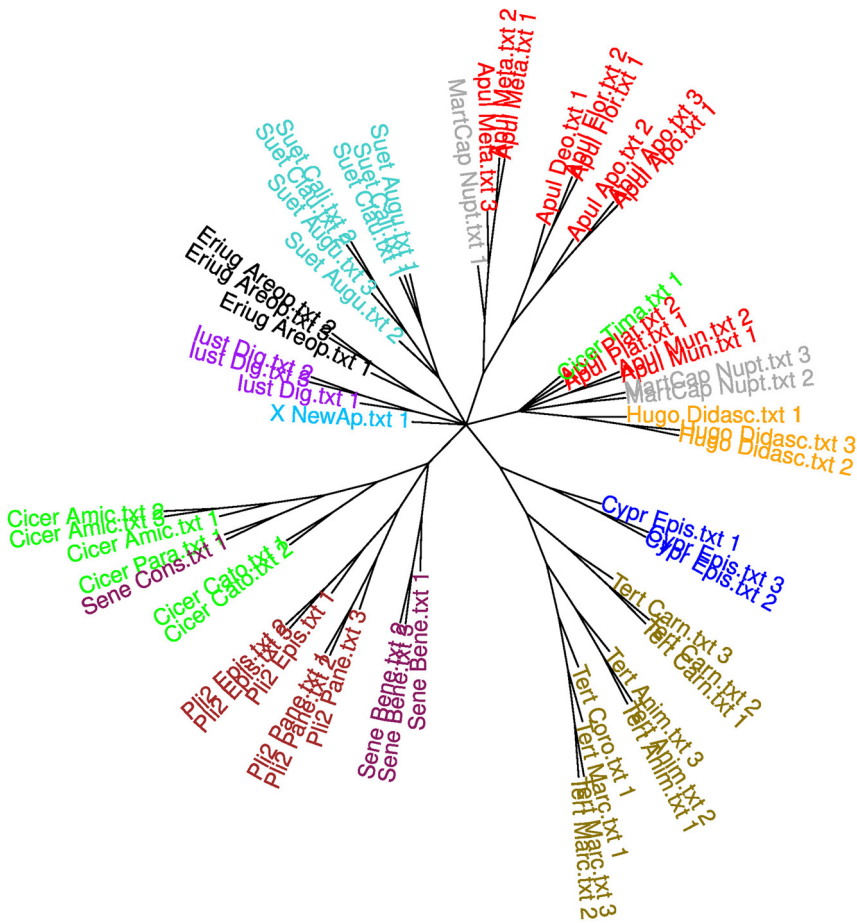

FIG. 4: Bootstrap Consensus Tree for the enlarged corpus. (This figure is in colour in the online version of our article.)

what the proponents of the algorithm proclaim.[49] Similar problems that were already present in fig. 1 are also evident here: in particular, the *Timaeus* is again separated from the rest of Ciceronian texts, and Seneca's *De constantia sapientis* is again detached from the rest of Seneca's works and coupled with other works by Cicero. The algorithm based on most frequent words minus pronouns is clearly not sensitive enough for a reliable identification of a work as belonging to a given author.

Finally, we repeated the same procedure with manual selection of function words, following the general approach adopted in Stover's third experiment but using our list of 150 words described above. As in the previous subsection, we carried out this experiment twice: first with the 3,000-word sample length and then with the

[49] Cf. in particular Stover and Kestemont (n. 8), 656: 'Since this analysis rests on the MFW, most of which are inconspicuous function words, we can rule out the possibility of deliberate skilled imitation.'

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

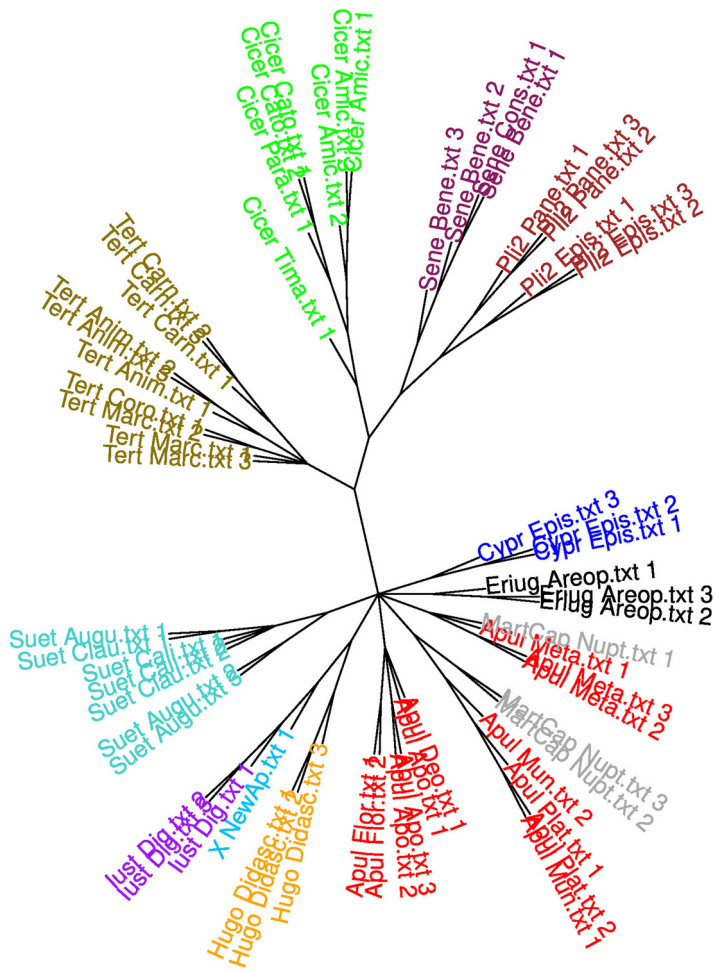

FIG. 5: Bootstrap Consensus Tree for the enlarged corpus with a manual selection of non-content-related words (3,000-word samples). (This figure is in colour in the online version of our article.)

4,500-word sample length (in the latter case, we had to leave out texts that are shorter than 4,500 words and to abandon dividing Martianus Capella into narrative and technical parts). The results are presented on fig. 5 and fig. 6 respectively.

Using manually selected function words again proved to be a superior approach. In figs. 5 and 6, works by Cicero (including the *Timaeus* in the case of fig. 5) and Seneca are reliably grouped together, like all the other works by the same author, except for the notoriously stylistically non-uniform Apuleius (with the threefold division we have already encountered) and, again, Martianus Capella (in fig. 5). Obviously, the narrative part of Martianus' text is too skilled an imitation of Apuleius' *Metamorphoses* even for this version of the algorithm to distinguish the imitation from the model. A tendency also now becomes discernible for the technical parts of Martianus Capella to group consistently with the *De dogmate Platonis* and the *De mundo*: should we perhaps

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

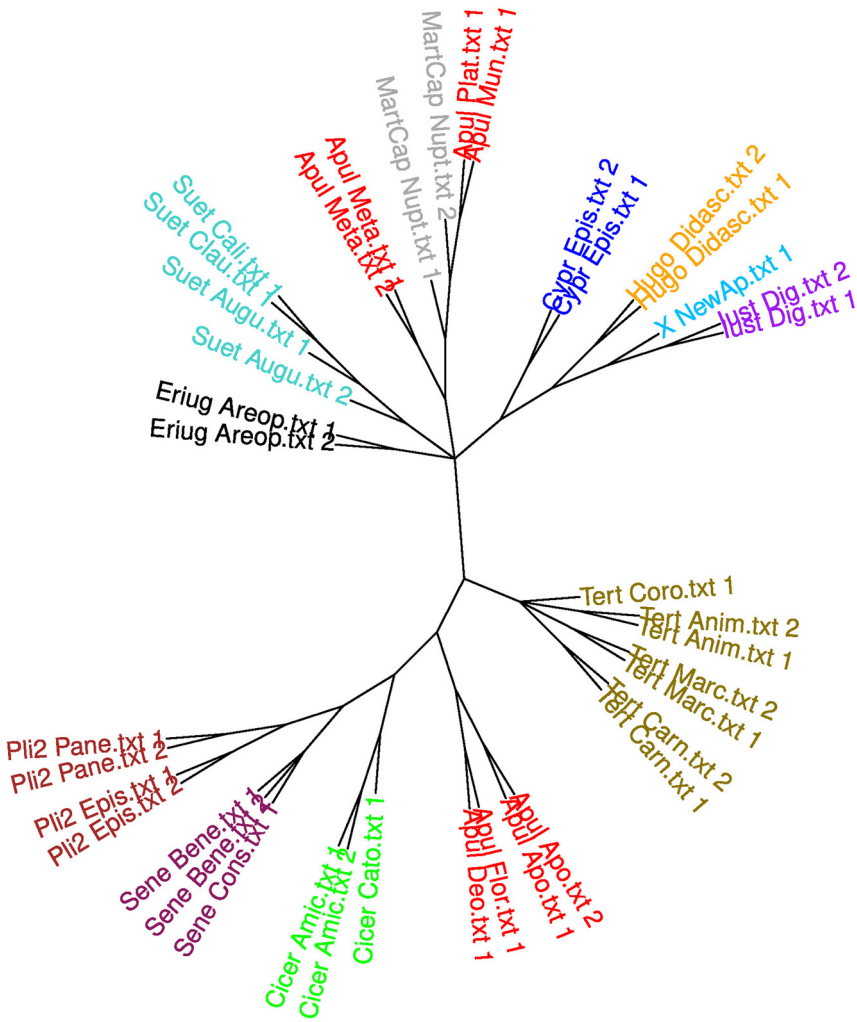

FIG. 6: Bootstrap Consensus Tree for the enlarged corpus with a manual selection of non-content-related words (4,500-word samples). (This figure is in colour in the online version of our article.)

suppose that Martianus in his technical passages tried to imitate in particular what he considered Apuleius' technical works?

As for the 'New Apuleius', the correspondence to our initial hypothesis is striking. In fig. 5 it is placed in the same subtree as the Apuleian corpus, but now it is also accompanied there by the technical/didactic works that we added to Stover's corpus plus Cyprian's letters, which we have seen entering the Apuleius + the 'New Apuleius' cluster in fig. 3, and for some reason also the works by Suetonius. Virtually the same subtree is discernible in fig. 6, though this time the undisputedly Apuleian works other than the *Metamorphoses* are detached from it. The 'New Apuleius' is in both cases particularly closely associated with the *Digesta*.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

Whether or not these results are reliable enough to confirm our hypothesis (this crucially depends on higher validity of function words vis-à-vis raw, most frequent words as the basis for analysis), our experiments probably make it certain that the results published in Stover's works are highly dependent on the particular algorithm settings and on the composition of the corpus, at least with respect to the grouping of the Apuleian corpus with the 'New Apuleius'. It may also be added that some features of our trees, in particular the behaviour of the samples from Martianus, clearly point toward general vulnerability of the approach to skilful imitation, although it is possible that another choice of diagnostic words for computing Burrows's Delta may lead to better and more stable results.

Other methods

In order to provide further corroboration for his hypothesis, in addition to clustering methods, Stover and his co-authors use two other methods: Principal Component Analysis and the so-called Impostor Method. We will briefly survey them in the next two subsections.

Principal Component Analysis

In order to provide another perspective on the stylistic relationship between texts in the corpus, Stover and Kestemont[50] use Principal Component Analysis (PCA), a time-honoured technique for providing lower-dimensional (and therefore more interpretable) approximations of multivariate data. Unlike BCTs, PCA, when applied to textual fragments characterized by word frequencies, is able to tell us which individual words contribute to the division of texts into groups.

However, PCA has its limitations. As the scholars point out, 'PCA has the drawback that, as a visualisation technique, it can only be reliably applied to a relatively small set of authors at the same time (typically three or four).'[51] They further clarify in a footnote: 'Because only so much information can be captured in a two-dimensional analysis, including more than three *œuvres* in a PCA should be generally avoided. The underlying theoretical assumption is that, because of this restriction, each dimension has the potential to contrast one author with the other authors included.'[52] Consequently, Stover and Kestemont compare Apuleius' writings together with the 'New Apuleius' to works by Cicero and Seneca[53] and point out that in these plots the 'New Apuleius' clusters with the works by Apuleius (when we look at the PCA plot for the whole corpus, which we created after replicating Stover and Kestemont's subcorpora plots [cf. fig. 7], we indeed see a lot of confusion, and the position of the 'New Apuleius', referred to as *X_Exp*, is ambiguous).

The scholars also give interpretation of the word frequencies that produce the reported configurations.

These analyses are highly selective and therefore, in our opinion, not really convincing. In the next section, we use a more powerful dimensionality-reduction technique (UMAP),

[50] Stover and Kestemont (n. 8); the argument is referenced in Stover (n. 1), 41–2 n. 31.
[51] Stover and Kestemont (n. 8), 657.
[52] Stover and Kestemont (n. 8), 657 n. 39.
[53] Figs. 14 and 16 in Stover and Kestemont (n. 8), 666, 668.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

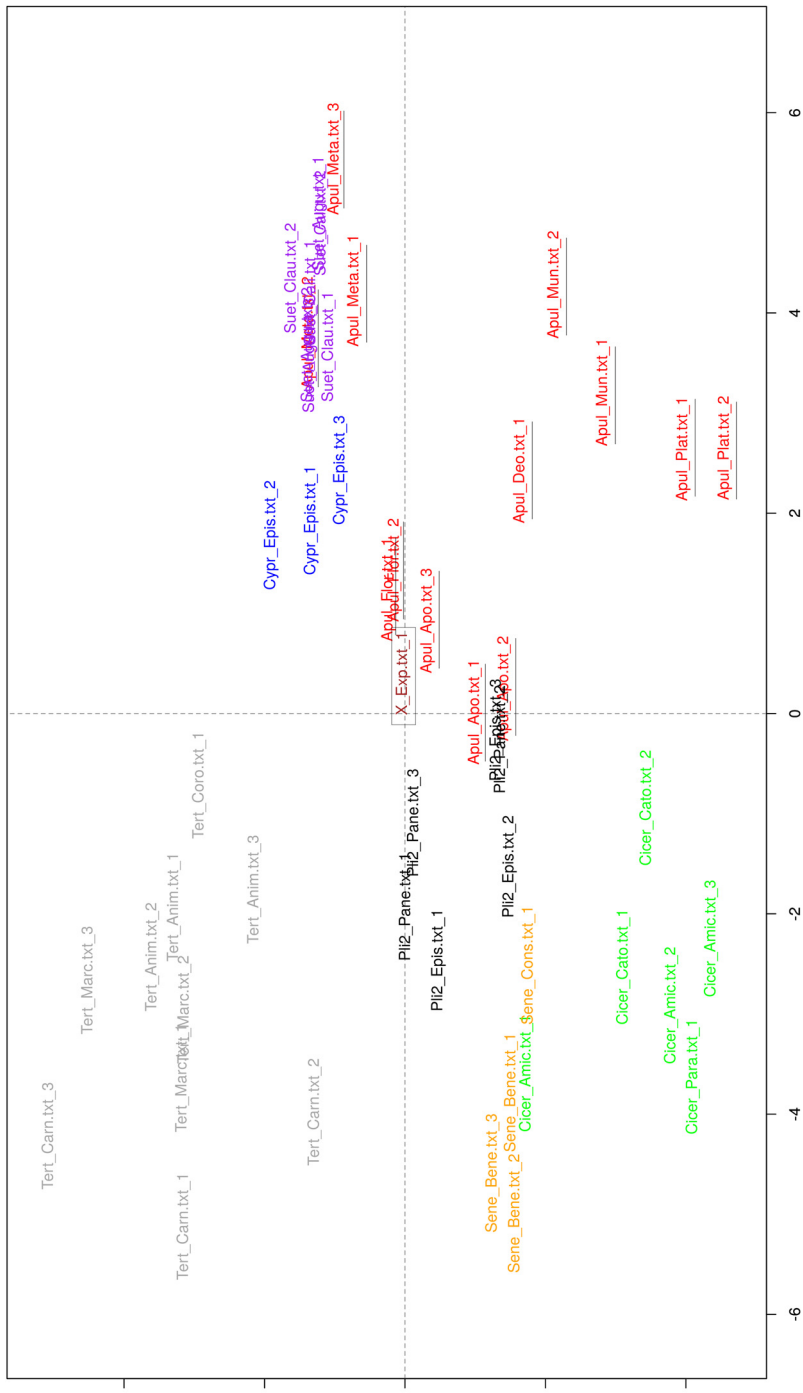

FIG. 7: PCA plot of the reference corpus based on 50 MFWs. (This figure is in colour in the online version of our article.)

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

which provides a very clear and nuanced view of the corpus and does not corroborate Stover's theory.

Impostor Method

The final technique that Stover mentions in the monograph as substantiating his claims about the authorship of the 'New Apuleius' is the Impostor Method originally proposed by Moshe Koppel and Yaron Winter[54] and applied to the Latin corpus in the 2016 brief communication by Stover and his co-authors.[55] The crux of the approach is to replace the traditional stylometric problem of choosing from a closed set (Which from among a given set of authors is the most likely creator of a given textual fragment?) with a binary-decision problem (Were textual fragments X and Y written by the same author?). In order to solve the latter problem, '[w]e systematically produce a set of "impostor" documents and—in a matter reminiscent of a police line-up—ask if X is sufficiently more similar to Y than to any of the generated impostors. The trick is using the proper methods to select the impostors and, more important, to measure document similarity.'[56]

The similarity measure used by the scholars is constructed over counts of 4-grams (sequences of characters of length 4 without spaces or sequences of characters of length less than 4 if they were surrounded by spaces). The scholars note that their method 'works most naturally with a very large and homogeneous feature set';[57] 100,000 most frequent 4-grams were used as diagnostic features in the original paper.

The counts of 4-grams are normalized using the technique called *tf–idf* (for 'term frequency–inverted document frequency'), in which the raw or normalized frequency of a word in a given text is multiplied by the logarithm of 1 divided by the number of documents in the corpus this word appears in. In other words, *tf–idf* is doing nearly the exact opposite of the task performed by the normalization component of Burrows's Delta: rare words or 4-grams found with high frequencies in only several texts are considered extremely informative while subtle fluctuations in the frequencies of common words are nearly completely ironed out by small inverse-document-frequency values.

The differences between documents are then computed based on the individual scores of the 4-grams using the minmax formula (the sum of minimal elements of pairs of corresponding 4-gram scores divided by the sum of maximal elements of those pairs).

In order to make the algorithm more robust, Koppel and Winter also resort to bootstrapping: they 100 times randomly select half of the features from the diagnostic set and then based on this subset find the best match for the analysed fragment from the set of candidate texts (consisting of the purported sibling text and impostors). To make the method symmetrical, given a pair of texts X and Y, they (i) generate a set of impostors for Y and compute the score of Y (that is, how often it is chosen as the best matching candidate for X), (ii) repeat the same procedure with impostors generated for X competing with X itself, and (iii) assign X and Y to the same author if the average

[54] M. Koppel and Y. Winter, 'Determining if two documents are written by the same author', *Journal of the Association for Information Science and Technology* 65 (2014), 178–87.
[55] Stover, Winter, Koppel and Kestemont (n. 8); the argument is referenced in Stover (n. 1), 41–2 n. 31.
[56] Koppel and Winter (n. 54), 178.
[57] Koppel and Winter (n. 54), 179.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

of the scores for X and Y produced in this way is greater than some threshold σ^*, which can be tuned based on a separate development corpus.

Koppel and Winter report two sets of test results. One set of results with accuracy surpassing 90% was achieved on a corpus of posts from blogger.com, where it may be assumed that different authors have different thematic preferences and therefore the performance of the algorithm may be boosted by thematic coherence. In order to control for that, the scholars conducted an additional test based on a corpus of 2,000 pairs of student essays where the second element of the pair was always from a different essay subgenre. Impostors in this case were other texts from the same subgenre but written by different authors, and the algorithm achieved 73.1% accuracy in making authorship attribution decisions, or was wrong *more than once out of four times*, which is a decent result for an exploratory method but is remotely not enough for a kind of forensic authorship testing that is aimed at in the case of the 'New Apuleius'. Koppel and Winter note: 'This reflects the fact that same-author pairs in this corpus differ by design both in terms of subgenre and topic, leaving many fewer common features to exploit.'[58] Frankly, this begs the question whether this method really has any place in stylometric applications, where subgenre and topic are assumed to be not only unavailable as props but chosen in an adversarial fashion.

In Stover and his co-authors, the section 'Verifying the Apuleian authorship of the *Expositio*'[59] reports a result of the application of this method to the 'New Apuleius' with frequencies of 125,000 word unigrams and bigrams (that is, individual words or successive pairs of words) used as diagnostic features and 180 background texts by 36 authors 'writing in similar genres and/or periods as the texts in our corpus'[60] used as impostors. The scholars then report that the pair *Expositio* and *De dogmate Platonis* obtains an 'exceptionally high score' of .73 with only one different-author pair in the test corpus achieving a score above .50 and no different-author pairs achieving a score above or equal to .73. The scholars note further that 'the *Expositio* does not yield high scores when paired with other texts by Apuleius than *De Platone*, most of which are in different genres than the *Expositio* (Platonic philosophy)'.[61] Moreover, the *Metamorphoses* obtains a score above .50 with the *Florida*, but with no other text by Apuleius.

Then Stover and his co-authors suggest that authorship attribution should be conducted in a transitive fashion: if we establish with some confidence that the same author wrote A and B and that the same author wrote B and C, we are forced to conclude that the same author wrote A and C. Transitivity holds, however, only when both links were actually proven. As things stand, neither the *De dogmate Platonis* is proven beyond doubt to be a text by Apuleius, nor the 'New Apuleius' is proven beyond doubt to have been written by the same person who composed the *De dogmate Platonis*: only with a large control corpus of Latin Platonic philosophy contemporary to Apuleius can we even begin seriously discussing this question using this methodology. Therefore, instead of adhering to transitive logical reasoning, we should essentially multiply the uncertainty of both postulated links, which leaves the end result very far from definite.[62]

[58] Koppel and Winter (n. 54), 186.
[59] Stover, Winter, Koppel and Kestemont (n. 8), 240–2.
[60] Stover, Winter, Koppel and Kestemont (n. 8), 241.
[61] Stover, Winter, Koppel and Kestemont (n. 8), 241.
[62] An anonymous *CQ* reader pointed out further that the precision-recall curve in fig. 2 in the paper

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

IV. GENERALIST APPROACHES

In this section, we use generalist methods for multivariate data analysis and classification to analyse Stover's corpus and our augmented reference corpora. We show that these methods are superior to *ad hoc* stylometric algorithms such as BCT and weak linear methods such as PCA in that they provide much more insight into the relationships between textual fragments and are better suited for author identification. These methods do not corroborate Stover's hypothesis.

Contemporary dimensionality reduction: UMAP

Stover and Kestemont's remarks on PCA quoted above reflect a rather pessimistic view of dimensionality reduction. It has been convincingly proven that two dimensions are entirely sufficient to disambiguate any number of clusters (texts written by different authors in this case), given that they form a pattern that can be projected on a 2-D surface; cf. the large-scale topic classification of texts conducted by Hinton and Salakhutdinov.[63] The trick is in recovering these patterns, and here PCA is unfortunately deficient.

The convenient interpretability of PCA is tied to its fundamental limitation: it is able to separate objects into well-defined groups only if this can be done by means of linear combinations of the variables (which means, in this case, summing normalized frequencies of different words multiplied by some coefficients). Real-world data, however, more often than not display non-linear dependencies, and textual data are especially complex in this regard, which makes PCA an inadequate method for tackling them.[64] Stover and Kestemont point to the success of PCA-based stylometric analyses performed by Burrows on a corpus of English literary texts more than 30 years ago,[65] but there is really no excuse for sticking with non-performant crude methods today or at least for not verifying them using more powerful ones.

In order to show that there is no need to impose artificial limitations on the number of authors in the analysis, we used UMAP,[66] a contemporary method for dimensionality reduction, which allows for custom distance measures. The results of its application to frequencies of 100 MFWs with Burrows's Delta used as the distance measure, shown on fig. 8, are remarkably clear. There is no confusion at all between different authors; Apuleian texts in different genres tend to cluster together, and the 'New Apuleius' (again referred to as *X_Exp* following Stover's corpus conventions) is essentially on its own although it is not far from Apuleius' philosophical works.

The UMAP analysis of our extended corpus, shown on fig. 9, again mostly succeeds in correctly separating the known authors, except for a narrative sample from Martianus Capella, placed deep in the area defined by the *Metamorphoses*, and the third sample

is not smooth and that the precision does not go down with the addition of more impostor texts, indicating an insufficient number of data points.

[63] G.E. Hinton and R.R. Salakhutdinov, 'Reducing the dimensionality of data with neural networks', *Science* 313 (2006), 504–7.

[64] Hinton and Salakhutdinov (n. 63), 505–6.

[65] J. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford, 1987).

[66] L. McInnes, J. Healy, J. Melville, 'UMAP: Uniform manifold approximation and projection for dimension reduction', *ArXiv e-prints* 2018 (https://arxiv.org/abs/1802.03426). Cf. https://github.com/jlmelville/uwot for an R package.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

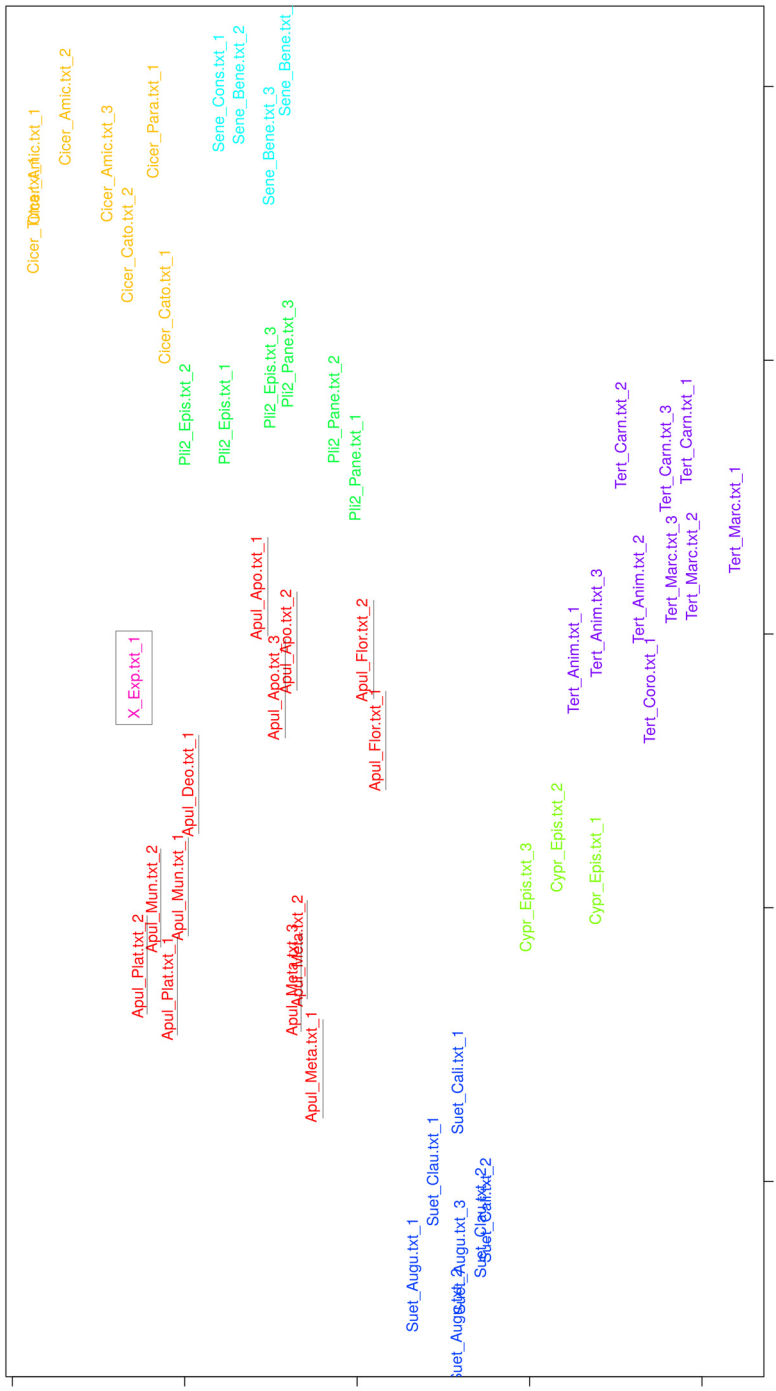

FIG. 8: UMAP plot of the reference corpus based on 100 MFWs and Burrows's Delta. (This figure is in colour in the online version of our article.)

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

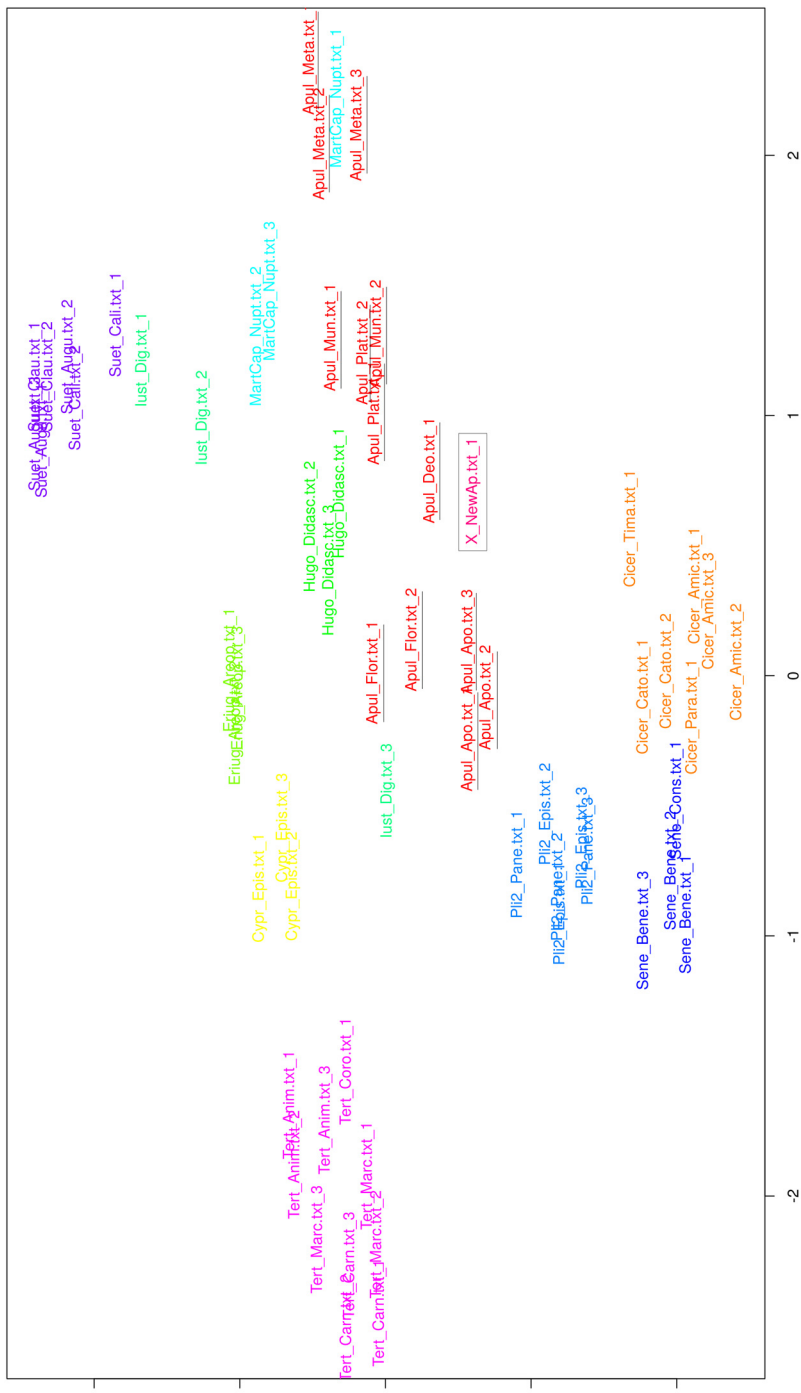

FIG. 9: UMAP plot of the enlarged corpus based on 100 MFWs and Burrows's Delta. (This figure is in colour in the online version of our article.)

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

from the *Digesta* (which is not a text written by a single author). Texts by Apuleius form a diffuse elongated cluster largely parallel to and partly intertwined with the area occupied by the *Digesta* together with the works by Martianus, Hugh of Saint Victor and Eriugena. The 'New Apuleius' is also a part of this pack: it is placed close to some Apuleian texts, but is not more apparently similar to them than the sample from the *Digesta* or works by Hugh of Saint Victor.

This shows that there are absolutely no grounds for artificially reducing the power of dimensionality-reduction analyses in stylometry by using small unrepresentative text samples. It may also be noted that UMAP by and large makes BCTs redundant: using a narrow frequency band of MFWs and Burrows's Delta, it immediately produces highly interpretable and, as far as we can verify them, essentially correct results.

Gradient-boosting classifier

Dimensionality reduction and clustering techniques are well suited for exploratory analysis, but they cannot be set up to provide an answer to a straightforward question: was this fragment written by a given author or not? Even the Impostor Method, which is a binary classifier, is deficient in this regard: it gives a yes/no answer regarding a *pair* of textual fragments. As we showed in the previous section, this can lead to contradictions and demands strong transitivity assumptions. A more logical approach would be to use a binary-classification model that can be fitted on a corpus of textual fragments some of which belong to a given author and then ask this model to assess the probability of a new fragment's being written by the same author.

We propose to use the gradient-boosting machine as such a model. The gradient-boosting machine[67] is probably the most powerful classifier used by the statistical-learning community that is not a neural network or a support-vector machine. Its main advantage over neural networks is that it does not need an enormous training data-set to converge, and this makes it particularly suitable for our task. Support-vector machines do not have such data requirements, but they take time proportional to the cube of the number of observations in order to converge, which makes it impractical to use them with cross-validation (see below) requiring dozens or hundreds of model reruns.

The gradient-boosting machine is an *ensemble classifier*: when learning from the data, it iteratively adds new predictors in such a way that each successive predictor corrects part of the remaining error left after using all of the previous ones. Each new predictor is a decision tree that assigns the data-point to a particular class based on its characteristics.

For example, if we use relative frequencies of words A, B and C as predictors for the authorship of a text, the classifier may first use the relative frequency of A as a defining feature, then correct the intermediate result by using the relative frequency of C, then refine the result even further by using a combination of the frequencies of A and B, etc. The number of predictors and the complexity of interactions (how many variables each predictor can take into account simultaneously) are chosen by the researcher.

It has been shown that in most cases it is advantageous not to use any interactions at all, because otherwise the classifier becomes too sensitive to noise in the training dataset

[67] Y. Freund and R.E. Schapire, 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences* 55 (1997), 119–39.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

and cannot generalize to new data. As for the number of predictors in the ensemble, the standard procedure is to divide the training corpus into a training set proper and a 'held-out' set (also called 'development set') and then check what number of predictors obtains the best classification results on the held-out set after the model was trained on the training set. In order to make the results more robust, each value for the number of predictors can be tested on several different versions of training and held-out sets (this is called *cross-validation*), and the results of these tests are averaged.

In our experiment, we used the implementation of the gradient-boosting machine provided in the R package *gbm*.[68] We used relative frequencies of the same set of function words as in the BCT experiments as features of 3,000-word-long textual fragments. We treated the *De dogmate Platonis* and the *De mundo* as texts written by Apuleius to make our assumptions more accommodating to Stover's hypothesis.

First, we checked if the gradient-boosting classifier is able to tell apart fragments by Apuleius from fragments by other authors in Stover's initial corpus. We adopted the following procedure for each fragment in the corpus:

(1) The fragment in question was held out for testing.
(2) The remaining fragments were repeatedly split into the development set consisting of one fragment and all other fragments (this method is called *leave-one-out cross validation*). These splits were used to determine the optimal number of predictors with the upper bound of 200. The fragments written by Apuleius were assigned the label 1, and all other fragments received the label 0.
(3) The model with the number of predictors chosen using leave-one-out cross validation was trained on all fragments from the training set. It was then asked to give a score from 0 to 1 to the held-out fragment.

The scores provided by the classifier are not informative individually, and the crucial question is whether we can find a threshold such that all fragments by Apuleius will lie above this threshold and all fragments by other authors will lie below this threshold. This will mean that the classifier achieved perfect discrimination. This indeed was nearly the case. The top of the list of the fragments sorted in the decreasing order of their ranks is shown in table 3 (below).

We see that the fragments by Apuleius occupy the top of the list and that the only mistake made by the classifier is the inflated score of the first fragment of Pliny's *Panegyric*, which is slightly higher than the score of the second fragment of the *Metamorphoses*. Further down in the list the scores drop sharply.

We may therefore hypothesize that, after training our model on all fragments with known authorship, given a score provided by the model to a new textual fragment, this fragment may be considered 'Apuleian with high confidence' if the score is above 0.31 and 'possibly Apuleian' if the score is somewhere near 0.3. The score of the fragment consisting of the first 3,000 words of the 'New Apuleius' provided by the model is 0.048, which is more than 6 times less than 0.3.

We repeated the same experiment on the augmented corpus that we used for BCT experiments, again using the relative frequencies of manually selected function words as features of 3,000-word-long textual fragments. The results, presented in table 4 (below), show that, given more data, the classifier can now confidently conclude that

[68] B. Greenwell, B. Boehmke, J. Cunningham, GBM Developers, *gbm: Generalized Boosted Regression Models. R package version 2.1.5*, 2019 (https://CRAN.R-project.org/package=gbm).

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

Table 3: The ranking of fragments from Stover's corpus according to their probability of being authored by Apuleius as estimated by the gradient-boosting algorithm based on relative frequencies of selected function words

Apul_Mun.txt_2	0.99
Apul_Plat.txt_2	0.95
Apul_Mun.txt_1	0.79
Apul_Apo.txt_3	0.79
Apul_Deo.txt_1	0.78
Apul_Apo.txt_2	0.72
Apul_Meta.txt_3	0.71
Apul_Plat.txt_1	0.68
Apul_Flor.txt_1	0.57
Apul_Apo.txt_1	0.47
Apul_Flor.txt_2	0.41
Apul_Meta.txt_1	0.33
Pli2_Pane.txt_1	**0.30**
Apul_Meta.txt_2	0.30
Cicer_Amic.txt_3	0.17
Suet_Clau.txt_2	0.09

Table 4: The ranking of fragments from the augmented corpus according to their probability of being authored by Apuleius as estimated by the gradient-boosting algorithm based on relative frequencies of selected function words

Apul_Meta.txt_3	0.97
Apul_Apo.txt_2	0.95
Apul_Plat.txt_2	0.94
Apul_Mun.txt_2	0.94
Apul_Apo.txt_3	0.92
Apul_Deo.txt_1	0.92
MartCap_Nupt.txt_1	**0.88**
Apul_Plat.txt_1	0.77
Apul_Meta.txt_1	0.68
Apul_Meta.txt_2	0.65
Apul_Apo.txt_1	0.43
Apul_Mun.txt_1	0.40
Apul_Flor.txt_2	0.38
Apul_Flor.txt_1	0.28
Cicer_Amic.txt_3	0.20
<…>	
Pli2_Pane.txt_2	0.03
<…>	
Pli2_Pane.txt_1	0.01

Pliny's *Panegyric* was not written by Apuleius. At the same time, the first fragment by Martianus Capella again proves itself to be indistinguishable from the fragments by Apuleius.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

The threshold of the scores distinguishing 'probably Apuleian' texts from all the rest is now around 0.28. The score of the first 3,000 words of the 'New Apuleius' given by the model trained on all fragments from the augmented corpus is below 0.002.[69]

Testing on a more representative corpus

In order to further test that our results do not depend on the choice of corpus, either by ourselves or by Stover, we added all sufficiently long Latin texts from the PHI5 database and the DigilibLT repository to the corpus we used before and reran the dimensionality-reduction and binary-classification analyses on the resulting, much bigger collection of 298 texts, which we take in their entirety. We did not try to apply BCT to this corpus, as this algorithm has already been shown to be unstable (moreover, trees with thousands of branches are no longer easily interpretable).

The results of applying UMAP to the bigger corpus are shown in fig. 10 (below). In order to make the figure more readable, we replaced names of textual fragments with grey crosses, except for the 'New Apuleius' and the Apuleian fragments. The 'New Apuleius' is firmly embedded in the region occupied by fragments of the *Digesta*. Apuleian fragments do not hang together, with some of his text samples gravitating towards works of Cicero, Petronius, Aulus Gellius and other authors and the samples from the *Metamorphoses* being more independent.

When training a binary classifier, we encountered the problem of overfitting: owing to the huge disparity between the volume of positive and the volume of negative training data, the gradient-boosting algorithm becomes too sensitive to random fluctuations and does not generalize well. Instead, we trained a binary classifier based on the random-forest algorithm,[70] which by design avoids overfitting. Similar to the gradient-boosting machine, the random-forest classifier consists of a set of decision trees with each tree trained on a subset of features. The difference is that, whereas in the gradient-boosting machine trees are fitted sequentially with each one trying to improve the performance by looking at hitherto neglected features (thus potentially leading to overfitting), in the random-forest classifier all trees are independent and the result is produced using a simple average. This still gives rise to strong performance in most cases while not letting a small subset of features have an undue influence.

In the resulting ranking, 24 out of 31 fragments by Apuleius appear above the 'New Apuleius', which occupies the thirty-second position. Non-Apuleian fragments in the top thirty positions include samples written by Columella, Pseudo-Quintilian, Julius Valerius and Zeno of Verona. Thus the random-forest algorithm, given a larger corpus, does not forcefully disprove Apuleius' authorship, but clearly does not support it either.

Based on the results of applying (i) UMAP and the gradient-boosting classifier to Stover's original corpus and to our augmented corpus and (ii) UMAP and the random-forest classifier to a bigger validation corpus, we may conclude that

[69] An anonymous *CQ* reader pointed out that, if we consider possible authors for the 'New Apuleius' to be an open set (not restricted to the authors in the corpus), the decision boundary for the classifier becomes ill-defined. This is true. However, given that stylometry assumes that fragments written by different authors must form coherent clusters, the only real danger here is that of *false positives*: the 'New Apuleius' may be more similar to the Apuleian fragments in the corpus than to fragments written by different authors but at the same time different from them in some way not recovered by the algorithm. All our results are negative.

[70] L. Breiman, 'Random forests', *Machine learning* 45 (2001), 5–32. We used the R implementation described in A. Liaw and M. Wiener, 'Classification and regression by randomForest', *R News* 2 (2002), 18–22.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

FIG. 10: UMAP plot of the big corpus based on 100 MFWs and Burrows's Delta

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press

(1) Powerful general methods for statistical learning and dimensionality reduction outperform *ad hoc* methods and older tools used by Stover and his co-authors: they provide better insight into stylistic differences between textual fragments and in most cases convincingly separate Apuleian fragments from fragments by other authors, given enough training data.
(2) Careful stylistic imitation can defeat even the most powerful stylometric tools.
(3) As far as computer-assisted analyses can be relied upon, there are no indications that the 'New Apuleius' is a text by Apuleius.

V. CONCLUSION[71]

Although our observations do not completely preclude the Apuleian authorship of the 'New Apuleius', they imply that in fact we have no reasons to suppose that he indeed composed it since the only reliable connections of this text with Apuleius appear to be the facts that

(1) the 'New Apuleius' is found in miscellaneous manuscripts containing works by Apuleius,
(2) there is some general similarity in the way in which Apuleius and the author of the 'New Apuleius' use function words, although the overall impression the style of the new text leaves is rather completely un-Apuleian,[72] and
(3) there are textual parallels between the 'New Apuleius' and the *De dogmate Platonis*.[73]

This is obviously not enough for a conclusive attribution. The hypothesis that the text in question is *De dogmate Platonis* Book 3 appears to be especially weak.[74] Nevertheless, Stover is certainly to be thanked for drawing attention to the intriguing problem of stylistic affinities between the new text and the Apuleian corpus.

Stockholm University, Stockholm
DMITRY NIKOLAEV
dnikolaev@fastmail.com

The Russian Presidential Academy of National Economy and Public Administration, Moscow
A.M. Gorky Institute of World Literature of the Russian Academy of Sciences, Moscow
MIKHAIL SHUMILIN
mvlshumilin@gmail.com

[71] The arguments in the last of the sections on stylometry in Stover's book, the one dedicated to prose rhythm (Stover [n. 1], 42–3), are only supposed to support Stover's dating of the text, not his attribution, and anyway cf. O. Zwierlein, 'Augustins quantitierender Klauselrhythmus', *ZPE* 138 (2002), 43–70, at 46–7 for problems connected with the methodology adopted there.

[72] See, for example, Jones (n. 2): 'Granted that Apuleius' style varies, still no one could mistake the style of the *Expositio* for that of the Apuleius; it has no style at all.' Nevertheless, a possibility remains that the author attempted to imitate Apuleius' style in certain aspects, as an anonymous *CQ* reader points out.

[73] Stover (n. 1), 31–4.

[74] In addition to what we say above, see in particular Jones (n. 2), Magnaldi (n. 1), 375–6, Dillon (n. 2), 192.

https://doi.org/10.1017/S0009838821000987 Published online by Cambridge University Press