The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1236–1252
https://doi.org/10.1214/21-AOAS1544

HIGHER CRITICISM FOR DISCRIMINATING WORD-FREQUENCY TABLES AND AUTHORSHIP ATTRIBUTION

BY ALON KIPNIS[a]

Department of Statistics, Stanford University, [a]*kipnisal@stanford.edu*

We adapt the higher criticism (HC) goodness-of-fit test to measure the closeness between word-frequency tables. We apply this measure to authorship attribution challenges, where the goal is to identify the author of a document using other documents whose authorship is known. The method is simple yet performs well without handcrafting and tuning, reporting accuracy at the state-of-the-art level in various current challenges. As an inherent side effect, the HC calculation identifies a subset of discriminating words. In practice, the identified words have low variance across documents belonging to a corpus of homogeneous authorship. We conclude that in comparing the similarity of a new document and a corpus of a single author, HC is mostly affected by words characteristic of the author and is relatively unaffected by topic structure.

1. Introduction. The unprecedented abundance and availability of text data in our age generate many *authorship attribution problems* of the following form. We obtain a new document of unknown authorship; we would like to determine its author. We also have data: several corpora of documents, each of homogeneous authorship. We believe the unknown author of the new document is represented among our corpora, and we wish to attribute authorship to the new document based on our data. Existing approaches for such problems usually construct a set of handcrafted features to discriminate between potential candidate authors (Glickman, Brown and Song (2019), Holmes (1985), Juola (2008), Mosteller and Wallace (1963), Thisted and Efron (1987), Tilahun, Feuerverger and Gervers (2012), Zheng et al. (2006)). Typically, these features originate from linguistic heuristics, such as rate of use of certain words and length of sentences, and are often first constructed by trial and error or based on domain expertise or historical tradition.

While this process sometimes achieves convincing and widely accepted results, it is not automatic. The discriminating features and test statistics are crafted for each specific problem, and it is unclear whether these features or tuned parameters can be reused in other problem domains. A famous example that demonstrates these limitations is Mosteller and Wallace's work on authorship in the Federalist Papers (Mosteller and Wallace (1963)), a collection of articles explaining the nascent U.S. constitution—written between October 1787 and September 1788 by Alexander Hamilton, James Madison, and John Jay. All articles were published under a single pseudonym, regardless of actual authorship. The identities of the three authors as well as the specifics of who wrote each article were revealed or claimed in subsequent years. Among the first 77 articles, historical sources agree that Jay wrote five articles, Hamilton wrote 43, Madison wrote 14, three articles were written jointly by all three, while the authorship of the remaining 12 is disputed between Hamilton and Madison. Mosteller and Wallace determined that all 12 disputed papers are the sole work of Madison. Their process involves two major steps:

Received October 2020; revised August 2021.
Key words and phrases. Higher criticism, two-sample testing, nonparametric methods, unsupervised learning, feature selection, authorship attribution.

(i) Identifying discriminating words, that is, words whose frequencies in known Hamilton texts are different from those of Madison's.

(ii) Combining frequencies of these words in articles of known authorship and disputed ones to a single test statistic.

The specifics of these steps are described in Mosteller and Wallace (1963) and Mosteller and Wallace (1984). In a nutshell, step (i) relied on linguistic assumptions for considering an initial list of 176 "noncontextual" words and some selection procedure to reduce this list. Step (ii) involved various Bayesian modeling decisions as well as some heuristics for estimating the parameters of these models. In particular, it appears that the overall procedure obtained from (i) and (ii) cannot be applied to other authorship challenges without significant modifications. Indeed, we are unaware of other authorship studies that have applied word elimination processes or modeling choices akin to (Mosteller and Wallace (1963)).

In this paper we describe a technique of authorship attribution that can be used "out-of-the-box." When applied to standard authorship challenges, it performs about as well as other approaches but without handcrafting and tuning.

Our technique relies on a relatively simple statistical tool: it uses the Donoho–Jin–Tukey higher criticism (HC) statistic as a measure of closeness between word-frequency tables (viz. bag-of-words) (Donoho and Jin (2004)). We select the likely author using proximity under this measure. The resulting procedure is automatic, in the sense that it does not require prior screening for discriminating words or features. In fact, it inherently identifies a set of likely discriminating features during the calculation of the HC statistic. As we show below, the set thus identified often corresponds to words whose counts exhibit low variance across documents within a corpus of homogeneous authorship. Consequently, for comparing a new document with the corpus of a known author, this proximity measure seems most affected by the words characteristic of that author and is relatively unaffected by the topic structure of the text.

The basic tool we develop in this paper is a technique to discriminate between two word-frequency tables which might both be sampled from the same source frequencies or else perhaps not. Here, "word frequencies" extend in an obvious fashion to n-gram or frequencies of other features of the text that can be summarized as entries in a frequency table. Aside from the authorship attribution problem, n-gram frequency tables have been proven to be a useful summary of textual data more broadly in information retrieval and linguistics (Manning, Raghavan and Schütze (2010), Roberts, Stewart and Airoldi (2016)). It is straightforward to adapt the approach described here to other text classification problems besides authorship attribution.

1.1. *Discriminating word-frequency tables.* Figure 1 illustrates word frequencies in the first 77 Federalist Papers studied in Mosteller and Wallace (1963), divided into three corpora: the circles and triangles represent the frequencies of words in Hamilton's and Madison's known documents, respectively. The squares represent word frequencies from one of the 12 disputed papers. Our goal is to determine which of the two word-frequency tables of known authors best resembles the word-frequency table of the unknown author's document.

Standard approaches to this problem include two-sample tests for homogeneity of discrete multivariate data, such as power divergence tests (Bishop, Fienberg and Holland (1975), Read and Cressie (2012)). It has long been observed, however, that these tests are not optimal in the high-dimensional setting where the number of entries in the table is large, compared to the size of the sample (Balakrishnan and Wasserman (2018), Hoeffding (1965)), or when the frequencies are imbalanced (Arias-Castro, Candès and Plan (2011)). This high-dimensional setting is the typical situation in word frequency tables representing natural text. Moreover,

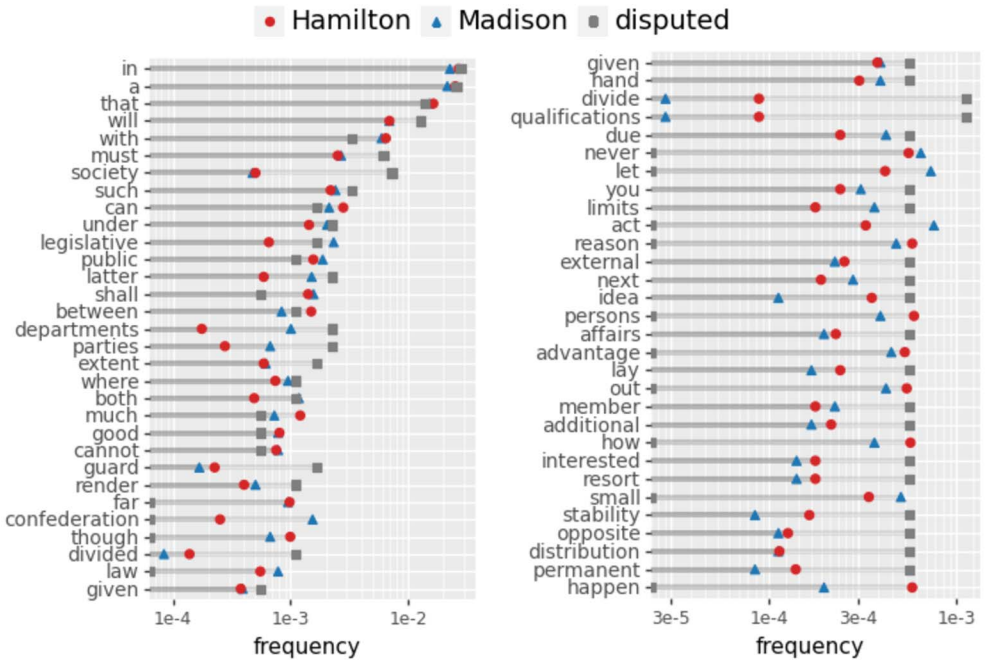

FIG. 1. *Word frequencies of three authorship sources. The two panels show a random sample of* 60 *words out of the* 500 *most common ones in one of the disputed articles* (*gray*), *the corpus of known Hamilton articles* (*blue*), *and the corpus of known Madison articles* (*red*) *out of the first* 77 *Federalist Papers. We attribute the disputed article by measuring the global discrepancy between its word frequencies to each corpus of known authorship.*

the form of alternatives considered in analyzing and developing classical tests of homogeneity is quite general, whereas the important differences in word frequencies between authors may be concentrated on a sparse subset. Namely, relatively few words, out of possibly thousands, may indicate a change of authorship. Consequently, a test that adapts well to sparsity seems promising in this application. In addition to the rareness of discriminating words, the evidence that each such word provides is weak; no single word serves as a decisive discriminating feature. To summarize, we are facing the problem of detecting a *rare* change in the distribution of a large set of possibly *weak* features. HC has long been known to detect signals of a rare/weak nature (Donoho and Jin (2004, 2015), Arias-Castro, Candès and Plan (2011), Mukherjee, Pillai and Lin (2015), Li and Siegmund (2015), Jin and Ke (2016)). This motivates us to adapt HC to our purpose of detecting changes between word frequency tables.

1.2. *Binomial allocation model.* We think about a document as an ordered list of words. Given a vocabulary W, the word-frequency table associated with the document D is denoted by $\{N(w|D), w \in W\}$, where $N(w|D)$ records the total number of occurrences the word w in D.

Consider two documents D_1 and D_2. For each occurrence of a word $w \in W$ in either document, place in a database the labelling pair (w, l) where w denotes the word and l the label "1" or "2," according to which document contains that occurrence. Suppose that, under the null hypothesis, different occurrences are independent and that each is equally likely to originate from "1" (respectively, "2"), only accounting for the relative size of D_1 compared to D_2 minus occurrences of w. Equivalently, occurrences of w are obtained by sprinkling the records in the database with the labels removed across the remaining locations in the large document obtained by concatenating D_1 and D_2. In this case,

$$N(w|D_1) \sim \mathrm{Bin}(n, p),$$

where[1]

$$n = N(w|D_1) + N(w|D_2), \qquad p = \frac{\sum_{w' \in W, w' \neq w} N(w'|D_1)}{\sum_{w' \in W, w' \neq w} (N(w'|D_1) + N(w'|D_2))}. \tag{1}$$

The hypothesis test

$$\begin{cases} H_0: & N(w|D_1) \sim \mathrm{Bin}(n, p), \\ H_1: & N(w|D_1) \sim \text{not } \mathrm{Bin}(n, p) \end{cases} \tag{2}$$

has an exact P-value under the null hypothesis,[2] roughly,

$$\pi(w|D_1, D_2) \equiv \mathrm{Prob}\big(|\mathrm{Bin}(n, p) - np| \geq |N(w|D_1) - np|\big). \tag{3}$$

Applying this test word-by-word, we obtain a large number of P-values $\{\pi(w|D_1, D_2)\}_{w \in W}$. We apply the higher-criticism (HC) statistic to these P-values, obtaining a global test against the null hypothesis that all obey the binomial allocation model outlined above.

1.3. *The higher criticism.* The HC of the P-values $\{p_i\}_{i=1}^N$ is defined as

$$\mathrm{HC}^* \equiv \max_{1 \leq i \leq \gamma_0 N} \sqrt{N} \frac{i/N - p_{(i)}}{\sqrt{\frac{i}{N}(1 - \frac{i}{N})}}, \tag{4}$$

where $p_{(i)}$ is the ith P-value among $\{p_i, i = 1, \ldots, N\}$ and γ_0 is a tunable parameter.[3] The HC test takes a large batch of P-values and returns a single number, indicating the global significance of the body of P-values (Donoho and Jin (2004, 2008)). The idea behind HC goes back to Tukey, who proposed a way to measure the global significance of many level-α independent tests by considering the difference between the standardized z-scores of the observed fraction of tests that are significant to their expected fraction under the joint null. Donoho and Jin proposed to use HC*, the maximized z-scores over the range of significance levels $0 \leq \alpha \leq \gamma_0$, as a global test against the joint null (Donoho and Jin (2004)). Their proposal has shown to be effective in resolving several challenging testing problems (Arias-Castro, Candès and Plan (2011), Arias-Castro and Wang (2015), Cai, Jeng and Jin (2011), Cai, Jin and Low (2007), Delaigle and Hall (2009), Hall and Jin (2010), Ingster, Tsybakov and Verzelen (2010), Jager and Wellner (2007), Mukherjee, Pillai and Lin (2015)).

In our adaptation of the HC test to word-frequency tables, we define the HC-discrepancy $\mathrm{d}_{\mathrm{HC}}(D_1, D_2)$ of documents D_1 and D_2 using the following variant of the HC test statistic:

$$\mathrm{d}_{\mathrm{HC}}(D_1, D_2) \equiv \mathrm{HC}^\dagger \equiv \max_{\substack{1 \leq i \leq \gamma_0 N \\ 1/N \leq \pi_{(i)}}} \sqrt{N} \frac{i/N - \pi_{(i)}}{\sqrt{\frac{i}{N}(1 - \frac{i}{N})}}, \tag{5}$$

where $\pi_{(i)}$ is the ith P-value among $\{\pi(w|D_1, D_2)\}_{w \in W}$ and $N = |W|$ is the size of the vocabulary W (note the symmetry $\mathrm{d}(D_1, D_2) = \mathrm{d}(D_2, D_1)$). The statistic $\mathrm{HC}^\dagger$ is based on a proposal of Donoho and Jin (2004) for improving the numerical stability of HC*. Our experience shows that $\mathrm{HC}^\dagger$ performs slightly better than HC* in authorship challenges; see the results in Table 2.

[1] The binomial model would not be correct unless we omit occurrences of w when considering the relative size of D_1 to define p.

[2] For example, see the `R` function `binom.test`.

[3] HC* and $\mathrm{HC}^\dagger$ appear to be insensitive to the choice of γ_0, provided $|W|$ is large enough. Our experience shows that the choice $\gamma_0 \in (0.2, 0.35)$ provides good results in moderate sample sizes where $|W| > 100$.

The procedure for obtaining the HC-discrepancy of two documents is summarized in Algorithm 1. We extend this procedure to measure the discrepancy between a document and a corpus by thinking about this corpus as the concatenation of all documents within it. It is generally challenging to use the HC-discrepancy to conduct a level-α test against a null hypothesis of the form "the given document and corpus are of the same author," as we consider in (Kipnis (2020)) in the context of the authorship verification challenge of (Kestemont et al. (2020)). Our experience shows that applying HC in this setting requires large amounts of calibration data that is unavailable in most real-world cases. Instead, in this paper we focus on the authorship attribution problem; that is, we associate a document of unknown authorship to one author among several candidates.

Figure 2 illustrates HC-discrepancies of Hamilton's corpus vs. Madison's in the Federalist Papers: each point indicates the HC-discrepancy of one document compared to either corpus. This figure suggests that it is possible to correctly attribute authorship with high accuracy by using HC as an index of discrepancy.

We emphasize that we are not relying on an assumption that the underlying generative model of binomial word allocation is exactly true; there may well be departures such as correlations and overdispersion. Again, HC is here being used as an index of discrepancy.

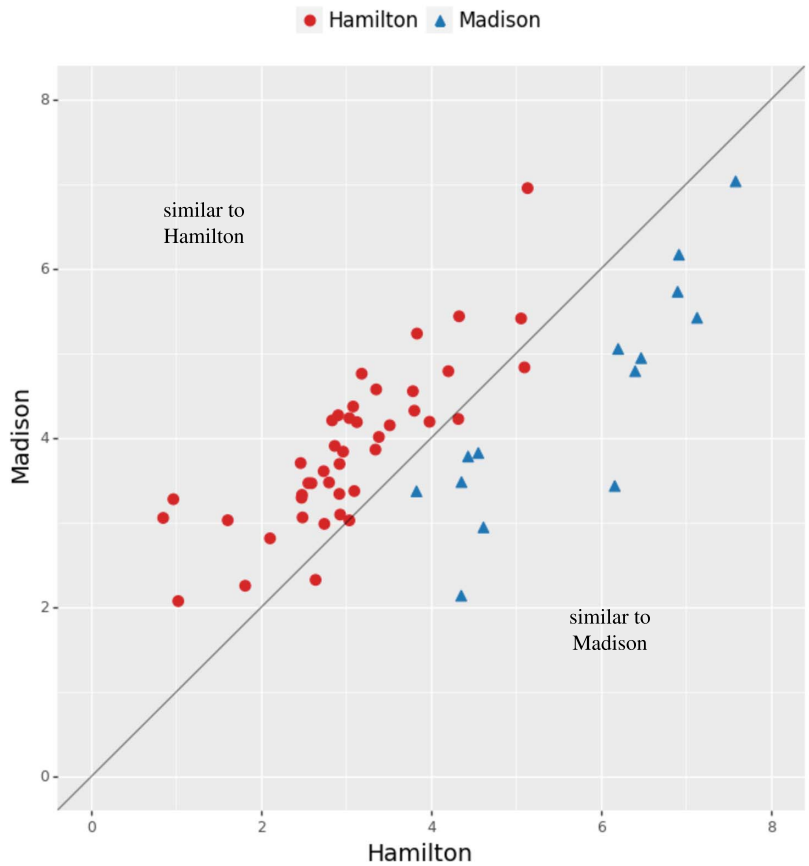

FIG. 2. *Authorship in the Federalist Papers. Each point indicates HC-discrepancy of a document by Madison (red) or Hamilton (blue) with respect to Hamilton's corpus of* 43 *papers (x-axis) and Madison's corpus of* 14 *papers (y-axis) (in comparing a document to the corpus of its own author, this document is left out of the corpus). The diagonal line $y = x$ is indicated. With a few exceptions, Madison's articles lie below the diagonal, while Hamilton's lie above the diagonal.*

Algorithm 1 HC-discrepancy of word-frequency tables

Input: Two word-frequency tables $\{N(w|D_1), w \in W\}$ and $\{N(w|D_2), w \in W\}$.

procedure HC-DISCREPANCY
$n_1 \leftarrow \sum_{w\in W} N(w|D_1)$;
$n_2 \leftarrow \sum_{w\in W} N(w|D_2)$;
for $w \in W$ **do**
$x \leftarrow N(w|D_1)$;
$n_w \leftarrow N(w|D_1) + N(w|D_2)$;
$p_w \leftarrow \frac{n_1 - x}{n_1 + n_2 - n_w}$;
$\pi(w|D_1, D_2) \leftarrow$ binomial test (x, n_w, p_w);
end for
$N \leftarrow |W|$;
$(\pi_{(1)}, \ldots, \pi_{(N)}) \leftarrow \text{Sort}(\{\pi(w|D_1, D_2)\}_{w\in W})$;
$z_i \leftarrow \sqrt{N}\frac{i/N - \pi_{(i)}}{\sqrt{\frac{i}{N}(1-\frac{i}{N})}}$, $i = 1, \ldots, N$;
$i_{\min} \leftarrow \text{argmin}\{i, 1/N \le \pi_{(i)}\}$;
$i^* \leftarrow \text{argmax}\{z_i, i_{\min} \le i \le \gamma_0 N\}$;
$\text{HC}^\dagger \leftarrow z_{i^*}$;
$W^\dagger \leftarrow \{w \in W : \pi(w|D_1, D_2) \le \pi_{(i^*)}\}$;
return $\text{HC}^\dagger$, $W^\dagger$
end procedure

1.4. *The HC threshold.* Associated with the HC statistic (4) is the *HC threshold* t_{HC}, defined as

$$t_{\text{HC}} \equiv \pi_{(i^*)}, \quad i^* \equiv \underset{\substack{1\le i\le \gamma_0 N \\ 1/N \le \pi_{(i)}}}{\text{argmax}} \frac{i/N - \pi_{(i)}}{\sqrt{\frac{i}{N}(1-\frac{i}{N})}}. \tag{6}$$

Roughly speaking, the HC statistic describes the maximal deviation of the collection of P-values $\{\pi(w|D_1, D_2), w \in W\}$ from the uniform distribution over $(0, 1)$. This deviation is mostly affected by P-values that fall below t_{HC}. Donoho and Jin (2009) showed that the HC threshold leads to an optimal feature selection procedure in some classification settings. Jin and Wang (2016) applied HC thresholding for selecting features in a specific clustering setting. In our context, we propose to use the HC threshold to identify words distinguishing between the two word-frequency tables and, consequently, between the author associated with each of these tables.

The procedure outlined above for measuring the discrepancy of two word-frequency tables and obtaining a set of discriminating words is summarized in Algorithm 1.

1.5. *Analyzing ingredients for success.* Textual data serves as a channel to deliver information in multiple contexts. It is, therefore, challenging, or perhaps impossible, to provide a comprehensive theory for the performance of the HC-discrepancy that covers all authorship attribution scenarios. Instead, in order to understand the empirical success of our test in attributing authorship, we analyze the properties of the words that fall below the HC threshold since these are the words affecting the HC-discrepancy most. For this purpose we apply variance-stabilizing transformations to word-counts and compare the variance associated with the same word across a corpus of homogeneous authorship to the P-value associated with this word under a binomial allocation model. By examining a large number of pairs of authors, we discover that words having the most influence on the value of the HC statistic are associated with small variances across documents in each author's corpus. This finding

shows that the HC-discrepancy is not heavily affected by the topic structure of the text. It seems that words contributing to the HC statistic are characteristic of the author's style rather than the characteristic of a particular topic.

1.6. *Related works*. The problem of testing hypotheses based on frequencies or contingency tables dates back, at least, to Pearson (Pearson (1900)), whose chi-square test is still the standard choice in this problem. We refer to the classical book (Bishop, Fienberg and Holland (1975)) as an introduction to the topic. The one-sample version of the problem in which the observed frequencies are replaced by the true underlying frequencies in one of the tables, appears under the names: testing multinomial, goodness-of-fit with categorical data, and, in computer science, distribution identity testing. In accordance with modern challenges in data analysis, there is much recent interest in the high-dimensional version of this problem in which the number of samples is small compared to the size of the vocabulary, the number of categories, or the support of the distribution; see Balakrishnan and Wasserman (2019) and the related review paper (Balakrishnan and Wasserman (2018)). Furthermore, the work of Donoho and Kipnis (2020) studies the asymptotic properties of HC under rare and weak perturbations of the categories. We note that choices for combining the binomial allocation P-values other than (5) may be preferable in some cases; see (Li and Siegmund (2015)) for a discussion. Our choice of HC here is largely motivated by its well-understood feature selection mechanism using the HC threshold (Donoho and Jin (2009)).

Authorship studies in the statistical literature include, most notably, the case of the Federalist Papers (Mosteller and Wallace (1963), Mosteller and Wallace (1984)). The surveys by Holmes (1985) and Juola (2008) provide wide coverage of the topic. Another line of statistical works concerning authorship first identifies some regularity property of the text and then uses deviations from the regular behavior to attribute or refute authorship (Cox and Brandwood (1959), Sichel (1974), Wake (1957)). This practice was also adopted by Efron and Thisted (Efron and Thisted (1976), Thisted and Efron (1987)), who applied their estimator of the number of unseen species to determine if the number of novel words in a disputed text matches the degree of novelty had Shakespeare been the author of the text. Ross (2020) addressed the possibility that the style of an author changes over time and suggested ways to account for this change in authorship studies. Tilahun, Feuerverger and Gervers (2012) tracked changes in word-frequencies over time to date medieval charters. Very recently, Glickman, Brown and Song (2019) considered harmonic and melodic features to determine the degree of collaboration in a few famous songs by The Beatles.

1.7. *Structure of the paper.* The paper is organized as follows: In Section 2 we develop a method for attributing the authorship based on the procedure outlined in Algorithm 1 and applying it in various authorship attribution challenges. In Section 3 we explain why our method works. Concluding remarks are provided in Section 4.

2. Authorship attribution. In this section we develop a method for text classification and authorship attribution using d_{HC} and evaluate its performance in several authorship attribution challenges.

Given a document D and a corpus $\mathcal{C}$ not containing D, the HC-discrepancy $\mathrm{d}_{\mathrm{HC}}(D,\mathcal{C})$, associated with D and $\mathcal{C}$, provides an index of discrepancy between the document and the corpus. We suggest using this index of discrepancy to solve the following classification problem: Let $\mathcal{C}_A = \{D_i, i \in I_A\}$ and $\mathcal{C}_B = \{D_i, i \in I_B\}$ be two disjoint corpora. Upon introducing a new document D that is neither a member of $\mathcal{C}_A$ nor $\mathcal{C}_B$, associate D with one of corpus A or B.

Henceforth, we identify a corpus $\mathcal{C}$ with the document formed by concatenating all documents in $\mathcal{C}$. We also use the notation

$$\mathcal{C}_{(D)} \equiv \{D' \in \mathcal{C}, D' \neq D\}$$

to denote the corpus $\mathcal{C}$ with the document D removed.

2.1. *Discrepancy between a document and a corpus.* Considering a corpus as one large document, we can naturally extend the HC-discrepancy between document-pair to discrepancy between a document D a corpus $\mathcal{C}$. In this case we set

$$\mathrm{HC}_{D|\mathcal{C}} \equiv \mathrm{d}_{\mathrm{HC}}(D, \mathcal{C}).$$

Figure 2 depicts an $x - y$ scatter plot in which each point represents a document in the combined set $\mathcal{C}_{\mathrm{Hamilton}} \cup \mathcal{C}_{\mathrm{Madison}}$. For $D \in \mathcal{C}_{\mathrm{Hamilton}}$,

$$(x, y) = (\mathrm{HC}_{D|\mathcal{C}_{\mathrm{Hamilton}}(D)}, \mathrm{HC}_{D|\mathcal{C}_{\mathrm{Madison}}}),$$

while for $D \in \mathcal{C}_{\mathrm{Madison}}$,

$$(x, y) = (\mathrm{HC}_{D|\mathcal{C}_{\mathrm{Hamilton}}}, \mathrm{HC}_{D|\mathcal{C}_{\mathrm{Madison}}(D)}).$$

It follows from Figure 2 that the HC-discrepancy between $D \in \mathcal{C}$ and $\mathcal{C}(D)$ is small compared to the HC-discrepancy between D and the corpus of the other author. Since points corresponding to documents of opposing authorship are largely separated by the identity line $(y = x)$, HC-discrepancy can determine the true author of a new document with high accuracy.

2.2. *Rank-based calibration in authorship attribution.* Due to the complicated structure of most texts, we do not expect the binomial model underlying our method to be strictly correct, nor do we expect the identity line to be the best discriminator between the two corpora. Instead, we deploy the HC statistic using a rank-based calibration. Consider the rank of the HC-discrepancy of a new document relative to the HC-discrepancy obtained from other documents within a corpus. This rank furnishes a *calibrated* index of discrepancy between the disputed document and the corpus. We assign the document D to whichever corpus gives the smallest normalized rank. We formalize this process using a rank-based testing procedure (Lehmann (1975)): For each corpus $\mathcal{C}_\alpha = \{D_j, j \in I_\alpha\}$ and document $D_i, i \notin I_\alpha$, consider the extended corpus $\mathcal{C}_{\alpha+i} \equiv \{D_j, j \in I_\alpha \cup \{i\}\}$. The null hypothesis $H_{0,\alpha+i}$ states that all scores

$$\mathrm{HC}_{\mathcal{C}_{\alpha+i}} \equiv \{\mathrm{HC}_{D_j|\mathcal{C}_\alpha(D_j)}\}_{j \in I_\alpha \cup \{i\}}, \qquad i \notin I_\alpha$$

are sampled independently from the same continuous distribution over the reals. A P-value with respect to $H_{0,\alpha+i}$ is $1 - \hat{r}_{D_i|\mathcal{C}_\alpha}$, where

$$\hat{r}_{D_i|\mathcal{C}_\alpha} \equiv \frac{\mathrm{rank}(\mathrm{HC}_{D_i|\mathcal{C}_\alpha} \mid \mathrm{HC}_{\mathcal{C}_{\alpha+i}})}{|I_\alpha| + 1} \tag{7}$$

is the rank of $\mathrm{HC}_{D_i|\mathcal{C}_\alpha}$ in the sample $\mathrm{HC}_{\mathcal{C}_{\alpha+i}}$. We consider large values of $\hat{r}_{D_i|\mathcal{C}_\alpha}$ to be evidence against the hypothesis that D_i and the other documents in $\mathcal{C}_\alpha$ were written by the same author. Consequently, we associate the document D_i to whichever corpus has a smaller $\hat{r}_{D_i|\mathcal{C}_\alpha}$.

2.3. *Performance in authorship attribution.* We consider the performance of our HC-based method in the authorship attributing challenges listed in Table 1. The code and data for obtaining the results described in this section are available in the Supplementary Material (Kipnis (2022)).

TABLE 1
Three authorship attribution challenges

collection	# authors	# documents per author	# words per doc (range)	# words per doc (average)
The Federalist	2	Hamilton 53 Madison 14	958–3.5k	2k
11,050 literary works	488	10–100 (average 23)	10k–2600k	74k
PAN2018 authorship attribution challenge (problems 1–4)	5, 10, 15, 20 (per problem)	7 (in training set)	600–1k	970

2.3.1. *The federalist papers*. Figure 3 illustrates the HC-discrepancy of each of the 12 disputed Federalist papers (Numbers 49–58, 62, and 63) with respect to Madison's and Hamilton's corpus, respectively.

Also shown in this figure are the normalized ranks $\hat{p}_{D_i|\mathcal{C}_\alpha}$ of the ith disputed paper for $\alpha \in \{\text{Hamilton}, \text{Madison}\}$. Based on our rank-based calibration, each of the disputed documents seems to be written by Madison rather than Hamilton. In testing all of Hamilton's corpus against Madison's, the binomial allocation P-values of 378 words fall below the HC threshold. The bottom of Figure 3 indicates the P-values of 63 of these words.

2.3.2. *Large collection of potential authors*. We now assess the performance of our HC-based method in determining authorship from among many authors. We use the Gutenberg Project[4] to form a collection of texts by 488 authors satisfying our inclusion criterion: At least 10 works with at least 10,000 words. We use a vocabulary consisting of the N most common words in English Google books, according to the list (Norvig (2013)), for $N \in \{250, 1000, 3000\}$. For each vocabulary size, we measure the discrepancy of each work and each of the 488 corpora associated with each author in a 10-fold cross-validation procedure: the entire dataset containing 11,050 works is randomly split into 10 disjoint subsets. For $i = 1, \ldots, 10$, all subsets, except subset i, are used as the training set, and accuracy is evaluated for attributing authorship of works in subset i. The reported average accuracy and standard error are over all 10 cases. We attribute the work to the author whose corpus attained the smallest HC-discrepancy. In this evaluation we also consider the HC variant HC^* of (4) in addition to $\text{HC}^\dagger$ of (5). The average accuracy in this procedure is reported in Table 2. We also used an analogous attribution procedure based on several other discrepancy measures:

- *Cosine discrepancy*. The cosine discrepancy between documents D_1 and D_2 is defined as

$$\mathrm{d}_{\cos}(D_1, D_2) \equiv 1 - \cos(D_1, D_2),$$
$$\cos(D_1, D_2) \equiv \frac{\sum_{w \in W} N(w|D_1) N(w|D_2)}{\sqrt{\sum_{w \in W} (N(w|D_1))^2} \sqrt{\sum_{w \in W} (N(w|D_2))^2}}.$$

  We also considered a nearest neighbor (NN) classifier with $\mathrm{d}_{\cos}(D_1, D_2)$ as its underlying metric.

[4]Project Gutenberg (n.d.), retrieved September 10, 2019, from www.gutenberg.org

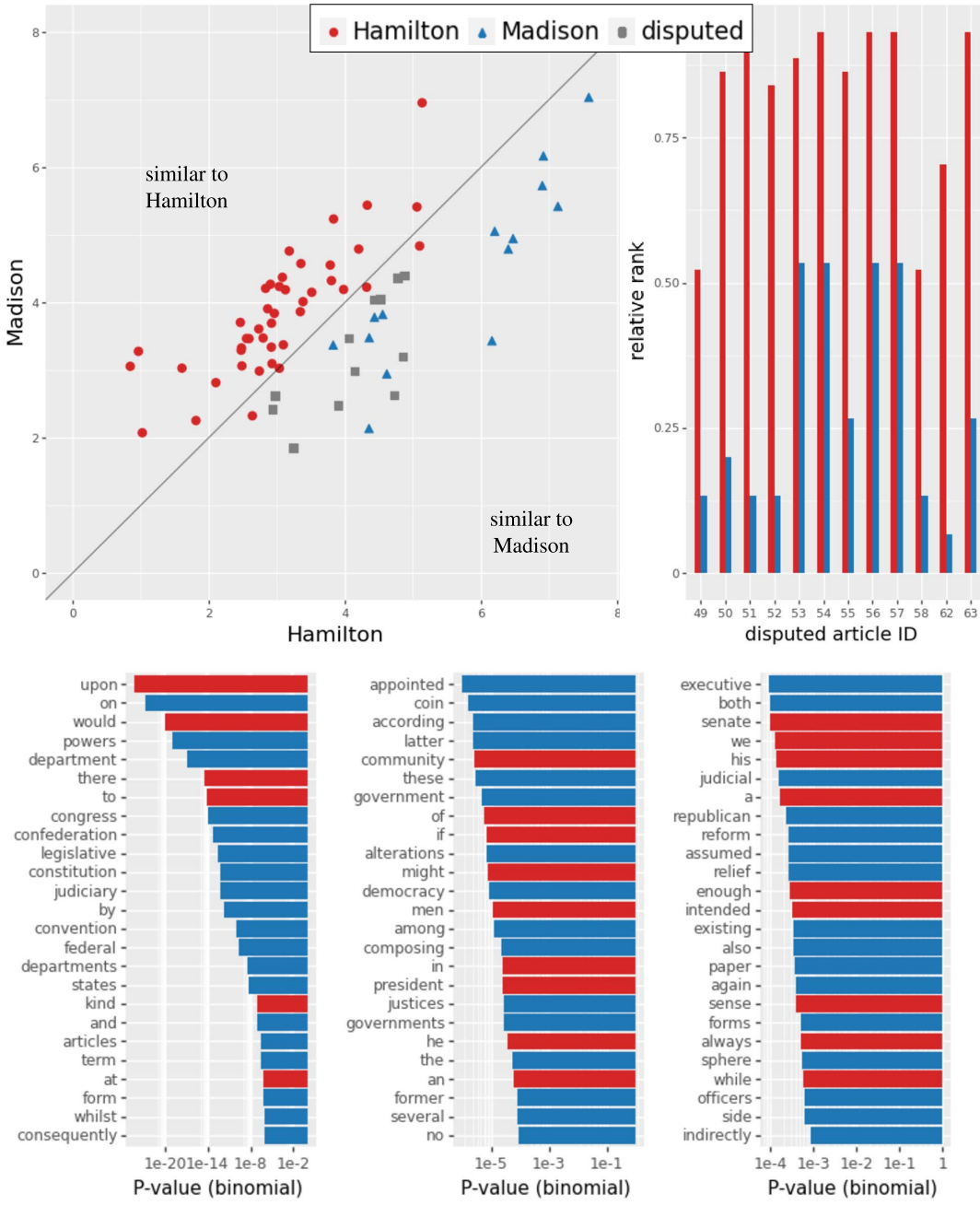

FIG. 3. *Authorship in the Federalist Papers. Top left*: *HC-discrepancies of each paper with respect to Hamilton's corpus of* 43 *papers* (x*-axis*) *and Madison's corpus of* 14 *papers* (y*-axis*). *The diagonal line* $y = x$ *is indicated. This scatter plot is an extended version of Figure* 2 *with additional points corresponding to the* 14 *disputed papers. Top right*: *The relative rank of each disputed article with respect to each corpus* (*lower rank indicates better similarity*). *Bottom*: *Words and their associate P-value in the set* $W^{\dagger}$ *returned by* HC − DISCREPANCY *when the entire Hamilton's corpus is compared to the entire Madison's corpus. The P-value for each of these words obtained from the test* (2) *falls below the HC threshold. The color of the bar signals the corpus in which the word is more frequent*: *Red* = *Hamilton*; *Blue* = *Madison. The vocabulary is the union of the set of* 1500 *most common words by each of Hamilton and Madison in the Federalist collection with proper names and cardinal numbers removed.*

TABLE 2

Accuracy in determining the authorship of 11,050 *literary works among* 488 *writers using several statistics (interpreted as indices of discrepancy) and vocabulary sizes. Each k-size dictionary consists of the k most frequent English words, according to the list in Norvig (2013). Reported accuracy is obtained using a* 10*-fold cross validation procedure; standard errors are in brackets*

vocabulary size	250	1000	3000
$HC^{\dagger}$	**0.791** (0.0092)	**0.833** (0.0093)	**0.851** (0.0097)
HC^{*}	**0.787** (0.0063)	**0.831** (0.0074)	0.840 (0.0103)
$d_{\lambda 0}$ (G^2)	0.724 (0.0158)	0.794 (0.0010)	0.829 (0.0121)
$d_{\lambda 2/3}$ (Cressie-Reed)	0.723 (0.0115)	0.774 (0.0088)	0.780 (0.0179)
$d_{\lambda 2}$ (Pearson's χ^2)	0.718 (0.0142)	0.747 (0.0130)	0.717 (0.0117)
cosine discrepancy	0.529 (0.00096)	0.553 (0.0145)	0.571 (0.0120)
5-nearest neighbors cosine discrepancy	0.697 (0.0147)	0.712 (0.0084)	0.722 (0.0122)

- *Power divergence*. The power-divergence test statistic with a real parameter λ is defined as (Read and Cressie (2012))

$$\mathrm{d}_{\lambda}(D_1, D_2) \equiv \sum_{\substack{w \in W' \\ i \in \{1,2\}}} N(w|D_i)\left(\left(\frac{N(w|D_i)}{\mathrm{T}(w|D_1, D_2)}\right)^{\lambda} - 1\right).$$

Here, W' is the set of words in W such that $N(w|D_1) + N(w|D_2) > 0$, and

$$\mathrm{T}(w|D_1, D_2) \equiv \alpha_1 N(w|D_1) + (1 - \alpha_1) N(w|D_2),$$

where

$$\alpha_1 \equiv \frac{\sum_{w \in W} N(w|D_2)}{\sum_{w \in W} N(w|D_1) + N(w|D_2)}.$$

We considered the cases $\lambda = 1$ (Pearson's chi-squared test statistic), $\lambda = 2/3$ suggested in Cressie and Read (1984), and $\lambda \to 0$ corresponding to the likelihood ratio statistic G^2. The index of discrepancy corresponding to the power-divergence test statistics is $\mathrm{d}_{\lambda}/(N' - 1)$, where N' is the number of words in W such that $N(w|N_1) + N(w|D_2) > 0$.

The average accuracy and standard error of each authorship attribution method in the Gutenberg authorship challenge are provided in Table 2. It follows that HC-discrepancy attains the best accuracy among all the discrepancy methods we tried. Welch's t-test implies that all differences between the accuracies of HC-discrepancy (based on either HC^{*} or $HC^{\dagger}$) and the other methods are *significant* at the level 0.05. The difference between $HC^{\dagger}$ and HC^{*} is significant only for vocabulary size $= 3000$.

2.3.3. *Authorship attribution challenge*. We evaluated the performance of our technique on the English-language part of the cross-domain authorship attribution challenge (Kestemont et al. (2018)). This challenge involves four independent authorship attribution problems with k candidate authors for $k \in \{5, 10, 15, 20\}$. For each author in each problem, a corpus containing seven different labeled documents is provided. Each problem is also provided with a set of unlabeled documents. The goal is to correctly attribute the authorship of each document in the test set to one of the k candidate authors in each problem.

We used our HC-based approach to solve each problem by attributing each document from the test set to whichever author has the smallest index of discrepancy between this document and the corpus of that author in the training set. The vocabulary W was formed

for each specific problem using 3000 of the most common words, word bigrams, and word trigrams over all documents in the training set. Before counting word n-grams, we removed proper names and cardinal numbers and converted words to their dictionary form using the lemmatizer described in Qi et al. (2018). For the problems with 5, 10, 15, and 20 authors, our technique attained accuracies of 0.75, 0.775, 0.5, and 0.43, respectively. These accuracies corresponding to an average F1 score of 0.75 — the second-best score reported for this part of the challenge according to Kestemont et al. (2018).

3. Analyzing success in authorship attribution. In this section we suggest an explanation for the observed success of the HC statistics as an authorship discriminator. We consider word counts under a variance-stabilizing transformation and analyze the variance of the transformed counts across documents within a corpus of homogeneous authorship. We observe that the HC-discrepancy between a document and the corpus of an author is mostly affected by words characteristic of the author and not by words characteristic of topics in the text.

3.1. *Author-characteristic words.* We propose that a word truly *characteristic* of an author would be used consistently across documents by that author. In contrast, a *topic-related* word will occur very frequently in documents associated with that topic but not frequently in documents associated with unrelated topics. A simple model articulating this distinction says that words characteristic of an author are sampled independently from a multinomial distribution that is fixed across the corpus, whereas topic-related words are sampled via more structured mechanisms (Blei and Lafferty (2007), Chang and Blei (2010), Deng, Geng and Liu (2014), Griffiths et al. (2004), Roberts, Stewart and Airoldi (2016), Ross (2020)). Therefore, if $\mu(w)$ is the underlying frequency of the characteristic word w in this multinomial distribution, the count of w in the document D is modeled by a Poisson distribution with parameter $\lambda(w|D) = \mu(w)|D|$, where $|D|$ denotes the total number of words in D. In contrast, the count of a topic-related word may follow a Poisson mixture distribution or may even be affected by stochastic dependence structures among words associated with the same topic (Blei and Lafferty (2007)). Such effects increase the variance of counts of topic-related words across documents within a corpus, resulting in overdispersion with respect to the Poisson sampling model (Breslow (1984), McCullagh and Nelder (1989, Chapter 6.2.3)). Admittedly, the Poisson sampling model for words that are not topic-related does not match observed word counts well (Mosteller and Wallace (1984), Church and Gale (1995)). Nevertheless, a prediction of the model—relative variance as a measure for topic-relatedness—seems to hold in the cases we examined.

3.2. *Variance-stabilizing transformation.* We transform word counts using the transformation

$$r(w|D) \equiv 2\sqrt{\frac{N(w|D) + \frac{1}{4}}{|D|}}. \tag{8}$$

This version of the variance-stabilizing transformation is based on a suggestion in Brown, Zhang and Zhao (2001). If $N(w|D)$ follows a Poisson distribution with parameter $\mu(w)|D|$ where $|D| \gg \mu(w)$, then the distribution of $r(w|D)$ is approximately normal with mean $2\sqrt{\mu(w)}$ and variance $1/|D|$. By considering documents of roughly equal number of words within the same corpus, we assume that this variance is constant across documents belonging to a corpus. Overdispersion, with respect to the Poisson sampling model of a word w, implies that the variance of $r(w|D)$ across a corpus is larger than the naively-expected variance $1/|D|$. Therefore, we think of this variance as a measure of topic-relatedness of w. According

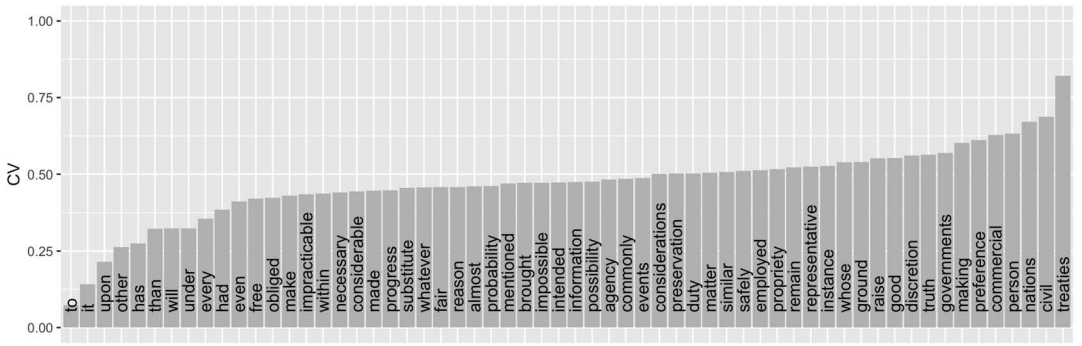

FIG. 4. *Examples of coefficient of variation* $\mathrm{CV}(w|\mathcal{C})$ *for selected words within a single corpus of homogeneous authorship* $\mathcal{C}$. *We assume that words characteristic of the author mostly appear at the left-hand side of this plot. Proper names and cardinal numbers were removed.*

to standard linear discriminant analysis principles, we consider a coefficient of variation (CV) based on the ratio of the squared root of this variance to the mean of $r(w|D)$. Specifically, define the sample mean and variance across documents within a corpus $\mathcal{C}$ as

$$\mu(w|\mathcal{C}) \equiv \mathrm{Ave}\big(\{r(w|D)\}_{D\in\mathcal{C}}\big),$$
$$\sigma^2(w|\mathcal{C}) \equiv \mathrm{Var}\big(\{r(w|D)\}_{D\in\mathcal{C}}\big),$$

respectively. We use the CV, defined as

$$\mathrm{CV}(w|\mathcal{C}) \equiv \frac{\sigma(w|\mathcal{C})}{\mu(w|\mathcal{C})},$$

as a measure for the variability of the word w within the corpus $\mathcal{C}$. Figure 4 illustrates examples of $\mathrm{CV}(w|\mathcal{C})$ for randomly selected words from Hamilton's corpus in the Federalist Papers. The multinomial sampling model for author characteristic words predicts that such words typically appear on the left-hand side of Figure 4. In what follows, we verify that HC is mostly affected by such words.

3.3. *Across-corpus coefficient of variation vs. P-value.* In order to identify words likely to influence the HC-discrepancy heavily, we perform many two-sample HC tests involving document-corpus pairs and quantify the properties of the words in the word-list selected by the HC calculation. Given a document D, a corpus $\mathcal{C}$, and a vocabulary W of words, the P-values with respect to the binomial allocation model between the two word-frequency tables provide an ordering of the words in W. For each position in this ordering, we record $\mathrm{CV}(w|\mathcal{C})$ of the word w appearing at that position and average the result over multiple document-corpus pairs. For a document-corpus pair in which the document happens to be a member of the corpus, we remove the document from the corpus before applying the HC test. Figure 5 illustrates the results of this evaluation: The top panel shows values of $\mathrm{CV}(w|\mathcal{C})$ ordered according to the P-value of w obtained in a single test of one document against the corpus $\mathcal{C}$. The middle frame reveals the trend seen in the top panel by showing the average of $\mathrm{CV}(w|\mathcal{C})$ at each location across multiple document-corpus pairs. It follows that, on average, a word w associated with a small P-value is also associated with a small $\mathrm{CV}(w|\mathcal{C})$. This CV serves as a measure of the degree to which a word is author-characteristic vs. topic-characteristic. We conclude that *words associated with small P-values are typically author-characteristic*, suggesting that the HC-discrepancy is, mostly, affected by author-characteristic words and explaining to some degree why it discriminates well between documents and corpora of different authorship.

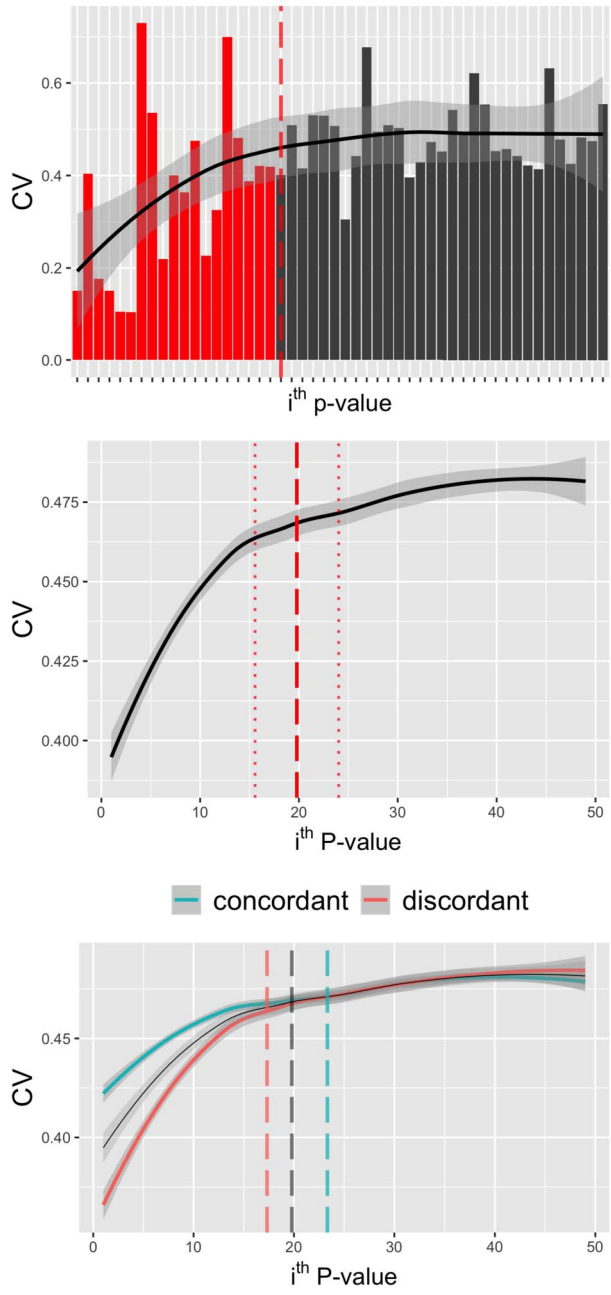

FIG. 5. *Coefficient of variation* $\mathrm{CV}(w|\mathcal{C})$ *within a corpus ordered according to the rank of the P-value of the word* w *in testing individual documents against a corpus* $\mathcal{C}$. *Proper names and cardinal numbers were removed. Top*: *Results from a single test. Words falling below the HC threshold are indicated in red. The smooth line is a LOESS curve fit. Middle*: *Average of* $\mathrm{CV}(w|\mathcal{C})$ *for each rank over* 1000 *document-corpus pairs. The vertical line indicates the mean value of the HC threshold, while the vertical dashed lines indicate the range of the HC threshold in* 95% *of the cases. Bottom*: *Average of* $\mathrm{CV}(w|\mathcal{C})$ *for each rank over* 1000 *document-corpus pairs while distinguishing cases when tested document is from the corpus of the same author* (*concordant*) *and the corpus of a different author* (*discordant*). *Vertical lines indicate the mean value of the HC threshold. The dark line is the global average, also given in the middle panel.*

3.4. *Concordant and discordant tests*. As a final illustration for our proposed interpretation of factors driving the success of HC in authorship attribution challenges, we distinguish between the case where the document and the corpus in each case have the same author (concordant) or not (discordant). Namely, we repeat the testing and averaging procedure outlined above, but, in addition, we mark whether the document and corpus are concordant or discordant. The results of this procedure are illustrated in the bottom panel of Figure 5. This figure shows that the averaged CV is significantly smaller in discordant pairs and that the HC threshold appears to derive the change between the curves. Since smaller P-values mean larger HC, this situation is in agreement with the observation that HC is affected by P-values of words associated with smaller CVs across a corpus of homogeneous authorship.

4. Conclusions. We developed a technique to measure the similarity/discrepancy of two word-frequency tables and applied it to authorship attribution challenges. Our measure, d_{HC}, uses the HC of P-values obtained from a word-level binomial allocation model. The HC calculation also identifies a set of words where there seem to be notable differences between the two tables.

When applied to authorship attribution challenges, we measure the value of the novel document's d_{HC} score relative to each corpus, attributing authorship based on the smallest rank-score. This automated procedure gives results comparable to previous studies but without handcrafting or tuning. In analyzing the ingredients for the success of our technique in authorship attribution, we found that, in practice, our discrepancy measure is mostly affected by words associated with low variance within a corpus of homogeneous authorship.

Acknowledgments. The author would like to thank David Donoho for fruitful discussions and three anonymous reviewers for providing comments that have greatly improved this paper.

Funding. This work is supported in parts by a fellowship from the Koret Foundation.

SUPPLEMENTARY MATERIAL

HCAuthorship (DOI: 10.1214/21-AOAS1544SUPP; .zip). Code and data for generating the figures and evaluating the performance in authorship attribution challenges.

REFERENCES

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877 https://doi.org/10.1214/11-AOS910

ARIAS-CASTRO, E. and WANG, M. (2015). The sparse Poisson means model. *Electron. J. Stat.* **9** 2170–2201. MR3406276 https://doi.org/10.1214/15-EJS1066

BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.* **12** 727–749. MR3834283 https://doi.org/10.1214/18-AOAS1155SF

BALAKRISHNAN, S. and WASSERMAN, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.* **47** 1893–1927. MR3953439 https://doi.org/10.1214/18-AOS1729

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*: *Theory and Practice*. The MIT Press, Cambridge, MA–London. With the collaboration of Richard J. Light and Frederick Mosteller. MR0381130

BLEI, D. M. and LAFFERTY, J. D. (2007). A correlated topic model of *Science*. *Ann. Appl. Stat.* **1** 17–35. MR2393839 https://doi.org/10.1214/07-AOAS114

BRESLOW, N. E. (1984). Extra-Poisson variation in log-linear models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **33** 38–44.

BROWN, L. D., ZHANG, R. and ZHAO, L. (2001). Root un-root methodology for nonparametric density estimation. Techical Report, The Wharton School, Univ. Pennsylvania.

CAI, T. T., JENG, X. J. and JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 629–662. MR2867452 https://doi.org/10.1111/j.1467-9868.2011.00778.x

CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653 https://doi.org/10.1214/009053607000000334

CHANG, J. and BLEI, D. M. (2010). Hierarchical relational models for document networks. *Ann. Appl. Stat.* **4** 124–150. MR2758167 https://doi.org/10.1214/09-AOAS309

CHURCH, K. W. and GALE, W. A. (1995). Poisson mixtures. *Nat. Lang. Eng.* **1** 163–190.

COX, D. R. and BRANDWOOD, L. (1959). On a discriminatory problem connected with the works of Plato. *J. Roy. Statist. Soc. Ser. B* **21** 195–200. MR0109102

CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464. MR0790631

DELAIGLE, A. and HALL, P. (2009). Higher criticism in the context of unknown distribution, non-independence and classification. In *Perspectives in Mathematical Sciences. I. Stat. Sci. Interdiscip. Res.* **7** 109–138. World Sci. Publ., Hackensack, NJ. MR2581742 https://doi.org/10.1142/9789814273633_0006

DENG, K., GENG, Z. and LIU, J. S. (2014). Association pattern discovery via theme dictionary models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 319–347. MR3164869 https://doi.org/10.1111/rssb.12032

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

DONOHO, D. and JIN, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4449–4470. With electronic supplementary materials available online. MR2546396 https://doi.org/10.1098/rsta.2009.0129

DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751 https://doi.org/10.1214/14-STS506

DONOHO, D. L. and KIPNIS, A. (2020). Higher criticism to compare two large frequency tables, with sensitivity to possible rare and weak differences.

EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

GLICKMAN, M., BROWN, J. and SONG, R. (2019). (A) data in the life: Authorship attribution in Lennon–McCartney songs. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.130f856e

GRIFFITHS, T. L., JORDAN, M. I., TENENBAUM, J. B. and BLEI, D. M. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems* 17–24.

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764

HAMILTON, A., MADISON, J. and JAY, J. (1961). The federalist papers, ed. Clinton Rossiter (New York: New American Library, 1961), 301. *Federalism*, *Citizenship*, *and Community* **207**.

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **36** 369–408. MR0173322 https://doi.org/10.1214/aoms/1177700150

HOLMES, D. I. (1985). The analysis of literary style—a review. *J. R. Stat. Soc., A* **148** 328–341.

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131 https://doi.org/10.1214/10-EJS589

JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. MR2363962 https://doi.org/10.1214/0009053607000000244

JIN, J. and KE, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statist. Sinica* **26** 1–34. MR3468343

JIN, J. and WANG, W. (2016). Influential features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359. MR3576543 https://doi.org/10.1214/15-AOS1423

JUOLA, P. (2008). Authorship attribution. *Found. Trends Inf. Retr.* **1** 233–334.

KESTEMONT, M., TSCHUGGNALL, M., STAMATATOS, E., DAELEMANS, W., SPECHT, G., STEIN, B. and POTTHAST, M. (2018). Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF* 2018 *Evaluation Labs. Avignon*, *France*, *September* 10–14, 2018/*Cappellato*, *Linda* [*edit.*]; *et al.* 1–25.

KESTEMONT, M., MANJAVACAS, E., MARKOV, I., BEVENDORFF, J., WIEGMANN, M., STAMATATOS, E., POTTHAST, M. and STEIN, B. (2020). Overview of the cross-domain authorship verification task at PAN 2020. In *CLEF* (*Working Notes*).

KIPNIS, A. (2020). Higher criticism as an unsupervised authorship discriminator. In *CLEF* (*Working Notes*).

KIPNIS, A. (2022). Supplement to "Higher criticism for discriminating word-frequency tables and authorship attribution." https://doi.org/10.1214/21-AOAS1544SUPP

LEHMANN, E. L. (1975). *Nonparametrics*: *Statistical Methods Based on Ranks*. *Holden-Day Series in Probability and Statistics*. Holden-Day, Inc., San Francisco, CA; McGraw-Hill International Book Co., New York–Düsseldorf. With the special assistance of H. J. M. d'Abrera. MR0395032

LI, J. and SIEGMUND, D. (2015). Higher criticism: p-values and criticism. *Ann*. *Statist*. **43** 1323–1350. MR3346705 https://doi.org/10.1214/15-AOS1312

MANNING, C., RAGHAVAN, P. and SCHÜTZE, H. (2010). Introduction to information retrieval. *Nat*. *Lang*. *Eng*. **16** 100–103.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. Second edition [of MR0727836]. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MOSTELLER, F. and WALLACE, D. L. (1963). Inference in an authorship problem. *J. Amer. Statist*. *Assoc*. **58** 275–309.

MOSTELLER, F. and WALLACE, D. L. (1984). *Applied Bayesian and Classical Inference*: *The Case of The Federalist Papers*. *Springer Series in Statistics*. Springer, New York. Second edition of *Inference and disputed authorship: the Federalist*. MR0766742 https://doi.org/10.1007/978-1-4612-5256-6

MUKHERJEE, R., PILLAI, N. S. and LIN, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *Ann*. *Statist*. **43** 352–381. MR3311863 https://doi.org/10.1214/14-AOS1279

NORVIG, P. (2013). Common words in Google books. http://norvig.com/mayzner.html.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos*. *Mag*. **50** 157–175.

QI, P., DOZAT, T., ZHANG, Y. and MANNING, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL* 2018 *Shared Task*: *Multilingual Parsing from Raw Text to Universal Dependencies* 160–170. Association for Computational Linguistics, Brussels, Belgium.

READ, T. R. and CRESSIE, N. A. (2012). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer Science & Business Media. MR0955054 https://doi.org/10.1007/978-1-4612-4578-0

ROBERTS, M. E., STEWART, B. M. and AIROLDI, E. M. (2016). A model of text for experimentation in the social sciences. *J*. *Amer*. *Statist*. *Assoc*. **111** 988–1003. MR3561924 https://doi.org/10.1080/01621459.2016.1141684

ROSS, G. J. (2020). Tracking the evolution of literary style via Dirichlet-multinomial change point regression. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **183** 149–167. MR4049658

SICHEL, H. S. (1974). On a distribution representing sentence-length in written prose. *J*. *R*. *Stat*. *Soc*., *A* **137** 25–34.

THISTED, R. and EFRON, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74** 445–455. MR0909350 https://doi.org/10.1093/biomet/74.3.445

TILAHUN, G., FEUERVERGER, A. and GERVERS, M. (2012). Dating medieval English charters. *Ann*. *Appl*. *Stat*. **6** 1615–1640. MR3058677 https://doi.org/10.1214/12-AOAS566

WAKE, W. C. (1957). Sentence-length distributions of Greek authors. *J*. *R*. *Stat*. *Soc*., *A* **120** 331–346.

ZHENG, R., LI, J., CHEN, H. and HUANG, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *J*. *Am*. *Soc*. *Inf*. *Sci*. *Technol*. **57** 378–393.