NOTES

WALT WHITMAN AT THE *AURORA*: A MODEL FOR JOURNALISTIC ATTRIBUTION

Relatively little manuscript material exists to definitively tie Walt Whitman to the bulk of the journalistic writing attributed to him, particularly the writing in the early years of his career. Because the vast majority of his early journalistic work was unsigned, attribution is most often based on the knowledge of Whitman's involvement with a given paper, coupled with the identification of some sort of Whitmanic voice or tone in a given piece of writing. However, a writer's style and tone are often affected by the form and context in which they are writing, meaning that Whitman's journalistic voice is often quite different than his poetic voice, which is in turn different than his prose fiction voice. Furthermore, certain similarities of style and tone are found across a given genre; many nineteenth-century newspaper editorials sound quite similar, for example, making any discussion of authorship in nineteenth-century periodicals rife with uncertainty. Therefore, even for the most knowledgeable scholars, a belief that Whitman was the author of a given piece of journalism generally rests upon a trust in the *tradition* of attributing a piece to Whitman, with skepticism arising only in the face of strong evidence to the contrary.

Last year, the editorial team overseeing the treatment of Whitman's journalism for the *Walt Whitman Archive* decided to add an editorial note to the metadata at the top of each text file, explaining the *Archive*'s rationale for attributing a piece to Whitman. In the note, we lay out all of the factors—including the piece's attribution history—that influenced our decision to present the piece as likely authored by Whitman. We also embedded, within the TEI encoding, an expression of our level of certainty in Whitman's authorship. Finally, we noted in the metadata whether and how the piece was signed by Whitman in the original publication. These measures are an attempt to foreground for users the inherent uncertainty of authorship in nineteenth-century periodical materials. But they also offer the opportunity to begin thinking about how we might move beyond traditional methods of

attribution.

Lately, staff members who work on Whitman's journalism have begun to apply the tools of computational linguistics to bolster attribution claims. To test these tools, we chose a corpus of texts generally ascribed to Whitman by traditional attribution methods—the editorials from the *New York Aurora*—and applied a bootstrapped classification task, using the classify-function of *stylo* for R.[1] On its most fundamental level, our method constitutes a comparison of texts turned into ranked lists. A reader might think of these as a comparison of shopping lists: if some nebulous entity were to collect a life's worth of said lists and compile them into an inventory of "most frequently bought groceries," one could statistically assess how close a set of newly-discovered shopping lists is to this assembled list. While scholarship in the field tends to rely on lists of most frequent words (as "items" on our "shopping lists"),[2] our assessment uses most frequent character trigrams (strings of three characters as they appear in text). This particular technique was used previously by members of our group to assess Whitman's contribution to the *Brooklyn Daily Times*[3] and has shown promising results, especially for shorter corpora.[4] The authorial corpus for Whitman consisted of a version of "Manly Health and Training" with potentially plagiarized passages excised[5] as well as Whitman's confirmed contributions to the *Brooklyn Daily Eagle*. Against this corpus, and the corpora of fifteen other contemporary authors, we then compared the *Aurora* pieces. To ensure accuracy and reduce false positives, we not only relied on a sizable pool of candidates[6] but each attribution was also repeated thousands of times, each time with minor variations to the assessment, with only consensus results considered positive attributions.

For an initial round of assessment, all contributions identified by Bergman, Noverr, and Recchia in *The Collected Writings of Walt Whitman: The Journalism,* vol. I (1834-1846),[7] spanning from February 28 to April 28 of 1842, were grouped together and a bootstrap consensus tree was produced. In this method, proximity of authorial voice is illustrated as lines radiating from a center, with each "branch" representing a significant difference in authorial voice.

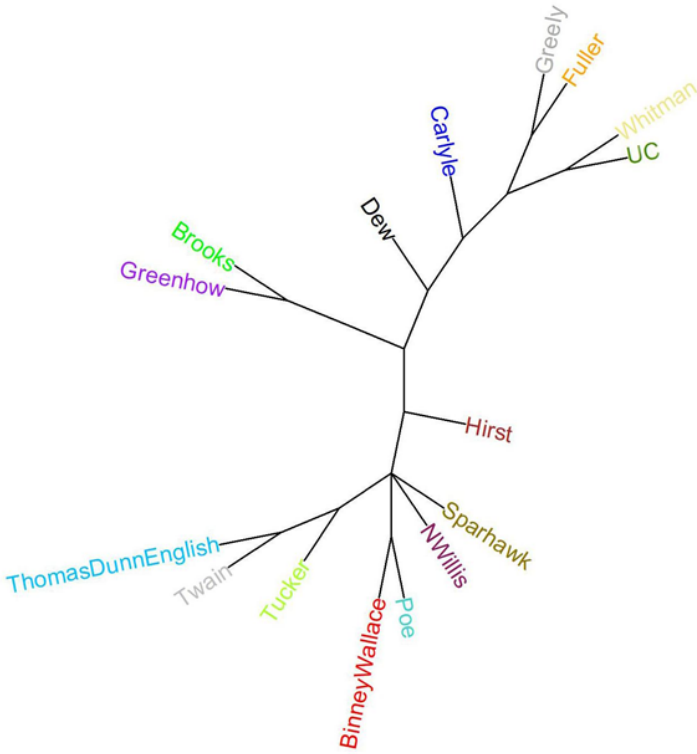

Figure 1. Bootstrap consensus tree for authorial corpora and Aurora texts, marked as "UC" ("unknown corpus") here (1801 Burrows' Delta-attributions on most frequent character trigram-lists from, incrementally growing from top 200 to 2000).

Figure 1 clearly shows that from all authorial corpora provided, Whitman's voice is by far the most similar (i.e. "least distant" in terms of compared lists) to the voice present in the *Aurora* corpus ("UC"): both voices have their own branch on the upper right corner of the consensus tree.

To get a more detailed view, we then grouped the writings in question by calendar week, and attributed based on lists of most frequent character trigrams, growing from top 200 to top 2,000 in increments of 1, and employing three difference measure of distance (Burrows' delta, Nearest Shrunken Centroid, Support Vector Machines). The results can be seen in Figure 2.

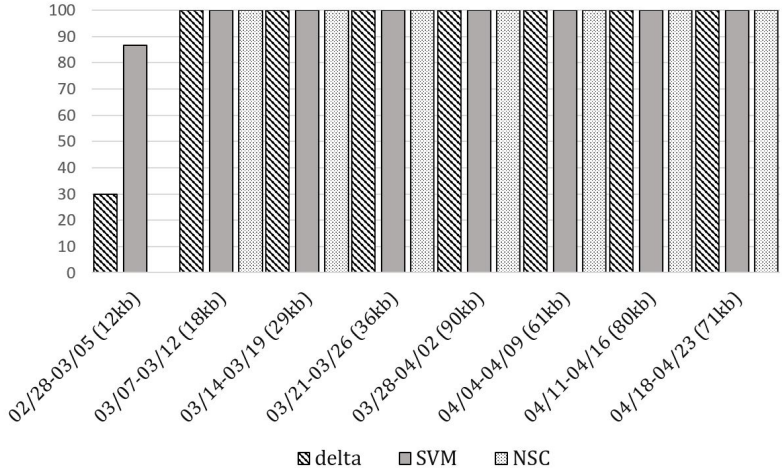

Figure 2. Percentage of attributions to Whitman of New York Aurora editorials grouped by week (in 1842), using NSC, SVM and Burrows' Delta (1801 attributions on most frequent character trigram-lists, incrementally growing from top 200 to 2000). Corpus size is listed in kb.

The data clearly suggests Whitman as the author of the vast majority of editorials attributed to him by Bergman, Noverr, and Recchia between March 7 and April 23 of 1842. Since Whitman was only announced as editor in the March 28 issue of the *Aurora*, it appears that the owners of the paper, Anson Herrick and John Ropes, assigned Whitman editorial duties soon after Thomas Low Nichols was fired as editor in February, but only announced Whitman's new position at the end of the month.

Given the limited data available for the first week (12kb / two editorials), a second round of assessment was performed, in which we multiplied the existing data by a factor of ten, which can help improve attribution success (see Figure 3).

While attribution rates for the first-week corpus went up slightly when its corpus was multiplied, it still could not comfortably be attributed to Whitman: NSC still failed to assign it to Whitman in any of its 1,801 attributions. Still, this does not *exclude* Whitman as the author but merely underscores a need for more data. With more texts available at the margins of Whitman's likely editorship, the exact period of his tenure could be narrowed down more precisely. The current state of the data does not allow us to make a clear determination as

to the beginning or the end of this period.

Additionally, we decided to assess a particular set of texts as a thematic group, namely those that exhibited strong nativist sentiments towards Irish Catholics in New York. These editorials centered around the funding debate over the Public School Society in which Whig Governor William Seward supported public funding for Catholic schools in an attempt to pull Irish Catholic New Yorkers away from the Democratic Party.[8] Considering their varying placement in the paper, we also assessed these themed texts grouped first as "leaders"—providing commentary on the most pertinent topic of the day in the first column on page two of each issue—and secondly as regular editorials. As expected, given the overall positive attribution, all three groupings were clearly attributed to the authorial signal of Whitman.

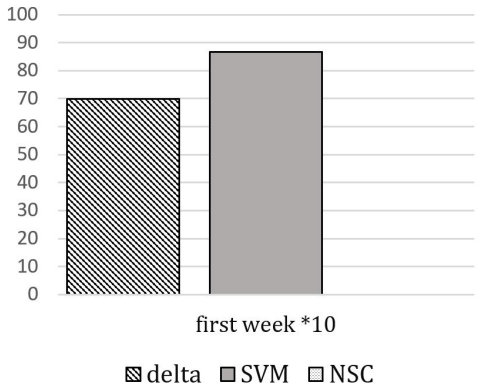

Figure 3. Additional bootstrapping performed on first (02/28-03/05) week of assessment, using NSC, SVM and Burrows' Delta (1801 attributions on most frequent character trigram-lists, incrementally growing from top 200 to 2000).

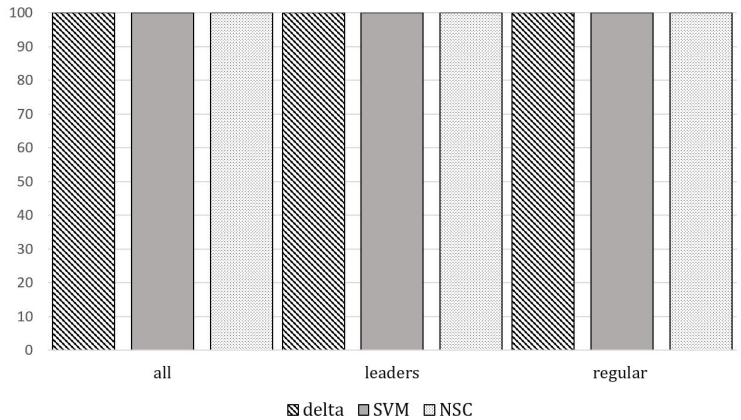

Figure 4. Assessment of nativist editorials, in percent, divided into leaders and regular contributions,[9] using NSC, SVM and Burrows' delta (1801 attributions on most frequent character 3gram-lists, incrementally growing from top 200 to 2000).

To ensure that the attribution has no accidental Whitman-bias or any other observable data-distortion, we included a text that should clearly fail attribution: Emerson's "American Scholar."[10] Since Emerson's voice was not included among our comparison corpora, "American Scholar" should not pass our benchmark for positive attribution (positive attributions for all three statistical methods). The results were a scattershot attribution across various authors: Delta attributed to Fuller (90.4%), Tucker (8.2%), or Brooks (1.4%), SVM to Greeley (68.4%), English (18.3%), or Starhawk (13.3%), while NSC attributed to Greeley (68.4%), Fuller (23.9%), or English (7.6%). There was no consistent misattribution across different measures of distance: the real author could not be identified, and the results reflect this clearly. In addition to what was gleaned from previous assessments of the method (see footnote 4), this on-the-fly test confirmed that non-attribution did not result in misattribution.

Our analysis of Whitman's *Aurora* editorials therefore confirms the scholarly consensus that Whitman was author of most of the *Aurora* material attributed to him, and perhaps resolves the longstanding debate as to whether the nativist articles found in the *Aurora* in the spring of 1842 were also penned by Whitman. Until now, the traditional attribution model of interpreting style and applying historical inference has led to conflicting conclusions about Whitman's authorship of these nativist pieces. The differences lie in determining both whether Whitman actually wrote these pieces, and, if he did, whether he believed them. David Reynolds calls Whitman's anti-Irish statements during this period a "strange dance" where the future poet "took the nativist side on several key questions," but ultimately "resisted thoroughgoing nativism" by rejecting the platform of the Native American Party and calling for benevolence toward newcomers.[11] Jerome Loving, on the other hand, finds it "difficult to believe that [Whitman] participated comfortably in the xenophobic 'Native American' campaign the *Aurora* launched in March and April" and reminds us that the co-owners of the newspaper, Anson Herrick and John Ropes, "probably did more than simply 'inspire' its opinions." [12] For Loving, the *Aurora*'s nativism "matches more with the language of Herrick and Ropes's [later] denunciation of Whitman," making "it . . . beyond even

the power of miraculous transformation . . . to think the spouter of these xenophobic editorials is the same person who . . . wrote *Leaves of Grass*." Joann Krieg argues that the "language Whitman employed was that of his readers" and implies that the young Whitman parroted the nativism of his subscribers.[13] While it is unlikely scholars will ever determine whether Whitman *meant* his critique of "insidious traitors from abroad,"[14] we can now definitively say, at least according to the computational model employed in this analysis, that Walt Whitman was the author of the nativist editorials published in the *Aurora* in the spring of 1842.

These computational assessments are the *Archive*'s latest effort to bring both renewed focus and a measure of clarity to Whitman's journalism—a vast, often neglected, but tremendously rich collection of writings that span the length of his career. We eventually plan to apply these methods to all of Whitman's journalistic material, incorporating our findings into the encoding as well as the editorial notes that accompany each piece. Ultimately, we hope that our approach can serve as a model for attribution work on other writers and lend some certainty to the long tradition of author attribution.

Kevin McMullen — *University of Nebraska-Lincoln*
Stefan Schöberlein — *Marshall University*
Jason Stacy — *Southern Illinois University-Edwardsville*

NOTES

1 Maciej Eder, Mike Kestemont, and Jan Rybicki, "Stylometry with R: a Suite of Tools," *Digital Humanities 2013: Conference Abstracts*, University of Nebraska-Lincoln, Lincoln (2013), 487- 489.

2 For an overview, see Efstathios Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology* 60 (2009), 538-556.

3 Walt Whitman, *The Complete Writings of Walt Whitman: The Journalism, vol. III 1848-1858*, Douglas Noverr, Jason Stacy, Zachary Turpin, eds, (New York:

Peter Lang, forthcoming). Stefan Schöberlein analyzed the corpus of *Brooklyn Daily Times* editorials between 1857 and 1858 for this project and described his methods in the introduction.

4 For a proof-of-concept assessment of the particular approach used here, see Stefan Schöberlein, "Poe or Not Poe? A Stylometric Analysis of Edgar Allan Poe's Disputed Writings," *Digital Scholarship in the Humanities* 73 (2017), 644-646.

5 The authors would like to thank Stephanie M. Blalock for information on potential plagiarism in "Manly Health."

6 The comparison pool is an updated comparison corpus, based on the authorial corpora employed in the Poe-assessment cited in the previous note. It consists of authorial corpora by Mark Twain, Nathaniel B. Tucker, Thomas Dunn English, Edward V. Sparhawk, Edgar Allan Poe, Nathaniel P. Willis, Henry B, Hirst, Horace Greeley, Robert Greenhow, Margaret Fuller, Thomas R. Dew, Thomas Carlyle, N. C. Brooks, and Horace B. Wallace. The smallest authorial corpus is Sparhawk's—with 164kb of data.

7 February 28-April 23, 1842. Whitman was announced as editor of the *Aurora* on March 28, 1842; his tenure as editor likely ended in late April.

8 See Joann P. Krieg, *Whitman and the Irish*, (Iowa City: University of Iowa Press, 2000), 38-45; David Reynolds, *Walt Whitman's America: A Cultural Biography*, (New York: Alfred A. Knopf, 1995), 99-101; Jason Stacy, *Walt Whitman's Multitudes: Labor Reform and Persona in Whitman's Journalism and the First Edition of* Leaves of Grass, *1840-1855* (New York: Peter Lang, 2008), 59-67.

9 The editorials included in the group of nativist pieces were: "Sectarianism and Our Public Schools" (03/07/1842); "The Schools" (03/10/1842); "The Schools" (03/15/1842); "Insult to American Citizenship!" (03/17/1842); "The Aurora and the School Question" (03/18/1842); "Public Schools" (03/21/1842); "Americanism" (03/23/1842); "Tammany in Trouble" (03/24/1842); "Tammany's 'Family Jars'" (03/26/1842); "Organs of the Democracy" (03/29/1842); "The School Bill" (03/29/1842); "Defining 'Our Position'" (03/30/1842); "Dissensions of Tammany" (04/01/1842); "Tammany Meeting Last Night" (04/06/1842); "The Mask Thrown Off" (04/07/1842); "The School Bill" (04/08/1842); ["On Saturday night"] (04/11/1842); ["It is a fearful thing"] (04/12/1842);) "More Catholic Insolence!" (04/12/1842); "Result of the Election" (04/13/1842); "Incidents of Last Night" (04/13/1842); ["According to the best authenticated"] (04/14/1842); "Plots of the Jesuits!" (04/14/1842); "The Late Riots" (04/15/1842); "Where Will Tammany Have to Stop?" (04/15/1842); "The Catholic Rows Not Ended" (04/16/1842); ["The Aurora has been roaring"] (04/18/1842).

10 Essentially, the method used here only identifies a "most likely candidate" but does not show us how likely said candidate is. There is, thus, a risk that, depending on the measure used to identify this candidate, the classify function might pick a wrong one. Varying these measures allows us to spot such cases. Metaphorically

speaking, we have a fruit basket and a banana and we are trying to figure out, which fruit in the basket is most similar to the banana: if there is an actual banana in the basket our method should identify it no matter how this "similarity" is calculated—but in the absence of banana in the basket, different ways to calculate "similarity" should not result in a consensus candidate (resulting in something like: 40% apple, 60% raspberry for one method; 100% grapefruit for another; 20% apple, 80% grapefruit for a third, for example. If all three methods were to come back with 51% to 100% apple, this approach would be invalid).

11 Reynolds, 99. Jason Stacy likewise argues that in this case Whitman was a "single issue" nativist, and that his anti-immigrant sentiments were a reflection of a broader contemporary concern over the perceived threat of the Catholic Church's influence over American institutions, in this case, the quasi-public Public School Society of New York City, Stacy, 61.

12 Jerome Loving, *Walt Whitman: The Song of Himself* (Berkeley: University of California Press, 1999), 62-63.

13 Krieg, 44.

14 "Sectarianism and Our Public Schools," *New York Aurora*, 3/7/1842, Bergman, et al., *The Collected Writings of Walt Whitman: The Journalism, vol. I 1834-1846*, (New York: Peter Lang), 43.