Customer Needs and Solutions (2021) 8:66–83
https://doi.org/10.1007/s40547-021-00116-x

RESEARCH ARTICLE

A Two-Stage Authorship Attribution Method Using Text and Structured Data for De-Anonymizing User-Generated Content

Matthew J. Schneider[1] · Shawn Mankad[2]

Accepted: 5 July 2021 / Published online: 6 August 2021

Abstract
User-generated content (UGC) is an important source of information on products and services for consumers and firms. Although incentivizing high-quality UGC is an important business objective for any content platform, we show that it is also possible to identify anonymous posters by exploiting the characteristics of posted content. We present a novel two-stage authorship attribution methodology that combines structured and text data by identifying an author first by the amount and granularity of structured data (e.g., location, first name) posted with the UGC and second by the author's writing style. As a case study, we show that 75% of the 1.3 million users in data publicly released by Yelp are uniquely identified by three structured variable combinations. For the remaining 25%, when the number of potential authors with (nearly) identically structured data ranges from 100 to 5 and sufficient training data exists for text analysis, the average probabilities of identification range from 40 to 81%. Our findings suggest that UGC platforms concerned with the potential negative effects of privacy-related incidents should limit or generalize their posters' structured data when it is adjoined with textual content or mentioned in the text itself. We also show that although protection policies that focus on structured data remove the most predictive elements of authorship, they also have a small negative effect on the usefulness of content.

Keywords Data privacy · De-anonymization · Stylometry · User identity linkage · Data linkage

1 Introduction

Authors of content on platforms such as Yelp, Twitter, Reddit, or Glassdoor often assume that the posts they contribute are anonymous. However, the subjects of such posts may seek to de-anonymize such content. For example, a Yelp user named "Birdzeye B" posted a negative one-star review about a California law firm. The maligned legal team identified and sued Ava Bird, obtaining a default judgment for over half a million dollars [51]. ZL technologies also filed suit against Glassdoor to obtain the identities of anonymous authors by claiming they wrote libelous reviews about the mismanagement of their company [14]. In another case, an accountant in California won an appeals court decision to confirm his belief that an author on Yelp named "Alex M" was a known female customer [30]. Yelp's senior director of litigation states that such cases "would really inhibit a website's ability to provide a balanced spectrum of views online and make it more doubtful that people would get the information they need to make informed decisions" [48].

Firms may want to identify an author for many other reasons. For example, a review platform may want to protect its reputation and identify high-volume reviewers with multiple accounts writing fake reviews [2]). Also, a vendor may want to respond directly to a user's concern on an external platform. For example, AMZ Tracker [6] suggested vendors identify reviewers to respond to complaints by "comparing the date and time of the review with your sales records, or asking publicly by replying to the review." In addition to commercial reasons, other motives for authorship attribution include "doxing" authors [18, 59] or exposing privacy vulnerabilities in the review platform [47].

We show in this paper that it is possible to identify an author using a combination of structured data (e.g., location and first name) and textual content (e.g., word choice and

✉ Matthew J. Schneider
mjs624@drexel.edu

Shawn Mankad
spm263@cornell.edu

[1] Drexel University, Philadelphia, PA 19104, USA

[2] Cornell University, Ithaca, NY 14853, USA

writing style) in the online review. Narrowing the identity of an author to a single individual is usually quite difficult because a multiplicity of potential authors matches the posted data; such a search has a remote chance of succeeding. However, more granular structured data quickly narrows the number of potential authors and greatly improves the chances of authorship attribution. We find that once the number is small enough, the textual content further helps to identify an author.

Protecting the anonymity of authors is very important for platforms that host user-generated content. Anonymity prevents retaliation for truth telling and is a prerequisite for other related domains such as peer reviews of academic journal articles, teaching evaluations, consumer surveys, and criminal complaints. Many academic journals attempt to have both anonymous referees and authors, yet research shows that authors can be identified using citations [27] or other textual content in the article [49]. Teaching evaluations by students are also assumed anonymous, and evidence shows that non-anonymous students rate their professors more positively than anonymous students [62]. Consumer surveys can also be designed to protect anonymity through randomized response surveys to help "induce respondents to answer more truthfully" ([15], p. 21). Further, strong privacy management practices provide consumers with a positive brand experience and lead to competitive advantage for platforms [24, 43], while consumer privacy concerns can translate into negative brand effects [39]. Moreover, the impact of information privacy and brand effects can be substantial, as platforms are famously sensitive to network effects [54], where different sides of the market can impact each other. If reviewers have privacy concerns, they may stop publishing certain content, which, in turn, impacts the utility derived by other consumers, firms, and advertisers.

To be concrete, consider a *data intruder* whose goal is to identify the author of an online post hosted by the *data host* (i.e., the content platform). The data intruder could be the subject of the review such as a restaurant or vendor whose objective is to address customer service concerns or collect more data on their consumer profiles. Such a data intruder would have access to their own sales data, including transaction timestamps, customer names (for credit card sales), details of past transactions, and potentially some customer-level demographics. The data intruder could be the general public as was the case with "online sleuths" attempting to de-anonymize right-wing Parler users through a combination of their structured metadata and user-generated content (UGC) [34]. The data intruder could also be another platform. For example, Facebook once attempted to match its user data to "anonymous" patient data from several US hospitals (data hosts) in order to figure out which patients needed special treatment [21]. It is also conceivable that a company like Amazon may want to identify which of its customers reviewed Amazon's own private-label products on an external review site. We consider a scenario in which the data intruder attempts to identify an author of a particular post by comparing their consumer data to the external content published on the review platform. On the other hand, the data host may try to prevent this from happening to protect their users' privacy and thereby increase the brand value of their platform [43]. A key privacy decision for the data host is to determine which user information (e.g., first name, review date, location, email address, IP address) is publicly displayed. For instance, Angie's List requires that all reviewers are not anonymous, Yelp allows relatively anonymous reviewers, and the review platform Consumer Affairs allows users to use initials for their name or greater geographic areas for their location. Draconian protection policies that reveal nothing about the author would reduce utility for data users, the general public that uses online reviews, and related user-generated content (UGC) to make decisions. In fact, Yelp's [75] user agreement highlights this trade-off explicitly by stating, "Other users may be able to identify you, or associate you with your account, if you include personal information in the content you post publicly. You can reduce the risk of being personally identified by using the Service pseudonymously, though doing so could detract from the credibility of your contributions to the Service."

We show how the data intruder's ability to attribute authorship changes according to the number of potential authors (i.e., collection of individuals who share characteristics) and secondarily by the volume of writing samples that the data intruder can use as training data for text analysis. We find that 75% of the 1.3 million reviewers on Yelp are uniquely identified by only three structured data variable combinations. The remaining reviewers can usually be identified by their writing style when the number of potential authors is less than 70 and sufficient training data is available. Specifically, with 100 training documents on each reviewer, the average probabilities of identification range from 40 to 81% when the number of potential authors ranges from 100 to 5, respectively. With only 5 training data points on each reviewer, the average probabilities of identification range from 18 to 54% when the number of potential authors ranges from 100 to 5, respectively. We also show that these results hold for several other types of content from student essays, Twitter posts, and redacted emails.

We then propose data-protection guidelines for the data host that would reduce the ability of the data intruder to identify the author of their UGC. In particular, we demonstrate how content platforms can protect the anonymity of their content contributors by generalizing their structured data, thereby enlarging the number of potential authors. For example, the data host may generalize the information presented publicly by changing an author's first name to first initial or an author's city to state. Importantly, generalization of structured data is

easy to implement, does not fundamentally affect content, and has been successfully applied to protect privacy in other contexts [65]. As a result of data generalization, the protected data continue to preserve the utility of the textual content contained in the UGC. Furthermore, we also show how structured data can also be found in the text itself and how the data host can use automated (or suggested) redaction to prevent this at a slight cost to the usefulness of an online review. Overall, our findings suggest that platforms concerned with brand effects due to privacy should limit or generalize their users' structured data when adjoined with textual content.

Our main contributions to the marketing, authorship attribution, and privacy literatures are the following.

1. Establishing the lack of privacy in UGC — We demonstrate an emerging privacy concern in UGC [7] using a general two-stage authorship attribution algorithm that combines structured and text content. We find that the chances for authorship attribution are very high in certain circumstances and directly proportional to the amount of structured data publicly posted. Importantly, the chances of authorship attribution do not seem to be greatly affected by the variety of products or services being reviewed because our two-stage approach leverages information separate from the review content itself, such as demographics and the style of writing. Second, based on the results of the authorship attribution algorithm, we investigate a realistic and straightforward generalization and text-redaction-protection strategy that the data host could use to mitigate the chances of identification. We find that the generalization of structured data is effective to a limited degree. Last, we study the externalities that potentially spill over to the broader public that would result from these protection strategies. We use the usefulness of the review to measure general utility and find evidence that generalization and redaction tend to lower usefulness. Thus, a trade-off exists for a platform between protecting its users from being identified and having content provide utility to the general public.
2. Authorship attribution using structured and unstructured data — As discussed in detail subsequently, the extant literature on de-anonymization tends to focus on either structured data or textual content, not both. Our paper differs from previous methodological literature by combining both in a systematic way for authorship attribution. Our approach newly combines ideas from stylometry, entity resolution, and database linkage.
3. Authorship attribution with limited training data — To our knowledge, we are the first to show that standard machine-learning methods are precise enough to accurately draw conclusions from a small number of text observations on potential authors. Our results show that as few as five training documents (or, equivalently, hundreds of words per author) can be enough to attribute authorship at a level several multiples higher than random guessing, even when the number of potential authors is large. Note that previous studies investigated larger training datasets: at least 6500 words per author [1, 10, 31, 32] and larger numbers of training documents per author [36, 53]. Our findings raise significant privacy concerns given that the bar to de-anonymize content through textual content is much lower than previous studies demonstrated.

2 Literature Review

In Section 2.1, we provide a brief discussion of how our paper fits into the large literature on privacy in Marketing. Section 2.2 discusses the authorship attribution methodology with respect to several related research streams.

2.1 Privacy and Online Reviews

Privacy as an area of study spans multiple fields, from economics and public policy [13] to Computer Science [47] and Marketing [43]. It has been the focus of scholarly work with electronic health records [40], chat rooms and forum posting [1], blogs [46], and the (supposedly) blind academic peer-review system [27, 49], among several other application areas.

As discussed in Wedel and Kannan [71] and Martin and Murphy [43], analytics in the context of customers' privacy has important implications for organizations, and consequently, several recent scholarly works have emerged on the issue. For example, Schneider et al. [55] studied how companies can share highly granular customer-level data (e.g., customer segment or point of sale) with its partners while minimizing privacy risks. Extensive literature describes database-protection methods to allow for safer data sharing (see, e.g., [41, 44]). When working with structured data, standard approaches for data protection commonly hide the gender or location of individuals, minimize the amount of data released [71], or change locations to greater geographical regions. For further discussion, Aggarwal and Philip [3] provided a summary of extant data-protection methods for structured data.

Note that in contrast to traditional settings, in online reviews and UGC, most data generated are unstructured in nature. Yet, with few exceptions [40], the vast majority of research on privacy and protection of data addresses structured data stored in relational databases [3]. Recent studies published within Computer Science have attempted to link users across platforms using their user-generated content and "friend" network linkages.[1] These

[1] As reviewed in Shu et al. [57], this problem is known by a number of names, including User Identity Linkage, Social Identity Linkage, User Identity Resolution, Social Network Reconciliation, User Account Linkage Inference, Profile Linkage, Anchor Link Prediction, and Detecting me edges.

works tend to focus on the network data (e.g., friend relationships) and bag of words analysis of the text [57]. With respect to this stream of work, the context of studying identifiability within online reviews and our methodological approach that combines structured and unstructured data are novel. Within the privacy and de-anonymization literature, key issues are still not well understood, such as (i) how much (training) data are necessary to reverse engineer the true author of an anonymous text document, (ii) what the best modeling approaches are and which features are most important to extract from the text for de-anonymization, and (iii) what are the resultant implications for protection strategies that can mitigate the risk of privacy breaches. This paper aims to fill these gaps in the literature in the context of online reviews to mitigate the paucity of research on privacy.

Online review and UGC platforms have received extensive scholarly focus, though typically not for privacy-related problems. For example, recent studies focus on the effects of textual content on product sales [23], users' decisions to post content [4, 45], the impact of management responses [26, 52], the rebroadcasting of Twitter posts [77], the effects of fake content [78], and the combination of structured data and textual content for forecasting and brand analysis [12, 56, 66].

A novelty of our work is in the context of studying identifiability within online platforms. The context motivates our methodological approach that newly combines structured and unstructured data. For example, we pursue a first stage matching and filtering that resembles methods developed in the entity resolution and data linkage literature, followed by text stylometric analysis in a second stage. We discuss how our algorithm fits into these fields next.

2.2 Authorship Attribution Algorithm

In the first stage, we utilize matching methods on structured data variables that have been studied and pursued extensively for database quality where the goal is to identify different records in a table that refers to the same underlying entity. Typically, the methods for solving this problem fall under the headings of "entity resolution" or "database linkage" [11, 20, 22, 72]. After narrowing down the set of potential authors using the structured data, the second stage of our algorithm uses text analysis to rank order candidate authors. It is important to note that the context of online reviews creates conceptual differences when compared to the aforementioned database quality field. Because a large percentage of an organization's data can be textual, many methods to identify duplicate records in databases also employ text-based techniques [11]. For example, a canonical problem in these subdomains is to identify different representations of the same entity by comparing a street address field: "44 West Fourth Street" versus "44 W. 4th St." [20]. Though technically the data are textual, they are also relatively structured, allowing for effective use of matching methods such as exact matching between two text entries, partial string matching through distances measures, or cosine similarity [25]. This distinction highlights a key difference between our methodology and those traditionally used in the database field. Specifically, our second stage approach uses thousands of input features that capture more nuanced aspects of the given text, such as readability, style, and general content. The high-dimensional approach is necessary for accurate results, given the truly unstructured nature of user-generated content versus structured text fields in curated relational databases. Note also that the entity-resolution and database-linkage methods are traditionally applied for database cleansing of a single, internal company database. Our focus is the integration between separate databases, one of which is external to the data intruder.

Our second stage approach is closer to the field of stylometry (also known as authorship attribution), which aims to identify authors based on their textual content [35]. Our proposed method is consistent with the authorship attribution literature in that the second stage is supervised, i.e., the anonymous text is assigned to an author, given a set of candidate authors for whom text samples (i.e., training data) are available [5, 53, 60]. Previous works investigated larger training datasets: at least 10 training documents per author [36, 53] or 6500 words per author [1, 10, 31, 32]. Motivated by online reviews, we are the first to accurately draw conclusions from a limited number of writing samples that can be used for training a classifier (e.g., 5 training documents per author or equivalently hundreds of words per author). We also innovate on the methodology by incorporating both the structured data (first stage) and textual content (second stage) of a review.

Focusing on the evaluation of our authorship attribution method, a database is said to be *k*-anonymous when each person contained in the database cannot be distinguished from at least *k-1* individuals also in the data. This seminal concept was introduced by Sweeney [64], and we evaluate our results using an adaptation of it. Specifically, in our context, some users are uniquely identifiable (*k* equals one) simply due to the nature of user-generated content. To understand the scope of the potential privacy problem, it is also insightful to examine the most common reviewer profiles, for example, those who have the most common name or location in the dataset. Thus, when evaluating the effectiveness of the authorship attribution method and protection policies, we examine the full distribution of the sizes of groups of users that cannot be distinguished from each other.

3 Authorship Attribution Method

Here, we present an overview of the proposed authorship attribution methodology to identify authors with the combination of structured data and textual content. As illustrated in

Fig. 1, the process begins with an online review being published subject to the platform's policy regarding which user information (e.g., first name, review date, location) to publicly display. Once the data intruder sees the external review that it wants to de-anonymize, it uses its own customer database with structured data (full name, email address, some demographics, and details of past transactions) and optionally writing samples for each customer to perform the authorship attribution. In the first stage, the data intruder retains consumers as potential authors that match the observed structured data in the target review. The second stage rank orders the candidate authors by the probability of authorship estimated through a supervised text-based analysis. Next, we discuss the stages in detail, followed by how the content and style of the observed text are summarized for use in the second stage.

3.1 First Stage

To formally discuss the authorship attribution method, we first introduce some notation. We label the data intruder database as D. Each row, D_i contains structured data (username, first name, or location) for customer i. Let t be the target review published on the review platform with accompanying structured data (e.g., different username, location). The data intruder then computes N_t as the total number of authors in D that (approximately) matches the structured data of author t. N_t increases when the structured data on the review platform is generalized (e.g., larger geographical areas) by the data host. In our empirical work, we perform exact matching. However, when the data intruder is worried about the author strategically using fake information (like a random location), then soft matching can be used to cast a wider net while still greatly reducing the number of potential authors in D. The probability that author i is the author of target review t for the first stage is equal to $\frac{1}{N_t}$ for all authors, and the set of N_t authors in D is defined as N_t^*.

3.2 Second Stage

The second stage of the authorship attribution algorithm relies on the data intruder having access to at least a single writing sample for each candidate author that can be used for stylometric analysis. If such data is not available, then the algorithm concludes with the first stage with each potential author estimated as equally likely to have written the target review. When the data intruder is a multinational firm such as Facebook or Amazon, then writing samples can be compiled from a number of sources, including customer emails, surveys, voice transcriptions, reviews or posts on their own website, and so on. For other types of entities (e.g., restaurants, small businesses), activity on social networking sites can be harvested; we also note the recent proliferation of data brokers [38] whose service is to provide data (e.g., writing samples) for a list of individuals based on each individual's digital activity.

Having narrowed down the number of potential authors in the first-stage filtering and given the presence of training data, in the second stage, we view the problem of inferring authorship from the text of a document as a supervised learning problem. Formally, we predict the author Y_t for document t and $Y_t \in N_t^*$. Let x_t denote a vector of features extracted from the text, such as the count variables described in Section 3.3 for the text. In our experiments, we compare two different machine-learning methods: random forest and support vector machines (SVM).

For SVM [28], we use function ϕ with kernel $K(x_i, x_k) = \phi(x_i)^T \phi(x_k)$ to optimize $m = 1, \ldots, N_t$ SVM models. The mth SVM model minimizes

$$\underset{w^m, b^m, \xi^m}{min} \frac{1}{2} w^{mT} w^m + C \sum_{i=1}^{n} \xi_i^m,$$

where $i = 1, \ldots, n$ indexes the training data, C is the penalty parameter of the error term, $\frac{1}{2} w^{mT} w^m$ is the regularization

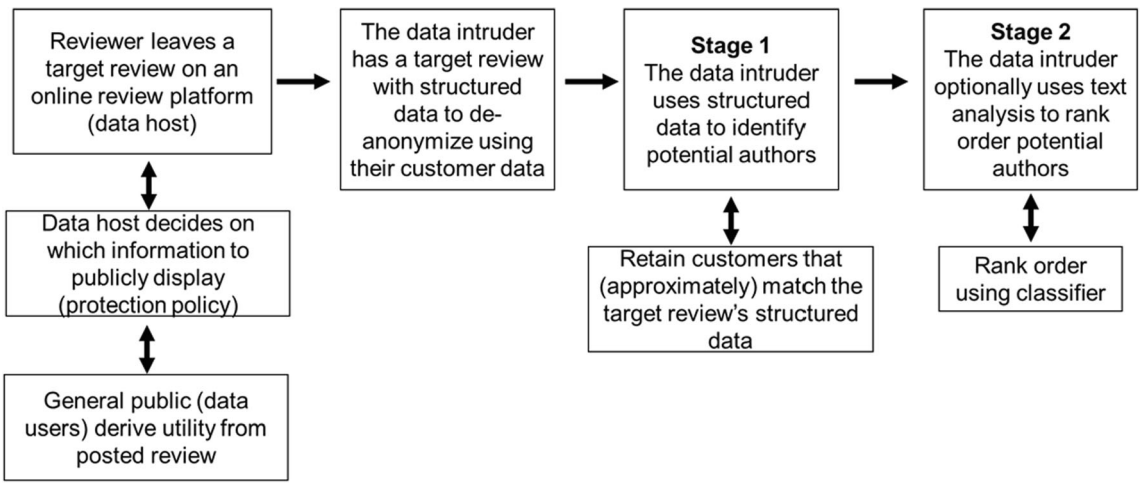

Fig. 1 Proposed authorship attribution model

term for the textual feature vectors x_i, and ${\xi_i}^{Y_i}$ is the smallest nonnegative number satisfying the constraints $(w^{mT}\phi(x_i) + b^m) \geq 1 - \xi_i^m$ if $Y_i = m$, else $(w^{mT}\phi(x_i) + b^m) \leq -1 + \xi_i^m$ if $Y_i \neq m$. For each SVM m, $w^{mT}\phi(x_i) + b^m$ represents the hyperplane separating the feature vectors x_i for author Y_i. To classify the target document t, we choose Y_t, corresponding to $\text{arg}max_m\, w^{mT}\phi(x_t) + b^m$.

For random forest [9], we use $m = 1, \ldots, N_t$ tree-based classifiers, $h(x = (x_1, x_2, \ldots, x_n), \Theta_m)$, where x_i is the textual feature vector and Θ_m are independent identically distributed random vectors generated from the tree construction. For each of 500 trees, we sample the number of features equal to $\sqrt{p}$ and sample x_i with replacement. To classify the target document t, we output the vector of probabilities for each input x_t and choose Y_t with the plurality vote.

These two techniques are both highly scalable and so well suited to the problem of predicting the author of a review from a moderate-sized set of potential authors. Although recent authorship attribution studies have used regularized multinomial logistic regression [46, 49], the computational cost of random forest and SVM, when faced with a large number of potential authors (classes), is much smaller in comparison to these alternative classification methods. The computational complexity of training random forest is $O\left(N_t * n_{tree} * \sqrt{p} * nlog(n)\right)$, growing linearly in the number of potential authors. For SVM, such a precise expression of computational complexity is difficult to derive. As discussed in Bordes et al. [8], regardless of the exact algorithm used to solve the quadratic optimization problem, the asymptotic computational cost grows, depending on the penalty parameter C and the number of samples n, that is, the main computational cost for training SVM does not align with the number of potential authors (classes). In addition to the scalability, in our experiments, the predictive capability of regularized multinomial regression was dominated by random forest. These two classifiers also cover the spectrum of assumptions with respect to the classification boundaries. SVM is naturally viewed as making a linear boundary assumption.[2] In contrast, random forest is a nonparametric methodology.

3.3 Text-Based Features for the Second Stage

The variables used in the second stage capture different aspects of content and style of writing, consistent with the literature. Specifically, we calculate four groups of variables: bag of words, stop words, punctuation, and stylometry (see Table 1).

The bag-of-words group consists of a set of count variables for every word used in any text document. For instance, the sentence "Technology is outpacing existing privacy protections in our laws" would create the variables "technology," "outpacing," "existing," and so on, where each variable is a count of the number of times the corresponding word appears in any document in the dataset. Such counts play a key role in text mining generally and are the input for a large number of content-analysis methods, including probabilistic topic modeling [61]. Because the dimensionality of bag-of-words features can easily be in the thousands, if not larger, this group of variables would require the data intruder to have sophisticated technical proficiency.

Most applications of text processing in marketing have focused on the bag of words, but the author identification literature suggests using other predictor variables. The three other sets of variables do not pose technical difficulties because they are lower dimensional constructs that measure aspects of style rather than content. The stop-word variables consist of 174 separate count variables of semantically empty words, such as "a," "and," and "the," that are necessary constructs of the English language. As such, these words help capture the formality and style of the author's writing.

The punctuation group consists of 14 separate count variables for different punctuation (e.g., "!," "#," "$") to again capture the style of writing. Finally, the stylometry variables, following Narayanan et al. [46] and Payer et al. [49], consist of just under 75 separate count variables, including seminal measures, such as the Flesch–Kincaid Grade Level and Yule's K, in addition to many variables that collectively help discriminate between authors in other contexts but are less interpretable. Measures of readability (also known as reading levels) are numerical scores that indicate how easy or difficult a given piece of text is to comprehend. The text has various measures of readability [19], though work in similar contexts (online reviews; [42]) focuses primarily on the reading level measure called the Flesch–Kincaid Grade Level [33]. This measure of readability combines factors such as word count and the number of syllables to estimate the comprehension difficulty of a given piece of text. Although originally developed by the US Navy to screen enlisted personnel, the Flesch–Kincaid Grade Level has remained popular due to its fast calculation—it is an automated statistic in Microsoft Word—and history of reliable results [33]. Yule's K is similar in spirit to reading levels and aims to compute the lexical diversity in a given text [76]. Specifically, it is a measure of vocabulary richness (see [68] for the exact mathematical definition) based on the rate at which words repeat. The other stylometry variables include the number of words of length K (for K = 1, 2, ..., to the largest observed word length), number of uppercase words, number of lowercase words, number of mixed-case words, how often numerical digits are used, counts of how often each letter was used, and so on.

[2] We explored several different kernels for SVM, including polynomial (2nd and 3rd order) and other nonlinear specifications. A linear kernel achieved the best results and is therefore presented throughout the paper.

Table 1 Predictor variables calculated from the text of each document

Category	Technical proficiency	Description	Count
Bag of words (Content)	Sophisticated	Frequency of any word that appears in the text	> 1000 (Depends on text corpus)
Punctuation	Simple	Frequency of ! . ? # @ $ % & , ; : ( )	14
Stop words	Moderate	Frequency of words like "the", "of," "and"	174
Stylometry	Simple	Frequency of *a* to *z* (ignoring case)	26
		Frequency of 0 to 9	10
		Number of words	1
	Moderate	Frequency of words with a mix of upper and lower case letters	1
		Frequency of words that have 1–30 characters	30
		Vocabulary richness (Yule's *K*)	1
		Flesch–Kincaid reading level	1

4 Empirical Application

We begin by summarizing the main dataset in Section 4.1, followed by an evaluation of each stage of the authorship attribution algorithm in Sections 4.2 and 4.3. In Section 4.4, we show the robustness of the algorithm by demonstrating consistent performance on several other datasets across different domains. In Section 4.5, we consider the problem from the data host's point of view and analyze the impact of generalization as a protection policy on the usefulness of reviews for the general public.

4.1 Data Description

As summarized in Table 2, we use the data from the Yelp academic challenge dataset available at http://www.yelp.com/dataset to illustrate the effectiveness of our authorship attribution method. It includes more than 5 million reviews from approximately 1.3 million reviewers. For each reviewer, the dataset provides a self-reported name field that we parse into first and last name and treat as structured data. Of reviews, 89.6% are posted under a first name that appears in data published by the US Social Security Administration [70] on the frequency of given first names in the population of US births where the individual has a Social Security number. This strong overlap between the two databases provides evidence that Yelp users often post under real first names. We also note that the percentage of actual first names being used is likely higher than 89.6%, as the remaining 10% of reviews tend to contain names written in different languages (e.g., with Chinese characters), which do not appear in the Social Security Administration's database. For last names, users are much less likely to provide reliable information: the percentage of reviews without any self-reported last name is 98.9%. When the last name is listed, it appears in the US Census Bureau's [69] surname database 69.9% of the time. When a user chooses to not self-report a structured data variable, such as the last name, we code this as "Unknown" for our empirical analysis.

4.2 Performance of the First Stage of the Authorship Attribution Algorithm

As mentioned previously, the data intruder may attempt to identify a reviewer by comparing the review to their customer database D to see if a unique match exists among common variables such as name or email address. To validate the first stage of the authorship attribution algorithm, we study how the number of potential customers varies depending on the characteristics D'_t of a hypothetical target review t, treating the Yelp database as the data intruder's customer database D. It would be ideal to have access to a gold-standard dataset that links reviewers across at least two separate platforms. Given regulations around the identification of customer data

Table 2 Summary of the investigated dataset

Dataset	Number of potential authors	Structured data variables	Number of documents	Average word count
Yelp online reviews	1,326,116	First name, last name, product/service	5,261,687	114.10

(e.g., General Data Protection Regulation and California Consumer Privacy Act rules), this remains a challenge. In this paper, we study the problem of privacy on UGC platforms by simulating the exercise through the use of data released by Yelp — a common strategy pursued in several scholarly works to overcome the lack of access to a gold standard dataset [57, 60].

When the first name with the target review was a common one, such as "John," we expect a large number of candidate reviewers to match this value. In fact, this is the most common first name in the Yelp dataset with 12,120 instances. In contrast, for a more uncommon name, such as "Yuqing," the number of matching potential authors should be smaller. There are 10 reviews by Yuqings in the Yelp data. Table 3 shows exactly this type of pattern. For example, when matching against last names only, distinctive last names quickly narrow the number of potential authors, whereas a user not entering any last name keeps the number of potential authors at 1.31 million. Table 3 also shows that from over 1.3 million users, only two pieces of structured data (first name + last name) result in $N_t = 1$ for 50% of combinations and three pieces of structured data (first name + last name + business/service purchase) result in $N_t = 1$ for 75% of the combinations. In terms of k-anonymity [64], Table 3 shows that k-anonymity = 1 for the Yelp dataset, since there exists at least one user who is uniquely identified by their structured data. We expect the performance of the first stage to hold even if the number of customers in the data intruder's database is larger than the dataset we analyze here: the number of candidate authors N_t will decrease as matching is performed with more granular structured data. When N_t is greater than 1, the data intruder proceeds to the second stage of the authorship attribution algorithm.

4.3 Performance of the Second Stage of the Authorship Attribution Algorithm

We investigate the efficacy of the second stage by dividing the reviews into training and test sets. Importantly, we vary the number of potential authors and the amount of available training data, both of which control the general difficulty of the problem. For example, if there are only two potential authors, then the problem of author identification reduces to a standard binary classification problem. For larger potential author sets, developing accurate predictions is more challenging because the cardinality of potential authors is much larger. To derive implications about the real-world feasibility to attribute authorship using statistical methods, we also vary the number of training samples from 5 training documents per author to 100 training documents per author. For each dataset, from the text of each document, we calculate all text-based variables in Table 1 and vary the data intruder's technical sophistication by training different machine-learning models with the three nested categories of features (simple, moderate, and sophisticated).

Finally, when testing the proposed algorithms, we compare out-of-sample predictions against withheld true authorship labels. Out-of-sample accuracy results appear in Fig. 2. Figure 4 in the Appendix provides expanded results for when the size of the training sample per author is between 5 and 100.

Random guessing (the green dotted line) assumes that the data intruder uses the structured data only with a probability of identification equal to $\frac{1}{N_t}$. Compared to random guessing, the probabilities of identification using machine-learning methods on text-based features are much larger for any level of sophistication, including "simple," which uses only frequencies of punctuation, letters, and numbers. We vary the input features to quantify the amount of discriminative information contained in each set of predictors and understand the practical

Table 3 First stage results for the Yelp dataset

Features for matching	Number of potential authors (N_t)					
	Avg.	50% quantile	75% quantile	95% quantile	99% quantile	Max
None (millions)	1.32 MM	1.32 MM	1.32 MM	1.32 MM	1.32 MM	1.32 MM
First name	13.94	1	2	16	179.01	12,120
Last name	269.97	1	2	7	37.89	1.32 MM
First name + last name	12.46	1	2	14	150	11,976
First name + last name initial	12.81	1	2	14	158	11,976
First name initial + last name initial	1719.98	10	27	123.5	69,822.9	165,687
First name + business/service purchase	1.15	1	1	2	4	65
First name initial + business/service purchase	3.57	1	3	12	33	956
First name + last name + business/service purchase	1.15	1	1	2	4	65
First name initial + last name initial + business/service purchase	3.45	1	3	12	32	944

difficulty of the prediction problem. The results show that the most informative variable group is the bag of words because the sophisticated data intruder consistently achieves the highest out-of-sample accuracy rate. It is noteworthy that even a simple or moderately sophisticated data intruder that uses variables that require less technical sophistication to measure achieves a reasonably high level of accuracy. In other words, we find evidence that even a simple data intruder, one that does not use bag of words or other content-driven features, will still be able to perform authorship attribution when given enough writing samples on each potential author.

Figure 3 shows that the data intruder will "usually" identify an anonymous reviewer by plotting the points on Fig. 2 that have an average probability of identification equal to or greater than 50%. When the training data size is large, a sophisticated data intruder using all the textual features can, on average, identify an anonymous reviewer correctly from over 70 potential authors.

We note that data privacy can be poor even if the average accuracy rates in Fig. 2, which are for a single target review, are relatively low. One reason is that the data intruder may have multiple target reviews written on the online review platform by the same author and the data intruder only has to identify the author of one such review. For example, Fig. 2 shows that the average out-of-sample probability of identification with 5 training data and 50 anonymous authors is about 25% (random guessing is 2%). Although this may not seem large, if there are ten reviews written by the same author, Y_t, on the online review platform, then the data intruder has ten separate chances to identify the author. Conservatively, this yields a $1 - 0.75^{10} = 94\%$ chance of identifying the true author of at least one of the ten reviews. Moreover, if one of the target reviews is particularly revealing, then the rest of the reviews are also compromised.

4.4 Second Stage Robustness Results with Additional Datasets

We analyze four additional datasets summarized in Table 4 to show the robustness of our authorship attribution algorithm. The data vary in terms of the document type, length, formality, and style of writing. For example, Twitter is by construction short and informal, whereas the two Brennan–Greenstadt corpora are long-form student essays.

The Brennan–Greenstadt and Extended Brennan–Greenstadt corpora were published by Brennan et al. [10]. The Brennan–Greenstadt corpus is based on a study conducted through Drexel University and contains 12 student authors who volunteered their time. The Extended Brennan–Greenstadt corpus contains authors solicited

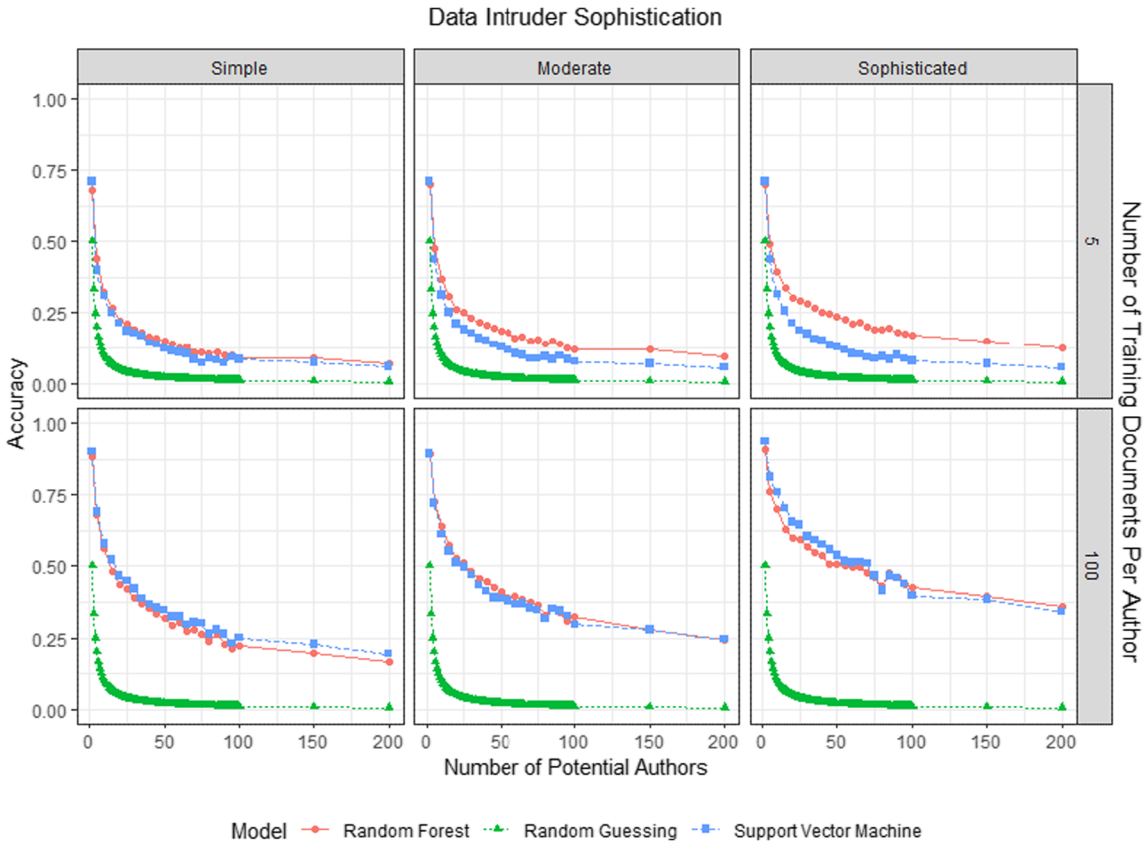

Fig. 2 Second stage accuracy results for the Yelp dataset

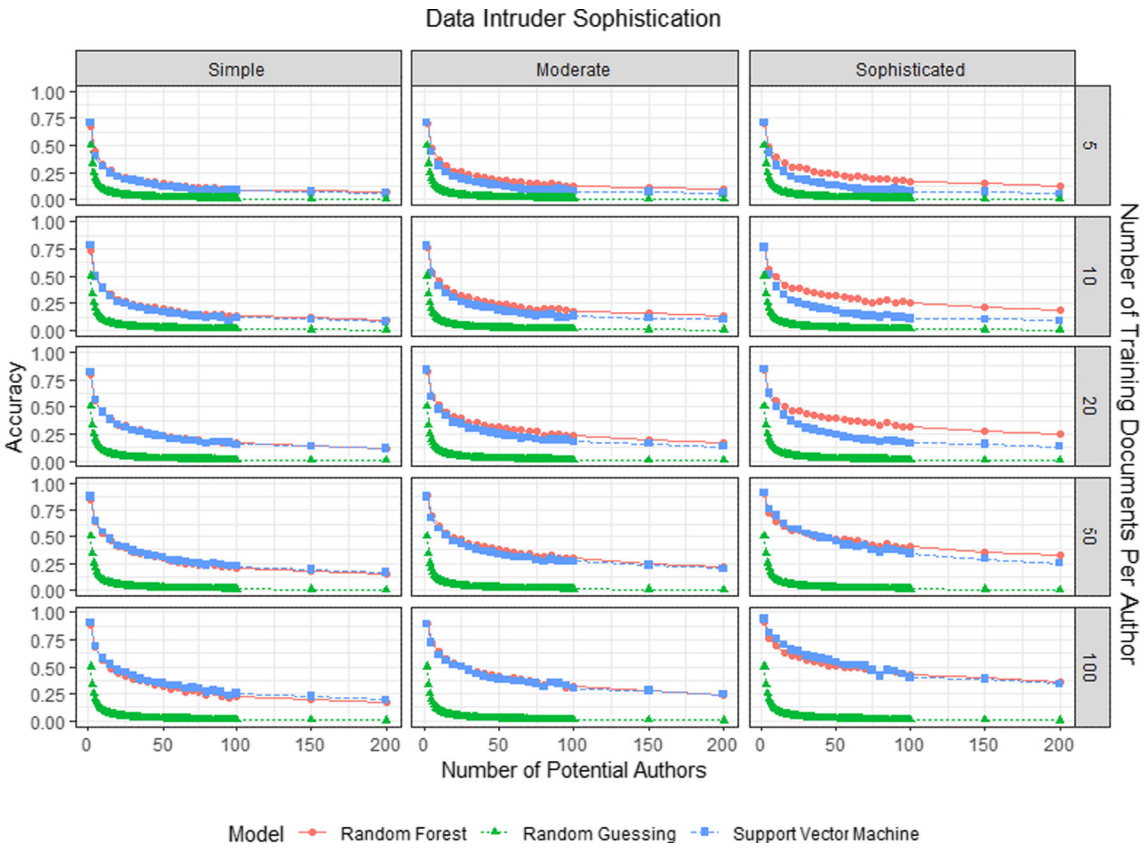

Fig. 4 Full second stage accuracy results for the Yelp dataset

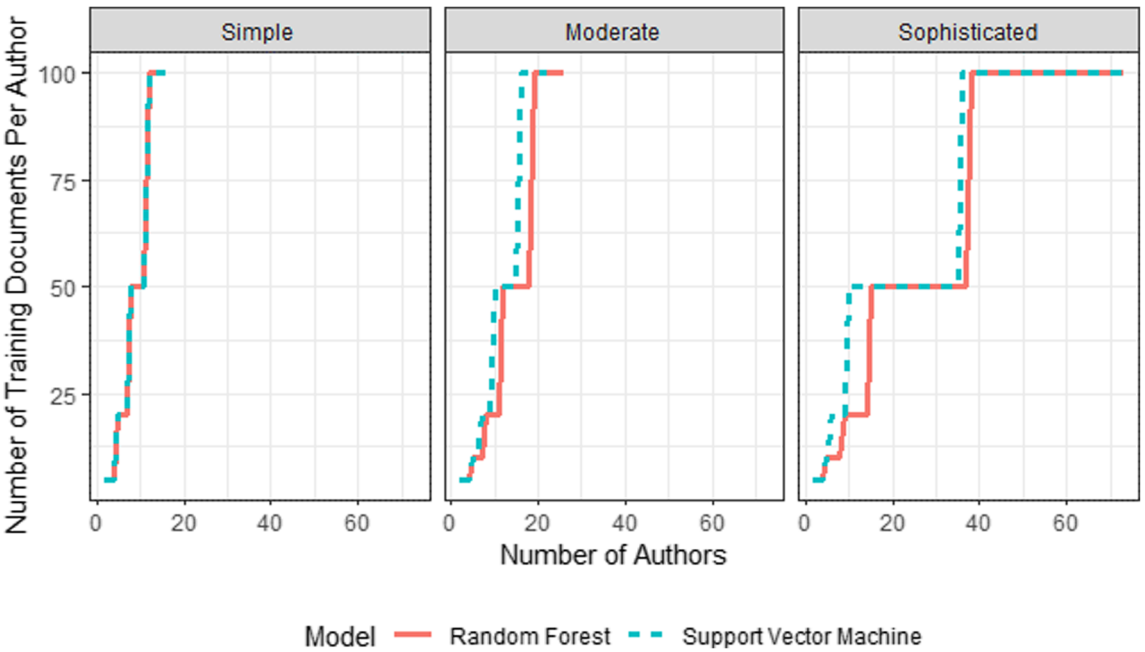

Fig. 3 The number of candidate authors by the amount of training data needed to attribute authorship with at least 50% accuracy

Table 4 Summary of investigated datasets

Dataset	Document type	Number of potential authors	Number of documents	Average word count
Brennan–Greenstadt [10]	Essays	12	113	508.92
Extended Brennan–Greenstadt [10]	Essays	45	699	497.62
US Senator Tweets [73]	Tweets	100	198,265	17.55
Enron Emails (www.nuix.com/enron)	Redacted emails	106	107,623	192.70

through the Amazon Mechanical Turk platform. Using a rigorous quality control protocol, 45 out of 100 authors were kept in the final dataset (see [10], for details). To form the corpora for each author, the writing samples were combined and split into approximately 500-word passages.

The US Senator Tweets data comes from Xia et al. [73] who collected in July 2014 using Twitter's API for official accounts belonging to US Senators. Specifically, accounts in the data are registered to 54 Democratic Senators and 46 Republican Senators. The data consist of approximately 200,000 tweets (documents), including the most recent 3000 from any single user within the set of 100 accounts. While at first glance Twitter may seem like an unlikely platform on which to apply authorship attribution algorithms, the dataset has distinct advantages. First, we illustrate the generality of the methodology by considering a dataset that is still based on user-generated web content but from a different information system. Second, we know the true author of each document, providing an objective testing ground for our general methodology. Third, driven by accusations of organized campaigns on Twitter (and competitor platforms such as Parler) to spread misinformation to influence elections, there is growing interest in methodology to link several distinct Twitter accounts that may be actually controlled by a single entity. Our research has relevance to this problem.

Our fourth dataset is the Enron email corpus, consisting of email messages sent by Enron staff in the course of day-to-day business along with the email's authorship (sender) label from 168 different addresses. The dataset is originally sourced from the US Federal Energy Regulatory Commission's investigation into the collapsed energy firm. The particular version of the data we use comes from a firm Nuix, which volunteered to examine the prevalence of personally identifiable information (PII) within the Enron data. To demonstrate their ability to remove personal information from the text of corporate communications, Nuix published a redacted version of the email dataset, where all personally identifiable information has been removed. Despite their redaction efforts, we find that Enron staff are still identifiable based on their writing style.

Results in Tables 5, 6, and 7 show stable and consistent findings across datasets. As with the Yelp results, we see Random Forest tends to be the most accurate classifier and that accuracy rates can be meaningfully high even with few training documents.

4.5 The Effect of Data Protection on Review Usefulness

Our results presented earlier indicate that a UGC platform should consider carefully which elements of their users' structured data are publicly visible, as it directly relates to user identifiability and viability of the second stage of the authorship attribution algorithm. Yet certain structured data may be critical for fostering trust and enhancing the user experience. In this section, we show that there exists a trade-off for a UGC platform in protecting its users from being identified and having its content provide utility to the general public. Our experimental question is similar to Xu and Ding [74] which performed an online experiment to investigate the effects of sharing sensitive personal information (maximum number of humans a person is willing to harm to save himself/herself) to third parties in response to the regulatory needs of autonomous vehicles. Specifically, using the useful vote count for each review (available for each review in the Yelp dataset) and our authorship attribution methodology, we show that reviews that reveal less structured data are harder to identify and also receive fewer useful votes on average.

With our data, the same structured data fields are posted alongside every review, making it difficult to estimate the impact of changing the available fields. To overcome this challenge, we use structured data revealed within the text itself as a source of variation, allowing us to estimate (albeit imperfectly) the impact of structured data on the review's perceived usefulness. Moreover, beyond an online review platform reducing the amount and granularity of the structured data posted alongside a review, it would also become necessary to protect the review text itself when this information is revealed.

First, we address the question of whether reviewers reveal structured data in the text itself. This is important because more structured data could reduce N_t from the first stage of our authorship attribution algorithm. For instance, the sentences "I've had my scarf for 10 days and I love it! It is warm and stylish for a 25-year old woman in Chicago" would

Table 5 Out of sample accuracy for a "simple" data intruder, averaged over varying number of potential authors

# Training documents	Model	Brennan–Greenstadt corpus	Extended Brennan–Greenstadt corpus	Twitter	Enron
5	Classification Tree	0.373	0.244	0.177	0.254
	Random Forest	0.889*	0.577*	0.250*	0.329
	Reg. Multinomial Regression	0.788	0.492	0.180	0.232
	Support Vector Machine	0.765	0.576	0.223	0.333*
10	Classification Tree	NA	0.362	0.219	0.319
	Random Forest	NA	0.690*	0.295*	0.393*
	Reg. Multinomial Regression	NA	0.633	0.245	0.264
	Support Vector Machine	NA	0.685	0.255	0.379
100	Classification Tree	NA	NA	0.311	0.487
	Random Forest	NA	NA	0.446*	0.686*
	Reg. Multinomial Regression	NA	NA	0.415	0.585
	Support Vector Machine	NA	NA	0.380	0.591

Note: * denotes the best performing method

create the structured data variables "gender," "age," "transaction date," and "location." We randomly select one review from each of 1035 unique reviewers on Yelp and ask MTurk workers yes/no questions about whether the review text contained an age, transaction date within 7 days, gender, race/ethnicity, or specific geographical location. We send each review to three MTurk workers and take the majority share vote. Consistent with best practices, as discussed in Peer et al. [50], we require workers to be Masters (workers who have demonstrated high performance over time) and to be located in the USA. Results are shown in Table 8 with approximately 16% of textual reviews exposing at least one structured

Table 6 Out of sample accuracy for a "moderate" data intruder, averaged over varying number of potential authors

# Training documents	Model	Brennan–Greenstadt corpus	Extended Brennan–Greenstadt corpus	Twitter	Enron
5	Classification Tree	0.357	0.249	0.176	0.253
	Random Forest	0.928*	0.689	0.263*	0.337*
	Reg. Multinomial Regression	0.730	0.450	0.170	0.232
	Support Vector Machine	0.832	0.703*	0.217	0.334
10	Classification Tree	NA	0.385	0.219	0.327
	Random Forest	NA	0.799	0.311*	0.406*
	Reg. Multinomial Regression	NA	0.670	0.222	0.257
	Support Vector Machine	NA	0.824*	0.252	0.392
100	Classification Tree	NA	NA	0.317	0.498
	Random Forest	NA	NA	0.486*	0.700*
	Reg. Multinomial Regression	NA	NA	0.456	0.572
	Support Vector Machine	NA	NA	0.394	0.600

Note: * denotes the best performing method

Table 7 Out of sample accuracy for a "sophisticated" data intruder, averaged over varying number of potential authors

# Training documents	Model	Brennan–Greenstadt corpus	Extended Brennan–Greenstadt corpus	Twitter	Enron
5	Classification Tree	0.355	0.268	0.178	0.286
	Random Forest	0.962*	0.821*	0.313*	0.509
	Reg. Multinomial Regression	0.643	0.495	0.266	0.628*
	Support Vector Machine	0.839	0.807	0.219	0.339
10	Classification Tree	NA	0.466	0.257	0.676
	Random Forest	NA	0.914*	0.385*	0.669
	Reg. Multinomial Regression	NA	0.724	0.362	0.706*
	Support Vector Machine	NA	0.920	0.258	0.406
100	Classification Tree	NA	NA	0.415	0.838
	Random Forest	NA	NA	0.608*	0.901*
	Reg. Multinomial Regression	NA	NA	0.602	0.863
	Support Vector Machine	NA	NA	0.537	0.741

Note: * denotes the best performing method

variable in the text. Location, which could be matched to billing or delivery address, and gender are the two most often revealed pieces of information.

Next, we examine how the amount of structured data revealed in the review text relates to data privacy (probability of identification) and data utility (usefulness of reviews). We create a variable named Text Revelation, defined as the count of structured data revealed in the text. Table 9 shows that as the review text reveals more personal information (i.e., Text Revelation increases), the probability of identification of a reviewer and the average usefulness of reviews increase. Quantiles are reported in the table because the exact value of N_t depends on the uniqueness of the structured data and the level of protection pursued by an online review platform or the author.

Table 8 Proportion of reviews that reveal structured data in the text based on MTurk labeling of 1035 reviews

Statistic	% of reviews
Age revealed	0.2
Transaction date revealed to within 7 days	0.6
Gender revealed	8.2
Race revealed	0.2
Home state or more precise location revealed	11.1
At least one of the above revealed	16.1

With respect to the authorship attribution results, regardless of whether the information is posted alongside a review or in the text itself, as the amount and granularity of structured data decreases, the number of potential authors from the first stage increases. The number of candidate authors from the first stage impacts the ultimate second-stage accuracies. If the review is sufficiently protected to remove all structured data, then the candidate number of authors is so large that authorship analysis becomes computationally infeasible. In contrast, if structured data is revealed, in a majority of cases, the data intruder will have a substantial chance at identifying the author of the target review.

Table 9 also shows that an increase in Text Revelation aligns with an increase in the usefulness of reviews. To test this more rigorously, similar to Krishnamoorthy [37] and Ghose and Ipeirotis [23], we use regression to explain review usefulness with a variety of standard features [58], such as the Flesch–Kincaid reading level, a measure of the readability of the review. Summary statistics for the variables used appear in Table 10. We previously discussed all variables with the exception of polarity (e.g., sentiment or tone), which is measured using a weighted average of words that are tagged by their sentiment (positive, negative, or neutral) according to the dictionary in Hu and Liu [29]. The exact implementation we use is contained in the "qdap" library in R.

Results for three regression models appear in Table 11, where we consistently find that an additional count of structured data aligns with a higher level of usefulness. Therefore, we can surmise that protection policies that reduce the amount

Table 9 Usefulness and probability of identification as a function of the count of structured data that is revealed in the text

Text Revelation	Average useful votes received		Quantile				
			50%	75%	95%	99%	100%
		First and second stage results					
0	0.84	1st stage N_t	1.3 M-M	1.3 M-M	1.3 M-M	1.3 M-M	1.3 M-M
		2nd stage accuracy	0.00	0.00	0.00	0.00	0.00
1	1.01	1st stage N_t	1	2	16	179.01	12,120
		2nd stage accuracy	1.00	0.94	0.68	0.36	0.00
2	1.21	1st stage N_t	1	2	14	150	11,976
		2nd stage accuracy	1.00	0.94	0.69	0.39	0.00
3	4.00	1st stage N_t	1	1	2	4	65
		2nd stage accuracy	1.00	1.00	0.94	0.86	0.51

Note: 1st and 2nd stage results are based on Table 3, Fig. 2, and a sophisticated data intruder with 100 training documents

of structured data would result in less useful reviews — though it is important to note that the effect sizes are relatively small. Reducing Text Revelation by one would lower the average useful count by only 0.2 votes.

Note that these models control for word count, which is important because higher Text Revelation usually implies a longer review. Moreover, the interaction term between Text Revelation and word count reveals that additional structured data in the text becomes slightly less impactful as the review increases in length and readers may focus on other content. We also find that more critical reviews tend to be seen as more useful (negative and significant effect for Review Rating) and that lexical diversity also improves perceived usefulness (vis-à-vis the effect for Yule's *K*).

In summary, the results in this section reinforce the notion that protecting structured data, whether posted alongside a review or within the review text itself, improves reviewer privacy. While there is a negative externality associated with a generalization or redacting structured data, the negative impact on review usefulness appears to be relatively small. This suggests that the benefits outweigh the costs for online review platforms pursuing stronger privacy management practices that reduce privacy concerns.

Table 10 Summary statistics for predictors of review usefulness

Statistic	*N*	Mean	St. Dev.	Min	Max
Useful (DV)	1035	0.879	2.244	0	28
Text Revelation	1035	0.203	0.441	0	3
Word Count	1035	99.014	98.954	15	808
Flesch–Kincaid	1035	5.881	5.055	−1.438	85.392
Review Rating	1035	3.756	1.576	1	5
Polarity	1035	0.465	0.589	−1.568	6.132
Yule's *K*	1035	103.299	54.320	0.000	625.000

5 Conclusions

In this paper, we conducted a detailed analysis of privacy with UGC and online review data in particular. First, we studied the potential to reverse engineer the true author of an anonymous online review. We presented a two-stage authorship attribution algorithm that newly combines structured and textual data wherein a set of potential authors are found by matching the structured demographic information posted alongside or in the online review to a customer database belonging to the party interested in identification. The optional second stage of the algorithm further rank orders the set of candidate authors by analyzing the text with feature engineering and multiclass supervised learning techniques. Using data from Yelp and other platforms, we demonstrated that this two-stage approach is effective in that the chances of identification can be practically meaningful when the target reviewer has distinctive structured data, such as an unusual name, or several structured data values. These results are consistent with previous scholarly work in different domains showing that as few as four spatio-temporal points are enough to de-identify 95% of 1.5 million individuals for mobile phone data [16] and 90% of 1.1 million individuals for credit card transaction data [17]. Sweeney [63] demonstrated that over 87% of individuals in the entire population of the USA are uniquely identifiable based on a

Table 11 Using generalized linear models to explain review usefulness on Yelp. Reviews that reveal more structured data tend to be more useful to the general public

	Logistic		Negative binomial		OLS	
	DV = Useful > 0?		*DV* = Useful		*DV* = log(1 + Useful)	
Text Revelation	0.286***	(0.101)	0.201**	(0.089)	0.070***	(0.027)
Word Count	0.006***	(0.001)	0.006***	(0.001)	0.002***	(0.0003)
Polarity	−0.036	(0.151)	−0.093	(0.128)	−0.034	(0.038)
Yule's *K*	0.003**	(0.001)	0.002**	(0.001)	0.001*	(0.0003)
Flesch–Kincaid	−0.019	(0.016)	−0.003	(0.011)	−0.005	(0.004)
Review Rating	−0.119**	(0.056)	−0.108**	(0.049)	−0.032**	(0.015)
Text Revelation ∗ Word Count	−0.001***	(0.0005)	−0.001**	(0.0004)	−0.0004***	(0.0001)
Constant	−0.974***	(0.268)	−0.723***	(0.236)	0.253***	(0.070)
Observations	1035		1035		1035	
R^2					0.115	
Adjusted R^2					0.109	
Log Likelihood	−633.362		−1207.748			
Theta			0.436***	(0.041)		
Akaike Inf. Crit.	1282.724		2431.497			
Residual Std. Error (*df* = 1027)					0.562	
F Statistic (*df* = 7; 1027)					14.922***	

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

combination of only three variables: their date of birth, ZIP code, and gender. We also studied generalization and redaction as protection policies that can be used by an online review platform to protect the identity of the review. These policies are realistic, technically feasible, and effective, though we uncovered small negative externalities that potentially spillover to the online review ecosystem. Protecting the demographic information of reviewers removed the most predictive elements of authorship and also aligned with a decrease in the usefulness of reviews for the general public.

We make a number of contributions to the marketing, authorship attribution, and privacy literatures. To our knowledge, we are the first to study the lack of privacy and protection in the online reviews domain. We demonstrated how authorship attribution methods and protection policies can impact the different constituents in the UGC ecosystem and found trade-offs and externalities. For example, increased privacy for users by protecting their structured data aligns with reviews that are perceived to be less useful. Protection could also negatively impact customer service, as it becomes harder for a company to respond directly to the dissatisfied consumer. In contrast, a complete lack of protection has led to a proliferation of negative incidents, as highlighted in our introduction. These trade-offs in the UGC ecosystem with respect to privacy, protection, and utility have implications for the management and design of review platforms and merit further study.

Methodologically, beyond systematically combining both structured and text data for authorship attribution and data linkage, another contribution is the performance assessment of the second stage of the algorithm which is stratified by the amount of effort put into feature engineering and when training data is scarce rather than plentiful. We found that as few as five training documents (or, equivalently, hundreds of words per author) can be enough to attribute authorship at a level several multiples higher than random guessing, even when the number of potential authors is large. To demonstrate that our results have relevance beyond the online review setting, we provided results from applying the algorithm to four other datasets that vary in terms of the document type, length, formality, and style of writing. We demonstrate consistent results using Tweets, corporate emails with redacted PII, and long-form student essays.

Our work is subject to limitations and caveats. First, we acknowledge that our results demonstrate reviewers are, on average, identifiable to a certain accuracy level. For an author who strategically writes a single review under a random username with completely random demographics, the only information that can reliably be used to identify the author is embedded in the text. We do not address this situation directly; however, our results do yield insights into this phenomenon. For example, for such a review to be considered helpful by others, the author may need to write a detailed review to the extent that key covariates are revealed which can be used for identification. Further, many online reviews and UGC platforms actively persuade users to include personal information to enhance data utility for the general public. As Yelp's [75] user agreement states, "You can reduce the risk of being

personally identified by using the Service pseudonymously, though doing so could detract from the credibility of your contributions to the Service."

An important area of future work is extending the authorship attribution methodology to further inform protection policies regarding other forms of unstructured data. For example, photos are often deleted for privacy concerns [67] and it is becoming increasingly common for users to post pictures of themselves or a product alongside their review. These photographs, through simple reverse image searches or more sophisticated computer vision algorithms, can potentially reveal a tremendous amount of information that helps identify the supposedly anonymous author. For example, automated image analysis can estimate demographic information (age, gender, etc.). Although the current implementation of these techniques requires a high degree of technical sophistication, we expect this to become an important extension to the proposed methodology as computer vision implementations streamline and associated costs decline.

Appendix. Expanded results for the yelp data

Figure 4 provides out-of-sample accuracy results for the Yelp data. Accuracy consistently improves as the sophistication of the data intruder and the size of the training data increase.

Acknowledgements We are thankful to Elea Feit, Sachin Gupta, Cameron Bale, and Sharan Jagpal for their helpful comments on earlier versions of this paper.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Abbasi A, Chen H (2008) Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transact Inform Syst (TOIS) 26(2):1–29
2. Abbasi A, Chen H, Nunamaker JF (2008) Stylometric identification in electronic markets: scalability and robustness. J Manag Inf Syst 25(1):49–78
3. Aggarwal CC, Philip SY (2008) A general survey of privacy-preserving data mining models and algorithms. In: In Privacy-preserving data mining. Springer, Boston, pp 11–52
4. Ahn D-Y, Duan JA, Mela CF (2015) Managing user-generated content: a dynamic rational expectations equilibrium approach. *Mark Sci* 35(2):284–303
5. Almishari M, Tsudik G (2012) Exploring linkability of user reviews. In: In European Symposium on Research in Computer Security. Springer, Berlin, pp 307–324
6. AMZ Tracker, 2018. How to deal with negative reviews. URL: https://www.amztracker.com/blog/deal-negative-reviews/. Accessed: July 24, 2020.
7. André Q, Carmon Z, Wertenbroch K, Crum A, Frank D, Goldstein W, Huber J, Van Boven L, Weber B, Yang H (2018) Consumer choice and autonomy in the age of artificial intelligence and big data. Cust Needs Solut 5(1):28–37
8. Bordes A, Ertekin S, Weston J, Bottou L (2005) Fast kernel classifiers with online and active learning. J Mach Learn Res 6(Sep): 1579–1619
9. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
10. Brennan M, Afroz S, Greenstadt R (2012) Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity. ACM Transac Inform Syst Secur (TISSEC) 15(3):1–22
11. Brizan DG, Tansel AU (2006) A. survey of entity resolution and record linkage methodologies. Commun IIMA 6(3):5
12. Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Mark Sci* 35(6):953–975
13. Campbell J, Goldfarb A, Tucker C (2015) Privacy regulation and market structure. J Econ Manag Strateg 24(1):47–73
14. Caselaw, (2017). ZL TECHNOLOGIES INC v. GLASSDOOR INC. Court of Appeal, First District, Division 4, California. URL: https://caselaw.findlaw.com/ca-court-of-appeal/1868279.html. Accessed July 24, 2020.
15. De Jong MG, Pieters R, Fox JP (2010) Reducing social desirability bias through item randomized response: an application to measure underreported desires. J Mark Res 47(1):14–27
16. De Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3(1):1376
17. De Montjoye YA, Radaelli L, Singh VK (2015) Unique in the shopping mall: on the reidentifiability of credit card metadata. Science 347(6221):536–539
18. Douglas DM (2016) Doxing: a conceptual analysis. Ethics Inf Technol 18(3):199–210
19. Du Bay WH, (2004). The principles of readability. Accessed April 7, 2020. http://en.copian.ca/library/research/readab/readab.pdf.
20. Elmagarmid AK, Ipeirotis PG, Verykios VS (2006) Duplicate record detection: a survey. IEEE Trans Knowl Data Eng 19(1):1–16
21. Farr C, (2018). Facebook sent a doctor on a secret mission to ask hospitals to share data. CNBC. URL: https://www.cnbc.com/2018/04/05/facebook-building-8-explored-data-sharing-agreement-with-hospitals.html. Accessed: July 24, 2020.
22. Getoor L, Machanavajjhala A (2012) Entity resolution: theory, practice & open challenges. Proc VLDB Endowment 5(12):2018–2019
23. Ghose A, Ipeirotis PG (2010) Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. IEEE Trans Knowl Data Eng 23(10):1498–1512
24. Goldfarb A, Tucker C (2013) Why managing consumer privacy can be an opportunity. MIT Sloan Manag Rev 54(3):10
25. Gravano L, Ipeirotis PG, Koudas N and Srivastava D, (2003). Text joins for data cleansing and integration in an rdbms. In Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405) (pp. 729-731). IEEE.
26. Hewett K, Rand W, Rust RT, van Heerde HJ (2016) Brand buzz in the echoverse. *J Mark* 80(3):1–24
27. Hill S, Provost F (2003) The myth of the double-blind review? Author identification using only citations. Acm Sigkdd Explor Newslett 5(2):179–184
28. Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425
29. Hu M and Liu B, 2004. Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).
30. Jones R, (2017). Court rules Yelp must identify anonymous user in defamation case. Gizmodo. URL: https://gizmodo.com/court-rules-yelp-must-identify-anonymous-user-in-defama-1820433103. Accessed: July 24, 2020.

31. Juola P (2012) Large-scale experiments in authorship attribution. Engl Stud 93(3):275–283
32. Juola P and Vescovi D, (2010). Empirical evaluation of authorship obfuscation using JGAAP. In Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security (pp. 14-18).
33. Kincaid JP, Fishburne Jr RP, Rogers RL and Chissom BS, (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch.
34. Klemko R (2021) A small group of sleuths has been identifying right-wing extremists long before the attack on the Capitol. URL: https://www.washingtonpost.com/national-security/antifa-far-right-doxing-identities/2021/01/10/41721de0-4dd7-11eb-bda4-615aaefd0555_story.html. Accessed January 2, 2021.
35. Koppel M, Schler J, Argamon S (2009) Computational methods in authorship attribution. J Am Soc Inf Sci Technol 60(1):9–26
36. Kourtis I, Stamatatos E (2011) Author identification using semi-supervised learning. In: In CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers). The Netherlands, Amsterdam
37. Krishnamoorthy S (2015) Linguistic features for review helpfulness prediction. Expert Syst Appl 42(7):3751–3759
38. Kroft S, (2014). The data brokers: selling your personal information. 60 Minutes. URL: https://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/. Accessed: July 24, 2020.
39. Kumar V, Reinartz W (2018) Customer privacy concerns and privacy protective responses. In: In Customer relationship management. Springer, Berlin, pp 285–309
40. Li XB, Qin J (2017) Anonymizing and sharing medical text records. Inf Syst Res 28(2):332–352
41. Li XB, Sarkar S (2006) Privacy protection in data mining: a perturbation approach for categorical data. Inf Syst Res 17(3):254–270
42. Mankad S, Han HS, Goh J, Gavirneni S (2016) Understanding online hotel reviews through automated text analysis. Serv Sci 8(2):124–138
43. Martin KD, Murphy PE (2017) The role of data privacy in marketing. J Acad Mark Sci 45(2):135–155
44. Menon S, Sarkar S (2016) Privacy and big data: scalable approaches to sanitize large transactional databases for sharing. MIS Q 40(4):963–981
45. Moe WW, Schweidel DA (2012) Online product opinions: incidence, evaluation, and evolution. *Mark Sci* 31(3):372–386
46. Narayanan A, Paskov H, Gong NZ, Bethencourt J, Stefanov E, Shin ECR and Song D, (2012). On the feasibility of internet-scale author identification. In 2012 IEEE Symposium on Security and Privacy (pp. 300-314). IEEE.
47. Narayanan A and Shmatikov V, 2008, May. Robust de-anonymization of large datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy.
48. The Associated Press, (2017). Yelp says lawsuit might eliminate all negative reviews. *New York Daily News*. URL: https://www.nydailynews.com/news/national/yelp-lawsuit-eliminate-negative-reviews-article-1.2796087. Accessed July 24, 2020.
49. Payer M, Huang L, Gong NZ, Borgolte K, Frank M (2014) What you submit is who you are: a multimodal approach for deanonymizing scientific publications. IEEE Transact Inform Forensics Secur 10(1):200–212
50. Peer E, Vosgerau J, Acquisti A (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behav Res Methods 46(4):1023–1031
51. Porter J, (2019). Fraudulent Yelp posting protected under the law, ridiculous. Tahoe Daily Tribune, May 20, 2019. URL: https://www.tahoedailytribune.com/news/jim-porter-fraudulent-yelp-posting-protected-under-the-law-ridiculous/. Accessed July 24, 2020.
52. Proserpio D, Zervas G (2017) Online reputation management: estimating the impact of management responses on consumer reviews. *Mark Sci* 36(5):645–665
53. Qian T, Liu B, Chen L and Peng, Z., (2014). Tri-training for authorship attribution with limited training data. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 345-351).
54. Rochet JC, Tirole J (2003) Platform competition in two-sided markets. J Eur Econ Assoc 1(4):990–1029
55. Schneider MJ, Jagpal S, Gupta S, Li S, Yu Y (2017) Protecting customer privacy when marketing with second-party data. *Int J Res Mark* 34(3):593–603
56. Schneider MJ, Jagpal S, Gupta S, Li S, Yu Y (2018) A flexible method for protecting marketing data: an application to point-of-sale data. *Mark Sci*. ePub ahead of print Jan 8 37:153–171. https://doi.org/10.1287/mksc.2017.1064
57. Shu K, Wang S, Tang J, Zafarani R, Liu H (2017) User identity linkage across online social networks: a review. Acm Sigkdd Explor Newslett 18(2):5–17
58. Singh JP, Irani S, Rana NP, Dwivedi YK, Saumya S, Roy PK (2017) Predicting the "helpfulness" of online consumer reviews. J Bus Res 70:346–355
59. Snyder P, Doerfler P, Kanich C and McCoy D, (2017). Fifteen minutes of unwanted fame: detecting and characterizing doxing. In Proceedings of the 2017 Internet Measurement Conference (pp. 432-444).
60. Stamatatos E (2009) A survey of modern authorship attribution methods. J Am Soc Inf Sci Technol 60(3):538–556
61. Steyvers M, Griffiths T (2007) Probabilistic topic models. Handb Latent Semantic Anal 427(7):424–440
62. Stone EF, Spool MD, Rabinowitz S (1977) Effects of anonymity and retaliatory potential on student evaluations of faculty performance. Res High Educ 6(4):313–325
63. Sweeney L (2000) Simple demographics often identify people uniquely. Health (San Francisco) 671(2000):1–34
64. Sweeney L (2002a) k-anonymity: a model for protecting privacy. Int J Uncertaint Fuzziness Knowl-Based Syst 10(05):557–570
65. Sweeney L (2002b) Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertaint Fuzziness Knowl-Based Syst 10(05):571–588
66. Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation. *J Mark Res* 51(4):463–479
67. Turjeman D and Feinberg FM, (2019). When the data are out: measuring behavioral changes following a data breach. Available at SSRN 3427254.
68. Tweedie FJ, Baayen RH (1998) How variable may a constant be? Measures of lexical richness in perspective. Comput Hum 32(5):323–352
69. US Census Bureau, (2016). Decennial Census Surname Files (2010, 2000). URL: https://www.census.gov/data/developers/data-sets/surnames.html. Accessed July 24, 2020.
70. US Social Security Administration, (2019). Baby names from social security card applications - national data. Data.gov. URL: https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data. Accessed: July 24, 2020.
71. Wedel M, Kannan PK (2016) Marketing analytics for data-rich environments. *J Mark* 80(6):97–121
72. Winkler WE, (1999). The state of record linkage and current research problems. In Statistical Research Division, US Census Bureau.
73. Xia D, Mankad S, Michailidis G (2016) Measuring influence of users in Twitter ecosystems using a counting process modeling framework. Technometrics 58(3):360–370

74. Xu J, Ding M (2019) Using the double transparency of autonomous vehicles to increase fairness and social welfare. Cust Needs Solut 6(1):26–35
75. Yelp, 2020. https://terms.yelp.com/privacy/en_us/20200101_en_us/#Controlling-Your-Personal-Data. .
76. Yule, G.U., 1944. The statistical study of literary vocabulary. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 42, pp. b1-9).
77. Zhang Y, Moe WW, Schweidel DA (2017) Modeling the role of message content and influencers in social media rebroadcasting. *Int J Res Mark* 34(1):100–119
78. Zhao Y, Yang S, Narayan V, Zhao Y (2013) Modeling consumer learning from online product reviews. *Mark Sci* 32(1):153–169