Fighting Authorship Linkability with Crowdsourcing

Mishari Almishari
College of Computer and Information Sciences
King Saud University
mialmishari@ksu.edu.sa

Ekin Oguz
Computer Science Department
University of California, Irvine
eoguz@uci.edu

Gene Tsudik
Computer Science Department
University of California, Irvine
gts@ics.uci.edu

ABSTRACT

Massive amounts of contributed content – including traditional literature, blogs, music, videos, reviews and tweets – are available on the Internet today, with authors numbering in many millions. Textual information, such as product or service reviews, is an important and increasingly popular type of content that is being used as a foundation of many trendy community-based reviewing sites, such as TripAdvisor and Yelp. Some recent results have shown that, due partly to their specialized/topical nature, sets of reviews authored by the same person are readily linkable based on simple stylometric features. In practice, this means that individuals who author more than a few reviews under different accounts (whether within one site or across multiple sites) can be linked, which represents a significant loss of privacy.

In this paper, we start by showing that the problem is actually worse than previously believed. We then explore ways to mitigate authorship linkability in community-based reviewing. We first attempt to harness the global power of crowdsourcing by engaging random strangers into the process of re-writing reviews. As our empirical results (obtained from Amazon Mechanical Turk) clearly demonstrate, crowdsourcing yields impressively sensible reviews that reflect sufficiently different stylometric characteristics such that prior stylometric linkability techniques become largely ineffective. We also consider using machine translation to automatically re-write reviews. Contrary to what was previously believed, our results show that translation decreases authorship linkability as the number of intermediate languages grows. Finally, we explore the combination of crowdsourcing and machine translation and report on results.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

COSN'14, October 1–2, 2014, Dublin, Ireland.

ACM 978-1-4503-3198-2/14/10 ...$15.00.
http://dx.doi.org/10.1145/2660460.2660486.

Keywords

authorship attribution; author linkability; author identification; author anonymization; crowdsourcing; stylometry

1. INTRODUCTION

The Internet has become a tremendous world-wide bazaar where massive amounts of information (much of it of dubious quality and value) are being disseminated and consumed on a constant basis. Sharing of multimedia content is one of the major contributors to Internet's growth and popularity. Another prominent source of shared information is textual, e.g., blogs, tweets and various discussion fora. Among those, community reviewing has carved out an important niche. This category includes well-known sites, such as: Yelp, CitySearch, UrbanSpoon, Google Places and TripAdvisor. There are also many others that include customer-based reviewing as a side-bar, e.g., Amazon or Ebay.

Regardless of their primary mission and subject coverage, community reviewing sites are popular, since many are free and contain lots of useful content voluntarily contributed by multitudes of regular people who document their experience with products, services, destinations, and attractions. Larger sites, e.g., TripAdvisor and Yelp, have tens of millions of users (readers) and millions of contributors [9].

Certain features distinguish community reviewing sites from other contributory Internet services:

- Discussion Fora: these vary from product or topic discussions to comment sections in on-line news media. They are often short and not very informative (even hostile).
- Body of Knowledge: the best-known and most popular example is Wikipedia – a huge amalgamation of communal knowledge on a very wide range of subjects. However, unlike reviewing sites where each review is atomic and discernable, related contributions to body-of-knowledge sites are usually mashed together, thus (by design) obscuring individual prose.
- Online Social Networks (OSNs): such sites are essentially free-for-all as far as the type and the amount of contributed information. Since most OSNs restrict access to content provided by a user to "friends" (or "colleagues") of that user, opinions and reviews do not propagate to the rest of Internet users.

Some recent work [20] has shown that many contributors to community reviewing sites accumulate a body of authored content that is sufficient for creating their stylometric profiles, based on rather simple features (e.g., digram frequency). A stylometric profile allows probabilistic linkage

among reviews generated by the same person. This could be used to link reviews from different accounts (within a site or across sites) operated by the same user. On one hand, tracking authors of spam reviews can be viewed as a useful service. On the other hand, the ease of highly accurate linkage between different accounts is disconcerting and ultimately detrimental to privacy. Consider, for example, a vindictive merchant who, offended by reviews emanating from one account, attempts to link it to other accounts held by the same person, e.g., for the purpose of harrassment. We consider both sides of this debate to be equally valid and do not choose sides. However, we believe that the privacy argument deserves to be considered, which triggers the motivation for this paper:

What can be done to mitigate linkability of reviews authored by the same contributor?

Roadmap:.

Our goal is to develop techniques that mitigate review account linkability. To assess efficacy of proposed techniques, we need accurate review linkage models. To this end, we first improve state-of-art author review linkage methods. We construct a specific technique, that offers 90% accuracy, even for a small number of identified reviews (e.g., 95) and a smaller set (e.g., 5) of anonymous reviews.

Our second direction is the exploration of techniques that decrease authorship linkability. We start by considering crowdsourcing, which entails engaging random strangers in rewriting reviews. As it turns out, our experiments using Amazon MTurk [1] clearly demonstrate that authorship linkability can be significantly inhibited by crowdsourced rewriting. Meanwhile, somewhat surprisingly, crowd-rewritten reviews remain meaningful and generally faithful to the originals. We then focus on machine translation tools and show that, by randomly selecting languages to (and from) which to translate, we can substantially decrease linkability.

Organization:.

The next section summarizes related work. Then, Section 3 overviews some preliminaries, followed by Section 4 which describes the experimental dataset and review selection process for subsequent experiments. Next, Section 5 discusses our linkability study and its outcomes. The centerpiece of the paper is Section 6, which presents crowdsourcing and translation experiments. It is followed by Section 7 where we discuss possible questions associated with the use of crowdsourcing. Finally, summary and future work appear in Section 8.

2. RELATED WORK

Related work generally falls into two categories: Authorship Attribution/Identification and Author Anonymization.

Authorship Attribution:.

There are many studies in the literature. For example, [20] shows that many Yelp's reviewers are linkable using only very simple feature set. While the setting is similar to ours, there are some notable differences. First, we obtain high linkability using very few reviews per author. Second, we only rely on features extracted from review text. A study of blog posts achieves 80% linkability accuracy [22]. Author identification is also studied in the context of academic paper reviews achieving accuracy of 90% [21]. One major difference between these studies and our work is that we use reviews, which are shorter, less formal and less restrictive in choice of words than blogs and academic papers. Abbasi and Chen propose a well-known author attribution technique based on Karhunen-Loeve transforms to extract a large list of Writeprint features (assessed in Section 5) [10]. Lastly, Stamatatos provides a comprehensive overview of authorship attribution studies [26].

Author Anonymization:.

There are several well-known studies in author anonymization [24, 17, 19]. Rao and Rohatgi are among the first to address authorship anonymity by proposing using round-trip machine translation, e.g., English → Spanish → English, to obfuscate authors [24]. Other researchers apply round-trip translation, with a maximum of two intermediate languages and show that it does not provide noticeable anonymizing effect [15, 13]. In contrast, we explore effects (on privacy) of increasing and/or randomizing the number of intermediate languages.

Kacmarcik and Gamon show how to anonymize documents via obfuscating writing style, by proposing adjustment to document features to reduce the effectiveness of authorship attribution tools [17]. The main limitation of this technique is that it is only applicable to authors with a fairly large text corpus, whereas, our approach is applicable to authors with limited number of reviews.

Other practical-counter-measures for authorship recognition techniques such as obfuscation and imitation attacks are explored [14]. However, it is shown that such stylistic deception can be detected with 96.6% accuracy [11].

The most recent relevant work is Anonymouth [19] – a framework that captures the most effective features of documents for linkability and identifies how these feature values should be changed to achieve anonymization. Our main advantage over Anonymouth is usability. Anonymouth requires the author to have two additional sets of documents, on top of the original document to be anonymized: 1) sample documents written by the same author and 2) a corpus of sample documents written by other authors. Whereas, our approach does not require any such sets.

3. BACKGROUND

This section overviews stylometry, stylometric characteristics and statistical techniques used in our study.

Merriam-Webster dictionary defines `Stylometry` as: *the study of the chronology and development of an author's work based especially on the recurrence of particular turns of expression or trends of thought* [7]. We use stylometry in conjunction with the following two tools:

Writeprints feature set: well-known stylometric features used to analyze author's writing style.

Chi-Squared test: a technique that computes the distance between each author's review in order to assess linkability.

3.1 Writeprints

Writeprints is essentially a combination of static and dynamic stylometric features that capture lexical, syntactic, structural, content and idiosyncratic properties of a given body of text [10]. Some features include:

- Average Character Per Word: Total number of characters divided by total number of words.
- Top Letter Trigrams: Frequency of contiguous sequence of 3 characters, e.g. $aaa, aab, aac, ..., zzy, zzz$. There are 17576 ($26^3$) possible permutation of letter trigrams in English.
- Part of Speech (POS) Tag Bigrams: POS tags are the mapping of words to their syntactic behaviour within sentence, e.g. noun or verb. POS tag bigrams denotes 2 consecutive parts of speech tags. We used Stanford POS Maxent Tagger [27] to label each word with one of 45 possible POS tags.
- Function Words: Set of 512 common words, e.g. *again*, *could*, *himself* and etc, used by Koppel et al. in Koppel, 2005.

Writeprints has been used in several stylometric studies [10, 22, 21]. It has been shown to be an effective means for identifying authors because of its capability to capture even smallest nuances in writing.

We use Writeprints implementation from JStylo – a Java library that includes 22 stylometric features [19].

3.2 Chi-Squared Test

Chi-Squared (CS) test is used to measure the distance between two distributions [25]. For any two distributions P and Q, it is defined as:

$$CS_d(P,Q) = \sum_i \frac{(P(i)-Q(i))^2}{P(i)+Q(i)}$$

CS_d is a symmetric measure, i.e., $CS_d(P,Q) = CS_d(Q,P)$. Also, it is always non-negative; a value of zero denotes that P and Q are identical distributions. We employ Chi-Squared test to compute the distance between contributor's anonymous and identified reviews.

4. LINKABILITY STUDY PARAMETERS

This section describes the dataset and the problem setting for subsequent linkability analysis.

4.1 Dataset

We use a large dataset of reviews from Yelp[1] that contains $1,076,850$ reviews authored by $1,997$ distinct contributors. We selected this particular dataset for two reasons:

1. Large number of authors with widely varying numbers of reviews: average number of reviews per author is 539, with a standard deviation of 354.
2. Relatively small average review size – 149 words – which should make linkability analysis more challenging.

4.2 Problem Setting

Informally, our main goal is to link a given set of anonymous set reviews R to a set of identified reviews, perhaps with a known author.[2] The problem is more challenging when the sets of anonymous and identified reviews are relatively small. The exact problem setting is as follows:

We first select 40 authors at random. Although this is a relatively small number, we pick it in order to make subsequent crowdsourcing experiments feasible, as described in Section 6. Then, we randomly shuffle each author's reviews and select the first N. Next, we split selected reviews into two sets:

- First X reviews form the **Anonymous Record** (AR) set. We experiment with AR sets of varying sizes.
- Subsequent $(N-X)$ reviews form the **Identified Record** (IR) set.

Our problem is then reduced to linking ARs to their corresponding IRs. We set N=100 and vary X from 1 to 5. This makes IRs and ARs quite small compared to an average of 539 reviews per author in the original dataset. As a result, the linking problem becomes very challenging.

Next, we attempt to link ARs to their corresponding IRs. Specifically, for each AR, we rank – in descending order of likelihood – all possible authors, i.e., IRs. Then, the top-ranked IR (author) is the one most similar to the given AR. If the correct author is among top-ranked T IRs, we say that the linking model has a hit; otherwise, it is a miss. For a given value of T, the fraction of hits of all ARs (over the total of 40) is referred as Top-T linkability ratio (LR). The linkability analysis boils down to finding a model that maximizes LR for different T and AR sizes. We consider two integer values of T: 1 denotes a perfect-hit and 4 stands for an almost-hit.

LR	Linkability Ratio
AR	Anonymous Records
IR	Identified Records
CS	Chi-Squared Distance Model
F	A feature
F_T	The set of tokens in feature F
S_F	Set of selected features
WP_i	Writeprint feature i
WP_{all}	Combination of all Writeprints
$CS_d(IR, AR)$	CS distance between IR and AR

Table 1: Notation and abbreviations

5. LINKABILITY ANALYSIS

We first apply a subset of the popular Writeprints feature set[3] to convert each AR and IR into a token set. We then use Chi-Square[4] to compute distances between these token sets.

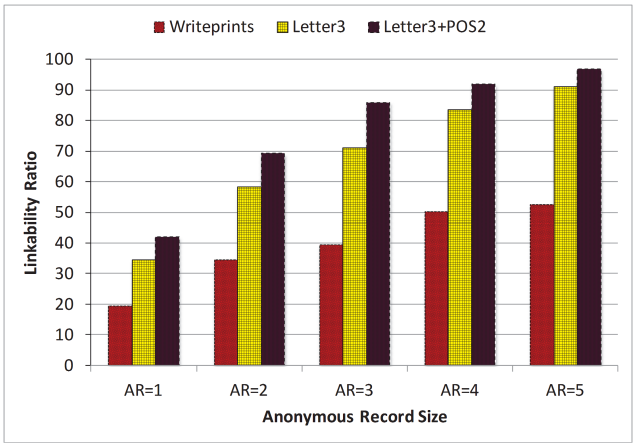
We now describe experimental methodology in more detail. Notation and abbreviations are reflected in Table 1.

[1]See: `www.yelp.com`.

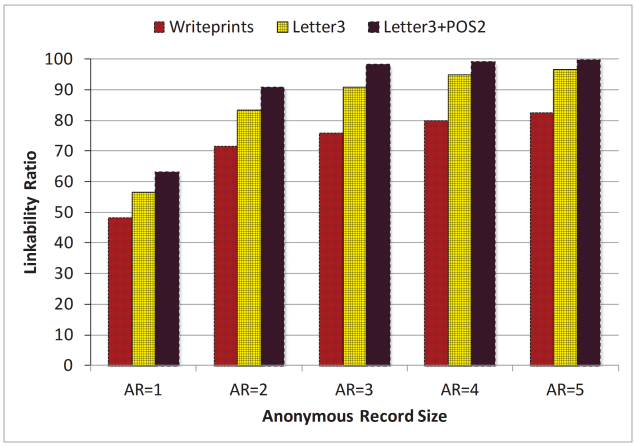
[2]Identification of the author might not be the only goal. It might suffice to simply link two disparate bodies of reviews, thus establishing that they were authored by the same person.

[3]We initially experimented with the Basic-9 feature set, which is known to provide useful information for author identification for less than 10 potential authors [13]. However, its performance was really poor, since we have 40 authors in the smallest set.

[4]We tried others tests including: Cosine, Euclidean, Manhattan, and Kullback-Leibler Divergence. However, Chi-Squared Test outperformed them all.



(a) Top-1



(b) Top-4

Figure 1: LRs of Writeprints, Letter3 and Letter3+POS2

5.1 Methodology

First, we tokenize each AR and IR sets using every feature – F – in our set of selected features – S_F – to obtain a set of tokens $F_T = \{F_{T_1}, F_{T_2}, ..., F_{T_n}\}$, where F_{T_i} denotes the i-th token in F_T. Then, we compute distributions for all tokens. Next, we use CS model to compute the distance between AR and IR using respective token distributions. Specifically, to link AR with respect to some feature F, we compute CS_d between the distribution of tokens in F_T for AR and the distribution of tokens in F_T for each IR. Next, we sort the distances in ascending order of $CS_d(IR, AR)$ and return the resulting list. First entry corresponds to the IR with the closest distance to AR, i.e., the most likely match. For the sake of generality, we repeat this experiment 3 times, randomly picking different AR and IR sets each time. Then, we average the results. Note that S_F is initially empty and features are gradually added to it, as described next.

5.2 Feature Selection

We use a general heuristics – a version of greedy hill-climbing algorithm – for feature selection [23]. The main idea is to identify most influential features and gradually combine them in S_F, until encountering a high LR.

5.2.1 WP_{all}

As a benchmark, we start with setting S_F to WP_{all}, which combines all 22 Writeprint features. We compute LR using WP_{all} in CS model with $|AR| = 5$. Unfortunately, WP_{all} results in low LRs – only 52.5% in Top-1 and 82.5% in Top-4. We believe that, because of the small AR set, the combination of many features increases noise, which, in turn, lowers linkability.

5.2.2 *Improving* WP_{all}

Next, we use each feature from WP_{all} individually, i.e., we try each WP_i with $|AR| = 5$. Table 2 shows the best five features together with WP_{all} after ranking LRs in Top-1 and Top-4. First five features perform significantly better than all others, especially, better than WP_{all} which landed in 9-th place. Interestingly, LR increases drastically – from 52.5% to 91% in Top-1 – with the best feature. Since Top Letter Trigrams performs best individually, we add it to S_F. Then we proceed to considering the combination of other four features with Top Letter Trigrams.

		Linkability Ratio	
Ranking	**Feature**	**Top-1(%)**	**Top-4(%)**
1	Top Letter Trigrams	91	96
2	POS Bigrams	89	96
3	Top Letter Bigrams	86	94
4	Words	79	94
5	POS Tags	78	90
9	WP_{all}	52.5	82.5

Table 2: LRs of best five Writeprint features individually and WP_{all}, with $|AR| = 5$

5.2.3 *Improving Top Letter Trigrams*

Next, we combine each feature from the set {POS Bigrams, Top Letter Bigrams, Words, POS Tags} with S_F to see whether it yields a higher LR. It turns out that combining POS Bigrams yields the best LR gain: from 91% to 96% in Top-1, and 96% to 100% in Top-4. Since we achieve 100% LR in Top-4, we set S_F as {Top Letter Trigrams, POS Bigrams}.

Figures 1(a) and 1(b). show LR comparisons of experimented features with varying AR sizes. For all AR sizes, there is a significant improvement with Top Letter Trigrams over Writeprints. A similar trend occurs with {Top Letter Trigrams, POS Bigrams} over only Top Letter Trigrams in both Top-1 and Top-4.

5.3 Scalability of the Linkability Technique

So far, our small-scale study assessed linkability of 40 authors within a set of 40 possible authors. This is partly because computation of WP_{all} in bigger author sets is very expensive. However, 40 is a very small number in a real-world scenario. Therefore, we need to verify that high LRs identified with S_F still hold for larger number of possible authors. To this end, we vary author set size between 40 and 1000. In particular, we consider set sizes of $[40, 100, 250, 500, 750, 1000]$ authors. In each experiment, we assess linkability of 40 authors when mixing them with others.

Figure 2 shows Top-1 and Top-4 LR of S_F with $|AR| = 5$. Our preferred selection of features – Top Letter Trigram and

POS Bigrams – achieves high LR, 77.5% in Top-1 and 90% in Top-4, even in a set of 1000 possible authors.

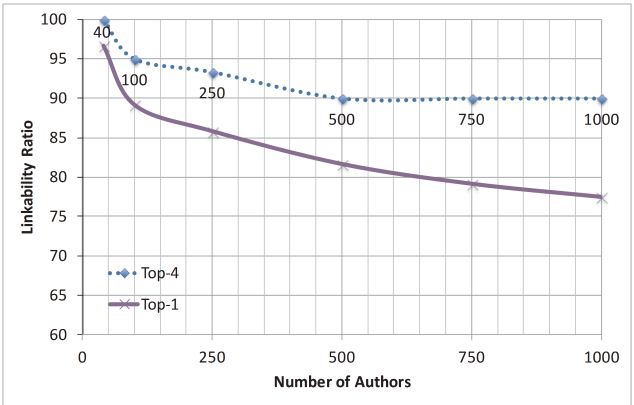


Figure 2: LRs of Letter3+POS2 in varying size of author sets

5.4 Summary

To summarize, key results are:

1. Starting with a well-known Writeprints feature set we achieved modest LRs of up to 52.5% in Top-1 and 82.5% in Top-4 using the CS model. (See Section 5.2.1)

2. Then, we tried each Writeprint feature individually with the intuition that the combination of multiple features would have more noise, thus decreasing linkability. Surprisingly, using only Top Letter Trigrams or POS Bigrams, we achieved significantly better LR than all Writeprints features. (See Section 5.2.2)

3. Next, we selected Top Letter Trigrams, which yields 91% and 96% LR in Top-1 and Top-4, as our rising main. Then, we increased linkability to 96% in Top-1, and 100% in Top-4 by adding POS Bigrams. (See Section 5.2.3)

4. Even when assessing linkability within a large number of possible authors sets, the preferred combination of features maintains high LR, e.g. 77.5% in Top-1 and 90% in Top-4 among 1000 possible authors (See Section 5.3). Thus, we end up setting S_F as {Top Letter Trigrams, POS Bigrams}, which will be used for evaluation of anonymization techniques.

6. FIGHTING AUTHORSHIP LINKABILITY

We now move on to the main topic of this paper: techniques that could mitigate authorship linkability. We consider two general approaches:

1. Crowdsourcing: described in Section 6.1.

2. Machine Translation: described in Section 6.2.

6.1 Crowdsourcing to the Rescue

We begin by considering what it might take, in principle, to anonymize reviews. Ideally, an anonymized review would exhibit stylometric features that are not linkable, with high accuracy, to any other review or a set thereof. At the same time, an anonymized review must be as meaningful as the original review and must remain faithful or "congruent" to it. (We will come back to this issue later in the paper). We believe that such perfect anonymization is probably impossible. This is because stylometry is not the only means of linking reviews. For example, if a TripAdvisor contributor travels exclusively to Antarctica and her reviews cover only specialized cruise-ship lines and related products (e.g., arctic-quality clothes), then no anonymization technique can prevent linkability by topic without grossly distorting the original review. Similarly, temporal aspects of reviews might aid linkability[5]. Therefore, we do not strive for perfect anonymization and instead confine the problem to the more manageable scope of reducing stylometric linkability. We believe that this degree of anonymization can be achieved by rewriting.

6.1.1 How to Rewrite Reviews?

There are many ways of rewriting reviews in order to reduce or obfuscate stylometric linkability. One intuitive approach is to construct a piece of software, e.g., a browser plug-in, that alerts the author about highly linkable features in the prospective review. This could be done in real time, as the review is being written, similarly to a spell-checker running in the background. Alternatively, the same check can be done once the review is fully written. The software might even proactively recommend some changes, e.g., suggest synonyms, and partition long, or join short, sentences. In general, this might be a viable and effective approach. However, we do not pursue it in this paper, partly because of software complexity and partly due to the difficulty of conducting sufficient experiments needed to evaluate it.

Our approach is based on a hypothesis that the enormous power of global crowd-sourcing can be leveraged to efficiently rewrite large numbers of reviews, such that:

(1) Stylometric authorship linkability is appreciably reduced, and

(2) Resulting reviews remain sensible and faithful to the originals.

The rest of this section overviews crowdsourcing, describes our experimental setup and reports on the results.

6.1.2 Crowdsourcing

Definition: according to the Merriam-Webster dictionary, `Crowdsourcing` is defined as: *the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers* [7].

There are numerous crowdsourcing services ranging in size, scope and popularity. Some are very topical, such as `kickstarter` (creative idea/project funding) or `microworkers` (web site promotion), while others are fairly general, e.g., `taskrabbit` (off-line jobs) or `clickworker` (on-line tasks). We selected the most popular and the largest general crowdsourcing service – Amazon's Mechanical Turk (MTurk) [1]. This choice was made for several reasons:

- We checked the types of on-going tasks in various general crowdsourcing services and MTurk was the only one where we encountered numerous on-going text rewriting tasks.

[5]Here we mean time expressed (or referred to) within a review, not only time of posting of a review.

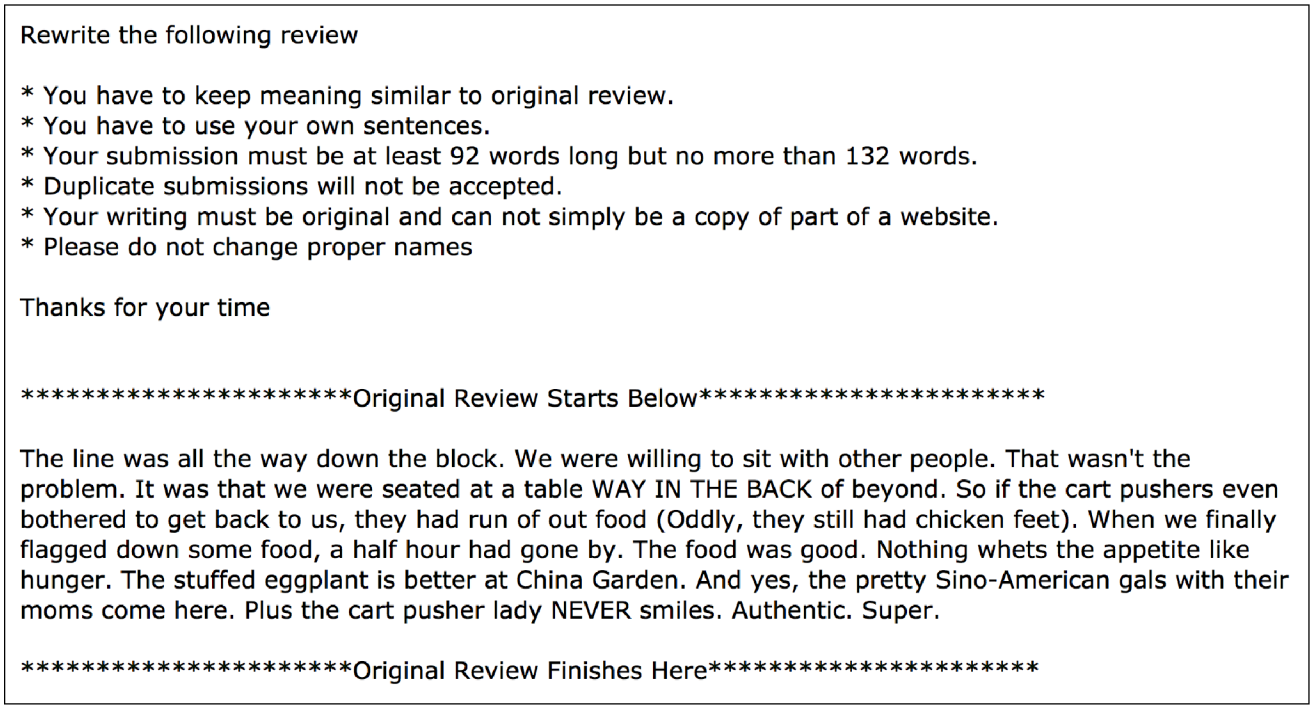


Figure 3: Sample rewriting task in MTurk

- We need solid API support in order to publish numerous rewriting tasks. We also need a stable and intuitive web interface, so that the crowdsourcing service can be easily used. Fortunately, MTurk offers a user-friendly web interface for isolated users and API support to automate a large number of tasks.
- Some recent research efforts have used MTurk in similar studies [28, 16, 18].

In general, we need crowdsourcing for two phases: (1) rewriting original reviews, and (2) conducting a readability and faithfulness evaluation between original and rewritten reviews. More than 400 random MTurkers participated in both phases.

6.1.3 *Rewriting Phase*

Out of three randomly created AR and IR review sets we used in Section 5, we randomly selected one as the target for anonymization experiments. We then uploaded all reviews in this AR set to the crowdsourcing service and asked MTurkers to rewrite them using their own words. We asked 5 MTurkers to rewrite each review, in order to obtain more comprehensive and randomized data for the subsequent linkability study. While rewriting, we explicitly instructed participants to keep the meaning similar and not to change proper names from the original review. Moreover, we checked whether the number of words in each new review is close to that of the original before accepting a rewritten submission; divergent rewrites were rejected. A sample rewriting task and its submission are shown in Figure 3.

We published reviews on a weekly basis in order to vary the speed of gathering rewrites. Interestingly, most tasks were completed during the first 3 days of week, and the remaining 4 days were spent reviewing submissions. We finished the rewriting phase in 4 months. Given 40 authors and AR size of 5 (200 total original reviews), each review was rewritten by 5 MTurkers, resulting in $1,000$ total submissions. Of these, we accepted 882. The rest were too short or too long, not meaningful, not faithful enough, or too similar, to the original. Moreover, out of 200 originals, 139 were rewritten 5 times. All original and rewritten reviews can be found at our publicly shared folder [6].

We paid US$0.12, on average, for each rewriting task. Ideally, a crowdsourcing-based review rewriting system would be free, with peer reviewers writing their own reviews and helping to re-writing others. However, since there was no such luxury at our disposal, we decided to settle on a low-cost approach[6]. Initially, we offered to pay US$0.10 per rewritten review. However, because review size ranges between 2 and 892 words, we came up with a sliding-price formula: $0.10 for every 250 words or a fraction thereof, e.g., a 490-word review pays $0.20 while a 180-word one pays $0.10. In addition, Amazon MTurk charges a 10% fee for each task.

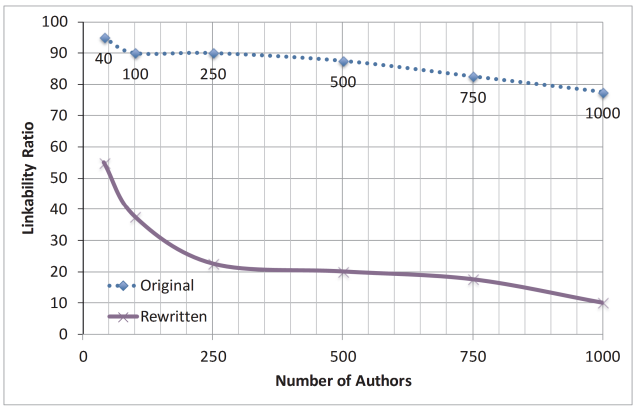
One of our secondary goals was assessment of efficacy and usability of the crowdsourcing service itself. We published one set of 40 reviews via the user interface on the MTurk website, and the second set of 160 reviews – using MTurk API. We found both means to be practical, error-free and easy to use. Overall, anyone capable of using a web browser can easily publish their reviews on MTurk for rewriting.

After completing the rewriting phase, we continued with a readability study to assess sensibility of rewritten reviews and their correspondence to the originals.

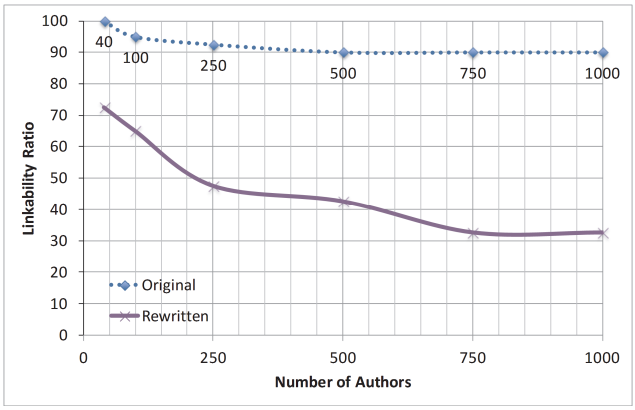
6.1.4 *Readability Study*

Readability study proceeded as follows: First, we pick, at random, 100 reviews from 200 reviews in the AR set. Then,

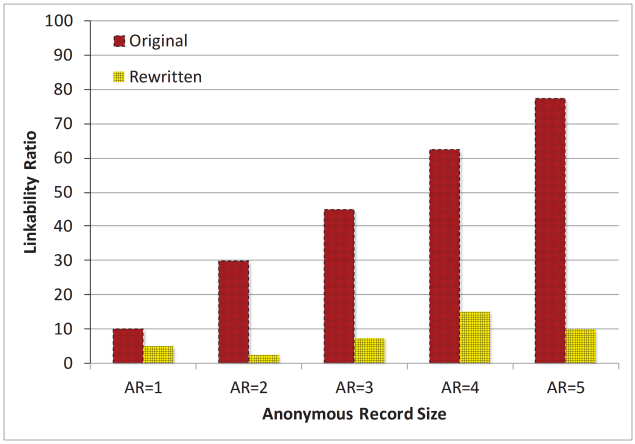
[6]We consider the average of US$0.12 to be very low per review cost.



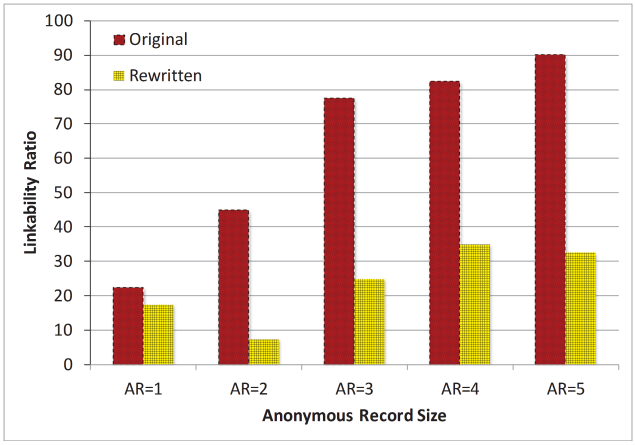
(a) Top-1 while varying the size of author set



(b) Top-4 while varying the size of author set



(c) Top-1 in a set of 1000 authors



(d) Top-4 in a set of 1000 authors

Figure 4: LRs of Original and Rewritten Reviews

for each review, we randomly select one rewritten version. Next, for every $[original, rewritten]$ review-pair, we publish a readability task on MTurk. In those tasks, we ask two distinct MTurkers to score rewritten reviews by comparing its similarity and sensibility to the original one. We define the scores as Poor(1), Fair(2), Average(3), Good(4), Excellent(5), where Poor means that the two reviews are completely different, and Excellent means they are essentially the same meaning-wise. We also ask MTurkers to write a comprehensive result which explains the differences (if any) between original and rewritten counterparts. A sample readability study task and its submission are given in Figure 5.

This study took one week and yielded 142 valid submissions. Results are reflected in Figure 6. The average readability score turns out to be 4.29/5, while 87% of reviews are given scores of Good or Excellent. This shows that rewritten reviews generally retain the meaning of the originals. Next, we proceed to re-assess stylometric linkability of rewritten reviews.

6.1.5 *Linkability of Rewritten Reviews*

Recall that the study in Section 5 involved 3 review sets each with 100 reviews per author. For the present study, we only consider the first set since we published anonymous reviews from first set to MTurk. In this first set, we replace AR with the corresponding set of MTurk-rewritten reviews where we pick a random rewritten version of each review, while each author's IR remains the same.

Figures 4(a) and 4(b) compare LRs between original - rewritten reviews with varying number of authors. Interestingly, we notice a substantial decrease in LRs for all author sizes. For $|AR| = 5$ in a set of 1000 authors, Top-1 and Top-4 LR drop from 77.5% to 10% and from 90% to 32.5% respectively. Even only in 40 authors set, Top-1 LR decreases to 55%, which is significantly lower than 95% achieved with original reviews.

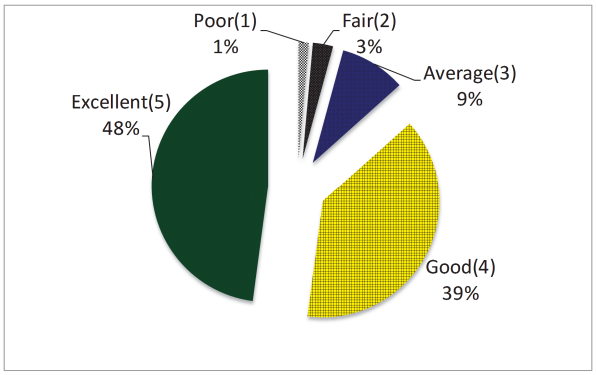


Figure 6: Readability Results of Rewritten Reviews

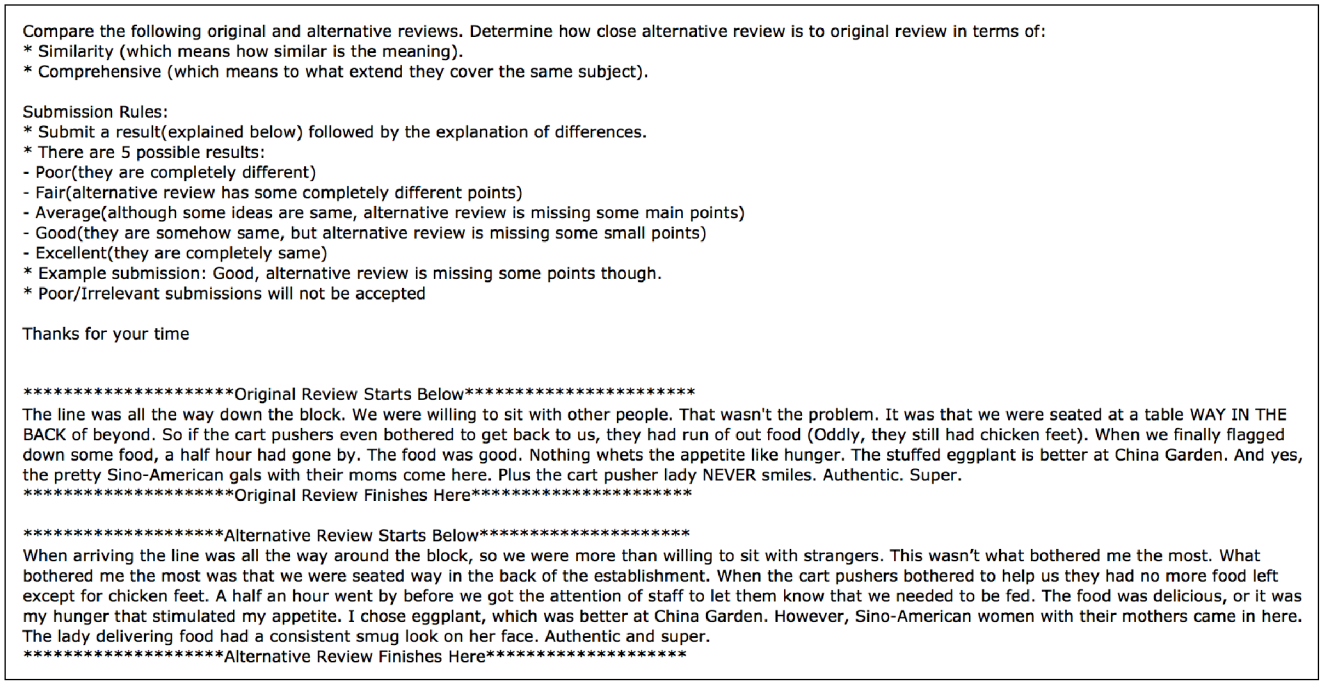


Figure 5: Sample readability task in MTurk

We also present a detailed comparison of original and rewritten reviews' LRs with different AR sizes in Figures 4(c) and 4(d). Notably, both Top-1 and Top-4 LR decrease dramatically for all AR sizes. 35% is the highest LR obtained with rewritten reviews, which is substantially less than those achieved with original counterparts.

After experiencing this significant decrease in linkability, we analyze rewritten reviews to see what might have helped increase anonymity. We notice that most MTurkers do not change the skeleton of original review. Instead, they change the structure of individual sentences by modifying the order of subject, noun and verb, converting an active sentence into a passive one, or vice versa. We also observe that MTurkers swap words with synonyms. We believe that these findings can be combined into an automated tool, which can help authors rewrite their own reviews. This is one of the items for future work, discussed in more detail in Section 7.

6.1.6 Crowdsourcing Summary

We now summarize key findings from the crowdsourcing experiment.

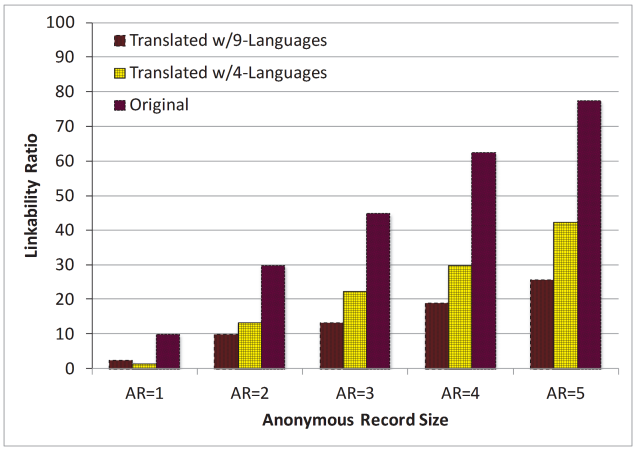
1. MTurk based crowdsourcing yielded rewritten reviews that were:
   - **Low-cost** – we paid only $0.12 including 10% service fee for rewriting each 250-word review.
   - **Fast** – we received submissions within 3-4 days, on average.
   - **Easy-to-use** – based on experiences with both user-interface and API of MTurk, an average person who is comfortable using a browser, Facebook or Yelp can easily publish reviews to MTurk.
2. As the readability study shows, crowdsourcing produces meaningful results: rewrites remain faithful to originals. (See Section 6.1.4).
3. Most importantly, rewrites substantially reduce linkability. For an $|AR| = 5$ where we previously witnessed the highest LR, Top-1 LR shrunk from 95% to 55% in a set of 40 authors and from 77.5% to 10% in a set of 1000 authors. (See Section 6.1.5).

6.2 Translation Experiments

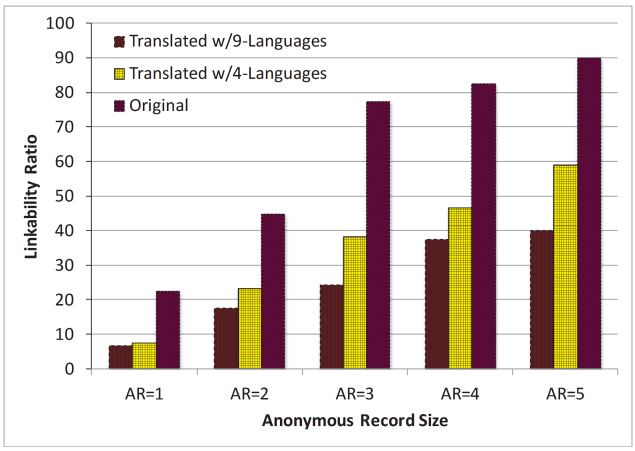
We now consider an alternative approach that uses on-line translation to mitigate linkability discussed in Section 5. The goal is to assess the efficacy of translation for stylometric obfuscation and check whether translation and crowdsourcing can be blended into a single socio-technological linkability mitigation technique.

It is both natural and intuitive to consider machine (automated, on-line) translation for obfuscating stylometric features of reviews. One well-known technique is to incrementally translate the text into a sequence of languages and then translate back to the original language. For example, translating a review from (and to) English using three levels of translation (two intermediate languages) could be done as follows: English → German → Japanese → English. The main intuition is to use the on-line translator as an external re-writer, so that stylometric characteristics would change as the translator introduces its own characteristics.

Using a translator to anonymize writing style has been attempted in prior work [13, 15]. However, prior studies did not go beyond three levels of translation and did not show significant decreases in linkability. Also, it was shown that that translation often yields non-sensical results, quite divergent from the original text [13]. Due to recent advances in this area, we revisit and reexamine the use of translation. Specifically, we explore effects of the number of intermediate languages on linkability and assess readability of translated outputs. In the process, we discover that translators are actually effective in mitigating linkability, while readability



(a) Top-1



(b) Top-4

Figure 7: Comparison of Original-Translated Reviews LRs in a set of 1000 authors

is (though not great) is reasonable and can be easily fixed by crowdsourcing.

6.2.1 Translation Framework

We begin by building a translation framework to perform a large number of translations using any number of languages. Currently, Google [4] and Bing [2] offer the most popular machine translation services. Both use statistical machine translation techniques to dynamically translate text between thousands of language pairs. Therefore, given the same text, they usually return a different translated version. Even though there are no significant differences between them, we decided to use Google Translator. It supports more languages: 64 at the time of this writing [5], while Bing supports 41 [3]).

Google provides a translation API as a free service to researchers with a daily character quota, which can be increased upon request. The API provides the following two functions:

- $translate(text, sourceLanguage, targetLanguage)$: Translates given text from source language to target language.
- $languages()$: Returns the set of source and target languages supported in the *translate* function.

Using these functions, we implement the algorithm, shown in Algorithm 1. We first select N languages at random. Then, we consecutively translate text into each of the languages, one after the other. At the end, we translate the result to its original language, English, in our case. We consider the final translated review as the anonymized version of the original.

We also could have used a fixed list of destination languages. However, it is easy to see that translated reviews might then retain some stylometric features of the original (This is somewhat analogous to deterministic encryption.). Thus, we randomize the list of languages hoping that it would make it improbable to retain stylometric patterns. For example, since Google translator supports 64 languages, we have more than $\prod_{n=0}^{N-1}(64-n) \approx 2^{53}$ distinct lists of languages for $N = 9$.

After implementing the translation framework, we proceed to assessing linkability of the results.

Algorithm 1 Round-Translation of *Review* with N random languages

```
Obtain all supported languages via languages()
RandomLanguages ← select N languages randomly
Source ← "English"
for Language language in RandomLanguages do
    Review ← translate(Review, Source, language)
    Source ← language
end for
Translated ← translate(Review, Source, "English")
return Translated
```

6.2.2 Linkability of Translated Reviews

Using Algorithm 1, we anonymized the AR review set[7]. We varied N from 1 to 9 and re-ran linkability analysis with translated reviews as the AR. In doing so, we used S_F identified in Section 5. To assert generality of linkability of translated texts, we performed the above procedure 3 times, each time with a different list of random languages and then ran linkability analysis 3 times as well. Average linkability results of all 3 runs are plotted in Figures 8, 7(a) and 7(b).

For the number of intermediate languages, our intuition is that increasing the number of levels of translation (i.e., intermediate languages) causes greater changes in stylometric characteristics of original text. Interestingly, Figure 8 supports this intuition: larger N values yield larger decreases of linkability. While the decrease is not significant in Top-4 for N: $[1, 2]$, it becomes more noticeable after 3 languages. For $|AR| = 5$, we have Top-1 & Top-4 linkabilities of 42.5% & 59% with 4 languages, 31% & 47% with 7 languages and 25% & 40% with 9 languages, respectively. These are considerably lower than 77.5% & 90% achieved with original ARs. Because Top-1 linkability decreases to 25% after 9 languages, we stop increasing N and settle on 9.

[7]Translated example reviews are shown in Appendix B.

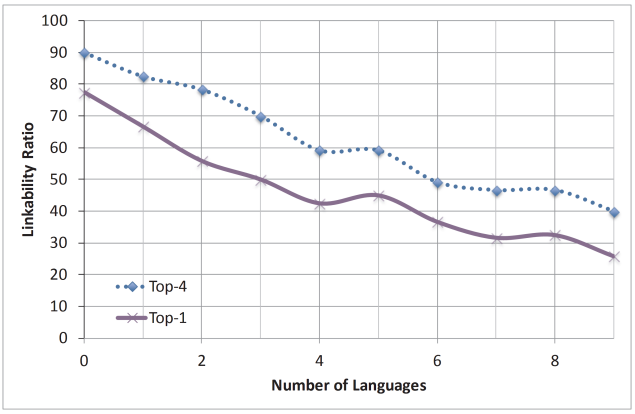


Figure 8: LRs with $|AR| = 5$ while varying number of languages in a set of 1000 authors

Figures 7(a) and 7(b) show reduction in Top-1 and Top-4 linkability for varying AR sizes. In all of them, original reviews have higher LRs than ones translated with 4 languages; which in turn have higher LRs than those translated with 9 languages. This clearly demonstrates that when more translations are done, the more translator manipulates the stylometric characteristics of a review.

6.2.3 *Readability of Translated Reviews*

So far, we analyzed the impact of using on-line translation on decreasing stylometric linkability. However, we need to make sure that the final result is readable. To this end, we conducted a readability study. We randomly selected a sample of translated reviews for $N = 9$. We have 3 sets of translated reviews, each corresponding to a random selection of 9 languages. From each set, we randomly selected 20 translated reviews, which totals up to 60 translated reviews. Then, for each $[original, translated]$ review-pair, we published readability tasks on MTurk (as in Section 6.1.4) and had it assessed by 2 distinct MTurkers, resulting in 120 total submissions.

Results are shown in Figure 9. As expected, results are not as good as those in Section 6.1.4. However, a number of reviews preserve the original meaning to some extent. The average score is 2.85 out of 5 and most scores were at least "Fair".

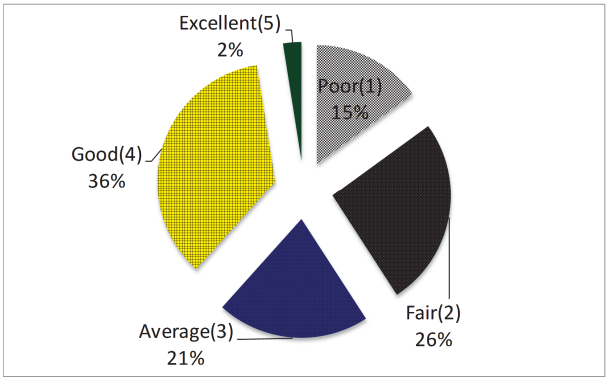


Figure 9: Readability Results of Translated Reviews

6.2.4 *Fixing the Translated Reviews*

Even though machine translation is continuously getting better at producing readable output, the state-of-the-art is far from ideal. After manually reviewing some $[original, translated]$ pairs, we realized that most translated reviews retained the main idea of the original. However, because of: (1) frequently weird translation of proper nouns, (2) misorganization of sentences, and (3) failure of translating terms not in the dictionary, translated review are not easy to read. We decided to provide translated reviews along with their original versions to MTurkers and asked them to fix unreadable parts[8]. As a task, this is easier and less time-consuming than rewriting the entire review.

Out of 3 translated review sets, we selected one at random and published all 200 ($|AR| = 5$ for 40 authors) translated reviews from our AR set to MTurk. We received 189 submissions; only 31 authors had their full AR's translated reviews completely fixed. We then performed the same linkability assessment with these 31 authors while we update their AR's by translated-fixed reviews.

Comparison of linkability ratios between original, translated, and fixed version of the same translated reviews is plotted in Figure 10(a). It demonstrates that, fixing translations does not significantly influence linkability. In AR-5, Top-1 linkability of fixed translation is 19% while non-fixed translations 25%. Meanwhile, both are significantly lower than 74% LR of original counterparts.

Finally, we perform a readability study on fixed translations. Out of 189 submissions, we select 20 randomly and publish to MTurk as a readability task. Average readability score increased from 2.85 to 4.12 after fixing the machine translation. Detailed comparison of readability studies between translated and translated-fixed reviews is given in Figure 11. We notice high percentage of translated reviews has Average score, while fixed counterparts mostly score as Good or Excellent. Results are really promising since they show that the meaning of a machine translated review can be fixed while keeping it unlinkable.

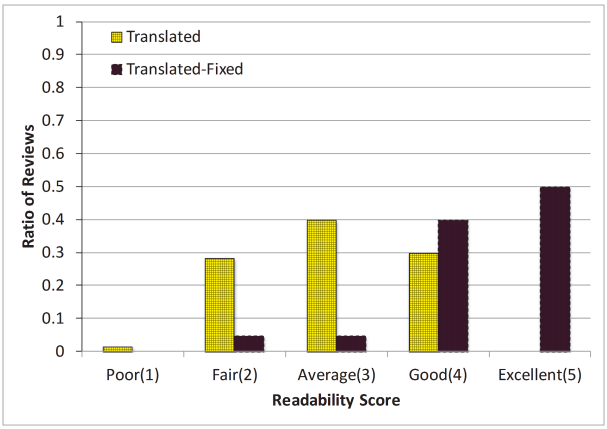
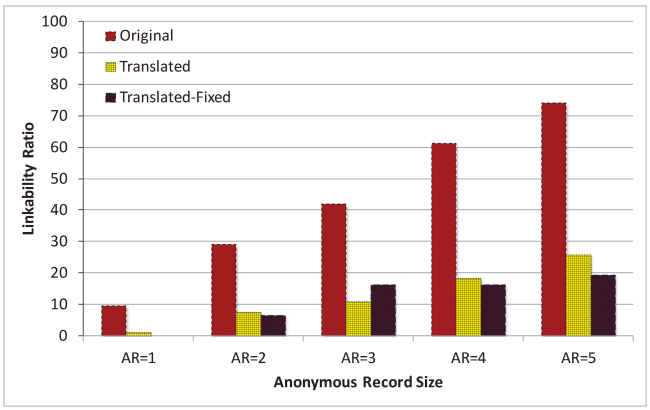
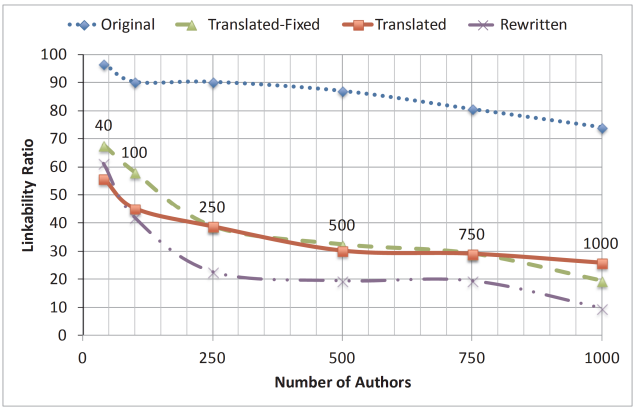


Figure 11: Readability study comparison between Translated and Translated-Fixed Reviews

[8]Sample submission to a translation-fix task is presented in Appendix B.



(a) LRs in a set of 1000 authors while varying the AR size



(b) LRs with $|AR| = 5$ while varying the number of authors

Figure 10: Top-1 LRs for Original, Translated, Translated-Fixed and Rewritten Reviews

6.2.5 *Comparison of Anonymization Techniques*

We present the comparison of linkability results achieved using crowdsourcing, machine translation and combination of both [9]. in Figure 10(b). Regardless of the size of author set, we achieve substantial decrease in linkability. Our techniques show that people are good at rewriting and correcting reviews while introducing their own style, keeping the meaning similar, and, most importantly, reducing linkability. While purely rewritten reviews have the lowest linkability, both translated and translated-fixed reviews perform comparable to each other. As far as readability, crowdsourcing (mean score of 4.29/5) performed much better than translation (mean score of 2.85/5). However, results show that low readability scores can be fixed (resulting in a mean score of 4.12/5) using crowdsourcing while keeping linkability low. We summarize results as follows:

- Crowdsourcing: Achieves better anonymity and readability. However, it takes longer than translation since it is not an automated solution. Moreover, though not expensive, it is clearly not free.
- Machine Translation: Completely automated and cost-free approach which takes less time than crowdsourcing. However, poor readability is the main disadvantage.

7. DISCUSSION

In spite of its usefulness in decreasing linkability and enhancing readability, there are some open questions associated with the use of crowdsourcing.

1. **How applicable is crowdsourcing to other OSNs?** In some other OSNs penalties for deanonymization would be higher than Yelp. However we chose Yelp dataset for the reasons given in Section 4.1. The same technique can be presumably applied to other settings, e.g., anonymous activist blogs, tweets in Twitter and TripAdvisor reviews.

2. **How might authors get their reviews rewritten?** This could be addressed by integrating a plug-in into a browser. When an author visits an OSN and writes a review, this plug-in can ease posting of a task to a crowdsourcing platform and return the result back to the author via one-time or temporary email address. On the system side, plug-in would create a rewriting task and relay it to the crowdsourcing system. A possible building block can be the recent work in [12] that proposes a crowdsourcing task automation system. It automates task scheduling, pricing and quality control, and allows tasks to be incorporated into the system as a function call.

3. **How feasible is crowdsourcing in terms of latency and cost?** We believe that a delay of couple of days would not pose an inconvenience since review posting does not need to occur in real time. Many popular OSNs does not publish reviews instantly, e.g., TripAdvisor screens each review to make sure it meets certain guidelines. This moderation can take as long as several weeks [8].

   As far as costs, we paid US$0.12, on average for each rewriting task. We consider this amount is extremely low which can be easily subsidized by the advertizing revenue, with ads in the plug-in.

4. **Is there a privacy risk in posting reviews to strangers?** It is difficult to assess whether there is a privacy risk since an adversary does not learn both posted and rewritten reviews, unless she is registered as a worker, completes the task, and her submission gets published. However, this clearly does not scale for the adversary when the number of posted reviews is large and requires manual follow-up with the posts. Also, MTurk Participation Agreement[10] involves conditions that protect privacy of both worker and requester.

5. **Is there a chance of having a rewriter's writing style recognized?** We believe that this is not the case. First, there are many workers to choose from and we can force the system not to select the same worker more than a specific number of times. Second, we expect that a worker would rewrite many reviews

[9]See: https://github.com/ekinoguz/JGAAP-Sprout and https://github.com/ekinoguz/hiding for the source code of our experiments and anonymization techniques.

[10]See: https://www.mturk.com/mturk/conditionsofuse

from different sources. This will widen the range of topics that rewritten reviews would cover and would make rewritten reviews more difficult to recognize. Finally, the identities of workers are expected to remain private since the only party who can see worker details for a given task is the person who posted it.

6. **Is there a chance of using crowdsourcing to generate fake content?** If our work is adopted, it might actually help spammers to make fake content easier to generate and avoid detection. However, the real effect is yet to be tested. Although real and fake reviews are generated by the same means, they are generated for different purposes with different content. Real reviews are generated to preserve anonymity and they hold on the meaning of the original review, whereas fake ones are generated to give fake assessments. Despite stylometric similarities, fake reviews could possibly hold many features that would distinguish them from real ones; such as multiple copies, similarities in ratings with other fake reviews, exaggerations in assessments, etc. But the real effect is yet to be assessed.

8. CONCLUSIONS AND FUTURE WORK

This paper investigated authorship linkability in community reviewing and explored some means of mitigating it. First, we showed, via a linkability study using a proper subset of the Writeprints feature set, that authorship linkability is higher than previously reported. Then, using the power of global crowdsourcing on the Amazon MTurk platform, we published reviews and asked random strangers to rewrite them for a nominal fee. After that, we conducted a readability study showing that rewritten reviews are meaningful and remain similar to the originals. Then, we re-assessed linkability of rewritten reviews and discovered that it decreases substantially. Next, we considered using translation to rewrite reviews and showed that linkability decreases while number of intermediary languages increases. After that, we evaluated readability of translated reviews, and realized that on-line translation does not yield results as readable as those from rewritings. Next, we take advantage of crowdsourcing to fix poorly readable translations and still achieve low linkability.

This line of work is far from being complete and many issues remain for future consideration:

- We need to explore detailed and sophisticated evaluation techniques in order to understand stylometric differences between original, rewritten and translated reviews. If this succeeds, more practical recommendations can be given to review authors.
- As discussed in Section 7, we want to parlay the results of our study into a piece of software or a plug-in intended for authors.
- We need to conduct the same kind of study in the context of review sites other than Yelp, e.g., Amazon, TripAdvisor or Ebay. Also, cross-site studies should be undertaken, e.g., using a combination of Amazon and Yelp reviews.

Acknowledgments

This research was conducted as part of NSF CSR Award 1213140: "Collaborative Research: Enabling Privacy-Utility Trade-offs in Pervasive Computing Systems".

9. REFERENCES

[1] Amazon Mechanical Turk. `https://www.mturk.com/mturk/`.
[2] Bing Translator. `http://www.bing.com/translator`.
[3] Bing Translator Language Codes. `http://msdn.microsoft.com/en-us/library/hh456380.aspx`.
[4] Google Translate. `http://translate.google.com/`.
[5] Google Translator API. `https://developers.google.com/translate/`.
[6] Original, Rewritten, Translated and Translated-Fixed Reviews. `http://sprout.ics.uci.edu/projects/aaa/dataset_userhiding.tar.gz`.
[7] Reference book and online dictionaries. `http://www.merriam-webster.com/`.
[8] TripAdvisor Review Moderation. `http://www.tripadvisor.com/vpages/review_mod_fraud_detect.html`.
[9] Yelp By The Numbers. `http://officialblog.yelp.com/2010/12/2010-yelp-by-the-numbers.html`.
[10] A. Abbasi and H. Chen. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. In *ACM Transactions on Information Systems*, 2008.
[11] S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *IEEE Symposium on Security and Privacy*, 2012.
[12] D. W. Barowy, C. Curtsinger, E. D. Berger, and A. McGregor. Automan: A platform for integrating human-based and digital computation. In *OOPSLA*, 2012.
[13] M. Brennan, S. Afroz, and R. Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 2012.
[14] M. R. Brennan and R. Greenstadt. Practical attacks against authorship recognition techniques. In *IAAI*, 2009.
[15] A. Caliskan and R. Greenstadt. Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In *ICSC*, 2012.
[16] E. Hayashi, J. Hong, and N. Christin. Security through a different kind of obscurity: evaluating distortion in graphical authentication schemes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, 2011.
[17] G. Kacmarcik and M. Gamon. Obfuscating document stylometry to preserve author anonymity. In *ACL*, 2006.
[18] P. G. Kelley. Conducting usable privacy & security studies with amazon's mechanical turk. In *Symposium on Usable Privacy and Security (SOUPS)(Redmond, WA*, 2010.
[19] A. W. E. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt. Use fewer instances of the letter "i": Toward writing style anonymization. In *Privacy Enhancing Technologies*, 2012.
[20] M. A. Mishari and G. Tsudik. Exploring linkability of user reviews. In *ESORICS*, 2012.

[21] M. Nanavati, N. Taylor, W. Aiello, and A. Warfield. Herbert West – Deanonymizer. In *6th USENIX Workshop on Hot Topics in Security*, 2011.

[22] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song. On the Feasibility of Internet-Scale Author Identification. In *IEEE Symposium on Security and Privacy*, 2012.

[23] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[24] J. R. Rao and P. Rohatgi. Can pseudonymity really guarantee privacy. In *Proceedings of the Ninth USENIX Security Symposium*, 2000.

[25] R. Schumacker and S. Tomek. Chi-square test. In *Understanding Statistics Using R*, pages 169–175. Springer, 2013.

[26] E. Stamatatos. A Survey of Modern Authorship Attribution Methods. In *Journal of the American Society for Information Science and Technology*, 2009.

[27] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[28] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, S. Egelman, and J. López. Helping users create better passwords. *USENIX*, 2012.

APPENDIX

A. CROWDSOURCING EXAMPLES

We present two example submissions from our rewriting and readability tasks in MTurk. Note that the full collection of original and rewritten reviews can be accessed in [6].

A.1: Rewriting Example

Sample rewritten review for the task given in Figure 3: *"When arriving the line was all the way around the block, so we were more than willing to sit with strangers. This wasn't what bothered me the most. What bothered me the most was that we were seated way in the back of the establishment. When the cart pushers bothered to help us they had no more food left except for chicken feet. A half an hour went by before we got the attention of staff to let them know that we needed to be fed. The food was delicious, or it was my hunger that stimulated my appetite. I chose eggplant, which was better at China Garden. However, Sino-American women with their mothers came in here. The lady delivering food had a consistent smug look on her face. Authentic and super."*

A.2 Readability Example

Sample readability study submission for a task given in Figure 5:
"Good, they are similar on the main points. When discussing the Sino-American women and their mothers, the alternative left out 'pretty' which seemed like a point that appealed to original reviewer. Also, by using the word 'smug' to describe the lady delivering food, the alternative review depicts her more negatively than the original."

B. TRANSLATION EXAMPLES

We present the round-trip machine translation of original review in Figure 3 with three, six and nine random selection of intermediate languages, and also translated-fixed version by MTurk below:

Random round-trip machine translation with three languages $English \rightarrow Welsh \rightarrow Irish \rightarrow Ukrainian \rightarrow English$**:** *"The line was all the way down the block. We are ready to sit with other people. This was not a problem. It was that we were sitting at the table, and the way in the back longer. So if the cart pushers even bother to get back to us, they run out of food (surprisingly, still chicken feet). When we finally gave up a little food half an hour passed. Food was good. Nothing stimulates the appetite like hunger. Best stuffed eggplants in the garden China. And many Chinese-American girls and their mothers to come here. Plus a woman pushing a cart ever smiles. Authentic. Super."*

Random round-trip machine translation with six languages $English \rightarrow Welsh \rightarrow Irish \rightarrow Ukrainian \rightarrow Yiddish \rightarrow French \rightarrow Korean \rightarrow English$**:** *"The line at the bottom of the box is We sit down with other people who are prepared. It was not a problem. Pusherz basket even if you do not bother coming back to us, we have it sitting on the table, and how I was back in May, they (surprisingly, many chicken feet) to food shortages. Some food when we passed an hour and a half. The food was good. Nothing like hunger and appetite stimulation. China stuffed in the garden of the. Many Chinese-American women and their mothers come here. The woman pushing a cart does not smile. True. More."*

Random round-trip machine translation with nine languages $English \rightarrow Welsh \rightarrow Irish \rightarrow Ukrainian \rightarrow Yiddish \rightarrow French \rightarrow Korean \rightarrow Esperanto \rightarrow Norwegian \rightarrow Malay \rightarrow English$**:** *"Line at the bottom of the box, we sit down with other people who are willing. It's not a problem. Pusherz cart even if you do not bother to come back to us, we have it sitting on the table, and as I was back in May, (surprisingly, a lot of chicken feet) to food shortages. Some foods when they pass an hour and a half. Food is good. Nothing like hunger and appetite stimulation. China filling the park. Many Chinese-American women and their mothers come here. Woman pushing a trolley is not smiling. Sant. More."*

Fixed version of previous machine translated review with nine languages: *"The problem wasn't that we weren't willing to sit with other people but the line was still all the way down the block. The problem was that we were seated a table as far away as could be. Even if the servers made it to our table, their trays were empty (except for the chicken feet). It was a half an hour before we were able to get any of the food but it tasted good because we were so hungry; although, the China Garden has better stuffed eggplant. The pretty Sino-American girls come here with their mothers and the woman server never smiles. Traditional. Awesome."*

Note that the full collection of translated reviews can be found in [6].