Investigation of Biases in Identity Linkage DataSets

Rishabh Kaushal
rishabhk@iiitd.ac.in
IGDTUW & IIIT, Delhi

Shubham Gupta
shubhamg@iiitd.ac.in
IIIT, Delhi

Ponnurangam Kumaraguru
pk@iiitd.ac.in
IIIT, Delhi

ABSTRACT

In social networks, the problem of identity linkage is to find whether a pair of user identities on two social networks belong to the same individual or not. Prior works typically first collect ground truth datasets of user identities across social networks belonging to the same individuals and then build a machine learning model driven by features from user identities. User behaviors in different social networks drive the construction of these datasets, and as a consequence, behavioral biases get manifested in them. Our work performs a detailed investigation into these dataset biases, a work which has mostly remained under-explored in the identity linkage research. More specifically, we characterize, detect, and quantify behavioral biases in the dataset that manifest in the form of lexical differences in user-generated content, particularly in usernames and display names configured by users. We study these biases on more than 1 million user identity pairs obtained by leveraging two user behaviors, namely cross-posting and self-disclosure. We find that users who self-disclose their usernames and display names on different social networks show higher lexical similarity than users who cross-post. These behavioral biases lower down the performance (precision and recall) of learning models by 5-20%. Inspired by discrimination measurement metrics, we propose and implement a framework to quantify the extent of these biases and find that 15-20% of test data get affected.

KEYWORDS

Bias Detection, Online Social Networks, Data Mining

ACM Reference Format:
Rishabh Kaushal, Shubham Gupta, and Ponnurangam Kumaraguru. 2020. Investigation of Biases in Identity Linkage DataSets. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341105.3374015

1 INTRODUCTION

A growing trend in online social media (OSM) landscape is that users join multiple OSMs to get access to diverse content and friend networks offered by these different OSM platforms. Statistics show that from 2013 to 2017, the average accounts on OSMs per online user have risen from 4.3 to 7.6.[1] It means that the same individual maintains multiple accounts across different OSMs, referred to as *user identities* in this work. This scenario leads to the problem of *identity linkage*, where the goal is to find whether two input user identities belong to the same individual or not. If the two user identities belong to the same individual, we refer them as *linked identity pair* else *non-linked identity pair*. Linked identities present more comprehensive coverage of user behavior, thereby, helping in better recommendations.

Prior works address this problem of identity linkage in two steps, as depicted in Fig 1. The first step involves the collection of linked identity pairs on two OSMs using a well-defined *data collection method*, which we refer to in this work as a *data source*. The second step involves learning of a data-driven classification model over handcrafted features extracted from the three dimensions of a user identity namely profile information [14, 25], content posted (and interacted) [4] and the friend network [27]. More formally, given two identities I_a and I_b from two social networks a and b, respectively, the goal is to learn a *classifier function* f defined in eq 1, such that it returns 1 if I_a and I_b belongs to the same individual else it returns 0.

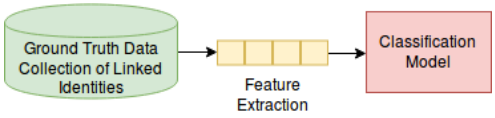


Figure 1: Basic Framework for Identity Linkage

$$f(I_a, I_b) = \begin{cases} 1, & \text{if } I_a \text{ \& } I_b \text{ belong to same user.} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Numerous data collection approaches have been proposed in the past to collect linked user identities. Each of them relies on the typical behaviors of users who maintain identities across multiple social networks. As a consequence, these behavioral biases exhibited by users get infested in these identity linkage datasets. Given that we build classification models on features derived from attributes of this collected data, the biases affect these models as well. Although biases, in general, are being extensively studied, however, the study of behavioral biases that manifest in the linked user identity datasets have not been explored. In this work, we fill this gap by investigating the *impact* of behavioral biases in user identities on the performance of an identity linkage solution by addressing three research questions.

(1) **Detection:** *Does behavioral bias exist in identity linkage datasets collected using different approaches?*

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3374015

[1] https://www.statista.com/statistics/788084/number-of-social-media-accounts/

(a) Cross-Platform Sharing: Instagram post is cross posted on Twitter. Link to th Instagram post appears on Twitter post (tweet).

(b) Self-Disclosure: Instagram identity is mentioned in the bio-field of Twitter identity.

Figure 2: Illustrations of two user behaviors namely cross-sharing and self-disclosure which are leveraged to collect ground truth linked user identities for construction of identity linkage models

(2) **Quantification:** *Whether the performance of an identity linkage model is affected by behavioral biases in the dataset? What is the severity with which these behavioral biases impact the decision making capability of identity linkage models?*

To address these questions, we consider two different approaches for collecting linked user identity pairs (in other words, two *data sources*), based on two user behaviors, namely cross-platform sharing (CPS) and self-disclosure (SD). In this work, we focus our attention on *two* elementary behaviors that a user performs to maintain their identities on a social network, namely *username creation* and *display name configuration*. To study the impact of these biases on the performance of identity linkage model, we follow the typical approach adopted by prior works ([4, 14, 25, 27]) in building a learning function based on lexical features derived from usernames and display names. Ideally, the identity linkage model ought to be *generalizable*, which means that it's performance should not get affected by the data source. However, our study shows that the model trained on CPS dataset and evaluated on SD dataset (and vice-versa) performs *5-20% poorly in terms of precision and recall* than the model which is trained and evaluated on the same dataset. This clearly indicates that behavioral biases that exist in the dataset are impacting the performance of the model. To address the second research question, we leverage the works on discrimination studies and biases [1, 19, 28], to propose a *framework* that uses discrimination discovery metrics to quantify the extent of damage caused by behavioral biases in the dataset. More specifically, we apply *situational testing* proposed by Luong et al. [17] for measuring individual biases to quantify the behavioral biases in user identity datasets. Code and related dataset of our work can be found at https://github.com/precog-iiitd/identity-linkage-biases. The key contributions of our work are as follows.

- Ours is the *first* work which investigates, detects, and measures the impact of data source bias on the performance of identity linkage models.
- We propose a *novel* methodology to apply discrimination studies to quantify the extent of damage caused by data source biases.

2 RELATED WORK

Prior works on the problem of identity linkage have been well documented [23]. The common underlying approach behind these works are first to perform data collection and then follow it up by construction of machine learning-based model on features extracted from the user's content, network, and profile attributes. We briefly summarize the data collection methods.

- *Crawling*: A web crawler is typically developed in some works [5], which crawls user pages on different OSMs platforms to collect user information. This is done to collect more comprehensive user information which is otherwise not possible due to API restrictions.
- *Social Aggregators*: There are many social platforms (like About.me) and blogs which allow users to specify their identities on various OSMs. Previous works [15, 18, 20, 25] use these platforms to collect linked user identities.
- *Self Disclosure*: Many OSMs allow users to explicitly *link* their account with their other identities on different OSMs, which have been leveraged by some works [12, 14]. At times, users also specifically do this linking by mentioning their identities on other OSMs in their profile (or bio field, in case of Twitter) attribute.
- *Cross-Platform Sharing*: It is observed that sometimes users cross-post the same information across multiple OSM platforms where they maintain their identities. Prior works [8, 9] leverage this cross-posting user behavior to collect linked user identities.
- *Email based Friend Finder*: Many OSMs provide this feature of suggesting friends based on the contact list in the email which was used in registering the account. This option is exploited in some prior work [4].

While multiple methods for data collection have been used in the past, however, an investigation into the dataset biases have remained largely under-explored. Once data is collected, prior works typically build a machine learning based model that leverages features derived from three dimensions of user identity, namely content, network and profile, and their combination. These works are briefly summarized below.

- *Content*: Some works [3, 4] extracts features from the posts made by users on OSM platforms like time of post, length of post, and ocation of a post.
- *Network*: Same individuals are likely to have common friends across multiple social identities that they keep, prior works [27] exploit this aspect to construct features derived from user's network like common friends.
- *Profile*: It is natural for users to maintain multiple identities across OSM platforms to have a similar profile image, display name, location, etc in their profile information which is exploited in previous works [14, 15, 25].

3 DATA COLLECTION

Given that we are to study the impact of data sources, we use two data collection approaches, which are based on two user behaviors, namely cross-platform sharing (CPS) and self-disclosure (SD), to collect linked identity pairs [11]. Both these methods exploit two different behaviors exhibited by users who maintain multiple identities across social networks. In both the methods, the two social networks from where linked identity pairs were collected are Twitter and Instagram.

3.1 Cross Platform Sharing

This data collection method leverages on the user behavior wherein a user *cross posts* i.e. share a post made on one (referred to as *source*) social network on two or more target social networks, thereby, revealing his identities on the source and target social networks. We take Instagram as the source social network and Twitter as the target social network. Steps followed in data collection are as below.

(1) Using Twitter Search API, we search for pattern *instagram.com/p/* which is present whenever users cross-posts their posts made on Instagram over Twitter.
(2) Every tweet has a source parameter associated with it that provides details related to the origin of the tweet. Using this parameter, we filter those tweets that originate from Instagram and discard the rest. This ensures that only posts cross posted from Instagram remain and filters out those posts in which Instagram post URLs are pasted manually while creating tweets.
(3) For each of the tweets obtained in the previous step, we obtain their twitter username and display names using the Twitter API.
(4) On expanding the *instagram.com/p/* URL obtained from each of the tweets in step 2, we land on the web page that contains the username of the user on Instagram using which we obtain the display name.

3.2 Self Disclosure

In this data collection method, we look for user behavior in which users, while configuring their profile information on one social network, explicitly *mentions or self-discloses* details of identities on other social networks. More specifically, we focus on *bio* field of Twitter users to extract whether they have shared their identity on Instagram. Steps followed in data collection are as below.

Table 1: DataSet Details.

Class Label	Collection Method	#Pairs
Linked	Cross-Platform Sharing	253,791
Linked	Self-Disclosure	253,791
Unlinked - Random Pairs	Cross-Platform Sharing	190,343
Unlinked - Random Pairs	Self-Disclosure	190,360
Unlinked - Similar Pairs	Cross-Platform Sharing	63,448
Unlinked - Similar Pairs	Self-Disclosure	63,454
	Total Identity Pairs	**1,015,187**

(1) Using Twitter API, we leverage the sequential nature of twitter user IDs to extract *bios* of Twitter users.
(2) Using the obtained collection of bios, we use regular expressions to filter for bios that contain Instagram usernames or URLs that we save as the Instagram identity.
(3) Using this Instagram URL or username, we obtain the corresponding display name. Moreover, using Twitter API, we obtain the username and display name configured by these users on Twitter.

3.3 Negative Sample Generation

We note that both the above methods provide us with positive samples, which are *linked user identities* in the context of our problem. However, to train a classification model, it should be able to see negative samples as well i.e. user identity pairs on Twitter-Instagram which do not belong to the same individual referred to as *unlinked user identities*. We follow two approaches to generate negative samples detailed below.

(1) **Random Pairing**: We generate negative samples by randomly pairing Instagram and Twitter identities obtained in the two data collection approaches. In general, if (I^1_{tw},I^1_{in}) and (I^2_{tw},I^2_{in}) are two known linked identity pairs obtained on Twitter-Instagram social networks using either of the data collection approach, then we create unlinked user identity pairs as (I^1_{tw},I^2_{in}) and (I^2_{tw},I^1_{in}).
(2) **Similar Pairing**: While random pairing will guarantee us negative samples, in the real world, we do find identities which are quite similar to each, at least in terms of names. For instance, Perito et al. [20] studied the uniqueness of names, and found that some names are rare, while others are quite common. To factor this, we create negative samples using this method of similar pairing. As per this method, for a linked identity pair (I^1_{tw},I^1_{in}), we first obtain display name of I^1_{in} in Instagram and then use it to perform user search in Twitter using the Twittter Search API to find *top-k* identities on Twitter who have a *similar* display name, $I^1_{tw},I^2_{tw},I^3_{tw},....,I^k_{tw}$. Since (I^1_{tw},I^1_{in}) is known linked identity pair, we ignore it and keep the rest of the pairs i.e. (I^2_{tw},I^1_{in}), $(I^3_{tw},I^1_{in}),....,(I^k_{tw},I^1_{in})$ as unlinked user identity pairs.

Table 1 gives a detailed distribution of our dataset after implementing these two methods for generating negative samples.

4 METHODOLOGY & RESULTS

In this section, we explain our approach and present results for the following. (1) Study, detect and characterize behavioral biases in user identities on Twitter and Instagram. (2) Propose a framework to quantify the severity of behavioral biases. (3) Lastly, explain the process of mitigation of behavioral biases.

4.1 User Behavioral Features

Users exhibit a multitude of behaviors on social networks which can be categorized broadly into three types. (1) Content-related behaviors like what kind of posts are made by the user, etc. (2) Network-related behaviors like friends maintained by the user, etc. (3) Profile-related behaviors like configuring details of profile attribute (say city), making it visible, etc. Out of these three categories of user behaviors, we focus our attention to profile related behavior. Within profile configuration, we restrict ourselves to only *username* and *display name*. In order words, we are interested to study user behavior in terms of usernames and display names that users configure in their identities across multiple social networks, particularly Instagram and Twitter. We have made this decision for three reasons. First, the support for programmatic access through APIs to content, network and profile information of user identities in social networks has declined considerably. While Twitter does grant access, Instagram has restricted access to content and network. Second, among the various profile information, username and display name are the *elementary* profile attributes that are always configured by users and are publicly available. Third, on both the social networks Instagram and Twitter, users have the flexibility of modifying both usernames and display names, thereby making them suitable for studying user behaviors.

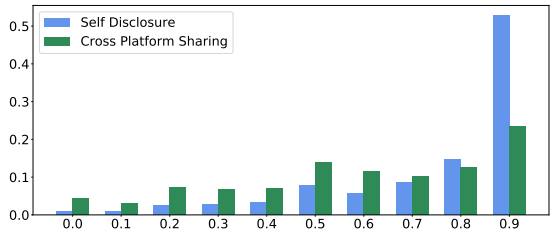
Having decided the user behaviors to study, our next step is to define lexical similarity based features that help us in measuring the differences and similarities in the username and display name configured by users across multiple social networks. In terms of lexical analysis, it may be observed that username can be considered as a string and features are derived from individual characters that appear in the username. On the other hand, the display name can be considered as a set of words (strings) and features are derived at word-level. We consider following features lexical features derived from username and display names.

- Longest Common Subsequence (LCS): For two given sequences (usernames in our case) UN^i_{tw} and UN^j_{in} from Twitter and Instagram, we find length of longest common subsequence at the character level.
- Jaccard Distance: It is based on jaccard similarity which considers two sets as input and returns their union divided by their intersection. We compute jaccard distance on two usernames obtained from two identities on Twitter and Instagram.
- Normalized Levenshtein Distance: For two given usernames UN^i_{tw} and UN^j_{in}, we compute the levenshtein distance as the minimum number of edits at the character level. Types of edits allowed are insertion, deletion or substitution of single character. We divide the distance by the length of the shorter username to normalize. For instance, levenshtein distance between two usernames namely *rishabhk_* and *rk.iiit* is 8 and the shorter username length among them is 7 so the resulting normalized Levenshtein distance of 1.142 is obtained.
- Edit Similarity: This metric is similar to normalized Levenshtein distance but instead of dividing by the length of the shorter username, the Levenshtein distance is divided by thelength of the longer username.
- Keyboard Typing Distance: The approximate distance traversed on a standard QWERTY keyboard while typing out the username. This metric is obtained by calculating the euclidean distance between each character in the username with row and column of the key serving as its coordinates.
- LCS Similarity: Given two strings (display names) this metric is defined as the ratio of the length of the longest common string to the minimum length among the two strings. Its value lies between 0 and 1 and a greater value indicate a higher degree of similarity between the two display names.

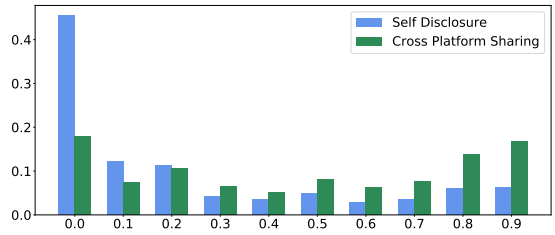
4.2 Behavioral Bias Characterization

To detect the existence of user behavioral biases in CPS and SD datasets, we study the distribution of lexical similarity features (jaccard similarity and edit distance) measured on user behaviors in terms of their configuration of username and display name. From the distribution of jaccard similarity values for usernames in linked identities obtained from CPS and SD in Fig 3a, it is evidently clear that almost 50% of linked identities in SD dataset have jaccard similarity on usernames value greater than or equal to 0.9 as compared to only 24% in CPS dataset. This clearly shows that users who self-disclose their usernames are *lexically more similar* than those who cross-post. When we change the measurement metric from jaccard similarity to edit distance, the trend of usernames (Fig 3b) coming from SD dataset exhibiting higher lexical similarity continues. This clearly shows evidence for existence of user behavioral biases in terms of configuring of their usernames and display names.

Next, we study the cumulative distribution of feature values for both linked and unlinked user identities obtained from cross-posting, self-disclosure and negative sampling. We first study the length of the longest common subsequence (LCS) in usernames at character-level. As depicted in Fig 4a, most (90%) of the unlinked user identities have LCS length less than 6. More proportion of linked user identities in CPS dataset have higher LCS length than those from SD dataset, a trend which is reversed in jaccard distance on usernames (Fig 4b) and normalized leveinshtein distance on usernames (Fig 4c). Given that both are distance variants, so the proportion of linked identities having higher distance in their usernames and display names is less. However, a significant gap between the blue and orange curves depicting linked identities from SD and CPS datasets clearly indicate presence of behavioral biases in these datasets. We say significant because when two sample S-K test was performed on these two distributions (blue and orange), then the p-value turned out to be less than 0.01 at significance level of α=0.05 with large D-statistic. This proves that distributions of jaccard distance and normalized leveinshtein distance for linked user identities in SD and CPS dataset are drawn from different distributions, consequently it establishes behavioral biases manifested through these lexical similarity metrics.

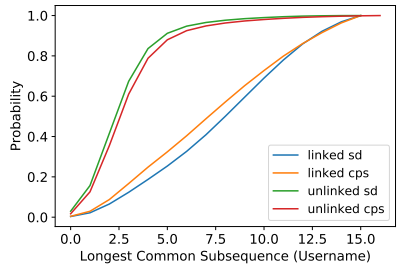


(a) Jaccard Similarity (JS) on User Names. 50% of user identity pairs obtained from self-disclosure have JS value in their usernames 0.9 as opposed to only 23% from cross-posting.

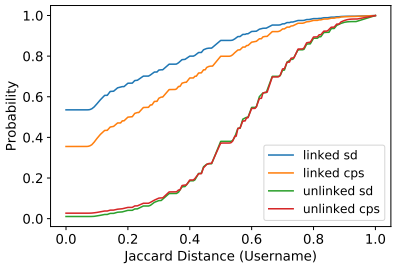


(b) Edit Distance (ED) on User Names. 45% usernames of user identity pair from self-disclosure have 0.0 ED than only 18% from those obtained from cross-posting.

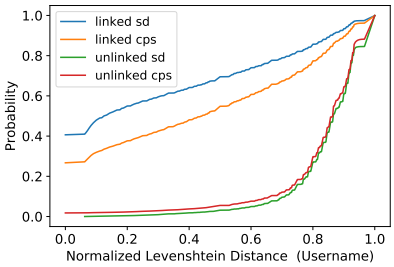
Figure 3: Distributions of Lexical Features (JS and ED) which clearly depict that usernames of user identity pair obtained using self-disclosure method exhibit higher lexical similarity than those obtained using cross-posting. Similar trends are observed in case of display names.



(a) CDF of Longest Common Subsequence on Usernames

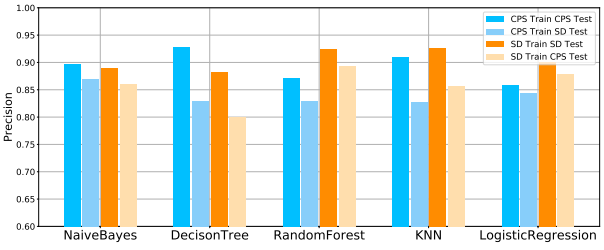


(b) CDF of Jaccard Distance on Usernames

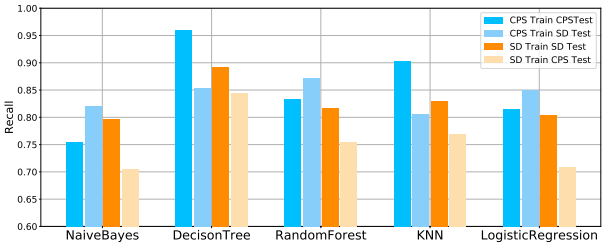


(c) CDF of Normalized Leveinshtein Distance on Usernames

Figure 4: Cumulative Frequency Distribution Plots of Lexical Features on User Names from Cross-Platform Sharing and Self-Disclosure Datasets. Similar trends are observed on Display Names.



(a) Precision Values of CPS and SD driven models



(b) Recall Values of CPS and SD driven models

Figure 5: Impact of Behavioral Biases in CPS and SD dataset on performance (precision and recall) of Classification Models

4.3 Impact of Behavioral Biases on Identity Linkage Models

In this section, we propose a methodology to study the impact of behavioral biases in classification models for identity linkage. After having proven the presence of behavioral biases and characterized them, we measure their impact on the decision making capability of identity linkage models. Recall from eq 1 that we solve identity linkage problem by constructing a classification model. A robust classification model ought to be *generalizable*, in other words, it is expected to perform equally well irrespective of the source of training data. To test the robustness of identity linkage models built in presence of behavioral biases that are manifested in lexical features derived from usernames and display names, we design four experiments. In *first* and *second*, we train the model using the same CPS dataset, but perform testing using CPS dataset & SD dataset, respectively. In *third* and *fourth*, we train the model using the same SD dataset and test it using SD dataset & CPS dataset, respectively. We consider five classification algorithms namely Naive Bayes,

Decision Tree, Random Forest, KNN and Logistic Regression in our experiments.

Fig 5a depicts the precision of these algorithms for all four experimental scenarios in predicting a linked identity pair to be correctly belonging to the same individual. As evident from Fig 5a, irrespective of the learning algorithm adopted, precision of models trained on CPS dataset & tested on CPS dataset is better than those tested on SD dataset. Similarly, the precision of models trained on SD dataset & tested on SD dataset is far better than those test on CPS dataset. This proves that the classification models get significantly biased with the dataset used to perform training. This could only happen when there are biases that exist in the dataset. In case of recall, depicted in Fig 5b, similar trend is observed in the case of models trained on SD dataset & tested on SD dataset outperforms those tested on CPS dataset. However, no conclusive trend is obtained in the case when models are trained on CPS dataset & tested on CPS and SD dataset, separately. This is due to the fact that linked identities in CPS dataset exhibit lower lexical similarities in the features than SD dataset. Consequently, when models are trained on CPS dataset, the training dataset is unable to provide the necessary discriminative training required for the model to be decisive.

4.4 Quantification of Bias

After detecting behavioral biases in user identities, characterizing them and measuring their impact on identity linkage models, we propose a *novel* methodological design that quantifies biases by leveraging from a well-established discrimination measurement approach namely *situational testing* [17]. Before explaining our approach, we briefly explain the concept of situational testing from the perspective of discrimination studies.

4.4.1 Situational Testing. In the context of discrimination studies, we refer a specific group of users as a *protected group* based on values of one or more *protected attributes* (like gender, race, locality, etc.) and the goal is to *protect* this group from discrimination based on protected attributes. As per situational testing, a data record (representing a user in the real world) is considered to be *discriminated* if a significant difference is observed in its treatment (prediction of a label in case of learning model-based decision making) with respect of its neighbors in protected group and neighbors not in the protected group. For illustration, consider a job suitable for both males and females, in which both males and females apply. And the job application process involves a stage in which a learning model-based applicant screening is adopted. Given that learning model is to be trained on historical decisions, so the biases (if any) that exist in the training data (in this case, say more males were offered a job in the past), are going to impact the learning model, make the decision outcomes of the model biased as well. Situational testing is an approach that quantifies such biases by leveraging K-Nearest Neighbor (KNN) classification technique. More formally, the following steps are performed in situational testing.

(1) Consider a dataset D of decision records, having n data instances $d_1, d_2, ...d_n$ and the class attribute represented by $class(d_i)$. In the above example, $class(d_i)$ can be either *accept* or *reject* job application.
(2) Consider a single protected attribute represented by *proc* which takes on categorical values. In the above example, *proc* attribute is *gender* taking on two values *male* and $female$ and the protected group ($P(D)$) is all females.

$$P(D) = d_i : proc(d_i) = female, \forall i = 1...n \quad (2)$$

Similarly, the unprotected group ($UP(D)$) becomes.

$$UP(D) = d_j : proc(d_j) = male, \forall j = 1...n \quad (3)$$

(3) Take a suitable distance function as required in KNN algorithm as f_{dist}, but define it only for non-protected attributes.
(4) For each test record d_{test}, find its K-nearest neighbors using f_{dist} in both protected group $P(D)$ and unprotected group $UP(D)$ in the training data. Accordingly, we define two variables for each d_{test} as below.
p_1: proportion of records in $P(D)$ with same decision as d_{test}
p_2: proportion of records in $UP(D)$ with same decision as d_{test}
(5) Lastly, we define $t = p_1 - p_2$, $-1 \leq t \leq 1$ and find the distribution of values of t which indicates the amount of discrimination with which each d_{test} gets affected. If $t = 0$, there is no discrimination but higher the values of t towards either -1 or $+1$, more is the severity with which the test record d_{test} is affected.

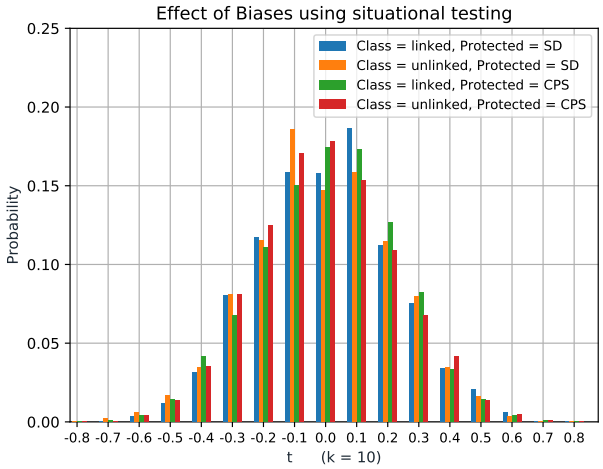


Figure 6: Effect of Biases on Linked and Unlinked User Identities in both scenarios when CPS and SD was taken protected group separately.

4.4.2 Novel Methodological Design. In this section, we discuss how we *apply* situational testing in the context of our problem of quantifying biases. To the best of our knowledge, this is the first work which is leveraging situational testing, which was originally proposed as a measurement methodology to study discrimination, to quantify biases. To adopt situational testing, we propose the following in our design.

(1) Create a combined dataset D which is drawn from two datasets of linked user identities namely D_{CPS} and D_{SD} obtained by leveraging cross-posting and self-disclosure user behavior, respectively. Alongside the linked user identities,

we also take the unlinked user identities as per the negative sample generation procedure explained earlier.

(2) While combining, create a new attribute *data_source* which would take values *CPS* or *SD*, and consider this new attribute as a *protected attribute* in order to apply discrimination measures in general and in particular situational testing.

(3) Perform two sets of experiments, first by treating records containing $data\_source = SD$ as a protected group, and second by treating records containing $data\_source = CPS$ as a protected group.

(4) The decision to be taken in the context of identity linkage problem is whether user identity pair belong to the same individual, referred as *linked* or different individuals, referred as *unlinked*. In all our experiments, we focus our attention on the decision of *linked*, unless otherwise stated.

Using the above methodological design, we are able to apply the concept of situational testing to study the quantification of behavioral biases in user identities. Next, we explain the results for different experimental designs.

4.4.3 Results. We design experiments to answer three questions, (1) *Are both decision classes (linked and unlinked) equally affected by biases?* To address this, we measure the impact of biases using the situational testing framework (keeping K=10 in KNN algorithm). We plot (Fig 6) probability distributions of $t - values$, on both class values namely *linked* and *unlinked* user identities and in both scenarios where $data\_source = SD$ and $data\_source = CPS$ are taken as a protected group, separately. It is clearly evident that probability distributions of $t - values$ are spread on both positive ($t > 0$) and negative ($t < 0$) sides which indicates that behavioral biases affect many test records.

(2) *What is the effect of the number of nearest neighbors (K in K-NN algorithm) on the severity of biases?* To understand the impact that K-Nearest Neighbors as per KNN algorithm have on the extent of biases, we plot (Fig 7) cumulative distribution of $p_1 - p_2$ values for different values of K. From Fig 7, it is observed that as the value of K increase, the amount of biases measured through cumulative distribution of $p_1 - p_2$ values decrease. This trend is consistent with every increase in the value of K. The intuitive explanation for this observation is that as we increase the value of K, we increase the probability of obtaining instances in training which belongs to both *linked* and *unlinked* classes.

(3) *Does the amount of training has an impact on the amount of biases suffered by test data instances?* We plot (Fig 8) the amount of biases suffered by test instances through probability distribution of $t - values$ for the varying amount of training data size, keeping the value of $K = 10$ and treating SD dataset ($data\_source = SD$) as a protected group. Looking at Fig 8, one concludes that biases exist

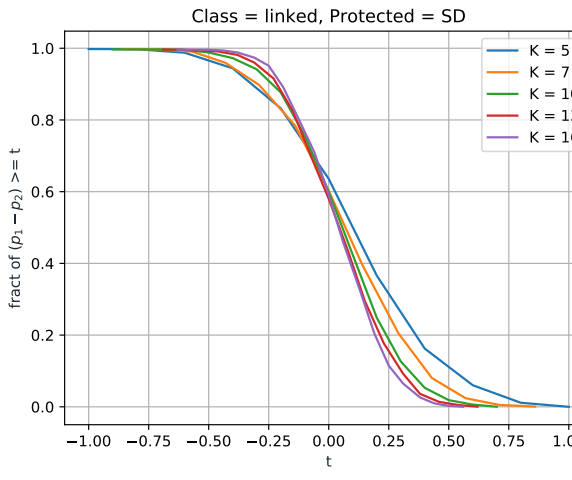


Figure 7: Effect of K-Nearest Neighbors on Biases on Linked User Identities when SD was taken protected group.

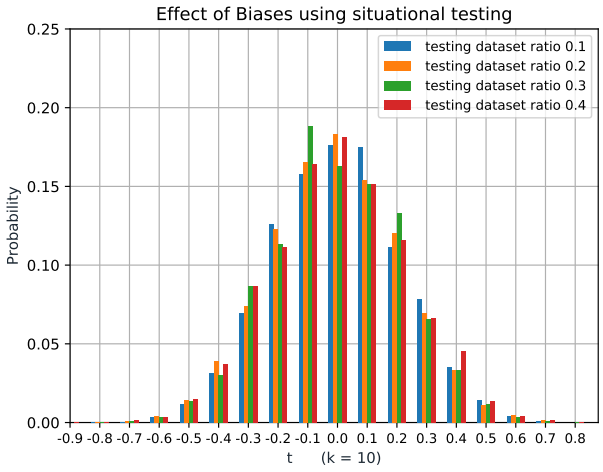


Figure 8: Effect of Biases on Linked User Identities with varying Training Sizes, considering SD as protected group.

across all scenarios irrespective of the amount of training dataset and the $t - values$ distribution follows a *normal distribution*. This rejects the intuitive notion that large amount of training would nullify the effect of biases.

5 DISCUSSION & LIMITATION

In this section, we explain our key observations and takeaways from our work.

User Behavioral Characterization: In regard to the presence of behavioral biases in user identities datasets, we find that users who indulge in self-disclosure intend to keep their usernames and display names more lexical similar as compared to those who cross-post. Behaviorally speaking, users who disclose [16] by mentioning details of their identities on other social networks in the bio-field of their Twitter identity, are conscious about making their identities appear similar and hence, their usernames & display names exhibit high lexical similarity across social networks. Users who cross-post, typically do it occasionally, and are not likely to be conscious to explicitly keep their identities similar, therefore their usernames & display names exhibit lessor lexical similarity. From the privacy standpoint, it would be good to have a system that nudges [10]

such users about linkability of their identities across multiple social networks so that they can make an informed decision about cross-posting.

Adoption of Best Practices: Further, it has been observed that researchers solving the problem of identity linkage have mostly relied upon a single data source to evaluate their proposed models. Therefore, a key takeaway from our work would be to re-evaluate the prior works in the light of biases that could be manifested in their dataset. Detection and neutralizing biases in user identity datasets need to become an essential pre-processing step before going ahead in evaluating proposed solutions.

Application of Discrimination Measures: Through this work, we have proposed an effective strategy to leverage discrimination measurement metrics to detect and quantify biases in dataset which are collected relying upon human behaviors. Just as we have leveraged discrimination measurement framework by considering *data_source* attribute as a protected attribute, we believe that works in solving problems in other domains through data-driven approaches would also benefit similarly.

Limitations in this work can be observed at two levels. At the *first* level, the fact that we study user behavior only in terms of usernames and display names, can be further extended to other profile attributes and also other behaviors in terms of content posting and networks that users keep. Given the increasing restrictiveness in the social network APIs, obtaining information about content and network would be a challenge, nevertheless. At the *second* level, the methodology for detecting, quantifying and preventing biases can be further strengthened by drawing more ideas from bias studies[6, 21], discrimination studies [7, 22] and fairness preserving algorithmic studies [2, 13, 24, 26].

6 CONCLUSION

We conclude that behavioral biases exist in user identity datasets obtained by leveraging cross-posting and self-disclosure. We study two user behaviors namely username and display name configuration, and find biases are present which get manifested in the form of lexical similarity features. Identity linkage models get affected by 5-20% when trained on data collected using the cross-posting method and tested on data obtained from self-disclosure and vice-versa. We quantify the extent of damage caused by these biases in terms of the number of biased decisions (15-20%) made by the classification models. Lastly, we implement an approach of data pre-processing which helps in mitigating the adverse affects of these biases, however at a significant cost of reducing model performance.

REFERENCES

[1] Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*. Springer, 43–57.

[2] Carlos Castillo. 2019. Fairness and Transparency in Ranking. In *ACM SIGIR Forum*, Vol. 52. ACM, 64–71.

[3] Wei Chen, Hongzhi Yin, Weiqing Wang, Lei Zhao, Wen Hua, and Xiaofang Zhou. 2017. Exploiting spatio-temporal user behaviors for user linkage. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 517–526.

[4] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 447–458.

[5] Oana Goga, Daniele Perito, Howard Lei, Renata Teixeira, and Robin Sommer. 2013. Large-scale correlation of accounts across social networks. *University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002* (2013).

[6] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Social Networks* 38 (2014), 16–27.

[7] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[8] Paridhi Jain and Ponnurangam Kumaraguru. 2012. Finding Nemo: searching and resolving identities of users across online social networks. *arXiv preprint arXiv:1212.6147* (2012).

[9] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. @ i seek'fb. me': Identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1259–1268.

[10] Rishabh Kaushal, Srishti Chandok, Paridhi Jain, Prateek Dewan, Nalin Gupta, and Ponnurangam Kumaraguru. 2017. Nudging nemo: Helping users control linkability across social networks. In *International Conference on Social Informatics*. Springer, 477–490.

[11] Rishabh Kaushal, Vasundhara Ghose, and Ponnurangam Kumaraguru. 2019. Methods for User Profiling Across Social Networks. In *Proceedings of the 12th IEEE International Conference On Social Computing (SocialCom 2019)*. IEEE, 100–108.

[12] Xiangnan Kong, Jiawei Zhang, and Philip S Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 179–188.

[13] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[14] Yongjun Li, You Peng, Zhen Zhang, Quanqing Xu, and Hongzhi Yin. 2017. Understanding the user display names across social networks. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1319–1326.

[15] Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. What's in a name?: an unsupervised approach to link users across communities. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 495–504.

[16] George Loewenstein, Cass R Sunstein, and Russell Golman. 2014. Disclosure: Psychology changes everything. *Annu. Rev. Econ.* 6, 1 (2014), 391–419.

[17] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 502–510.

[18] Anshu Malhotra, Luam Totti, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. 2012. Studying user footprints in different online social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 1065–1070.

[19] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[20] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. 2011. How unique and traceable are usernames?. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 1–17.

[21] Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *ACM SIGMOD Record* 46, 4 (2018), 16–21.

[22] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[23] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter* 18, 2 (2017), 5–17.

[24] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* (2015).

[25] Reza Zafarani and Huan Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 41–49.

[26] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1569–1578.

[27] Xiaoping Zhou, Xun Liang, Xiaoyong Du, and Jichao Zhao. 2018. Structure based user identification across social networks. *IEEE Transactions on Knowledge and Data Engineering* 30, 6 (2018), 1178–1191.

[28] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).