Other Times, Other Values: Leveraging Attribute History to Link User Profiles across Online Social Networks*

Paridhi Jain*, Ponnurangam Kumaraguru*, Anupam Joshi†

*Indraprastha Institute of Information Technology (IIIT-Delhi)
{paridhij, pk}@iiitd.ac.in
†University of Maryland, Baltimore County (UMBC)
joshi@cs.umbc.edu

Abstract. Profile linking is the ability to connect profiles of a user on different social networks. Linked profiles can help companies to build psychographics of its potential customers and segment them for targeted marketing in a cost-effective way, can help advertisers target personalized ads and can help security practitioners capture detailed characteristics of malicious / fraudulent users. Existing methods link profiles by observing high similarity between most recent (current) values of the attributes like name and username. However, for a section of users who are observed to evolve their attributes over time and choose dissimilar values across their profiles, these current values have low similarity. Existing methods then falsely conclude that profiles refer to different users. To reduce such false conclusions, we suggest to gather rich history of values assigned to an attribute over time and compare attribute histories to link user profiles across networks. We believe that attribute history highlights user preferences and behavior while creating attribute values on a social network. Co-existence of these preferences across profiles on different social networks result in alike attribute histories that suggests profiles potentially refer to a single user. Through this study, we quantify the importance of attribute history for profile linking on a dataset of real-world users with profiles on Twitter, Facebook, Instagram and Tumblr. We show that attribute history correctly links 48% more profile pairs with non-matching current values that are incorrectly unlinked by existing methods. We further explore if factors such as longevity and availability of attribute history on either profiles affect linking performance. To the best of our knowledge, this is the first study that explores viability of using attribute history to link profiles on social networks.

1 Introduction

Over the past decade, Online Social Networks (OSNs) have become a vital ingredient of an individual's online life. They help her exercise freedom of speech,

* An early version of this manuscript appeared in the 2015 ACM Conference on Hypertext and Social Media (HT) [1].

relations, hobbies, interests, and creativity. For instance, Twitter helps her express opinions about news events and campaigns, Facebook helps reviving and building personal relations, Instagram and Pinterest promote creativity by allowing her to share pictures. In order to enjoy these services simultaneously, a user creates profile on each of these OSNs. During registration, she creates an identity for herself listing personal information and connections. Due to distinct purpose, requirements and policy of each OSN, quality and veracity of her identity vary with the OSN. This results in dissimilar identities of the same user, scattered across Internet, with no explicit links directing to one another.

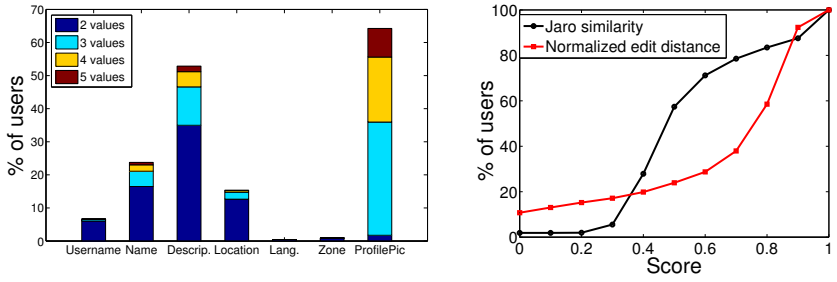
Disparate profiles, scattered across multiple OSNs, are often needed to be linked together to benefit various stakeholders. An individual can understand privacy issues associated with aggregated and inferred information from her linked profiles. On knowing the privacy leaks that are exploited by proposed and existing profile linking approaches, the user can then patch the leaks to avoid the inferred connection. Social media marketers can estimate correct audience size by avoiding double-counting users engaged in their campaigns via multiple OSN profiles. [1] Organisations like Disney and PepsiCo, that carry out psychographic segmentation based upon customers' activities, interests, opinions and lifestyles to adapt marketing strategies on their needs and wants, can benefit from linked profiles. [2] It is shown that psychographic segmentation is the most effective segmentation citing a rise of 24% in business performance [2]; however includes high cost in both time and money [3]. The cost of constructing psychographics of each customer can be brought down with the use of her linked social profiles. [3] Beyond bringing the cost down, linked social profiles also help organizations segment their market on unknown attributes such as location, gender or age. These attributes are rarely available within one social network for all users but can be collected by linking user profiles across multiple social networks. Therefore, we believe that profile linking is an important and relevant for businesses, security agencies and the users themselves. We therefore, in this work, aim to build methods that connect profiles of a user across multiple OSNs and term the process as *profile linking*.

Existing profile linking methods compare attributes like username and name to find connection between a pair of profiles. However, challenges like dissonant social platforms with partially overlapping list of supported attributes and heterogenous attributes holding veracious values impede effective profile linking. Literature suggests various methodologies that compare common attributes between examined user profiles and evaluate similarity between corresponding values on different metrics. Similarity between text attributes like name is estimated using Jaro similarity while media attributes like profile-picture are compared using face detection algorithms and histogram matching [4–8]. These methodologies consider most recent (current) values of the attributes and assume high

[1] http://www.clickz.com/clickz/column/1716119/the-five-biggest-mistakes-measuring-social-media

[2] http://thewaltdisneyco.blogspot.in/2011/11/chapter8segmentingtargetingmarkets.htm

[3] http://www.marketingtechnews.net/news/2012/mar/16/how-social-media-influencing-marketing-segmentation/



(a) (b)

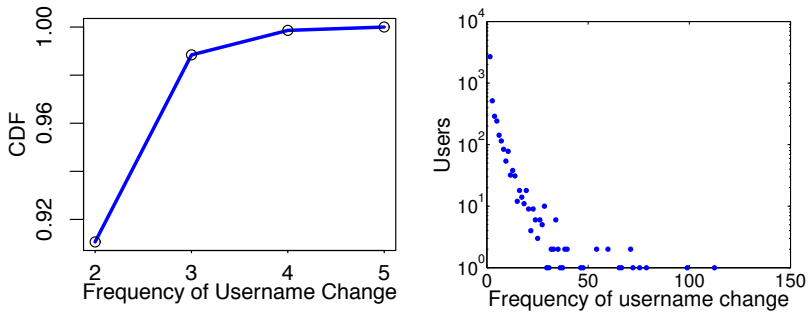
Fig. 1. Attribute evolution on Twitter. (a) Around 73.21% users tend to change their attributes on Twitter. (b) Users who evolve their username have low similarity between usernames across their profiles. For these users, attribute history can be leveraged for profile linking.

similarity to infer a link between respective profiles. However, current values may have low similarity for reasons such as user's choice to maintain privacy or attribute evolution over time as described now [9, 10].

User's choice: Characterization studies on OSNs suggest that users consistently keep same values for their attributes like their name, gender, location, across OSNs [11, 12]. Zafarani *et. al* shows that 59% users create similar usernames across their profiles for reasons like to represent a universal identity in online space or to ease remembering [13]. Remaining 41% users choose dissimilar usernames for reasons such as to maintain privacy and avoid de-anonymization [14]. For this section of users, existing profile linking methodologies that assume high similarity between current values of the attributes across OSN profiles may fail to conclude that profiles refer to a single user.

Attribute evolution: Recent studies that examine temporal nature of OSNs suggest that users exhibit a tendency to evolve their attributes over time [10, 15, 16]. Consider the following scenario – A user registers on Twitter and Facebook with the same username value; she updates her Twitter profile more frequently than her Facebook profile; she chooses a new username on Twitter, not similar to the old one but makes no such changes on Facebook. Due to evolution of username over time on a favored social network, she now owns dissimilar usernames on her profiles. On observing dissimilarity, existing methods that match only username falsely conclude that Twitter and Facebook profiles refer to different users.

To validate if a significant section of users change attributes, we deploy an automated system to track 8.7 million Twitter users every fortnight and record changes to their attributes. Figure 1(a) shows the distribution of users that evolve over time and hold distinct values for their attributes. On a two month period, we observe that 73.21% users changes their attributes and assign distinct values.



(a) (b)

Fig. 2. Frequency of username change on Twitter (a) for 10% of 8.7 million users tracked for 2 months, (b) for random sample of 10K users tracked for 14 months.

Thereby, we gather that attribute evolution is an evident phenomenon. Further, we test if evolution causes dissimilar current values across profiles of users and hence, filter users who evolve their usernames. We compute Jaro similarity and Edit distance between current usernames on their profiles and plot the user distribution (see Figure 1(b)). Observe that 78% users have usernames with Jaro similarity < 0.7 and 62% users with Edit distance > 0.7 implying dissimilar current usernames across profiles for a majority section of users due to username evolution. Low similarity between current usernames can be falsely manipulated by existing methods as different users.

We also examine the frequency of username change and observe that most users changed their username atleast once, in the tracking duration of two months (see Figure 2(a)). For another set of 10K users, tracked for about 14 months, we observe that few users changed multiple times while many a few times. We observe that username changing behavior follows Pareto Principle (see Figure 2(b)).

For users who evolve and select dissimilar attribute values across profiles, we propose to asset rich information created due to their tendency towards evolution i.e. past values. These past values created by a user, termed as *attribute history*, reveal her preferences and consistent behavior responsible for structuring the values. Preferences like her choice of length, characters, lexical and morphological structure, frequency of reuse of the values can co-exist across her profiles on different OSNs, thereby creating similar attribute histories in terms of syntactic, stylistic and temporal characteristics. Similarities between attribute histories across OSN profiles can suggest a potential link to a single user.

Scope: A user profile is composed of multiple attributes; each signifies a unique characteristic of the user. Among the attributes, literature suggests username to be an important and discriminating attribute for profile linking [17–19]. Though a considerate section of users change username on Twitter (≈10%), it is the most common publicly available attribute across OSNs that can uniquely iden-

tify users within an OSN. In addition to availability and uniqueness, usernames can only contain alphanumeric and special characters irrespective of the preferred language of the user profile, thereby allowing clean string comparisons. We, therefore, choose to track changes to *username*, collect a set of values, and use the value set for profile linking. History of other attributes like name, description and profile-picture can further help in identifying user profiles of the same user; however lack of their universal support, availability across social platforms, and API restrictions on their access direct us to limit our scope to only usernames. For this study, we ask following research question: *Given two user profiles and respective username histories on a pair of OSNs, can we predict that profiles belong to the same user?*

Contributions: On a labelled dataset of 128,251 pairs of username histories accessed from users with profiles on four popular social networks – Twitter, Instagram, Facebook and Tumblr, we examine viability of using username history for profile linking and impact of various factors that govern its effectiveness. We show that:

- Out of 89.34% profile pairs that current values fail to link, a comparison of username histories correctly links 48.47% profiles pairs while keeping a false positive rate of 3.71%. Therefore, attribute history helps profile linking.
- Out of 40.87% profile pairs that were misclassified by syntactic, stylistic and temporal features of username histories, 95.84% attribute to Twitter-Tumblr profile comparisons while 5.50% results from Twitter-Instagram and Twitter-Facebook profile comparisons.
- Availability of username history only on one profile increases false linkings by 12% as compared to its availability on both.
- Importance of username history for profile linking increases with history longevity.

To the best of our knowledge, this is the first study that provides insights in estimating the use of attribute history of user profiles on social networks for profile linking. We believe that attribute history can also help other applications that build on derived behavioral characteristics of users.

2 Problem Statement

We now formally define the research question using following definitions and notations. User profiles under examination belong to a pair of social networks, SN_A and SN_B, termed as *source profile* S and *candidate profile* C, respectively. An evolved username set U is a set of pairs, where each pair contains new value and time of evolution of the attribute, ordered on the time of evolution i.e. $U = \{(u_1, t_1), (u_2, t_2), \cdots, (u_L, t_L)\}$, where $t_i < t_{i+1}$. Here, L denotes the length of the username set, t_1 denotes the time when first username change is recorded and t_L denotes the time when the last username change is recorded; u_L represents the most recent (current) value. Username sets on source and candidate

profiles are denoted by U_S and U_C, respectively. If past usernames of the candidate profile are *not* available, set U_C is replaced by the current username u_c. We define our problem as –

Problem Statement: *Given a source profile S on SN_A, a candidate profile C on SN_B and their respective username sets U_S and U_C, each composed of pairs of usernames and their respective evolution timestamps, find if U_S and U_C refer to the same user $\mathcal{I}$.*

A collection of methods can solve the problem. Heuristic approaches like rule based methods, collaborative approaches like crowd sourcing and manual tagging, and algorithmic approaches like machine learning can look for similarities between username sets and infer the potential link between them. We model profile linking as a classification problem with three phases – feature extraction, labelled dataset collection and supervised machine learning framework for correct profile identification (see Figure 3). Features extract similarities between usernames across username sets by capturing unique behavioral characteristics and consistent preferences that a user exhibits while choosing usernames across her profiles over time (Section 3); Labelled datasets collect users who evolve with profiles on popular social networks (Section 5) followed by supervised classification by an ensemble of classifiers organized in a framework (Section 6). Section 7 describes relevance of our study to related literature. Finally, Section 8 presents discussion and Section 9 concludes this research with directions for future work.

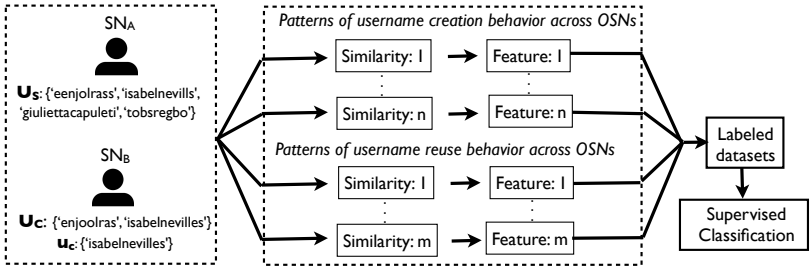


Fig. 3. Proposed methodology to compare username sets and capture similarities based on unique behavioral patterns while creating and reusing usernames over time.

3 Features

Individuals often maintain unique preferences and consistent behavior while creating attribute values across their profiles on different social networks. Cross-OSN analysis of users on social media shows that 85% users have more then 50% matching attribute values across different OSNs [12]. These attributes, however,

evolve over time, leading to matching histories than current values. Further, a recent study shows that users exhibit similar choices while selecting usernames across OSNs [19]. We believe that such choices may repeat over time and can co-exist across OSNs. On a granular note, choices can be segmented further in three categories – *syntactic*, *stylistic*, and *temporal*.

Syntactic choices govern the composition of the usernames like choice of length, characters, or arrangement, stylistic choices regulate the liguistic structure of the usernames like choice of abusive words, slangs, leetspeak, upper and lowercase characters, while temporal preferences supervise timely reuse of the usernames in either exact or modified form across OSNs. Co-existence of these choices within and across OSNs leads to similar username histories.

3.1 Syntactic features

Syntax choices while creating usernames on one's profiles are affected by self-bias and limited memory. These push an individual to deploy similar username compositions across her profiles resulting in username creation patterns. These can either remain static or change with time as per the need of the users. We capture both static and evolutionary username creation patterns, and list methods to quantify them into features.

Static creation: On OSNs, users converse by tagging another user's username with '@' tag. Tagged user specifies username properties that aid these interactions. For instance, a user chooses short usernames on OSNs that restrict message length in order to help her friends post more content when tagging her[4]. Properties that do not change over time for new usernames constitute static patterns. We capture three string properties – length, choice of characters, and the arrangement of characters. It is likely for a user to create usernames of similar length with a limited set of characters compiled in similar fashion. For both source and candidate username sets, we calculate these properties and compare using different methods.

Length of a username l_{u_i} is calculated by counting alphanumeric characters in the username. Length distribution of usernames in source $\mathcal{L}_S$ is compared with that of usernames in candidate username set $\mathcal{L}_C$ using JS divergence. The low divergence hints use of similar username lengths across OSNs.

To compare choice of characters, we compare character distribution of usernames in source $\mathcal{C}_S$ with that of usernames in candidate username set $\mathcal{C}_C$ using Jaccard similarity index J and cosine similarity cos. The best value at '1' for both metrics implies same choice of characters on username sets, made by the same user.

To compare the arrangement of characters, we compute string similarity between usernames of different sets. We calculate normalized Longest Common Subsequence (LCS) similarity score between u_i and u_j such that u_i, u_j belong to different sets and estimate mean, median and standard deviation of score

[4] www.mediabistro.com/alltwitter/twitter-username-tips_b12367

distribution $\mathcal{A}$. The low standard deviation of the distribution hints similar arrangement of characters likely to be made by the same user while high mean and median values denote the high similarity among usernames in the two sets. In a nutshell, static features are:

$$F_{static} : (JS(\mathcal{L}_S||\mathcal{L}_C), J(\mathcal{C}_S, \mathcal{C}_C), cos(\mathcal{C}_S, \mathcal{C}_C), \mathbb{E}(\mathcal{A}), med(\mathcal{A}), \sigma(\mathcal{A}))$$

Evolutionary creation: With changing requirements on an OSN like privacy concerns, a user can consider to change a few properties of new usernames she creates within an OSN. For instance, user can start using initials over full name in her username, thereby anonymizing and shortening its length. It is likely that her new preferences influence usernames created on other OSNs as well. Similar transitions in the properties of usernames created across OSNs result in similar evolutionary patterns of properties. We capture such patterns by comparing evolution sequence of the username properties computed for each username set.

Consecutive usernames of each username set are compared on length, character distribution and arrangement of characters, resulting in three comparison vectors for each set – length, character, and arrangement vector. Length vector $\mathbb{L}$ is a sequence of lengths l_{u_i}, character vector $\mathbb{C}$ is a sequence of Jaccard index and cosine similarity scores between character distribution while arrangement vector $\mathbb{A}$ is a sequence of string similarity scores between consecutive usernames of a username set. For the arrangement vector, we use four string similarity metrics – Edit distance, Jaro similarity, LCS similarity and Longest Common Substring similarity (LCSub). Multiple similarity metrics ensures different penalties for character insertion, deletion and replacement. Normalized versions of string similarity scores are used in the arrangement vectors.

Length, character and arrangement vectors for two username sets are compared to find any correlation between the two sets. We use normalized cross-correlation (NCC) to compute the correlation, whose values ranges from -1 to 1. This metric is used to find correlation between two time series data lists as a function of lag τ at which the time series best align each other, also used for temporal analysis on Twitter in [20]. A positive correlation implies similar pattern of evolution of the username property on both username sets, from which we may link username sets to the same user. In a nutshell,

$$F_{evolution} : (NCC(\mathbb{L}_S, \mathbb{L}_C), NCC(\mathbb{C}_S, \mathbb{C}_C), NCC(\mathbb{A}_S, \mathbb{A}_C))$$

3.2 Stylistic features

Literature suggests that users create non-similar profiles across OSNs in order to maintain privacy and anonymity. These users avoid using a rule or a syntax to create usernames across OSNs, rather choose usernames that sync with their projected identity on the OSN. Extensive work in authorship analysis suggests that in cases where the text differs, users often maintain their writing style. Grant *et. al* shows how writing styles can help link anonymized SMS texts to known authors / users [21]. With this motivation, we believe that user's style

choices can still repeat across distinct and dissimilar usernames. We capture similarities between linguistic styles with which a user creates her usernames across OSNs. Before extracting features, we split each compound username into a set of words that constitute it. We describe our stylistic features as follows:

- **Case (Cs)** captures the use of UPPERCASE, Titlecase, ToGgLeCaSe, and PascalCase in a username as a binary vector. Maximum jaccard index between binary vectors of two usernames belonging to different username sets returns a stylistic feature.
- **LeetSpeak** (LS) captures the use of leet in username. Users can choose to replace a character in their username with a leet symbol for reasons such as to make an available version of wished username or to avoid keyword search with username. We identify 20 leet symbols in a username, some of them are listed in Table I. Usage of any leet symbol in two usernames belonging to different username sets compose a stylistic feature.
- **Emphasizer** (Em) captures the user's style of stressing on certain alphabet in her username. Two stylistic features, one captures if the user consistently stress on an alphabet in her usernames across OSNs and other captures if the user stresses on the same set of characters in most usernames she creates.
- **Prefix (Pf) / Suffix** (Sf) captures user's tendency to start or end her usernames in a specific way. A stylistic feature that captures if common prefixes or suffixes are just in creation of the usernames across OSNs. Further, a match between prefixes of one username with suffix of another username indicates that user intends to use same words to either start or end a username. We capture three features here.
- **Slangs** (Sw) use denote the tendency of user to use short forms, acronyms, internet chat jargons in their username for reasons like space limitation or non-availability of wished username. User's choice to use same slangs or any slang across her usernames indicates her stylistic consistency across OSNs. We capture the set of slangs commonly used as well as the presence of slangs in two stylistic features.
- **Bad words** (Bw) in a username imply the user behavior of abusing or expressing aggression towards a topic or a user. Presence and choice of bad words is captured using two stylistic features.
- **Function words** (Fw) imply the use of common stop words to mark association between words. A frequent and consistent use of same function words across usernames on OSNs highlight the user's way of writing. We capture presence of function words and the common use of same function words as stylistic features.
- **Phonetic replacement** (Pr) is often a choice of users when they wish to amend the spelling of a word with its phonetic equivalent. Another stylistic feature captures this tendency.
- **Grammar** (G) is an important linguistic feature of text. It denotes a user's tendency towards use of specific grammatical elements such as nouns, adjectives, etc. in the username. A binary vector captures the presence of 36 elements and a jaccard index calculates their consistent use across usernames.

Table II list examples of each stylistic feature we capture. In summary, we list possible similarities between username sets resulting from synchronous user behavior when selecting usernames within and across OSNs over time. Discussed methods quantify these similarities into a set of 15 stylistic features; all features are normalized between [0, 1]. In a nutshell, features are:

$$F_{stylistic} : (J_{max}(Cs_{S_i}, Cs_{C_j}), LS_{\{0,1\}}, Em_{\{0,1\}}, J_{max}(Em_{S_i}, Em_{C_j}), J_{max}(Pf_{S_i}, Pf_{C_j}), J_{max}(Sf_{S_i}, Sf_{C_j}), J_{max}(Pf_{S_i}, Sf_{C_j}), Sw_{\{0,1\}}, J_{max}(Sw_{S_i}, Sw_{C_j}), Bw_{\{0,1\}}, J_{max}(Bw_{S_i}, Bw_{C_j}), Fw_{\{0,1\}}, J_{max}(Fw_{S_i}, Fw_{C_j}), Pr_{\{0,1\}}, J_{max}(G_{S_i}, G_{C_j}))$$

Table I. Few of the leet symbols identified in a username.

Leet symbol	0	1	3	4	5	7	8	9	z	x	0rs	xck	0rz
Corresponding character	o	i	e	a	s	t	b	g	s	a	s	uck	ers

Table II. Examples of consistent user style during usernames created over time and across OSNs. Here, each pair of usernames belong to a single user. Italicised letters highlight the feature presence.

Case	'*C*upcake*G*awd','*F*oodie*F*luency'	**Slangs**	'*idk*narryisperf','um*idk*isabel'
LeetSpeak	'JLSInspireM*3*','dissapp0int*33*d'	**Bad words**	'_You*Fuck*Up_','*ugly*group2014'
Emphasizer	'Februha*rryy*','fvcck-yo*uu*'	**Function words**	'*the*dazefaze','fucc*the*hype'
Prefix	'*0dd*ace','*odd*fuckingace'	**Phonetic repl.**	'homiesexuall','aerogance'
Suffix	'TDush*Cox*', 'tonyd*cox*'	**Grammar**	'kissmetravis','givemelov3'

3.3 Temporal features

With an increasing number of OSNs and evolving preferences, a user struggles to remember her latest usernames on all OSNs in order to sign in or use the usernames for interactions. However, a naive reuse of a username borrowed from her other OSN profiles can ease her cognitive load [19]. Reused username can either be a latest username or an old username from any of her OSN profiles. Frequent tendency to reuse a username from other profiles results in a set of common usernames appearing in the same order at the same time across user profiles indicating user synchronous behavior across her profiles.

Occasional reuse: User's choice of reusing a username from her other profiles at least once results in observing that username on different profiles at different times. To find the common username, we intersect username lists extracted from

each username set. If the intersection results in an empty set, there is a possibility that the username she wants to use is already taken by a different user within the OSN. In that case, user can make minor modifications to the selected username to create an available version and use the available version on the OSN. With minor modifications, selected username and its available version have a high string similarity score. We, therefore, perform pairwise comparisons between usernames from different sets to find best matching username pair,

$$\max_{(u_i,t_i)\in U_S,(u_j,t_j)\in U_C} Sim(u_i, u_j)$$

We compute the similarity based on four string based metrics – edit distance, jaro similarity, LCSub similarity and LCS similarity. We acknowledge that the existence of a common username or a pair of similar usernames between two username sets can be co-incidental. It is likely that different users pick the same username at some point in their past. This can happen to usernames derived from celebrity, brand or popular names. Therefore, we calculate second best similarity score between usernames from different sets. A low second best similarity indicates that the best similarity can be an outlier, implying that username sets refer to different persons.

Frequent reuse: Repeated use of borrowed usernames results in a set of common usernames between profiles of a user. We examine if there exists a set of common usernames and compute a boolean feature. We estimate the ratio of common usernames to the size of smaller username set which denotes if all (or few) usernames are copied from other OSN profiles. A sequential and simultaneous use of common usernames across OSNs lends support to the belief that username sets refer to the same user. It is highly unlikely for different users to choose same usernames in the same order at the same time across multiple OSNs. Further, similar sequential ordering of common usernames in both sets is an indicator of a single user consistently choosing same usernames over time across her profiles. Earlier research suggests Smith-Waterman algorithm as an effective algorithm to measure sequential ordering [4], originally proposed to perform sequence alignment in protein sequences [22]. We use Smith-Waterman similarity to estimate sequential ordering between common usernames in the username sets. To capture temporal synchrony, we use timestamps of evolution to find if same usernames are used on both sets at the same time.

As described earlier, users may make minor modifications to a selected username, in order to create an available version to use on the OSN. We incorporate such minor modifications while calculating set of common usernames. We consider two usernames as variations of the same username, if LCS string similarity is above a threshold. We adjust the threshold from 0.8 to 1 and compute set of common usernames and other features accordingly. Comparing username sets on common usernames, their ordering and concurrent use, we calculate five features – boolean feature capturing if a username set is a (partial) subset of another set, common usernames, ratio of common usernames to smaller set size, boolean

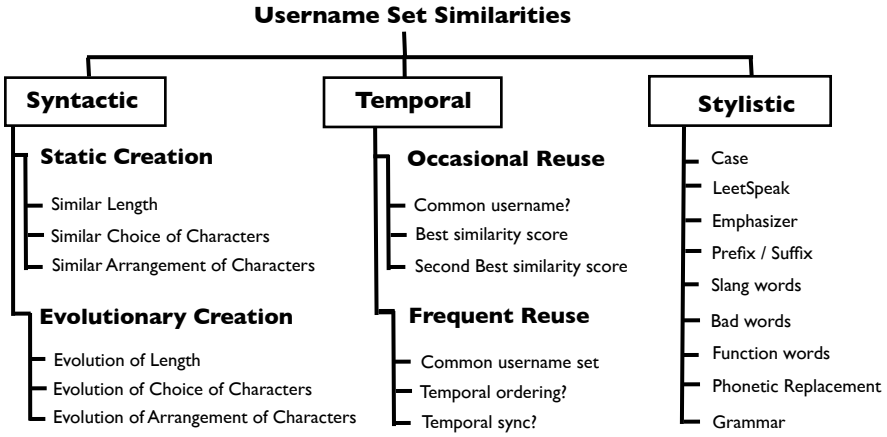


Fig. 4. Syntactic, Stylistic and Temporal similarities captured between username sets corresponding to examined user profiles.

feature capturing sequence alignment, and boolean feature estimating temporal synchronization.

We extract 13 syntactic, 15 stylistic and 13 temporal features from a labeled dataset of username sets, learn a supervised classifier and use it to predict connection between test username sets. In scenarios where past usernames are accessible only on one user profile, we compute syntactic static, occasional reuse and stylistic features between the source username set on source profile and candidate current username.

4 Framework

We experiment with three plausible supervised frameworks – Independent, Fusion and Cascaded framework.

4.1 Independent framework

Most profile linking approaches use a feature set, labelled datasets and a single classifier to predict link between test profiles [6, 17–19]. Classifier decision is not revised further either manually or computationally. We experiment with such a framework by learning a supervised classifier on proposed features extracted from username sets in the labelled datasets (see Figure 5(a)). However, we suspect the dominance of a subset of features that extract similarities between histories than current values. Hence, trained classifier can be biased towards finding similar histories and can falsely label username sets with dissimilar past but similar current values as negative. To avoid this, we suggest fusion and cascaded frameworks.

4.2 Fusion Framework

Fusion framework is an ensemble of four classifiers, one trained on current username features and three trained on syntactic, stylistic and temporal username set features. Each classifier is learned using a common training split and evaluated on a testing split. Decision of each classifier is then either 'ORed' or fed into a weighing scheme to predict the label of username sets derived from two examined user profiles (see Figure 5(b)). Ensemble frameworks are proven to be efficient classifiers though we suspect that a single training to fusion framework can result in overfitting. The reason is that training instances vary their richness with the genre of features considered. To avoid the same, we formulate a cascaded framework, thereby enriching training at each step.

4.3 Cascaded framework

Cascaded framework is an ensemble of two classifiers trained on different features to uncover link between two profiles and is extensively used in machine learning domain [23]. **Classifier I** extracts current username features and uses an existing method to classify username sets while **Classifier II** extracts syntactic, stylistic and temporal features from username sets and uses a supervised classifier to re-classify username sets labelled as negative by **Classifier I** (see Figure 5(c)). We train **Classifier II** with the false negatives of **Classifier I**, thus ensuring the richness of the training instances in features required for the accurate classification. We further experiment with two existing profile linking methods as **Classifier I** and different supervised classification techniques as **Classifier II** of the framework. These existing methods act as baselines, also used in [17, 19] to evaluate performance of the suggested features:

- **Exact matching (b1):** Links two username sets if current usernames are an exact match.
- **Substring matching (b2):** Links two username sets if substring similarity score between respective current usernames is beyond a threshold. We use Jaro similarity score to compute substring similarity, and vary the threshold to report best achieved accuracy.

5 Data Collection

For a positive dataset, we need to know accounts of a user on multiple OSNs. We start with Twitter, choose a random set of users and find their profiles on three popular social networks that contain quality information about a user[5] – Facebook, Instagram, and Tumblr. All networks, except Facebook, allow multiple changes to username. Facebook allows username change only once.

[5] http://mashable.com/2013/04/12/social-media-demographic-breakdown/

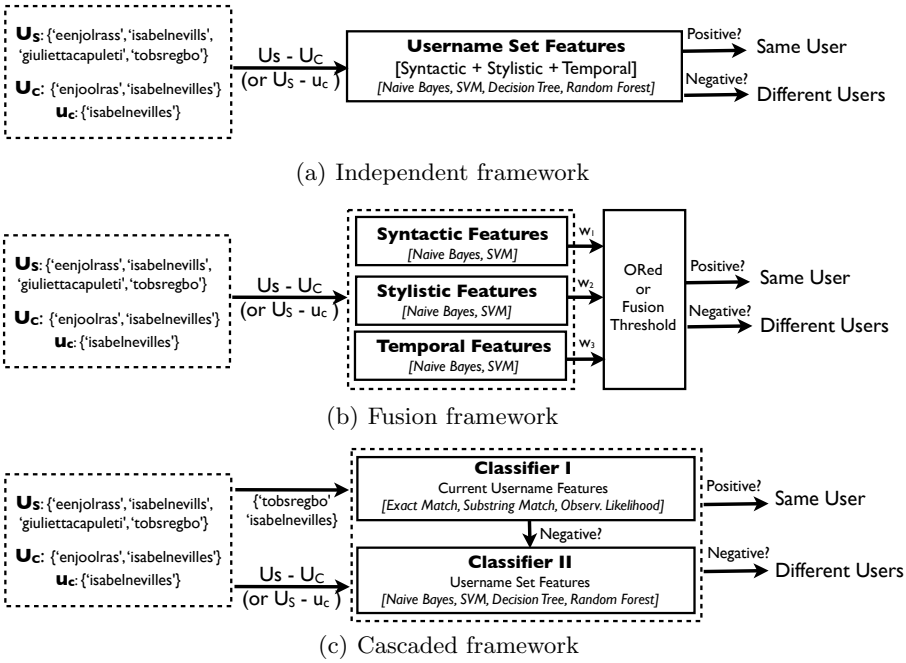


Fig. 5. Independent, Fusion and Cascaded framework. Independent framework uses proposed features independently; fusion framework uses weighted decisions of classifiers trained on different sets of proposed features and cascaded framework uses proposed features for re-classification.

Ground Truth: One way to identify other OSN profiles of selected Twitter users is manual, which is cumbersome and time-consuming. Another way is to exploit the tendency of users to broadcast hyperlinks to other OSN profiles via URL attribute of their Twitter profiles [24]. Such users *self-identify* themselves on other OSNs. For instance, a user posts *www.facebook.com/username* or *www.facebook.com/username/pictures/01* on her URL attribute, thereby informing other Twitter users about her Facebook profile. Similar methods are used in literature to create positive datasets either from social aggregation sites, forums or social networks where users *self-list* their OSN accounts [19]. To further validate the truthfulness of the ground truth, we manually checked user accounts of about 100 users to confirm that their Twitter URL refers to their other own OSN accounts.

Username History: Once user profiles are identified across OSNs, we collect past usernames owned by the user profiles. We build an independent tracking system for Twitter to monitor any changes to 8.7 million randomly chosen Twitter profiles as on October 2013. Tracking system repeatedly query Twitter Search API with *user_id* of the user profile after every fortnight and store responses

mentioning username, name, URL and similar details user owns at the time of the query. The system then compare consecutive API-responses to take a note of any changes to usernames, names, URLs, etc. Unique usernames chosen by the queried user profile over the tracking period of October 1, 2013 to November 26, 2013, constitute past username set on Twitter. Note that, the system collects only publicly available data available on social networks and does not engage in any user authorization asking for private data.

To gather past usernames used on other OSN profiles of the user, one can deploy a similar independent tracking system to track each OSN profile. However, configuring and deploying a tracking system for each OSN requires extensive infrastructure.[6] To reduce infrastructure costs, we use an alternate way to record username changes on other OSNs while tracking Twitter. We record any changes to URL attribute of the Twitter user profile to mark any changes to her username on other OSN. For instance, a Twitter user changes her URL attribute from *www.instagram.com/happygu!* to *www.instagram.com/gulben!* to notify Twitter followers (or others) about the username change on Instagram. We exploit this method to record username changes on users' Facebook, Instagram or Tumblr profiles. We also time username changes on the social networks. Other methods to collect past usernames are discussed in Section 8.

Pre-processing: Recorded usernames on Twitter, Facebook, Instagram and Tumblr profiles are processed prior comparison. Usernames on most social networks are case-insensitive, therefore, usernames are converted to lower case. Further, different OSNs allow a different set of special characters in the usernames. Twitter allows underscore '_', Tumblr allows the hyphen '-', Instagram and Facebook allow dot '.'. A user's wish to reuse a past username on other OSN in its exact form can be restricted by the use of special characters. She needs to replace the special characters with those allowed on the other OSN. To avoid low similarities or miss exact username matches between two username sets, we remove special characters from the usernames. Since no feature captures choice of special characters, their removal will not affect our results.

Dataset: For experiment purposes, we use Twitter profile as a source profile and the corresponding username set as a source username set U_S. We use other OSN profile (Tumblr, Facebook or Instagram) as a candidate profile and the respective username set as a candidate username set U_C. If candidate usernames set is not accessible, current username of the candidate profile is used as u_c. Post processing, we collect 18,959 $U_S - U_C$ username set pairs and 109,292 $U_S - u_c$ pairs, totaling 128,251 instances whose username sets are known to belong to a single user and hence are positive instances (see Table III). We create an equal number of negative instances, by randomly pairing a username set of a positive instance with a username (set) of a different positive instance, which are known to belong to different users. We extract features from positive and

[6] Tumblr API does not share a unique *user_id* of a user to keep track of changes to her Tumblr profile, hence development of an automated tracking system is challenging.

Table III. Datasets capture username changes of 128,251 users within two months on source and candidate networks.

	Tumblr	Facebook	Instagram	Total
U_S - U_C	14,301	1,166	3,492	18,959
U_S - u_c	58,285	31,076	19,931	109,292

negative instances and use features in an engineered framework that effectively classifies username sets as same or different users.

6 Evaluation

We evaluate listed frameworks on two genre of instances: $U_S - U_C$ instances (18,959 positive; 18,959 negative) and $U_S - u_c$ instances (109,292 positive; 109,292 negative) and on three metrics – *accuracy*, *false negative rate* (FNR) and *false positive rate* (FPR). Accuracy shows number of username sets correctly classified. False negative rate shows number of username sets falsely classified as unlinked while false positive rate shows the number of username sets falsely classified as linked.

Table IV details 10-fold cross validated accuracy, FNR and FPR of the baselines and the three frameworks. Classifying $U_S - U_C$ instances with only **b1** results in false negative rate of 89.34% and an accuracy of 55.38%. The high false negative rate alerts that most users have non-matching current usernames across their OSN profiles. When instances are (re-)classified using suggested features by either of the frameworks, we observe a drop in false negative rate by 35% or more, thus boosting the profile linking accuracy. A significant reduction denotes the importance of username history in linking user profiles, when current usernames do not match. To further boost the FNR and the accuracy, we evaluate and compare the performance the three frameworks. With Naive Bayes as a basic classifier, we observe that cascaded framework gives a slight better accuracy and false negative rate than independent and fusion framework while maintaining a low false positive rate.

Performance of Cascaded Framework: We now experiment with different baselines used as **Classifier I** and different supervised machine learning algorithms as **Classifier II** in the cascaded framework. **Classifier II** learned using Naive Bayes technique exploits username set features of **b1** negative predictions and reclassifies them. Reclassification reduces false negative rate to 48.87% thereby boosting accuracy to 73.12% leading to a significant reduction in false negative rate by 40%. We experiment with other supervised methods to learn the classifier, and achieve best accuracy with SVM (reduction by 48.47%) while maintaining a low FPR. With baseline **b2** as **Classifier I**, the framework achieves best accuracy of 76.93% and reduction in false negative rate by 37.59% with SVM classifier learned on username set features as **Classifier II**. ROC curves in Figure 6 shows that in order to gain higher TPR with **Classifier II**,

which directly contributes to the reduction in FNR of the framework, we need to compromise on FPR of the framework.

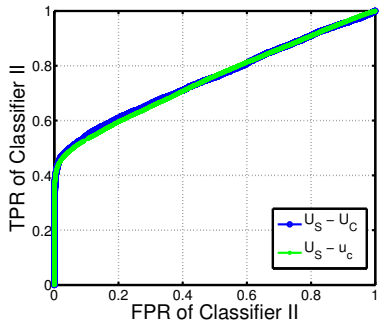


Fig. 6. ROC curve when SVM with RBF kernel is used as **Classifier II**. 40% TPR at low FPR implies that FNR of the framework reduces by 40%.

Significant reductions in FNR of the framework imply that the username history helps in linking user profiles and is an important feature for profile linking methods. An example where baselines fail to link with current usernames but cascaded framework compares the username sets and finds the link is two chronologically ordered sets – { U_S: [*'eenjolrass', 'isabelnevills', 'giuliettacapuleti', 'tobsregbo'*], U_C: [*'enjoolras', 'isabelnevilles'*] }. We see that current usernames do not match, however two of the past usernames are similar.

Classification of $U_S - u_c$ instances shows similar trends. On comparing classification accuracies of $U_S - U_C$ and $U_S - u_c$ instances, we observe that without access to candidate's past usernames, framework achieves a little less but similar accuracies. Lower linking accuracies for $U_S - u_c$ can be attributed to a slight increase in FPR. We, therefore, investigate if history availability on both profiles is beneficial for profile linking. Using $U_S - U_C$ instances, we create another dataset where we access only the current username of the candidate profile. With **b1** and SVM classifier (linear), we achieve an accuracy of 70.43% (FNR: 45.25%, FPR: 13.77%). Observe that due to increased FPR, profile linking accuracy fall from 76.97%, when username history on both profiles is available, to 70.43%, when username history is available only on source profile. Therefore, a comparison of a single username with a set may lead to higher FPR than a comparison of two username sets.

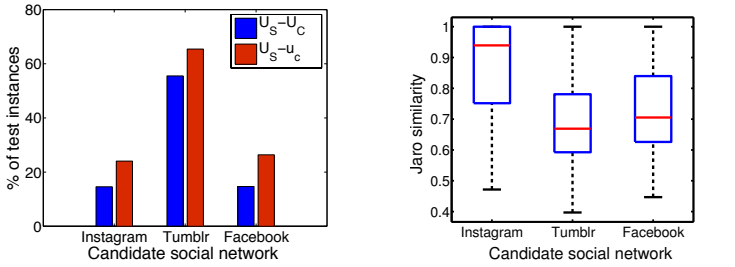
Impact of choice of OSNs: Though cascaded framework significantly reduces false negative rates, we are curious why false negative rates are still high (∼40%). To answer the question, we plot a distribution of false negative instances among the three candidate social networks (see Figure 7(a)). We find that an enormous 55.52% Twitter-Tumblr username set comparisons are misclassified (69% for Twitter set U_C-Tumblr username u_c). A high false negative rate on Tumblr

Table IV. Accuracy, FNR and FPR of supervised frameworks, baselines and their integration with another classifier learned using proposed feature set extracted for users tracked for two months and different supervised classification techniques.

	$U_S - U_C$			$U_S - u_c$		
Framework Config.	**Acc.**	**FNR**	**FPR**	**Acc.**	**FNR**	**FPR**
Exact Match (b1)	55.38%	89.34%	0.00%	52.79%	90.10%	0.00%
Substring Match (b2)	60.99%	78.46%	0.00%	56.44%	83.03%	0.00%
Independent [Naive Bayes]	72.10%	53.81%	1.91%	74.31%	47.38%	1.78%
Fusion [Naive Bayes]	72.93%	51.89%	0.19%	73.72%	49.19%	1.04%
Cascaded [b1→Naive Bayes]	73.12%	48.87%	3.07%	74.66%	45.97%	2.61%
b1 → Naive Bayes	73.12%	48.87%	3.07%	74.66%	45.97%	2.61%
b1 → SVM [Linear]	**76.97%**	**40.87%**	3.71%	75.60%	43.79%	3.03%
b1 → SVM [RBF]	76.57%	42.12%	3.21%	75.55%	44.72%	2.12%
b1 → Decision Tree	70.56%	27.19%	31.85%	68.46%	29.76%	33.48%
b1 → Random Forest	76.14%	34.71%	12.11%	74.25%	37.90%	12.36%
b2 → Naive Bayes	73.27%	48.52%	3.14%	74.81%	45.43%	2.90%
b2 → SVM [Linear]	**76.93%**	**40.87%**	3.78%	77.21%	39.42%	2.41%
b2 → SVM [RBF]	76.57%	42.12%	3.20%	75.33%	41.85%	3.55%
b2 → Decision Tree	71.18%	27.07%	30.70%	68.34%	29.60%	33.92%
b2 → Random Forest	75.21%	36.55%	12.05%	74.11%	38.15%	12.39%
Fusion [Weighted SVM-Linear]	76.05%	43.06%	3.27%	74.66%	45.97%	2.61%
b1 w/o Tumblr	60.49%	78.10%	0.00%	53.48%	87.10%	0.00%
(b1 → SVM [Linear]) w/o Tumblr	**92.56%**	**14.38%**	0.33%	86.10%	23.82%	2.50%
b2 w/o Tumblr	67.27%	64.70%	0.00%	59.53%	75.64%	0.00%
(b2 → SVM [Linear]) w/o Tumblr	92.56%	14.38%	0.33%	86.10%	23.28%	2.51%

can be attributed to the lowest Jaro similarity between most similar usernames from Tumblr and Twitter username sets (see Figure 7(b)). For instance, a user's usernames on Twitter – [*'articulatedan'*, *'radicaliguori'*, *'satanichowell'*] do not hold any similarity with her usernames on Tumblr – [*'ptvkitty'*, *'piercethecait'*, *'ptvcait'*]. Best Jaro similarity score for the username sets is 0.56. For instances like this, we need support of other attributes like name, location to find link between the two profiles. We then evaluate cascaded framework only on instances with candidate profile on either Facebook or Instagram. We achieve an accuracy of 92.56% on 4,658 $U_S - U_C$ instances (FNR: 14.38%, FPR: 0.33%) and 86.10% on 51,007 $U_S - u_c$ instances (FNR: 23.82%, FPR: 2.50%). On removal of candidate network Tumblr, a significant improvement in the accuracy shows that proposed cascaded framework accurately can find links between two user profiles given the username sets resemble and are created with similar behavioral characteristics.

Feature importance We now detail features that help the most during classification of usernames sets. We examine feature weights to estimate their importances for the most accurate framework configuration – Exact matching (**b1**) followed by temporal matching using SVM and compute them by squaring coefficients of features returned by **Classifier II** as suggested in [25]. Top-10 features, calculated between source and candidate username sets, are –



(a) $U_S - U_C$ instances (b) Jaro similarity distribution

Fig. 7. False negatives distribution among three candidate networks; Tumblr results in most false negatives. On further analysis, we observe that among the three candidate networks, Tumblr usernames have least Jaro similarity with corresponding Twitter usernames.

- Maximum normalized LCSub similarity.
- Second best normalized LCS similarity.
- Minimum normalized edit distance.
- Maximum normalized jaro similarity.
- Median of LCS similarity between source and candidate username pairs.
- Standard deviation of LCS similarity between source and candidate username pairs.
- Mean Jaccard similarity between alphabet distribution of source and candidate username pairs.
- Second best normalized edit distance.
- Maximum normalized LCS similarity.
- Second best normalized LCSub similarity.

Note that, top-10 features capture username creation behavior of a user. Username creation behavior play an important role for classification, but username evolutionary features and reuse behavior have relatively weaker roles. We analyze if evolutionary and frequent reuse patterns can contribute better given a longer history to find connections between the user profiles in Section 8.

To summarize, the key observations are – i) Cascaded framework performs better than an independent framework, ii) A comparison of username history reduces false predictions by 48% which are caused by the only comparison of current usernames, iii) Cascaded framework identifies links between 8% more instances with availability of past usernames on both profiles than on only one, iv) Success of the framework relies on the platforms to which examined profiles belong to; 55.52% misclassified Twitter-Tumblr username sets while approx. 14% Twitter-Instagram and Twitter-Facebook for U_S - U_C instances. Our experiments on fairly large datasets give a detailed proof of concept on the importance of using attribute history for profile linking. However, as observed, profile linking accuracy varies with the choice of OSNs to which profiles belong to, we next investigate on why and how username creation processes differs for OSNs like Tumblr but not for Twitter, Instagram or Facebook.

7 Related Work

Profile linking is a well studied problem in literature. Existing literature addresses the problem of connecting user profiles across social networks by comparing current values of the attributes of user profiles. The suggested methods explore combination of attributes to compare and techniques to measure the similarity. Profile attributes such as username and name are compared using string similarity measures [4, 6, 18, 19], content attributes such as posts and message length are compared using language models [4, 6, 26], and network attributes such as number of friends and nature of ties with friends are compared using graph algorithms [6, 27, 28]. Few studies suggest crowd-sourced mechanisms to match user profiles across OSNs [29]. Others have examined the effectiveness of using only usernames to connect user profiles across OSNs [18, 19]. The state-of-the-art method MOBIUS compares a candidate username with a set of usernames owned by a user profile on other OSNs. MOBIUS assumes that user's unique behavior often leads to redundancies / similarities among the usernames across OSNs, which can be captured into features. Supervised classification techniques then predict if a candidate username and usernames on other OSNs are linked [19].

Most profile linking methods compare user profiles based on current values of the attributes observed at the time of executing the method. Existing methods are successful when users do not evolve their attribute values over time. However, recent studies show that users frequently change their attributes to suit their changing preferences on different OSNs [10, 30], similar to our observations in the study. In these scenarios, current values of the attributes on multiple profiles of a user may not match, thereby leading existing methods to falsely infer that user profiles as different users.

To address the limitations of existing methods and complement MOBIUS, we suggest considering attribute history to find links between user profiles. We propose to compare a candidate username with a set of past usernames of multiple user profiles across networks, not just with current values. We re-implement MOBIUS and build a framework with **Classifier I** extracting top-10 features by comparing candidate username with a set of current usernames on other OSNs, as proposed by the authors and **Classifier II** extracting username set features by comparing candidate username history with other profiles' username histories as proposed in this work. On a dataset of 8,997 users who have profiles on more than two social networks as well past history on all the social networks, 42.67% instances are false negatives i.e. **Classifier I** miss the link among profiles. **Classifier II** identifies links among 30.72% more instances, reducing false negatives to 11.95%. Therefore, we see that attribute history complements state-of-the-art method and extends support to existing profile linking methods.

8 Discussion

On a dataset of real-world users, we show that username history holds its significance by extending performance to existing methods for profile linking. However,

its effectiveness varies with the choice of OSNs. We observe that majority users create different usernames on Tumblr as compared to their profiles on Twitter, Facebook or Instagram. Differences between the username sets hints disparate user needs and choices across OSNs. We think that profile linking strategies need to tune according to the nature and genre of the OSN with a prior knowledge of popular user behavior on that OSN. Now, we discuss applicability of attribute history along with other dependencies of the framework that uses attribute history for linking.

8.1 Applicability

Apart from observing users over time on OSNs, one can get user history archived by external services like DataSift[7] or Gnip[8]. We further suggest other two methods to collect past usernames – via timeline and via public datasets.

Via timeline On social networks like Twitter and Instagram, users converse by tagging another user's username with '@' tag. When a user changes her username, old tweets and replies where others tagged her with her old username stay on her timeline. By listing old posts with replies and extracting mentions from the tweets, one may list her past usernames. We believe that a recent history of past usernames can be captured by this method.

Via public datasets Multiple researchers collect private and public posts related to a topic, event or a campaign ranging over a period of time. They often store information about authors who created these posts. One may query these databases with the *user_id* of a user and find posts created by her at different times. If the author details are recorded with each post, one may list unique usernames used by the user in the past. With this methodology, we find past usernames of 4% of 128,251 Twitter users, via datasets shared by an event monitoring tool, MultiOSN [31].

With these methods, applicability of the proposed profile linking framework can be extended to random users who are not tracked continuously over time.

8.2 Dependency

We test the proposed framework for dependency on the grounds of understanding how much history is required for efficient profile linking. In other words, does a longer history on source username set impact framework accuracy? To answer the question, we create a dataset of $U_S - u_c$ username sets with 502 users from the dataset of 109,292 users who had changed their Twitter username maximum number of times (5 times) within tracking period of two months. We further partitioned 502 $U_S - u_c$ sets into 4 datasets $(d_i)_{i=2}^{5}$, where dataset d_i contains

[7] http://datasift.com/platform/historics/
[8] https://gnip.com/products/historical/

instances with first i past usernames from their respective U_C sets. For instance, d_2 contains 502 $U_S - u_c$ instances, where each U_C contains only first two usernames of the five usernames in the username set. FNR by cascaded framework with respect to the baseline **b1** on the derived datasets with varying set sizes is shown in Figure 8.

Observe that as past username set size increases, difference between FNR of the framework and FNR of the baseline increases, thereby indicating that longer the username history of a Twitter user, better the matching with a candidate username or set.

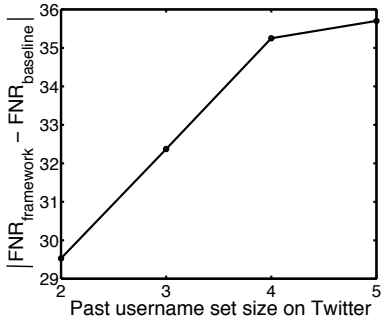


Fig. 8. Higher FNR reduction with increasing source username set size.

8.3 Importance of evolutionary creation and temporal reuse behavior

List of important features suggests that username creation behavior helps better than other behavioral patterns to suggest if username sets refer to a single user. We suspect that a user's evolutionary behavior or her tendency to reuse usernames across social networks over time are of little help to the classification process due to fewer instances with these features. We, therefore, repeat feature importance analysis for another dataset with longer username history. We randomly sample a set of 10,000 users from 128,251 users on Twitter and record their attributes every fifteen minutes for 12 months (November 26, 2013 - November 28, 2014). Out of 10,000 users, 47% users change their username at least once during the tracking period. To create ground truth dataset, we filter users who self-identify themselves on at least one of the candidate social networks – Instagram, Tumblr or Facebook. For 682 users, we retrieve their current username on either of the candidate networks while for 155 users we retrieve their past usernames on both Twitter and one of the candidate networks. SVM classifier with linear kernel, used as **Classifier II**, ranked 'username reuse' features above 'username creation' features deemed important earlier – ratio of common usernames to candidate set size, and number of common usernames found between source and candidate username sets are ranked above than mean Jaccard similarity between two sets. Therefore, we gather that relative importance of behavioral patterns to reveal a potential link between two user profiles

varies with the longevity of the attribute history on either profiles. Username reuse behavior can be only observed over a prolonged track of history, however has proven useful feature for profile linking.

8.4 Implications to Privacy

We understand that tracking a user to gather history of attribute values may issue privacy threats. A track of location can reveal mobility patterns of the user while description can reveal her changing likes, favorites, or professions, revealing instability in the user's financials. History of other attributes like messages, interactions and friends can further affirm patterns that intrude a user's privacy. For our research, we track only descriptive characteristics like profile attributes. We have not asked for user authorization or Twitter authorization (except the OAuth query tokens) to request collection of any of their public data and characteristics. However, we agree and believe that a periodic tracking of a user's data may have serious implications. The data can be easily gathered by anyone without proper permission channels, including spammers and attackers. In our system that uses the framework, we approve of proper credentials in order to allow anyone to track any user for a period of time. For future, we recommend that such a frequent and periodic data collection methods should include prior notification to the social network and the user herself. If she insists not to be part of the data-collection, she can be removed. Another suggestion is to anonymise the data gathered such that public sharing of insights and results can not be linked to individual users. We have tried to make sure that none of the users mentioned in the study are uniquely identifiable.

8.5 Extension to other attributes

Not just username, but other attributes evolve over time on Twitter. Evolution leads to distinct set of values ever assigned to an attribute which can be used for comparison during profile linking. Figure 1(a) shows a distribution of 5.5 million out of 8.7 million Twitter users who had changed one of their profile attributes during our observation period of two months. Observe that, apart from username, majority users change their description and profile picture, thereby creating a distinct set of values to be compared with other candidate profiles. One is likely to reuse a picture or describe her in a similar fashion on different networks. A new set of features capturing distinct similarities between these attributes can be devised in future to help profile linking.

9 Conclusion and Future Work

In this work, we emphasize that attribute values evolve over time on one or multiple profiles of a user. This causes non-matching current attribute values on the profiles. Existing methods do not consider attribute evolution and falsely predict non-matching user profiles as different users. To avoid such mistakes, we

propose to compare attribute's past values, not just current values, to revise the prediction. Focusing on username, we find similarities between username sets, comprised of past and current username of a user profile. We assume that user behavior of username creation and its reuse across her profiles is unique; these user behavioral patterns can remain static or vary over time and are captured into a set of features. Novel cascaded framework of classifiers uses proposed features to rectify erroneous classifications based only on current values. We show that our framework outperforms existing profile linking methods by reducing false negative errors by 48%. In conclusion, comparing username history along with current values help in effective profile linking.

In future, we plan to extend our work on two fronts. First, further reduce the false negative rate with the help of other features such as name, location, description and profile picture. We believe that inclusion of distinguishing features like locations, which are static in nature compared to attributes like username, can help strengthen / revise the binary decision made by the cascaded classifiers. Second, evaluate the proposed framework on a random set of users rather than users who self-identify themselves. We believe that an accumulation and knowledge of a user's past profiles can also help in other research domains other than profile linking. Our work can be extended to profile search frameworks, which is an important part of profile resolution process [24], however is not the focus of this study.

10 Acknowledgment

We would like to thank members of Precog, a research group at IIIT-Delhi, and members of Cybersecurity Education and Research Centre (CERC), IIIT-Delhi for their constant feedback and support. The research presented is funded by TCS Research Labs, India and the first author is the awardee of TCS research fellowship.

References

1. Jain, P., Kumaraguru, P., Joshi, A.: Other times, other values: Leveraging attribute history to link user profiles across online social networks. In: Hypertext (HT). (2015)
2. Weinstein, A.: Handbook of Market Segmentation: Strategic Targeting for Business and Technology firms. Haworth Press (2004)
3. Cockerell, G.: Making Marketing Meaningful. Kendall Hunt Publishing Company (2010)
4. Iofciu, T., Fankhauser, P., Abel, F., Bischoff, K.: Identifying Users Across Social Tagging Systems. In: ICWSM. (2011)
5. Irani, D., Webb, S., Li, K., Pu, C.: Large Online Social Footprints–An Emerging Threat. In: CSE. (2009)
6. Liu, S., Wang, S., Zhu, F., Zhang, J., Krishnan, R.: HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling. In: SIGMOD. (2014)

7. Motoyama, M., Varghese, G.: I seek You: Searching and Matching Individuals in Social Networks. In: WIDM. (2009)
8. Szomszor, M., Cantador, I., Superior, E.P., Alani, H.: Correlating User Profiles from Multiple Folksonomies. In: Hypertext (HT). (2008)
9. Li, P., Dong, X.L., Maurino, A., Srivastava, D.: Linking Temporal Records. VLDB (2011)
10. Liu, Y., Kliman-Silver, C., Mislove, A.: The Tweets They are a-Changin': Evolution of Twitter Users and Behavior. In: ICWSM. (2014)
11. Chen, Y., Zhuang, C., Cao, Q., Hui, P.: Understanding cross-site linking in online social networks. In: WOSN. (2014)
12. Chen, T., Kaafar, M.A., Friedman, A., Boreli, R.: Is More always Merrier?: A Deep Dive into Online Social Footprints. In: WOSN. (2012)
13. Zafarani, R., Liu, H.: Connecting Corresponding Identities across Communities. In: ICWSM. (2009)
14. Liu, J., Zhang, F., Song, X., Song, Y.I., Lin, C.Y., Hon, H.W.: What's in a Name?: An Unsupervised Approach to Link Users Across Communities. In: WSDM '13. (2013)
15. Feizy, R., Wakeman, I., Chalmers, D.: Transformation of Online Representation through Time. In: ASONAM. (2009)
16. Jain, P., Kumaraguru, P.: @I to@ Me: An Anatomy of Username Changing Behavior on Twitter. arXiv preprint arXiv:1405.6539 (2014)
17. Malhotra, A., Totti, L., Meira, W., Kumaraguru, P., Almeida, V.: Studying User Footprints in Different Online Social Networks. In: ASONAM. (2012)
18. Perito, D., Castelluccia, C., Kâafar, M.A., Manils, P.: How Unique and Traceable Are Usernames? In: PETS. (2011)
19. Zafarani, R., Liu, H.: Connecting Users across Social Media Sites: A Behavioral-modeling Approach. In: KDD. (2013)
20. Shi, X., Nallapati, R., Leskovec, J., McFarland, D., Jurafsky, D.: Who Leads Whom: Topical Lead-lag Analysis across Corpora. In: NIPS Workshop. (2010)
21. Grant, T.: TXT 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages. In: JL & Pol'y. (2012)
22. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of molecular biology (1981)
23. Heitz, G., Gould, S., Saxena, A., Koller, D.: Cascaded Classification Models: Combining Models for Holistic Scene Understanding. In: Advances in Neural Information Processing Systems. (2009)
24. Jain, P., Kumaraguru, P., Joshi, A.: @ I seek 'fb. me': Identifying Users across Multiple Online Social Networks. In: WWW Companion. (2013)
25. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine learning (2002)
26. Goga, O., Lei, H., Parthasarathi, S.H.K., Friedland, G., Sommer, R., Teixeira, R.: Exploiting Innocuous Activity for Correlating Users across Sites. In: WWW. (2013)
27. Bartunov, S., Korshunov, A., Park, S., Ryu, W., Lee, H.: Joint Link-Attribute User Identity Resolution in Online Social Networks. In: SNAKDD. (2012)
28. Narayanan, A., Shmatikov, V.: De-anonymizing Social Networks. In: SP. (2009)
29. Shehab, M., Ko, M.N., Touati, H.: Social Networks Profile Mapping using Games. In: USENIX. (2012)
30. Zhang, J., Wang, C., Wang, J.: Learning Temporal Dynamics of Behavior Propagation in Social Networks. In: ICWSM. (2014)

31. Dewan, P., Gupta, M., Goyal, K., Kumaraguru, P.: Multiosn: Realtime monitoring of real world events on multiple online social media. In: I-CARE. (2013)