Automated Methods for Identity Resolution across Online Social Networks

By

Paridhi Jain

Under the supervision of Dr. Ponnurangam Kumaraguru

Indraprastha Institute of Information Technology Delhi

April, 2016

Automated Methods for Identity Resolution across Online Social Networks

By
Paridhi Jain

Submitted

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

to the

Indraprastha Institute of Information Technology Delhi
April, 2016

Certificate

This is to certify that the thesis titled "**Automated Methods for Identity Resolution across Online Social Networks**" being submitted by **Paridhi Jain** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Supervisor Name: Dr. Ponnurangam Kumaraguru, 'PK'
April, 2016
Department of Computer Science
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Keywords: Online social networks, identity, identity resolution, profile linking, identity search, historic profiling

Abstract

Today, more than two hundred Online Social Networks (OSNs) exist where each OSN extends to offer distinct services to its users such as eased access to news or better business opportunities. To enjoy each distinct service, a user innocuously registers herself on multiple OSNs. For each OSN, she defines her identity with a different set of attributes, genre of content and friends to suit the purpose of using that OSN. Thus, the quality, quantity and veracity of the identity varies with the OSN. This results in dissimilar identities of the same user, scattered across Internet, with no explicit links directing to one another. These disparate unlinked identities worry various stakeholders. For instance, security practitioners find it difficult to verify attributes across unlinked identities; enterprises fail to create a holistic overview of their customers.

Research that finds and links disconnected identities of a user across OSNs is termed as *identity resolution*. Accessibility to unique and private attributes of a user like 'email' makes the task trivial, however in absence of such attributes, identity resolution is challenging. In this dissertation, we make an effort to leverage intelligent cues and patterns extracted from partially overlapping list of *public* attributes of compared identities. These patterns emerge due to consistent user behavior like sharing same mobile number, content or profile picture across OSNs. Translating these patterns into features, we devise novel heuristic, unsupervised and supervised frameworks to *search* and *link* user identities across social networks. Proposed search methods use an exhaustive set of public attributes looking for consistent behavior patterns and fetch correct identity of the searched user in the candidate set for an additional 13% users. An improvement on the proposed search mechanisms further optimizes time and space complexity. Suggested linking method compares past attribute value sets and correctly connect identities of an additional 48% users, earlier missed by literature methods that compare only current values. Evaluations on popular OSNs like Twitter, Instagram and Facebook prove significance and generalizability of the linking method.

Proposed search and linking methods are applicable to users that exhibit *evolutionary* and *consistent* behavior on OSNs. To understand the dynamics and reasons for such behavior, we conduct two independent in-depth studies. For user evolutionary behavior, specifically for username, we observe that username evolution leads to broken link (404 page) to a user profile. Yet, 10% of 8.7 million tracked Twitter users changed their username in two months. Investigation reveals that reasons to change include malign intentions like fraudulent username promotion and benign ones like express support to events. We believe that Twitter can monitor frequent username changes, derive malign intentions and suspend accounts if needed. Study of sharing information consistently across OSNs, e.g. mobile number, highlights why users share a personally identifiable information online and how can it be used with auxiliary information sources to derive details of a user.

In summary, this dissertation encashes previously unused public user information available on a social network for identity resolution via novel methods. The thesis work makes following advancements: a) Propose search frameworks that aim to fetch correct identity of a user in the candidate set by searching with public and discriminative attributes, b) Propose a supervised classification framework for linking identities that compares respective attribute histories in situations where state-of-the-art methods fail to predict the link, c) Study username evolution on Twitter, and d) Study mobile number sharing behavior across OSNs. Proposed methods require no user authorization for data access, yet successfully leverage innocuous user public activity and details, find her accounts across OSNs and help stakeholders with better insights on user's likings or her suspicious intentions.

Dedicated to my parents

Acknowledgements

These wonderful six years with my advisor, Prof. Ponnurangam Kumaraguru, have been the most cherished and exciting years of my life. He has not only inspired, guided and supported but also introduced me to multiple opportunities of collaboration in the community. I am grateful to him for providing a cultivating environment that supports liberty, open discussions, and critical feedbacks, without which this work would not have been possible. I am indebted to the long 'wisdom talks', which will benefit me throughout my future. I am extremely fortunate to have him as my advisor.

I would like to express my sincere gratitude to Prof. Anupam Joshi and Prof. Rahul Purandare. I thank you for trusting and believing in me. You inspire me to set the research standards high and not to compromise with them. You have cheered for me during my ups and cheered me up during my downs. I am extremely honored to have worked and interacted with you, on both professional and personal fronts.

I am thankful to my thesis monitoring committee members, Prof. Vikram Goyal, Dr. Gaurav Gupta, and Dr. Mohan Dhawan, for giving me honest and timely reviews on my work that helped me correct my mistakes at the right time. I would like to thank Prof. Pankaj Jalote and Prof. Dheeraj Sanghi for being my patrons and inspirited me to pursue Ph.D. at the first place.

I have been lucky to have worked with Prachi, Rohan and Swati. I thank you for your efforts and contribution to this work. I am also thankful to Prateek, Anshu, and Luam for sharing the datasets. This thesis would not have shaped as it is without the long healthy discussions with my mates at Precog@IIIT-Delhi and my professional siblings, Aditi, Niharika, Prateek, Anupama, and Srishti. Each one of them has motivated and supported me in their own ways. I learned formulating and pitching research problem from Aditi, Niharika has always been generous and a critical reviewer of my work, Prateek finds his way of humor to de-stress, Anupama has been a 'go-to' place for any technical advise, and Srishti has always energized the environment with her hard work and 'never-give-up' attitude. I thank each one of you.

I am grateful to my partners-in-journey, Tejas, Anush, Shruti, Siddhartha, Venkatesh, Himanshu, Samarth, Kuldeep. From time to time, they have made me laugh, helped me sustain through hard times and intervened on my mistakes. I am extremely fortunate to have found a friend in them.

Finally, I would like to thank my supportive and affectionate family for its support. I am grateful to my parents for a wonderful education, my brother, Rahul, and my parents-in-law for supporting my education and my stay-out-of-home duration for about two long years. I want to thank my 'partner-in-crime', Aniket, for his support and care. During many months when we stayed in different cities and in different timezones, I have always felt him near me. Aniket is the most understanding person I have ever met and I am blessed to have married to him for the rest of our times in health and sickness, failure and success.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Online Social Networks (OSNs) are easily accessible and usable digital platforms that expand opportunities to meet new people, build communities, discuss topics and diffuse information. Today, over 2.9 billion individuals across the globe use Internet and about 42% of these users actively participate on OSNs [12]. Recent statistics list about 209 active OSNs in 2015 [107]. Few OSNs are more popular than others. Facebook witnesses 1.5B monthly active users [37], Instagram has 400M monthly active users [50] and Twitter sees 316M users sending 500M tweets per day [100]. Each OSN offers distinct set of innovative services that ease access to information. For instance, Twitter's retweet feature enables quick access to news, campaigns, and crisis information, while pin boards of Pinterest facilitate reach to the work of artists, photographers, and fashion designers.

In order to enjoy these services simultaneously, a user innocuously registers herself on multiple OSNs [14, 15, 69, 108]. Recent study by Pew Research Center shows that 91% users registered themselves on both Twitter and Facebook; 52% users on Twitter and Instagram [15]. The overlap of users across OSNs has increased from the similar study conducted in 2013 by the Pew Research Center [14] and is expected to increase in future years, as OSNs introduce new features with time to attract users [97]. During registration on any OSN, a user creates an identity for herself listing personal information and connections. Due to varying policy and purpose of the identity creation on each OSN, quality, quantity, 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 (see Figure 1.1). These disparate identities liberate her from any privacy concerns that could emerge if the identities were implicitly collated. However, disparate unlinked identities is a concern for various stakeholders.

Enterprises like multinationals and news companies, non-profit organisations, and political parties spend resources to seek user sentiment towards their organisation, events or products via social media, hence create accounts on multiple OSNs. They ask users to 'like' or 'follow' their accounts and

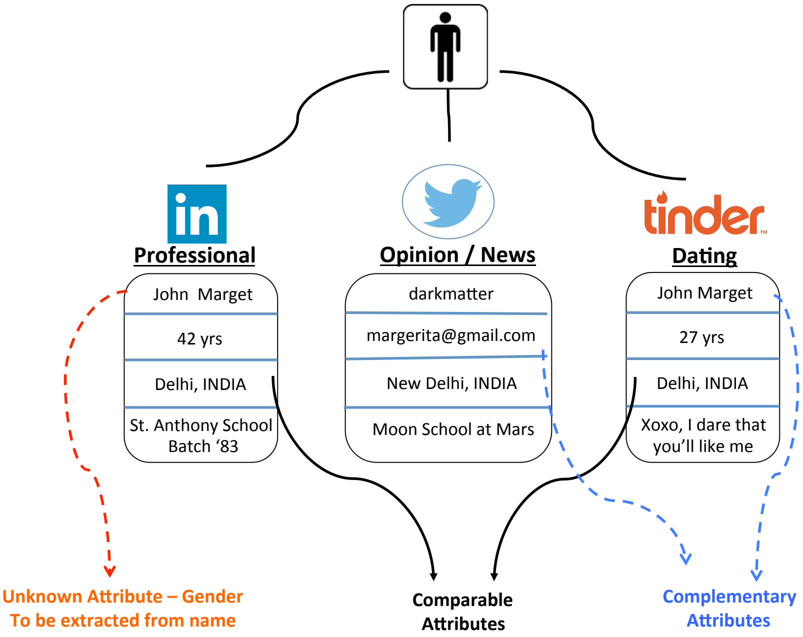


Figure 1.1: A user named John Marget varies his details depending on the purpose of the OSN.

request users to share their feedback on these accounts. Constant efforts create a social audience i.e. a section of online population targeted by product campaigns via social media. To calculate their audience reach, enterprises count number of users liking or talking about their products. With enterprise accounts on multiple OSNs, it is difficult to estimate correct social audience. This is because a single user can participate in the same activity via her multiple OSN accounts. For example, a user with an account on Facebook and Twitter can like 'Disney' official account on Facebook as well as 'follow' Disney official account on Twitter. Therefore, an arithmetic sum of 'followers' or 'viewers' from organisation's each OSN account can inflate the real correct audience size [41, 95]. It is necessary to deduplicate users by linking their multiple OSN identities before counting. Further, enterprises carry out psychographic segmentation based upon customers' activities, interests, opinions and lifestyles to adapt marketing strategies to their needs [26]. It is the most effective segmentation citing a rise of 24% in business performance [105]; however includes high cost in both time and money [24]. The cost of constructing psychographics of each customer can be brought down if a user's personality can be inferred using aggregated data from her linked social identities on OSNs [40].

Security practitioners often need to verify individual's characteristics to discover her real identity and mark untrustworthy attributes. Recently, a police report says that 'Skout', a mobile social networking app, found that three adults masqueraded as 13 to 17-year olds and contacted kids and sexually harassed them [84]. In such scenarios, security practitioners[1] need to verify portrayed identity of a user. Within the limits of a social network, the task is non-trivial, which raises concerns in the community. However, attribute verification is plausible with identity resolution. One can draw links between a user's multiple identities and aggregate information to effectively highlight attribute value discrepancies, thereby contributing to identity verification. Figure 1.1 shows an example where age can be matched to find any discrepancy and if any, possible malign intentions to fake age.

E-commerce sites pursue to provide personalized recommendations to their customers. However, most suggestions are made based on prior purchases of the user. Often, it is required to know the likes, dislikes and plausible interests of the user, predominately available via her activity on social networks. For instance, Foursquare check-ins, reviews and tips suggest the cuisine a user likes, while Instagram pictures uploaded by the user can suggest places she likes to travel. Within a social platform boundary, the knowledge about the user is rather incomplete. Explicit links connecting unlinked user identities can aggregate complementary user information from different platforms and create a comprehensive user profile for effective personalization and targeting.

The task of finding and linking disconnected identities of a user across multiple social platforms is termed as *identity resolution* across online social networks. Challenges like dissonant social platforms with partially overlapping list of supported attributes, missing or veracious attribute information, and restricted API information sharing impede effective identity resolution. We discuss each of these challenges below:

- **Heterogeneous OSNs:** OSNs ask users to define their profile attributes during registration in order to uniquely identify them as well as help others on the network to connect to them. The quantity and granularity of the information asked varies with each OSN, with few OSNs demanding descriptive attributes, while few needing a valid email and chosen username. The reason for such heterogeneity is intuitive. For content based OSNs like Flickr, Instagram, Twitter and Pinterest, a user's interests constitute her identity on the network than her real personality, while for link based networks like LinkedIn, and Facebook, a user's real identity extends to others either to revive old or create new connections. List of supported user attributes, the genre of the content created, and the strength of the ties thus vary among OSNs. Such heterogeneity challenges identity resolution methods by making it difficult to spot overlapping list of attributes supported universally across OSNs [71].

- **Un-verified accounts:** Each OSN deploys a distinct identity verification mechanism to ensure the veracity of the account and the details entered. For instance, Facebook enforces the policy of using 'real' name on the network followed by mobile number verification to ensure

[1] Security practitioners can be a police official, security team of the OSN or a qualified individual.

registration request of a human [36], while Twitter allows users to register under pseudonyms followed by an email verification [99]. No OSN advocates the need of verifying specified attributes to the real attributes of the user. In such a scenario, a user can choose to hide or lie or maliciously copy others on few or all her characteristics described on the OSN. With little knowledge on the degree of veracity of available attributes, the confidence on similarity based matching for identity resolution can be challenged.

- **Missing Information:** For factors like recent developments in privacy awareness, users either restrict or skip to mention few of their attributes on OSNs, thus leading to incomplete identities carrying insufficient information for resolution methods to compare and link them across OSNs.

- **Attribute Evolution:** OSNs have recently observed a temporal variance in the user behavior. Users prefer to update their profile attributes across their accounts in an asynschronous manner [73]. Thus, two identities of a single user across OSNs can hold different values at the same time. State-of-the-art resolution methods restrictively compare snapshots of the user identities taken at the same time, thus are challenged with frequently changing attribute values. Further, most OSN APIs do not store past details of their users, further limiting introduction of better resolution methods that compare past versions of the examined user identities.

- **Limited access:** OSNs offer Application Program Interface (API) to query details of a user. Private details can be obtained via a user authorisation and permission to use her details for research purposes, while public details are available without her permission. Often, convincing a user to share her details is a challenging task, convincing APIs to share all public details is another limitation. Few OSN APIs like Twitter REST API shares most details of a user, while other APIs like Facebook or Instagram share few of the public details, which further disbalance the quantity and quality of data identity resolution methods can use from each OSN to derive similarity among unlinked identities.

1.1 Thesis Statement

In purview of these challenges, this thesis aims to develop novel methods for identity resolution across online social networks. This work makes following assumptions on the type of users searched for identity resolution – a) the user has registered an identity on multiple social networks, b) the user maintains a single identity on each social network, and c) the user behavior is redundant; she performs a set of similar activities across her identities. Limiting the scope of this thesis to attributes that are publicly accessible via the APIs, the thesis statement is:

A user's identities across online social networks can be searched and linked using past and present values of the identifiable and discriminative public attributes.

1.2 Thesis Contribution

This thesis propose novel methods to search correct identity of a user on different OSNs with the sole use of public attributes of the user and link potential candidate with the known identity based on history of attribute values. I believe that findings of this research can aid stakeholders like enterprises, political organisations and security practitioners to build scientific methods for identity resolution, disambiguation and verification with the lone use of public attributes. Methods based on public attributes are free from any form of user authorization, data collection restrictions, legal, and privacy issues that may occur.

Also, we conduct in-depth investigation of user behavior exploited by the proposed methods for identity resolution. Questions like *how and why users publicly share a mobile number* or *create a common attribute history across OSNs* are addressed in independent studies. Insights on user behavior enrich identity resolution methods further and can be leveraged for various other applications. We now describe the contributions in detail.

1.2.1 Methods for creating candidate set by exploiting public attributes

Given a user identity on a social network, we devise methods to find a set of similar (candidate) identities, possibly containing the user's correct identity, on other network. Approaches in literature are limited to using only profile attributes of an identity to create a set of candidate identities on the searched social network. Search by profile has inherent disadvantages; it is highly restrictive and dependent on the availability of same profile attributes across networks. For example, 'gender' profile attribute is available on Facebook, while no such attribute exists on Twitter. For users who use significantly different values for their profile attributes across social networks either purposely or unintentionally, search by profile will retrieve a candidate set without containing the correct identity. Missing out the correct user identity in the candidate set lowers the accuracy of identity resolution.

We observe that content and network attributes are discussed fewer times in literature, to be used as search parameters for generating candidates. Therefore, we introduce heuristic identity search methods based on content and network attributes of a user and improve on traditional search by profile method. Search based on content and network attributes are motivated by user behavior of cross-posting and connecting to same individuals across platforms. Given a user's identity on Twitter, we evaluate proposed identity search methods to find her identity on Facebook. We report that a combination of proposed identity search algorithms found Facebook identity for 40.5% of Twitter users searched, while traditional method based on profile attributes found Facebook identity for only 27.4%. Each proposed identity search algorithm access public attributes of a user on any social network. We conclude that inclusion of more than one identity search algorithm, each exploiting distinct personality attributes of an identity, helps in improving the accuracy of identity resolution. To address scalability, we also devise an unsupervised method for candidate set creation using canopy clustering on available and discriminative attributes. Evaluating on four different social networks – LinkedIn, Quora, Twitter and Facebook, we show the effectiveness of using a cost-effective prior filtering of candidates.

1.2.2 Method for effective identity linking by leveraging attribute history

Given two user identities from different OSNs (a candidate and a known identity), we build a novel framework to predict a link between the identities and infer their connection to a single user. Methods in literature link identities by observing high similarity between most recent (current) values of the attributes like name and username. However, for a section of users observed to evolve their attributes over time and choose dissimilar values across their identities, these current values have low similarity. Existing methods then falsely conclude that identities 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 of respective user identities across networks to predict a link between them. We believe that attribute history highlights user preferences for creating attribute values on a social network. Co-existence of these preferences across identities on different social networks result in alike attribute histories that suggests they potentially refer to a single user. Through a focused study on *username*, we quantify the importance of username history for identity linking on a dataset of real-world users with users on Twitter, Facebook, Instagram and Tumblr. We show that username history correctly linked 48% more identity pairs with non-matching current values that are incorrectly missed by existing methods.

1.2.3 Study of username changing behavior

Proposed method for identity linking is applicable to a set of users who change their attributes over time. Our observations during formulation of the identity linking method suggest that around 10% of Twitter users changed their attribute like username on Twitter within a tracking duration of two months. To understand how and why do these users undergo changes to their username, we characterize username changing behavior of carefully selected Twitter users and find that majority users changed username frequently after short time intervals (a month) and chose new username dissimilar to the old one. Few favored a username by repeatedly choosing it multiple times. We report few of the many reasons for username change; benign reasons like space gain, suit a trending event, gain / lose anonymity, adjust to real-life events, avoid boredom and malicious intentions like obscured username promotion and username squatting. We believe that this work will not only help identify and tag fraud users but also promote researchers to devise new linking algorithms that capture learned username creation patterns.

1.2.4 Study of mobile number sharing behavior

Mobile number is a unique and personal identifiable attribute of a user, yet users share their mobile number publicly on OSNs. This helps one of the proposed identity search methods to find a user's online identities, however little is known on why users exhibit such behavior. Through an in-depth study of 2,997 Indian mobile numbers shared on OSNs, we find that most users shared their own mobile numbers to spread urgent information and to market products, IT facilities and escort business. Users resorted to applications like Twitterfeed and TweetDeck to post and popularize mobile numbers on multiple OSNs. We also show that a mobile number can further reveal personal accurate information about a real-world user when augmented with auxilary information sources like Voter ID rolls. To the best of the knowledge, this is the first work to understand the mobile number sharing behavior of users and implied repercussions of the sharing.

1.3 Implications of the Contributions

To enterprises: As discussed in our motivation of the thesis, the contributions on the front of identity search and linking will help enterprises and marketers to de-duplicate audience reached through their campaigns on online social media and estimate the audience size correctly. These enterprises can further merge identified multiple profiles of their potential customers, build aggregated comprehensive profiles, collect previously unknown information about them and run psychographics analysis to segment them for targeted advertising. Security professionals will be able to verify the authenticity and credibility of common attributes among the identified multiple profiles of users and may further infer hidden characteristics of the users like location and age using existing methods in literature. Social platforms themselves benefit by uncovering hidden malicious intentions reflected in temporal user behavior on social media. In various ways in addition to these, the research contributions in this thesis help industry to make advancements towards better customer services.

To user: Our contributions can further enrich an over-the-top application that guides users to identify all possible leaks in their public information that aid identity resolution. Our studies on mobile number sharing behavior and username change behavior show that users innocuously share their private information on a social network and engage in exactly same behavior across their profiles. Timely interventions with an appropriate notification can guide user to patch or hide sensitive information like phone number and posts (if shared across multiple profiles). Such interventions can suggest users to avoid keeping same history of activity across profiles or behave in a distinct but ubiquitous manner. This essentially challenges the identity resolution methods we develop but helps secure the online privacy of an individual.

1.4 Organization of the Thesis

This document is organized as follows. Chapter 2 provides a background to identity resolution, motivates its need and discusses state of the art methods to resolve identities across social networks. Chapter 3 elaborates on novel methods to search for a user identity across platforms, while Chapter 4 presents automated framework to predict a link between user identities. Lastly, Chapter 5 and Chapter 6 discusses the characterization studies of user behavior, Chapter 7 concludes with the future extensions of the work.

Chapter 2

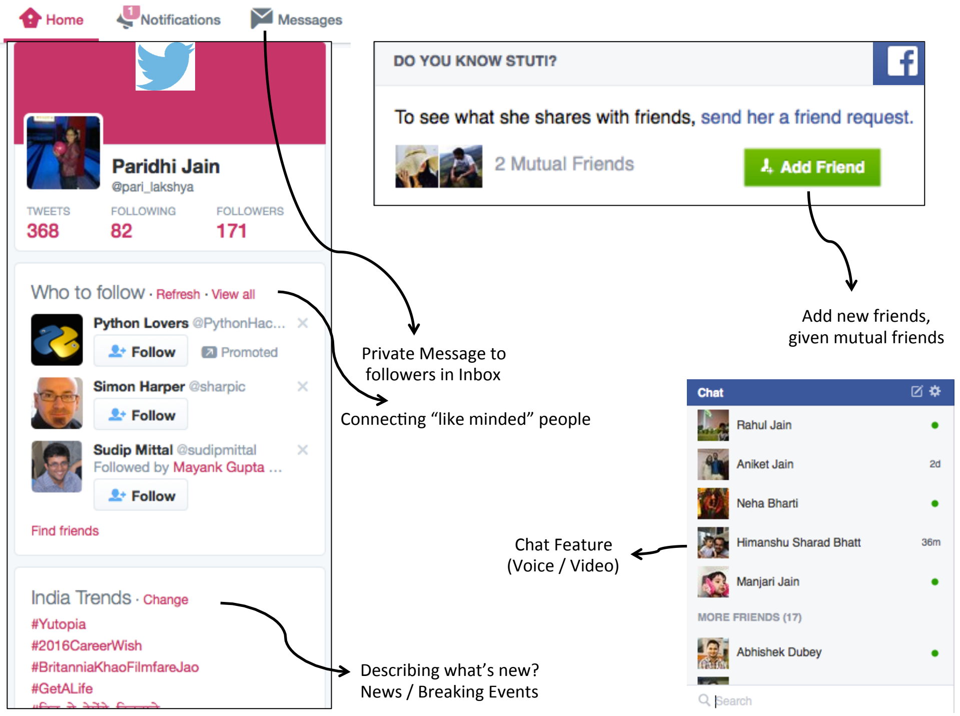
Background

2.1 Online Social Networks

Over the last decade, technology has thrived to provide better, quicker and effective platforms to help individuals connect and disseminate information to other individuals. Starting from Internet Messengers (IMs), websites, emails and blogs, Online Social Networks (OSNs) have emerged as a popular media. According to Boyd *et al.*, a social network is defined as "*a platform to build social relations among people who share similar interests, activities, backgrounds or real-life connections. Social networks are web-based services that allow individuals to create a public profile, create a list of users with whom to share connections, and view and cross the connections within the system.*" OSNs like MySpace and Friendster were restricted to only interactions with friends but today OSNs like Instagram, Pinterest, LinkedIn, Twitter, Facebook, Quora, Tumblr, further help to showcase talent, reach to businesses, discuss opinions and share knowledge.

Today, about 209 active OSNs, popular in different parts of the world, are targetting different sections of population. For instance, Weibo and Twitter are alike OSNs in terms of their services however their popularity is restricted within certain geographical regions. Elaborating on services provided by the OSN, each offers a unique feature to attract users to register on their platform. Location based OSNs like Foursquare and Yelp allow users to check in at the places they have been to, rate the place, leave a tip or even write a review. Relation based OSNs like Facebook and LinkedIn focus on engaging users in personal and professional connections with voice / video chat, help find a friend and connect instantly through email services within the platform. OSNs like Twitter is proved to be used as a platform to voice opinions, share news, breaking events and initiate relief efforts during disasters. Figure 2.1 shows an example of OSN specific features that enable usage of their unique services.

In order to enjoy any of the listed services offered by OSNs, an individual needs to register herself on



To see what she shares with friends, send her a friend request.

Private Message to followers in Inbox

Connecting "like minded" people

Describing what's new? News / Breaking Events

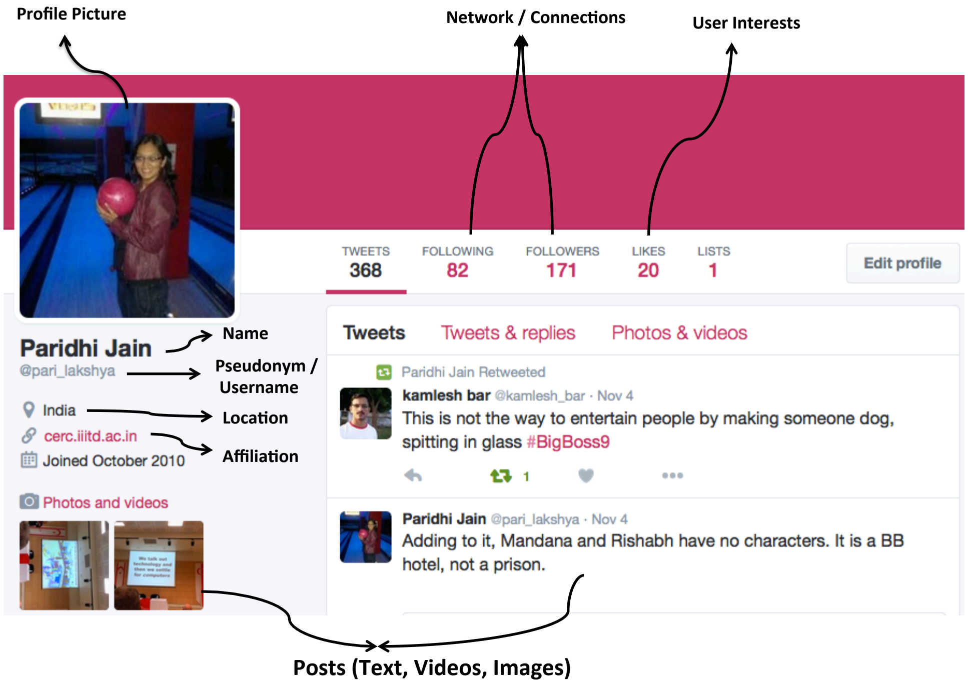
Add new friends, given mutual friends

Chat Feature (Voice / Video)

Figure 2.1: Few OSN specific features that enable unique services to its registered users.

the OSN and create a public identity for others on the platform. A public identity of an individual thus contains her personal information, referred to as *profile attributes*, her connections, referred to as *network attributes*, and her feed of posts created by or shared with her, referred to as *content attributes*. Figure 2.2 shows a snapshot of a user public identity on a social network.

Often, the details entered by an individual on her public identity varies with the purpose of the network. For instance, Foursquare and Yelp mandate users to share their location (or home address) but store a little about their education, while Facebook and LinkedIn ask users to share their education and not location in order to suggest friends and new connections. Social networks serving a common purpose are homogenous, while social networks offering a different set of services are heterogenous in terms of quantity, quantity, veracity and nature of attributes they ask from their registered users. Now, we discuss how can we exploit homogenous networks to aid attribute verification and heterogenous networks to help infer unknown attributes of an individual as we suggest these applications to enterprises and security practitioners. The following description motivates the importance and need of identity resolution across social networks.



Paridhi Jain
@pari_lakshya
India
cerc.iiitd.ac.in
Joined October 2010

Paridhi Jain Retweeted
kamlesh bar @kamlesh_bar · Nov 4
This is not the way to entertain people by making someone dog, spitting in glass #BigBoss9

Paridhi Jain @pari_lakshya · Nov 4
Adding to it, Mandana and Rishabh have no characters. It is a BB hotel, not a prison.

Figure 2.2: Example of a user's public identity on a social network.

2.1.1 Verifying consistent and extracting complementary attributes

Homogenous social networks ask a significantly overlapping set, while heterogenous social networks capture a distinct set of attributes from their users. Labitzke *et al.* present a list of common as well as complementary attributes among Facebook, MySpace, StudiVZ and Xing [65], Irani *et al.* present a similar list for Delicious, Digg, Flickr, Last.fm, LinkedIn, MyJournal, MySpace, Technorati, Twitter and YouTube. *Common attributes* like hometown are consistently available only on Delicious, Flickr, Last.fm, MySpace and YouTube, while birthday is available on Digg, LiveJournal, MySpace and YouTube [52]. *Complementary attributes* like gender and birthday are available only on Facebook but not on Twitter. Further, few OSNs enforce similar policies on the veracity of the information. Chen *et al.* shows that relation based networks like Facebook and Google enforce “real-name" policy and 90% users identity themselves with their real names, while pseudonyms are prevalent on blog based networks like LiveJournal and Blogger [16].

With homogenous OSNs, common attribute values can be checked against each other for consistency if identities of a user on these OSNs can be collated. It is observed that users put the same values on same attributes across OSNs, however if the value differs from the value mentioned on most OSNs, the value can be deemed unreliable and false. Across heterogenous OSNs, unknown attributes of a user can be identified from her collated identities on these OSNs, thus helping in aggregated user profile [38].

2.1.2 Inferring unknown attributes

In many cases, important attributes like sexual orientation, age, gender, political affiliation remain unknown even after collating user identities across networks. Few reasons include – a) no support for these attributes on either of the networks, and b) users intentionally disallow public sharing of these attributes. In these scenarios, literature suggests methods to infer attributes based on user name [72], activity [48, 85, 86], content produced [110] and friend connections [53, 78]. Inferred attributes strengthen the collated identity of a user, further giving away her detailed characteristics.

2.2 Identity Resolution across Online Social Networks

Given the need of identity resolution on social networks, we now formally define identity, identity resolution and the related tasks. *Identity* of a user on an OSN refers to a collective set of her attributes namely, profile, content and network, defined as follows.

- *Profile attributes* describe her persona like username, name, age, location.
- *Content attributes* describe the content she creates or is shared with her such as text, time of post.
- *Network attributes* refer to the connections of the user like number of friends, number of followers.

An individual is denoted by I and her identity on a social network SN_A is denoted by I_A. The task of identity resolution can be formally defined as follows.

Problem Definition 1: *Identity Resolution: Given an identity I_A of user I on social network SN_A, find her identity I_B on social network SN_B using a search function S and a linking function L.*

$$I_B = \max_{1 \leq j \leq N} (L(I_A, I_{Bj})) \quad \mathit{where} \quad I_{Bj} \in S(I_A))$$

Observing the two functions involved, the process of identity resolution in online social networks can be divided into two subprocesses – identity search and identity linking. *Identity search* lists a set of candidate identities on SN_B, which are similar to the given known identity I_A in accordance to the search function S and are suspected to belong to user I. Such a set of candidate identities is represented as $S(I_A)$ and its size is denoted by N. The search function S inputs I_A's attribute value, a defined similarity metric sim_S, and search space (SN_B in this scenario) as arguments, and selects all identities ($I_{B1} \cdots I_{Bj} \cdots I_{BN}$) from the search space for whom similarity sim_S between the candidate's attribute value and I_A attribute values is greater than a threshold. The threshold can be computed empirically for the best precision and recall. The function S cannot be applied on missing but wok with partial information. Time complexity depends on the size of search space.

Identity linking calculates the link-score between I_A and every candidate identity I_{Bj} returned by identity search using a linking function L. The link function inputs attributes of a given identity and a candidate identity, and computes a link-score. L here can be a supervised classifier or a rule based heuristic function and thus link-score can be calculated on one or multiple attributes at the same time. The link-score can thus either be a probability or normalised similarity score. The function L is designed to be efficiently computable given variety of attributes – text, numbers, date and image, and works with both available complete and partial set of values of attributes. Candidate identities are then ranked on the basis of link-score, and the candidate identity with 'maximum' link-score is returned as I_B. Figure 2.3 illustrates an identity resolution process.

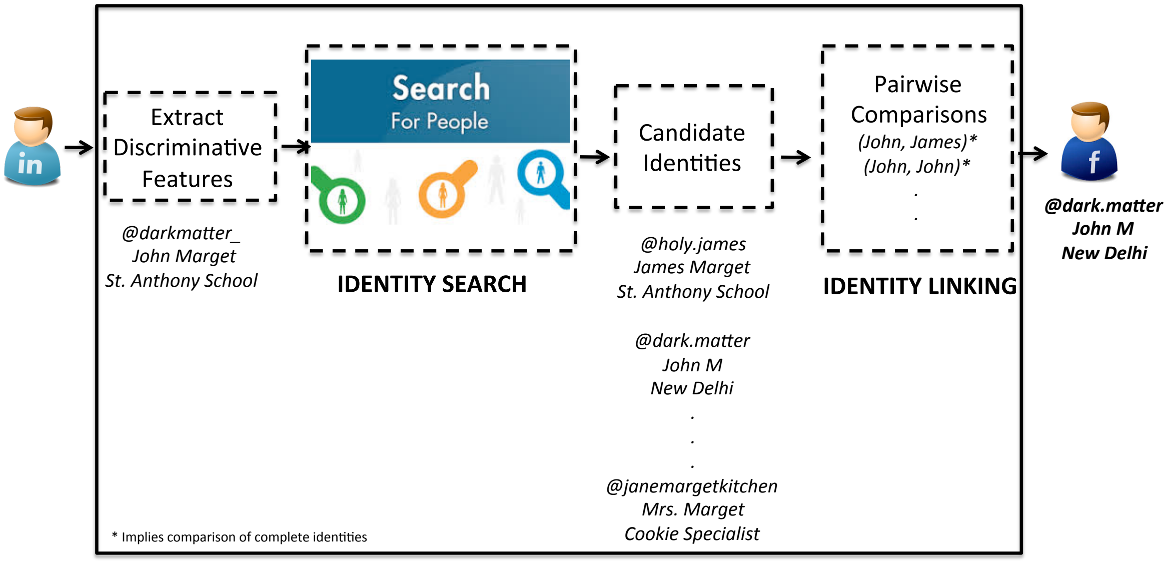


Figure 2.3: Architecture of an identity resolution process.

2.2.1 Identity Search

Problem Definition 2: *For a user I, given her identity I_A on social network SN_A and a search function S, find a set of identities I_{Bj} on social network SN_B such that $sim_S(I_A, I_{Bj}) \geq \theta$, on defined similarity metric sim_S and empirically calculated threshold θ.*

$$\{I_{B1}, \ldots, I_{Bj}, \ldots, I_{BN}\} = S(I_A) \quad s.t. \quad sim_S(I_A, I_{Bj}) \geq \theta$$

Each identity I_{Bj} in the set is termed as *candidate identity* and the set as *candidate set*. The size of the candidate set is termed as *candidate set size* and is denoted by N. Any search method takes a source I_A, a search function S, a search space SN_B, and a predefined set of similarity metrics sim_S as input. Search with I_A retrieves a set of candidate identities that hold similar values for the similarity metrics to the searched identity I_A. For an identity search algorithm, search function can be applied to I_A's attributes defined on her three identity attributes namely profile, content, and network, described below.

- *Identity Search by profile*, implies searching for candidate identities on SN_B with profile attributes of I_A. The candidate identities I_{Bj} are similar to I_A in terms of profile attributes as username, name, gender, school, education, etc.
- *Identity Search by content*, implies searching for candidate identities on SN_B with content attributes of I_A. The candidate identities I_{Bj} are similar to I_A in terms of content creation, URLs posted, platform used for content creation, timestamp, etc.
- *Identity Search by network*, implies searching for candidate identities on SN_B by network attributes of I_A. The candidate identities I_{Bj} are similar to I_A in terms of friends, network in-degree, network out-degree, etc.

A variety of search functions S applied on each of the above set of attributes and numerous similarity metrics sim_S have been discussed in literature. Each of these search methods, criteria and similarity metrics are discussed now.

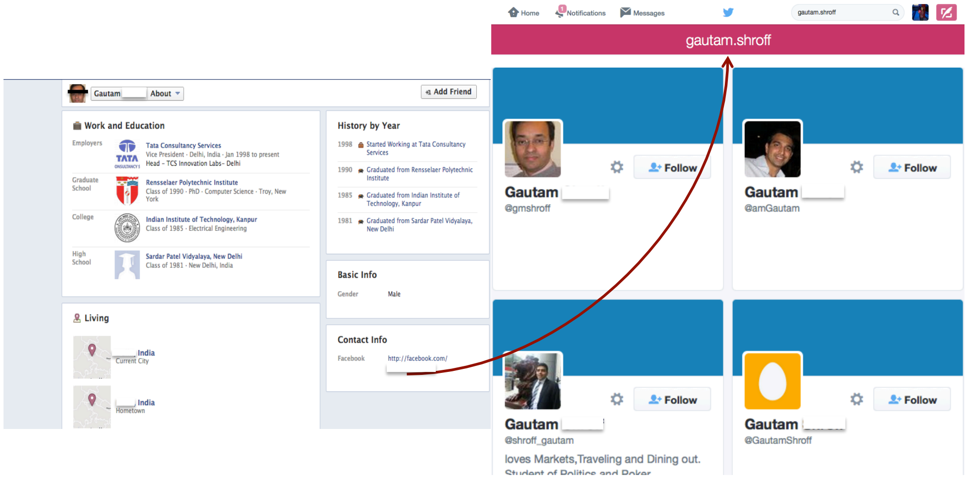
Profile Search

Profile attributes of a user describes her basic characteristics and include name, city, age, date-of-birth, location, bio, photo, interests, and many other attributes. Each of these attributes hold a string, numeric or date value. Profile attributes of a user are invariable across social networks unless a user fakes them or enters false, empty, or erroneous data. Research shows that even though a user has huge control on defining her profile attributes, most users tend to repeat the profile attribute values across networks [16, 27]. Therefore, researchers suggest to use profile attributes as a strong feature to search for candidate identities of a user [13, 65, 74, 79]. All methods employ heuristic approaches to query.

A major challenge to use profile attributes for identity search is the non-homogeneity and non-availability of common set of attributes across all online social networks (or the social networks considered). For example, gender attribute of an identity is available on Facebook, but not on Twitter. A user is uniquely identified by her username on Facebook but is uniquely identified by an integer ID in Google+. Further, certain attributes like 'username' across social networks could be public, while others like 'gender' might be restricted to certain audience only due to privacy concerns.

Many researchers exploited only publicly accessible profile attributes to make comparison between two identities of a user, while others used private information (via user authorization) to make the comparison. Comparisons are mostly made using syntactic and semantic methods [28, 33, 45]. We now discuss each of the profile attributes considered in literature for identity search on online social networks.

- **Username:** Also known as pseudonym and screen name, it is a publicly accessible profile attribute of a user which uniquely identifies her within a social network. A user can choose a username, which may be a compressed form of her name or any nickname. Research shows that around 40% users tend to keep same or similar username across social networks, therefore for such users, username can be used to resolve identities across social networks [16, 27]. Carmagnola *et al.* and Motoyama *et al.* searched with user's pseudonym to find a user's identity on other networks via their APIs [13, 79, 113]. Figure 2.4 shows an example where an individual's username is used to generate candidate set.



Employers: Tata Consultancy Services
Vice President · Delhi, India · Jan 1998 to present
Head – TCS Innovation Labs– Delhi

Graduate School: Rensselaer Polytechnic Institute
Class of 1990 · PhD · Computer Science · Troy, New York

College: Indian Institute of Technology, Kanpur
Class of 1985 · Electrical Engineering

High School: Sardar Patel Vidyalaya, New Delhi
Class of 1981 · New Delhi, India

1998 Started Working at Tata Consultancy Services
1990 Graduated from Rensselaer Polytechnic Institute
1985 Graduated from Indian Institute of Technology, Kanpur
1981 Graduated from Sardar Patel Vidyalaya, New Delhi

Gender: Male

Facebook: http://facebook.com/[redacted]

[redacted] India
Current City

[redacted] India
Hometown

Gautam [redacted]
@gmshroff

Gautam [redacted]
@amGautam

Gautam [redacted]
@shroff_gautam
loves Markets,Traveling and Dining out. Student of Politics and Poker

Gautam [redacted]
@GautamShroff

Figure 2.4: Example of using profile attributes to fetch correct identity of the user on other social network.

- **Name:** This is a descriptive attribute of a user that is publicly accessible on social networks. Similar to username, users can choose to name their identity, as per the choice. Few OSNs like Facebook enforce "real-name policy" and ask users to name their Facebook account as their real-name, while OSNs like Twitter have no such restrictions. Under the assumption that benign users put same name on their accounts, researchers suggest methods to filter similar candidate identities based on search by name [13, 65, 74, 79].

- **AboutMe Attributes:** These attributes of a user on a social network describe the remaining characteristics of a user apart from username and name. For example, education, school, email, picture, description (bio), city, work, etc., are AboutMe attributes. Such attributes can also be used for searching candidates set on social networks. The assumption is that a user reuses her certain AboutMe attributes across social networks and is proven to hold true for certain social networks [16]. AboutMe attributes of two identities can be compared either on the basis of syntactic similarity methods [13, 52, 65, 74, 79] or semantic similarity methods [28]. Syntactic based similarity methods measure if the user has mentioned same value for a AboutMe attribute on both social networks, while semantic similarity further captures the similarity between attributes semantically. For example, string based comparison of city attribute value - New Delhi on one social network and ND, India on other social network, imply that the values are different, while semantically they refer to the same city. Motoyama *et al.* searched with a given user's account location, age, gender, hometown, education, description and school and filtered out candidates for identity linking. Note that, in order to access AboutMe attributes, one needs user authorization to access the information, since the attributes are private.

Though profile attributes are effective in surfacing relevant similar identities of the searched user on other OSNs, they are valid only under the assumption that the user reuses her profile attributes across social networks. But, the assumption is not generalizable for any user. A user is free to use very different variations of the same information across social networks, which string based comparison and semantic based comparison methods may fail to capture. As earlier stated, OSNs also vary in their list of profile attributes. To the best of our knowledge, little has contributed to address these challenges and drawbacks of profile search.

2.2.2 Identity Linking

Problem Definition 3: *Given an identity I_A of user I on social network SN_A, a set of candidate identities $Q = S(I_A) = \{I_{B1}, \ldots, I_{Bj}, \ldots, I_{BN}\}$ on social network SN_B and a linking function L, locate an identity pair (I_A, I_{Bj}) such that $L(I_A\ , I_{Bj}) = \max\{L(I_A, I_{B1}), \ldots, L(I_A, I_{BN})\}$. I_{Bj} with highest link-score is inferred as I_B.*

$$I_B = \max_{1 \leq j \leq N}(L(I_A, I_{Bj})) \quad where \quad I_{Bj} \in Q)$$

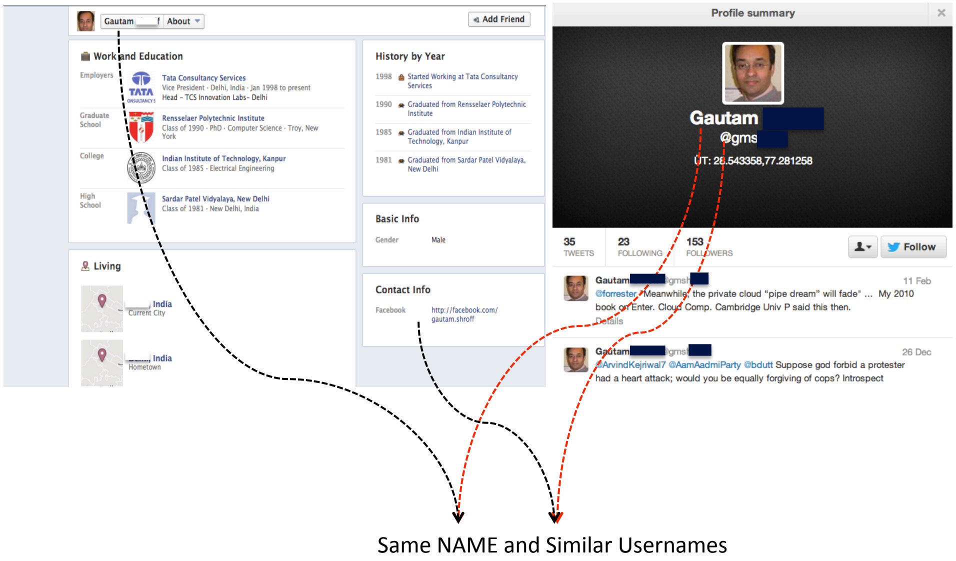
An identity linking method estimates the correspondence between identity I_A and each candidate identity I_{Bj} by calculating a link-score $L(I_A, I_{Bj})$ between their respective attributes and then rank the candidate set on the basis of link-score. Candidate identity I_{Bj} with highest link-score is concluded, as I_B. The function L can be computed for all variety of data – text, date, image and location. The function can either be a supervised classifier decision boundary or a heuristic rule, in both scenarios, the function can be computed with partial and complete information.

Link-score between two identities can be calculated by methods as syntactic similarity methods, semantic similarity methods, image similarity methods, graph matching methods and crowd-sourced methods, applied on attributes such as profile, content and connection network. Syntactic and Semantic similarity methods calculate metrics as edit distance, jaccard distance, jaro distance, soundex, ontology matching, language model similarity etc. on text based profile or content attributes of two given identities I_A and I_{Bj} (e.g., name, username, location, school, posts, tags). Image similarity algorithms calculate similarity between profile (background) images used by two identities. Graph matching methods calculate structural similarities between connection networks of the two identities. Crowd-sourced methods generate human intelligence tasks (HIT) to associate a link-score to each candidate identity, on the basis of human background knowledge and apprehension. We discuss the use of each set of attributes for identity linking in literature now.

Profile Linking

Profile attributes, both public and private, have been extensively exploited to link two user profiles and infer if profiles belong to a single individual. Unique public attributes like username are universally available across OSNs, while private attributes like email, home address, location and gender are available post user permissions. Profile linking methods based on public attributes are scalable and applicable to identities belonging to any set of OSNs. However, methods that link profiles based on private attributes are limited to OSNs that support the attributes.

- **Username:** It is a unique attribute of a user with which she interacts with others and get identified in the network. Researchers argue that around 40% users choose same or similar usernames across their accounts. Perito *et al.*, in their study, further observed that levenshtein distance between usernames of the same individual is less than usernames of different persons [83]. Based on the observation, they build a probabilistic method to link Google and ebay profiles only using username. Likelihood of one username extracted from other is computed using Markov chains and is used as a similarity metric. Using supervised learning framework, authors test their framework on 10 million Google and Ebay profiles and report an accuracy of 71%. Meanwhile, many linking methods use syntactic similarity between usernames as one of the features in learned supervised classifiers to predict link between profiles [13,51,71,72,74,102,113] (see Figure 2.5 as an example). Zafarani *et al.* present state-of-the-art method to boost the linking accuracy to 99% using only usernames. They believe that when users create usernames across their accounts, inherent redundancies relating to user behavior and circumstances infuse. For instance, users of a geographic origin and residence use local language and the respective limited vocabulary words to create usernames and users' limited memory push them to use similar length usernames across networks. Based on similar behavioral characteristics and redundancies derived from a known set of usernames, authors learned a supervised classifier and test if a candidate username belongs to the user who owns the username set [112].



Gautam [redacted] About ▾

Add Friend

Work and Education

Employers — Tata Consultancy Services
Vice President · Delhi, India · Jan 1998 to present
Head – TCS Innovation Labs– Delhi

Graduate School — Rensselaer Polytechnic Institute
Class of 1990 · PhD · Computer Science · Troy, New York

College — Indian Institute of Technology, Kanpur
Class of 1985 · Electrical Engineering

High School — Sardar Patel Vidyalaya, New Delhi
Class of 1981 · New Delhi, India

History by Year

1998 Started Working at Tata Consultancy Services
1990 Graduated from Rensselaer Polytechnic Institute
1985 Graduated from Indian Institute of Technology, Kanpur
1981 Graduated from Sardar Patel Vidyalaya, New Delhi

Basic Info

Gender Male

Contact Info

Facebook http://facebook.com/gautam.shroff

Living

[redacted] India
Current City

[redacted] India
Hometown

Profile summary ✕

Gautam [redacted]
@gms[redacted]
ÜT: 28.543358,77.281258

35 TWEETS | 23 FOLLOWING | 153 FOLLOWERS

Follow

Gautam [redacted]@gmsh[redacted] 11 Feb
@forrester "Meanwhile, the private cloud “pipe dream” will fade" ... My 2010 book on Enter. Cloud Comp. Cambridge Univ P said this then.
Details

Gautam [redacted]@gmsh[redacted] 26 Dec
@ArvindKejriwal7 @AamAadmiParty @bdutt Suppose god forbid a protester had a heart attack; would you be equally forgiving of cops? Introspect

Figure 2.5: Example of using profile attributes as linking parameters to link identities of a single user.

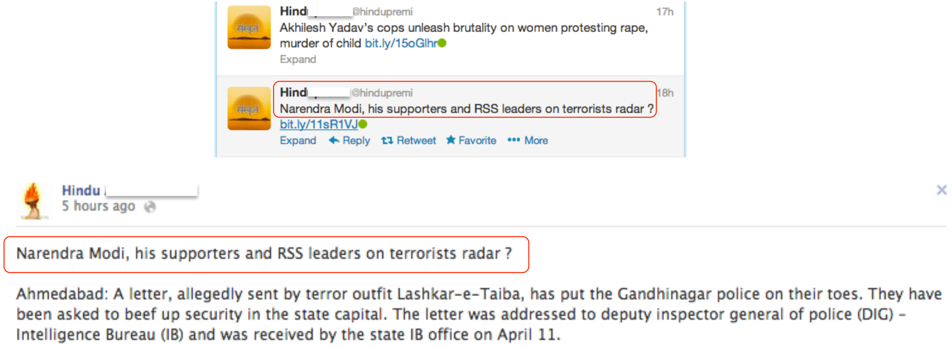
- **AboutMe:** Extension of features to include similarity between other profile attributes further aid profile linking. Researchers suggest methods to use include similarity between profiles' name, location, description and profile picture as metrics to infer link between examined user profiles [13, 71, 72, 74, 79]. Degree of similarity between names is calculated using jaro similarity [74, 79] and n-gram similarity [71] between location using euclidean distance [74] and postal codes set similarity [44], between descriptions by using bag-of-words models and between profile pictures using state-of-the-art face recognition algorithms like SIFT [74].

Note that profile linking methods are effective only when attributes are available and accessible (either privately or publicly) on both examined profiles. However, as discussed in Chapter 1, OSNs are heterogenous and thus availability and support of similar profiles attributes on profiles of different OSNs can be restrictive. However, we now discuss methods based on content and social structure to link identities i.e. using posts, tags and social connections.

Content Linking

By content, we collectively refer to the information created or shared by a user on an OSN in the form of posts. The semantics of a post varies with the OSN. A post can be a video on YouTube, a tweet on Twitter, a picture on Instagram, check-in post on Foursquare and .gif on Vine. Though each post has its own characteristics like the content itself, it is further associated with a set of descriptive features that describe the post i.e. *metadata*. For an instance, a tweet contains text-based attributes but also meta information like date of tweet, length of tweet, hashtags, emoticons, average word length, etc. Beyond profile attributes, efforts in the direction to use posts itself, as well as, metadata for profile linking are enormous, as discussed now.

- **Posts:** Methods discussed here derive similarities between posts created by examined user profile to infer their link to a single individual. Estimating post similarity is rather a non-trivial task. For text-based and location based OSNs (e.g. Twitter and Facebook), posts are tokenized, processed for stop-words removal and then compared using bag-of-words model, n-gram language models [44, 71, 72]. Content linking using posts assumes that a single user is bound to use same or similar words when discussing same or similar topics across their identities on OSNs. Figure 2.6 shows an example where an individual social accounts can be linking by tokenized matching of her posts on Twitter and Facebook. For location-based OSNs, posts are further filtered to extract embedded location. Similarity between locations posted by two OSN profiles is calculated using string metrics and geographic Euclidean distance [74], correspondingly similarity between sets of locations shared by profiles is calculated using K-L divergence of location histograms [44]. For video-based and image-based OSNs, image and video classification experts have made independent efforts [66]. Since, image and video matching methods are complex, sophisticated yet naive, metadata associated with the posts aid faster and efficient profile linking for these OSNs.



Hind [redacted] @hindupremi 17h
Akhilesh Yadav's cops unleash brutality on women protesting rape, murder of child bit.ly/15oGlhr

Hind [redacted] @hindupremi 18h
Narendra Modi, his supporters and RSS leaders on terrorists radar ?
bit.ly/11sR1VJ

Hindu [redacted]
5 hours ago

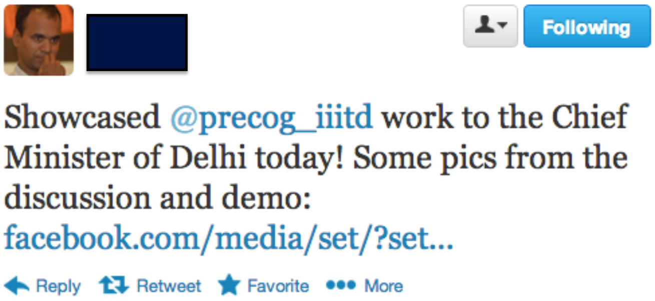
Narendra Modi, his supporters and RSS leaders on terrorists radar ?

Ahmedabad: A letter, allegedly sent by terror outfit Lashkar-e-Taiba, has put the Gandhinagar police on their toes. They have been asked to beef up security in the state capital. The letter was addressed to deputy inspector general of police (DIG) – Intelligence Bureau (IB) and was received by the state IB office on April 11.

Figure 2.6: Example of using content attributes as linking parameters to match identities of a user. Note the exact same posts made by her two identities implying match between the identities.

- **Metadata:** Posts have characteristics like author (who created the post), timestamp (time of creation of the post), location (from where the post is created), tags (describing the post) and other stylistic features of the post itself (e.g., use of short words, long sentences, etc.). Each of these features can be used to compare post created by two identities to resolve under the assumption that posts containing same data are described by similar metadata attributes thus removing the need to compare the posts themselves (image or video).

For text-based networks like Twitter and Facebook, metadata like timestamp, location and description of posts can be exploited. It is shown that the similarity between timestamps of content and language profiles (author's writing style from description of content) resolved 94.7% of Twitter and Yelp identities [44]. Extensive use of authorship analysis techniques (extract lexical features, syntactic features and idiosyncratic features) further help in identity linking [43]. Taking advantage of timing of the posts in addition to stylometric features have further proved to be effective in identity linking [59]. Use of other metadata features extracted from posts e.g. URLs has been proved as an important source of linking. Research shows that an online user has a tendency to cross-pollinate information on Online Social Media. For example, a user behavior is observed to post her uploaded video link from YouTube to Twitter by the authors [57]. Using that, Correa *et al.* finds a user's multiple identities on online social media using Twitter and tried to link Twitter account with them. Authors monitor the content a user posts on Twitter and observed that she posts URLs that point to either one of her Foursquare, YouTube, Flickr and last.fm account. Indirectly, the user is self-mentioning herself on other social networks via posts generated by her on Twitter (see Figure 2.7). In this way, a user's correct identity on multiple social networks can be revealed and the identities can be resolved more accurately (100% claimed accuracy). However, on Twitter there exists a huge user base who do not exhibit the self-mention behavior (assumption made), for when the method fails and is not generic.



[redacted]

Showcased @precog_iiitd work to the Chief Minister of Delhi today! Some pics from the discussion and demo: facebook.com/media/set/?set...

Figure 2.7: Example of self-mention behavior. Twitter user posts a tweet referencing to his Facebook account.

For social tagging networks like Delicious and Flickr, tags created describing the posts are matched to resolve two identities. For example, if a user uploads a picture on Flickr, she can tag the picture with tags like *me*, *fun*, *nature*, etc. The assumption is that a user creates same posts on both OSNs and tag it in the same (or similar) ways. Iofciu *et al.* suggests to match image tags and evaluate the approach on Flickr, Delicious and StumbleUpon [51]. Further, authors search for a user's identity on a third network by collating resolved identities and using the collective information to link the third identity of the user. When evaluated on the same dataset, the authors claimed to link an addition of 20% of the user identities across three networks correctly [51]. Tags are further exploited to correlate user identities, however the authors suggested to first filter (clean) the tags used by a user across social networks and then compare them. The idea is that users are not consistent in using conventions for tags and therefore represent same tag in variety of ways. Using raw tags for comparison may lead to true negatives, while using filtered tags help in removing noise and then compare [96].

Network Linking

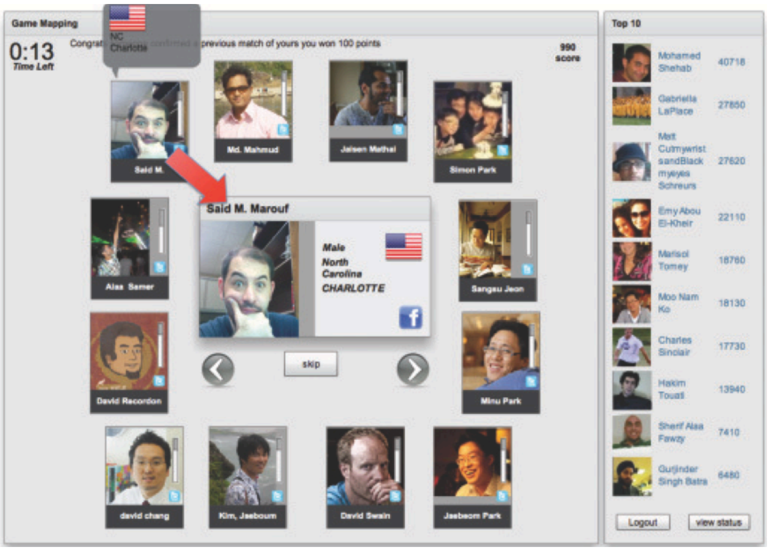
Multiple identities can also be linked via network attributes i.e. social connections. Few deanonymization techniques used network attributes to link one anonymized and one labeled user. The idea was to overlay two friends network of two identities and analyze the network similarity to claim whether two identities belong to the same user or not. Graph theoretic approaches have been proposed and discussed in literature for identity deanonymization and resolution [81,94]. Narayanan *et al.* used a graph theoretic approach to de-anonymize Twitter users with the use of labelled Flickr network [81]. Authors iteratively matched each node network using a set of seed users (pre-deanonymized users) to find to most similar node with the similar friend network and claimed 30.8% accuracy. However the method needed 150 labelled seed users in anonymized network and Flickr network, each having more than 80 friends. Recent research improvises on seed selection techniques by using unsupervised clustering methods on profile attributes [23], graph and subgraph matching methods to find social structure similarity [1,71,113]

Labitzke *et al.* followed a different approach of matching mutual friends between two identities (to be matched). Authors used string matching methods to link names of common friends of two identities. If there exist more than three mutual friends with same name, the two identities were marked as linked (belonging to same person) [65]. However, the approach had a gap of understanding that in real world, there could be multiple mutual friends between two users, or no mutual friends (in case when user used different social networks for different purposes).

Apart from the syntactic and semantic methods applied to each attribute set of a user on online social networks, researchers have devised techniques to use third party information sources to infer if two identities refer to the same entity. These third party information sources can be a Google Search Engine or a human annotator. Note that social networks are indexed by search engines and therefore users (entities) are searchable on search engines. The fact has been exploited by few researchers [8].

CrowdSourcing Linking

Parallel to resolving records in databases with human intelligence, resolving multiple identities of an entity across social networks has been proposed [90]. The idea is to present two online identities of a user to a human, and ask her judgement of whether the two identities belong to the same user. However sole exploitation of human judgement in identity resolution, increases the burden on a human as in huge social networks, there exists large number of identity pairs to resolve. To reduce the overhead, researchers suggested to pre-compute a set of candidate identities which satisfy a threshold as a proof to be possibly linked together and then present it to a human. For example, first profile attributes of identities are matched and if the similarity score is above a threshold, the pair is presented to the human to further resolve if the identities refer to the same entity. The pre-computation of candidate identity pairs to be presented to human annotators, saves time and human effort and also addresses the scalability issue. Such a semi-automated system solves the identity resolution problem by exploiting the power of automated techniques as well as human intelligence. To save the process from unwanted wrong human annotation or any guessing work from the human, the identity pair is presented to multiple humans before judging whether the references belong to the same entity (see Figure 2.8).



Congrat[illegible] previous match of yours you won 100 points

Said M. Marouf

Male
North Carolina
CHARLOTTE

Mohamed Shehab	40718
Gabriella LaPlace	27850
Matt Cutmywrist sandBlack myeyes Schreurs	27620
Emy Abou El-Kheir	22110
Marisol Tomey	18760
Moo Nam Ko	18130
Charles Sinclair	17730
Hakim Touati	13940
Sherif Alaa Fawzy	7410
Gurjinder Singh Batra	6480

Figure 2.8: Example of using crowdsourcing / human intelligence to match identities of a user.

Search Engine Linking

Bilge *et al.* used a "search and link" approach to link multiple identities. Authors suggested to pick few identity attributes as first name, last name, occupation and education from each identity and to query a search engine with the extracted attributes for each identity (see Figure 2.9). If the top three results turned out same, then the two identities belong to same entity [8].

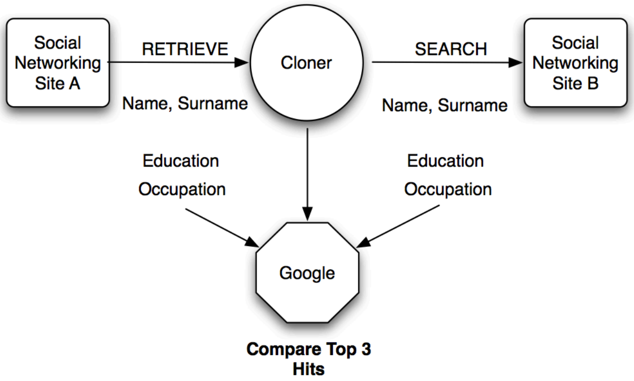


Figure 2.9: Methodology of search engine linking.

Real-world implementations and systems for identity resolution like `PeekYou`,[1] `Pipl` [2] and `Yasni` [3] are extensively used by people today. Given a user's name or location, these public online portals search in vast space of criminal records, court records, yellow page directory, news, publications, blogs, social profiles, IMs, phone directory to find relevant information about the searcher user in real-time. These portals are flexible in their query parameters. One can search for a user with her name, email, phone number, age, company, education or interests. Methods deployed to filter and rank the query results are not published in public domain. These portals extensively focus on user search but not linking, thus presents its users with a long list of candidates that might match with their searched users.

2.3 Entity Resolution across Databases

Prior to the problem of duplicate identities across OSNs, extensive research attempts to address the problem of duplicate entities across multiple data sources like databases and publication records (e.g. DBLP) [35]. Few of the profile linking methods discussed on OSNs are motivated by entity linking methods in databases. Relevant for this thesis background, we briefly discuss methods for entity resolution.

Extensive research in the field of study discusses automatic methods and techniques to resolve records across multiple tables / databases, to understand if they refer to the same entity, and therefore can be merged [11, 21, 35, 68]. Methods suggested rely on the attributes a record has and values the attributes hold. Each record may have variety of attributes, however each attribute is restricted to hold any value of one of the three types – characters, numbers and dates. In literature,

[1] www.peekyou.com
[2] www.pipl.com
[3] www.yasni.com

numbers and dates are considered as characters only. Therefore, record matching techniques in databases compare characters and strings (a group of characters). If the similarity score between attribute strings in two records is above a threshold, the strings are considered to be same and therefore the records belong to the same entity. Resolution techniques based on character and string matching are termed as *Syntactic Resolution Methods*.

Syntactic resolution methods are efficient in capturing small variations in different attribute values introduced by either spelling mistakes or human errors. However, syntactic resolution techniques suffers from two issues – different representation of same information by significantly different strings and decision on hard thresholds to verify if the two strings are similar. To overcome these issues, comparison of meaning of two strings is suggested, rather than the string themselves. Resolution techniques which compare the meaning of the attribute values are termed as *Semantic Resolution Methods*. Semantic resolution methods represent an entity in a predefined framework termed as ontology, and compare the two entities either represented in same ontology or different ontology, to understand if the two entities are instances of each other or vary minimally with each other [7, 9, 28, 30, 33, 34, 45, 98].

Both syntactic and semantic resolution techniques are automatic methods with no human intervention. However the applicability of each of the methods are highly dependent on the nature of data available (similar attribute values) and the support structure required (ontology) in databases. To overcome the dependency, researchers suggested to exploit human intelligence to resolve two records. The idea is to present a pair of records to be resolved by a human annotator. If more than two humans agree on the decision made on the record pair, the records are assigned that decision (refer to same entity or not). Resolution techniques based on human intelligence and annotation are termed as *Crowdsourcing Resolution Methods* [31, 103].

We now discuss each of the methods in detail in the following subsections. Note that, all the approaches discussed here are successful and applicable if no intentional obfuscation or manipulation of attribute values exists in the database. We also discuss the research work, which aims to make the entity resolution techniques fast and scalable.

2.3.1 Syntactic Resolution

If an attribute of two records hold same value, the records are easier to resolve. However, the attribute may hold the same information but is represented differently across records. For example, New Delhi and ND represents the same information about a city attribute. Syntactic resolution techniques captures such variations of same information, human errors, use of abbreviations and different spellings to infer if the attribute values are similar and help in resolving records. To capture string variations (most of the attributes hold string values), string based similarity metrics are proposed. The metrics calculate a similarity score between given two attribute values, and if the similarity score is above a threshold, the values are considered to be multiple representations of the same information [11, 35]. String based comparison of all attributes of one record with the attributes of other record, generates a similarity score between two records. If two records has high similarity score, the records are resolved to refer to same entity, else different entities. Following are the major string based similarity metrics, discussed in literature.

Edit Distance

Edit distance between two strings is defined as number of single character inserts, deletes and substitutions required to change one string into another. It is also known as Levenshtein distance [67]. If edit distance between two strings is less than a threshold, two strings are assumed to be possible variations of each other. Edit distance has been proposed to compare two attribute values of two records, to understand if the attribute values represent same information and therefore same entity. Edit distance fails when one of the two strings is an abbreviated or shortened version of the other.

Affine Gap

Affine gap between two strings takes into account the abbreviated variations of an information. It calculates the distance between two strings as edit distance, however allows two more operations – open a gap and extend a gap [104]. Opening a gap implies the checkpoint from which onwards one must start putting gaps rather than any other insertions, in order to convert an abbreviated string to other string. Extending a gap implies the operation of adding gaps rather than other character insertions. Affine gap penalizes extend a gap operations much lower than open a gap operations and therefore, for two strings in which one string is an abbreviated or shortened version of the other, affine gap is lower than edit distance.

However, affine gap fails when characters in two strings are exchanged from their positions. For example, a name “John Smith" and “SmithJ" might have large affine gap and edit distance score, which misleads that the two strings do not represent the same entity.

Jaro Distance

Jaro distance addresses the above concern by taking into account the number of characters overlap between two strings within allowed position shifts between the characters [58]. A mathematical definition is given by –

$$d_j = \begin{cases} 0, & \text{if m} = 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right), & \text{otherwise} \end{cases}$$

Here, m represents number of matching characters between two strings and t represents half the number of transpositions. Jaro distance penalizes a little for mismatched positions of the characters present in both strings. Other variants of Jaro distance have been proposed as Jaro-Winkler distance. Jaro-Winkler distance gives better scores to the strings which share a common prefix of length l. Jaro distance fails if the positional difference between two strings is beyond the allowed shifts. For example, “Alice bruce Bob" and “Bob bruce Alice", allowed positional shift is six, while ‘B’ in Bob is separated by 12 positions. Therefore, the two strings has only five matching characters ‘bruce’.

Smith-Waterman Distance

Originated for DNA sequencing, Smith-Waterman distance measures the local sequence alignment matches between two strings [93]. It first aligns the two strings in a manner so as to maximize local subsequence matches and then compare the two strings. Smith-Waterman distance penalizes for a mismatch of characters in the aligned strings, and generate a score on the basis of number of character matches in the two aligned strings.

Jaccard q-gram distance

Jaccard q-gram distance measures similarity between two strings in terms of the similarity between their smaller units termed as q-grams. A q-gram is defined as q characters coupled together in sequence from a string [101]. For example, q-grams of a string ‘hello’ with q=2, are ‘he’, ‘el’, ‘ll’, ‘lo’. Jaccard q-gram distance measures the number of q-grams present in both strings, irrespective of their position or frequency and is defined as –

$$Jaccard_{qgram} = \frac{\textit{q-grams(s1)} \cap \textit{q-grams(s2)}}{\textit{q-grams(s1)} \cup \textit{q-grams(s2)}}$$

Q-gram distance assigns high score to the strings with same or close spellings even when the strings vary largely. For example, Jaccard q-gram similarity score assign high score to “Chris" and “Rishi", even though the two strings are not same or similar. To overcome such scenarios, value of q is tuned to a higher value i.e. q = 3 or more. However, slight variations in any of the strings, may decrease the number of q-gram matches and therefore is given a low score.

Q-gram Tf-Idf Cosine similarity

Q-gram Tf-Idf Cosine similarity metric between two strings compute the cosine similarity between tf-idf vectors of q-grams extracted from each string [25]. Given two strings, q-grams are extracted and tf-idf score for each q-gram is calculated. Term frequency (Tf) and Idf (Inverse document frequency) are two parameters which weights a q-gram on the basis of its frequency and its rarity. Term frequency computes how many number of times a q-gram is repeated in the string. Inverse document frequency captures how infrequent / rate the q-gram is in the strings. Cosine similarity measures how similar the two strings are in terms of q-gram and is computed as –

$$cos\theta \ = \frac{\vec{S_1}.\vec{S_2}}{|\vec{S_1}||\vec{S_2}|}$$

where $\vec{S_1}$ and $\vec{S_2}$ are tf-idf weighted q-gram vectors.

Apart from character based similarity metrics, researchers proposed to use phonetic based similarity metrics to capture the possible spelling mistakes, and different phonetics of the same word in different languages and dictionary. Soundex is an example of phonetic based similarity metric. Further, to compare numbers in the attribute values of two records, very few approaches are suggested. Numbers are treated as strings and similar string based similarity metrics are applied to compare numbers.

2.3.2 Semantic Resolution

Certain entity attributes may be represented completely by different words and strings, however they may refer to the same information. For example, an entity with “United States" as location attribute and an entity with “Baltimore" as location attribute may refer to the same entity, since Baltimore is located in United States. To capture such semantically related information values, semantic methods are developed to connect multiple database records [30, 33]. Each record is reflected on an ontology, to capture its attributes and give a meaning to each of them. For example, entity X lives in Y located in Z. According to the ontology, Y is mapped to a city and Z is mapped to a country and according to the RDF rules, if X lives in Y then X lives in Z too. Records are then linked together on syntactic matching of semantically related attributes e.g. location in one record and city in other. The approach is successful, however needs a structured information support and an ontology predefined for every entity in the database.

Some of the popular entity resolution frameworks developed by researchers include Stanford Entity Resolution Framework (SERF) [2], D-Dupe [9], TAILOR [34], Matching Based Object Matching System (MOMA) [98], MARLIN [7], Self-Tuning Entity Matching (STEM) [62]. Each of the system exploits semantic approach to resolve records in a relational database using matching attribute set [63].

2.3.3 CrowdSourcing Resolution

Methods suggested to syntactically and semantically match the attribute values are threshold-dependent, complex and time consuming, with high false positive and true negative rates. As a solution to it, researchers proposed the approach of utilizing human intelligence and background knowledge to resolve entities in a database [6, 31, 76, 103, 106]. A pair of two records to resolve, is presented to a human, to mark if two records possibly belong to the same entity. However, with large databases and therefore large number of record pairs, crowd sourced resolution takes huge amount of time and human effort. To address this concern, an idea to first filter out non-matching records by syntactic matching and then present only confusing pair of records to a human was suggested [103]. This solution saves the human effort by pruning out explicitly non-matching records. Results show that the hybrid approach of exploiting crowd intelligence with computational power gives better accuracy in less time.

2.3.4 Other resolution methods

Apart from Syntactic, Semantic and Crowdsourcing based entity resolution methods, researchers have discussed the use of markov logic networks [92], graphical network structures [5,18] and iterative (collective) resolution [4,5] to solve the problem of entity resolution in databases. Singla *et al.* have exploited first order logic rules to understand if a record predicate or reverse predicate are equivalent, then an inference can be made that the records refer to the same entity [92]. Authors experimented their approach on two citation databases – Cora and BibServ. Chen *et al.* and Bhattacharya *et al.* approached the entity resolution problem by mapping each reference as a node and its relation with other references as an edge in a graph (co-occurence). Chen *et al.* suggested that if any two references are similar with high confidence, they are resolved into one, and the unknown mappings are then mapped using the network structure [18], while Bhattacharya *et al.* proposed to use network structures between references to find common entities co-occurring with each reference and if there exist large common network between two references, the two references tend to point the same real-world entity [5]. Further, same set of researchers proposed an iterative resolution methodology where a set of references are resolved given that the references that they are connected to (or with which they co-occur) gets resolved first [4]. The process is iterative to start with the references of most confident similar references and then continue with resolving the entire database. Chen *et al.* experimented on citation and movie databases, while Bhattacharya *et al.* experimented with arXiv and Elsevier BioBase citation databases.

2.4 Research Gaps

2.4.1 Restrictive database methods

Unfortunately, methods for entity resolution in databases are not directly applicable to identity resolution on social networks. Firstly, entity resolution methods find it difficult to scale with the size of the databases wherein datasets of social networks have billion of user records to be compared and linked. Various strategies like mapping large databases as graphs and then partitioning the large graph into smaller graphs, perform entity resolution on each of the smaller graphs and then integrate them to combine the resolved entities [82] or by proposing inverted indexing techniques facilitating fast linking [22] have been proposes to deal with large databases.

Secondly, databases used for entity resolution contain detailed and homogenous characteristics of the entities, while social networks for identity resolution are heterogenous and have a few characteristics of the identities to consider for linking. Thirdly, entities' network structure is limited i.e. entities have a fewer connections with other entities in databases such as DBLP and Citeseer [5]. Identities, on the other hand, share hundreds and thousands of connections with other user identities. Database methods do not take opportunistic advantage of the complex network structure for entity resolution. We, therefore, believe that there is a need to devise novel methods that cater to complexities, scalability, volume, veracity and variety of identities on social networks.

2.4.2 Limited search attributes

Identity Search algorithms based on profile attributes are effective but have limitations and have not been exploited to its potential [55]. Firstly, search by profile attributes is highly restrictive, and dependent on the availability of same profile attributes across networks. For example, 'gender' profile attribute is available on Facebook, while no such attribute exists on Twitter. Location profile attribute is public in Twitter, while is private on Facebook. Therefore, a search algorithm may have access to limited profile attributes to use for searching.

Secondly, search by limited profile attributes results in large number of candidate identities which have similar profile attributes e.g. same name, similar username or similar location. Matching large number of candidate identities becomes computationally expensive and time consuming.

Thirdly, search by profile attributes may miss identities for those users, who use significantly different profile attributes across social networks, either purposely or unintentionally. For such users, candidate set may never contain the correct identity of the user. This results in lower accuracy of complete identity resolution process.

Lastly, URL attribute of a profile has been discussed in literature but has not been exploited in any of the profile based identity search methods. We think that URLs mentioned as a profile attribute on one social network may help in locating a user's identity on other social networks. Search by limited profile attributes may not give satisfying results.

We observe that search methods on the basis of content and network attributes remain unexplored. Content and Network attributes are important aspects of a user's identity on a social network. Due to advanced services to push content simultaneously on multiple online social networks, users post same / similar content across networks. Search by content can help in finding such users' identities across networks. Further, a segment of users tend to connect with similar people across social networks [79] and therefore search by network, may also help in finding the identities of a user across networks. In this work, we attempt to understand if inclusion of search methods based on an identity's content and network attributes, along with search methods based on an identity's profile attributes can help in improving the accuracy of the identity resolution process in online social networks.

2.4.3 Ineffective linking limited to attributes' current values

Existing identity linking algorithms access only latest versions of the examined user identities. They assume high similarity among the current values of the attributes for respective users. However, current values may remain dissimilar for users' accounts for reasons such as attribute evolution over time [73] or a user's choice of dissimilar values for anonymization. In both scenarios, current values falsely direct existing identity linking algorithms to infer accounts of a single user as different users. Therefore, there is a need to consider history of values along with current values to examine a potential link between two user identities.

2.5 Summary

In nutshell, there exists a vast literature on identity search and identity linking methods exploiting private and public attribute information of an online user. Few of these methods are motivated and borrowed from entity resolution in databases. For example, similarity metrics to measure correspondence between a set of identities are used from the entity resolution literature. Building on the same, more sophisticated techniques like feature selection, unsupervised machine learning, and graph modelling are introduced to correlate user identities. Evaluations on simulated and real-world datasets prove significance and effectiveness of suggested methods.

In this thesis, we improvise on identity search and identity linking methods to gain better profile linking accuracy on a real-world dataset using only publicly available information. Research gaps, as discussed earlier, motivate our intention to use exhaustive search parameters and novel information for linking profiles, that we explore in this thesis work.

Chapter 3

Identity Search

Here, we address the research gap of limited search attributes. Content and network characteristics, along with profile attributes, are proposed to be used as search parameters to find relevant candidate set of identities on a social network. Given publicly available information of user I_A on social network SN_A, a candidate set of identities is generated on social network SN_B. First, we discuss heuristic methods based on unique user behavioral characteristics, followed by an unsupervised search method for candidate selection on a social network.

3.1 Heuristic Search on available attributes

Heuristic methods suggested in this work capture unique user behavioral characteristics of sharing few of her attributes publicly i.e., either her profile, posts or connections. Improving on the runtime of identity search, we then suggest unsupervised clustering methods to segment SN_B and allow apriori candidate set generation. Suggested methods access only publicly available data about a user as compared to other algorithms proposed in the literature that access detailed information about a user.

3.1.1 Publicly available attributes

Search based on attributes can succeed only if the attributes are publicly accessible on both social networks. To explore what these attributes are, we find the attribute availability matrix for four social networks – Twitter, Facebook, LinkedIn and Quora (see Table 3.1).

Table 3.1: Available features across social networks. A set of discriminative features are chosen from these as searching criteria in identity search. **Tw**: Twitter, **Fb**: Facebook, **Qu**: Quora, and **Ln**: Linkedin.

Type	**Raw feature**	**Availability**				**Extracted feature**
		Tw	**Fb**	**Qu**	**Ln**	
Profile	Username	✓	✓	✓	✓	Jaro distance LCS distance Levenshtein distance Length difference Char Bi-gram Jaccard index Char Bi-gram Cosine similarity Entropy difference
	Name	✓	✓	✓	✓	Jaro distance LCS distance Levenshtein distance Length difference Char Bi-gram Jaccard index Char Bi-gram Cosine similarity Entropy difference
	Location	✓	-	✓	✓	Ratio of common locations Ratio of common postal codes
	Bio	✓	-	✓	✓	Bio length difference Bio words distribution POS tags distribution Jaccard index of bio words Cosine similarity of bio words Ratio of miss-spelled words
	Profile Image	✓	✓	✓	✓	Histogram Similarity Face similarity
	Gender	-	✓	-	-	Boolean
	Language	✓	✓	-	-	Boolean
Content	Posts	✓	✓	✓	-	Activity distribution Device distribution Application distribution URLs distribution POS tags distribution Avg. # of multimedia shared Avg. # of words Avg. # of miss-spelled words Avg. length of words

3.1.2 Heuristic methods

Based on the common attributes that can be used for search, four search methods are defined: **Profile Search**, **Content Search**, **Self-mention Search** and **Network Search**. We discuss these methods now.

Profile Search

Profile attributes describe basic information about a user like name, location, gender, etc. If the user does not demonstrate any active obfuscation and does not create altogether a different identity, it is likely that she re-uses certain profile attributes' value on the social networks she joins. If the user demonstrates such behavior, profile attributes can be used in a search function S to find her identity on other social networks.

Algorithm 1 describes the method. First, we use I_A's username and query SN_A API to extract public details like her name, username, location, profile image, and URL. We use URL attribute first to observe if I_A herself has given her SN_B identity (I_B). We term this behavior of mentioning one's SN_B network identity (or any other network identity) on SN_A explicitly, as "*Self-Identification*". We observe two varieties of self-identification behavior – i) a user directly gives her other OSN identity on her URL attribute, ii) a user indirectly gives her OSN identity via referring to a webpage on her URL attribute, that contains her OSN identity. A user referring to her blog on Twitter URL with her blog having her Facebook identity is an example of indirect self-identification. If I_A has not identified herself via URL, we use her username, name and location attribute to query SN_B API to find identities with same or similar username / name having the same or similar location. Note that, since we use SN_B's API to search, we have limited control on selected threshold θ to claim similarity between names of a retrieved candidate and a given identity. The SN_B API only allows to specify the parameters to search for. The search function S, sim_S and θ is defined by the API and details are not shared.

Identities (users, pages and communities) who either have same name as the queried name or a part of queried name in their name and share the queried location are captured in the candidate set. We also search for a candidate identity who has the same username as I_A's username. The reason for the same username search is motivated by the previous research which shows that around 30%-40% users have same username across social networks [27]. We aggregate I_A's candidate identities and term the set as non-ranked candidate set.

Content Search

A user creates content to share her activities, interests and knowledge with others. Owing to the popularity of social aggregation sites and ways to link multiple networks together, a user is facilitated with a choice to push the same content on multiple networks simultaneously. For example, Twitter provides a functionality to connect Twitter and Facebook identity to post user's tweets on her Facebook timeline, Twitterfeed,[1] HootSuite,[2] Friendfeed[3] allows a user to connect Twitter, Face-

[1] http://www.twitterfeed.com

[2] http://www.hootsuite.com

[3] http://www.friendfeed.com

Algorithm 1 Heuristic Search Methods

```
procedure PROFILE SEARCH
    $I_A \leftarrow$ known identity on $SN_A$
    $S \leftarrow \{I_A.\text{url}, I_A.\text{username}, I_A.\text{name}, I_A.\text{location}\}$
    if $S[0]$ directs to $SN_B$ then
        $I_B \leftarrow S[0]$
        exit
    else
        delete S[0]
        for each $s$ in $S$ do
            query $SN_B$ API with $s$ and retrieve candidates
            $C_{xs} \leftarrow$ candidates
            add $C_{xs}$ to $C_x$
    add an identity on $SN_B$ with $I_A$.username to $C_{xs}$
    return $C_{xs}$
```

book, and LinkedIn to push feeds in three social networks simultaneously. Because of such services, it is likely that a user generates same content on multiple social networks. Such a user behavior can be exposed by using metadata like “source" of a post i.e., from which device / application the tweet is posted. Source can be exploited to reduce the search space for a user's online identities, if an analyst intends to save her efforts by searching for a user in only social networks where the user hints existence. Content Search method uses content in the search function S for users of these services. Again, since we query SN_B's API, we have little information on how API returns results for the queried content. Therefore, we implement another sim_S on top of the results from the API.

Algorithm 2 explains the pseudocode of content search method. We extract most recent 100 (or less)[4] posts by I_A on SN_A, and process each of the posts to limit the length to 75 characters and to remove non-ASCII characters. We query SN_B API with the processed post to search for the users who posted same or similar content on SN_B. API returns a candidate set of identities who posted similar content as queried content. We choose cosine similarity as our similarity metric sim_S and remove all such candidate identities that hold zero cosine similarity between words of their posts and I_A's posts. Cosine similarity between two posts is calculated as,

$$Cosine\_sim(I_A, I_{Bj}) = \frac{\overrightarrow{P_{I_A}}.\overrightarrow{P_{I_{Bj}}}}{|\overrightarrow{P_{I_A}}||\overrightarrow{P_{I_{Bj}}}|}$$

where $\overrightarrow{P_{I_A}}$ and $\overrightarrow{P_{I_{Bj}}}$ are word-frequency vector of post by I_A and post by candidate identity I_{Bj}, respectively.

[4] We limit to process most recent 100 posts to avoid long execution time. The intention to capture the timely interest of the user. Processing more than 500 posts can reveal mixed and temporal interests of the user and hinder identity resolution.

Algorithm 2 Heuristic Search Methods

```
procedure Content Search
    $I_A \leftarrow$ known identity on $SN_A$
    $S \leftarrow \{I_A.\text{source}, I_A.\text{posts}\}$
    if $S[0] \in \{\text{HootSuite, TwitterFeed, Facebook}\}$ then
        $posts \leftarrow S[1]$
        for each $m$ in $posts$ do
            remove stop-words and non-ascii characters from $m$
            limi to 75 characters
            query $SN_B$ API with $m$ and retrieve candidates with similar posts
            $C_{xs} \leftarrow$ candidates
            for each $c$ in $C_{xs}$ do
                if $sim(c.post, m) \leq 0$ then
                    delete $c$ from $C_{xs}$
            add $C_{xs}$ to $C_x$
    return $C_{xs}$
```

Self-mention Search

This method exploits a user's tendency to cross-pollinate information on Online Social Media [57] and was introduced by Correa *et al.* [27]. The method explores content attributes of I_A and assumes that if I_A has accounts on two or more networks, she might cross refer to her other account, in few of her posts. For example, I_A might post a tweet with a URL referring to an album on Flickr, indirectly revealing her Flickr identity. We term this behavior of posting URLs indirectly but consciously, pointing to user's other network identity as "*Self-mention*". Self-mention behavior allows identity leaks via content created in the form of URLs by the user. This method exploits self-mention behavior to search for identities of a user across networks.

Algorithm 3 illustrates the procedure. We query SN_A API to extract 100 (or less) recent posts by I_A and filter out the posts with URLs. Each URL is processed to verify if it directs to SN_B. We create a set of all URLs directing to SN_B, query SN_B API to process each URL and extract identity of the candidate user, thereby creating a set of candidate identities.

Another way of using content attributes of I_A as search parameter is by extracting user attributes which are innocuously shared in posts. Phone (Mobile) number is an identifiable information with which a real-world individual can be associated uniquely, in most cases [114]. We observe that large number of mobile numbers are shared publicly on OSNs via profile attributes [16] or via posts (see Figure 6.1). Using regular expressions, queried user's posts can be examined to understand if there exists a mobile number in her posts. A mobile number is then used to search the queried user's identity on other OSNs. Our choice of using the mobile number as a search parameter is based on our study of mobile number sharing behavior on OSNs (discussed in detail in Chapter 6).

Algorithm 3 Heuristic Search Methods

```
procedure SELF-MENTION SEARCH
    $I_A \leftarrow$ known identity on $SN_A$
    $S \leftarrow \{I_A.\text{posts}\}$
    for each $m$ in $I_A.posts$ do
        $urls \leftarrow$ Extract urls from $m$
        if $urls.\text{length} > 0$ then
            for $u$ in $urls$ do
                if $u$ redirects to $SN_B$ then
                    candidate $\leftarrow$ redirected $SN_B$ identity
                    $C_{xs} \leftarrow$ candidate
        add $C_{xs}$ to $C_x$
    return $C_{xs}$
```

We pre-define regular expression to filter a mobile number from a post. Based on the regular expressions and I_A's messages on SN_A, we filter out mobile numbers shared by I_A. We then search SN_B API with each mobile number and list candidates that shared the same mobile number in a post on SN_B. Candidate identities are then collated for all the mobile numbers found posted by I_A.

Network Search

Connections of a user on a social network define her network. A user needs other users to define her connection attributes. If a user leaks her identity on any other social network, identities of users connected with her risk a leak. Network Search method explores the possibility of a user's identity leak via her network attribute.

We search for I_A's identity on SN_B using her connection network. By exploiting self-identification behavior of users in connection network of I_A on SN_A, her *candidate neighborhood* on SN_B is identified. A candidate neighborhood of I_A is composed of SN_B users whose corresponding identities are connected to I_A on SN_A. Users in the candidate neighborhood of I_A are then queried via SN_B API to retrieve their connections. Any such connection similar to I_A is captured in the candidate set. The assumption is that I_A and I_B connects to the same subset of users on both social networks. We, thus, try to map I_A's identity from one social network to another via mapping her connection network on two social networks (see Algorithm 4). Note that the method is applicable, even when the incomplete neighborhood of any user is available, as compared to other graph-based search methods, which require complete neighborhood of multiple users to find I_B [81].

Algorithm 4 Heuristic Search Methods

```
procedure NETWORK SEARCH
    $I_A \leftarrow$ known identity on $SN_A$
    $S \leftarrow \{I_A$.name, $I_A$.connections$\}$
    for each $friend$ in $I_A.connections$ do
        $friend_B \leftarrow$ Get identity of $friend$ on $SN_B$ using self-identification
        if $friend_B$ is not $NULL$ then
            $friend_B$.connections $\leftarrow$ Get connections of $friend_B$ on $SN_B$
            for $u$ in $friend_B$.connections do
                if similarity($u$, $I_A.name > 0.8$) then
                    candidate $\leftarrow u$
                    $C_{xs} \leftarrow$ candidate
        add $C_{xs}$ to $C_x$
    return $C_{xs}$
```

3.1.3 Identity resolution framework

Figure 6.3 describes the identity resolution framework with the proposed search methods and state-of-the-art identity linking method. Candidate identities of each search method are collated together and ranked using standard approaches for identity linking. We use username syntactic linking and profile image linking for linking. We then rank the candidate set by the match-score associate with each candidate set. The aim of ranking is to retrieve the correct identity of the queried user within top results. The ranked candidate set is then presented to a human manual verifier to locate the correct identity among the candidate identities. We assume that the human verifier is 100% accurate, in making the inferences. In this work, authors are the human verifiers.

3.1.4 Evaluation

We evaluate the identity resolution framework on two popular social networks – Twitter and Facebook. We assume access to a Twitter profile, use it as a source identity of the user and look for candidate identities on Facebook, in order to retrieve her correct identity. We borrow the ground truth dataset from [74] collected from Social Graph API. The dataset consists of 543 users who self-identify themselves on both Twitter and Facebook. These users can be regular online individuals, community, organizations, and celebrities. We do not restrict our evaluation on any specific kind of users and believe that the methodology is independent of such a choice. With these 543 users, denoted by U_{total}, we query the framework and record the number of users for whom the framework retrieves correct Facebook identity in the candidate set, denoted by $U_{correct}$. Therefore, *accuracy* of the framework is the ratio of $U_{correct}$ to U_{total}.

We observe that for 220 Twitter users (40.5%), the system retrieves their correct Facebook identities. Table 3.2 lists the contribution of each search algorithm. We further compare our identity search

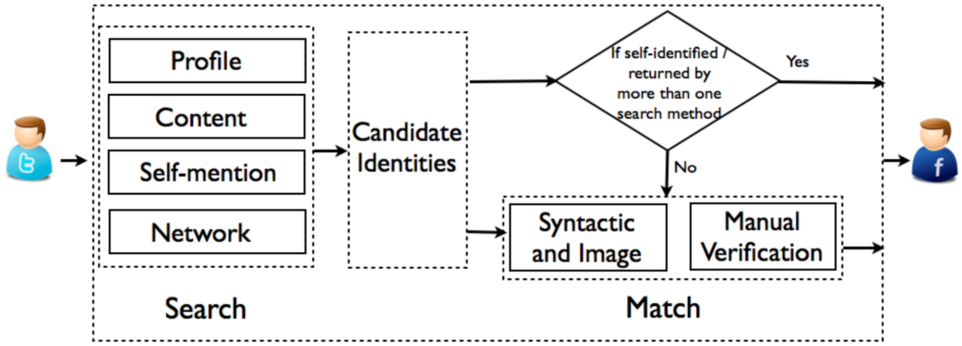


Figure 3.1: Architecture of the identity resolution framework using proposed heuristic search methods and linking methods from literature.

Table 3.2: Evaluation of the identity resolution framework with contribution of each search algorithm in the resolution accuracy. Search methods based on profile (url), content, self-mention and network attributes improve resolution accuracy by 13.1%.

Search Algorithm	$U_{correct}$	**Accuracy**
Profile Search (P)	205	37.7%
Content Search (C)	3	0.5%
Self-mention Search (SM)	31	5.7%
Network Search (N)	1	0.2%
Identity Search (P+C+SM+N)	220	**40.5%**
P (without URL)	149	27.4%
P (with URL) + (C+SM) + N	149+71	27.4% +**13.1%**

with the traditional profile search used in the literature, assuming access to only public profile attributes. Traditional profile search method finds candidate identities by search parameters – username, name, and location. To the best of our knowledge, no profile search method exploited an important profile attribute, URL attribute of an identity, to understand if a user herself has directly or indirectly self-identity themselves. We include the URL attribute and improvise profile search method. Table 3.2 shows a comparison of using traditional profile search methods with improvised and proposed identity search methods, to search for a user's Facebook identity. We observe that an additional 13.1% users are identified by the combination of improvised profile and proposed identity search methods.

In summary, we extend literature method of profile (name and username) based search to search methods based on profile (name, username, URL), content and network attributes. Our search methods assume user behavior like 'cross-posting' and successfully fetch I_B in a candidate set. Evaluation on real-world users show the improvement in identity resolution accuracy by 13%. Therefore, we conclude that multiple facets of a user's identity help in deriving and identifying her on other OSNs.

3.2 Unsupervised Search on discriminative attributes

Heuristic identity search methods are real-time i.e., given a user's known identity, the methods aim to fetch the candidate set via SN_B API. Thus, run-time of the identity resolution framework increases with increasing number of identities to search in on SN_B. To avoid such dependency, we suggest to segment SN_B into a set of clusters and given a user identity on SN_A, select her candidate identities by finding the most suitable cluster out of segmented SN_B. The suitable cluster is expected to contain identities similar to the searched user. After that, supervised approaches match each candidate and known user identity to assign link-scores. The novelty of this work lies in the choice of clustering parameters, clustering method and identification method of the most suitable cluster for a searched user, keeping low time and space complexity.

The choice of clustering parameters is based on the discriminative power of the parameter. Here, a parameter refers to a feature or an attribute of the user identity like age or gender. Discriminative features help in finding if identities refer to same or different users as these features have different values for both genres. Via feature analysis, we first aim to identify discriminative features among the profile and content attributes of a user. Taking the identified discriminative features, we then choose a clustering algorithm that create overlapping clusters. We need overlapping clusters of user identities on SN_B as an identity can intrinsically be a candidate for more than one searched user. Now, as each cluster is a set of user identities, estimating a centroid of the cluster and compare it with the searched user identity to find the similarity is another task. We start the description with the feature analysis now.

3.2.1 Finding discriminative attributes

We conduct an in–depth feature analysis for four heterogenous social networks – Twitter, Facebook, Quora and LinkedIn. Table 3.1 shows a variety of features that are publicly available for our analysis. To identify the discriminative power of a feature, we define three metrics that are calculated by processing pre–labelled ground truth data. These metrics are estimated using pre-labelled ground truth data. The data contains user identities that belong to a single user, termed as 'positive' or 'match' class, and user identities that belong to different users, termed as 'negative' or 'no-match' class. The metrics are –

- **Class Majority Index (CMI):** The purpose of this index is to empirically identify a threshold $CMI(Feature)$ that acts a distinguishing point for the data. For each feature, the CDF (cumulative distribution function) is plotted individually for both classes against the range of values of *Feature*. This gives us information about the distribution of the feature values. We define $CMI(Feature)$ as the intersection point of these graphs. Further, a feature is discriminative if $CMI(Feature)$ is a point that divides the data such that, majority (e.g. 80%) of the data points on either side of $CMI(Feature)$ belong to the same class. For features, where this doesn't hold true, we term those as non–discriminative. Figure 3.2 shows the CDF curves with different similarity metrics used for username. The intersection point of the CDF curves signify the class distribution at that particular distance. As seen in Figure 3.2, Levenshtein distance and Normalized LCS features are discriminative as majority data of the same class lies on either side of $CMI(Feature)$.

- **Encroachment Index (EI):** The metric aims to capture the percentage of feature values on either side of $CMI(Feature)$ that are confusing. In other words, EI signifies how deep one needs to go into the opposite class to reach one–class purity. We define encroachment index ($EI(Feature)$) for each class as

$$EI(Feature)_M = \frac{|min(ClassNM) - CMI(Feature)|}{|min(ClassM) - CMI(Feature)|} \tag{3.1}$$

where, $min(ClassNM)$ is the minimum value that the feature takes for the encroaching class, and $min(ClassM)$ is the minimum value that the feature takes for the calculating class. We do a similar thing for the other class,

$$EI(Feature)_{NM} = \frac{|max(ClassM) - CMI(Feature)|}{|max(ClassNM) - CMI(Feature)|} \tag{3.2}$$

This second index is of importance for the following reason. Say, we define a threshold above which we have 20% user–pairs of the non–match class. If we experimentally conclude that the non–match class completely encroaches into the match class. Then this means that some fraction of the number of non–match pairs have values equal to the highest (or for some metrics lowest) value of the match pair. This is an indicator of the metric not being discriminative in our study. For some features, the range of each class can be the same but the distribution of data across classes on either side of $CMI(Feature)$ differs. In this case we modify $EI(Feature)$ to capture the difference in the standard deviation and variance of the data distribution rather than just the range alone.

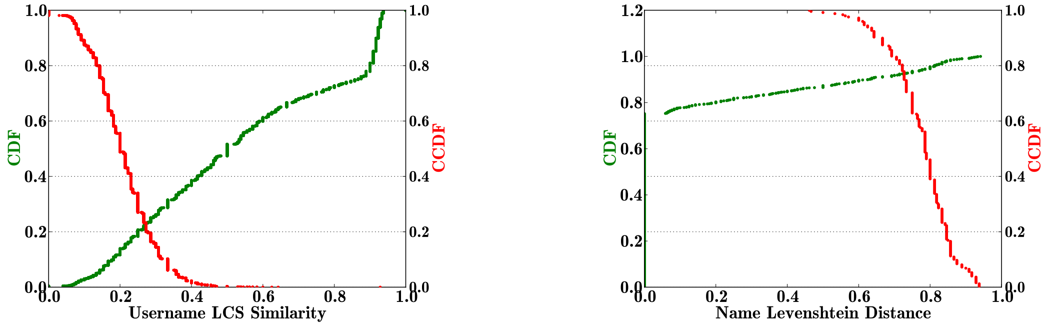
- **Error Index:** The metric captures the percentage of users that hold feature values beyond (or below) $CMI(Feature)$. In other words, the index captures the type-I error rate (false negatives) and type-II error rate (false positives). It is defined as:

$$ErI(Feature) = max(type - I,\, type - II) \tag{3.3}$$

We intend to balance the two types error here. Any feature that increases either error is considered as non-discriminative.

Table 3.3: Features evaluated on the three metrics to estimate discriminative power using pre-labelled dataset for Twitter-Quora users.

Feature	**Range**	**CMI**	**EI**	**ErI**
Username: Levenshtein distance	[0 - 1]	0.76	0.49 [M], 0.95 [NM]	25%
Username: Jaro similarity	[0 - 1]	0.54	0.59 [M], 0.97 [NM]	24%
Username: LCS similarity	[0 - 1]	0.27	0.46 [M], 0.92 [NM]	22%
Username: Char Jaccard index	[0 - 1]	0.33	0.58 [M], 0.98 [NM]	22%
Username: Keyboard distance	[0 - 100]	21	0.98 [M], 1.00 [NM]	40%
Username: Char cosine similarity	[0 - 1]	0.10	0.52 [M], 1.00 [NM]	20%
Name Levenshtein distance	[0 - 1]	0.64	0.09 [M], 0.97 [NM]	10%
Name: Jaro similarity	[0 - 1]	0.63	0.16 [M], 1.00 [NM]	9%
Name: LCS similarity	[0 - 1]	0.38	0.11 [M], 1.00 [NM]	9%
Name: Char Jaccard index	[0 - 1]	0.53	0.25 [M], 1.00 [NM]	11%
Name: Keyboard distance	[0 - 75]	17.8	0.36 [M], 0.54 [NM]	11%
Name: Char Cosine similarity	[0 - 1]	0.19	0.25 [M], 0.68 [NM]	7%
Bio: Char Jaccard Index	[0 - 1]	0	0.15 [M], 1.00 [NM]	100%
Bio: Ratio difference of mis-spelled to all words	[0 - 25]	2	1.00 [M], 0.04 [NM]	42%
Bio: # of words	[0 - 25]	12	0.68 [M], 1.00 [NM]	49%



(a) Username normalized LCS similarity

(b) Name normalized levenshtein distance

Figure 3.2: Feature analysis to find discriminative features. Discriminative features separate match class (green) and no-match class (red).

Table 3.3 shows the three metrics for few profile and content features out of 35 features listed in Table 3.1. Username and name features like levenshtein distance, LCS similarity and char cosine similarity returns the lowest **ErI** and **EI** among other features. Hence, both these features are the most discriminative and important features. Similarly, other features are ranked on the basis of decreasing **ErI** and **EI**. Note that, content features return the highest **ErI** and **EI** and hence, we filter out content features for any analysis in searching candidate set.

3.2.2 Canopy clustering

Discriminative features are used to identify the search space for candidate selection. In an ideal world, the candidate space for every user profile is the searched network SN_B. Searching through a network is computationally expensive and time consuming. To reduce the time complexity, the focus is on using unsupervised methods. The search network is pre–processed and segmented. Clustering approaches are used in this phase. A unique clustering algorithm used in our framework is an adaptation of canopy clustering [77]. Canopy clustering was introduced to reduce the computational overhead in clustering and process large scale data. Canopy clustering algorithm uses two thresholds to create clusters of user identities – loose threshold (T_1) and tight threshold (T_2). This can be interpreted as two user identities are in the same cluster if their chosen feature value (say name or username char cosine similarity) is greater than T_1 and are not be a part of any other cluster if the feature value is greater than T_2 (Algorithm 5). The intuition behind this algorithm is to create overlapping canopies of user identities (attributes) based on two thresholds. We need overlapping canopies of user identities on SN_B as an identity can intrinsically be a candidate for more than one searched user. Required clustering parameters are:

- **Thresholds T_1 and T_2**: Results of feature relevancy analysis help in deciding T_1 and T_2. For each pair of networks and for each feature, the thresholds change. CMI of features, e.g. username jaro similarity, is chosen T1, further a tighter threshold of 0.9 or 0.95 whichever is higher than T_1 is chosen as T_2. The reason for choosing CMI as T_1 is as follows. Fig 3.2(b) shows the plot for username across the match and the no-match class using LCS similarity measure for user profiles on Twitter and Quora. As is clear from the plots, for close to 80% of matched user profiles, the LCS similarity between usernames is greater than 0.27. For the no-match class on the other hand, around 80% of the user profiles in the no-match class have LCS similarity lesser than 0.27. Hence 0.27 act as a discriminative threshold value and is chosen as the threshold T_1 for comparing usernames across Twitter and Quora. A similar analysis is carried out for all pairs of networks and all metrics to find T_1 and T_2. Each pair of networks is observed to have different thresholds. This follows from the fact that there are some networks where users are more likely to give their accurate name than in others. The calculated threshold for name Jaro similarity feature in the case of Facebook and Linkedin is 0.9. This value is high as both these networks are used to maintain a friend network (Facebook) and a professional network (LinkedIn) where name is an important attribute. In case of Twitter, the threshold is lower (0.8). This is lower as both of these are not used for official purposes and users might use nick names and acronyms as their name attribute.

- **Cluster centroid**: Canopy clustering is a distribution–based clustering algorithm and hence does not have an inherently defined cluster centroid. As step is an intermediate towards matching, there is a need to define the cluster centroid to understand the belongingness of a given user profile. For numerical data, mean of all points in a cluster is used to identify the cluster center. However, for string–based data such as names and usernames, defining measures is non–trivial. We consider a centroid to be equivalent to a D-dimensional point where D is equal to the cardinality of the alphabet under consideration. We are working with the English alphabets and hence D= 26. Figuring out the value of the i^{th} coordinate of the centroid is equivalent to determining the average frequency of the i^{th} character of the alphabet ($i = 1$ corresponds to the character 'a' in the English alphabet). The formula for determining the value of this i^{th} coordinate is as follows:

$$C_i = \sum_{j=1}^{N} \frac{f_{ij}}{N} \tag{3.4}$$

Where f_{ij} is the frequency of the i_{th} character in the j_{th} member of the cluster, N is the total number of members in the cluster.

3.2.3 Unsupervised search to find most suitable cluster

In order to determine the appropriate cluster for the user, we compute the square of the Euclidean distance between the frequency distribution of the user profile we are looking for and each cluster representative. The minimum distance is then chosen as the suitable cluster and we get our candidate set. For example, as per the representation of cluster centroid, each centroid is represented by a 26-dimensional point. The user identity for whom we have to find the cluster is also converted to a 26-dimensional representation. So, I_A is represented as a point with a=2, d=1 and m=1. In the steps stated below, we take the example for 3-dimensional points. This can be easily extended to D-dimensional points.

1. We represent I_A as (1,1,1) and three clusters C_1, C_2, and C_3 with centroids as (1,0,1), (2,0,0) and (1,3,4) respectively.
2. Consider cluster C_1. Compute the distance as a square difference. In this case, it is equal to $[\ (1-1)^2 + (1-0)^2 + (1-1)^2\ ] = 1$.
3. Repeat (b) for each cluster.
4. Determine the cluster that has minimum distance to our user. We call this the most suitable cluster for our user. In this case, C_1 is the most suitable cluster.

The general formula to determine the distance between two D–dimensional points p_x and p_y is given as follows:

$$Distance = \sum_{i=1}^{D} (p_{xi} - p_{yi})^2 \tag{3.5}$$

Algorithm 5 Algorithm for Canopy Clustering

```
procedure CANOPY-CLUSTERING
$U \leftarrow$ set of user profiles on the network
$T_1 \leftarrow$ *loose threshold*
$T_2 \leftarrow$ *tight threshold*
$d(x, y) \leftarrow$ *distance measure*
*for each user-profile $x$ in $U$*:
    create canopy $C_x$ such that
        for each user-profile y in U:
            insert y into $C_x$ if $d(x, y) > T_1$;
for each user-profile y selected in the previous step
    remove y from U if $d(x, y) > T_2$;
```

Here, p_{xi} and p_{yi} denote the value of the i^{th} coordinate for points p_x and p_y. This searching step to find the most suitable cluster is repeated for each feature on which we make clusters in SN_B. For instance, we find the most suitable cluster for user identity I_A on each feature. This gives us a set of overlapping clusters C_1, C_2, C_3, ..., C_m. We then take a union of these clusters to determine a candidate set for I_A on SN_B. Note that, this approach may not scale to larger networks due to the space complexity of canopy clustering i.e.$O(n^2)$. Hence, we propose an alternate approach.

3.2.4 Modified canopy clustering and search

Canopy clustering produces overlapping clusters and is O(n^2) in time complexity. The modified algorithm uses a single threshold and produces non overlapping clusters. The space complexity now decreases to $O(n)$. The search algorithm is modified and a concept of 'sibling' clusters is introduced. As non–overlapping clustering tend to miss out some probable candidates, extending this constrained set with siblings results in higher accuracy. The algorithm is given as Algorithm 6.

Algorithm 6 Modification to the Canopies

```
procedure MOD-CANOPIES
$U \leftarrow$ set of user-profiles on the network
$T \leftarrow$ *threshold*
$d(x, y) \leftarrow$ *distance measure*
*for each user-profile $x$ in $U$*:
    create canopy $C_x$ such that
        for each user-profile y in U,
        insert y into $C_x$ if $d(x, y) <$ T;
    Remove all user profiles y added in the previous step from U.
    **loop while U is not empty**;
```

The algorithm is similar to canopy clustering and its time complexity is still O(n^2) in the worst case. Space complexity however reduced to $O(n)$, since clusters are non-overlapping.

The search algorithm to identify the candidate set for a given user is given by Algorithm 7 and is explained as follows. After determining the closest cluster, we determine 'siblings' of this cluster. These are clusters that are similar to the cluster. We evaluate the distance of this cluster to other clusters and define similar clusters as those which are closer than a specific threshold.[5] We then evaluate the distance of our user profile to these siblings and expand the candidate set obtained thus far.

The union of the candidate clusters is used to define the candidate set. This process is repeated for all features. In the Algorithm 7, the distance measure d(C_x,C_y) is calculated as follows. We represent each canopy C_x by its centroid (which as we explained earlier is nothing but a 26-dimensional point). We then calculate the square of the Euclidean distance between the points to determine their distance. We experimente with different values of threshold T to determine the most optimal one. With a very small value, we cannot be able to expand our candidate set since we will not find any sibling clusters whereas with a extremely value, the candidate set can be too large making the algorithm computationally expensive. The empirical threshold T for our dataset is set to 12.

Algorithm 7 Unsupervised search method

```
procedure MODIFIED-SEARCH
    $U \leftarrow$ User profile we are looking for
    $C \leftarrow$ set of non overlapping clusters
    $T \leftarrow$ threshold
    $d(C_x, C_y) \leftarrow$ distance measure
for each cluster $C_x$ in $C$:
    compute the distance d(U, $C_x$)
select cluster $C_m$ such that
    d(U, $C_m$) is minimum of all distances
    computed above, this is the most suitable cluster;
$L \leftarrow$ List of suitable clusters, initially empty
for each cluster $C_x$ in $C$:
    if d($C_m$,$C_x$) $<$ T then
        if d(U, $C_x$) $<$ T then
            append $C_x$ to L
L holds our list of candidate clusters
```

3.2.5 Identity resolution framework

In this section, we present the identity resolution framework with our proposed search methods and supervised linking method (see Figure 3.3). The identity linking stage starts with the candidate set and helps identify a single identity for the queried user. The linking phase of our framework has

[5] After experimenting over a range of thresholds, we find the best results for threshold equal to 12.

two steps: (1) Assign match probabilities (scores) to each user in the candidate set and (2) rank the candidate set based on the score.

As stated earlier, identity linking can be approached as a binary classification problem with Match and No–match class. *Match* class shows that two profiles belong to the same individual. *No–match* class represents different users. The matching function is learned based on all 35 features listed in Table 3.1 using state-of-the-art supervised machine learning algorithms. Once the matching function is learned, each identity in the candidate set is compared with the known identity and assigned a match probability that signifies the confidence of matching. The goodness of the classifier is evaluated based on ROC curves and precision–recall graphs. Our experiments show that Random Forest and Naive Bayes are best suited for this framework. Here, we report results for Random forest used as a classifier in our experiments.

The next task is to pick one from the candidate set. One can pick the best match candidate identity (with highest match probability) and declare it as the correct identity. This results in exactly one identity for the queried user. However, there is a possibility that the best match probability is low. Hence, the output is not the correct identity. To overcome this disadvantage, we define a threshold above which the candidate is considered a match. Using this approach, we may not get any matched identities or at times get multiple match identities. A fall back option can be using human annotations to verify manually a candidate as a match. This approach relies on the goodness of the threshold. Another approach is to use human feedback to identify the correct candidate. This approach is useful only when the candidate set is small and an expert is interacting with the system. We explore both the best–pick and the threshold–approach in this work.

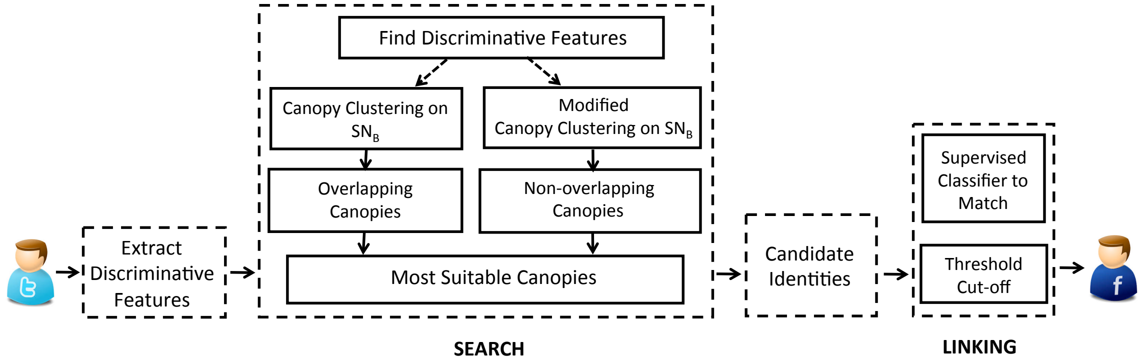

Find Discriminative Features
Canopy Clustering on SN_B
Modified Canopy Clustering on SN_B
Overlapping Canopies
Non-overlapping Canopies
Most Suitable Canopies
Extract Discriminative Features
Candidate Identities
Supervised Classifier to Match
Threshold Cut-off
SEARCH
LINKING

Figure 3.3: Architecture of the identity resolution framework using unsupervised search method and supervised linking method.

3.2.6 Evaluation

This section presents the various experiments that show comparison for different combinations of features, clustering, and classification parameters. Here, we focus on the effect of change in accuracy of the overall identity resolution rather than focusing on individual tasks. We first present the assumptions and the dataset details, and then discuss different accuracy levels within a pair of networks and across networks.

The experiments are carried out on datasets of varying sizes across different networks. For the smaller size dataset, users are split in the ratio 70:30.(i.e., 70% of the users from the positive set and 30% from the negative dataset.) Here, M denotes the users in match class and NM denotes the users in the no–match class. Table 3.4 shows the data split across networks and Table 3.5 describes the results of identity resolution. Note that, the choice of discriminative features like username and name led to achieving comparable precision and recall, otherwise achieved with all features. Clearly, they are the most discriminative features. Therefore, the stage of feature analysis is important to find such features. We also observe that precision and recall values are dependent on OSNs to which examined identities belong to. As seen from the Table 3.5, the system achieves the best recall for Facebook-Quora (0.85) and highest precision for Quora-LinkedIn. It is important to note here, that the state-of-the-art in the industry for this problem is around 20% accuracy and in academia other researchers have not attempted looking at a diversity in networks.

Table 3.4: Class splits used for experiments with different OSN pairs.

Network-Pair	**Total Users**	**M**	**NM**
Quora-Linkedin	731	512	219
Facebook-Quora	1027	719	308
Facebook-Linkedin	2377	1664	713
Facebook-Twitter	2267	1587	680
Twitter-Linkedin	657	460	197
Twitter-Quora	1000	700	300

To generalize the results, we experiment on larger datasets for Facebook and Twitter. In these experiments, data is split evenly between the match and the no–match class. The probability threshold for classifier prediction score on a pair in the identity linking phase is varied to see the change in precision and recall. The results are shown in Table 3.6. On increasing the threshold values, precision increases and recall decreases. With **MOD-CANOPY**, precision and recall improve as compared to the original version. We further test if modified canopy algorithm is successful, because it can fetch the correct user identity of I_A in the candidate set from SN_B.

Canopy v/s Modified Canopy Clustering

Two approaches based on canopy clustering have been presented so far. Table 3.7 shows a comparison between their performance. Success is defined as match if the target user appeared in the candidate set, failure is otherwise.

Table 3.5: Precision and Recall for the small dataset experiments for each OSN pair.

Network Pair	**Matches**	**All features**		**Only names and username**		**Less name and username**	
		Precision	Recall	Precision	Recall	Precision	Recall
Quora-Linkedin	512	0.67	0.71	0.71	0.72	0.55	0.71
Facebook-Quora	719	0.69	0.85	0.68	0.84	0.5	0.85
Facebook-Linkedin	1664	0.45	0.78	0.43	0.77	0.36	0.76
Facebook-Twitter	1587	0.51	0.78	0.51	0.78	0.6	0.71
Twitter-Linkedin	460	0.33	0.75	0.31	0.75	0.24	0.74
Twitter-Quora	700	0.28	0.72	0.29	0.72	0.15	0.72

Table 3.6: Experimental results for a larger database (Facebook and Twitter) using canopy and modified canopy clustering as the search step. Total users include both M and NM users present in ratio 1:1.

Users	**Threshold**	**Precision (Canopy)**	**Recall (Canopy)**	**Precision (MOD-Canopy)**	**Recall (MOD-Canopy)**
20000	0.95	0.15	0.9	0.25	0.79
20000	0.97	0.2	0.7	0.3	0.55
20000	0.98	0.24	0.62	0.33	0.69

As shown in the results, the two methods are fairly close w.r.t accuracy. While clustering on a very large dataset (5000 users), the modified method is more suitable owing to its lower space complexity. However, searching is slower. For a very small dataset, the traditional canopy clustering approach is also efficient. We also observe the effect of the use of clustering at all. The precision for identity resolution for a dataset of 1000 users (500M, 500NM) before clustering was 0.44, while it increased to 0.52 after clustering. The recall was unaltered at 0.68 in both cases. Hence, along with reduction in computational complexity, we also achieve better accuracy due to the introduction of this step.

Effect of varying feature combinations

Another parameter that was varied through the experiments is the feature set. We experimented with using only name and username features and less name and username features (i.e., we applied a single metric for name and username). We compared this with the results obtained using all features. Table 3.6 shows the precision and recall obtained from the different feature sets used for each pair of networks. We see that the precision obtained by using all features and by using only name and username features is comparable; though using all features gives better results. Also, precision obtained using less names and usernames is slightly lower than the other two. Hence, we can conclude that name and username are clearly the most important features that give satisfactory results. However, the addition of extra features improves precision to an extent. So, these features, while not too useful by themselves, are useful in conjunction with name and username for identity aggregation.

In summary, we present a detailed system comprising of in-depth feature analysis methods, can-

Users	$CANOPY$		$MOD-CANOPY$	
	Success	% Success	Success	% Success
100	80	80	83	83
1000	708	70.8	706	70.6
2000	1290	65	1220	61
5000	3368	68	3176	64

Table 3.7: Comparing the original and modified clustering methods.

didate selection approach, and a unique framework that aims at identity resolution. Through experiments and results, we conclude that doing prior feature selection using relevancy metrics aids in the selection of better candidate set. Further, we also show that having an intermediate candidate selection step reduces the computational overhead for identity linking and use of unsupervised clustering reduces the complexity of identity search itself. A modification to clustering introduced to suit the application and our comparisons show that choosing between clustering and the modified version, there is a tradeoff between computational efficiency and accuracy.

3.3 Discussion

Identity search methods serve an important task in the identity resolution process. Given a user identity on one OSN, it is important to fetch her identity on other OSN in a set of shortlisted / candidate identities. To do so, one must remember that a user can not be assumed to: a) reveal all her attributes, b) demonstrate similarity of only one facet of the personality say, name or username. Since OSNs serve different purposes and enforce different privacy policies, the user may choose to either repeat self-representation (profile), views (content) and friends (connections) or choose completely different attributes to define her identities across OSNs. Even if the user repeats herself, she may expose the same to a limited audience (no public access). In these scenarios, it is necessary to find methods that explore public attributes and capture many similarities across the user identities. Our deployed methods prove the idea of using all public attributes beneficial, however in scenarios where the identities do not have any similarity or the identity of the user does not exist on the searched network, our methods may fail and result in false positives.

To avoid false positives, we suggest following improvisations in future: a) weighing search parameters, b) including history of attributes to search. Candidate identities can further be pruned by weighing search parameters. Candidates returned on less-confident (low weight) search criteria like gender can be pruned to reduce the candidate set size while maintaining its quality.[6] Weights can be decided either empirically or based on domain knowledge. Past values of attributes like name or username can also be used to search for candidates. Literature studies show that users change their

[6] Quality of a candidate set is determined by the inclusion of the most similar and correct identities of the searcher user.

attributes over time. Users who keep dis-similar identities now may have similar identities across OSNs in the past. We believe that these extensions can strengthen our methods against returning bad and false candidates.

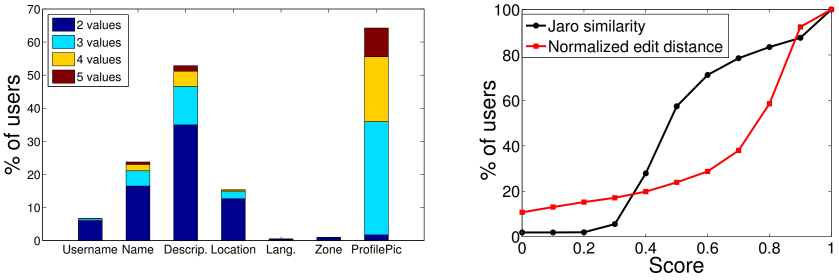
Chapter 4

Identity Linking

We now work on devising methods to predict if user identities (given and a candidate) belonging to different networks refer to a single individual i.e. identity linking. Existing linking methods compare attributes like 'username' and 'name' to find the connection between a pair of identities. However, challenges like dissonant social platforms with a partially overlapping list of supported attributes and heterogenous attributes holding veracious values impede efficient identity linking. Literature suggests various methodologies equipped with tools that compare common attributes between examined user identities 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 [51, 52, 71, 79, 96]. These methodologies consider most recent (current) values of the attributes and assume high similarity to infer a link between respective identities. However, current values may have low similarity for reasons such as user's choice to maintain privacy or attribute evolution over time as described below [68, 73]. In this chapter, we study only profile attributes and hence, refer to an identity as a profile interchangeably.

User's choice: Characterization studies on OSNs suggest that users consistently keep same values for their attributes like their name, gender, location, across OSNs [16, 17]. Zafarani *et. al* shows that 59% users create similar usernames across their profiles for reasons such as to represent a universal identity in online space or to ease remembering [111]. Remaining 41% users choose dissimilar usernames for reasons such as to maintain privacy and avoid de-anonymization [70]. 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



(a)

(b)

Figure 4.1: 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.

exhibit a tendency to evolve their attributes over time [39, 54, 73]. Consider the following scenario – A user registers on Twitter and Facebook with the same username value; she favors Twitter and updates her Twitter profile more frequently than her Facebook profile. After a few weeks, 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 the 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 4.1(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. 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 4.1(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. Thus, low similarity between current usernames can be falsely manipulated by existing methods as different users.

For users who evolve and select dissimilar attribute values across profiles, we propose to take advantage of 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 for an attribute, say username, 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 essential and discriminating attribute for profile linking [74,83,112]. Though a considerate section of users changes username on Twitter (≈10%), it is the most common publicly available attribute across OSNs that can uniquely identify 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?*

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.

We now formally define the research question for this work using following definitions and notations. User profiles under examination, that belong to a pair of social networks, SN_A and SN_B, are 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 receptive evolution timestamps, find if U_S and U_C refer to the same user $\mathcal{I}$.*

4.1 Methodology

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. Features extracts 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; Labelled datasets collect users who evolve with profiles on popular social networks followed by supervised classification by an ensemble of classifiers organized in a framework.

4.2 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 than 50% matching attribute values across different OSNs [16]. These attributes, however, evolve over time, leading to matching histories (i.e. overlapping / similar past values) than current values. Further, a recent study shows that users exhibit similar choices while selecting usernames across OSNs [112]. 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 linguistic structure of the usernames like choice of abu-

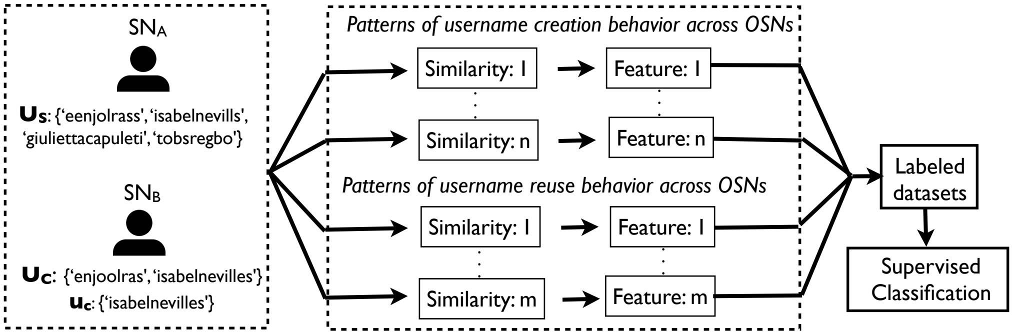

SN_A
U_S: {'eenjolrass', 'isabelnevills', 'giuliettacapuleti', 'tobsregbo'}
SN_B
U_C: {'enjoolras', 'isabelnevilles'}
u_c: {'isabelnevilles'}
Patterns of username creation behavior across OSNs
Similarity: 1
Feature: 1
Similarity: n
Feature: n
Patterns of username reuse behavior across OSNs
Similarity: 1
Feature: 1
Similarity: m
Feature: m
Labeled datasets
Supervised Classification

Figure 4.2: Architecture of the identity linking framework to compare username sets and capture similarities based on unique behavioral patterns while creating and reusing usernames over time.

sive 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.

4.2.1 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 [3]. 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 the 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 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 changing 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,Jaccard 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 [91]. 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,

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

4.2.2 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 [46]. 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 [10]. 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 4.1. 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 stresses 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) 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 essential 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 4.2 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:

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

Table 4.1: Few of the many leet symbols identified in usernames.

Leet symbol [10]	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 4.2: Examples of consistent username creation style of a user across OSNs.

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’

4.2.3 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 [112]. 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 [51], originally proposed to perform sequence alignment in protein sequences [93]. 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 feature capturing sequence alignment, and boolean feature estimating temporal synchronization.

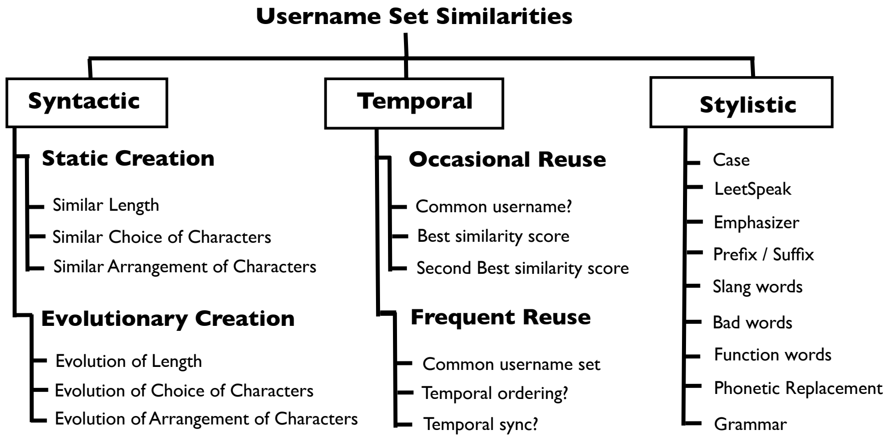


Figure 4.3: Syntactic, Stylistic and Temporal similarities captured between username sets corresponding to examined user profiles.

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.

4.3 Identity linking framework

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

4.3.1 Independent framework

Most profile linking approaches use a feature set, labelled datasets and a single classifier to predict link between test profiles [71, 74, 83, 112]. 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 4.4(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.

4.3.2 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 4.4(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.

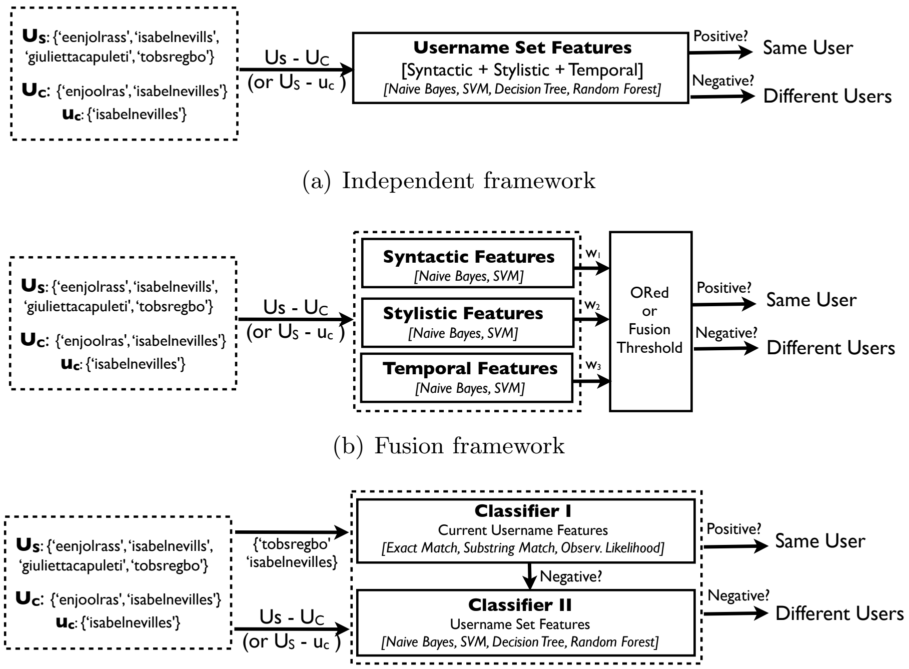
4.3.3 Cascaded framework

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 [49]. **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 4.4(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 [74, 112] 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.

4.4 Data Collection

For a positive dataset, we need to know accounts of a user on multiple OSNs. To start with, we choose a random set of 8.7 million users from Twitter. We followed a similar methodology to [73] where random Twitter *user_ids* are generated between 1 and 1,918,524,009 (the largest *user_id* authors observed). The random selection of Twitter users with a seeded random number generator avoids any bias towards specific set of Twitter users while training and evaluating our methodology.



(a) Independent framework

(b) Fusion framework

(c) Cascaded framework

Figure 4.4: 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.

We then build a tracking system on October 2013 and track any changes made to these user profiles till November 26, 2014. Tracking system repeatedly queries 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 compares consecutive API-responses to take a note of any changes to usernames, names, URLs, etc. 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.

Out of 8.7 million tracked users, we then select users for whom we can find their profiles on other three popular social networks – Facebook, Instagram and Tumblr, within our tracked dataset. The three networks are shown to contain qualitative and descriptive information about the user[1], and the choice of selecting them. All networks, except Facebook, allow multiple changes to username. Facebook allows username change only once.

Ground Truth: One way to find 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

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

hyperlinks to other OSN profiles via URL attribute of their Twitter profiles [55]. Such users *self-identify* themselves on other OSNs. For instance, a user posts *www.facebook.com/username* 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 [112].

Username History: Once user profiles are identified across OSNs, we collect past usernames owned by the user profiles. Using our independent tracking system for Twitter to monitor any changes to 8.7 million randomly chosen Twitter profiles, we gather the past usernames of the user on Twitter. 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.[2] 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 record the time of each username change made by a tracked user on social networks. Other methods to collect past usernames are discussed in Section 4.6.

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, totalling

[2] 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.

128,251 instances whose username sets are known to belong to a single user and hence are positive instances (see Table 4.3). 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 negative instances and use features in an engineered framework that effectively classifies username sets as same or different users.

Table 4.3: 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

4.5 Evaluation

We evaluate listed frameworks on two genres 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 4.4 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 30% 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 of the three frameworks. With Naive Bayes as a basic classifier, we observe that cascaded framework gives a slightly better accuracy and false negative rate than independent and fusion framework and 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%.

Table 4.4: 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%

We experiment with other supervised methods to learn the classifier, and achieve best accuracy with SVM (reduction by 48.47%) and 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 4.5 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.

Significant reductions in FNR of the framework imply that the username history helps in linking user profiles and is an necessary feature for profile linking methods. An example where baselines failed 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

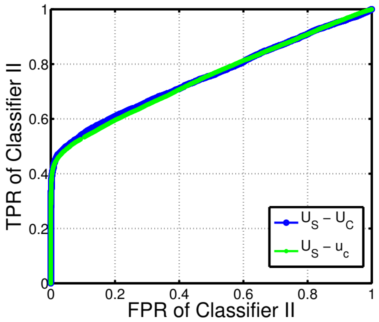
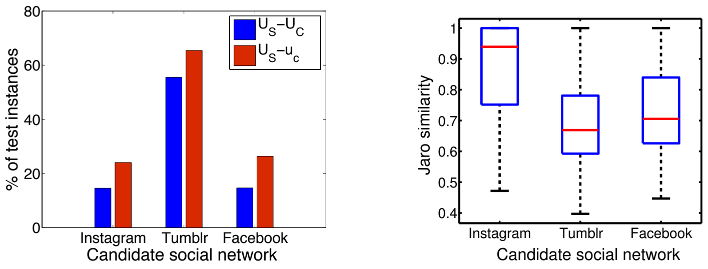


Figure 4.5: ROC curve of Classifier II trained with SVM with RBF kernel.

profiles is beneficial for profile linking. Using $U_S - U_C$ instances, we create another dataset where we consciously 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 4.6(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 can be attributed to the lowest Jaro similarity between most similar usernames from Tumblr and Twitter username sets (see Figure 4.6(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



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

Figure 4.6: .
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.

by squaring coefficients of features returned by **Classifier II** as suggested in [47]. Top-10 features, calculated between source and candidate username sets, are –

- 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 plays 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 4.6.

In summary, the key findings of this work 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) Availability of username history only on one profile increases false linkings by 12% as compared to its availability on both, iv) Success of the framework relies on the platforms to which examined profiles belong to; 95.84% misclassified Twitter-Tumblr username sets, while approx. 5.50% 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.

Comparison with prior research The state-of-the-art method **M**odeling **B**ehavior for **I**dentifying Users across **S**ites (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 [112]. To compare our methodology with MOBIUS, 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.

4.6 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 hint 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.

4.6.1 Applicability

Apart from observing users over time on OSNs, one can get user history archived by external services like DataSift[3] or Gnip[4]. We further suggest other two methods to collect past usernames – via timeline and 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 [32].

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

4.6.2 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 a 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 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

[3] http://datasift.com/platform/historics/

[4] https://gnip.com/products/historical/

cascaded framework with respect to the baseline **b1** on the derived datasets with varying set sizes is shown in Figure 4.7.

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

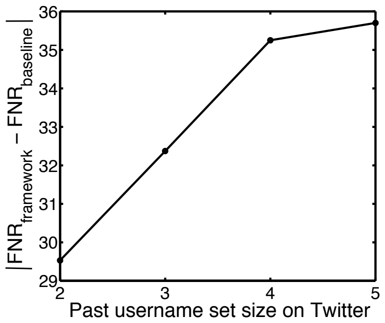


Figure 4.7: Higher False Negative rate (FNR) reduction with increasing source username set size.

4.6.3 Importance of evolutionary creation and temporal reuse behavior

List of important features (Page 65) 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 profile. Username reuse behavior can be only observed over a prolonged track of history, however, has proven useful feature for profile linking.

4.6.4 Implications to Privacy

We understand that tracking a user to gather the 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.

4.6.5 Extension to other attributes

Not just username, but other attributes evolve over time on Twitter. Evolution leads to the distinct set of values ever assigned to an attribute which can be used for comparison during profile linking. Figure 4.1(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 the 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.

The proposed identity linking framework is helpful to identify users who evolve their attributes over time. We next investigate on why and how username change occurs and what are the characteristics of such user behavior. We also investigate if username changing behavior can help tag users as benign or malicious, in the next chapter.

Chapter 5

Study of username changing behavior

Questions on *how*, *why* and *who* are these user accounts that frequently change their username on Twitter are, so far, unanswered. We believe that answering *how users create usernames over time* can aid in finding a user's account on other social networks. Username creation methods to create usernames over time can replicate and re-occur while creating usernames on her other accounts. Literature and our prior work devise profile linking methods that can link user accounts assuming usernames are created in the similar fashion within and across networks over time [56,112]. Answering *who are these users* and *why do they change usernames* can help us understand if the username changing behavior is a characteristic of a specific set of users. Finding reasons for username change can indicate if the intentions are benign and valid or fraudulent.

We make the first attempt to answer these questions and characterize username changing behavior on Twitter. We carefully create a dataset of 10K users, randomly sampled from 8.7 million users, and track them for a duration of 14 months every fifteen minutes. Our work with recorded past usernames of the users can help Twitter to effectively redirect user search queries rather than either serving with a dead link or different user. With an understanding of patterns and reasons of username change, Twitter can also develop tailored username suggestion algorithms for its benign users during the sign-up process and later.

5.1 Data Collection

We use the same data collected for linking user identities across OSNs, described in detail in Section 4.4. In summary, we query 10K users via Twitter API every 15 minutes. We term the faster scan of 10K users as *Fifteen-minute scan.* Fifteen-minute scan starts on November 22, 2013; we bookmark the scan till January 22, 2015 and use 14 months scan for our analysis.[1]

[1] We continue to scan 10K users after link-score 22 '15 and record any username change.

With fifteen-minute scan we have an advantage; we could record the exact time when user changed her username, with an error limit of 15 minutes. Further, we observe that while our regular fortnight scan took only one snapshot, fifteen-minute scan took 794 snapshots of 10K users, during November 22, 2013 - November 26, 2013. Regular fortnight scan missed 712 username changes triggered by 607 users, well captured by fifteen-minute scan. Therefore, fifteen-minute scan is successful in capturing most username changes made by the tracked users. All analysis here examines 10K users that switched usernames within 14 months on Twitter.

Table 5.1: Fortnight scan tracked over 8.7 million users every fortnight and missed many username changes due to long scan stretches of four to five days. We then initiated a fifteen-minute scan with a random sample of 10,000 users out of 10% of 8.7M users who changed their usernames at least once, and tracked them every fifteen minutes for more than a year.

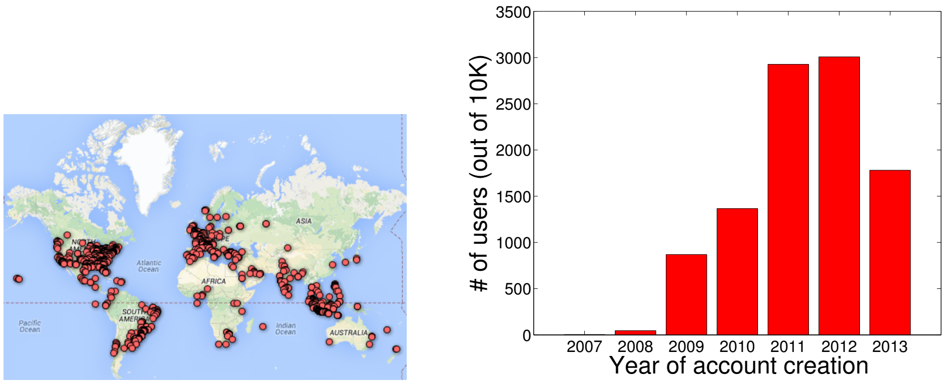
Name of scan	**Period of scan**	**# users**
Fortnight scan	October 16, 2013 - November 26, 2013	8,767,576
Fifteen-minute scan	November 22, 2013 - November 22, 2015	10,000

5.1.1 Representativeness

As mentioned earlier, it is necessary that users in 10K dataset span across diverse locations and different registration years on Twitter, to avoid any bias towards a special section of users. We examine geographical locations of 10K users to understand if they span across diverse locations. We use geo-tagged tweets by the users to record their location. We map 1,849 unique latitude, longitude pairs from where 926 users (9% of 10K) have posted their tweets (see Figure 5.1(a)). We observe that users in our dataset tweet from different locations around the world and not biased to only a few locations. Therefore, our analysis and results can be generalized to Twitter population from various global locations, who change their usernames over time. We further examine if these 10K users mainly contain newly registered users and hence they prefer to change their username to adjust to their requirements on Twitter. We note the year of account creation for these users and show the distribution in Figure 5.1(b). We observe that users in our dataset registered themselves on Twitter in years ranging from 2007 to 2013. Therefore, our dataset captures users from different levels of familiarity to Twitter.

5.2 Characterization

Before analyzing the characteristics of the entities involved in username changing process – usernames and users, we estimate the frequency of the behavior. Out of 10K users, 4,198 users changed



(a)

(b)

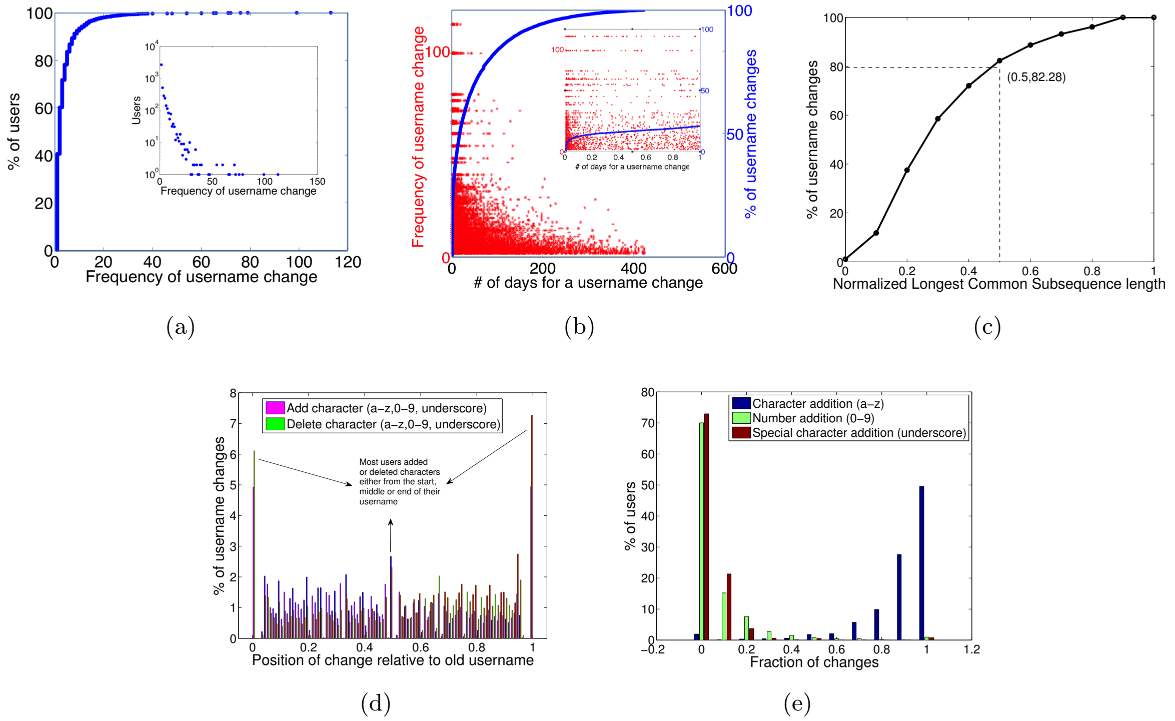
Figure 5.1: Tracked users tweeted from different geographical locations and registered on Twitter at different times.

their usernames at least once in 14 months, constituting 14,880 username changes. About 20% users changed five times or more triggering around 12,648 (85% of all) username changes (see Figure 5.2(a)). One user changed her username 113 times in 14 months which on manual inspection, turned out to be an inorganic user [20] with half completed tweets, tweets with same text, and frequent posts in short duration. We also examine the number of days after which users trigger the change (see Figure 5.2(b)). Around 20% of username changes were triggered within a day of the previous username change. We, therefore, observe a Pareto distribution with 20% users frequently changing usernames in short intervals and 80% users rarely changing after long durations (see inset figure in Figure 5.2(a)).

5.2.1 Usernames

An action of username change involves dumping an old username and creating a new one. Often, users favorite a username and repeatedly use that username. In our dataset, we find that around 35% of users reuse an old username later, while 65% never do so. For the 65% users, it is important to understand how users create their usernames over time. This information can be helpful in predicting their username creation patterns on other social networks as well, thus helping connecting multiple profiles of a single user. In order to understand username creation patterns, we filtered out reused usernames and considered unique usernames used by the users over time. We first investigate how the usernames differ from each other. To measure the similarity between two consecutive usernames used by a user, we use Longest Common Subsequence (LCS) similarity, a well-defined metric in literature.

LCS similarity estimates the sequence of characters that appear together without penalizing for insertions made. Figure 5.2(c) shows the cumulative distribution of username changes v/s LCS



(a) (b) (c)

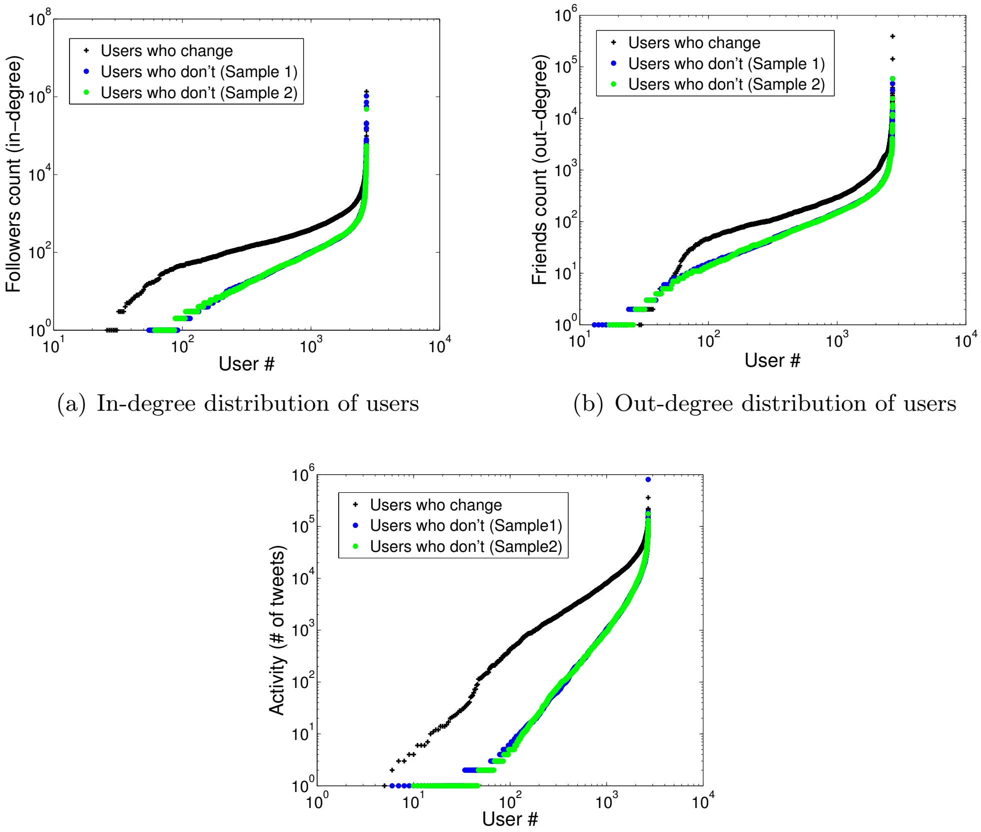
(d) (e)

Figure 5.2: Distribution of users and username changes among different username creation strategies. A section of users change their username multiple times within short intervals and choose un-related new usernames. Users are likely to alter old username by adding or deleting only characters at preferred positions.

matching length between the usernames (old and new) associated with the change. LCS matching length is normalized by length longer of the two, as suggested in literature [112]. For approximately 82% of username changes, new username is un-related to the old username (length ≤ 0.5), while for around less than 10% changes have new usernames highly similar and derived from the old ones (length ≥ 0.8). The observation indicates that majority of users select dissimilar usernames over the time within one social network, which is a complementary observation to literature which suggests that users create similar usernames across other social network sites [83, 112]. Note that such a user behavior may repeat as well across social networks and hence could challenge profile linking methods that use traditional string matching algorithms to match usernames in order to find connection between two user profiles.

We now examine the kind of change that users make to their usernames i.e. do they add special characters, numbers or alphabets, at what positions they make changes, and how they vary length changes. For the analysis, we selected usernames with LCS length ≥ 0.5. Figure 5.2 shows distribution of users and username changes among various username alteration methods. Changes to usernames are preferred either at the beginning, at the center or in the end of the username. Possibly, users add suffixes and prefixes to create a new username out of the old one. Users prefer to use characters rather than numbers or special characters every time they create a new username (100%

changes include only characters). We conclude that users exhibit certain preferences of username creation and alteration, which can be captured to modify the search with the outdated username and for profile linking across networks.



(a) In-degree distribution of users

(b) Out-degree distribution of users

(c) Activity distribution of users

Figure 5.3: Comparison between 4,198 users who changed their usernames with two random samples of 4,198 users who never changed their usernames. Users who change usernames demonstrate statistically significant superiority in terms of popularity and activity.

5.2.2 Users

We now explore the characteristics of users who opt to change their usernames. Does their popularity or activity or familiarity with network govern the frequency of username change? We answer these questions now.

Normal v/s Changing Users

On examining if users who undergo username changes are different from others who don't, we compare the activity and popularity of both genre of users. From 90% proportion of users from 8.7 million users that does not engage in this behavior, we extract an equal number of users as those who change usernames. A comparison of in-degree, out-degree and activity of 4,198 users with characteristics of two random samples of size 4,198 users,[2] are shown in Figure 5.3.

Both random samples do not differ in their properties but differ from the users who change their usernames over time. Users who change have higher popularity and activity as compared to users who do not change their username. Statistical significance of the difference is tested using Kolmogorov - Smirnov (KS) test. We make two pairwise comparisons of distributions and report test result for one of them. For in-degree, the value of KS D statistic is 0.3943 with *p-value* < 0.0001, for out-degree, the D value is 0.2654 with *p-value* < 0.0001 and for activity, D value is 0.4143 with *p-value* < 0.0001). We receive similar values for the pairwise comparison between second normal set and changing users. Higher values of D with a lower p-value for the comparisons signify that normal and changing users differ from each other in Twitter's *activity*, *popularity*, and *following* behavior.

Popularity v/s Frequency of Change

On Twitter, users tweet, reply or converse with their username. Changing usernames by a popular user may lead to confusion among her followers or may lead to loss of tweets in case someone else picks the username. For instance, the Indian Prime Minister Office's Twitter handle '*@PMOIndia*' was acquired by a teenager for 30 minutes during the transfer of social accounts from the earlier government [89]. In such a scenario, we speculate that users with higher number of followers avoid any username changes. We measure popularity of 4,198 users using followers (in-degree) and plot it against frequency of username change (see Figure 5.4(a)). To find correlation between the two, we remove users with too many followers ($>$ 1 million) or too less (< 1). We observe that username change frequency is weakly yet positively correlated to the in-degree of the user (*Pearson correlation: 0.1153, p-value* < 0.00001, α*: 0.05*). A significant positive correlation implies that higher the popularity, higher is the frequency of change, however, weak correlation does not guarantee the same.

Activity v/s Frequency of Change

An active user on Twitter, who engages herself in conversations and group chats, may change her username less frequently to avoid confusion during tagging / replying in a tweet. We conjecture

[2] In order to justify the generalizability of observations, we pick two random samples.

that active users change their usernames less frequently. We analyze 4,198 users and measure their activity with the number of created tweets. Figure 5.4(b) shows the frequency of username change with the user's activity. To find correlation between the two, we removed users with too many tweets (> 100K) or too few (< 1). We observe a weak and positive correlation between the two (*Pearson correlation: 0.1045, p-value < 0.0001, α: 0.05*). A positive yet weak correlation implies that users with high activity are inclined towards frequent username changes, however, activity does not guarantee frequency of change.

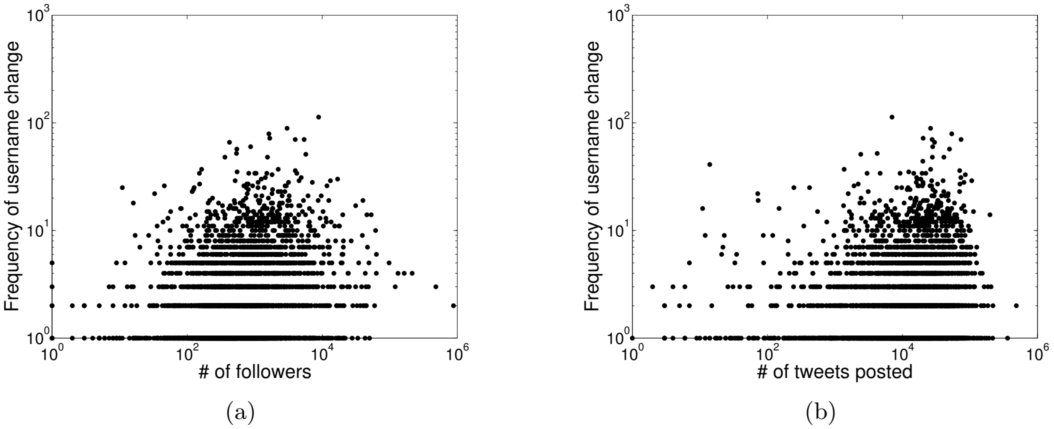


Figure 5.4: Frequency of username change v/s user popularity and activity. Weak correlations imply that popularity and activity has a little impact on the choice of changing username.

Familiarity v/s Frequency of Change

Intuitively, users who registered long time ago are familiar with Twitter and must have chosen stable and beneficial username for themselves than users who have registered recently and are still in exploratory stage. We examine if old user accounts engage themselves in username changing behavior, or only new users change their usernames multiple times. Figure 5.5 shows the frequency of username change with the age of the account for 4,198 users. We observe negative and very weak correlation between the age of the Twitter account and the frequency with which the account changes username (*Pearson correlation: -0.0942, p-value < 0.0001, α: 0.05*). Negative and weak correlation implies that both older and newer accounts engage in this behavior.

5.3 Plausible reasons

So far it in unclear on reasons that encourage users to change their usernames. Note that, users put in efforts to create a suitable username to converse with others on the network. A sudden change

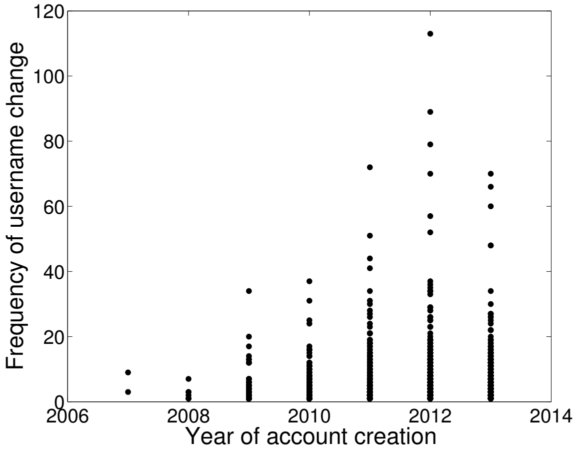


Figure 5.5: Frequency of username change v/s account age. Very weak implies that both old and new user accounts actively change their username.

to the username may direct users to a broken link or to a different user altogether who now owns the dumped username. We, therefore, explore a set of reasons for this change based on literature and observations using data analysis and talking informally to tracked users via tweets. On manual inspection, we classify listed reasons as benign and malicious.

5.3.1 Benign Reasons

Reasons for change that are motivated by user requirements and natural behavior are marked as benign. These are:

Space Gain

On Twitter, a user can converse with another user by tagging her *'@<username>'* in 140-character tweet. Since the tweet length is limited and maximum character limit for username is 15 characters, long usernames imply short message. We speculate that users with long old usernames may change to short new usernames to allow other users (followers) to post more content than before and benefit from space gain. This reason is motivated by the introduction of shortened URLs and RT symbol in Twitter to save space in a tweet [61]. We calculate the length difference between new and old username of users and separately represent users with old usernames less than and greater than the median length ($\geq$ 11). We observe that 75.19% of long usernames moved to short or same length new usernames, while 60.87% short usernames picked long new usernames (see Figure 5.6). In other words, most users with old usernames of length $<$ 11 tend to add characters in their new usernames, while most users with old usernames of length $\geq$ 11 prefer to remove characters for their new usernames. With this observation, we infer that creating shorter usernames is an incentive for users to change usernames.

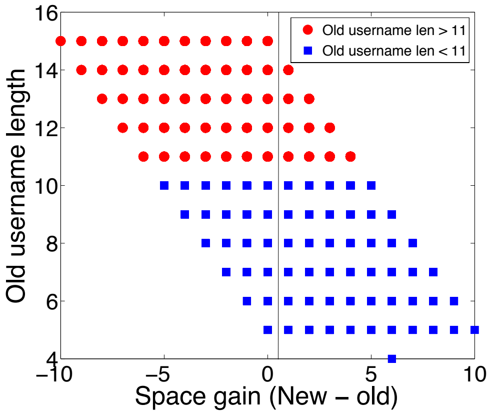
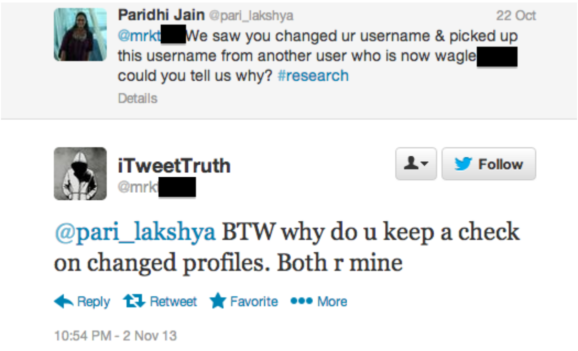


Figure 5.6: Length difference between new and old username v/s length of old username. Users with long usernames pick shorter new usernames (higher negative space gain), while users with short usernames pick longer new usernames.

Maintain Multiple Accounts

On Twitter, a user is allowed can create multiple accounts, each with a different email address.[3] On observation, we see that few changed username in order to exchange usernames among their multiple accounts (see Figure 5.7). We think that by tracing shared username's owners over time may help link multiple accounts of a single user within Twitter.



Paridhi Jain @pari_lakshya 22 Oct

@mrkt[redacted]We saw you changed ur username & picked up this username from another user who is now wagle[redacted] could you tell us why? #research

Details

iTweetTruth
@mrk[redacted]

@pari_lakshya BTW why do u keep a check on changed profiles. Both r mine

10:54 PM - 2 Nov 13

Figure 5.7: Shared username is exchanged among user's multiple accounts on Twitter.

[3] https://support.twitter.com/articles/20169956

Change Username Identifiability

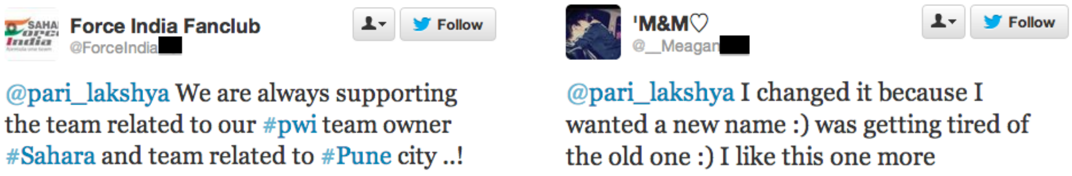
Few users in our dataset changed usernames to reverse the identifiability of the usernames i.e. either to make them personal or anonymous. For instance, a user named 'loried ligarreto' changed her username from 'loriedligarreto' to 'sienteteotravez' (translated: 'feel again') implying that user possibly intended to make her username anonymous. In other instances, we observe users who previously picked less identifiable usernames, made them personal later. For example, a user named 'rodrigo' changed her username from 'unosojosverdes' (translated: 'green eyes') to 'rodrigothomas_', thereby implicating that user probably wished to associate her real identity to her username.

Adjust to Events

Another user told us in a tweet that she represents Sahara India FanClub. She has supported Sahara's Pune Warriors team in IPL event with username 'pwifanclub' and then Sahara F1 team with username 'ForceIndia@!' and therefore has changed her username (see Figure 5.8(a)).

No specific reason

Few users responded that they changed their usernames without any specific reasons, that they got bored of the earlier one (see Figure 5.8(b)).



Force India Fanclub
@ForceIndia[redacted]

@pari_lakshya We are always supporting the team related to our #pwi team owner #Sahara and team related to #Pune city ..!

(a)

'M&M♡
@_Meagan[redacted]

@pari_lakshya I changed it because I wanted a new name :) was getting tired of the old one :) I like this one more

(b)

Figure 5.8: Username change due to change of events over time and boredom.

5.3.2 Malicious Reasons

Username changes that trick and misguide users violates Twitter usage policy and result from unnatural user behavior are marked as malicious. These are:

Obscured Username Promotion

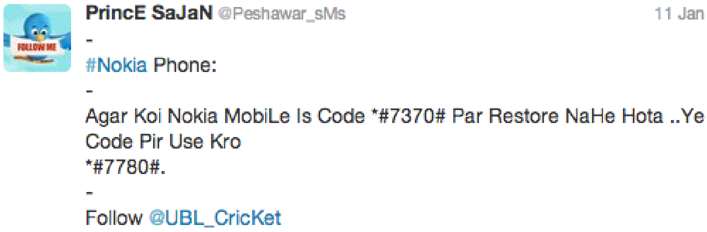
Owing to a limited number of users in fifteen-minute scan, we use fortnight scans of 8.7M users for this analysis. To our surprise, we find that a few user profiles collaboratively picked the same

Table 5.2: Rotational use of a shared username among users belonging to a group on Twitter. Numbers in the brackets represent follower count as recorded in the scan. We observe that users of a group (partner accounts) collaboratively share and promote the username via tweets / description to evenly distribute followers among themselves.

ID	**Scan - I**	**Scan - II**	**Scan - III**	**Group**	**Observation Date**
12xx62463x	**CollaGe_InFo** (61)	Dictionary_ID	Dictionary_ID	Sajan	2013-04-01
11xx79686x	DaiLy_GK (292)	**CollaGe_InFo** (2,372)	Geo_Account	Sajan	2013-10-02
95xx1822x	Geonewspak9	DictioNary_GK (1,279)	**CollaGe_InFo**	Sajan	2013-10-25
19xx56472x	-	-	**CollaGe_InFo**	Sajan	2013-12-04
60xx2762x	**Peshawar_sMs** (1,282)	MoBile_TricKes	BBC_PAK_NEWS	Sajan	2013-04-08
11xx37099x	Vip_Wife (180)	**Peshawar_sMs** (4,325)	UBL_Cricket	Sajan	2013-10-25
28xx1645x	NFS002cric	NaKaaM_LiFe (3,880)	**Peshawar_sMs** (4,044)	Sajan	2013-11-08
70xx9502x	**FuNNy_SardaR** (1,406)	MaST_DuLHaN (4,175)	KaiNaT_LipS	Khan	2013-04-01
99xx9356x	SaIrA_JoX (1,009)	**FuNNy_SardaR** (1,841)	MaST_DuLHaN (1,900)	Khan	2013-10-02
12xx73970x	-	-	**FuNNy_SardaR**	Khan	2013-12-04

username at different timestamps. Table 5.2 shows two such groups and the rotation of a username among the profiles, as observed in four scans. Usernames ‘Collage_InFo’, ‘Peshawar_sMs’ and ‘FuNNy_SardaR’ were used by different user IDs at different times. All these users claimed to belong to a group, either in their name or bio attribute. We term the username which is shared by multiple accounts as *shared username* and profiles who picked the shared username at different times as *partner accounts*. We observe 70 other shared usernames in our fortnight scans. We inspect the intentions for such a behavior in following ways.

We analyze tweets and description of the partner accounts mentioned in Table 5.2. We calculate the number of ‘@’ tags mentioned in their tweets and description. It was surprising to see that irrespective of the group, the partner accounts promoted a shared username by posting “Follow @<username>" in their tweets (or in description) multiple times (see Figure 5.9). Altogether for the two groups under observation, ten accounts promoted 30 other usernames. Seventeen percent (5 out of 30) usernames are promoted by more than one user. We think that by asking other Twitter users to follow a shared username and then keep exchanging the username with each other, the intention is to obscure the real identity of the user behind the free flowing shared username and distribute the followers evenly across the partner accounts.



PrincE SaJaN @Peshawar_sMs 11 Jan

-

#Nokia Phone:

-

Agar Koi Nokia MobiLe Is Code *#7370# Par Restore NaHe Hota ..Ye Code Pir Use Kro
*#7780#.

-

Follow @UBL_CricKet

Figure 5.9: Example post on Twitter where one partner account promotes another in her tweets.

We explain the username promotion methodology as: an account holds a shared username u_s, while other partner accounts promote the username by asking users to follow the account with username u_s. The account gains followers and decides to let her partner accounts gain further. She then releases her username to be picked by any of her partner accounts, and picks another (shared) username. She starts the promotion of the username u_s, along with other partner accounts. Her partner account, which picked u_s then gain followers. For the accounts mentioned in Table 5.2, we observe that a username is picked by the partner accounts with fewer followers. A similar modus operandi was observed when Recorded Future[4] analyzed Twitter accounts of a terrorist organization, Islamic State (IS). A single username was promoted by multiple ISIS-related accounts or followers either via bio or tweets, thereby tricking and gaining followers [42]. We suspect that the accounts listed in Table 5.2 engage in similar malicious activities.

[4] https://www.recordedfuture.com/

Username Squatting

On Twitter, there are four pools to which a username can belong to – *free-username pool*, *taken-username pool*,[5] *suspended / deactivated-username pool*[6] and *squatted-username pool.*[7] A username belongs to a free-username pool if no one else uses it. The username moves to taken-username pool if taken by an account. If the account gets deactivated or is suspended by Twitter, the username is blocked forever (for now), which thereby moves the username in suspended / deactivated-username pool. If an inactive user account keeps the username, in order to block or preserve that username, and not to allow others to use it, the username belongs to squatted-username pool.

Username squatting is against Twitter Rules.[8] Squatted usernames on OSNs have been investigated as a challenge in literature by researchers [29,88] to investigate cases of trademark infringement. We are curious to find if users change usernames in order to squat interesting usernames or usernames that represent an organization or an entity. Method to squatting here is to create profiles that either show no activity (i.e. no tweets) or have zero followers. For our fifteen-minute scan, we observe that for around 12% of 4,198 users, at least one of their vacated usernames are blocked by inactive Twitter profiles, either created by themselves or others. Without the access to emails, used to create user accounts, we have a little information to find if users themselves created the accounts to squat the usernames or others were opportunistic to find a free username and block the username with an inactive account. We think that future research can add these observations as features to find malign / phoney users on platforms like Twitter.

5.4 Discussion

This work aims at finding how and why users change their usernames within a social network like Twitter. Based on literature which suggests that when users *create* new usernames on different sites, they tend to create similar ones, we speculated that when users *change* their username within a network, they will pick new usernames similar to the old ones [74, 83, 112]. However, our analysis suggests the contrary. Most users created new username un-related to the old username when they changed username within a social network. We think that un-related usernames over time could be credited to the absence of cognitive load to remember a past dumped username [112]. When creating username across social networks, a user needs to remember all usernames, but when creating usernames within a network, she needs to remember only the latest one. Therefore, she has the liberty to choose it to be different from others. Un-relatedness between old and new usernames may challenge people on the network to derive new username from a user's old username and use

[5] https://support.twitter.com/groups/51-me/topics/205-account-settings/articles/14609-changing-your-username

[6] https://support.twitter.com/articles/15348-my-account-information-is-already-taken#deactivatedaccount

[7] https://support.twitter.com/articles/18370-username-squatting-policy

[8] https://support.twitter.com/articles/18311

Twitter search engine to find her.

Reasons for change on other platforms: We also inquired if similar reasons for change exist for other social networks that allow username change any number of times. Wikipedia is a moderated platform which allows changes to usernames. Every time a Wiki member wants to change username, she needs to request a moderator with her old username, wished new username and the reason for change. We collated 16,167 reasons from 15,288 Wiki members listed within six years i.e., from December 20, 2007 to December 20, 2013, publicly available here.[9] Reasons are described as free text, so we used grounded theory and classified the reasons in categories based on Wiki policies of username creation.[10] Figure 5.10 shows the categories and the distribution of reasons within each category. We observe that 22% users request a username change as their old username is not in accordance with Wikipedia's username policy, 30% users change to gain anonymity and avoid abuse, few for unified identity, adjust to spelling errors and capitalization, rest change for no specific reason. Few examples for username change are mentioned in Table 5.3. Study of username change on two networks, Twitter and Wikipedia, show that few users are concerned about their privacy, while others want to establish their unique identity across platforms they use. Few other reasons are platform specific e.g. username promotion on Twitter to gain followers while username change on Wikipedia to adjust to platform's policy.

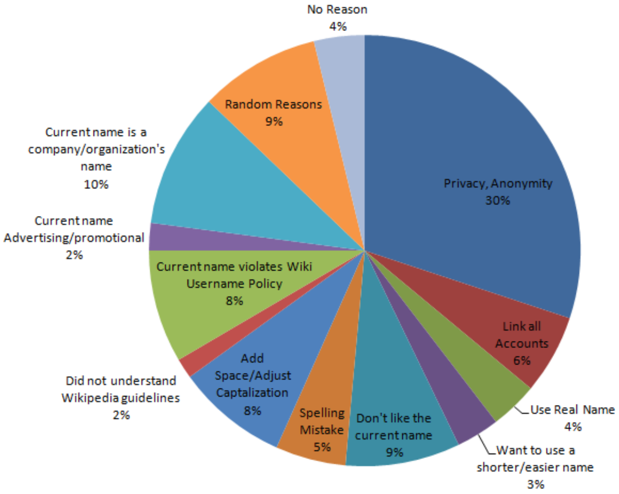


Figure 5.10: Reasons for username change mentioned by Wiki members as part of their request.

[9] https://en.wikipedia.org/wiki/Wikipedia:Changing_username

[10] https://en.wikipedia.org/wiki/Wikipedia:Username_policy

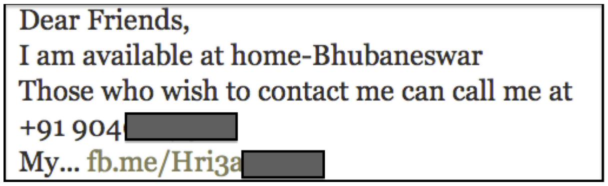
Reason Category	**Example**
Privacy	*"For privacy, since Sstrieu has my initials and part of my full name."*
Privacy and Abuse	*"For privacy. Ive attracted the attention of online bullies lately and theyve been trying to harass me anywhere they can find the name FazzMunkle."*
Link All Accounts	*"Consistency with other logins across the whole range of places where one can login, including some publicly accessible online profiles (e.g. Twitter). Having these consistent usernames allows my online identity to be consistent to. Plus, I keep forgetting my Wikipedia login info, having to guess quite a lot every time and having to request it be sent to me. Thanks in advance."*
Use Real Name	*"Changing my account from my nickname to my real name"*
Violates Wiki Policy [Promotional]	*"I have received a message from one of the administrators that my username is promotional/advertising for [[Roblox]] and he said either change your username or make a new account, so I am requesting to change my name to this."*
Violates Wiki Policy [Group Usage]	*"Current username represents organisation"*
Violates Wiki Policy [Religious Connotation]	*"My current name is apparently too ethnic for some editors, leading to inappropriate talk page speculation about my religion."*
Violates Wiki Policy [Bot]	*"Didnt read username policy, not allowed to have Bot in username."*
Violates Wiki Policy [Offensive]	*"Was told username may be offensive to some, and therefore a violation of username policy."*
Random Reason	*"My current name was used for something else I was using at the time of a low point in my life and Id like to move on from it. Every time I look at my user name I remember that day and how bad I felt and I dont want to be reminded of it just from my Wikipedia user name. Please let me change it."*

Table 5.3: Examples of few reasons for username change listed on Wikipedia.

Chapter 6

Study of mobile number sharing behavior

Phone (Mobile) number is an example of identifiable information with which a real-world individual can be associated uniquely, in most cases [114]. Though, public sharing of mobile numbers can help identity resolution, the associated individual can become an easy target for SMS and phone-based phishing scams.[1] Mobile numbers are observed to be shared either via profile attributes [16] or via posts (see Figure 6.1). Auxiliary details of mobile number owners shared along with the mobile numbers, or collected otherwise, can help attackers to launch targeted attacks against them. To examine the necessity of safeguard methods to prevent public exposure of users' mobile numbers either via profile or posts, there is a need to comprehend mobile number sharing behavior on OSNs, and the gravity of associated risks.



Dear Friends,
I am available at home-Bhubaneswar
Those who wish to contact me can call me at
+91 904[redacted]
My... fb.me/Hri3a[redacted]

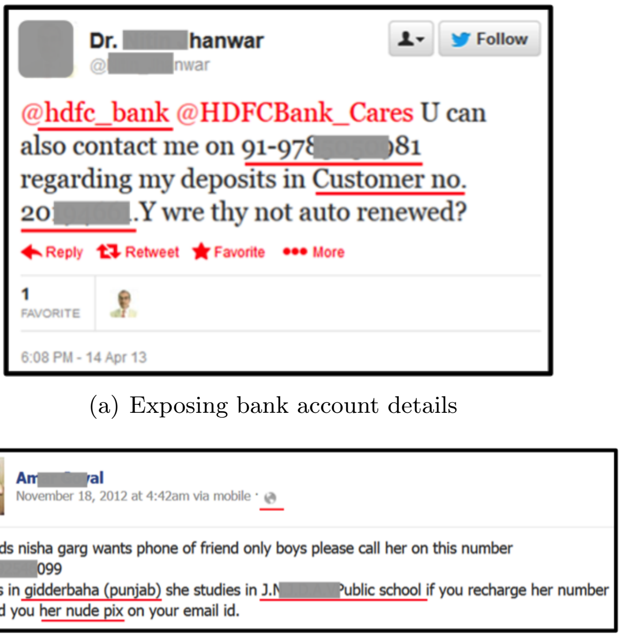
Figure 6.1: Example of exposed mobile number on Twitter along with auxiliary information such as location and Facebook account of the user to whom it belongs.

India has been a popular venue for mobile and phone frauds owing to the huge telecom industry. India has the second largest mobile network in the world, with 919.17 million subscribers by February, 2013.[2] We, therefore, focus on exposure of Indian mobile numbers in this study. We explore reasons, modes and whereabouts of Indian mobile numbers shared on two most popular OSNs – Facebook and Twitter. Further, an Indian mobile number can be used to reveal critical information about its owner such as name, age, location, which may invite targeted identity attacks (see Figure 6.2). We communicate the risks of sharing mobile numbers online to their owners by calling them on their

[1] http://www.scmagazine.com/fbi-warns-of-sms-and-phone-based-phishing-scams/article/191565/

[2] http://www.trai.gov.in/

numbers and note their reactions. We now describe our study and the findings of the study.



Dr. [redacted]hanwar
@[redacted]nwar

@hdfc_bank @HDFCBank_Cares U can also contact me on 91-978[redacted]81 regarding my deposits in Customer no. 20[redacted].Y wre thy not auto renewed?

6:08 PM - 14 Apr 13

(a) Exposing bank account details

Am[redacted]al
November 18, 2012 at 4:42am via mobile ·

Hii friends nisha garg wants phone of friend only boys please call her on this number
+91-[redacted]099
she lives in gidderbaha (punjab) she studies in J.N[redacted]Public school if you recharge her number
she send you her nude pix on your email id.

(b) Cyber-harrassment of a girl using her mobile number

Figure 6.2: Sample posts on OSNs mentioning a user's mobile number and personal sensitive information.

6.1 Data Collection

We deployed a three stage data collection methodology – keyword selection, data collection and data validation (see Figure 6.3). We collected Indian mobile numbers shared on two popular OSNs – Facebook and Twitter. We use snowball sampling strategy based on a seed set of keywords to collect posts that share mobile number rather than collect a random set of posts. The choice is made to ensure that the datasets capture the missed posts that mentions a phone number but not a known keyword but also does not include irrelevant posts in the dataset.

6.1.1 Keyword selection

To collect public relevant posts and tweets with a mobile number, we need to select a set of relevant keywords [75]. To create the keyword list, we surveyed OSN users at IIIT-Delhi to determine possible words they would use while sharing a mobile number on OSNs. We selected most commonly listed words for our initial set of 50 keywords, such as *mobile number*, *contact us*, *call me*. With the initial set of keywords, we collected 1,525 public tweets using Twitter Streaming API[3] and 1,000

[3] https://dev.twitter.com

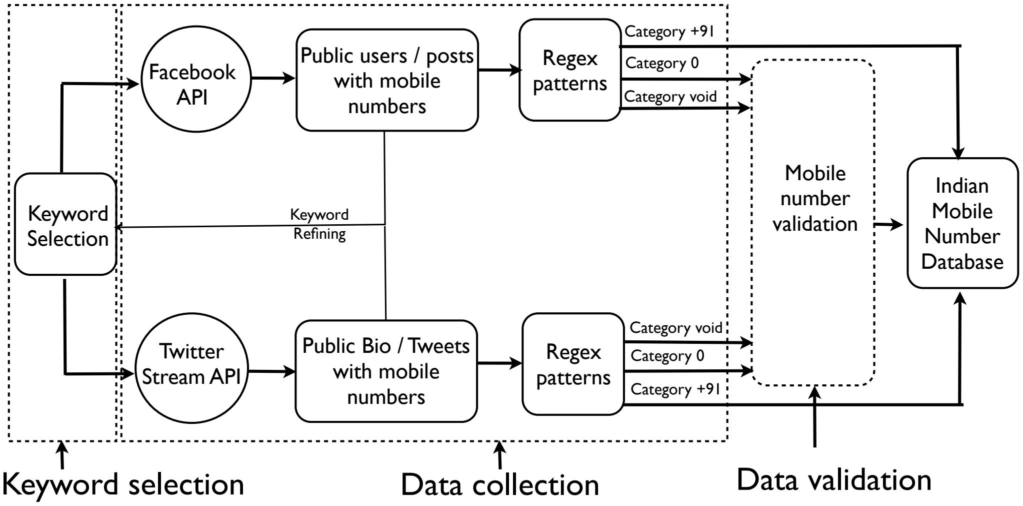


Figure 6.3: Data collection methodology to gather public profiles and posts which shared Indian mobile numbers on OSNs.

public posts using Facebook Graph API.[4] We used the collected posts to identify other common keywords when mobile numbers were shared (adapting a standard technique of query expansion from Information Retrieval [109]). We tokenized the posts, removed stop words and added most frequent words to expand the seed keyword set size to 278. A similar approach was used by Mao *et al.* to gather tweets with required contexts [75].

6.1.2 Mobile number data collection

We used the final set of keywords to collect public English posts and bio[5] which shared mobile numbers, using Twitter Streaming API and Facebook Graph API. We started our data collection from Facebook on November 16, 2012, and ended on April 20, 2013, while from Twitter on October 12, 2012, and ended on April 20, 2013. We stored public bio and posts which shared mobile numbers on OSNs, along with profiles of the users who shared the number either via bio or public post. We stored user bio, and public posts leaking mobile numbers as well as profiles of users sharing those public posts, in a MySQL database.

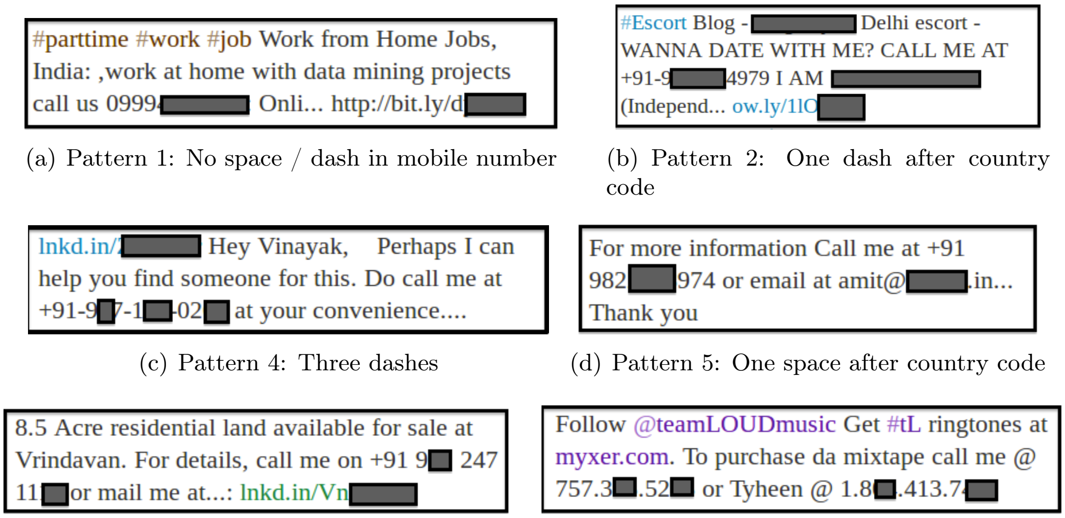
To tag Indian mobile numbers in users' posts and users' bio, we exploited the standard convention and structure of an Indian mobile number. It is a 10 digit number, where the first digit should start with either 9 or 8 or 7. It can be prefixed with a country code (+91) or trunk code (0).[6] We used rule-based named entity recognition [80] and created a set of regular expression rules which captured Indian mobile number structure to filter out Indian mobile numbers from posts and bio of

[4] https://developers.facebook.com/docs/reference/api

[5] referred to as "description" in Twitter API

[6] http://www.dot.gov.in/numbering_plan/nnp2003.pdf

users. We further observed that most users post Indian mobile numbers in different patterns (see Figure 6.4).



#parttime #work #job Work from Home Jobs, India: ,work at home with data mining projects call us 0999[redacted] Onli... http://bit.ly/d[redacted]

(a) Pattern 1: No space / dash in mobile number

#Escort Blog - [redacted] Delhi escort - WANNA DATE WITH ME? CALL ME AT +91-9[redacted]4979 I AM [redacted] (Independ... ow.ly/1lO[redacted]

(b) Pattern 2: One dash after country code

lnkd.in/[redacted] Hey Vinayak, Perhaps I can help you find someone for this. Do call me at +91-9[redacted]7-1[redacted]-02[redacted] at your convenience....

(c) Pattern 4: Three dashes

For more information Call me at +91 982[redacted]974 or email at amit@[redacted].in... Thank you

(d) Pattern 5: One space after country code

8.5 Acre residential land available for sale at Vrindavan. For details, call me on +91 9[redacted] 247 11[redacted]or mail me at...: lnkd.in/Vn[redacted]

(e) Pattern 6: Two spaces

Follow @teamLOUDmusic Get #tL ringtones at myxer.com. To purchase da mixtape call me @ 757.3[redacted].52[redacted] or Tyheen @ 1.8[redacted].413.7[redacted]

(f) Pattern 7: Dots between numbers

Figure 6.4: Various formats and patterns in which users posted Indian mobile numbers on OSNs. Mobile numbers were prefixed with either trunk code '0' and country '91' while others had no prefix.

Some of the sample patterns are, numbers with no space/dash in mobile number (0999xxxxxxx), one dash after country code (+91-9xxxxx4979), two dashes (+91-99xx-79xxxx), three dashes (+91-9x7-1xx-02xx), one space after country code (+91 982xxxx974), two spaces (+91 9x 24711xxx), and dots between digits (757.3x.52xxx). We modified our regular expressions to capture all possible ways of posting an Indian mobile number on social networks. We categorized Indian numbers prefixed with +91 as "Category +91" numbers (+91-9x7-1xx-02xx), prefixed with 0 as "Category 0" (09x71xx02xx), and prefixed with nothing "Category void" (9x7-1xx-02xx). Table 6.1 shows the count of mobile numbers collected from tweets or bio on Twitter and public posts or names on Facebook.[7]

6.1.3 Data validation

Rule-based named entity recognition used to extract Indian mobile numbers from public posts and bio in the earlier stage, relied on a set of regular expressions and, therefore, misinterpreted certain other country numbers as Indian mobile numbers. For instance, Figure 6.5 shows an example of a tweet where a user's card number is in a similar format as of an Indian mobile number.

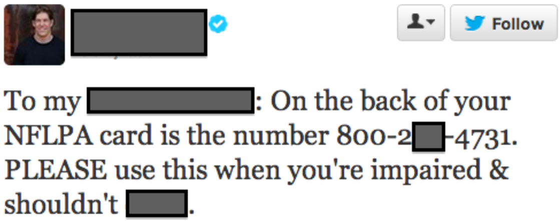
[7] Description of a user on Facebook is not publicly accessible

Mobile number format for few countries (The United Kingdom,[8] and USA[9]) is similar to that of an Indian mobile number. Mobile numbers from the UK are also 10-digit numbers starting with 07, which were confused as Indian mobile numbers prefixed with 0 and starting with 7. Mobile numbers from the USA also follow 10-digit format with first three digits representing area code, ranging from 2-9, therefore, the USA mobile numbers without country code and with area codes starting with 7, 8, 9 are similar to an Indian mobile number.

To avoid any noise in our database, we ran a validation check for the Category 0 and Category void numbers. Category +91 numbers were confirmed to belong to India as they were prefixed with Indian country code. We used a service[10] which checked if a number's first four digits belonged to a valid Indian mobile number series. However, the service was not updated. We observed that 19,934 mobile numbers out of 23,405 in Category 0 (85%), and 42,360 numbers out of 49,946 in Category void (85%), were confirmed to be Indian numbers by the service. After manual verification, we observed some non-Indian numbers were marked as Indian numbers by the service. We, therefore, considered *only* Category +91 mobile numbers for our analysis, which were confirmed to be Indian mobile numbers. Our intent was to avoid any bias or noisy inferences by including Category 0 and Category void numbers.

Table 6.1: Descriptive statistics of the mobile numbers collected from Twitter and Facebook.

Numbers	**Category +91**		**Category 0**		**Category void**		**Total**	
	Twitter	Facebook	Twitter	Facebook	Twitter	Facebook	Twitter	Facebook
Mobile numbers	885	2,191	14,909	8,873	25,566	25,294	41,360	36,358
User profiles	1,074	2,663	17,913	9,028	31,149	25,406	49,817	36,588



[redacted]

To my [redacted]: On the back of your NFLPA card is the number 800-2[redacted]-4731. PLEASE use this when you're impaired & shouldn't [redacted].

Figure 6.5: Example of a card number with a similar regex pattern as an Indian mobile number. Validation phase aim to remove all such numbers from the dataset.

[8] http://stakeholders.ofcom.org.uk/binaries/telecoms/numbering/numbering-plan201212.pdf

[9] http://nanpa.com/enas/npaDialingPlansReport.do

[10] http://trackmobileonline.co.in

6.2 Behavioral Analysis

6.2.1 Context analysis

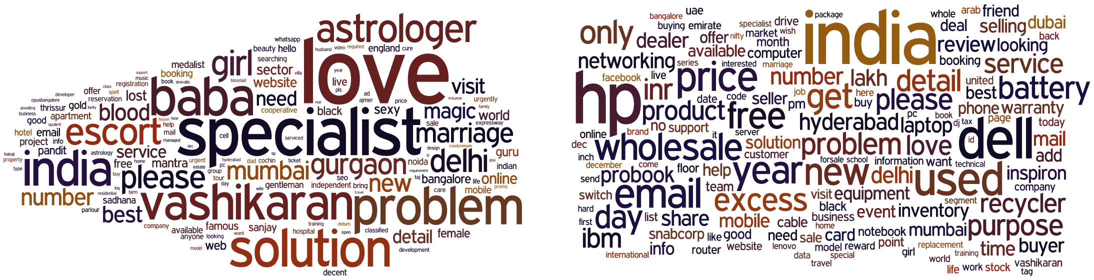
To understand the context, we extracted most frequent words from the bio, descriptions, and collected posts which shared the number. We removed stop words and performed stemming [87] to avoid repeated forms of the same root word.

We manually analyze word-clouds of the most frequent words (see Figure 6.6(a) and 6.6(b)). We observe words such as *blood, specialist, hospital, love, sexy, escort, girl, music, movie, fun, offer, reservation, ticket, hotel, seo, sale, astrologer, business* in Figure 6.6(a). We infer that on Twitter, users post Indian mobile numbers, majorly to ask for blood donations / aid, help in emergency situations, to promote escort business, to promote entertainment, to market for travel, holiday, hotel packages, and to buy / sell products, etc. Such a behavior is understandable since Twitter is used as a news media, and marketing platform [64]. On Facebook, users post Indian mobile numbers majorly in the context of Information Technology (IT) facilities and education related products, evident by the presence of words such as *price, hp, battery, dell, laptop, ibm, email, notebook, computer* (see Figure 6.6(b)). We infer that users post mobile numbers on social media platforms in order to benefit from social network structure and promote their business by spreading the contact information (mobile number) to a We, therefore,large number of users.

6.2.2 Ownership analysis

Exposure of mobile numbers by non-owners might lead to unwanted privacy leaks and annoyance to their owners.[11] however, analyze if owners of the mobile numbers themselves leaked their numbers at the first place or other users posted them. For each mobile number collected from Twitter (885)

[11] http://thenextweb.com/media/2011/07/10/supposed-phone-number-of-news-internationals-chief-executive-leaked-on-twitter



(a) Twitter Tag-cloud

(b) Facebook Tag-cloud

Figure 6.6: Extracted contexts in which users shared mobile numbers on Twitter and Facebook.

and Facebook (2,191), we retrieved the first tweet (post) in our dataset sharing that mobile number on Twitter (or Facebook). The mobile number was marked as 'leaked by its owner', if the tweet (post) included a first person pronoun such as *me, my, us, mera (my in English)* along with most frequent action verbs such as *call, text, sms, ping, whatsapp, message, contact*. For instance we check for the presence of phrases like - *"call us", "text us"*. The mobile number was marked as 'leaked by a non-owner', if the tweet (post) included second person pronoun such as *you, your, yours* or third person pronoun such as *his, her, them* along with same action verbs used with first person pronouns. Researchers used only pronouns to check for ownership [75], this may give false positives like - *"You may call me at xxx"*, however, we avoid it by using phrases here. Figure 6.7 details the procedure to determine if the number is leaked by owner. We also assume that mobile numbers shared on Twitter via bio or on Facebook via name are users' mobile numbers.

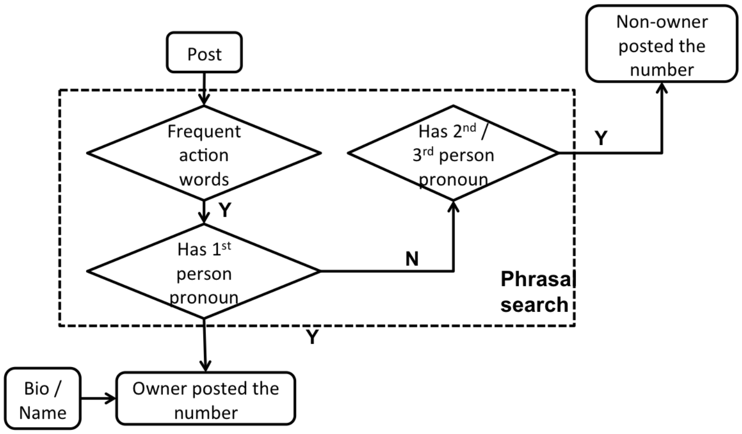


Figure 6.7: Procedure to determine if the mobile number is posted by the owner herself.

Table 6.2 shows the descriptive statistics of mobile numbers which were leaked by their owners and non-owners. Two hundred and ninety-one mobile numbers (32.8%) were shared by their owners, while only 18 mobile numbers (2.0%) were shared by non-owners on Twitter. Four hundred and eighty-five mobile numbers (22%) were shared by owners, and 25 mobile numbers (1.1%) were shared by non-owners on Facebook. Example post where owner shared his mobile number is *"F1 INR 2500/- tickets are available with me..!! Limited stocks..!! Ping me or call me up on +91 989 xxx xxxx asap!*" Example post where non-owner shared the mobile number is *"@VodafoneIN My friend Debasrita took a new connection (+91-73816xxxxx), she is having issues. Please contact her at +91-9556xxxxxx"*. For remaining mobile numbers, the methodology used could not infer if the numbers were shared by the owners or non-owners. Example post is "*Need a male punjabi artist of age 35 for a ad in #chennai pls contact +91 98-41-xxxxxx*".

Table 6.2: Mobile numbers shared by owners and non-owners on Twitter and Facebook. Most mobile numbers were leaked by owners themselves; though few were leaked by non-owners.

Social Network	**Mechanism**	**Mobile numbers**
Twitter - Owner	Bio Tweet	155 136
Twitter - Non-owner	Tweet	18
Facebook - Owner	Post Name	468 17
Facebook - Non-owner	Message	25

6.2.3 Source analysis

We inquire the source or application by which most mobile numbers were posted on OSNs. To extract application used to post the number, we extracted 'source' attribute of the tweet, available from Twitter API,[12] and 'application' attribute of the post, available from Facebook Graph API.[13] On Twitter, apart from the web (234), mobile numbers were largely posted from social aggregators and other social networks such as Facebook (148), Twitterfeed (121), Google (121), LinkedIn (50), TweetDeck (22). We observe the major use of social aggregators and other social networks to post mobile numbers on Twitter. Users might be sharing same mobile number not only on one OSN but multiple OSNs simultaneously. On Facebook, most numbers were posted by Facebook mobile applications such as Facebook mobile (125), Facebook for iPhone (36 numbers), Photos (34), Facebook for Android (19), and few by social aggregators such as HootSuite (31), and Twitterfeed (3). We observe the major use of OS based Facebook mobile applications to post numbers on Facebook with comparatively less exploitation of social aggregators.

6.3 Collating Auxillary Information

We now turn our focus to understand how publicly shared mobile numbers can be exploited to gather critical and sensitive information about the owners. We used two online services – Truecaller[14] and OCEAN.[15] Truecaller allows to query a mobile number and returns the name of the owner as well as the network operator. OCEAN allows to query a name of a person and returns matching entries from publicly available e-government data sources, listing Voter ID, family details, age, home address, and father's name. OCEAN has data only for Delhi citizens.

[12] dev.twitter.com/docs/platform-objects/tweets

[13] developers.facebook.com/docs/reference/api/post/

[14] http://truecaller.com

[15] http://precog.iiitd.edu.in/research/ocean/OCEAN.pdf

We got manual annotators to extract data from Truecaller and OCEAN for Category +91 mobile numbers. For each number, they were asked to observe name of the owner, her location, and mobile number operator from Truecaller,[16] along with the name of the owner, and her location from public posts and profiles on OSNs, sharing the same number. Possible names of the mobile number owner and her possible locations were inferred for 2,997 Category +91 numbers. Name of the owners whose inferred location was *Delhi* were then used to query OCEAN and matching set of citizens were recorded. Surprisingly, out of annotated 94 Delhi mobile numbers, we were able to identify uniquely eight users with details like name, age, father's name, home location, gender, and voter ID (see Table 6.3). We identified a professional Indian singer[17] as he posted his number on Facebook, and the number revealed other sensitive information.

Aggregation of information extracted from OSNs with the otherwise collected information about a Delhi mobile number owner may lead to a convenient identity theft.[18]

Table 6.3: Anonymized mobile number, name, age, gender, father's name, address, Voter ID of Delhi residents who shared their mobile number on OSNs.

Number	**Details**	**Shared by owner?**
9873xxxxxx	X Kakrania, 24, Male, X Kakrania, "B-***, B-block, X Vihar Ph-I, Delhi", WHC17xxx63	Yes
+9199xxxx2708	X Gambhir, 23, Male, X X Gambhir, "***, xxxx Bagh, Delhi", NLNxxx5696	No
8447xxxxxx	X Singh Nagi, 33, Male, X Singh Nagi, "D-**-b, Block- D, X Vihar, X Ext., Nangloi",IPN13xxx17	Yes
+9198xxxx5485	X X Jeswani Pankaj, 53, Male, X X Jeswani, "***, Mig Flats, *-block, xxxxx Vihar Phase-", DL/04/xxx/222668	Yes

We also experiment with an Android application, Whatsapp,[19] to understand if we can add leaked mobile numbers and hence abuse the Address Book Matching feature of the application and get access to their status messages [19]. We add leaked mobile numbers to a phone's contact directory and run Whatsapp application from the phone. Users leak variety of sensitive information via their Whatsapp status updates such as travel plans, social network profile, BBM Pins. Few examples of status updates are *"100% Single"*; *"No longer in India. UK: # +44 75xx 81xxxx US#610xx xxxxx as of June 10"*; *"www.facebook.com/iakrfilms"*; *"New BBM Pin: 25C7xxxx"*. We infer that an accidental / unintentional leak of the mobile number on OSNs is capable of exposing other sensitive information and thus creating a larger user's digital footprint.

[16] As per Truecaller policy, we did not store content.

[17] http://www.pankajjeswanimusic.com/home.html

[18] http://www.hindustantimes.com/India-news/Gurgaon/Identity-theft-cases-on-the-rise/Article1-931638.aspx

[19] http://www.whatsapp.com/

6.4 Implications to Privacy: Communication Strategy and Reaction

With risks of exposing user details using a leaked mobile number online, we attempt to communicate the observed risks to mobile number owners. Researchers have suggested various channels for risk communication, e.g., Short Message Service (SMS) [60], and Interactive Voice Response (IVR) system,[20] to communicate awareness information to its users. Online bloggers have also deployed automated tools to display partially obfuscated mobile numbers onto a public web page[21] and SMS with random texts, to publicly shared mobile numbers.[22] We deployed an IVR system and communicated the risks associated with posting mobile number online by calling the owners of the numbers. We choose IVR to ensure the reach to the owners and to convince the credibility of the message to them. We now discuss the IVR deployment details, calling procedure and users' reactions to the calls.

6.4.1 IVR system design and implementation

We set up an IVR system using FreeSWITCH[23] and a Java application (see Figure 6.8). We called 2,492 mobile numbers from Category +91 collected from earlier mentioned methodology until February 28, 2013. In India, we are not required to go through an Institutional Review Board (IRB)-type approval process before calling the users. However, one of the authors of this paper had previously been involved in studies with U.S. IRB approvals, and we apply similar practices in this work. Before the actual risk communication part of the message, we inform the user that an audio recording of the call will be taken only for research purposes. Furthermore, participants are given options to disconnect the call and request the deletion of the audio recording, at any given point of time during the call.

When a callee answered the call, for credibility purposes, we introduce ourselves as researchers from New Delhi. We then play the risk communicating message - “We found your number on X”, where X was either “Facebook” or “Twitter” or “Facebook and Twitter”, depending on the source from where we extracted the number of the callee. We then prompted a voice message “Posting your number online is not a good practice. Doing so will make you fall prey to various phone number frauds. Keep yourself safe and consider removing your number from the Internet.” We intentionally keep the language simple as English is not a native language of India and we have minimal information about the expertise level of the callee. We then present callee with the following options: “Press 1, If you did not know that your number can be leaked, and now you will remove it from the Internet;

[20] http://www.ddm.gov.bd/ivr.php

[21] http://www.weknowwhatyouredoing.com/

[22] http://textastrophe.com/

[23] https://wiki.freeswitch.org/wiki/IVR

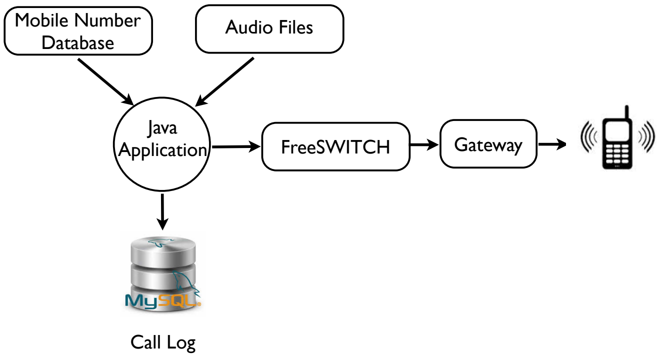


Figure 6.8: IVR System Design implemented using FreeSWITCH and Java application.

Press 2, If you posted it purposefully, and you will not remove it from the Internet; Press 3, if you want to hear the message again." If the user pressed either 1 or 2, we request him to leave us a feedback and later give him the option to end the call. We inform the user that we record the call for research purposes and log all responses and activities of callees in a database. We make the calls during weekdays from 1100hrs IST to 1600hrs IST.

6.4.2 User reactions

Figure 6.9 shows how callees collectively reacted at each stage during the call. Sixty-one percent of callees who picked the call opted to listen to the message and six percent chose to remove their mobile numbers from OSNs. An equivalent percentage (6.2%) chose *not* to remove their numbers. Forty-seven users from the 2,492 numbers that we called left feedback on our IVR system. A few are: *"Thank you for information, I have deleted, I will not post my number online"*, *"I want to know how to remove my number and I don't know, I haven't put my number purposely but if it is there, where exactly it is there I would also like to know that. Please get in touch with me asap. Thank you"*. Some callers showed their concerns and some even requested us for help to remove their numbers from the Internet. Such user reactions urge the necessity for a safeguard solution to control the spread of personal and sensitive information on OSNs. We also receive feedback saying *"I posted my number purposely for my website promotion, I usually do deal in web hosting business so that is why I want someone to contact me for hosting services"* implying intentional sharing of mobile numbers.

In summary, we examine the sharing behavior of Indian mobile numbers on OSNs via profile and public posts and investigate its association with other details of the user. We analyze Indian mobile numbers, shared on Twitter via tweets or bio and on Facebook via public posts or names. Most mobile numbers are shared to ask for blood help, to market astrology business, IT facilities, and escort services. We observe few posts where numbers are shared in personal contexts like *"My contact no in India is +91-9958xxxxxx"*, however, posts used for personal contexts had few context-specific

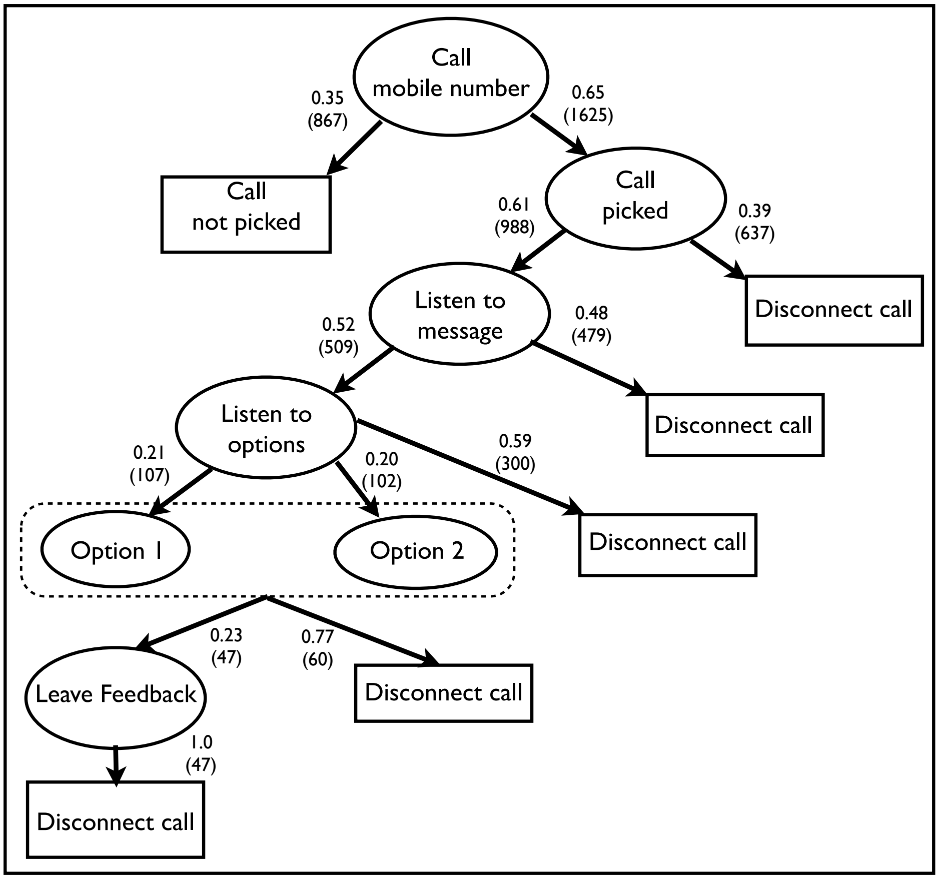


Figure 6.9: Callee Decision Tree. Each stage in the call is associated with a probability and the number of users who chose that stage.

keywords. Therefore, personal contexts are difficult to highlight. Users exploit social aggregators to popularize the same number on multiple OSNs. Sensitive and identifiable information such as Voter ID could be extracted with the use of mobile number and other information sources. There is a need to let users know that their unique personal identifiable information is shared online for public use, which may include spammers and scammers. We attempt to communicate the risk to users. As a result, few remove their numbers, while few justify that the numbers bring business to them, thus avoiding any future implications.

6.5 Discussion

Sharing mobile number online can help in identity resolution and building the holistic and comprehensive profile of a real-world user. An aggregated profile of a user is beneficial for marketers and business that intent to server their potential customers better. However, risks and implications of sharing identifiable information online are scary. Not just mobile number, our research finds other identifiable information on OSNs such as Blackberry Messenger Pins (BBM), and email addresses, that can also help in accurate identification of their owners. With little awareness on the risks, OSNs need to provide safeguard mechanisms to disallow (warn) exposure of sensitive and identifiable information via either profile or public posts. There is a need to build technological, people and process-oriented solutions to forewarn users and raise the awareness towards risks of PII leaks, so that users can make better decisions.

Chapter 7

Conclusions and Future work

Social data of users has been helping industry in exciting ways to enrich user experience and services. An important premise of many new services and research is that it is possible to search and link the different accounts of a user. The main contribution of my thesis is the development and analysis of automated identity resolution methods, both for searching and linking user accounts that correspond to the same individual in popular social networks. Matching accounts across OSNs allows marketers, enterprises, business and security professionals to work on comprehensive user profiles. This may, however, raise privacy concerns, in particular when we can link the accounts of users who deliberately keep the information disparate across their profiles to maintain separate personas. However, the resolution process can aware the users in turn to help identify leaks and cover them to maintain their privacy online. Also, we believe that malicious users can be effectively tagged using our methods.

7.1 Summary of Contributions

This thesis makes following contributions:

- **Observe and highlight user behavior across OSNs**: We highlight that users engage in redundant behavior across their identities on different OSNs – *Self-Identification*, *Self-Mention*, *Self-sensitive sharing*. Self-identification implies that users explicitly mention their identities on other OSNs or on webpage using hyperlinks. Self-mention refers to indirect exposure of their identities on other OSNs via hyperlinks embedded in their posts. Not only their accounts' information, users are observed to post sensitive and identifiable information about themselves across OSNs like mobile number, BBPins, credit card numbers, etc. Redundant information posted by users on their unlinked identities of different OSNs help linking the identities.
- **A study of mobile number sharing behavior**: With an in-depth study of sharing sensitive

information about a user like mobile number, we provide methods, reasons, patterns and risks associated with such user behavior. Scoping to verified Indian mobile numbers, we observe that mobile numbers are shared in emergency, relocation, business and escort calling / services. These numbers are shared either via the description of the user profile or the posts made by the user. Various patterns exist to share numbers which often confuse regex expressions with valid credit card numbers. Augmenting a mobile number with auxiliary information sources help us find name, age, family details, VoterID, location, and other sensitive details about the owner. A framework to communicate these risks to mobile number owners show that most users are either unaware of its online existence or possible threats to privacy. We emphasize that though mobile numbers can help business in collecting comprehensive user profile, it is important to make sure that users are aware of the risks associated with such sharing.

- **Methods for identity search that exploit public attributes and user behavior across OSNs**: We assume that observed user behaviors can help us create a quality candidate set for a searched user. Hence, we devise heuristic and unsupervised identity search methods based only on public and discriminative user attributes that avoid the need of any user authorization or privacy breach. Most literature search methods, on the other hand, rely on availability on private and public attributes. Evaluation of real world users and popular social networks show that methods effectively fetch the correct identity for 13.1% more users.

- **Observe and highlight user behavior over time**: Not only do users create redundant information across OSNs, but also over time. On tracking 8.7 million users on a popular OSN, Twitter, we find that about 73% users changed atleast one of their profile attributes within a short duration of two months. Such frequent changes to one's profile need attention to understand if the changes are benign or carried out with malicious intentions. Not limiting to Twitter, other OSNs like Facebook, Instagram and Tumblr observe users to change their unique attributes like username. Such unique attributes help others find a user, however frequent changes to username leave them with broken links. Yet 10% tracked Twitter users change their usernames. Such evolution of attributes over time can implications, however can be leveraged for identity resolution.

- **A study of username evolution**: Frequent changes to a unique attribute of a user, username, draws attention as to why users change their usernames so frequently and how do they create new usernames. Such study can help tag malicious users as well as understand redundant username creation patterns of a user that may extend to other OSNs. We analyze Twitter users tracked every fifteen minutes and find that 20% users constitute 80% of username changes recorded; users further create dissimilar new usernames. We learn username creation patterns over time. Further investigation on intentions for change reveals that few users change for benign reasons like supporting an event or to avoid boredom, but few change to trick users, abscond suspension, and squat good usernames.

- **Method for identity linking that leverage user evolution over time**: On observing that users change their profiles over time, it is plausible that current snapshots of user identities may fail to match. In such cases, current identity linking methods falsely predict that user identities refer to different individuals. To revise such false predictions, we suggest to compare current as well as pas snapshots of the user identities. Experimenting with username, we compare username sets, each composed of past and present usernames of each user identity, and match username creation patterns learned in the earlier study. With username histories, we could revise predictions for 48% users, thereby reducing missed links (false negatives) of an identity linking framework. We emphasize here that attribute history can help various applications, including identity resolution.

7.2 Limitations

- **Identity search: Dependency on API**: Our heuristic search methods rely on the API search endpoints of other OSNs. If a search parameter is not supported by the API, it is challenging to retrieve candidate identities similar to a searched user identity on the mentioned parameter. Therefore, the methods are asymmetric i.e. the methods need to be modified to start with a Facebook user identity and find the corresponding Twitter identity.

- **Identity linking: Using username only**: The proposed framework for identity linking is evaluated on username sets only. History of other attributes like profile picture and description, that change more frequently than username, can be used further to link user profiles. We could not do so because of non-availablility of data of the respective attributes on other OSNs. However, note that, our aim is to devise an identity linking framework that leverage attribute history and evaluation shows that with only username, we achieve a reduction in false negative rate by 48%. Other attributes can further improve the false negative rate. On a second note, achieving a significantly better false negative rate with a single universal attribute indicates the importance and impact of username.

- **Evaluation: On self-identified users**: Ground truth datasets of real-world users, used for evaluation of identity search and linking methods, contain those users who explicitly self-identify their identities on multiple social networks i.e. who expose themselves voluntarily. A validated dataset of users and their identities across OSNs who do not explicitly identify their own accounts is challenging to gather. Therefore, applicability and performance of our methods on non self-identified users is difficult to examine. However, our methods capture redundancies across OSNs and over time, if any user exhibits similarities across her identities, our methods can fairly link the identities.

- **Username evolution study**: Reasons for username change are verified empirically and with user responses to our tweets. To ask and validate reason for each username change, we crafted a survey to be distributed to users via tweets. The survey intends to collect reasons for which each user changed her username over time. In spite of our various methods to disperse the survey and attract users to fill it, only a few responded. Hence, we might have missed few valid and important reasons for which users evolve their usernames over time and choose dissimilar new values.

7.3 List of Publications

- Our work in Section 3.1 is published as:
  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*, WWW '13 Companion. ACM, New York, NY, USA, 1259-1268. DOI=http://dx.doi.org/10.1145/2487788.2488160.

- Our work in Section 3.2 is a collaborated work and published as:
  Niyati Chhaya, Dhwanit Agarwal, Nikaash Puri, Paridhi Jain, Deepak Pai, and Ponnurangam Kumaraguru. 2015. EnTwine: Feature Analysis and Candidate Selection for Social User Identity Aggregation. *In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '15. ACM, New York, NY, USA, 1575-1576, DOI=http://dx.doi.org/10.1145/2808797.2809340.

- Our work in Chapter 4 is published as the following and a journal version is under submission:
  Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. 2015. Other Times, Other Values: Leveraging Attribute History to Link User Profiles across Online Social Networks. *In Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15. ACM, New York, NY, USA, 247-255. DOI=http://dx.doi.org/10.1145/2700171.2791040.

- Our work in Chapter 5 is published as:
  Paridhi Jain and Ponnurangam Kumaraguru. 2016. On the Dynamics of Username Changing Behavior on Twitter. *In Proceedings of the 3rd IKDD Conference on Data Science, 2016*, CODS '16. ACM, New York, NY, USA, , Article 6 , 6 pages. DOI=http://dx.doi.org/10.1145/2888451.2888452.

- Our work in Chapter 6 is published as:
  Prachi Jain, Paridhi Jain, and Ponnurangam Kumaraguru. 2013. Call me Maybe: Understanding Nature and Risks of sharing Mobile Numbers on Online Social Networks. *In*

Proceedings of the first ACM Conference on Online social networks, COSN '13. ACM, New York, NY, USA, 101-106, DOI=http://dx.doi.org/10.1145/2512938.2512959.

- Other publication that is not included in the thesis:
  Paridhi Jain, Tiago Rodrigues, Gabriel Magno, Ponnurangam Kumaraguru, and Virgilio Almeida. Cross-Pollination of Information in Online Social Media: A Case Study on Popular Social Networks. *In Proceedings of the 2011 IEEE 3rd International Conference on Social Computing*, SocialCom ′11, pages 477–482, Oct 2011.

7.4 Future work

We believe that insights gathered from this dissertation can help stakeholders to build aggregated user profiles, derive their likings, and interests and find bad malicious users on the network. Based on our experience so far, we suggest the following directions:

Improve resolution methods with comprehensive list of public attributes: Existing and proposed methods can be improved based on attributes of a user like time profiles (time when the posts are shared across OSNs), activity, locations and stylometric features, and past values of the attributes can help in effective identity search. Specifically, identifiable information like location profiles constituted with geo-tags of the posts and user mentioned locations within the post, can enhance the identity search and linking accuracy. Also, for identity linking, we recommend extensive tracking of all attributes of a random sample of users across OSNs, so that importance of history of other attributes can be proved. We also believe that research can monitor users and find patterns that are indicative of malicious users.

Address the challenge of fake identities: Beyond improving accuracy of identity resolution methods, research can focus on understanding the possibility of a malicious user (attacker) cloning other user's identity (victim e.g. "paridhij"). In scenarios like these, identity resolution methods may incorrectly link cloned identity (created by the attacker e.g. "paridhis) to a real identity of the victim (i.e. "paridhij"), thus may hurt the online reputation of the victim. Future identity resolution methods should cater to the need of identifying cloned identities first and later perform the task of connecting identities.

Build privacy nudge technologies: We believe that future technologies can nudge users and suggest appropriate measures to avoid identity leaks, depending on the their choice of disclosing or restraining the connection among her identities. The nudge can take help from the insights developed from resolution methods suggested in this work. For instance, this work highlights that consistent sharing of attributes give away information about a user's presence across multiple OSNs. If she wishes to restrain linking among her identities, what attributes should she generalize / hide / skip / lie / restrict to share across OSNs. Else, if she wishes to be discoverable by her friends and contacts, what attributes should she re-post / share / describe. Thus in future, tools built on identity resolution methods, can also support users to secure their privacy.

Bibliography

[1] Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. Joint Link-attribute User Identity Resolution in Online Social Networks. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis*, SNAKDD '12. ACM, 2012.

[2] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: A Generic Approach to Entity Resolution. volume 18, pages 255–276, Secaucus, NJ, USA, January 2009. Springer-Verlag New York, Inc.

[3] Shea Bennett. 4 Reasons Why You Need to Change Your Username on Twitter. `http://www.adweek.com/socialtimes/twitter-username-tips/453851`, 2012. [Online; accessed 21-July-2015].

[4] Indrajit Bhattacharya and Lise Getoor. Online Collective Entity Resolution. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, AAAI '07.

[5] Indrajit Bhattacharya and Lise Getoor. Collective Entity Resolution in Relational Data. In *ACM Transactions on Knowledge Discovery from Data*, volume 1 of *TKDD*, New York, NY, USA, March 2007. ACM.

[6] Indrajit Bhattacharya, Lise Getoor, and Louis Licamele. Query-time Entity Resolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 529–534, New York, NY, USA, 2006. ACM.

[7] Mikhail Bilenko and Raymond J. Mooney. Adaptive Duplicate Detection Using Learnable String Similarity Measures. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 39–48, New York, NY, USA, 2003. ACM.

[8] Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *Proceedings of the*

18th International Conference on World Wide Web, WWW '09, pages 551–560, New York, NY, USA, 2009. ACM.

[9] Mustafa Bilgic, Louis Licamele, Lise Getoor, and Ben Shneiderman. D-dupe: An interactive tool for entity resolution in social networks. In *Proceedings of the IEEE Symposium On Visual Analytics Science And Technology, 2006*, pages 43–50. IEEE, 2006.

[10] Katherine Blashki and Sophie Nichol. Game geek's goss: Linguistic Creativity in Young Males within an Online University Forum. volume 3, pages 71–80. Australian Centre for Emerging Technologies and Society, 2005.

[11] David Guy Brizan and Abdullah Uz Tansel. A Survey of Entity Resolution and Record Linkage Methodologies. In *Communications of the International Information Management Association*, volume 6, pages 41–50, 2015.

[12] Jeff Bullas. 33 Social Media Facts and Statistics You Should Know in 2015. `http://www.jeffbullas.com/2015/04/08/33-social-media-facts-and-statistics-you-should-know-in-2015/`, 2015. [Online; accessed 19-January-2016].

[13] Francesca Carmagnola, Francesco Osborne, and Ilaria Torre. User Data Distributed on the Social Web: How to Identify Users on Different Social Systems and Collecting Data About Them. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, pages 9–15, New York, NY, USA, 2010. ACM.

[14] Pew Research Center. Social Media Matrix. `http://www.pewinternet.org/2013/12/30/social-media-matrix/`, 2013. [Online; accessed 22-January-2016].

[15] Pew Research Center. Social Media Matrix. `http://www.pewinternet.org/2015/01/09/social-media-update-2014/pi_2015-01-09_social-media_10/`, 2015. [Online; accessed 22-January-2016].

[16] Terence Chen, Mohamed Ali Kaafar, Arik Friedman, and Roksana Boreli. Is More Always Merrier?: A Deep Dive into Online Social Footprints. In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN'12, pages 67–72, New York, NY, USA, 2012. ACM.

[17] Yang Chen, Chenfan Zhuang, Qiang Cao, and Pan Hui. Understanding Cross-site Linking in Online Social Networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, SNAKDD '14, pages 61–69, New York, NY, USA, 2014. ACM.

[18] Zhaoqi Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra. Adaptive Graphical Approach to Entity Resolution. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 204–213, New York, NY, USA, 2007. ACM.

[19] Yao Cheng, Lingyun Ying, Sibei Jiao, Purui Su, and Dengguo Feng. Bind Your Phone Number with Caution: Automated User Profiling Through Address Book Matching on Smartphone. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, ASIA CCS '13, pages 335–340, New York, NY, USA, 2013. ACM.

[20] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi. sh/$oCiaL: the phishing landscape through short URLs. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 92–101. ACM, 2011.

[21] Yueh-Hsuan Chiang, AnHai Doan, and Jeffrey F. Naughton. Tracking Entities in the Dynamic World: A Fast Algorithm for Matching Temporal Records. In *Proceedings of the International Conference on Very Large Data Bases*, VLDB '14.

[22] Peter Christen and Ross Gayler. Towards Scalable Real-time Entity Resolution Using a Similarity-aware Inverted Index Approach. In *Proceedings of the 7th Australasian Data Mining Conference*, AusDM '08, pages 51–60. Australian Computer Society, Inc., 2008.

[23] Cheng Ta Chung, Chia Jui Lin, Chih Hung Lin, and Pu Jen Cheng. Person Identification Between Different Online Social Networks. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies - Volume 01*, WI-IAT '14, pages 94–101, Washington, DC, USA, 2014. IEEE Computer Society.

[24] G. Cockerell. *Making Marketing Meaningful*. Kendall Hunt Publishing Company, 2010.

[25] William W. Cohen. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pages 201–212, New York, NY, USA, 1998. ACM.

[26] The Walt Disney Company. Segmenting Target Markets. `http://thewaltdisneyco.blogspot.in/2011/11/chapter-8-segmenting-targeting-markets.html`, 2011. [Online; accessed 02-February-2015].

[27] D. Correa, A. Sureka, and R. Sethi. WhACKY! - What Anyone Could Know about You from Twitter. In *Proceedings of the 10th Annual International Conference on Privacy, Security and Trust, 2012*, PST'12, pages 43–50, 2012.

[28] Keith Cortis, Simon Scerri, Ismael Rivera, and Siegfried H. Discovering Semantic Equivalence of People behind Online Profiles. In *Proceedings of the Resource Discovery (RED) Workshop*, RED'12, pages 104–118, 2012.

[29] Tom Curtin. The Name Game: Cybersquatting and Trademark Infringement on Social Media Websites. In *Journal of Law and Policy*, volume 19. 2010.

[30] Souripriya Das, Eugene Inseok Chong, George Eadon, and Jaannathan Srinivasan. Supporting Ontology-based Semantic Matching in RDBMS. In *Proceedings of the 30th International Conference on Very Large Data Bases*, VLDB '04, pages 1054–1065. VLDB Endowment, 2004.

[31] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-scale Entity Linking. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 469–478. ACM, 2012.

[32] Prateek Dewan, Mayank Gupta, Kanika Goyal, and Ponnurangam Kumaraguru. MultiOSN: Realtime Monitoring of Real World Events on Multiple Online Social Media. In *Proceedings of the 5th IBM Collaborative Academia Research Exchange Workshop*, I-CARE '13, pages 61–64, New York, NY, USA, 2013. ACM.

[33] AnHai Doan and Alon Y. Halevy. Semantic-integration Research in the Database Community. In *AI Magazine - Volume 26*, volume 26, pages 83–94, Menlo Park, CA, USA, March 2005. American Association for Artificial Intelligence.

[34] M. G. Elfeky, V. S. Verykios, and A. K. Elmagarmid. TAILOR: A Record Linkage Toolbox. In *Proceedings of the 18th International Conference on Data Engineering, 2002*, ICDE '02, pages 17–28, 2002.

[35] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate Record Detection: A Survey. In *IEEE Transactions on Knowledge and Data Engineering*, volume 19 of *TKDD*, pages 1–16, Jan 2007.

[36] Facebook. What Names are Allowed on Facebook? `https://www.facebook.com/help/112146705538576`. [Online; accessed 06-February-2016].

[37] Facebook. Facebook User Statistics. `http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/`, 2015. [Online; accessed 21-January-2016].

[38] Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. Harvesting Multiple Sources for User Profile Learning: A Big Data Study. In *Proceedings of the 5th ACM on*

International Conference on Multimedia Retrieval, ICMR '15, pages 235–242, New York, NY, USA, 2015. ACM.

[39] R. Feizy, I. Wakeman, and D. Chalmers. Transformation of Online Representation through Time. In *Proceedings of the 2009 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '09, pages 273–278, July 2009.

[40] Paul Fennemore. How Social Media Influences Market Segmentation. `http://www.marketingtechnews.net/news/2012/mar/16/how-social-media-influencing-marketing-segmentation/`, 2012. [Online; accessed 07-February-2015].

[41] J.G. FitzGerald. Cross-media Interactivity Metrics, 2009.

[42] Recorded Future. ISIS Jumping from Account to Account, Twitter Trying to Keep Up . `https://www.recordedfuture.com/isis-twitter-activity/`, 2014. [Online; accessed 14-June-2015].

[43] Kahina Gani, Hakim Hacid, and Ryan Skraba. Towards Multiple Identity Detection in Social Networks. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 503–504, New York, NY, USA, 2012. ACM.

[44] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting Innocuous Activity for Correlating Users Across Sites. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 447–458, New York, NY, USA, 2013. ACM.

[45] Jennifer Golbeck and Matthew Rothstein. Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In *Proceedings of the 23rd National Conference on Artificial Intelligence*, AAAI '08, pages 1138–1143. AAAI Press, 2008.

[46] Tim Grant. Txt 4n6: Method, consistency, and distinctiveness in the analysis of SMS text messages. In *Journal of Law, Economics & Policy*, volume 21, page 467. HeinOnline, 2012.

[47] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. In *Machine Learning*, volume 46, pages 389–422, Hingham, MA, USA, March 2002. Kluwer Academic Publishers.

[48] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[49] Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. Cascaded Classification Models: Combining Models for Holistic Scene Understanding. In *Proceedings of the Neural Information Processing Systems, 2009*, NIPS '09, 2009.

[50] Instagram. Celebrating a Community of 400 million. `https://instagram.com/press/`, 2015. [Online; accessed 19-January-2016].

[51] Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. Identifying Users Across Social Tagging Systems. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11.

[52] Danesh Irani, Steve Webb, Kang Li, and Calton Pu. Large Online Social Footprints–An Emerging Threat. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 03*, CSE '09, pages 271–276, Washington, DC, USA, 2009. IEEE Computer Society.

[53] Jun Ito, Kyosuke Nishida, Takahide Hoshide, Hiroyuki Toda, and Tadasu Uchiyama. Demographic and Psychographic Estimation of Twitter Users Using Social Structures. In *Online Social Media Analysis and Visualization*, pages 27–46. Springer, 2014.

[54] Paridhi Jain and Ponnurangam Kumaraguru. On the Dynamics of Username Changing Behavior on Twitter. In *Proceedings of the 3rd IKDD Conference on Data Science, 2016*, CODS '16, pages 61–66, New York, NY, USA, 2016. ACM.

[55] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. @I Seek 'fb.me': Identifying Users Across Multiple Online Social Networks. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 1259–1268, New York, NY, USA, 2013. ACM.

[56] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. Other Times, Other Values: Leveraging Attribute History to Link User Profiles Across Online Social Networks. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 247–255, New York, NY, USA, 2015. ACM.

[57] Paridhi Jain, Tiago Rodrigues, Gabriel Magno, Ponnurangam Kumaraguru, and Virgilio Almeida. Cross-Pollination of Information in Online Social Media: A Case Study on Popular Social Networks. In *Proceedings of the 2011 IEEE 3rd International Conference on Social Computing*, SocialCom '11, pages 477–482, Oct 2011.

[58] Matthew A Jaro. *Unimatch: A Record Linkage System: Users manual*. Bureau of the Census, 1978.

[59] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. Timeprints for Identifying Social Media users with Multiple Aliases. In *Security Informatics*, volume 4, pages 1–11. Springer, 2015.

[60] Colin Keigher. Being an Avivore and data mining Twitter. `http://colin.keigher.ca/2013/04/being-avivore-and-data-mining-twitter.html`, 2013. [Online; accessed 17-March-2015].

[61] Farshad Kooti, Haeryun Yang, Meeyoung Cha, Krishna P Gummadi, and Winter A Mason. The Emergence of Conventions in Online Social Networks. In *Proceedings of the 6th International Conference on Weblogs and Social Media*, ICWSM '12, pages 194–201. AAAI, 2012.

[62] Hanna Kopcke and Erhard Rahm. Training Selection for Tuning Entity Matching. In *Proceedings of the International Conference on Very Large Data Bases*, VLDB '08, page 3. VLDB Endowment, 2008.

[63] Hanna Kopcke and Erhard Rahm. Frameworks for Entity Matching: A Comparison. In *Data and Knowledge Engineering*, volume 69, pages 197–210, Amsterdam, The Netherlands, The Netherlands, February 2010. Elsevier Science Publishers B. V.

[64] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[65] Sebastian Labitzke, Irina Taranu, and Hannes Hartenstein. What Your Friends Tell Others About You: Low Cost Linkability of Social Network Profiles. In *Proceedings of the 5th International ACM Workshop on Social Network Mining and Analysis*, SNAM '11, pages 1065–1070. ACM, 2011.

[66] Howard Lei, Jaeyoung Choi, Adam Janin, and Gerald Friedland. User Verification: Matching the Uploaders of Videos across Accounts. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2011*, ICASSP '11, pages 2404–2407. IEEE, 2011.

[67] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, 1966.

[68] Pei Li, Xin Dong, Andrea Maurino, and Divesh Srivastava. Linking Temporal Records. volume 4 of *VLDB '11*, pages 956–967. VLDB Endowment, 2011.

[69] Bang Hui Lim, Dongyuan Lu, Tao Chen, and Min-Yen Kan. # mytweet via Instagram: Exploring User Behaviour across Multiple Social Networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15.

[70] Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, and Hsiao-Wuen Hon. What's in a Name?: An Unsupervised Approach to Link Users Across Communities. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 495–504, New York, NY, USA, 2013. ACM.

[71] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 51–62, New York, NY, USA, 2014. ACM.

[72] Wendy Liu and Derek Ruths. What‘s in a Name? Using First Names as Features for Gender Inference in Twitter. In *Proceedings of the 2013 AAAI Spring Symposium Series*, 2013.

[73] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The Tweets They Are a-Changi′‘: Evolution of Twitter Users and Behavior. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM '14, pages 305–314, 2014.

[74] Anshu Malhotra, Luam Totti, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. Studying user footprints in different online social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, (ASONAM '12, pages 1065–1070. IEEE Computer Society, 2012.

[75] Huina Mao, Xin Shuai, and Apu Kapadia. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, pages 1–12, New York, NY, USA, 2011. ACM.

[76] Adam Marcus, Eugene Wu, David Karger, Samuel Madden, and Robert Miller. Human-powered Sorts and Joins. In *Proceedings of the International Conference of Very Large Data bases*, volume 5 of *VLDB '11*, pages 13–24. VLDB Endowment, September 2011.

[77] Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. Efficient Clustering of High-dimensional Data Sets with Application to Reference Matching. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 169–178, New York, NY, USA, 2000. ACM.

[78] Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 251–260, New York, NY, USA, 2010. ACM.

[79] Marti Motoyama and George Varghese. I Seek You: Searching and Matching Individuals in Social Networks. In *Proceedings of the 11th International Workshop on Web Information and Data Management*, WIDM '09, pages 67–75, New York, NY, USA, 2009. ACM.

[80] David Nadeau and Satoshi Sekine. A Survey of Named Entity Recognition and Classification. In *Lingvisticae Investigationes*, volume 30, pages 3–26. John Benjamins publishing company, 2007.

[81] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing Social Networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, pages 173–187, Washington, DC, USA, 2009. IEEE Computer Society.

[82] Byung-Won On, Ingyu Lee, and Dongwon Lee. Scalable Clustering Methods for the Name Disambiguation Problem. In *Journal of Knowledge and Information Systems*, volume 31, pages 129–151, New York, NY, USA, April 2012. Springer-Verlag New York, Inc.

[83] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. How Unique and Traceable Are Usernames? In *Proceedings of the 11th International Conference on Privacy Enhancing Technologies*, PETS '11, pages 1–17, Berlin, Heidelberg, 2011. Springer-Verlag.

[84] Nicole Perlroth. Verifying Ages Online Is a Daunting Task, Even for Experts. `http://www.nytimes.com/2012/06/18/technology/verifying-ages-online-is-a-daunting-task-even-for-experts.html`, 2012. [Online; accessed 02-February-2015].

[85] Ferran Pla and L Hurtado. Political Tendency Identification in Twitter using Sentiment Analysis Techniques. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING'14, pages 183–192, 2014.

[86] Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp'12, pages 898–905, New York, NY, USA, 2012. ACM.

[87] M. F. Porter. Readings in Information Retrieval. In Karen Sparck Jones and Peter Willett, editors, *Program*, volume 40, chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[88] Lisa P. Ramsey. Brandjacking on Social Networks: Trademark Infringement by Impersonation of Markholders. In *Buffalo Law Review*, volume 58, page 851. 2010.

[89] ET Bureau: Ravi Teja Sharma. Qaiser Ali: The Lucknow Boy Who Acquired PM's Twitter handle for 30 minutes. `http://articles.economictimes.indiatimes.com/2014-05-23/news/50055221_1_new-handle-mark-zuckerberg-pm-manmohan-singh`, 2014. [Online; accessed 02-June-2015].

[90] Mohamed Shehab, Moo Nam Ko, and Hakim Touati. Social Networks Profile Mapping Using Games. In *Proceedings of the 3rd USENIX Conference on Web Application Development*, WebApps '12, pages 27–38, 2012.

[91] Xiaolin Shi, Ramesh Nallapati, Jure Leskovec, Dan McFarland, and Dan Jurafsky. Who Leads Whom: Topical Lead-lag Analysis across Corpora. In *Proceedings of the Neural Information Processing Systems Workshop*, NIPS Workshop '10, 2010.

[92] Parag Singla and Pedro Domingos. Entity Resolution with Markov Logic. In *Proceedings of the 6th International Conference on Data Mining*, ICDM '06, pages 572–582, Washington, DC, USA, 2006. IEEE Computer Society.

[93] Temple F Smith and Michael S Waterman. Identification of Common Molecular Subsequences. In *Journal of molecular biology*, volume 147, pages 195–197. Elsevier, 1981.

[94] Mudhakar Srivatsa and Mike Hicks. Deanonymizing Mobility Traces: Using Social Network As a Side-channel. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 628–637, New York, NY, USA, 2012. ACM.

[95] Gary Stein. The Five Biggest Mistakes in Measuring Social Media. `https://www.clickz.com/clickz/column/1716119/the-five-biggest-mistakes-measuring-social-media`, 2009. [Online; accessed 07-February-2015].

[96] Martin N. Szomszor, Iván Cantador, and Harith Alani. Correlating User Profiles from Multiple Folksonomies. In *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, HT '08, pages 33–42, New York, NY, USA, 2008. ACM.

[97] Ryan Tate. Facebook introduces Hashtags. `http://www.wired.co.uk/news/archive/2013-06/13/facebook-hashtags`, 2013. [Online; accessed 20-January-2016].

[98] Andreas Thor and Erhard Rahm. MOMA - A Mapping-based Object Matching System. In *Proceedings of the Conference on Innovative Data Systems Research*, CIDR '07, pages 247–258, 2007.

[99] Twitter. Signing Up for Twitter. `https://support.twitter.com/articles/100990#`. [Online; accessed 02-February-2016].

[100] Twitter. Twitter usage: Company facts. `https://about.twitter.com/company`, 2015. [Online; accessed 21-January-2016].

[101] Esko Ukkonen. Approximate String-matching with Q-grams and Maximal Matches. In *Theoretical Computer Science*.

[102] Jan Vosecky, Dan Hong, and Vincent Y Shen. User Identification across Multiple Social Networks. In *Proceedings on the 1st International Conference on Networked Digital Technologies, 2009*, NDT '09, pages 360–365. IEEE, 2009.

[103] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. CrowdER: Crowdsourcing Entity Resolution. In *Proceedings of the International Conference on Very Large Data Bases*, volume 5 of *VLDB '12*, pages 1483–1494. VLDB Endowment, July 2012.

[104] M. Waterman, T. Smith, and W. Beyer. Some Biological Sequence Metrics. In *Advances in Mathematics*, pages 367–387, 1976.

[105] A. Weinstein. *Handbook of Market Segmentation: Strategic Targeting for Business and Technology firms.* Haworth Press, 2004.

[106] Steven Euijong Whang, Peter Lofgren, and Hector Garcia-Molina. Question Selection for Crowd Entity Resolution. Technical report, Stanford University, 2012.

[107] Wikipedia. List of Social Networking Sites. `https://en.wikipedia.org/wiki/List_of_social_networking_websites`, 2015. [Online; accessed 24-March-2015].

[108] Jiejun Xu, Tsai-Ching Lu, Ryan Compton, and David Allen. Quantifying Cross-platform Engagement Through Large-scale User Alignment. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 281–282, New York, NY, USA, 2014. ACM.

[109] Jinxi Xu and W. Bruce Croft. Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996. ACM.

[110] Quanzeng You, Sumit Bhatia, Tong Sun, and Jiebo Luo. The Eyes of the Beholder: Gender Prediction using Images posted in Online Social Networks. In *Proceedings of the IEEE International Conference on Data Mining Workshop, 2014*, ICDMW '14, pages 1026–1030.

[111] Reza Zafarani and Huan Liu. Connecting Corresponding Identities across Communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '09, pages 354–357, 2009.

[112] Reza Zafarani and Huan Liu. 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*, KDD '13, pages 41–49, New York, NY, USA, 2013. ACM.

[113] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S. Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1485–1494, New York, NY, USA, 2015. ACM.

[114] Elena Zheleva and Lise Getoor. Privacy in Social Networks: A Survey. In *Social network data analytics*, pages 277–306. Springer, 2011.