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
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. 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. It is the most effective segmentation citing a rise of 24% in business performance; however includes high cost in both time and money. 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.

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 \textsuperscript{36}, while Twitter allows users to register under pseudonyms followed by an email verification \textsuperscript{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 asynchronous manner \textsuperscript{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 auxiliary 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.