Entity Resolution Techniques

Comprehensive Report

Student Name: Paridhi Jain

March 22, 2013

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Dr. Ponnurangam Kumaraguru (Chair)
Dr. Anupam Joshi
Dr. Gaurav Gupta
Dr. Vikram Goyal

Advisor
Dr. Ponnurangam Kumaraguru

©2013 IIIT-Delhi
All rights reserved

Keywords: Identity resolution, identity search, profile matching, fake identity, identity cloning

Abstract

A real-world entity is defined by a set of attributes which describe her uniquely. Her attributes can be digitally recorded in contexts as databases or in online social networks. In databases, her identity attributes are recorded as a database record, while in online social networks, her identity attributes are recorded as profile, content or network attributes collectively. An entity plays different roles with varying attributes in different contexts, therefore verifying if personalities with different roles in different contexts belong to the same real-world entity or not, is a difficult task. In databases, the problem of identifying if two records refer to the same real-world entity is termed as "Entity Resolution in Databases". Researchers have devised multiple techniques to correlate two records on multiple grounds, to remove duplicates, as well as aggregate more data of an entity present across databases. On the similar lines, identities across online social networks are studied and analyzed to understand if a set of identities belong to the same user or not. Further, researchers have proposed techniques to not only match a given set of identities but to find possible identities referring to the same individual, within and across social networks. This problem is referred to as "Entity Resolution in Online Social Networks". The problem differs from Entity Resolution in databases, as in databases are maintained and created by administrators / managers and have duplicate, unlinked entities due to human errors, different databases etc. with no admin's malicious intent to manipulate the record, however in online social networks, a profile, content and network attributes are created and maintained by the user itself where user might have malicious intents and, therefore trustworthiness of the attributes is questionable. In this report, we present a comprehensive overview of the entity resolution techniques, tools and frameworks proposed in literature (till date) for both databases and online social networks, to comprehend the possible methodologies tested to solve the problem assuming both non-malicious attributes and malicious data attributes, to understand maturity of the domains to quest for research gaps and scope of further research in the domains.

Contents

1 Introduction **1**

2 Entity Resolution **4**

2.1 Problem 4

2.2 Application Domains 4

2.2.1 Databases 4

2.2.2 Online Social Networks 6

2.3 Challenges 7

3 Entity Resolution in Databases **9**

3.1 Syntactic Resolution 10

3.2 Semantic Resolution 12

3.3 CrowdSourcing Resolution 12

3.4 Other Resolution Methods 13

3.4.1 Challenges 13

4 Entity Resolution in Online Social Networks **15**

4.1 Database Methods 15

4.1.1 Profile Resolution 16

4.2 Online social network methods 18

4.2.1 Content Resolution 18

4.2.2 Network Resolution 19

4.2.3 Search Engine Resolution 19

4.2.4 CrowdSourcing Resolution 19

5 Malicious Entity Resolution Methods **22**

5.0.5 Social networks and Digital Forensics 24

6 Entity Resolution Tools **25**

7 Discussion **27**

List of Figures

1.1 Timeline of the literature of Entity Resolution in Databases. 2
1.2 Timeline of the literature of Identity Resolution in Online Social Networks. 2

2.1 Entity Resolution of online identities that refer to a unique user 7

4.1 A comprehensive view of Identity Resolution Methods 16

6.1 A comparison of open Identity Search and Resolution systems 25

List of Tables

2.1 An example table in an organizational database 5
2.2 Multiple data sources of information for an entity 5
2.3 An equi join example of multiple tables in a database 5

4.1 A comparison of the existing methodologies for entity resolution in online social networks. 21

Chapter 1

Introduction

An entity that holds a unique existence in the real world has different roles to play in different contexts. To suit each context, an entity features only a subset of its attributes, few of them are variable and few invariable. Variable attributes of an entity are restricted to the context and therefore as the context changes, they change. Invariable attributes of an entity are independent of the context and do not change as context changes. To understand the unique and all-around characteristics of an entity, different contexts are put together to aggregate an entity's attributes. Since an entity's invariable attributes hold same or similar value across contexts, they connect different contexts together, to create an overall identity of an entity. Such aggregation of information about an entity from multiple sources has been discussed in sociology and exploited in computer science. Researchers in computer science have tried to build methods to connect different information sources of an entity together to know more about the entity. For instance, different aspects of an employee can be known by aggregating personal information from an online service with professional information from an organization records. The process of aggregating an entity information from multiple sources and producing an unified comprehensive view is termed as **"Data Integration"** in Computer Science.

Data integration process is majorly carried out in a specific computer science application domain – Databases. Two databases (tables) are joined together to add more information to a single database, and to infer more details about an entity. Along with that, duplicate entries with slight variations introduced due to human errors, are suggested to remove to clean the databases. Data integration and Database cleaning introduces a classical problem of inferring whether two records (within or across databases) refer to a single entity. This problem is termed as **Entity Resolution** and is well researched in databases in computer science domain [1–10]. Timeline of the research done in this domain is given in Figure 1.1. Height does not hold any implications. We can observe that seminal work in entity resolution started a decade earlier.

When entity resolution was researched in databases, a new platform was emerging. Advancement of Web 2.0 gave birth to new ways of sharing and consuming information, making connections via the introduction of Online Social Networks (OSNs) with users at its core. Users join an online social network and create an identity for themselves. Note that, identity creation process on online social networks is not supported by any verification process, and users have huge control on what attributes should go as their identity attributes on social networks. It is therefore probable for users having varied identity attributes across various social networks, they become member of. To infer, if two identities within or across online social networks refer to the same individual, is an extension of the problem of entity resolution in databases

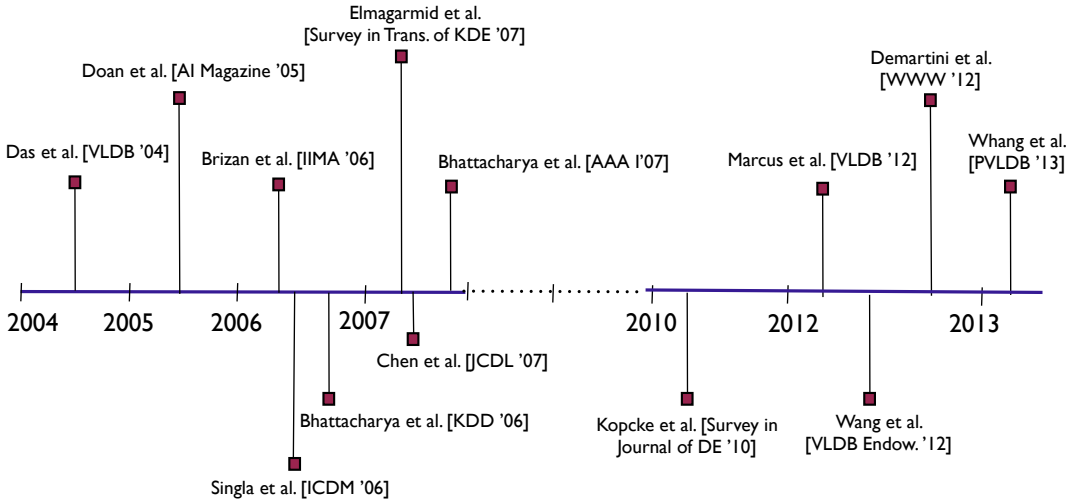

Figure 1.1: Timeline of the literature of Entity Resolution in Databases.

and is termed as **Identity Resolution** in Online Social Networks.

Since Identity Resolution in online social networks is an extension of entity resolution in databases, it can be assumed that entity resolution methods are applicable to solve identity resolution on social networks. However, the assumption is restricted and is valid in limited circumstances. In databases, records are not maintained by the entity itself but the administrators and entity has least control to manipulate its attributes, whereas in social networks, users have huge control on her attributes and can be easily manipulated by her. Along with that, social networks allow a variety of attributes (e.g. content, image, network, etc.) associated with an entity which have been exploited to devise better methods for identity resolution in online social networks [4, 11–19]. Timeline of the research done in this domain is given in Figure 1.2. Height does not hold any implications. We can observe that research community started working in this domain being motivated by entity resolution in databases.

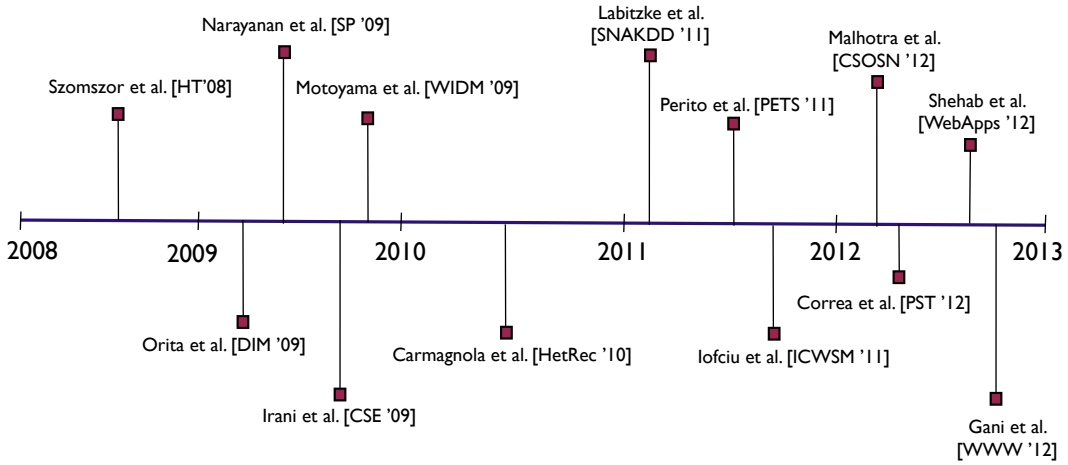

Figure 1.2: Timeline of the literature of Identity Resolution in Online Social Networks.

In this report, we present a comprehensive understanding of the entity (identity) resolution techniques discussed in literature till date applicable to databases and online social networks and suggest research

gaps in literature to promote future research. We discuss the Entity Resolution problem in detail in Chapter II. We then study methods devised to resolve entities in databases in Chapter III and then discuss techniques proposed to resolve entities across multiple online social networks in Chapter IV. Techniques discussed in Chapter III and Chapter IV assumes that entity does not intentionally obfuscate its attributes in different contexts and its certain invariable attributes hold same values irrespective of the context. However, certain entities intentionally features deceptive attributes. We discuss the deceptive entity resolution in databases and social networks in Chapter V. We then glance through some of the commercial tools available to solve the entity resolution problem and compare them on a certain set of parameters in Chapter VI. We summarize our understanding by discussing some of the open research problems in entity resolution domain specifically in computer science in Chapter VII.

Chapter 2

Entity Resolution

2.1 Problem

Multiple contexts in which an entity appear are put together to produce a comprehensive set of characteristics of an entity. Contexts may be completely different from each other as entity variable attributes may hold different values in different contexts. In such a case, an entity's invariable (constant) attributes bridge contexts to imply that the same entity is present in different contexts. Therefore, it is necessary for an entity to have a unique set of constant attributes in any context, it is a part of, to bridge multiple contexts. But this may not be true always. For the scenarios, where the entity's invariable attributes do not exist or if they exist but are not recorded, it is difficult to connect contexts of an entity and create a comprehensive view of her attributes. The problem of determining if a given set of entities with limited set of attributes available in different contexts, refer to a single real-world entity is termed as "**Entity Resolution**".

2.2 Application Domains

We focus on two major computer science domains where the problem significantly prevails – databases and online social networks. We first describe each of the domain and then understand the applications of the solutions to the problem in the domains.

2.2.1 Databases

With an explosive growth of digital data generated in multiple spheres (for example business, organizations, web), data storage and its fast retrieval became an important concern. As a solution to it, databases have been introduced in early sixties to store large amount of structured data [20]. Databases are extensively used in IT-based organizations and play an important role in information access in day-to-day transactions. Banking organizations, supermarkets, academic organizations, and e-governance organizations are few examples which highly rely on databases to provide efficient services to their users, build insights from the collective statistics of users and make better decisions to increase their productivity and customer base. Each database is a set of tables, which store data in the form of tuples or records. Each record represents an entity and is composed of a set of attributes which define her independent existence.

An entity may represent any name, place, thing, product or any object with virtual but unique existence. Each table or a database represent the context of which an entity is a part of. Attributes which uniquely identify a record in a table or database are termed as candidate keys, one of which is chosen as primary key for all the records in the table. Different tables may have different primary keys for its records. Table 2.1 shows a table in the database where records represents employees (entity) in an organization. Each employee (entity) is composed of the following attributes – Employee_ID (primary key), Name, Department, DOB, Place etc. Table represents a professional context.

To create a comprehensive view of an entity represented as a record, by aggregating its attributes from multiple tables or databases it is a part of, databases allow data integration by a set of operations termed as Join Operations. Join operations allow merging two records of two tables on a conditional clause. In most cases, two tables are merged via an equi-join operation on either primary or candidate key of the table, with the key available in both the tables to avoid duplicate entries. Table 2.3 shows an example of an equi-join operation on two tables in Table 2.2 with Employed_ID as a primary key for both the tables.

Employee_ID	Name	Department	DOB	Place	Position	Date_of_joining
319	Alice	Office	12-02-88	CA	Manager	13-11-11
320	Bob	Finance	23-11-86	CA	Manager	23-04-11
321	Bruce	Human Resources	02-03-91	CA	PA	04-09-07
322	Emily	Technical	03-09-84	CA	Technical Head	05-05-07

Table 2.1: An example table in an organizational database

Employee_ID	Name	DOB
319	Alice	12-02-88
320	Bob	23-11-86
321	Bruce	02-03-91
322	Emily	03-09-84

Employee_ID	Position	Date_of_joining
319	Manager	13-11-11
320	Manager	23-04-11
321	PA	04-09-07
322	Tech. Head	05-05-07

Table 2.2: Table 1 - Employee Personal Records, Table 2 - Employee Company Records

However, in the absence of any primary or candidate key available across all the tables to merge, it is difficult to make a decision if a record A and record B refer to the same entity. This problem of resolving if two records in two tables or databases, refer to same entity is termed as "*Entity Resolution in Databases*". The problem is also known by "Record Linkage", "Merge / Purge", "Record Resolution", and "Entity Disambiguation". Since most IT-based organizations deploy databases to store information, solution to this problem in databases domain can help in maintaining cleaner databases without duplicate entries, help in building better comprehensive view of an entity by aggregating its features from different databases, and may help the organizations to derive relationships between various entities in the databases, that can be further utilized to devise methods to improve productivity and services. From a security perspective, databases are used to store criminal identities and crime records. Entity Resolution techniques in databases may help in linking criminal's information stored across different databases and

Employee_ID	Name	DOB	Position	Date_of_joining
319	Alice	12-02-88	Manager	13-11-11
320	Bob	23-11-86	Manager	23-04-11
321	Bruce	02-03-91	PA	04-09-07
322	Emily	03-09-84	Technical Head	05-05-07

Table 2.3: Equi-join of Table 1 and Table 2 on Employee_ID attribute

may derive sensitive information from the merged information of each criminal.

2.2.2 Online Social Networks

Advent of Web 2.0 technology has revolutionized the ways in which people create, receive and share information among peers. Starting from Internet Messengers (IMs), websites, blogs, technology today empowers better information sharing services e.g. Online Social Networks. Online Social Network refers to an information sharing platform, which facilitates digital interactions among its users. Over the last decade, large number of online social networking sites have been introduced in webosphere. Each site offers different services as their main unique selling proposition (e.g. Flickr offers photo sharing, Youtube offers video sharing, Twitter offers micro-blogging). Variety of services offered by multiple online social networks facilitate various ways of information sharing, leading to the massive popularity of social networks. Users spend one-fourth of their Internet time on social networks and blogs [21]. Twitter claims 600,000 new users per day and 200 million tweets per day [22], and Facebook has 901 million users by March 2012 with an average user spending 8 hours a month on Facebook [23,24], implying the impact of social media on an online user.

Owing to a large number of online social networks, popular in different parts of the world, targeting different population and facilitating different services, users do not restrict their presence to one social network, but become a member of multiple social networks. Therefore a user as a real-world entity marks her presence in multiple online social networks which can be considered as contexts. A user's presence on multiple social networks helps her to control the nature and audience of information shared (e.g. professional and personal network) [14], and the rate of information dissemination (e.g. viral or restricted). A user defines her identity on each network which includes a set of attributes that describes her uniquely and differentiate it from others. A user's online identity includes her username, her profile, her friends network, and the content she creates or that is shared with her. Her identity creation process on each social network gives her huge control on how she can choose to give / hide / skip her identity attributes. Because of a larger user's control on her identity formulation, her identity attributes may vary largely across multiple social networks. With no handle / identifier / attribute for a user to mark her presence uniquely in Online Social Media, her multiple social network identities remain un-linked with each other. Because of varied and non-linked identities, it is difficult to reverse-map the identities back to a user. In such a scenario, given a set of identities on online social media, the process of establishing which identities represent the same real-world user and therefore, a single entity, is termed as *"Entity Resolution in Online Social Networks"*. The problem is also known as "Identity Resolution" since two identities on different social networks are resolved to understand if they belong to a single user.

Figure 2.1 explains the problem of entity resolution in social media domain. Figure shows a set of five online identities, where first two identities (paridhi.jain.399@Facebook and pari_lakshya @Twitter) refer to one individual while last three identities (PJ_YT, paridhi.jain08 and Paridhi Jain) belong to another user. The first two identities and last three identities are "Linked" identities. Identity 2 and 3 (pari_lakshya@Twitter and PJ_YT@YouTube) are identities of different users and therefore are termed as "Non-Linked" identities. Given the set of five identities with no information about their interconnectivity, identity resolution point to the problem of predicting that first two identities refer to one entity while other three refer to entity.

Solutions to the problem have multiple application domains. In security domain, entity resolution solutions can help searching for malicious user's multiple online identities. Malicious users exploit online

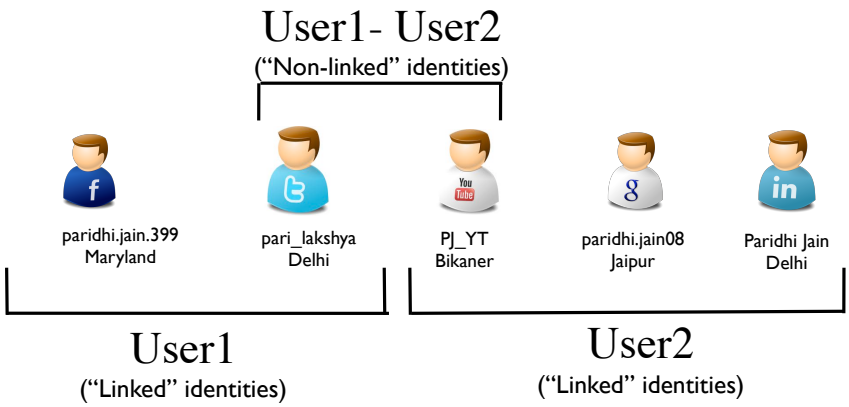

Figure 2.1: Entity Resolution of online identities that refer to a unique user. First two identities are to be resolved to refer to one entity while last three identities are to be resolved to refer to another.

social media for activities such as Phishing, Spam, Identity theft, and One click advertisements [25]. Such malicious users create multiple accounts on different networking sites to enhance reachability to targets (victims). To identify malicious users, security researchers have devised features on Twitter [26–29] on YouTube [30], Myspace [31] and other social networks. Solutions suggested to detect malicious user accounts on the sites mentioned are network dependent, hence security analysts need to identify malicious accounts on each networking site. In order to save identification cost and efforts, linking malicious user identities present on multiple online social networks is suggested. However in real world, malicious users demonstrate active obfuscation of their attributes to avoid detection and linkage of their multiple identities. To meet this challenge, behavioral based identity resolution solutions can help in linking malicious user's identities across social networks. In privacy domain, the problem finds its application in understanding the quantity and quality of the user's information leakages via either aggregation of user's information from multiple social networks or differences in privacy policies of multiple social networks [12, 32–34]. System analysts then can improve privacy policies and anonymization methods to preserve user's privacy. In system building domain, our solution can help in building recommendation feature for social aggregation sites. The recommendation feature can find a user's presence on multiple networks with the given user information on one network and suggest her to aggregate the suggested identities (probably belonging to her), hereby saving user effort. In marketing domain, the solution can be applied to aggregate customer data from multiple online shopping sites, to understand customer preferences and buying patterns and then recommend better deals to promote products. In e-commerce domain, with aggregating online networks with e-commerce sites, one can understand the stock market impression on public and correlate the ups and down in the stock market with the user's social behavior.

Note that, records in databases are analogous to identities in online social networks. Both represent an entity and is a set of attributes that describe a real-world entity. Therefore, techniques proposed for Record Resolution may be applied to Identity Resolution.

2.3 Challenges

Solving entity resolution problem in databases as well as on online social networks offers challenges. Firstly, in both domains, records and identities have limited attributes available for comparison, owing to limited data recorded in databases and strict privacy settings of users. Even with the few attributes

available, no guarantees are provided that the attributes are not manipulated by a user / analyst intentionally, and doesn't have a compulsion to represent a verified real-world entity. Therefore, there can be missing, inaccurate and inaccessible information which hinders a feasible and scalable entity resolution solution. Secondly, both domains suffer from the absence of a unique distinguishing attribute for each record in or across databases and an unique user identity across online social networks. Thirdly, entity resolution process is performed after searching for similar records (identities) across tables (online social networks). Therefore, given a record in a table (identity in a social network), an entity search on other tables (online social networks) is executed and a candidate set of records (identities) is generated. The candidate records (identities) showcase some level of similarity with the given record or identity. With limited accessible attributes for the given record (identity), entity search in a large database (big social networks), returns a long list of records (identities) possibly relating to the given search record (identity). Assigning scores to the candidate record (identity) set with limited attributes pose another challenge. Lastly, devising sophisticated methods to match different media e.g. text, audio, video, images in order to link multiple records (identities) is another big challenge.

Chapter 3

Entity Resolution in Databases

Researchers have proposed 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. 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, 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. 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 to 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*.

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 intend to survey the approachers discussed, however for the sake of completeness, we also discuss the research work, which aims to make the entity resolution techniques fast and scalable.

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 [1, 2]. 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 [35]. 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 the 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 [36]. 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 this concern by taking into account the number of characters overlap between two strings within allowed position shifts between the characters [37]. 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 for strings as"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

Smith-Waterman distance measures the substring matches between two strings [38]. The measure align the two strings in a manner so as to maximize local substring 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.

Q-gram distance

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 [39]. For example, q-grams of a string 'hello' with q=2, are 'he', 'el', 'll', 'lo'. Q-gram distance measures the number of q-grams present in both strings, irrespective of their position or frequency. Q-gram distance assigns high score to the strings with same or close spellings even when the strings vary largely. For example, 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 [40]. 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.

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 [3, 4]. 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 are SERF [41], D-Dupe [42], TAILOR [43], MOMA [44], MARLIN [45], STEM [46] , etc. Each of the system exploits semantic approach to resolve records in a relational database using matching attribute set [5].

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 [6–10]. 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 [6]. 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.

3.4 Other Resolution Methods

Apart from Syntactic, Semantic and Crowdsourcing based entity resolution methods, researchers have discussed the use of markov logic networks [47], graphical network structures [48, 49] and iterative (collective) resolution [49, 50] 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 [47]. 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 [48] 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 [49]. 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 [50]. 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. Very few approaches have been experimented on real-world and more generic databases.

Each of the methodology proposed to resolve database entities is independent and exclusively involves some inherent properties of entities e.g. connections, which helps in decision making of whether two entities with similar network structure with other entities. Further, decision made for an entity-entity pair can help in resolving other entity pairs connected to the resolved entities as discussed in collective entity resolution. However, few major disadvantages of such approaches includes domain dependency, and scalability. Citeseer (publication) databases may allow entities to form a network via their co-occurence in a research paper, while this may not hold true for customer databases where stand alone entities are stored without any reference of with whom they might co-occur or are connected to. Further, converting databases to graphs and then scaling the graph theoretic approach to work on large databases, may lead to serious scalability issues.

3.4.1 Challenges

Approaches discussed to address entity resolution in databases, face a major concern of generalizability and scalability. Owing to the entity resolution problem existence in multiple domains as citation matching, object consolidation, name disambiguation, etc., the complexity of the problem varies. Methods proposed to infer if two authors refer to the same real-world author for citation databases, may not be applicable for resolving actors in movie databases, since the basic assumption of the approach for author resolution is that a similar set of authors tend to work together, which might not hold true for actors. Therefore, Bhattacharya *et al.* approach may work for resolving authors but not actors. Further, owing to the continuous increase in the large size of databases, researchers have tried to minimize complexity by pruning out high confidence low matching record pairs, and then input only confusing record pairs to entity resolution engine [6, 48], by 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 [51], and by proposing inverted indexing techniques facilitating

fast matching [52]. However, as the database size increases, most entity resolution techniques tend to provide low accuracies, and researchers are in continuous endeavor of making entity resolution on large scale databases (near) realtime.

Chapter 4

Entity Resolution in Online Social Networks

As discussed in Chapter 2, an identity of a user on an online social network is composed of a set of attributes, e.g., name, city, education, etc. Her identity attributes can be classified in three categories – *Profile attributes*, *Content attributes*, and *Network attributes*. Profile attributes of a user are attributes which describes her basic characteristics. For example, name, age, gender, etc. Profile attributes can be considered as invariable attributes (discussed in Chapter 2), to an extent, since they tend to remain constant across online social networks. [1] Content attributes of a user are attributes which describes the content she creates or is shared with her. For example, language, timestamp, use of words, URLs, emoticons etc. Network attributes of a user are attributes which captures the properties of friend circle and group of a user. For example, number of friends, centrality of a user etc. Both content and network attributes are variable attributes of a user, which may not remain same across her social networks. Entity resolution techniques in online social networks exploit one or a combination of the attribute categories discussed.

4.1 Database Methods

Services provided by online social networks are facilitated by the underlying storage units, which are databases itself. Each online social network stores its user's information and attributes in big independent databases, where each record represents an identity in a database. Therefore, the problem of resolving two identities across online social networks can be mapped to the problem of resolving two records across databases deployed by each of the online social network. Resolution approaches discussed for entity resolution in databases, have been applied to social networks, however have exploited only profile attribute category and other attributes of a user. To explore the potential of other attributes sets and to analyze them with better granularity e.g. Network and Content, social network specific resolution methods are suggested. We first discuss the approaches suggested to exploit profile attributes of an identity with resolution techniques suggested primarily for databases.

[1] Unless a user intentionally obfuscate her attributes.

4.1.1 Profile Resolution

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 [11, 12]. Therefore, researchers suggested to use profile attributes as a strong feature to resolve identities of a user [13–17]. A major challenge to use profile attributes for entity resolution 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 across social networks could be public while others 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 (see Figure 4.1). Comparisons are mostly made using syntactic and semantic methods [4, 18, 19]. We now discuss each of the profile attribute, considered for identity resolution across online social networks.

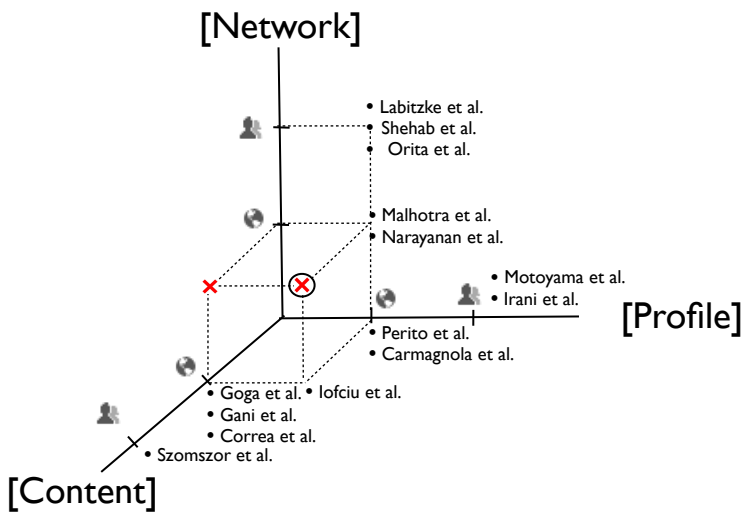

Figure 4.1: A comprehensive view of Identity Resolution Methods

Username attribute

Username, also termed as pseudonym, and screen name, is a publicly accessible profile attribute of a user on a social network. Username is an attribute of an online 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 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 [11, 12]. String comparison metrics (e.g. Jaccard coefficient), n-gram similarity matching, and Markov chain probability are few metrics discussed in database entity resolution literature to understand if any two usernames are same or are variant of each other [1, 13, 17]. String comparison metrics compare two strings (e.g., usernames) character by character. N-gram similarity score is calculated by matching the n-grams extracted from

each username. Markov chain probability measures the probability of one username being the variant of other. The motivation to use string based features of a username is to accommodate user behavior to use a "variant" of her username in case of unavailability of the same username on other social network or by choice, accommodate spelling mistakes and capture different ways of writing a same string. The state of the art accuracy of usernames to resolve online identities on Google and Ebay is 71% [13].

AboutMe Attributes

AboutMe attributes of a user on a social network are the attributes that describes the remaining characteristics of a user apart from username. For example, name, education, school, email, picture, description (bio), city, work, etc., are AboutMe attributes. Such attributes can also be used for identity resolution across 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 [12]. AboutMe attributes of two identities can be compared either on the basis of syntactic similarity methods [14–17] or semantic similarity methods [18]. 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. Further, research shows that few AboutMe attributes are more trustworthy and user-specific than other attributes [16]. Researchers argue that such attributes can be a strong indicator of a positive match. On the other hand, certain attributes are more distinguishing and are termed as strong indicator of a negative match. Examples of strong match indicator are homepage, work attribute, etc., which if matched for both identities then both identities belong to same user. Examples of a negative match indicator are 'gender', 'city' which if *not* matched for both identities, then both identities belong to different users. Experimented with different social networks, different set of AboutMe attributes are concluded to be strong identity resolvers. For example, Motoyama *et al.* resolve a user's Facebook and MySpace identity and observed 'name', 'name + school' and 'email' as strong identity resolvers, Irani *et al.* experimented with fifteen social networks and conclude 'last name', 'birth year' and 'country' as strong identity resolvers and linkers while Malhotra *et al.* experimented with Twitter and LinkedIn identified and concluded 'name' and 'username' as strong identity resolvers. The state of the art accuracy is achieved with AboutMe attributes to revolve identity on Flickr and MySpace, is 88% [16].

We observe that profile attributes are effective in resolving identities of a user who reuses her profile attributes across social networks. However the assumption is not valid to 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 needn't capture. Further few profile attributes may be present on one social network, while may not be present on other, thereby leaving no common metric to compare for two users. Further, the challenge still dominates when profile based resolution methods are applied to *any* set of social networks. To meet the challenge, researchers then suggested other identity resolution methods which exploit features specific to online social networks. They suggested the exploitation of content and network features available for an entity in social networks but not available in databases, and devised identity resolution techniques.

4.2 Online social network methods

Resolution methods specific to online social networks involves a different set of features i.e. content and network, however the basis similarity metrics for comparison still remains limited to syntactic and semantic matching methods. We now discuss how each attribute is used for identity resolution.

4.2.1 Content Resolution

In this section, we discuss the identity search and resolution methods devised on the basis of content created by or is shared with the user. Content type may vary from one social network to other. For example, in Youtube, content implies a video while on Twitter it implies the text or URL. For text based social networks (e.g., Twitter and Facebook), researchers have tried to devise methods to match content created by two identities to infer if they belong to same individual. For tag-based social networks (e.g Flickr and Delicious), approaches exploit tags to devise identity resolution methods.

Any kind of content created by a user posses certain characteristics – author (who created the content), timestamp (time of creation of the content), location (from where the content is created), description (what is the content), tags describing the content and other stylistic features of the content itself (e.g., use of short words, long sentences, etc.). Each of these features can be used to compare content created by two identities to resolve. Few research work used content features for identity resolution, since they noticed users exposing overlapping content attributes across social networks [34, 53–55].

For text-based networks like Twitter and Facebook, three content characteristics have been studied extensively to resolve online identities across social networks – Timestamp, location and description. The similarity between timestamps of content, language profiles (author's writing style from description of content), and location profiles (zip codes of places a user geotagged in the content), resolved 94.7% of Twitter and Yelp identities [34]. Extensive use of authorship analysis techniques (extract lexical features, syntactic features and idiosyncratic features) and additional use of friends network attribute, further helped in identity resolution [53].

Apart from authorship analysis of the content, researchers have proposed the use of other features from content e.g. URLs. 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 [56]. Using that, Correa *et al.* tried to find a user's multiple identities on online social media using Twitter and tried to link Twitter account with them. The authors monitored 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 content generated by her on Twitter. 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.

For social tagging networks like Delicious and Flickr, tags created describing the content 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 kind of content on both the networks and tag it in the same way. The approach is evaluated on Flickr, Delicious and StumbleUpon and gives an accuracy of 60%. Furthermore, as suggested by Irani *et al.*, the authors also implemented the concept of social footprint [15]. To search for a user's identity on a third network, they aggregated two

resolved identities and used the collective information to link the third identity with the social footprint of the user. When evaluated on the same dataset, the authors claimed to link 80% of the user identities across three networks correctly [54]. 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 [55].

4.2.2 Network Resolution

Multiple identities can also be linked via network attributes. In literature, 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 [57, 58]. Narayanan *et al.* used a graph theoretic approach to de-anonymize Twitter users with the use of labelled Flickr network [57]. The 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.

Labitzke *et al.* followed a different approach of matching mutual friends between two identities (to be matched). The 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) [59]. 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 other information sources to infer if two identities refer to the same entity. Such third party information source may 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 [60].

4.2.3 Search Engine Resolution

Bilge *et al.* used a "search and link" approach to link multiple identities. The 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. If the top three results turned out same, then the two identities belong to same entity [60].

4.2.4 CrowdSourcing Resolution

Parallel to resolving records in databases with human intelligence, resolving multiple identities of an entity across social networks has been proposed [61]. 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.

All the above mentioned techniques rely on the assumption that the user doesn't actively obfuscate / hide her correct social network attributes to avoid detection. However, the assumption does not hold true for malicious users present on online social networks. Malicious users also termed as criminal entities, are the users who misuse the social network platform to spread false / untrusted information and gain from the same. Each online social network declares strict policies against malicious users and deploy multiple strategies to detect and suspend malicious user identities. To avoid detection and therefore suspension, malicious users hide behind the false information and fake attributes declared by them on an online social network. To detect such malicious identities, researchers have devised methods for each social network [26–31], however no research has proposed techniques to resolve multiple malicious identities and identified features common for a criminal entity in different contexts. However, there has been some work in resolve criminal identity records in databases. In the next chapter, we discuss the entity resolution methods for malicious users in databases and within online social networks.

Paper	Social network	Dataset size	Search attr.	Match attr.	Match method	Acc.	Precision	Recall
Perito et al.	Google & ebay	10,000	-	Profile	Syntactic	71%	-	-
Carmagnola et al.	Myspace & Flickr	300	Profile	Profile	Probabilistic	-	86.7%	59%
Goga et al.	Yelp, Twitter & Flickr	-	Profile	Profile & Content	Syntactic & Probabilistic	29.8%	-	-
Narayanan et al.	Flickr & Twitter	27000	-	Network	Graph based	30.8%	-	-
Malhotra et al.	Twitter & LinkedIn	29129	-	Profile & Network	Syntactic	98%	96%	99%
Labitzke et al.	StudiVZ, Facebook, Myspace, Xing	300	-	Network	Syntactic	-	-	-
Motoyama et al.	Myspace & Facebook	900	Profile	Profile & Network	Syntactic	72%	-	-
Szomszor et al.	Delicious & Flickr	502	-	Content	Syntactic	-	-	-
Irani et al.	Myspace, Twitter, Blogspot, Delicious, Digg, Facebook, Flickr, Last.fm, LinkedIn, LiveJournal, Technorati, Tumblr, Wiki, Youtube	12674	-	Profile	Syntactic	-	-	-
Iofciu et al.	Flickr, Delicious & Stumbleupon	1788	-	Profile & Content	Syntactic	60 - 90%	-	-
Shehab et al.	Facebook, Myspace & Twitter	7411	Network	Profile	Crowd-sourcing	100%	-	-
Correa et al.	Flickr, Twitter, Foursquare, Youtube, & Last.fm	69,496	-	Content	Syntactic	100%	-	-

Table 4.1: A comparison of the existing methodologies for entity resolution in online social networks.

Chapter 5

Malicious Entity Resolution Methods

Entity resolution methods discussed for databases and online social networks assume that entities do not obfuscate their attributes or do not report their false or manipulated attributes. However, this may not be true always. Various police departments and security agencies store huge amount of data about the criminal and malicious entities. Such malicious users intend to change their identity and therefore their attributes intentionally over time and with each crime. Criminal databases may contain different information about the same criminal distributed over multiple tables. It is therefore necessary to identity which records belong to the same criminal. Traditional approaches discussed above are successful given slight variations in an entity's attribute values, which may not be true for criminal entities. Therefore, researchers introduced a different set of approaches to resolve criminal and malicious identities in a database.

Criminal entities may obfuscate their identity attributes however it is difficult to hide their behavioral characteristics, events in which they are involved and people they are connected to. Researchers suggested a set of features to capture the same i.e. role based personal features, social contextual features and social links [62]. Role based personal features extract the role of a criminal in an event i.e. suspect, arrestee, victim, social contextual features capture the social circle of the criminal i.e. what group the criminal belongs and social links capture the social network of the criminal. In criminal databases, extracting role based personal features is easy, however social contextual features are extracted via transitive metrics i.e. by first understanding the role based features for each criminal in the group, and then assigning the group characteristic to the criminal. Social links between any two criminals is extracted by understanding if two criminals appear together in the same crime incident (doing the crime together). Each elaborate and behavioral characteristic extracted for each criminal record pair is further compared and the pair is then given a score. Researchers proved that with the inclusion of social contextual features, entity resolution precision increases. However, researchers proved that role based personal features and group comparison between two criminals may lead to false negatives. Large number of cases appeared where different criminals have zero distance between their role based personal features and their group affiliations. Contextual features, on the other hand, proved to be better distinguishing feature between matching and non-matching criminal records.

To the best of our knowledge, no research work focus on resolving multiple identities of malicious users across social networks. Lately, researchers have started drifting towards resolving multiple identities within an online social network in either of the two classes – fake or legitimate. Due to loose privacy

settings of a user, and aggregation of a user's multiple accounts on multiple social networks following different privacy policies, private attributes can be leaked. An attacker can apparently exploit public and private attributes of a user to create a fake profile of the user [60]. Such identity attacks on a legitimate social network user are further used for variety of malicious purposes. For example, fake accounts misuse the established trust links between the victim user and her friends to spread false information which largely gets accepted owing to the trust built in victim's friends on her and the fake account created by malicious user gets better hit rate on the false information. Further, famous celebrities and popular personalities face the problem of multiple identities created on their name [63].

To approach this problem, researchers have suggested to verify a user's social network identity from her webpage [64]. The authors assume that legitimate users maintain websites / webpages and list their correct online social network identity on their websites. The assumption holds true for majority of the celebrities, and popular brands because most of them maintain websites for publicity listing their legitimate online social network accounts on their webpages. However, it might not hold true for all online social network users. Another approach suggested by [65] detect fake user accounts on the basis of IP address used by a fake and a legitimate user identity. If a social network user suspects a friend's fake account, the approach can help her to check for the IP addresses used by seemed legitimate account and suspected fake account. The approach is effective, however not applicable and helpful a normal user to verify fake accounts in her network, since IP addresses are not publicly accessible for any online social network user. To overcome this problem, researchers suggested another approach to help a user to list her possible fake identities on a social network [66]. Authors proposed three subsystems in order to find a user's fake accounts given her legitimate account– Information Distiller, Profile Crawler and Profile Verifier. Information Distiller extract user-specific (rare) attributes to search for her fake identities. The reason for using rare attributes is to return identities highly similar to the user and to avoid suspecting legitimate user identities with similar information as fake. The user-specific attributes are pipelined to Profile Crawler, which then search within social network to find possible identities of a user. The returned fake identities are then ranked on the basis of their similarity with the legitimate user account.

5.0.5 Social networks and Digital Forensics

Other approaches have been discussed by digital forensic researchers, not directly applicable but if tweaked, can be exploited to search for fake or malicious users on online social networks. Certain forensic methods and tools have been proposed by Chang *et al.* to retrieve specific user profile and contextual information e.g. IP address, MAC address, which can be used to compare fake versus a legitimate online identity [67]. Authors suggested to install visitor tracker on a user's social networking profile (given that online social network supports the same) and collect detailed statistics about user sessions, cookies, IP address, etc. To further retrieve hidden and detailed information about an online social network user, forensic experts suggested frameworks to extract information from social networks [68]. Huber *et al.* implemented an automated web-crawler along with a third party application to retrieve maximum data from an online social networking account (experimented on Facebook). They argued that a browser based web-crawler activated with a user's (stolen / identified) credentials or authentication cookies, would act as a human, and therefore most information available to the user would be available to the application. In this way, extracted information can then be used to create a social snapshot of a user. Such a methodology can be extended to match social snapshots of two entities, however require access to user authentication tokens / credentials which may not be available. Researchers then proposed various ways to analyze data via visualizations e.g. social-interaction graph, social-interconnection graph, timeline, etc [69]. Social-interconnection graph is an undirected graph where nodes represents the user and her friends and edges represents the friend connection between the nodes. A node with high degree imply a node with maximum friends in the network. Social-interaction graph is a directed graph where there exists an edge between two nodes given that two nodes have communicated (direct message, wall posts, likes, etc.) at least once. Timeline visualization can help the forensic investigator in understanding the activity of the user. Such methods may reveal detailed information about the interactions of the users within social network and with other users and can further be co-related with phone records database to verify the strong connections between real-world users. Such forensic analysis of users may be used to resolve if two profiles (users) exhibit similar patterns of activity and connectivity and if so, can be inferred as they refer to the single real-world entity. Further, it may also be used to infer fake profiles vs legitimate profiles since legitimate profiles tend to keep similar offline and online friend networks and therefore, a user with mapped online interactions with offline phone records are more legitimate and real as compared to others.

Inferences made by detailed and forensic analysis result in lower false positives and false alarms, however one major disadvantage of the forensic approaches discussed is that they require and assume active support from online social networking site involved as well as the user whose data is collected, since authors assume they have access to user credentials while collecting detailed information about the user, which might not be true in case of malicious identities to track or any suspected user.

Chapter 6

Entity Resolution Tools

Many commercial organizations have developed tools to locate an entity's multiple references across databases and social networks. Note that, any entity (person, place or organization) can be searched by the tools described in Figure 6.1. Each tool inputs an entity and its few known attributes and searches for the similar entities across different databases and online social networks.

Features / Identity Search System	Query attributes	Returned attributes	Search space*	Attribute Category	Scalability	Real-time	API Support
Pipl ['06 - '13]	Username,Name, Email, Phone, Location,Age	Address, Images, Social Profiles, Related People	Vital Records, criminal records, court records, adoption records, phone directories, publications, social profiles, blogs, address, emails	Profile	Yes	Yes	Yes
Yasni ['07 - '13]	Username, Name, Birthname, Location, Phone, Hobby, Company	Related tags, Occupation, Address, Social Profiles, Images	Yellow page directory, Doctors directory, Social profiles, E-commerce networks, news, books, publications, audio, video	Profile	Yes	Yes	No
Peekyou ['06 - '13]	Username, Name, Email, Phone, Location, Age, School, Business, City, Interests	Public Records, Social profiles, Spokeo, Phone, email address, news, patents, books, images	Public Records, Social profiles, Spokeo, Phone directory, email address, web, news, patents, books, images	Profile, Network	Yes	Yes	No
Spokeo^ ['10 - '13]	Username, Name, Location, Address, Phone, Email	Name, Address, Phone, Gender, Email, Photos, Family Tree, Social Profiles	All Public Sources	Profile	Yes	Yes	No
123people ['07 - '13]	Name, City, Location	Social Profiles, Email address, Business professionals, phone, IMs, documents	Blogs, Social profiles, news, videos, documents, IMs, Web, Criminal Records, Phone directory	Profile	Yes	Yes	No
OCEAN ['13]	Name, City, Location	Name, PAN Card, Driver's License, Family Tree, Home address, Social	E-government sites, Social sites	Profie, Network	Yes	Yes	No

Figure 6.1: A comparison of open Identity Search and Resolution systems

We compare entity resolution tools in terms of the query attributes needed to initiate a entity search,

entity attributes the tool outputs, search space of the tool, category of the attributes of an entity the tools searches with, the tool efficiency in terms of scalability, timely response and further usage. We observe that tools query with only certain profile attributes however, do not take into account content or network attributes to search for any queried entity. Most of the tools are scalable and return the query results within few seconds. Some results are more surprising since some tools returns old online group posts or pictures which are already taken down by the user. Efficient (unknown) methods have been implemented by the tool developers to parse all public sources in a standard format, store, search and return results within limited time. All tools are commercial and do not share data or provide any API, which hinders the future research and data availability to the research community.

Chapter 7

Discussion

Literature has used multiple terms for the entity resolution process both in databases and in online social networks, making it difficult for the new researchers in creating a comprehensive bibliography. We make a first attempt to summarize entity resolution methods specifically in the domain of online social networks, to help the researchers to create an absolute understanding of the area, and further comprehend the research gaps where they can contribute. We observe that entity resolution techniques in databases are matured however entity resolution techniques in online social networks are still developing. There is no foolproof algorithm or method proposed till now, which can be applied on any user to locate her multiple identities and connect them. Researchers have a scope here to contribute with better formal and better entity resolution techniques. We further argue towards a collaboration of multi-disciplinary researchers (graph-theoristicians, social scientists) to contribute to the research area, since the problem involves identities (references) of a real-world entity and huge networks and therefore it is important to understand behavioral characteristics and thought processes behind creating multiple varied identities across huge graph networks. We think that standard terminologies should be used to relate to entity resolution techniques therefore making it easier to search and relate to the existing literature, further make their datasets public, to allow validation of their research and usage to derive better understanding and analysis.

Bibliography

[1] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, 2007.

[2] D. G. Brizan and A. U. Tansel, "A Survey of Entity Resolution and Record Linkage Methodologies," 2006.

[3] S. Das, E. I. Chong, G. Eadon, and J. Srinivasan, "Supporting ontology-based semantic matching in RDBMS," in *Proceedings of the Thirtieth international conference on Very large data bases*, ser. VLDB, 2004.

[4] A. Doan and A. Y. Halevy, "Semantic-integration research in the database community," *AI Magazine.*, 2005.

[5] H. Köpcke and E. Rahm, "Frameworks for entity matching: A comparison," 2010.

[6] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "CrowdER: crowdsourcing entity resolution," *Proceedings of VLDB Endowment.*

[7] S. E. Whang, P. Lofgren, and H. Garcia-Molina, "Question selection for crowd entity resolution," Stanford University, Tech. Rep., 2012.

[8] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller, "Human-powered sorts and joins," ser. VLDB, 2012.

[9] G. Demartini, D. E. Difallah, and P. 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*, ser. WWW, 2012.

[10] I. Bhattacharya, L. Getoor, and L. Licamele, "Query-time entity resolution," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD, 2006.

[11] D. Correa, A. Sureka, and R. Sethi, "WhACKY! - What anyone could know about you from Twitter," in *PST*, 2012.

[12] T. Chen, M. A. Kaafar, A. Friedman, and R. Boreli, "Is More always Merrier?: a Deep Dive into Online Social Footprints," in *Proceedings of the ACM Workshop on online social networks*, ser. WOSN, 2012.

[13] D. Perito, C. Castelluccia, M. A. Kâafar, and P. Manils, "How Unique and Traceable Are Usernames?" in *PETS*, 2011.

[14] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," in *Proceedings of the eleventh international workshop on Web information and data management*, ser. WIDM, 2009.

[15] D. Irani, S. Webb, K. Li, and C. Pu, "Large Online Social Footprints–An Emerging Threat," in *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 03*, ser. CSE, 2009.

[16] F. Carmagnola, F. Osborne, and I. 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*, ser. HetRec, 2010.

[17] A. Malhotra, L. Totti, W. Meira, P. Kumaraguru, and V. Almeida, "Studying User Footprints in Different Online Social Networks," *International Workshop on Cybersecurity of Online Social Network (CSOSN)*, 2012.

[18] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh, "Discovering semantic equivalence of people behind online profiles," in *In Proceedings of the Resource Discovery (RED) Workshop*, ser. ESWC, 2012.

[19] J. Golbeck and M. Rothstein, "Linking social networks on the web with FOAF: a semantic web case study," in *Proceedings of the National conference on Artificial intelligence* -, ser. AAAI, 2008.

[20] D. Bitton, C. Turbyfill, D. J. Dewitt, and D. J. Dewitt, "Benchmarking database systems: A systematic approach," ser. VLDB, 1983.

[21] "Neilsen Social Media Report," 2011. [Online]. Available: http://blog.nielsen.com/nielsenwire/social/

[22] "Twitter Popularity Higher Than Ever in 2011," 2011. [Online]. Available: http://searchenginewatch.com/article/2095216/Twitter-Popularity-Higher-Than-Ever-in-2011-Check-Out-These-Stats-Cool-Uses

[23] "August 2011 Top US Web Brands," 2011. [Online]. Available: http://blog.nielsen.com/nielsenwire/online_mobile/august-2011-top-us-web-brands/

[24] "Facebook now has 901 million users," 2012. [Online]. Available: http://mashable.com/2012/04/23/facebook-now-has-901-million-users/

[25] E. Quayle and M. Taylor, "Social networking as a nexus for engagement and exploitation of young people," *Information Security Technical Report*.

[26] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: the underground on 140 characters or less," in *Proceedings of the ACM conference on Computer and communications security*, ser. CCS, 2010.

[27] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in Information Retrieval*, ser. SIGIR, 2010.

[28] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Proceedings of the Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[29] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: human, bot, or cyborg?" in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC, 2010.

[30] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves, "Detecting spammers and content promoters in online video social networks," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.

[31] D. Irani, S. Webb, and C. Pu, "Study of Static Classification of Social Spam Profiles in MySpace," in *ICWSM*, 2010.

[32] E. Zheleva and L. Getoor, "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles," in *Proceedings of the 18th international conference on World wide web*, ser. WWW, 2009.

[33] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," 2010.

[34] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, RobinSommer, and R. Teixeira, "On exploiting Innocuous User Activity for Correlating Accounts across Social Network Sites," 2012.

[35] V. I. Levenshtein, *Soviet Physics Doklady*, 1966.

[36] M. Waterman, T. Smith, and W. Beyer, "Some biological sequence metrics."

[37] M. A. Jaro, *Unimatch: A record linkage system: Users manual*. Bureau of the Census, 1978.

[38] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences." *Journal of molecular biology*, 1981.

[39] E. Ukkonen, "Approximate string-matching with q-grams and maximal matches," *Theoretical Computer Science*, 1992.

[40] W. W. Cohen, "Integration of heterogeneous databases without common domains using queries based on textual similarity," in *SIGMOD International Conference on Management of Data*, 1998.

[41] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: a generic approach to entity resolution," *The VLDB Journal*, 2009.

[42] M. Bilgic, L. Licamele, L. Getoor, and B. Shneiderman, "D-dupe: An interactive tool for entity resolution in social networks," in *IEEE VAST*, 2006.

[43] "TAILOR: A Record Linkage Tool Box," in *Proceedings of the 18th International Conference on Data Engineering*, ser. ICDE '02, 2002.

[44] A. Thor and E. Rahm, "MOMA - a mapping-based object matching system," in *CIDR*, 2007.

[45] M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '03, 2003.

[46] H. Köpcke and E. Rahm, "Training selection for tuning entity matching," in *QDB/MUD*, 2008.

[47] Parag Singla and Pedro Domingos, "Entity Resolution with Markov Logic," in *ICDM*, 2006.

[48] Chen, Zhaoqi and Kalashnikov, Dmitri V. and Mehrotra, Sharad, "Adaptive graphical approach to entity resolution," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ser. JCDL '07, 2007.

[49] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data (T-KDD)*, 2007.

[50] Bhattacharya, Indrajit and Getoor, Lise, "Online collective entity resolution," in *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, ser. AAAI'07, 2007.

[51] On, Byung-Won and Lee, Ingyu and Lee, Dongwon, "Scalable clustering methods for the name disambiguation problem," *Knowledge and Information Systems*, 2012.

[52] P. Christen and R. Gayler, "Towards scalable real-time entity resolution using a similarity-aware inverted index approach," in *Proceedings of the 7th Australasian Data Mining Conference - Volume 87*, ser. AusDM '08, 2008.

[53] K. Gani, H. Hacid, and R. Skraba, "Towards multiple identity detection in social networks," in *Proceedings of international conference companion on World Wide Web*, 2012.

[54] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying Users Across Social Tagging Systems," in *ICWSM*, 2011.

[55] M. Szomszor, I. Cantador, E. P. Superior, and H. Alani, "Correlating user profiles from multiple folksonomies," in *In Proceedings of International Conference Hypertext (HT '08)*, 2008.

[56] P. Jain, T. Rodrigues, G. Magno, P. Kumaraguru, and V. Almeida, "Cross-Pollination of Information in Online Social Media: A Case Study on Popular Social Networks," in *SocialCom/PASSAT*, 2011.

[57] A. Narayanan and V. Shmatikov, "De-anonymizing Social Networks," in *Proceedings of IEEE Symposium on Security and Privacy*, ser. SP, 2009.

[58] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: using social network as a side-channel," in *Proceedings of the ACM conference on Computer and communications security*, ser. CCS, 2012.

[59] S. Labitzke, I. Taranu, and H. Hartenstein, "What your friends tell others about you: Low cost linkability of social network profiles," ser. SNAKDD, 2011.

[60] L. Bilge, T. Strufe, D. Balzarotti, and E. 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*, ser. WWW, 2009.

[61] M. Shehab, M. N. Ko, and H. Touati, "Social networks Profile Mapping using Games," in *Proceedings of the 3rd USENIX conference on Web Application Development*, ser. WebApps, 2012.

[62] G. A. Wang, J. J. Xu, and H. Chen, "Using Social Contextual Information to Match Criminal Identities," *Hawaii International Conference on System Sciences*, 2006.

[63] "Twitter to block six fake PM accounts," 2012. [Online]. Available: http://articles.timesofindia.indiatimes.com/2012-08-23/social-media/33340785_1_fake-twitter-accounts-pachauri-communications-adviser

[64] A. Orita and H. Hada, "Is that really you?: an approach to assure identity without revealing real-name online," in *Proceedings of the 5th ACM workshop on Digital identity management*, ser. DIM, 2009.

[65] J. L. C. Y. A. L. Zifei Shan, Haowen Cao, "Enhancing and Identifying Cloning Attacks in Online Social Networks," ser. ICUIMC, 2013.

[66] G. Kontaxis, I. Polakis, S. Ioannidis, and E. P. Markatos, "Detecting social network profile cloning," in *PerCom Workshops'11*, 2011.

[67] Janet Cheng and Jennifer Hoffman and Therese Lamarche and Ahmet Tavil and Amit Yavad and Steve Kim, "Forensics tools for social network security solutions," 2009.

[68] Huber, Markus and Mulazzani, Martin and Leithner, Manuel and Schrittwieser, Sebastian and Wondracek, Gilbert and Weippl, Edgar, "Social snapshots: digital forensics for online social networks," in *Proceedings of the 27th Annual Computer Security Applications Conference*, ser. ACSAC '11, 2011.

[69] M. Mulazzani, M. Huber, and E. Weippl, "Social network forensics: Tapping the data pool of social networks," in *Eighth Annual IFIP WG 11.9 International Conference on Digital Forensics*, 2012.