International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Identity & Relationship Resolution Approach Supported with Sample and Real World 9/11 Case Study to Show Identity Mapping

Arvind Kumar[1]

[1]IT Program/Project Management Capacity and Worked for Various Major IT Companies and Clients,
PMI Certified Project Management Professional,
Fellow Member of Professional Body - IETE (The Institution of Electronics and Telecommunication Engineers)

Abstract - *Duplicate and false identity records are quite common in identity management systems due to unintentional errors or intentional deceptions. Identity resolution is to uncover identity records that are co-referent to the same real-world individual. In this paper we introduce a COTS (Commercial off-the-self) product based solution of identity resolution that discovers individual's identity and discovers relationships of individuals to other individuals or entities by analyzing disparate sets of data. Solution would be capable of determining relationships with individual and would provide the features like Semantic resolution of identities (means that heuristics are used to match records that refer to the same person, even when their name/address/social are misspelled, or have transposed digits, etc., Anonymous data sharing (ensures privacy and security), Non-obvious relationship, detection-including out to 30 degrees of separation, Generates alerts ensuring real-time actionable intelligence and Risk rating based upon its relationship.*

Key Words: — Identity, Identity Resolution, Social Relationship, Entity Resolution

1. INTRODUCTION

THE world is moving towards major shift of digitization from paper-based documents and recording everything as electronic records. Due to the ease of generating identity records and lack of sufficient verification or validation during data entry processes, duplicate and false identity records become quite common in electronic systems. In many practices of identity management and collection processing systems, especially those that require integrating multiple data sources, it is often inevitable and tedious to deal with the problem of identity duplication. In banking system, every day, we do number of payment transactions but we cannot identify how particular transaction used- whether it is used for personal use or to feed some sort of terrorism. Particularly, finding an effective solution to this problem is extremely critical in fighting crime and terrorism to enforce national security. Criminals and terrorists often assume fake identities using either fraudulent or legitimate means so as to hide their true identity. In a number of cases documented by government reports, terrorists in different countries are known to commit identity crimes, such as falsifying passports and baptismal certificates, to facilitate their financial operations and execution of attacks, either in the real world or in the cyber space [1, 2]. The problem of an individual having multiple identities can easily mislead intelligence and law enforcement investigators [3]. Therefore, to determine whether an individual is who they claim to be is essential in the mission of counterterrorism to identify potential terrorists and prevent terrorism acts from occurring [4]. Identity resolution is a process of semantic reconciliation that determines whether a single identity is the same when being described differently [5]. The goal of resolution is to detect duplicate identity records that refer to the same individual in the real world. Over the years, researchers in the areas of database and statistics have proposed many different techniques to tackle this problem. Traditional resolution techniques rely on key attributes such as identification numbers, names and date-of-birth to detect matches. These attributes are commonly used for they are simple describers of an individual and often available in most record management systems [6, 7]. However, such personal identity attributes also vary in terms of availability and reliability across different systems. These attributes are not always accurate due to various reasons such as unintentional entry errors and intentional deception [6]. In the context of cybersecurity, it is much easier and more common for criminals to fake identities to cover their traces.

The remainder of the paper is organized as follows. We first review identity theories and concept. Next, we introduce proposed resolution framework based on COTS product. We report benefits of solution using sample case study along with real 9/11 case study (based on facts available on public forums). Finally, we conclude the paper with a summary.

2. IDENTITY THEORIES AND CONCEPT

In this section, we review existing identity resolution approaches, with a focus on the identity attributes.

2.1 Entity resolution and identity resolution

Identity resolution is a special type of entity resolution that specializes in identity management. Entity resolution is also known as record linkage and deduplication in the areas of statistics and database management. Record linkage, originated in the statistics community, is used to identify those records in one or multiple datasets that refer to the same real-world entity [8].

Entity resolution has traditionally focused on attribute matching. Attribute matching is an approach whereby one examines the attributes (features) of two identities to see if they match up. However, the events of the world are now increasingly represented in digital form, allowing us to see not only the subjects in which we are interested, but also the objects with which those subjects interact. This means that in addition to having access to the attributes of subjects, we can capture their relationships to other objects, providing a richer view of the environment.

Relational entity resolution has been gaining more attention recent days in terms of study, however the majority of this work on relational entity resolution assumes that one knows the relationships between the entities. There have been multiple efforts [9], [10], [11] in the social media domain that show how entities can be resolved using lists of followers or friends. However, for multiple applications in the real world, the picture is not so neat. Fraudsters don't publish lists of their co-conspirators—their conspirators must be inferred by observing their behavior.

2.2 An Example of Entity Resolution vs. Identity Resolution

Entity resolution and identity resolution are terms associated with record linkage, data matching, and deduplication. Even though these terms are often used interchangeably, there is a subtle, but important difference, in meaning. The difference revolves around the use of the words "entity" and "identity."

In an entity resolution context, the definition of an entity is more narrowly defined than its general use. In entity resolution, "entity" denotes a real-world object that is distinguishable from other objects of the same type, i.e. objects having distinct identities. The same entities are often the subject of master data management (MDM) such as customers, patients, students, organizations, products, events, and locations.

Here is where the nuance comes in with the word "identity." In particular, it is the idea of "known identity" versus "distinct identity." Perhaps this is best illustrated by a simple example related to crime solving.

Suppose there has been a burglary at a business. The police investigate and find a number of fingerprints. When the police laboratory examines the prints, they discover that there are two distinct sets of fingerprints from two different people. We know the burglars are people, and these people have distinct identities, but at this point the police don't actually "know" those identities. However, after sending the fingerprints to the FBI, the burglars are identified because their prints are in a database of previously convicted criminals.

So, let's put this story in the context of data integration. The fingerprints are "references" to the persons who left them. The fingerprints are not themselves persons, but only references to persons. Similarly, the records we create in information systems to describe entities such as customers or patients are simply references to those customers or students, not the actual persons. Entity resolution is the process of determining whether two entity references are for the same entity, or for different entities. So, in our example, what the police did in their local laboratory was entity resolution. They sorted out the fingerprints into two groups referencing two different people.

On the other hand, identity resolution is resolving an entity reference against a collection of known identities. In our crime example, the FBI performed identity resolution. They took the fingerprints from the police and matched them against fingerprints of known criminals, i.e. known identities. Identity resolution is sometimes referred to as "recognition" as in "customer recognition." From this perspective, identity resolution can be considered a special case of entity resolution in which one of the two references being resolved is from a known identity.

2.3 Identity attributes

Based on the identity theories from the social science literature, an individual's identity is considered to have two basic components, namely a personal identity and a social identity. A personal identity is one's self-perception as an individual, whereas a social identity is one's biographical history that builds up over time [12]. In particular, one's personal identity may include personal information given at birth (e.g., name, date and place of birth), personal identifiers (e.g., social security number), physical descriptions (e.g., height, weight), and biometric information (e.g., fingerprint, DNA). In contrast, a social identity is concerned with one's existence in a social context. Social identity theories consist of psychological and sociological views. The psychological view defines a social identity as one's self-perception as a member of certain social groups such as nation, culture, gender identification, and employment [13, 14]. The sociological view focuses on "the relationships between social actors who perform mutually complementary roles (e.g., employer employee, doctor-patient)" [15]. While psychological view deals with large-scale groups, the sociological view emphasizes the role-based interpersonal relationships among people [16]. These two views combined together provide a more complete concept for understanding a social identity at levels of social context. Traditional identity resolution methods primarily rely on personal identity attributes such as name, gender, date of birth, and identification numbers mostly because they are commonly available as identifiers in record management systems.

These attributes, however, may suffer from data quality problems such as unintentional errors [17], intentional deception [6],

and missing data [7]. Biometric features such as fingerprints and DNA also belong to the category of personal attributes. Although they are considered as more reliable, they are not available or accessible due to issues such as high costs and confidentiality in most systems. A study conducted by the United Kingdom Home office [18] suggests that identity crimes usually involve the illegal use or alteration of those personal identity components. The low data quality in fact based personal attributes can severely jeopardize the performance of identity resolution [7]. Individuals are not isolated but interconnected to each another in a society. The social context associated with an individual can be clues that reveal his or her undeniable identity. Recognizing the limitations of personal attributes, many recent studies have started exploiting social context information such as social behaviors and relationships for identity resolution.

3. IDENTITY AND RELATIONSHIP RESOLUTION APPROACH

Solution proposes highly robust and scalable Identity and Relationship Resolution whereby system can connect multiple data sources (through collection processing system) with a view to understand possible identity matches and non-obvious relationships across the reports. Solution would analyze all of the information relating to individuals and/or entities, and would apply likelihood and probability scoring to determine which identities are a match and what, if any, non-obvious relationships exist between those identities.

Identity and relationship resolution engine would be used to uncover risk and fraudulent records by applying business intelligence to the resolved identities and their relationships. Solution would help below-

1) Matching with the pre-defined lists
2) Identify reports which pertain to same entity.
3) Capture all possible relationships.
4) Analyze relationships to form closely linked entities.
5) Perform search to retrieve records with similar parameters.

During the Collection Processing Solution, all the reports would be matched against the list of know entities with higher risks and based on the confidence threshold achieved in the match the report would be flagged for the priority processing.

Proposed solution addresses Identity and Relationship resolution by using IBM Entity analytics solution. The solution helps in identifying and analyzing all the relationships by answering two basic questions

3.1 Who is Who?

Identity Resolution resolves inconsistent, ambiguous identity and attributes information into a single resolved entity across multiple data sets, despite deliberate attempts to confuse or misrepresent individual identity. It does this by matching up information in common between supposedly-different individuals to discover that they are the same person.

3.2 Who knows Who?

Entity Analytics Relationship resolution establishes obscure relationships between distinguishes individual identities – thus improving fraud finding and reporting capabilities thus enhancing the capabilities. Relationship Resolution extends the value of existing information assets by linking unique resolved identities to outside obvious and non-obvious entities to establish relationships to individuals or organizations and uncover social, professional, value, and criminal networks. It does this by finding for example addresses and phone numbers in common between pairs of individuals and then traverses the network of related people to see who is a friend-of-a-friend as it were.

4. ARCHITECTURE OVERVIEW

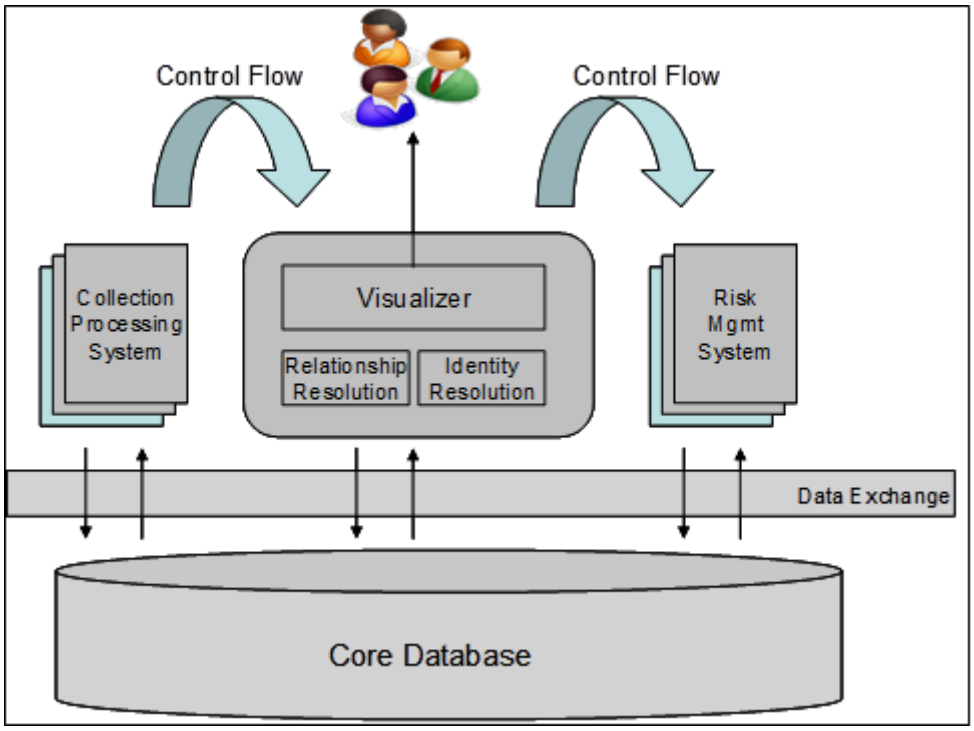
The below fig shows the details and interactions related to the Identity relations Resolution. Solution uses COTS Products for the IRR.

IRR solution is triggered from the Collection processing system. This system collects data from various data sources

Case 1: Any critical data is reported
Case 2: Any Report matches known entity
Case 3: After all the records are processed by the collection processing system
In all the three cases after the Collection processing script is processed. It triggers the script for IRR application.

In Case 1 & Case 2 only the flagged entities are processed on priority and the relationship is established with the known entities and IRR database is enriched. The script only updates the records flagged for priority processing by the Collection processing Application. Script is triggered immediately after the reports are matched against know entities.

After the IRR application enriches the data for all reports designated for priority processing it would trigger the scripts for Risk Management System (not discussed on this paper) to initiate the processing of pre-defined rules to associate risk with each report.



In Case 3 After all the records are processed by the collection processing system the script triggers the Identity and relationship resolution application to identify all the common entities, (legal and individual)and detail the relationship amongst the known entities. Subsequently it triggers the scripts to process the records based on rules in Risk Management Solution.

User Interactions: Users can interact with the IRR results and database via IRR visualization utility which GUI tool

- To configure the parameters for building relationship / clusters,
- Modify the parameters and associated weightings.
- Modify the identified relationships
- Define the degree of relationship to be established amongst the entities.
- Ability to traverse and find relationship through rich visual interface.
- Interface to define and configure search parameters and thresholds for matching names and address
- To access and retrieve the reports, clusters for manual identity and relationship resolution based on parameters
- Ability to search on group and cluster Modes

5. FEATURES

This section covers the key features of the proposed Identity and Relationship Resolution System though a step by step example. We first look at the key features & then move on to the example-

Identity Resolution (identified "Who is Who?")

5.1 Identify reports which pertain to same entity

It resolves inconsistent, ambiguous identity and attributes information into a single resolved entity across multiple data sets, despite deliberate attempts to confuse or misrepresent individual identity. It does this by matching up information in common between supposedly-different individuals to discover that they are the same person.

5.2 Uniquely Identify Single entity from multiple customer

Unlike conventional name and address matching, data quality, or CDI (customer data integration) technologies, which can be used to identify that Mr. A. Kumar is the same as Arvind Kumar, say, because they have the same address and/or telephone number, Identity Resolution resolves data against any of the attributes that pertain to the individual or organization under

consideration. The Identity Resolution process determines whether multiple records, which appear to describe different individuals or organizations, are actually records for a single resolved identity, even with different name variations.

5.3 More than Uniquely Identify Single entity from multiple records

Solution would be designed such that all records are retained separately in Identity Resolution thus it supports multiple views of an individual rather than the conventional single view; or perhaps it would be better to say that it supports every version of the truth. If one thinks in terms of criminals or terrorists, these people often have multiple identities and for system purposes it is important to understand the various aliases that a person may have.

5.4 Ability to modify the weightings to the matching parameters

Proposed solution includes an interface to configure and standardize the weight associated with each parameter used for matching the entities. Proposed Entity analytics Identity Resolution has capability for an advanced heuristic matching technique, designed to identity unique records.

5.5 Threshold Configuration

Thresholds are configured using the Configuration Console and are an integral part of a resolution rule. Thresholds are defined for each matching parameter are instrumental in defining a rule.

5.6 Real-time Report processing

Proposed COTS solution Identity Resolution, supports event stream processing, thus providing the real time capability such that whenever a record is received Identity Resolution immediately compares it with its database to see if it

1) matches an existing record,
2) Is new and requires the creation of a new record (which is done automatically)

5.7 Historical Database Enrichment

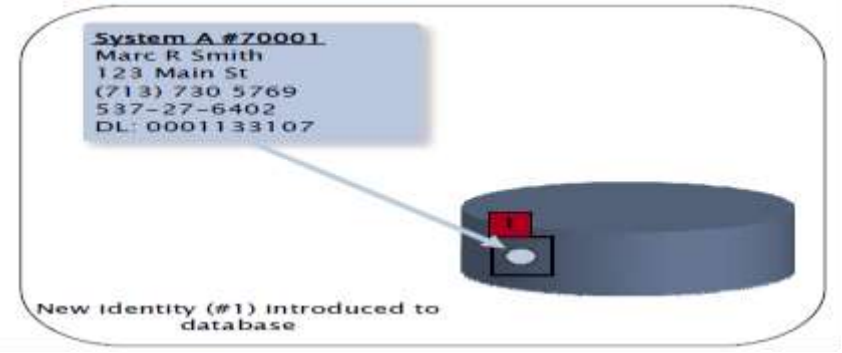
Proposed Identity Resolution solution retains all pertinent records it builds up a history or 'timeless view' as opposed to a snapshot of each individual or company in this database over time. This would result in better analysis along with better resolution process that itself becomes gradually more accurate, because there is more data to make comparisons with. Thus it will have all the entries of the attributes like address, phone number , PAN number etc. which were submitted by for example Arvind Kumar. Thus any time Arvind Kumar hits the system(s), the Identity resolution uses the attribute for resolution. Apart from the resolution benefit, it can also have a complete track of the information of Arvind Kumar over the years.

5.8 Rich library of standard rules, contexts and locale information

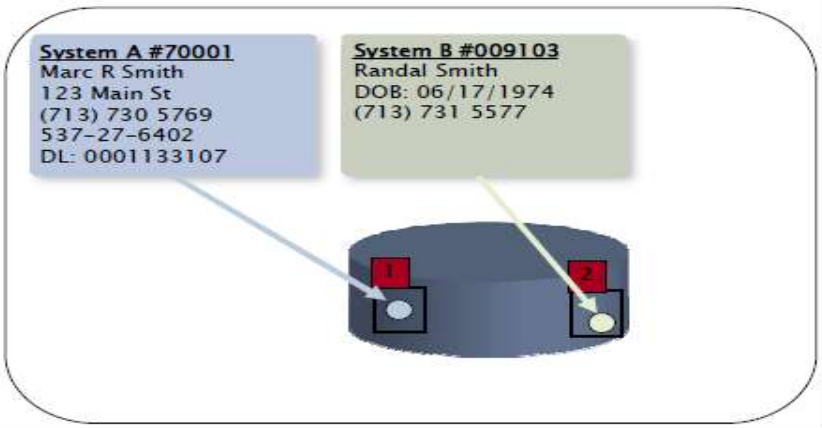
The matching technology in Identity Resolution is excellent. The reason for this is that, in simple terms, it understands context (as an analogy, this is comparable to the use of natural language search mechanisms as opposed to those that are purely statistical)

Identity relationship is explained in details using the following example [19]

Step 1: 1st Record in the Database

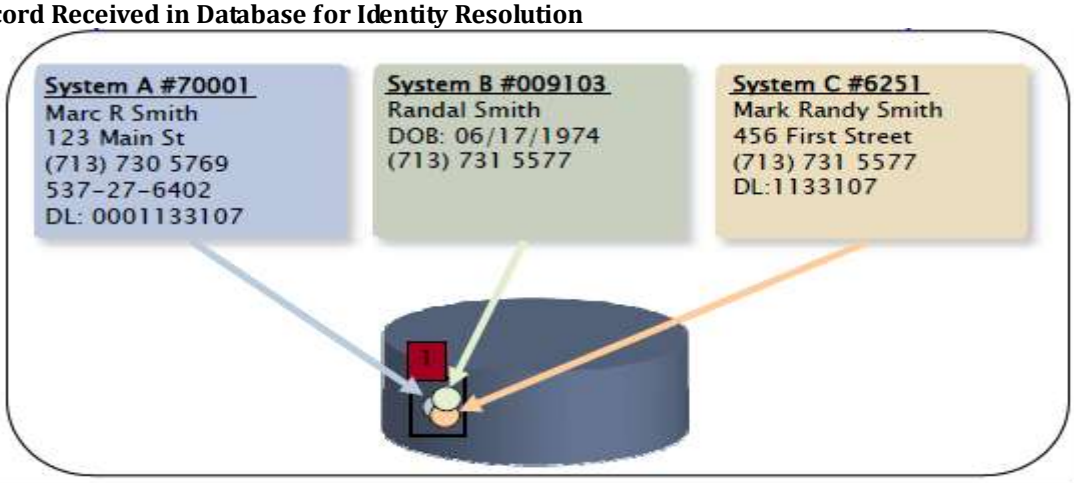


Step 2: 2nd Record Received in Database for Identity Resolution



Inference: No meaning correlation between records

Step 3: 3rd Record Received in Database for Identity Resolution

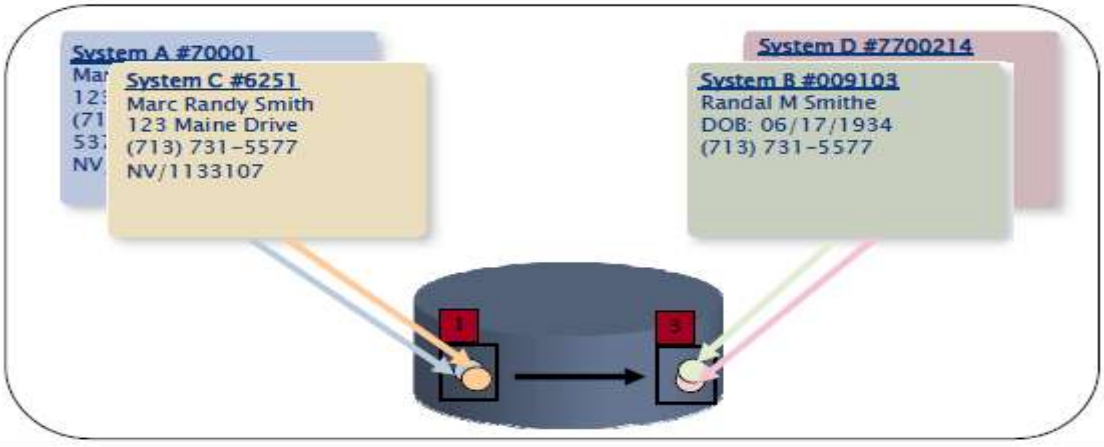


Inference: Identity #2 is identity #1 – Identities are conjoined

Step 4: IRR Engine Accumulates References and enhances database

Names	Marc R Smith	A-#70001
	Randal Smith	B-#009102
	Mark Randy Smith	C-#6251
Address	123 Main St.	A-#70001
	456 First Street	C-#6251
Phones	(713) 730-5769	A-#70001
	(713) 731-5577	B-#009102
	(713) 731-5577	C-#6251
SSN	537-27-6402	A-#70001
DL	0001133107	A-#70001
	1133107	C-#6251
DOB	06/17/1974	B-#009103

Step 5: 4th Record Received



Inference: Identity of #4 established

Relationship Resolution (identified "Who knows Who?")

The use of IBM Relationship Resolution which extends the value of existing information assets by linking unique resolved identities to outside obvious and non-obvious entities to establish relationships to individuals or organizations and uncover social, professional, value, and criminal networks.

It does this by finding for example addresses and phone numbers in common between pairs of individuals and then traverses the network of related people to see who is a friend-of-a-friend as it were. Relation Resolution is superset of Identity resolution.

The relationship resolution would define the relationships by comparing addresses, phone numbers, e-mail addresses, names, surnames and any other characteristics it can discover across multiple individuals. Identity Resolution leverages every identity attribute, it provides Relationship Resolution with multiple linkage points from which it can make connections or discover linkages between unique identities.

Relationship resolution is not just done on the basis of a snapshot (as conventional matching does) but it continues this process on the historical and multiple identities associated with the entity.

Relationship Resolution would be able to do things like determine that Mr. Naval and Ms Smitha live in the same house and share the same phone number. It may or not be able to determine whether they are partners, or that Sharma is the married name of Mr. Ajay' sister, or that they are simply flat-mates, depending on the other data that is available. However, Relationship Resolution is not limited to house holding. For example, you might establish that Mr. Naval and Ms Smitha work in the same office or that they attended the same school. As another example, the software could also determine that your employee, Mr Jason Quilby-Leach, is also the owner of JQL International Ltd from whom you purchase all your left-handed widgets, for example. This may or may not be an issue but you would certainly like to know about it. Similarly, Mr. Naval might be a person working for Prime Minister Security while Ms Smitha is facing charges of National Security.

IBM Relationship Resolution seeks out and integrates non-obvious relationships between service seekers, with historical passport records, transient passport service request data and any internal and external source to determine potential value or potential danger - even if an individual is trying to hide or disguise their identity.

With the help of IBM Relationship resolution product a case study post infamous Sep 11 was conducted. Solution was able to establish the relationship of all the engaged participants within third degree of relationship.

6. IBM Case Study –@ Work – Solving the 9/11 mystery

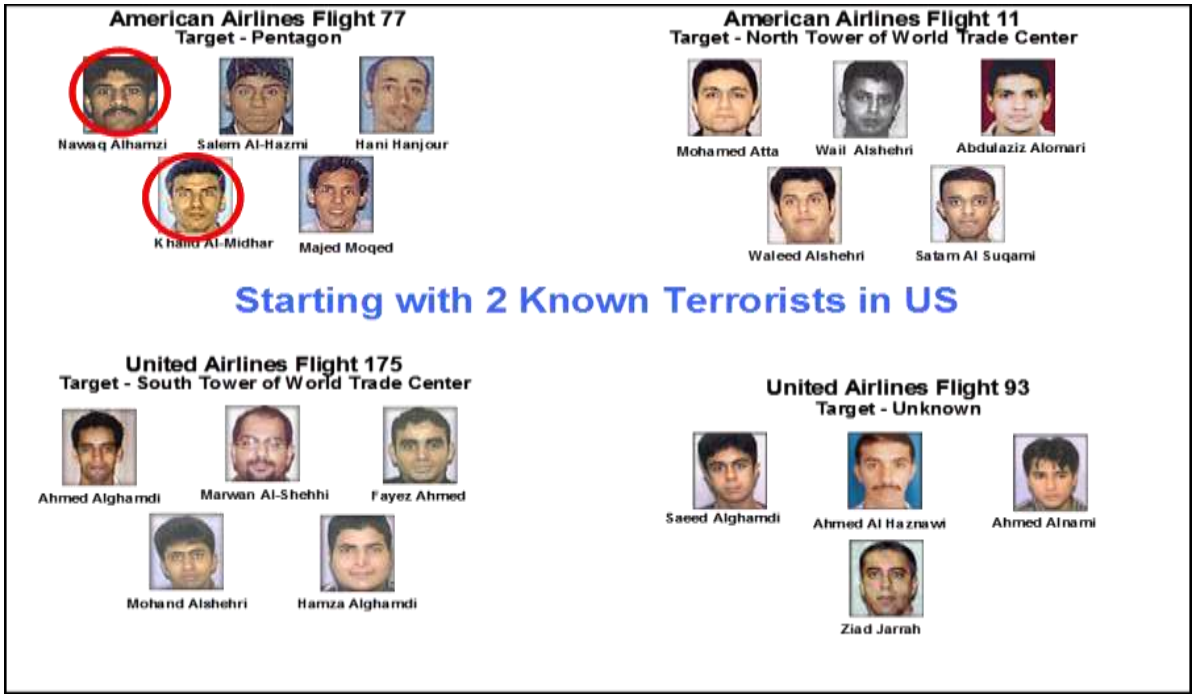
Once the data related to all the known people involved in the 9/11 were made available, IBM conducted a case study [19] to find the relationship between the individuals based on the public data available.

WATCH LIST: CIA/INS/FBI POSSIBLE TERRORISTS IN THE US:

- On or before August 23rd, 2001, Nawaq Alhamzi and Khalid Al-Midhar added to INS watchlist

MAKE PLANE RESERVATIONS USING SAME NAMES:

- On or about August 25, 2001, Khalid Al-Midhar purchases cash ticket for American Airlines flight #77 scheduled for September 11, 2001
- On or about August 27, 2001, Nawaq Alhamzi books a flight on American Airlines flight #77 scheduled for September 11, 2001



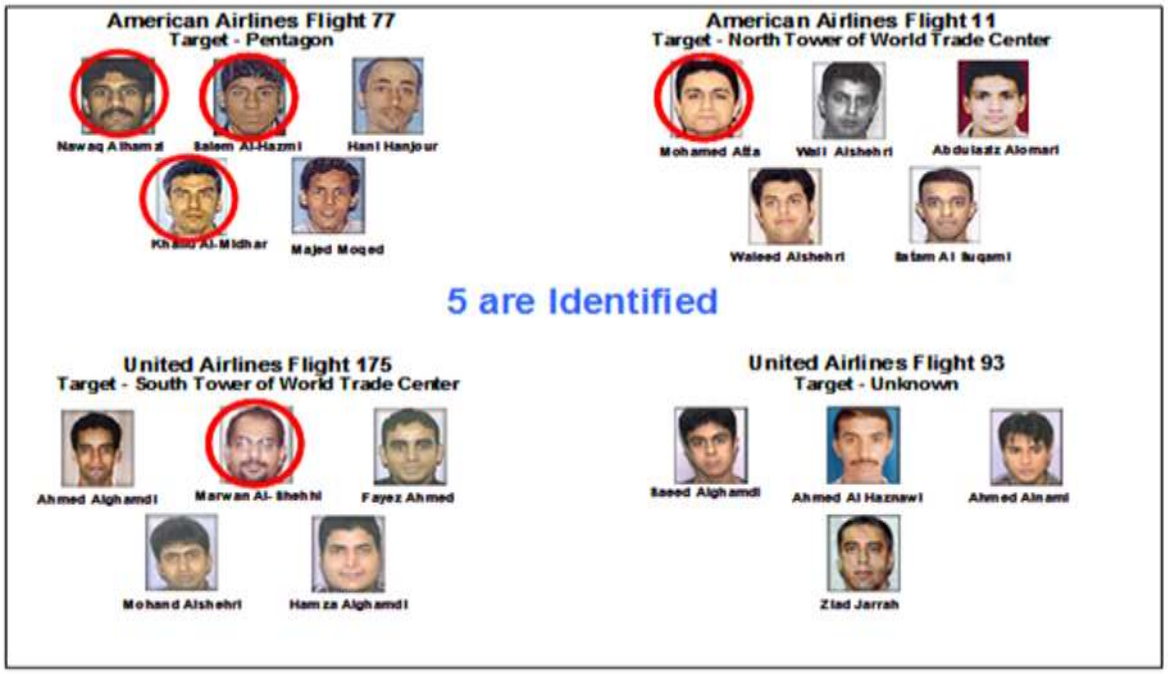
Address Connections

RESERVATIONS MADE WITH ADDRESS #1 AND ADDRESS #2

- On or about August 25, 2001, Khalid Al-Midhar makes a reservation on American Airlines flight #77 scheduled for September 11, 2001 using Common Address #1
- On or about August 27, 2001, Nawaq Alhamzi books flight on American Airlines flight #77 scheduled for September 11, 2001 using Common Address #2

ADDRESSES ARE USED BY THREE (3) ADDITIONAL PASSENGERS

- Mohamed Atta has reservation on American Airlines flight #11 scheduled for September 11, 2001 using Common Address #1 as a contact address
- Marwan al-Shehhi has reservation on United Airlines flight #175 scheduled for September 11, 2001 using Common Address #1 as a contact address
- Salem Alhamzi has reservation on American Airlines flight #77 scheduled for September 11, 2001 using Common Address #2 as a contact address



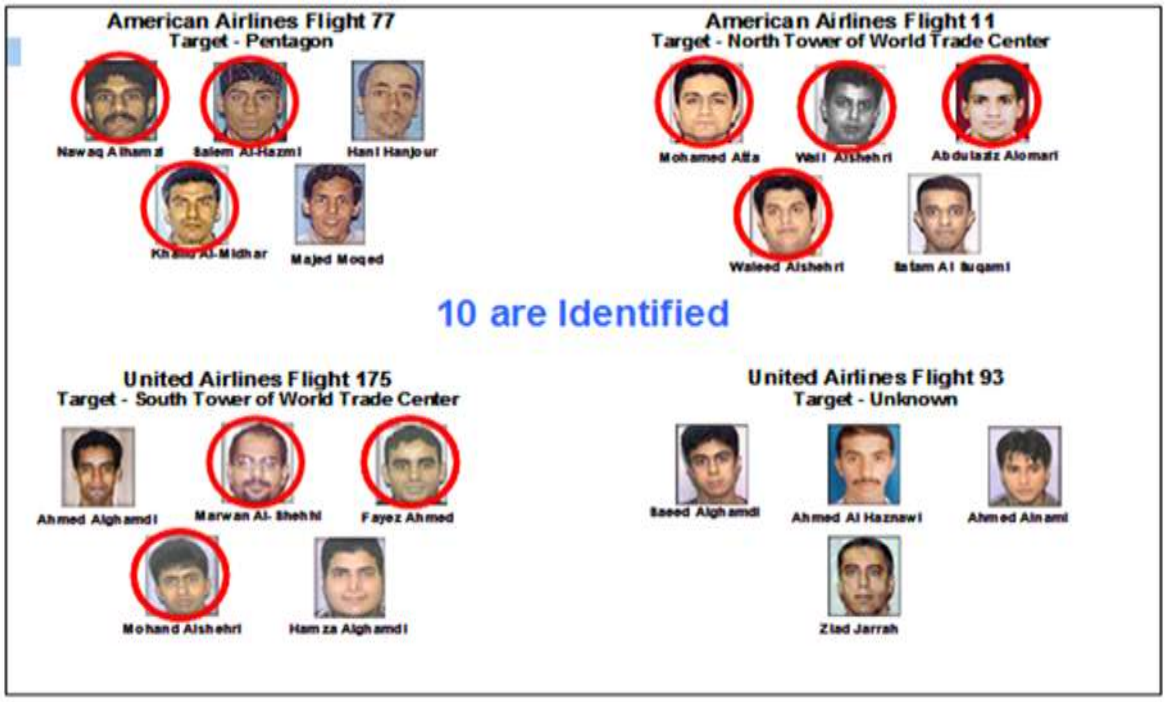
Phone Number Connections

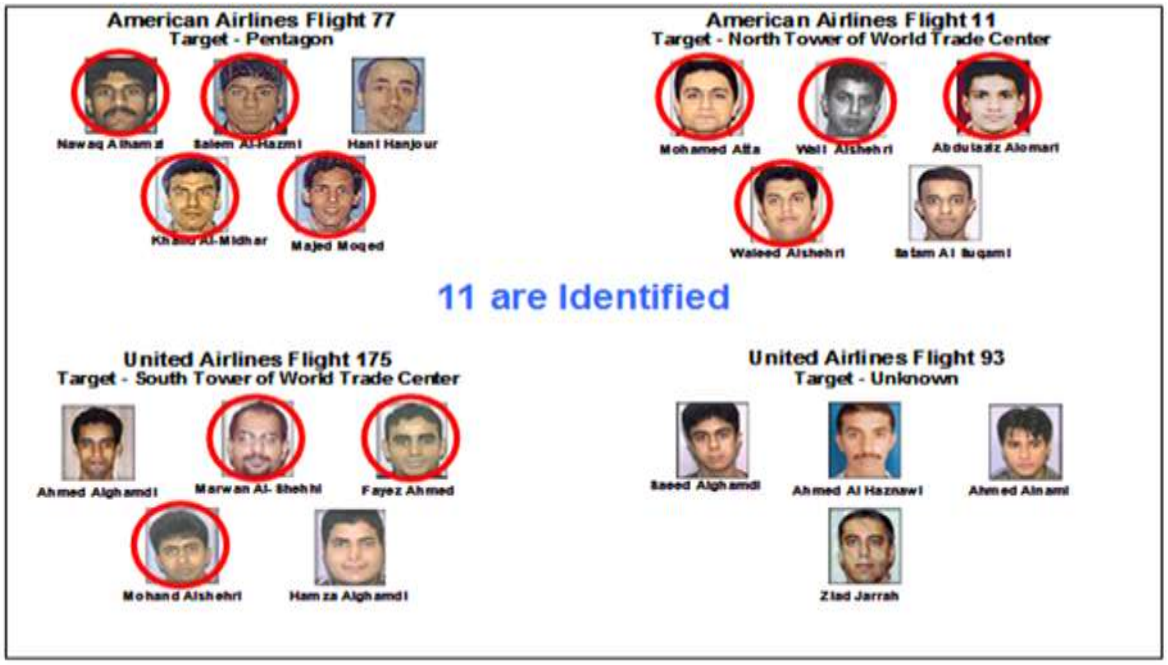
ONE (1) ALERTED PASSENGER MAKES RESERVATION USING COMMON TELEPHONE NUMBER

- On or about August 28, 2001, Mohamed Atta uses Florida Telephone #1 as a contact number when making reservations on American Airlines flight #11 scheduled for September 11, 2001

NUMBER IS USED BY FIVE (5) ADDITIONAL PASSENGERS

- On or about August 26, 2001, Waleed Alshehri and Wail Alshehri make reservations on American Airlines flight #77 scheduled for September 11, 2001 using Florida Telephone #1 as a contact number
- On or about August 27, 2001, reservations for electronic, one-way tickets were made for Fayez Ahmed and Mohand Alshehri for United Airlines flight #175 using Florida Telephone #1 as a contact number
- On or about August 28, 2001, Abdulaziz Alomari reserves a seat on American Airlines flight #11 using Florida Telephone #1 as a contact number

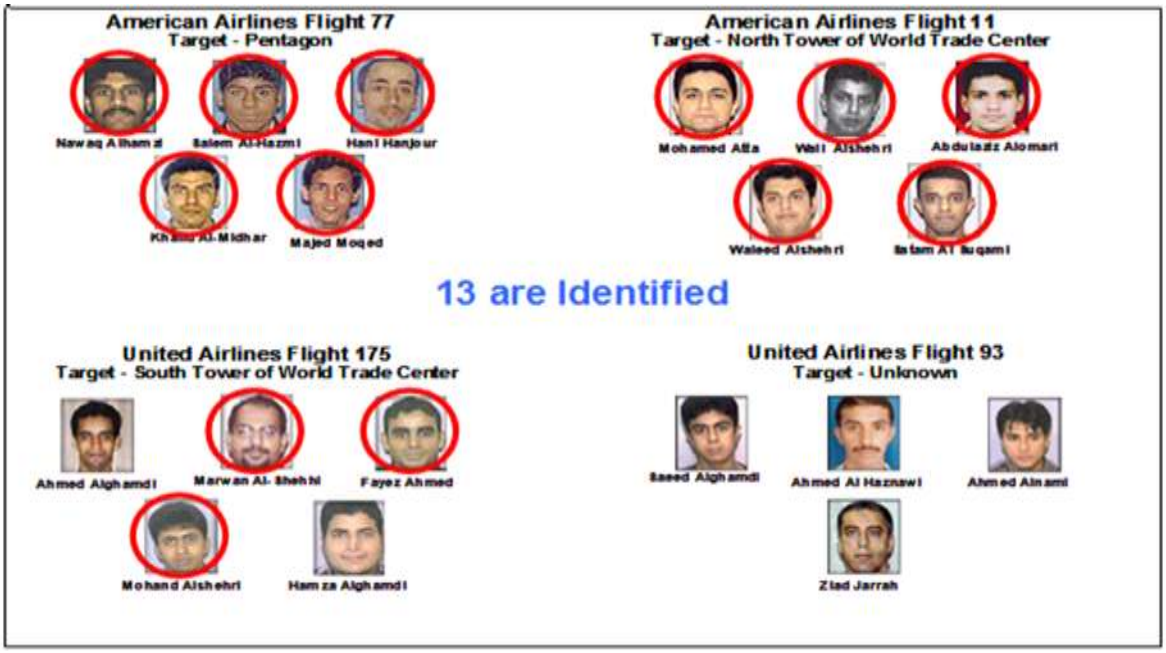




Public Record Connections

PUBLIC RECORDS

- Alerted subjects Nawaq Alhamzi and Khalid Al-Midhar lived with Hani Hanjour
- Alerted subject Wail Ashehri was roommates and shares PO Box with Satan Al Suqami



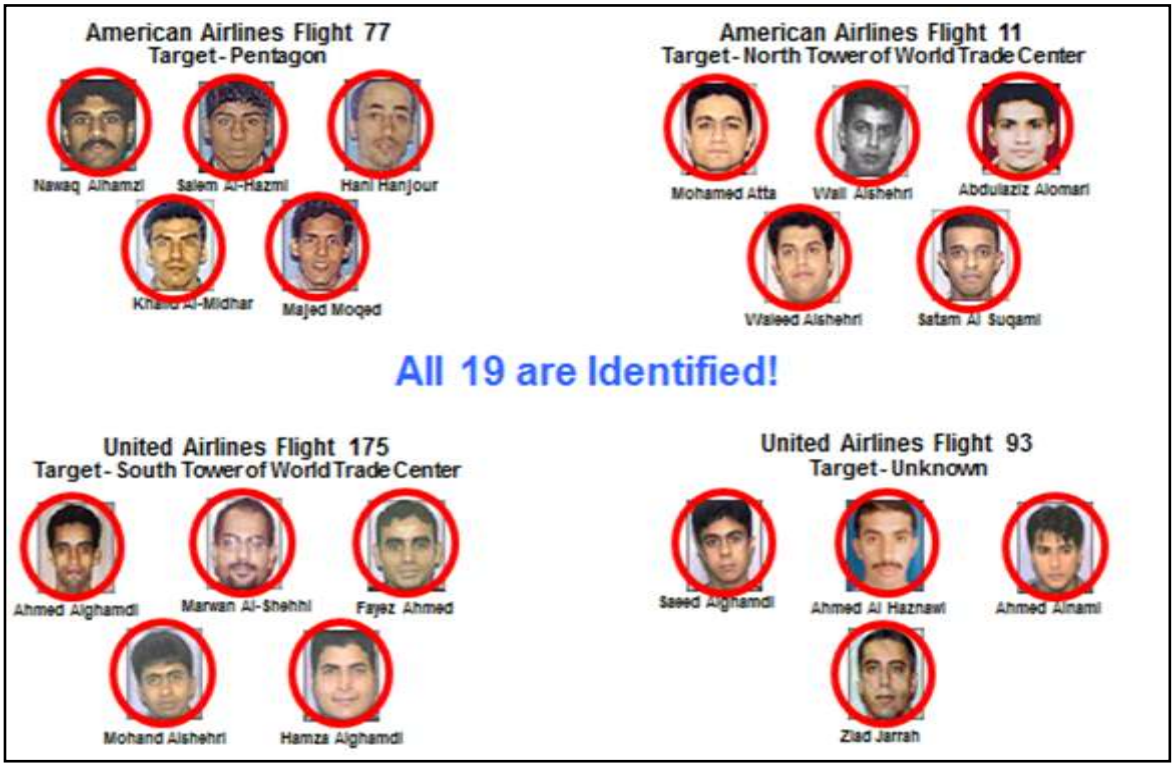
Remaining Connections

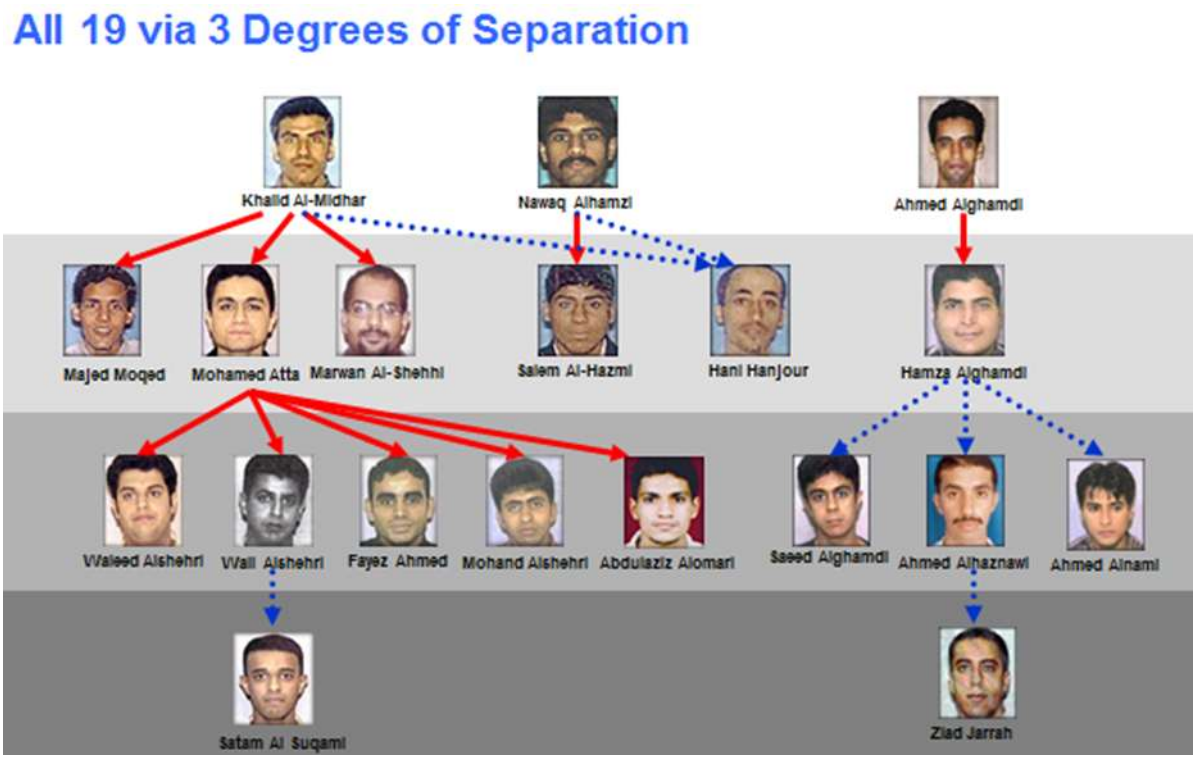
WATCH LIST: INS ILLEGAL/EXPIRED VISAS

- On or about August 29, 2001, Ahmed Alghamdi reserves an electronic one-way ticket on United Airlines flight #175 scheduled for September 11, 2001

FIVE (5) ADDITIONAL PASSENGERS:

- Alerted subject Ahmed Alghamdi and Hamza Alghamdi both use same address on their airline reservations
- Alerted subject Hamza Alghamdi has/does live with Saeed Alghamdi, Ahmed Alhaznawi, Ahmed Alnami
- Alerted subject Ahmed Alhaznawi lives/lived with Ziad Jarrah





7. Conclusion

As our ability to collect data about the world has expanded, identity and relationship resolution has become increasingly important for combating fraud and for understanding complex environments with multiple data sources. In this work, we focus on the task of entity resolution in relational environments. In relational environments, besides having attributes, entities relate to other entities, and those relations can be used to help resolve entities. We presented sample example of indent resolution along with IBM case of study on 9/11 incident. To present the benefits we used IBM tools for identity and relationship resolution which are built on industry standard algorithms and business rules. .

REFERENCES

[1] TH Kean, CA Kojm, P Zelikow, JR Thompson, S Gorton, TJ Roemer, JS Gorelick, JF Lehman, FF Fielding, B Kerrey, The 9/11 Commission Report. 2004. URL: http://govinfo.library.unt.edu/911/report/index.htm

[2] U.S. Department of State: Country Reports on Terrorism 2006. 2007. URL: http://www.state.gov/j/ct/rls/crt/2006/

[3] J Li, GA Wang, H Chen, Identity matching using personal and social identity features. Inf. Syst. Front. 13, 101–113 (2010)

[4] JS Pistole, Fraudulent Identification Documents and the Implications for Homeland Security. Statement Rec Before House Sel Comm Homel Secur. 2003. URL: https://www.fbi.gov/news/testimony/fraudulent-identificationdocuments-and-the-implications-for-homeland-security

[5] J Jonas, Identity resolution: 23 years of practical experience and observations at scale, in Proc 2006 ACM SIGMOD Int Conf Manage data - SIGMOD'06 (ACM Press, New York, NY, USA, 2006), p. 718 [SIGMOD'06]

[6] GA Wang, H Chen, H Atabakhsh, Automatically detecting deceptive criminal identities. Commun. ACM 47, 70–76 (2004)

[7] GA Wang, HC Chen, JJ Xu, H Atabakhsh, Automatically detecting criminal identity deception: an adaptive detection algorithm. IEEE Trans. Syst. Man.Cybern. Part a-Systems Humans 36, 988–999 (2006)

[8] IP Fellegi, AB Sunter, A theory for record linkage. J. Am. Stat. Assoc. 64,1183–1210 (1969)

[9] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, p. 5, 2007.

[10] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in Security and Privacy, 2009 30th IEEE Symposium on. IEEE, 2009, pp. 173–187.

[11] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Entity matching in online social networks," in Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013, pp. 339–344.

[12] JM Cheek, SR Briggs, Self-consciousness and aspects of identity. J. Res. Pers.16, 401–408 (1982)

[13] H Tajfel, JC Turner, The Social Identity Theory of Inter-Group Behavior (Nelson-Hall, Chicago, 1986)

[14] JC Turner, Some Current Issues in Research on Social Identity and Self-Categorization Theories (Blackwell, Oxford, 1999)

[15] K Deaux, D Martin, Interpersonal networks and social categories: specifying levels of context in identity processes. Soc. Psychol. Q. 66, 101–117 (2003)

[16] S Stryker, RT Serpe, Commitment, Identity Salience, and Role Behavior: Theory and Research Example (Springer-Verlap, New York, 1982)

[17] TC Redman, The impact of poor data quality on the typical enterprises. Commun. ACM 41, 79–82 (1998)

[18] United Kingdom Home Office: Identity Fraud: A Study. 2002. URL: http://www.homeoffice.gov.uk/cpd/id_fraud-report.pdf

[19] IBM Case Study – 9/11;http://www.cs.utsa.edu/~bylander/cs1023/SRD-IBM-Entity-Analytics-9-15-05.ppt

BIOGRAPHY

Arvind Kumar is Digital transformation leader with more than 14 years of IT experience and successfully implementing digital strategies/solution for highly visible large complex projects. Arvind is also a PMI certified Project Management professional and Fellow member of reputed professional body - IETE (The Institution of Electronics and Telecommunication Engineers)