A Rule and Graph-Based Approach for Targeted Identity Resolution on Policing Data

Michael Phillips
Intelligent Systems Research Centre
School of Computing and Digital Media
London Metropolitan University
London, UK
m.phillips@londonmet.ac.uk

Mohammad Hossein Amirhosseini
Department of Engineering and Computing
University of East London
London, UK
m.h.amirhosseini@uel.ac.uk

Hassan B. Kazemian
Intelligent Systems Research Centre
School of Computing and Digital Media
London Metropolitan University
London, UK
h.kazemian@londonmet.ac.uk

***Abstract*— In criminal records, intentional manipulation of data is prevalent to create ambiguous identity and mislead authorities. Registering data electronically can result in misspelled data, variations in naming order, case sensitive data and inconsistencies in abbreviations and terminology. Therefore, trying to obtain the true identity (or identities) of a suspect can be a challenge for law enforcement agencies. We have developed a targeted approach to identity resolution which uses a rule-based scoring system on physical and official identity attributes and a graph-based analysis on social identity attributes to interrogate policing data and resolve whether a specific target is using multiple identities. The approach has been tested on an anonymized policing dataset, used in the SPIRIT project, funded by the European Union's Horizon 2020. The dataset contains four 'known' identities using a total of five false identities. 23 targets were inputted into the methodology with no knowledge of how many or which had false identities. The rule-based scoring system ranked four of the five false identities with the joint highest score for the relevant target name with the remaining false identity holding the joint second highest score for its target. Moreover, when using graph analysis, 51 suspected false identities were found for the 23 targets with four of the five false identities linked through the crimes they had been involved in. Therefore, an identity resolution approach using both a rule-based scoring system and graph analysis, could be effective in facilitating the investigation process for law enforcement agencies and assisting them in finding criminals using false identities.**

***Keywords*— identity resolution, identity model, graph analysis, rule-based, policing data.**

I. Introduction

A. Identity Resolution

Identity can be described as a set of identifiable characteristics that can distinguish one individual from another. These characteristics include physical characteristics, such as gender or ethnicity, and social characteristics, such as known associates and organisations. Identity resolution is the process of being able to collect and match identifying attributes of a person to build a consistent identity of that individual. In policing, one of the main aims of identity resolution is to be able to determine the true identity of criminals using multiple identities to hide their consistent involvement in illegal activities.

Identity resolution has been used effectively to evaluate customers, and their interests, in the marketing sector and examine travel records but there are few implementations in the field of policing and criminal fraud detection [1]. Duplicate and false identity records are present in databases and electronic systems because of a lack of validation of data attributes or verification during data entry processes [2]. This can be a problem as some criminals try to hide their identity using fake names and other information. Many government reports detail terrorists in different countries committing differing identity crimes e.g. falsifying passports or birth certificates to facilitate their travel or financial operations [3,4]. In these cases trying to decipher the identities relating to the same person can be a particular challenge [5].

B. Identity Modelling

An individual's identity can be considered to be constructed of: (1) a personal identity, which is a person's self-perception as an individual, and (2) a social identity, which is a person's biographical history that builds over time [6]. Previously researchers have only considered personal identity, which considers a person in isolation, whereas in reality people interact with each other in society both in person or online virtually. The social context of an individual can be a useful factor when resolving identities along with personal data relating to both physical appearance and legal documentation. Given these different aspects of identity, we propose an identity model which considers four categories of information: (1) official identity, (2) physical identity, (3) social identity, and (4) virtual identity.

C. Methodologies for Identity Resolution

Methodologies of identity resolution typically use one or more of a combination of techniques including rule or score-based comparisons, distance measures and graph analysis. Rule-based methods use specific rule sets to match records of identities to distinguish similar records, for example, by matching forename, surname, and date of birth [7]. This method typically has high precision but low sensitivity in detecting true matches because of missing or incorrect data, however, this can be improved by creating an effective and robust rule set which can work in a variety of different contexts, although this can be time consuming [2].

Brown and Hagen [8] introduced a data association method linking criminal records that could potentially be the same suspect by using a weighted sum of distance measures comparing specific features of each record. A similar methodology comparing name, date of birth, social security number and address was used on criminal identity records has also been used [9]. Records are labelled by an expert, with a supervised learning method used to determine the threshold for defining a match. This method was criticised for its potential to fail if one or more of the considered attributes contained missing data [10].

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

Bhattacharya and Getoor [11] proposed a graph-based method for entity resolution which used a distance measure that combined corresponding attribute similarities with graph-based relational similarity for each entity pair. This work was later extended, and a methodology using collective entity resolution was proposed which could derive new social information and incorporate it into a further resolution process repeatedly, as opposed to making pair-wise entity comparisons. There were concerns that this methodology could not distinguish one person having several profiles on different social media platforms and techniques for matching user profiles were developed to solve this issue. A similar graph-based method used two graphs created from both social linkage and user profile attributes to improve the performance of identity resolution [12].

The development of social media in recent years, such as Facebook, Twitter and LinkedIn, has resulted in further research on user identity linkage across online networks i.e. trying to match one person to all their online identities [13]. This has led to several methodologies being used to look at user identity linkage and identity resolution on social media networks including neural networks and graph analysis [14-16]. This research, and the continued use of AI methodologies generally, has resulted in concerns over bias in algorithms, particularly in regard to ethnicity, so this was considered in developing the proposed methodology for use on the policing data where gender, ethnicity and age are important attributes [17,18].

For our methodology, we have developed a targeted identity resolution method which attempts to find false identities of a specific target, given simply a name to search for in a dataset. The method uses physical, official and social identity attributes using first a rule-based scoring system to search the database for possible alternate identities and calculating a score for each based on comparisons of different physical and official attributes in each criminal record. Following this, a social graph analysis is used, linking the crimes the target, and each possible false identity, has been involved with to see if a link can be made. Using both a combination of rule-based and graph-based systems to interrogate policing data has not been investigated as far as we are aware.

II. METHODOLOGY

A. Dataset

This research uses an anonymized policing dataset which is part of the SPIRIT project funded by European Union's Horizon 2020. The dataset contains arrest records distinguished by an identity and a crime (specified by reference numbers). A person can be related to multiple crimes as can a crime be related to multiple persons. The dataset is made up of 1,145,418 records containing 694,264 identities involved in 913,734 crimes. The attributes relating to each person include forename, surname, gender, date of birth, ethnicity, home address and the role they played in the crime e.g. victim, suspect, defendant. The attributes relating to each crime include description and category of offence, location and the date and time when the crime was committed. All location details include an address, postcode and geographic Cartesian coordinates. The attributes in the dataset used in the methodology proposed are detailed further in Table I.

B. Data Cleaning

For the methodology, it is important to have data for forename, surname, gender, date of birth, ethnicity and the locations of both the home address of the person and the crime location. 977 records contained no forename or surname, 9,790 records were without a valid date of birth and 26,261 had an unclear or blank ethnicity so these were removed from the dataset. Additionally, there were 454 records where gender was defined as 'Unknown' and these records were also removed. The cleaned dataset still contained 1,051,049 records.

As well as removing records, data cleaning involved categorizing data including grouping nine ethnicities with two genders to create categories e.g. 'White Skinned European Male' or "Asian Female'. As this categorization is done from arrest records, it eliminates the possibility of ethnicity bias in the scoring system or graph analysis. Types of offence were categorized based on descriptions and key words. This was done by an expert in the field and included categories such as fraud, theft, sexual offences, harassment, drug offences and violent attacks. Finally, the type of role a person played in the crime was split into two categories, either the victim of the crime or the defendant / suspect.

TABLE I. DETAILS OF USED ATTRIBUTES IN THE DATASET

Attribute	***Description***
nominal_ref	Unique ID for the person. Often one identity has multiple nominal_ref because they are not recognized in the database in further arrests.
crime_ref	Unique ID for the crime
forename	Forename of the person
surname	Surname of the person
gender	Gender of the person, either 'M', 'F' or 'U'
date_of_birth	Date of birth of the person
role_type	Role of the person in the crime e.g. 'VICT' (victim), 'DEFE' (defendant)
ea_desc	Ethnicity of the person e.g. 'Oriental', 'African-Carribean'
northing	Northing co-ordinate for person's home address
easting	Easting co-ordinate for person's home address
crime_easting	Northing co-ordinate for the location of the crime
crime_easting	Easting co-ordinate for the location of the crime
offence	Description of the crime committed

C. Identity Targets

Unknown before the methodology was applied, the dataset contained five 'known' false identities linked to four different people / names (one of which was using two of the false identities). A list of 23 names of possible suspects, including the four people using the false identities, was compiled by a SPIRIT project colleague who knew the identities which related to one another and those that did not. The other 19 people / names were included to test that the methodology would not bring back significant numbers of false positive results and were chosen with the number of arrest records considered to include people with similar data to the true targets. The names chosen had several records in the dataset as would be expected from serial offenders using false identities to escape capture.

The date of birth of each target was unknown so for several targets the methodology would split the name into individual identities where there was different ethnicity, gender and date of birth. The methodology was run for each identity of a named target with the aim of finding possible alternative identities which that target could be using or be an associate of.

D. Rule-Based Scoring System

The first phase of the methodology uses a rule-based scoring system to compare key attributes between records to decide which are most likely to be false identities of a given target. To do this each record with the name of the target is retrieved from the dataset. These records are then split depending on the date of birth of each target and a list of targets is created for each name given. For each, the gender, ethnicity and age / year of birth of the target is taken from the records returned, ensuring there are no discrepencies between records. The algorithm then brings back other records in the dataset which match the gender, ethnicity and age match of the target. These physical attributes were used because they are difficult to fake when arrested and they would typically narrow down the dataset to a more suitable number of records which could be compared with those of each target.

Each record containing the target's name and date of birth is then compared with the other records in the aforementioned list. Six further key attributes are then compared using differing comparisons and distance measures (summarised in Table II). These include the crime reference number in the record (x1), which is used to see if identities have been involved in the same crime, yielding a result of either a direct match (1) or no match (0). The role type (x2) the identity played in the crime (either a defendant or victim) were compared in the same way as either a match (1) or no match (0).

The identity's home address (x3) and the location of the crime (x4), based on easting and northing co-ordinates, were compared using straight line distance between the two co-ordinates up to a maximum of 100km and then normalised to fit in to a 1 (same location) and 0 (>=100km) decimal scale.

The distance between the two locations in km was calculated as:

$$km\ dist = \frac{\sqrt{(e1-e2)^2-(n1-n2)^2}}{1000} \qquad (1)$$

where e1 and n1 are the easting and northing co-ordinates for one location, e2 and n2 are the easting and northing co-ordinates for the other location and km dist is the distance in km between the two locations. To calculate as a decimal for the scoring system this value was divided by 100 and subtracted from 1. The straight line distances do not take into account the curvature of the Earth but with locations all typically in the UK it was thought that this effect would be negligible.

Offence (x5) was measured as a match if the two types of offence were in the same category, as categorized by the labelling of an expert using keywords and offence descriptions. Finally, the identity's birthday (x6) was compared and a score was given for an identical match (1), a close match with one number difference in day to account for typos (0.75) and no match (0) used to score a comparison.

All these attributes were compared using the simple summation below to produce a final score to the rule-based scoring system:

$$score = x^1 + x^2 + x^3 + x^4 + x^5 + x^6 \qquad (2)$$

With the scores calculated for each set of records, a threshold score of 4.8/6 was proposed to separate the records for the next stage of analysis. This score was chosen by an expert given the methodology noted and in trying to make sure potential matches were not lost before the graph-based analysis. This was done before any of the false identities were known. The threshold allows one attribute to have a score of 0 and gives a margin for other attributes to have very high scores close to a perfect match. This threshold was tested on a range of other identities not in the list of test subjects to ensure that the threshold brought back reasonable numbers of identities to fit in with the rest of the methodology and that the police have limited time to put towards investigating false identities.

TABLE II. SUMMARY OF COMPARED ATTRIBUTES AND MEASURES

Attribute	*Comparison / Distance Measure*
Crime Refence	Exact Match
Role	Exact Match
Home Location	Straight Line Distance
Crime Location	Straight Line Distance
Offence	Categorical Match
Birthday	Match Based on Potential Typo

Having removed any compared records below the threshold score, the remaining comparisons were ordered by the highest score. Multiple instances of the same identities being compared were removed leaving the one with the highest score. Finally, if the target identity is compared with itself then this was removed from the list.

E. Graph-Based Analysis

Following the rule-based scoring system phase, a graph analysis method is proposed to investigate any identities with a score of 4.8 or higher for each target. The graph analysis aims to use social networking graphs based on the crimes the target has been involved with to see if they can be linked to identities from the rule-based scoring system. The graph is constructed by starting with the node of the target and searching the dataset for all the records involving this target node / identity (based on the combined name and date of birth attributes). For each of these records, the crime reference number is recorded and any records with this crime reference are extracted from the dataset. Any identities involved in these sets of crimes are linked back to the target i.e. all identities become nodes in the graph and then edges are created between them if they can be linked in the same crime. This process is done for two cycles, or when identities cannot be linked to further crimes, to ensure consecutive links can be accounted for. The theory behind this analysis is that a target using false identities is still likely to be involved with the same criminal associates and can likely be linked through crimes they have been involved in.

A graph was constructed for each target and it was recorded which, if any, of the potential false identities was a node in the graph connecting them to the target. If there was a connection this was considered to suggest that there is more likely to be a link between them and the target and therefore it may be the target using a false identity or at least a criminal associate of the target.

Following the application of both the rule-based scoring system and graph analysis, the possible false identities were listed for each target. Those with the highest scores from the rule-based scoring system and found to be linked in the graph-based analysis were considered as the best matches and most likely to be false identities of the target. The full methodology has been illustrated in Fig. 1, where each phase and the inputs and outputs to and from each part can be clearly seen.

F. *Development of Methodology*

The methodology was implemented in Python 3.7 using Anaconda IDE. The dataset was stored and analyzed using the Python data analysis library (pandas) which was used heavily for data cleaning and implementing the rule-based scoring system. Graph analysis was done using the NetworkX library and visualized using pyplot from the matplotlib library.

III. RESULTS AND DISCUSSION

A. *Rule Based Scoring System Results*

The rule-based scoring system was run for all 23 targets on over 1,224 identities based on a combination of name and date of birth. Just over half the names given had six or fewer identities but more popular names had significantly more identities, the highest being 341 (mean = 53.21). In total, 4,357 records were analyzed for the targets, a mean of 3.56 for each identity.

Using the threshold score of 4.8, the rule-based methodology produced 6,173 possible false identities, a mean of 5.04 matches per identity. For some targets, with identities based on date of birth, this would be too many records for law enforcement agencies to investigate further, however, if the date of birth of the target was known it would narrow down the search considerably. It is worth noting that some of the identities matched with possible false identities seem to be the same identity spelt differently or with additional names e.g. ‘Ratar Nhung’ is matched with ‘Ratar Nan Nhung’ and it is fair to assume, given the same date of birth, that these identities are the same person, just with one a middle name has been added.

All five false identities are contained within the returned records for the respective target for the rule-based scoring system. Indeed, more encouragingly, four of them are listed with the joint highest score and the final false identity has the second joint highest score for the respective target. Furthermore, they all have a score of 5 or higher, suggesting the threshold could be moved up to improve specificity. If a threshold of 5 or above was used then the number of matches returned would be just 266 for all 23 targets, a mean of just 0.22 potential matches per identity.

B. *Graph-Based Results*

With the rule-based scoring system having removed records with scores of less than 4.8, the remaining targets and potential false identities are inputted into the graph-based analysis. An example of two graphs from this stage of the methodology are shown in Fig. 2 and Fig. 3. In both these examples the graphs correctly connected one of the targets with a false identity These graphs have the crime reference nodes removed but show the network between identities with the target and false identity labelled.

The graph-based analysis brought back a total of 49 links between the targets and potential false identities for the 23 names. Four of the five false identities were found in this analysis. The false identity not found did not contain as many links with other crimes which is likely why it was missed but this identity did have a high score for the rule-based scoring system which would still have allowed the law enforcement agency to investigate it.

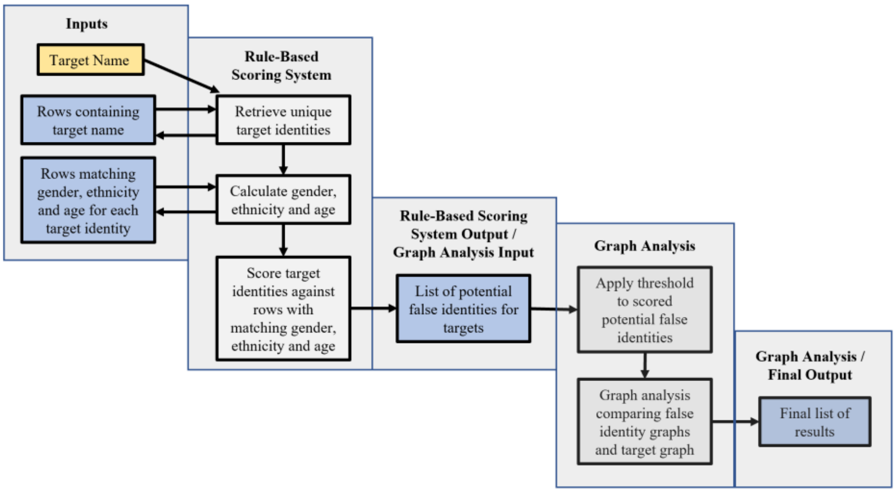
Inputs
Target Name
Rows containing target name
Rows matching gender, ethnicity and age for each target identity
Rule-Based Scoring System
Retrieve unique target identities
Calculate gender, ethnicity and age
Score target identities against rows with matching gender, ethnicity and age
Rule-Based Scoring System Output / Graph Analysis Input
List of potential false identities for targets
Graph Analysis
Apply threshold to scored potential false identities
Graph analysis comparing false identity graphs and target graph
Graph Analysis / Final Output
Final list of results

Fig. 1. Diagram of the full methodology showing each phase with summaries of the main tasks, the input / output from each, the data used (in the purple boxes) and the initial input of the target name (in the yellow box)

Fig. 4 and Fig. 5 show examples of graph-based analysis links that did not relate to false identities but could still show vital connections between criminals. They show the difference in the number of links in individual target identity graphs as the chains were broken earlier. In Fig. 4, a shared surname suggests the two identities could be twin brothers or linked associates. In Fig. 5, it appears the identities have a clear link and given the same date of birth it may be that these are the same person but that this is not 'known' in the dataset.

C. *Overall Results*

The overall results show that this methodology succeeds in giving law enforcement agencies potential identities that a specific target could be using as false identities or potentially involved with criminally. The ability to find the five false identities in over 500,000 identities without bringing back large amounts of data is a positive outcome. Even though one of the false identities could not be connected in the graph analysis, it scored highly in the rule-based scoring system.

D. *Discussion*

This work has shown that using a combination of a rule-based scoring system and graph analysis can be very effective for identifying false identities used by criminals in policing records. The rule-based scoring system has been shown to filter down relevant records well using a considered, conservative threshold of 4.8. The dataset in this study has shown that it is possible that this threshold could be raised to 5 or above. With more data, more experimenting could be done to look at this threshold to find a more optimal cutoff to ensure false identities would be detected while leaving out unwanted identities with no known connections. In addition to this, further data would allow the scoring system to be modified and other attributes compared which could be useful for detecting false identities. This would allow the

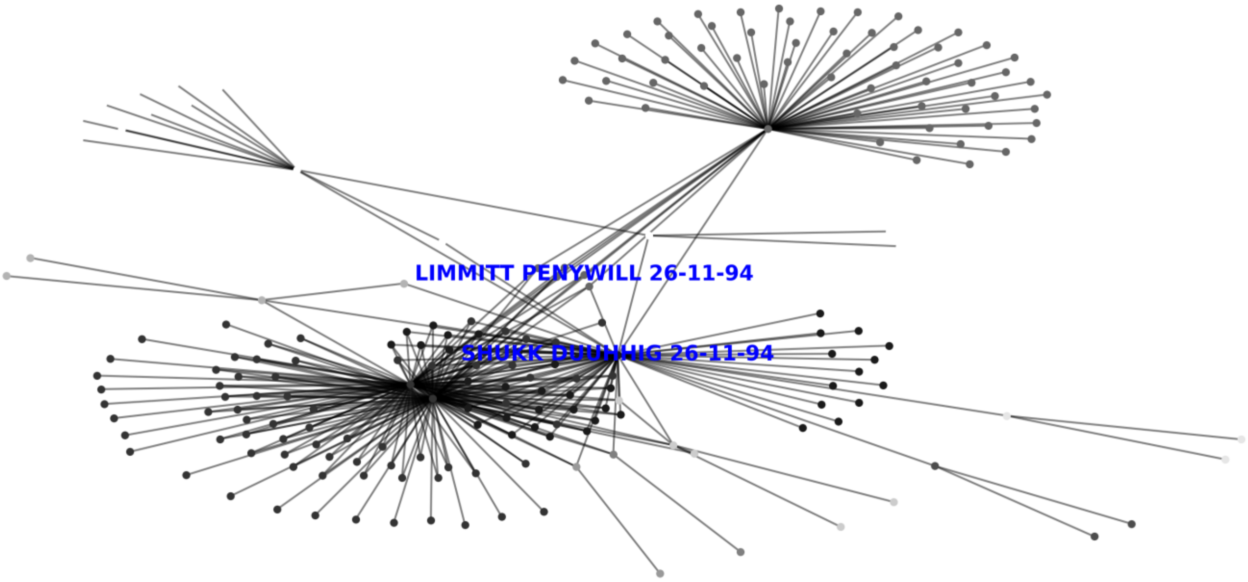

Fig. 2. Graph-based analysis on target Shukk Duuhhig (born 26/11/94) which reveals a link with Limmitt Penywill (born 26/11/94). These names represented the same person.

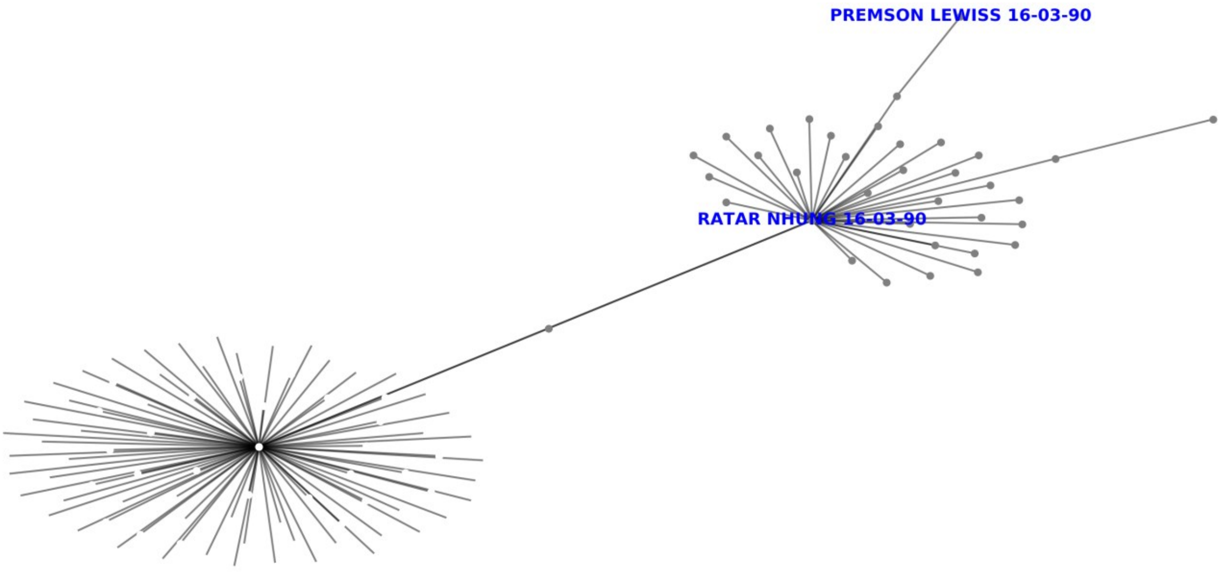

Fig. 3. Graph-based analysis on target Ratar Nhung (born 16/03/90) which reveals a link with Premson Lewiss (born 16/03/90). These names were not known to represent the same person but could well be connected still.

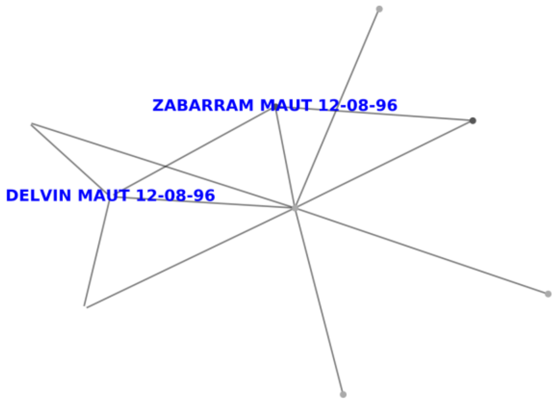

Fig. 4. Graph-based analysis on target Zabarram Maut (born 12/08/96) which reveals a link with Delvin Maut (born 12/08/96). These names were not known to represent the same person but could still be connected.

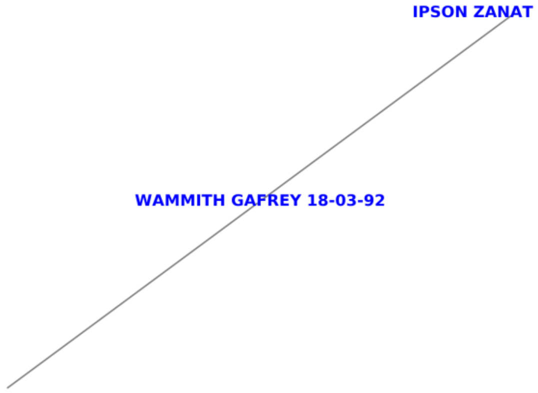

Fig. 5. Graph-based analysis on target Ipson Zanat (born 18/03/92) which reveals a link with Wammith Gafrey (born 18/03/92). These names were not known to represent the same person but could still be connected.

development of the model further and produce a more robust methodology to work across more data.

To improve the current methodology, it would be beneficial to look at string comparison methods, such as Jaro-Winkler and Soundex, to be able to match identities in the dataset which are likely to be the same person but with slight differences in name due to spelling errors or differences in data input. For example, the inclusion of middle names or a change in surname, particular for women who have gotten married, has been noticed throughout the analysis and this could provide an automatic step to reduce the number of identities the methodology returns which are not useful. Another concept to consider would be changing the influence of each individual attribute to the scoring system by using weightings for each to try and achieve better results by optimizing the weightings to improve the algorithm.

It would be good to compare the methodologies developed in this paper with AI techniques such as deep learning or neural networks. Given the large number of records in this policing dataset, a combination of the rule-based scoring system and a neural network could potentially work well, and this is something we hope to investigate in the future. While the graph analysis worked well with an 80% success rate, it would be interesting to look deeper into what happened with the false identity that was not detected. If this method could be improved to be able to find a connection then the combination of the two methodologies could be combined even further, potentially even using a lower threshold on the rule-based scoring system to ensure no possible false identities are missed. Investigating the number of degrees of separation used to connect nodes on the graph was not experimented with as part of this work and this would be an area of interest to look at in the future. As well as this, it may be possible to use other attributes to build graph-based analysis which could also aid the methodology for identity resolution.

We hope to be able to apply this methodology to more data in the future and from this produce clearer statistical analysis on the success of the method and how it may compares to other techniques as although our primary focus is the usefulness to law enforcement agencies, we acknowledge it is also important to be able to show the success of methodologies in a more quantifiable way.

IV. CONCLUSION

This research introduces a new rule-based scoring system combined with a graph-based analysis for identity resolution. The methodology was applied to an anonymized policing dataset used as part of the SPIRIT project funded by the European Union's Horizon 2020 initiative and contained 1,145,418 records. The rule-based scoring system used physical and official identity attributes to narrow down potential false identities of a target using gender, ethnicity and year of birth. It then directly compared the category of offence, person role, crime location, home location of the person and the person's birthday of the target identity record and potential false identities. The rule-based scoring system correctly included all the false identities with four of them having the joint highest score for their respective target and the remaining false identity scoring the joint second highest score.

Using a threshold score of 4.8, the graph-based analysis used the list of potential false identities from the rule-based scoring system. It found 49 links from targets to potential false identities including four of the five false identities. Other links also seem to reveal interesting identities in the dataset unnoticed for this work including possible twin brothers and another possible false identity. Given all these results, we believe, this identity resolution approach could be used to effectively facilitate the investigation process for law enforcement agencies and assist them in finding criminals using false identities.

ACKNOWLEDGMENT

This research was supported by European Union's Horizon 2020 SPIRIT project - Grant No. 786993 and Intelligent Systems Research Centre. We thank our colleagues from the SPIRIT project consortium who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] D. R. Kretz and R. W. Paulk, "Establishing Traveler Identity Using Collective Identity Resolution," in 2010 IEEE International Conference on Technologies for Homeland Security (HST), Waltham MA, 2010, pp. 308-313.

[2] J. Li and A. G. Wang, "A framework of identity resolution: evaluating identity attributes and matching algorithms," Secur. Inform., 4, 6, 2015.

[3] T. H. Kean, C. A. Kojm, P. Zelikow, J. R. Thompson, S. Gorton, T. J. Roemer, J. S. Gorelick, J. F. Lehman, F. F. Fielding, and B. Kerrey, "The 9/11 Commission Report," 2004, [online]. Available:

http://govinfo.library.unt.edu/911/report/index.htm. [Accessed Aug. 8, 2020].

[4] U.S. Department of State, “Country Reports on Terrorism,” 2006, [online]. Available: http://www.state.gov/j/ct/rls/crt/2006/. [Accessed Aug. 8, 2020].

[5] J. Li, A. G. Wang, and H. Chen, “Identity matching using personal and social identity features,” Inf. Syst. Front., vol. 13, pp. 101-113, 2010.

[6] J.M. Cheek and S.R. Briggs, “Self-consciousness and aspects of identity,” J. Res. Pers., vol. 16, pp. 401-408, 1982.

[7] B. Marshall, S. Kaza, J. Xu, H. Atabakhsh, T. Petersen, C. Violette, and H. Chen, “Cross-Jurisdictional Criminal Activity Networks to Support Border and Transportation Security,” in 7th Int IEEE Conference Intelligent Transportation Systems, Washington D.C., 2004, pp. 100-105.

[8] D.E. Brown and S.C. Hagen, “Data association methods with applications to law enforcement,” Decis. Support Syst., vol. 34, pp. 369-378, 2003.

[9] G.A. Wang, H. Chen, and H. Atabakhsh, “Automatically detecting deceptive criminal identities,” Commun. ACM, vol. 47, pp. 70-76, 2004.

[10] G.A. Wang, H.C. Chen, J.J. Xu, and H. Atabakhsh, “Automatically detecting criminal identity deception: an adaptive detection algorithm,” IEEE Trans. Syst. Man. Cybern. Part A- Systems Humans, Vol. 36, pp. 988-999, 2006.

[11] I. Bhattacharya and L. Getoor, “Entity resolution in graphs,” in Min graph data, Wiley-Blackwell, Hoboken, 2006.

[12] S. Bartunov, A. Korshunov, S. Park, W. Ryu, and H. Lee, “Joint Link-Attribute User Identity Resolution in Online Social Networks,” in Proc. 6th Work. Soc. Netw. Min. Anal., Beijing, China, 2012.

[13] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, “User identity linkage across online social networks: a review,” in SIGKDD Explorations Newsletter, pp. 5–17, ACM, 2017.

[14] F. Zhou, L. Liu, K. Zhang, G. Trajcevski, J. Wu, and T. Zhong, "DeepLink: A Deep Learning Approach for User Identity Linkage," IEEE Conference on Computer Communications, Honolulu, HI, pp. 1313-1321, 2018.

[15] S. Fu, G.Wang, S. Xia, and L. Lu, “Deep multi-granularity graph embedding for user identity linkage across social networks,” Knowledge-Based Systems, 193, 6, 2020.

[16] R. Wang, H. Zhu, L. Wang, Z. Chen, M. Gao, and Y. Xin, "User Identity Linkage Across Social Networks by Heterogeneous Graph Attention Network Modeling," Appl. Sci. 10, 16, 2020.

[17] S. Silva and M. Kenney, ‘Algorithms, platforms, and ethnic bias,’ Communications of the Association of Computing Machinery, 62, 11, pp. 37-39, 2019.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” ArXiv, abs/1908.09635, 2019.