Journal of Cyber Security Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tsec20

A graph-based method for identity resolution to assist police force investigative process

Mohammad Hossein Amirhosseini, Hassan Kazemian & Michael Phillips

To cite this article: Mohammad Hossein Amirhosseini, Hassan Kazemian & Michael Phillips (26 May 2024): A graph-based method for identity resolution to assist police force investigative process, Journal of Cyber Security Technology, DOI: 10.1080/23742917.2024.2354555

To link to this article: https://doi.org/10.1080/23742917.2024.2354555

Published online: 26 May 2024.

JOURNAL OF CYBER SECURITY TECHNOLOGY
https://doi.org/10.1080/23742917.2024.2354555

Taylor & Francis
Taylor & Francis Group

A graph-based method for identity resolution to assist police force investigative process

Mohammad Hossein Amirhosseini[a], Hassan Kazemian[b] and Michael Phillips[b]

[a]Department of Computer Science and Digital Technologies, University of East London, London, UK; [b]Intelligent Systems Research Centre, School of Computing and Digital Media, London Metropolitan University, London, UK

ABSTRACT
The ability to prove an individual identity has become crucial in social, economic, and legal aspects of life. Identity resolution is the process of semantic reconciliation that determines whether a single identity is the same when being described differently. This paper introduces a novel graph-based methodology for identity resolution, designed to reconcile identities by analysing the similarity of attribute values associated with different identities within a policing dataset. The proposed methodology employs graph analysis techniques, including centrality measurement and community detection, to enhance the identity resolution process. This paper also presents a new identity model for identity resolution. SPIRIT policing dataset was used for testing the proposed methodology. This dataset is an anonymised dataset used in the SPIRIT project funded by EU Horizon. It contains 892 identity records and among these, two 'known' identities utilize different names but actually represent the same individual. The presented method successfully recognised these two identities. Additionally, another experimental evaluation was conducted on a refined and extended version of the dataset and the false identities were successfully detected. This method can assist police forces in identifying criminals and fraudsters using fake identities and has applications across finance, marketing, and customer service.

ARTICLE HISTORY
Received 13 November 2023
Accepted 8 May 2024

KEYWORDS
Identity resolution; identity model; graph analysis; community detection; centrality measurement

1. Introduction

Identity can be described as a set of identifiable characteristics that can distinguish one individual from another. The characteristics that can describe an identity are usually categorised as physical characteristics and social

CONTACT Mohammad Hossein Amirhosseini m.h.amirhosseini@uel.ac.uk Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London, Docklands Campus, University Way, London E16 2RD, UK

characteristics. For instance, gender or ethnicity can be considered as physical characteristics, and known associates or organisations can be considered as social characteristics. Bilgrami [1] discussed that identity has two different aspects including subjective and objective. Subjective identity reflects an individual's self-perception, while objective identity concerns how others view an individual, independent of their self-perception. Essentially, objective identity defines individuals based on observable biological or social characteristics. Nowadays electronic records are replacing paper-based documents, and identity records can be generated easily. Therefore, due to inadequate verification or validation during data entry, duplicate and false identity records frequently occur in electronic systems and databases [2]. In this situation, finding an effective solution to address this issue is extremely critical, and it can facilitate fighting crime, terrorism or enforce national security. Li and Wang [2] pointed out that criminals and terrorists try to hide their true identity via using fake identities. There are some cases documented by government reports which are showing terrorists in different countries have committed different identity crimes such as falsifying passports or birth certificates to facilitate their travelling or their financial operations [3,4]. Moreover, the rapid advancements in artificial neural networks, bolstered by increased computing power, have led to the development of technologies capable of digitally manipulating content. This progress has sparked significant concerns over deepfake images and videos, which can be exploited for harmful purposes, including identity theft [5].

The problem of multiple identities for an individual can mislead police and law enforcement investigators [6]. Identity resolution is a pathway to tackle problems when it becomes intensely difficult to determine if the resultant identity is the same when criminals describe it differently. In other words, it is the process of collecting and matching identifying attributes of a person to be able to build a consistent identity of that individual. Identity resolution is one of the main police force investigative processes, which can be considered as a form of classification where two or more identities of a person should be compared based on the similarity of their attributes [7]. It has been used effectively in different disciplines such as marketing for evaluating customers and their interests in the marketing sector or transportation for examining travel records. However, it has not been widely used in the field of policing, and there are very few implementations for criminal fraud detection [8]. In fact, the main aim of identity resolution in the field of policing and fraud detection is to be able to determine the true identity of criminals who are using multiple identities to hide their involvement in illegal activities.

This research advances the field of identity resolution, particularly within policing and fraud detection, by presenting a novel identity model that integrates both physical identity and social identity, alongside additional factors including official identity and virtual identity that enhance the accuracy and reliability of identity verification. Traditional identity models have primarily focused on physical and social identity components, often overlooking the complex interplay between an individual's multiple identity facets and their social context. By acknowledging and incorporating a broader range of attributes, this study addresses the critical challenges of duplicate and fraudulent identities prevalent in electronic databases.

Furthermore, the research builds on and improves the existing identity resolution methodologies. While rule-based systems and distance measures have their merits, they often suffer from issues related to data quality and the inflexibility of rule sets across different contexts. By employing a combination of advanced graph analysis techniques and distance measures for the first time, this study demonstrates a more robust approach to identity resolution. The presented method not only enhances the detection of false identities but also adapts to the complexities of modern data environments, which include diverse and often incomplete data sources. The research's methodological contributions set a foundation for future advancements in the field, ensuring that identity resolution approaches keep pace with technological advancements and evolving societal norms.

2. Literature review

2.1. Identity model

There should be a clear identity model before starting the identity resolution process. Kidd and Teagle [9] explained that identity relates to how we think about ourselves as people, how we think about other people around us, and what we imagine others think of us. In other words, it means being able to figure out how we are as people. 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 [10]. Buckingham [11] discussed that a person's identity is something that can be uniquely possessed, and it distinguishes that person from other people. On the other hand, identity also implies a relationship with a social group. For instance, when someone talks about national identity or cultural identity, they imply that their identity is partly a matter of what they share with other people. In this case, identity is about identification with others whom we assume are similar to us at least in some significant ways.

These two aspects of identity (personal identity and social identity) have been considered by researchers for identity resolution. However the previous identity models have been suffering from some limitations which can affect the accuracy of the results. In fact, individuals are not isolated but interconnected to each other 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 for identity resolution.

For instance, Ananthakrishna, Chaudhuri and Ganti [12] introduced a method that eliminates duplicates in data warehouses using a dimensional hierarchy over the link relations. This method can improve the performance of the matching technique by only comparing those attribute values that have the same foreign key dependency. For instance, the similarity of two identity will be analysed only when both of them live in the same city. Afterwards, Kalashnikov, Mehrotra and Chen [13] combined co-affiliation and co-authorship relationships and created a new resolution model for reference disambiguation. In another research, Köpcke and Rahm [14] categorized entity resolution methods into context matchers and attribute value matchers. They explain that attribute value matchers rely on descriptive attributes, while context matchers consider information inferred from social interactions which is represented as linkages in a graph.

This research introduces a new identity model for identity resolution that incorporates four key aspects including physical, social, official, and virtual. By expanding the scope to include official and virtual dimensions, the model integrates a broader range of attributes, enhancing both the accuracy and reliability of identity resolution. Details of this identity model will be further elaborated in the methodology section.

2.2. *Identity resolution methods*

The techniques that are used in existing identity resolution methods can be categorized into three groups including (1) rule-based comparisons, (2) distance measures, and (3) graph analysis. Most of the rule-based identity resolution methods have been developed based on the matching rules. As an example, for a simple rule, two identity records match only if their first name, surname, and date of birth values are identical [15]. Li and Wang [2] explained that matching rules try to have high precision, but they usually suffer from low sensitivity in detecting true matches. This is because of data quality issues such as missing data, entry error and deceptions. They also discussed that the most important challenge for a rule-based method can be creation of the rule set because creating an effective and comprehensive rule set can be very complicated and time-consuming and the rules may not be portable and applicable across different contexts.

Developing a comprehensive set of rules for identity resolution can be both time-consuming and costly. Additionally, the applicability of these rules may be limited, as they often pertain to specific domains and may not be transferable to others. In light of these challenges, machine learning offers a viable alternative. By automatically identifying patterns in training datasets that contain matching pairs, machine learning facilitates the creation of adaptable models for resolving new identity records, thereby streamlining the process and enhancing flexibility across various domains. Li and Wang [2] explained that when there is a pair of identity records, distance measures can be defined for different descriptive attributes, and then they can be combined into an overall score. The overall distance score will be compared to a pre-defined threshold, and the pair should be considered as a match if this score is below or above the threshold.

One of the first identity resolution methods was a data association method for linking criminal records that possibly refer to the same suspect [16]. This method was comparing two different records and calculating an overall distance measure as a weighted sum of the distance measures of all corresponding feature values. In another attempt, Wang, Chen and Atabakhsh [17] proposed a record linkage method which was detecting misleading identities by comparing four attributes and combining them into an overall distance score. These attributes were as follows (1) name, (2) date of birth, (3) social security number, and (4) address. They used a supervised learning method to determine a threshold for match decisions. This was done via using a set of identity pairs which were labelled by an expert. Wang et al. [18] discussed that these methods perform based on a limited number of descriptive attributes and the most important issue in this case is that they tend to fail if one or more of the considered attributes contains missing values.

In another study, a graph-based method for entity resolution was proposed by Bhattacharya and Getoor [19]. This method established a distance measure that integrates graph-based relational similarity with corresponding attribute similarities between each pair of entity references. Subsequently, this approach was expanded to develop a collective entity resolution method. As a result, instead of simply making pair-wise entity comparisons, they could derive new social information and incorporate it into further resolution process repeatedly. However, there were concerns that this method might struggle to accurately identify an individual when multiple profiles across various social media platforms are involved. To mitigate this issue, various techniques were developed specifically for matching user profiles on social media. For instance, a CRF-based approach was proposed by Bartunov et al. [20]. They created two user graphs, one based on user profile attributes and the other on social connections, and then merged these graphs. This integration effectively demonstrated how social information can enhance the performance of identity resolution when included in matching algorithms. In recent years, using different social media platforms such as Facebook, Twitter and LinkedIn has been growing greatly, and researchers have realised that further research on user identity linkage across online networks is required to be able to match one individual to all their online identities [21]. As a result, researchers have started to develop different methodologies and use neural networks and graph analysis to tackle issues related to user identity linkage and identity resolution on social media [22,23]. However, in the digital age, privacy and human rights are increasingly under the spotlight. As countries navigate the complex landscape of surveillance and data privacy, they face the challenging task of balancing the protection of individual rights with the implementation of national security measures. This delicate equilibrium seeks to uphold personal freedoms while ensuring collective safety [24]. Considering privacy-related concerns, we made sure that the policing dataset used in this research is fully anonymised and the individuals in this dataset are not identifiable. The anonymisation had taken place by the police forces before providing the dataset to be used in this research. The use of AI-based methodologies has also raised concerns about bias in the developed algorithms, specifically regarding ethnicity. These concerns have been considered during the development process of the proposed methodology in this research which is going to be used on policing data where physical attributes such as gender or ethnicity are playing an important role.

3. Methodology

3.1. SPIRIT policing dataset

SPIRIT policing dataset is an anonymised dataset which has been used in the SPIRIT project funded by the European Union’s Horizon 2020. This dataset includes 891 identities, and each identity has 30 different attributes. Eight of these attributes will be considered in this research, which are as follows (1) postcode, (2) date of birth, (3) town, (4) offence, (5) gender, (6) street name, (7) district, and (8) ethnicity. These attributes have been selected as the most highly valued attributes in this dataset based on the advice from the end users in SPIRIT project consortium including (1) West Midland Police Authority (UK), (2) Crime Commissioner for Thames Valley (UK), (3) Ministry of Interior in Serbia, (4) Hellenic Police in Greece, (5) Police Academy in Szczytno (Poland), and (6) Antwerp Police in Belgium. There are two ‘known’ identities in this dataset who are using two different names, ‘Billy Smith’ and ‘Mariet Snehh’, but they both belong to the same person.

3.2. The proposed identity model

Identity refers to those attributes that enable us to recognise an individual from others. As discussed in section 1.2, this research introduces a new identity model that incorporates four categories of attributes. This model extends beyond the traditional physical and social aspects by including two additional aspects including official and virtual, thereby broadening the scope of attributes considered. This comprehensive approach enhances both the accuracy and reliability of identity resolution. The first category is physical identity which includes characteristics which an object or person is definitively recognizable or known by. O'Neill [25] defined physical identity as 'the innate human drive we are all born with to move our bodies through space'. In other words, the first thing we see when we look at someone could be the factors relevant to their physical appearance, such as their face, hair, height, weight, skin tone, eye color, and other physical traits [26]. These are all examples of physical characteristics or attributes. The second category is official identity which is the identity that carries a legal status, usually issued by governments to their citizens. Official identity encompasses essential personal identifiers such as name and date of birth. Typical examples include documents like birth certificates, national identity cards, social security numbers, and voter registration cards.

The third category is virtual identity which is the identity created by human user that acts as an interface between physical person and virtual person that other users see on their computer screen. It is a model for self-expression, and tools for virtual interaction and a representation of a user in a virtual world. In fact, when a virtual identity is shaped, ethical issues in social media begin as the user can create a picture of him or herself as he or she would like to be, not what he or she really is. It becomes even more extreme when people in the real world cannot or do not want to show themselves as they really are, while if they express their true opinions, they face penalties [27]. Virtual identity can be described by various elements, such as representative images or video content, a name, a detailed profile of the account holder, lists of friends, or the groups and communities an account is associated with.

Finally, the fourth category is social identity which is a set of behavioral or personal characteristics by which an individual is recognizable as a member of a group. In other words, social identity refers to conceptualising the identification with different social groups as an integrated part of the self [28]. In reality, people adopt the identity of a group they have categorised themselves as belonging to. For instance, if you categorise yourself as a student, most probably you will adopt the identity of a student and try to conform to the norms of the group and begin to act in the way you believe students act. As a result, there might be an emotional significance to your identification with a group, and your self-esteem will become bound up with group membership [29].

3.3. Graph creation

Eight graphs will be created after selecting eight highly valued attributes which were mentioned in section 3.1. In this casefor instance, if four identities have the same postcode, the graph shows that these four identities are connected to each other. In these graphs, the nodes will be presenting a person with his/her first name and surname, and each edge will be showing that there is a similarity between two nodes (person).

3.4. Community detection algorithm

After the graph creation step, the Louvain algorithm will be used for community detection based on the selected attributes. This modularity-based community detection algorithm can provide satisfactory performance for hierarchical community structures in terms of detection efficiency [30]. This method is a very efficient method for identifying communities in large networks. Blondel et al. [31] mentioned that the Louvain method has been used successfully for analyzing different types of networks and for sizes up to 100 million nodes and billions of links. They also pointed out that the analysis of a typical network of 2 million nodes takes 2 min on a standard PC. In fact, this method is a greedy optimization method which tries to optimize the modularity of a partition of the network [31]. Modularity is a metric that can be used to quantify the quality of an assignment of nodes to communities. In other words, modularity can be defined as a value between −1 and 1 that measures the density of links inside communities compared to links between communities [32]. For a weighted graph, modularity is defined as:

$$M = \frac{1}{2k} \sum_{xy} \left[Q_{xy} - \frac{p_x p_y}{2k} \right] \delta\left(n_x, n_y\right) \tag{1}$$

In this equation, Q_{xy} represents the edge weight between nodes x and y. p_x and p_y are the sum of the weights of the edges attached to nodes x and y. k is the sum of all the edge weights in the graph. n_x and n_y are the communities of the nodes and δ is a Kronecker delta which is a function of two variables. It is 1 if the variables are equal and it is 0 if the variables are not equal. Equation (2) explains this.

$$\delta_{\boldsymbol{wt}} \begin{cases} 0 \; xf \;\; w \neq t \\ 1 \; xf \;\; w = t \end{cases} \tag{2}$$

In the Louvain algorithm, optimisation will be performed in two steps. In the first step, small communities will be found by optimizing modularity locally. Then in the second step, the nodes which belong to the same community will be cumulated and a new network will be built where its nodes are the communities. These steps will be repeated until a maximum of modularity is achieved, and a hierarchy of communities is produced [31]. In other words, in the first step, each node in the network will be assigned to its own community. Then for each node x, the change in modularity will be calculated for removing node x from its own community and moving it into the community of each neighbor y of x. Equation (3) explains the process of inserting x to the community of y.

$$\Delta M = \left[\frac{\sum_{in} + 2p_{x,in}}{2k} - \left(\frac{\sum_{tot} + p_x}{2k} \right)^2 \right] - \left[\frac{\sum_{in}}{2k} - \left(\frac{\sum_{tot}}{2k} \right)^2 - \left(\frac{p_x}{2k} \right)^2 \right] \tag{3}$$

In this equation, ∑in is the sum of all the weights of the links inside the community that node x is moving into. ∑tot is the sum of all the weights of the links to nodes in the community that node x is moving into. Moreover, the weighted degree of node x is represented by p_x and the sum of the weights of the links between node x and other nodes in the community that x is moving into, is represented by p_x. Finally, k is the sum of the weights of all links in the network. Table 1 shows some of the most important studies used Louvain method for community detection.

3.5. Investigating potential targets

A key aspect of network analysis is identifying the most influential nodes within a graph. Newman [38] explained that 'centrality' is a term that can be used to describe the importance of individual nodes in a graph and 'degree of a node' is the number of edges that it has. The nodes with more connections are more influential and important in a network. As a result, the person with more friends in a social graph, is the one that is more central. Thus, in the next step of our method, eight different lists of names will be provided based on the measurement of centrality and the degree of nodes in each graph. Top 20 nodes based on their degree (number of connections that they have) will be recorded in each list. Then these lists will be compared with each other to find similar identities. If any identity is repeated in at least five lists, it will be recorded in a new list as a potential target. Following this step, a new list of all related identities for the potential targets will be provided.

3.6. Phonetic algorithms

In the next step, a cascading method will be used for applying three phonetic algorithms including (1) Soundex, (2) Metaphone and (3) Jaro-Winkler on the potential targets and their relevant identities in order to detect any possible human errors during data entry or wrong information given by the person. These three phonetic algorithms are designed to index names based on their

Table 1. Louvain method for community detection.

Project	Number of nodes	Source
Twitter social network	2.4M	Divide and Conquer: Partitioning Online Social Networks [33].
Mobile phone networks	4M	Tracking the Evolution of Communities in Dynamic Social Networks [34].
Flickr	1.8M	Real World Routing Using Virtual World Information [35].
LiveJournal	5.3M	
YouTube	1.1M	
Citation network	6M	Subject clustering analysis based on ISI category classification [36].
LinkedIn social network	21M	Mapping search relevance to social networks [37].

sound and will be sequentially applied to analyse the potential targets and their corresponding identities. This means that in the first cycle, Soundex method will be applied. Then in the second cycle Metaphone will be applied to the results of Soundex method. Finally, in the third cycle, the Jaro-Winkler method will be applied to the results of the Metaphone method. Thus, we narrow down the results to get the best output. As a result, all similar first names and surnames to the potential target identities and their relevant identities, with a potential of being manipulated will be detected to be considered in the next step.

3.6.1. Soundex phonetic algorithm

The Soundex phonetic algorithm can be used for indexing names by sound, as they are pronounced in English [39]. It has been mainly used in applications that involve searching for people's names and other tasks presenting typing errors due to phonetic similarity [40]. The main goal of this phonetic algorithm is to encode homophones into the same representation so that they can be matched despite minor differences in spelling. The algorithm mainly encodes consonants, and a vowel will not be encoded unless it is the first letter [41]. In other words, the algorithm evaluates each letter in the input word and then a numeric value will be assigned to that word. As a result, each word will be converted into a code made up of four elements [39]. Soundex uses numeric codes explained in Table 2, for each letter of the string to be codified.

In the first step, the algorithm replaces all but the first letter of the string by its phonetic code. Following this step, any adjacent reptations of codes will be eliminated. Moreover, all occurrences of code 0 will be eliminated which is for eliminating vowels. Then the first four characters of the resulting string will be returned [42].

3.6.2. Metaphone phonetic algorithm

The Metaphone phonetic algorithm contains several improvements over the Soundex metaphone algorithm. These improvements have been done by reducing English words to their basic sounds fundamentally. This algorithm has more sensitivity towards changes in the sequence of the letters as well as concerned information about inconsistencies and variations in spelling and pronunciation [43]. As a result, it can produce a more accurate

Table 2. Soundex phonetic codes for the English language [42].

Letter	Numeric code
a, e, i, o, u, y, h, w	0
b, p, f, v	1
c, g, j, k, q, s, x, z	2
d, t	3
l	4
m, n	5
r	6

encoding. The Metaphone algorithm uses an inventory of 16 consonants, 0BFHJKLMNPRSTWXY. In this inventory, 0 stands for/θ/and X for/ /or/t /. In fact, all 21 orthographic English consonants are mapped to these 16 consonants by collapsing some letters like <d> and <*t*> to <*t*> [44]. The vowels AEIOU are also used, but only at the beginning of the code. The main improvement in Metaphone algorithm compared to Soundex algorithm is about cases that a letter can be pronounced in two different ways. For instance, the letter <c> is sometimes pronounced as/s/and sometimes as/k/ and the Metaphone algorithm covers these kinds of cases, whereas Soundex algorithm does not because of its more simplistic mapping strategy [45].

3.6.3. Jaro-Winkler phonetic algorithm

The Jaro-Winkler phonetic algorithm is generally used for checking duplicate words. This algorithm is developed from the basic Jaro distance algorithm, which is a string-matching algorithm to find the similarity between two words. The basic Jaro distance algorithm was introduced by Jaro [46] and it works based on three steps including (1) Compute the length of words, (2) Find the number of similar characters between two words, and (3) Calculate the number of transpositions. The transposition is the measure of how many similar characters in the two words are out-of-order. In Jaro distance algorithm, the Equation (1) is used to measure the similarity score $(\boldsymbol{d_j})$ between two words $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

$$\boldsymbol{d_j} = \begin{cases} 0 \\ \frac{1}{3}\left(\frac{\boldsymbol{m}}{|\boldsymbol{S}_1|} + \frac{\boldsymbol{m}}{\boldsymbol{S}_2} + \frac{\boldsymbol{m}-\boldsymbol{t}}{\boldsymbol{m}}\right) \end{cases} \tag{4}$$

In Equation (4), m is the number of the same character with the same position both in the first and the second word where the distance is not more than 1 character. $|\boldsymbol{S}_1|$ and $|\boldsymbol{S}_2|$ are the length of the first and the second word, and $\boldsymbol{t}$ is $\frac{1}{2}$ of the transpose character number. $\boldsymbol{d_j}$ value will be 0 when $\boldsymbol{m} = 0$.

Later, the basic Jaro algorithm was modified by Winkler and Thibaudeau [47] and they stated that if the prefix is common in two words, then the similarity score is increased [48]. Yancey [49] explained that this enhancement in the basic Jaro algorithm is based on the observation that most common typographic variations occur towards the end of a string. The Jaro-Winkler similarity score $(\boldsymbol{d_w})$ can be calculated using Equation (5).

$$\boldsymbol{d_w} = \boldsymbol{d_j} + \left(\ell \boldsymbol{p}\left(1 - \boldsymbol{d_j}\right)\right) \tag{5}$$

In this equation, $\boldsymbol{d_j}$ is the Jaro similarity for words $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. ℓ is the length of common prefix at the start of the word up to a maximum of 4 characters. $\boldsymbol{p}$ (Prefix scale) is a constant value which is used to show how much the score is adjusted upwards for having common prefixes. The value for this constant scaling factor should not exceed 0.25 as the similarity could become larger than 1. The standard value for $\boldsymbol{p}$ is 0.1.

3.7. Comparison process

After applying Soundex, Metaphone and Jaro-Winkler algorithms using a cascading method, all potential targets, and their relevant identities as well as similar identities with the potential of using manipulated forenames and surnames will be added to a new dataset for comparison purpose. All attributes of the potential targets and their relevant identities in the new dataset will be compared separately, and similarity will be scored. This means that all 30 attributes for each one of these identities in the SPIRIT policing dataset will be considered for comparison and scoring process. The same identities will be investigated based on the similarity scores.

3.8. Implementation of the proposed methodology

The proposed methodology in this research has been implemented in Python 3.7 using Anaconda IDE. The Python data analysis library (pandas) has been used for sorting and analysing the dataset. Moreover, the NetworkX library has been used for graph analysis, and the results have been visualised using pyplot from the matplotlib library.

4. Results and discussion

As it was explained in the methodology, eight attributes in the SPIRIT policing dataset were selected, and they are including (1) postcode, (2) date of birth, (3) town, (4) offence, (5) gender, (6) street name, (7) district, and (8) ethnicity. As a result, in the first step 8 graphs were created. Figure 1 shows one of these graphs, which was created based on the same postcodes and it shows that Billy Smith, Lorret Denhart and Mariet Snehh have been using the same postcode. Moreover, Figure 2 shows all eight created graphs.

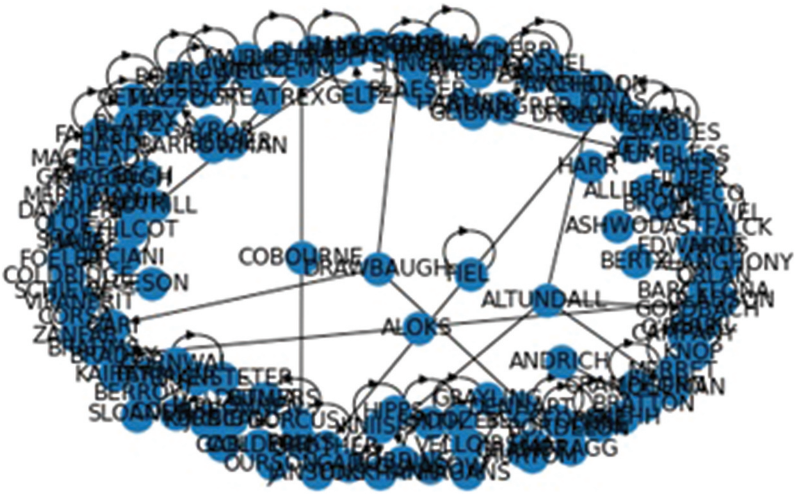
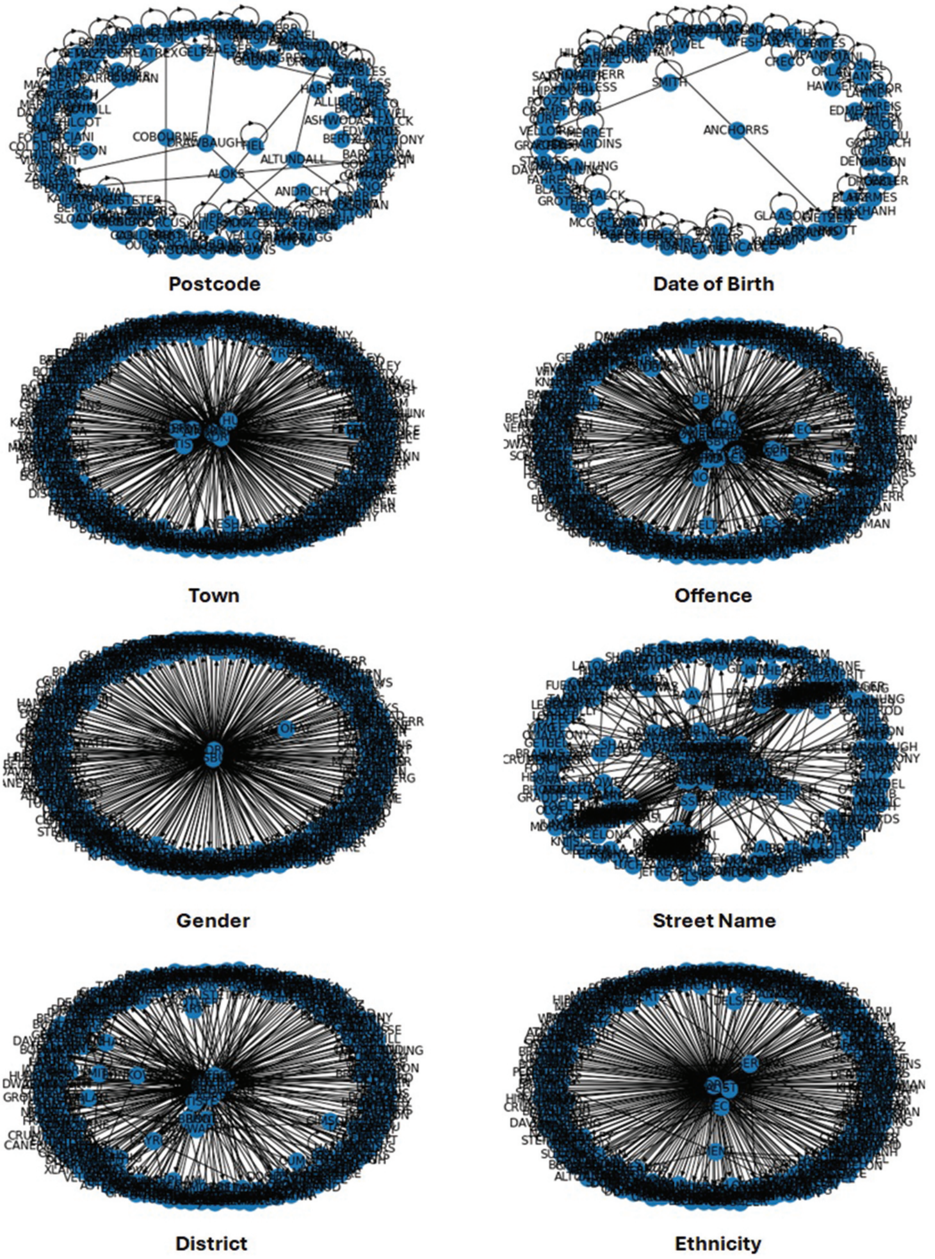


Figure 1. Graph based on the same postcode.



Postcode

Date of Birth

Town

Offence

Gender

Street Name

District

Ethnicity

Figure 2. Graphs for eight selected attributes.

In the second step, the Louvain algorithm was used for community detection based on the eight selected attributes. Figure 3 shows the community detection graphs.

Following this step, centrality, and the degree of nodes in each graph were measured, and top 20 nodes in each graph were recorded in eight different lists. These lists were compared, and those identities which were repeated in at least

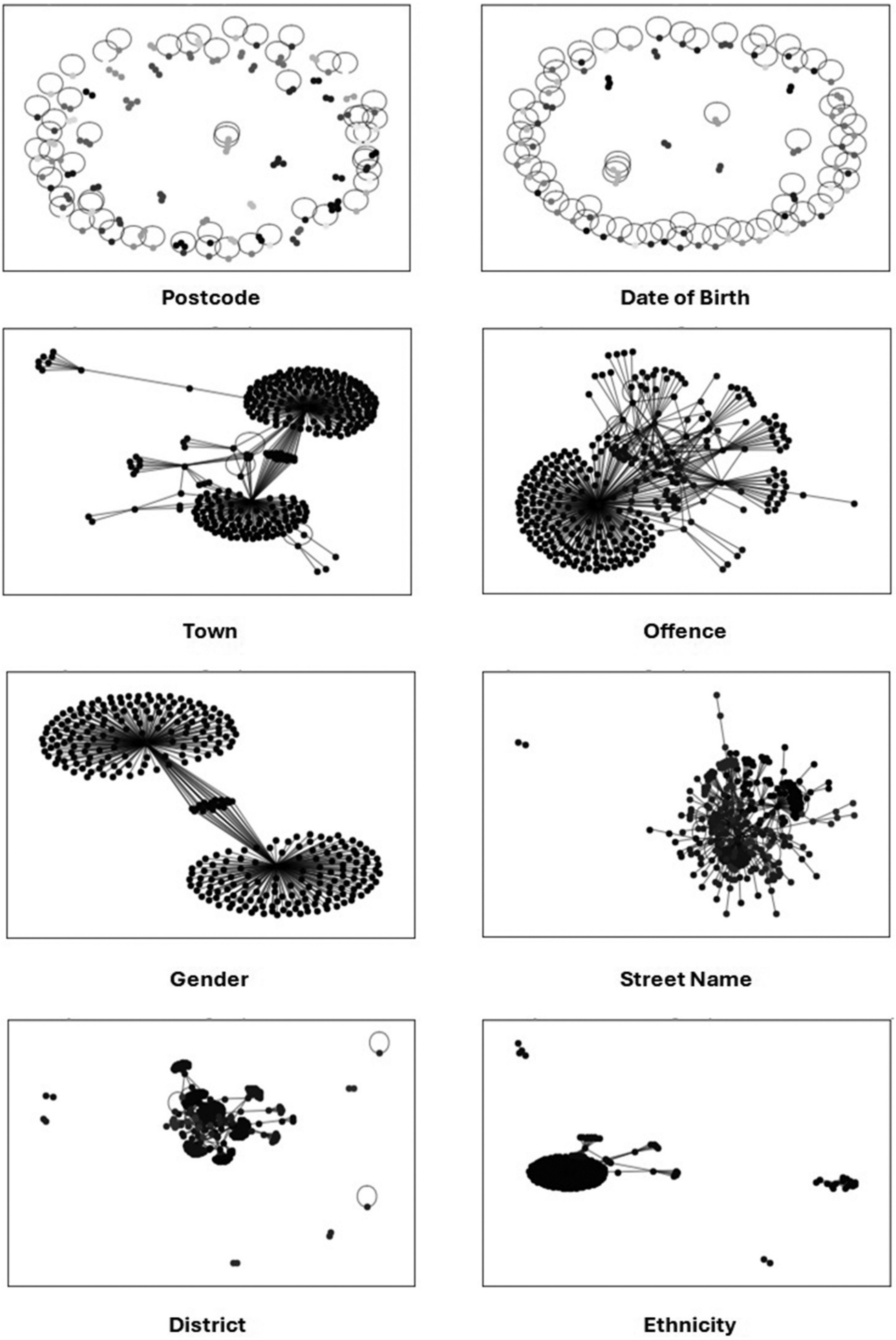


Figure 3. Community detection graphs.

five lists were recorded in a new list for potential targets. Table 3 shows the surnames which were repeated in at least five lists.

Table 3. Surnames which were repeated at least in 5 lists.

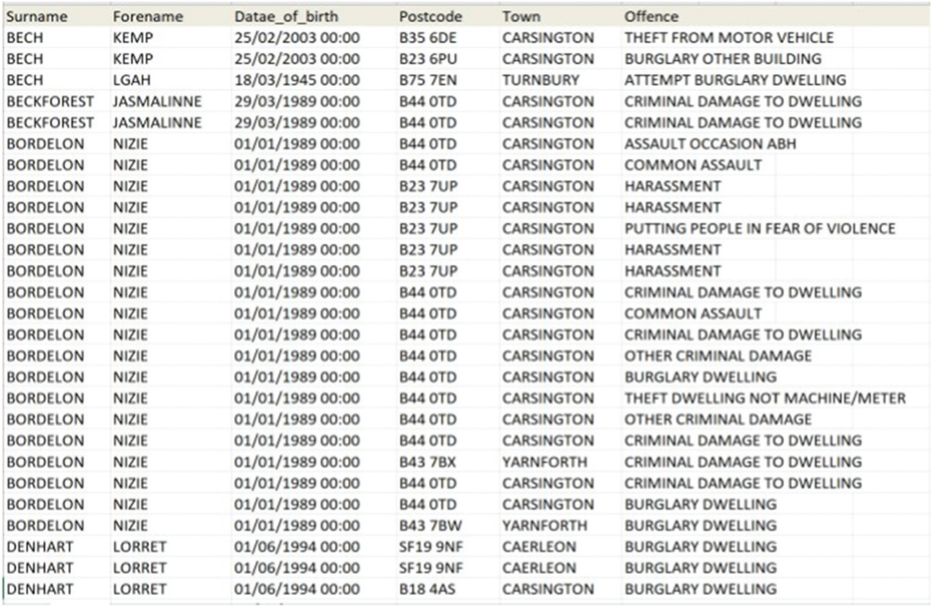
Top 20 nodes in				
Postcode graph	Town graph	Date of birth graph	Offence graph	Street name graph
ALTUNDALL	AINSBURY	**SMITH**	AINSBURY	BESSER
DROACH	ADRE	VELLOIRS	ALTUNDALL	ADRE
FIEL	ARTHURS	BOWLES	AMENT	GLAASOW
GROTHER	BATISTE	DENHART	ALOKS	DICIANI
BECKFOREST	**BORDELON**	GLAASOW	VELLOIRS	BALOW
BESSER	BOSSERMAN	KER	ADRE	**SMITH**
BLAESER	DENHART	MERRET	**SNEHH**	CRECO
BORDELON	GAYROR	ANCHORRS	ASTFALCK	**SNEHH**
DICIANI	VELLOIRS	ASTFALCK	BECKFOREST	FETTES
DORCUS	**SMITH**	**SNEHH**	BALOW	BRAHMS
DRAWBAUGH	**SNEHH**	BARCELONA	**BORDELON**	CRECO
GAYROR	BECH	BECH	ALOI	DORCUS
GELTZ	GROTHER	BECKFOREST	CRECO	FAHREN
GETEL	FETTES	BERROW	FAHREN	KER
JUNCALL	BHOTT	BHOTT	KER	**BORDELON**
KER	BUTLAND	BLAESER	**SMITH**	GETEL
SNEHH	CORSA	BLATZ	GOSNEL	JUNCALL
KNIISIS	DEGNER	**BORDELON**	BARCELONA	GAYROR
POOZEY	DORCUS	BRADWICK	GAYROR	KNIISIS
SMITH	FONES	BRAHMS	FILLINGHAM	SATVINGRER

Table 4. Potential target names and their identities.

Loret Denhart
Billy Smith
Nizie Bordelon
Mariet Snehh
Jasmalinne Beckforest
Kemp Bech

According to Table 3, three surnames including Smith, Snehh and Bordelon were appeared in at least five lists for top 20 nodes. These lists are related to (1) postcode graph, (2) town graph, (3) date of birth graph, (4) offence graph, and (5) street name graph. The relevant surnames to these surnames were detected in the next step based on their connections in different graphs, and they were added to a new dataset after applying the phonetic algorithms. Table 4 shows these identities.

This new dataset encompasses all 30 attributes for each identity. Figure 4 shows a screenshot from a part of the new dataset, which was created after applying Soundex, Metaphone and Jaro-Winkler algorithms. As Figure 4 shows, some of the names in this dataset are repeated. The reason is that some of the values of their different attributes are different. For instance, there are two rows for Kemp Bech. This is because there are two different postcodes related to this name and this person was committed two different offences. As a result, there are two rows related to this person with different values for two attributes including postcode and offence. Finally, every single row of this dataset was compared with other rows, and based on the number of attributes which had the same value,



Surname	Forename	Datae_of_birth	Postcode	Town	Offence
BECH	KEMP	25/02/2003 00:00	B35 6DE	CARSINGTON	THEFT FROM MOTOR VEHICLE
BECH	KEMP	25/02/2003 00:00	B23 6PU	CARSINGTON	BURGLARY OTHER BUILDING
BECH	LGAH	18/03/1945 00:00	B75 7EN	TURNBURY	ATTEMPT BURGLARY DWELLING
BECKFOREST	JASMALINNE	29/03/1989 00:00	B44 0TD	CARSINGTON	CRIMINAL DAMAGE TO DWELLING
BECKFOREST	JASMALINNE	29/03/1989 00:00	B44 0TD	CARSINGTON	CRIMINAL DAMAGE TO DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	ASSAULT OCCASION ABH
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	COMMON ASSAULT
BORDELON	NIZIE	01/01/1989 00:00	B23 7UP	CARSINGTON	HARASSMENT
BORDELON	NIZIE	01/01/1989 00:00	B23 7UP	CARSINGTON	HARASSMENT
BORDELON	NIZIE	01/01/1989 00:00	B23 7UP	CARSINGTON	PUTTING PEOPLE IN FEAR OF VIOLENCE
BORDELON	NIZIE	01/01/1989 00:00	B23 7UP	CARSINGTON	HARASSMENT
BORDELON	NIZIE	01/01/1989 00:00	B23 7UP	CARSINGTON	HARASSMENT
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	CRIMINAL DAMAGE TO DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	COMMON ASSAULT
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	CRIMINAL DAMAGE TO DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	OTHER CRIMINAL DAMAGE
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	BURGLARY DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	THEFT DWELLING NOT MACHINE/METER
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	OTHER CRIMINAL DAMAGE
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	CRIMINAL DAMAGE TO DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B43 7BX	YARNFORTH	CRIMINAL DAMAGE TO DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	CRIMINAL DAMAGE TO DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B44 0TD	CARSINGTON	BURGLARY DWELLING
BORDELON	NIZIE	01/01/1989 00:00	B43 7BW	YARNFORTH	BURGLARY DWELLING
DENHART	LORRET	01/06/1994 00:00	SF19 9NF	CAERLEON	BURGLARY DWELLING
DENHART	LORRET	01/06/1994 00:00	SF19 9NF	CAERLEON	BURGLARY DWELLING
DENHART	LORRET	01/06/1994 00:00	B18 4AS	CARSINGTON	BURGLARY DWELLING

Figure 4. New dataset including all potential targets and their relevant identities to be used for comparison process.

a score was assigned to show the similarity between each two identities. After comparing these scores, it was realized that two identities including Billy Smith and Mariet Snehh have the most similarity. As it was explained in section 3.1 of the methodology, these two were the 'known' identities in SPIRIT policing dataset who had two different names, but they both belonged to the same person. Thus, the system was successful in resolving their identities.

5. Additional experimental evaluation

To further assess the experimental outcomes, we refined the SPIRIT policing dataset by excluding two 'known' identities who had different names, but they both belonged to the same person. Subsequently, we introduced two new false identities by selecting an individual in the dataset with multiple records. We altered the first and last names in several of these records, thereby representing a single identity with two different names, including 'Alex Wilson' and 'Sarah Williams'. We also extended the dataset by adding additional true identities. It's crucial to note that due to the unavailability of other policing datasets, which may also differ structurally, we were compelled to utilize the same dataset, albeit with varied false identities and extended by new true identities.

As it was explained in the methodology, in the first step of the new experiment, eight graphs were created based on the eight selected attributes in the dataset. Figure 5 shows all eight created graphs in the new experiment.

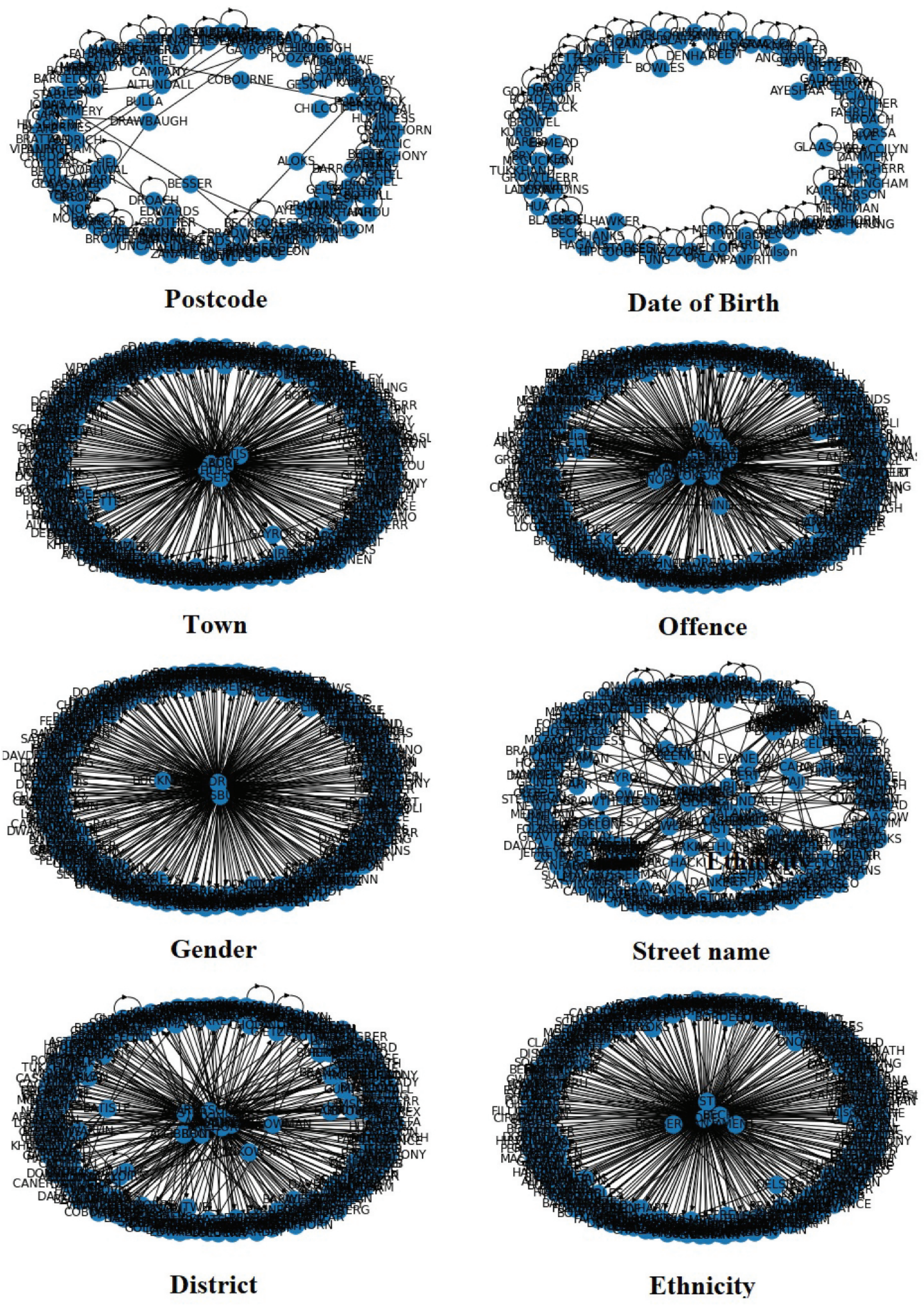
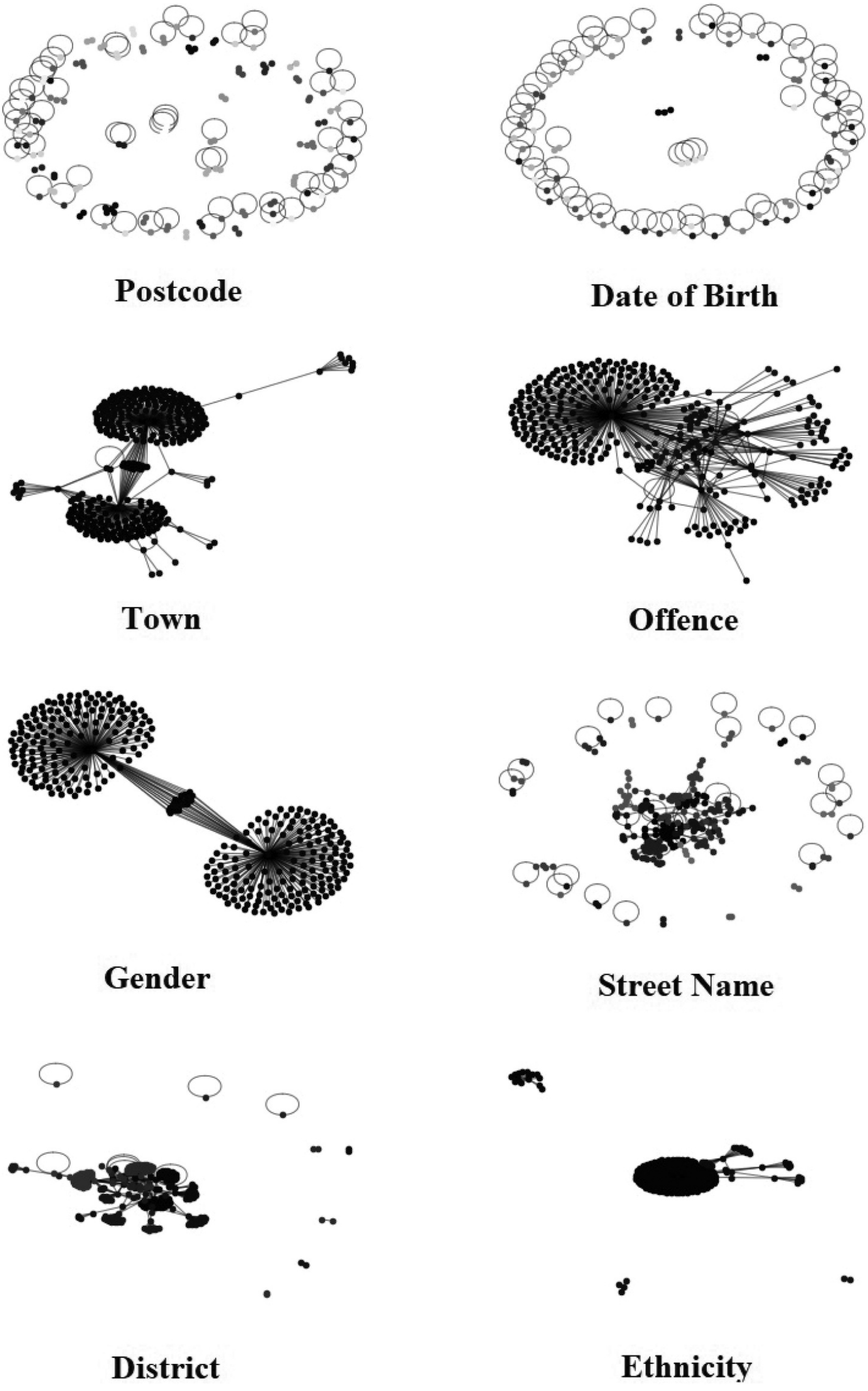


Figure 5. Graphs for eight selected attributes.

The Louvain algorithm was then used in the second step for community detection based on the eight selected attributes. Figure 6 shows the community detection graphs.

Following this step, centrality, and the degree of nodes in each graph were measured, and top 20 nodes in each graph were recorded in eight different lists.



Postcode

Date of Birth

Town

Offence

Gender

Street Name

District

Ethnicity

Figure 6. Community detection graphs.

Table 5. Surnames which were repeated at least in five lists.

Top 20 nodes in				
Postcode graph	Date of birth graph	Offence Graph	Street name graph	ethnicity graph
ALTUNDALL	BOWLES	AINSBURY	BARWISSE	ADRE
DROACH	**VELLOIRS**	ALTUNDALL	ALTUNDALL	BECH
FIEL	DENHART	AMENT	**Wilson**	ALISTER
GROTHER	GLAASOW KER	ALOKS	ASHWOD	BOSSERMAN
Williams	MERRET	ADRE	DENHART	AMENT
BECKFOREST	**Williams**	ASTFALCK	GOSNEL	DELSIE
BESSER BLAESER **Wilson**	ANCHORRS	**Williams**	MERRET	ZANFAR
DICIANI	ASTFALCK	**Wilson**	COLDRIDGE	BLAESER
DORCUS	BARCELONA	BECKFOREST	ADRE	BLATZ
DRAWBAUGH	BECH	BALOW	**VELLOIRS**	BROWEL
GAYROR	BECKFOREST	TUKKHANH	BERTZ	**VELLOIRS**
GELTZ	BERROW	**VELLOIRS**	BLAESER	CANTWEL
VELLOIRS	BHOTT	CRECO	DEGNER	DEGNER
GETEL JUNCALL	BLAESER BLATZ	ALOI	ALISTER	**Wilson**
KER	**Wilson** BRADWICK	FAHREN	**Williams**	GROWTHERR
KNIISIS POOZEY	BRAHMS BROWEL	KER	BARROWMAN	HARMES
		BARCELONA	BECH	HARPHAM
		DENHART	ALOI	TUKKHANH
		GOSNEL	BOWLES	**Williams**
		BRADWICK	GAYROR	DNOA

The lists were compared, and the repeated identities in at least five lists were recorded in a new list for potential targets. The surnames which were repeated in at least five lists are presented in Table 5.

Table 5 shows that three surnames, including Williams, Wilson and Velloirs were appeared in at least five lists for top 20 nodes. These lists are related to (1) postcode graph, (2) date of birth graph, (3) offence graph, (4) street name graph, and (5) ethnicity graph. In the next step, the relevant surnames to these surnames were detected based on their connections in different graphs. These detected surnames were then added to a new dataset after applying the phonetic algorithms. These identities are presented in Table 6.

A new dataset was constructed that incorporated all 30 attributes for each one of these identities. To assess similarities, algorithms such as Soundex, Metaphone, and Jaro-Winkler were applied to the data. Each record was then methodically compared against the others, with a similarity score assigned based on the number of attributes sharing identical values. This comparison revealed that two identities – Alex Wilson and Sarah Williams – exhibited the highest similarity. These identities are the new 'known' identities which were

Table 6. Potential target names and their identities.

Sarah Williams
Iezi Zanat
Dmytro Zanfar
Alex Wilson
Galit Velloirs

added to the refined and extended dataset, and they are related to the same individual. This discovery confirmed the efficacy of the applied method in successfully resolving identity discrepancies.

The policing dataset used in this research was expanded during the SPIRIT EU Horizon 2020 project, and a larger revised dataset of 1,145,418 records containing 694,264 identities was used in future research. Using the concept of a graph-based approach from this paper, a targeted rule and graph-based approach was developed and applied to this dataset with promising results – detecting four of five known false identities as part of 51 suspected false identities when investigating 23 targets [50]. The methodology works by first finding close matches based on a rule-based matching system with a known target and then narrowing down potential suspects by using graph analysis to link identities through the crimes that they have been involved in. We believe this research could also be continued by implementing deep learning algorithms and comparing their performance with the methodology used in this paper and in [50]. There is potential that combining this graph-based methodology with deep learning methods could improve the performance, and we are interested to investigate this in further steps of this research.

6. Conclusion

This paper presents a novel graph-based approach for identity resolution. Graph analysis techniques such as community detection and centrality measurement have been used in the proposed methodology. In addition, a new identity model for identity resolution has been presented in this research for the first time. This identity model represents four different types of attributes including (1) physical attributes, (2) social attributes, (3) official attributes, and (4) virtual attributes. The presented methodology has been tested on the SPIRIT policing dataset, which is an anonymised dataset used in the SPIRIT project funded by the European Union’s Horizon, Grant Number 786,993. This dataset comprises 892 identity records, including two ‘known’ identities belonging to the same individual but using different names. The methodology presented in this research successfully detected these two identities within the SPIRIT policing dataset. We also conducted an additional experimental evaluation on a refined and extended version of the dataset and the presented methodology successfully detected the new false identities which were added to the dataset. The presented identity resolution approach in this paper can effectively facilitate the investigation process for police forces and assist them to find criminals and individuals who committed fraud using a false identity. It can also be useful for other similar datasets which contain identity records related to other fields such as finance and banking, customer service or marketing.

Acknowledgments

This work was supported by the European Union Horizon SPIRIT project, Grant Number 786993. We would like to thank all colleagues who provided us with comments to improve the research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by the European Union's Horizon 2020 (grant number 786993).

ORCID

Mohammad Hossein Amirhosseini http://orcid.org/0000-0002-3404-084X

Data availability statement

The data that support the findings of this study are available from the second author upon reasonable request.

Author contributions

Conceptualization, First Author and Second Author; Data curation, First Author and Second Author; Formal analysis, First Author; Funding acquisition, Second Author; Investigation, First Author; Methodology, First Author; Software, First Author; Supervision, Second Author; Validation, First Author and Third Author; Visualization, First Author; Writing – original draft, First Author; Writing – review & editing, First Author, Second Author and Third Author.

Ethical approval

The anonymized dataset used in this research has been provided by the police forces involved in the EU Horizon 2020 SPIRIT project who also provided the required approvals.

References

[1] Bilgrami A. Notes toward the definition of "identity. Daedalus. 2006;135(4):5–14. doi: 10.1162/daed.2006.135.4.5

[2] Li J, Wang AG. A framework of identity resolution: evaluating identity attributes and matching algorithms. Secur Inform. 2015;4(6):015–0021–0. doi: 10.1186/s13388-015-0021-0

[3] Kean TH, Kojm CA, Zelikow P, et al. The 9/11 Commission report. 2004. http://govinfo.library.unt.edu/911/report/index.htm

[4] U.S. Department of State. Country reports on terrorism. 2006. http://www.state.gov/j/ct/rls/crt/2006/

[5] Stroebel L, Llewellyn M, Hartley T, et al. A systematic literature review on the effectiveness of deepfake detection techniques. J Cyber Secur. 2023;7(2):83–113. doi: 10.1080/23742917.2023.2192888

[6] Li J, Wang GA, Chen H. Identity matching using personal and social identity features. Inform Syst Front. 2010;13(1):101–113. doi: 10.1007/s10796-010-9270-0

[7] Vukovic I. Truth-value unconstrained face clustering for identity resolution in a distributed environment of criminal police information systems. Eng Appl Artif Intell. 2023;124:106576. doi: 10.1016/j.engappai.2023.106576

[8] Kretz DR, Paulk RW. Establishing traveler identity using collective identity resolution. In: 2010 IEEE International Conference on Technologies for Homeland Security (HST); Waltham MA; 2010. p. 308–313.

[9] Kidd W, Teagle A. Culture and identity, skill-based sociology. Bloomsbury Publishing; 2012. Available from: https://books.google.co.uk/books?hl=en&lr=&id=0SJHEAAAQBAJ&oi=fnd&pg=PR1&dq=identity&ots=h0t2TN9Fb1&sig=6uXDHJoy3NxafGZJrCStc_fidJw&redir_esc=y#v=onepage&q=identity&f=false

[10] Cheek JM, Briggs SR. Self-consciousness and aspects of identity. J Res Pers. 1982;16(4):401–408. doi: 10.1016/0092-6566(82)90001-0

[11] Buckingham D. Introducing identity." youth, identity, and digital media. In: Buckingham D, editor. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge (MA): The MIT Press; 2008. p. 1–24. doi: 10.1162/dmal.9780262524834.001

[12] Ananthakrishna R, Chaudhuri S, Ganti V. Eliminating fuzzy duplicates in data warehouses. In: Proceedings of 28th International Conference on Very Large Data Bases; Hong Kong, China; 2002. p. 586–597.

[13] Kalashnikov DV, Mehrotra S, Chen Z. Exploiting relationships for domain-independent data cleaning. In: Proceeding of 2005 SIAM International Conference on Data Mining; Newport Beach, CA; 2005. p. 262–273.

[14] Köpcke H, Rahm E. Frameworks for entity matching: a comparison. Data Knowledge Eng. 2010;69(2):197–210. doi: 10.1016/j.datak.2009.10.003

[15] Marshall B, Kaza S, Xu J, et al. Cross-jurisdictional criminal activity networks to support border and transportation security. In: Proceedings 7th Int IEEE Conference Intelligent Transportation Systems; Washington, D.C.; 2004. p. 100–105.

[16] Brown DE, Hagen SC. Data association methods with applications to law enforcement. Decis Support Syst. 2003;34(4):369–378. doi: 10.1016/S0167-9236(02)00064-7

[17] Wang GA, Chen H, Atabakhsh H. Automatically detecting deceptive criminal identities. Commun ACM. 2004;47(3):70–76. doi: 10.1145/971617.971618

[18] Wang GA, Chen HC, Xu JJ, et al. Automatically detecting criminal identity deception: an adaptive detection algorithm. IEEE Trans Syst Man Cybern - Part A: Systems Humans. 2006;36(5):988–999. doi: 10.1109/TSMCA.2006.871799

[19] Bhattacharya I, Getoor L. Entity resolution in graphs, in min graph data. Hoboken: Wiley-Blackwell; 2006.

[20] Bartunov S, Korshunov A, Park S, et al. Joint link-attribute user identity resolution in online social networks. In: Proceeding of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis; Beijing, China; 2012.

[21] Shu K, Wang S, Tang J, et al. User identity linkage across online social networks: a review. SIGKDD Explorations Newsl. 2017;18(2):5–17. doi: 10.1145/3068777.3068781

[22] Zhou F, Liu L, Zhang K, et al. DeepLink: a deep learning approach for user identity linkage. In: IEEE Conference on Computer Communications; Honolulu, HI; 2018. p. 1313–1321.

[23] Wang R, Zhu H, Wang L, et al. User identity linkage across social networks by heterogeneous graph attention network modeling. Appl Sci. 2020;10(16):16. doi: 10.3390/app10165478

[24] Radanliev P. Cyber diplomacy: defining the opportunities for cybersecurity and risks from artificial intelligence, IoT, blockchains, and quantum computing. J Cyber Secur. 2024;1–51. doi: 10.1080/23742917.2024.2312671

[25] O'Neil DF. Survival of the fit: how physical education ensures academic achievement and a healthy life. New York: Teachers College Press; 2020.

[26] White MG. Examples of physical characteristics in humans. In: Your Dictionary; 2019. Available from: https://examples.yourdictionary.com/examples-of-physical-characteristics.html

[27] Jamali R. Online Arab spring, social media and fundamental change. Chandos Publishing; 2014. Available from: https://www.sciencedirect.com/book/9781843347576/online-arab-spring

[28] Allert V, Reese G. Social identity-based motivation to engage in collective action supporting the redistribution of street space, transportation research part F. Traffic Psychol Behav. 2023;94:9–24. doi: 10.1016/j.trf.2023.01.009

[29] Tajfel H, Turner JC, Austin WG, et al. An integrative theory of intergroup conflict. Organizational Identity: A Reader. 1979:56–65. https://books.google.co.uk/books?hl=en&lr=&id=BgBREAAAQBAJ&oi=fnd&pg=PA56&dq=Tajfel,+H.,+Turner,+J.C.,+Austin,+W.G.+and+Worchel,+S.,+1979.+An+integrative+theory+of+intergroup+conflict.+Organizational+identity:+A+reader,+56(65),+pp.9780203505984-16.&ots=5sSeFdln5j&sig=EEc9VMFIrD2pNJtJCGTkBmDKU3I#v=onepage&q&f=false

[30] Eddin ME, Massaoudi M, Abu-Rub H, et al. Novel functional community detection in networked smart grid systems-based improved Louvain algorithm. In: 2023 IEEE Texas Power and Energy Conference (TPEC); College Station, TX, USA; 2023. p. 1–6. doi: 10.1109/TPEC56611.2023.10078573

[31] Blondel VD, Guillaume JL, Lambiotte R, et al. Fast unfolding of communities in large net- works. J Stat Mech. 2008;2008(10):P10008. doi: 10.1088/1742-5468/2008/10/P10008

[32] Hua F, Fang Z, Qiu T. Modeling ethylene cracking process by learning convolutional neural net- works. Comput Aided Chem Eng. 2018;44:841–846.

[33] Pujol JM, Erramilli V, Rodriguez P. Divide and conquer: partitioning online social networks. 2010.

[34] Greene D, Doyle D, Cunningham P. Tracking the evolution of communities in dynamic social net- works. In: International Conference on Advances in Social Networks Analysis and Mining; Odense, Denmark. 2010.

[35] Hui P, Sastry NR. Real world routing using virtual world information. In: International Conference on Computational Science and Engineering; Vancouver, British Columbia, Canada. 2009.

[36] Zhang L, Liu X, Janssens L, et al. Subject clustering analysis based on ISI category classification. J Inform. 2010;4(2):185–193. doi: 10.1016/j.joi.2009.11.005

[37] Haynes J, Perisic I. Mapping search relevance to social networks. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis; Paris, France. 2010.

[38] Newman M. Networks: an Introduction, chapter 7: measures and metrics. Oxford, England: Oxford University Press; 2010. p. 168–234.

[39] Raykar N, Kumbharkar P, Jayatilal DH. De-duplication avoidance in regional names using an approach based on pronunciation. Int J Adv Electr Eng. 2023;4(1):10–17. doi: 10.22271/27084574.2023.v4.i1a.32

[40] Knuth DE. The art of computer programming, sorting and searching. Redwood City, CA, USA: Addison Wesley Longman Publishing;1998. Vol. 3 p. 315–332. https://dl.acm.org/doi/book/10.5555/280635

[41] National Archives) "The Soundex indexing system. 2007. Available from: *https://www.archives.gov/research/census/soundex*

[42] Pinto D, Vilarino D, Aleman Y, et al. The Soundex phonetic algorithm revisited for SMS-based information retrieval. In: Spanish Conference on Information Retrieval (CERI 2012); Valencia, Spain; 2012.

[43] Anonthanasap O, Leelanupab T. iMnem: interactive mnemonic word suggestion using phonetic algorithms. In: The 20th International Society on Artificial Life and Robotics (AROB 2015); Beppu, Japan; 2015.

[44] Philips L. Hanging on the metaphone. Computer Language. 1990;7(12):39–43.

[45] Smiley C, Kubler S. Native language identification using phonetic algorithms. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications; Copenhagen, Denmark; 2017. p. 405–412. https://aclanthology.org/W17-5046.pdf

[46] Jaro MA. Probabilistic linkage of large public health data files. Stat Med. 1995;14(5–7):491–498. doi: 10.1002/sim.4780140510

[47] Winkler WE, Thibaudeau Y. An application of the fellegi-sunter model of record linkage. Technical report. U.S. Decennial Census, Bureau of the Census; 1990.

[48] Yancey WE. Evaluating string comparator performance for record linkage. Technical report. Statistical Research Division, U.S. Census Bureau; 2005.

[49] Yancey WE. An adaptive string comparator for record linkage. Technical report. Statistical Research Division, U.S. Bureau of the Census; 2004.

[50] Phillips M, Amirhosseini MH, Kazemian H. A Rule and Graph-Based Approach for Targeted Identity Resolution on Policing Data. 2020 IEEE Symposium Series on Computational Intelligence (SSCI); Canberra, ACT, Australia. 2020. p. 2077–2083. doi: 10.1109/SSCI47803.2020.9308182