An Interactive Human-Centered Data Science Approach towards Crime Pattern Analysis

Nadeem Qazi[1], B.L.William Wong[2]

[1] Brunel University, London UK
[2] Middlesex University, London UK

Abstract

The traditional machine learning systems lack a pathway for a human to integrate their domain knowledge into the underlying machine learning algorithms. The utilization of such systems, for domains where decisions can have serious consequences (e.g. medical decision-making and crime analysis), requires the incorporation of human experts' domain knowledge. The challenge, however, is how to effectively incorporate domain expert knowledge with machine learning algorithms to develop effective models for better decision making.

In crime analysis, the key challenge is to identify plausible linkages in unstructured crime reports for the hypothesis formulation. Crime analysts painstakingly perform time-consuming searches of many different structured and unstructured databases to collate these associations without any proper visualization. To tackle these challenges and aiming towards facilitating the crime analysis, in this paper, we examine unstructured crime reports through text mining to extract plausible associations. Specifically, we present associative questioning based searching model to elicit multi-level associations among crime entities. We coupled this model with partition clustering to develop an interactive, human-assisted knowledge discovery and data mining scheme.

The proposed human-centered knowledge discovery and data mining scheme for crime text mining is able to extract plausible associations between crimes, identifying crime pattern, grouping similar crimes, eliciting co-offender network and suspect list based on spatial-temporal and behavioral similarity. These similarities are quantified through calculating Cosine, Jacquard, and Euclidean distances. Additionally, each suspect is also ranked by a similarity score in the plausible suspect list. These associations are then visualized through creating a two-dimensional re-configurable crime cluster space along with a bipartite knowledge graph.

This proposed scheme also inspects the grand challenge of integrating effective human interaction with the machine learning algorithms through a visualization feedback loop. It allows the analyst to feed his/her domain knowledge including choosing of similarity functions for identifying associations, dynamic feature selection for interactive clustering of crimes and assigning weights to each component of the crime pattern to rank suspects for an unsolved crime.

We demonstrate the proposed scheme through a case study using the Anonymized burglary dataset. The scheme is found to facilitate human reasoning and analytic discourse for intelligence analysis

Keywords: Interactive Clustering, Linkage Analysis, Crime matching, Text mining, Human-centred searching, Network Visualization, Knowledge Graph, Data Science

1. Introduction

The intuition of human experts plays a vital role in solving complicated problems. Researchers have emphasized the human role in analyzing and exploring the dataset for the extraction of relationships in structured and unstructured content (Cao *et al.,* 2014). This integration of computer and human, where a computer assists the human design process and a human being is the in charge of an algorithmic process is termed as interactive data mining in information retrieval literature. It offers three benefits over black box algorithms of machine learning i.e. understanding (why one machine learning (ML) technique/algorithm is different than others), diagnosis (reasons for the failure of an ML technique/algorithm) and refinement (factors affecting the performance of an ML technique such as changing feature vector) (Shixia Liu *et al.,* 2017).

The data mining/machine learning applications, however, are mostly automatic with limited or without any human intervention and hence increase the risk of modeling artifacts. The grand challenge is to integrate effective human interaction with powerful machine intelligence through visual analytics to support both human insight and decision making (Holzinger, 2013). In another similar research, (Holzinger and Jurisica,2014) have emphasized that data mining techniques should keep the domain expert‘s intelligence into the loop.

This grand challenge is handled in many data science applications including health information (Holzinger, 2016), social media (Amershi, Fogarty and Weld, May 5, 2012), (Arın, Erpam and Saygın, 2018), image processing (Gonçalves, Guilherme and Pedronette, 2018), and web page analysis (Kulesza *et al.,* 2014), etc.

Answering this challenge of human-machine integration, we in this paper, present our approach of integrating human interaction with the machine learning intelligence for the crime analysis particularly focusing on crime matching. (Keyvanpour, Javideh and Ebrahimi, 2011) have defined crime matching as the process of “assigning crimes or criminals to [previously] solved or unsolved crime incidents”, while (Oatley, Zeleznikow and Ewart, 2004) have described crime matching as the ability to link or connect crimes in ways that enable the identification of potential suspects. Despite subtle differences, both refer to the use of machine learning based algorithms to (i) find similarities among crimes to discover potential suspects; and (ii) to develop offender profiles in a way that can be used to find matches with the profiles of offenders in unsolved crimes. The information-intensive querying process of crime matching requires establishing multi-level associations among crime entities to discover and reconstruct crimes through analysis of the evidence left at the crime scene.

1.1 Problem statement

A crime report text document is generally defined to be a logical unit of unstructured textual data holding details about the reported crime. It contains information about crime objects such as the name, type, spatial, temporal and modus operandi information of the committed crime along with any details of the suspect or associated offender. The modus operandi description in a crime report describes the method of the operation adopted to commit a crime. Early researchers (Cohen and Felson, 1979) had given a routine activity theory that defines the necessary conditions for a crime to happen. It includes a likely offender, a suitable target, and the absence of a capable guardian, coming together in time and space to form a crime triangle. Crime analysts seek to understand the motivations of the crimes looking into the available data from the perspective of the crime triangle and routine activity theory (Cohen and Felson, 1979).

More often during crime matching process, analysts spend a large amount of time reading crime reports to find the associations among the entities such as criminals, vehicles, weapons, bank accounts, and organizations. They ask a variety of questions based on associative questioning (Wong and Kodagoda, 2016) to learn more about the diverse nature of the context in which the crimes were committed to make the sense of the situation that would help to solve the crime.

Some analysts apply the 5WH (Who, What, When, Where, Why and How) structured analytic model to discover who else might have been involved in the crime, what other factors or events could be relevant, and how the crime and other similar crimes were committed? They seek information that could lead them to make associations with other concepts to create plausible hypotheses that can lead to solving a criminal case. Currently, the investigator has to painstakingly perform time-consuming searches of many different databases to collate such a comprehensive picture of the crime. The keyword and semantic base searches, however, do not leverage the power of associations of concepts in the search domain, as the former does not consider the meaning of the given query and later though looks for the meaning of the question, however, lacks the ability to elicit any association in the data. Additionally, due to lack of proper visualization analysts face a number of significant difficulties including making sense of collated data, distinguishing the relevance or similarities among the cases, identifying and understanding associations between criminal entities.

Crime analysts in addition to seeking extraction of the crime association from unstructured crime text, also follow fundamental task of analyzing the similarity of the criminal cases to identify common crime patterns and to reason about unsolved crimes. They perform spatial-temporal and behavioral grouping of the crimes to examine solved crimes that have similar characteristics as an unsolved crime to generate a new hypothesis. It also helps them in understanding trends of crime identifying criminal spatial and temporal hot spots.

While in recent years several data clustering techniques have been demonstrated to group similar items, however, they all depend the upon selection of appropriate similarity functions and feature vectors to produce good quality clusters. Most of these clustering algorithms are either automatic or provide limited human intervention. In addition to this crime clustering due to the lack of ground truth cannot be easily validated without human interaction. Thus clustering process in general and crimes clustering, in particular, requires a human role to achieve good quality clusters through the best possible combination of feature vectors and similarity functions. It, however, requires to handle the challenge of effectively integrating human interaction. Therefore it is valuable important and at the same time challenging to work towards a human-centered knowledge discovery pipeline that should provide efficient interactive visualization for clustering the similar cases, revealing associations between them to facilitate hypothesis formulation.

1.2 Solution

Therefore identifying this need for linkage based search mechanism and interactive visualization; we proposed a human-centered knowledge discovery and data mining (KDD) scheme for elicitation of temporal, spatial and behavioral crime associations from the unstructured text of crime reports. Our proposed scheme shown in Figure 1 is inspired by (Fayyad, Piatetsky-Shapiro and Smyth, 1996), however, it integrates human role with the data mining algorithms for the cognitive analysis of the crime reports. It offers extraction of possible associations between crimes and offenders, offender network and a plausible suspect list based on spatial-temporal

and behavioral similarities observed in crime reports. Hierarchical panorama visualization is also utilized to show links among the crime objects.

The knowledge discovery process in our proposed scheme is based on the associative search (Qazi *et al.,* 2016). We defined associative search as the cognitive thinking process consisting of associative questioning based on the crime triangle and the routine activity theory (Cohen and Felson, 1979). It, unlike semantic search, extends the scope of the search to the networks of objects including people, places, organizations, products, events, services, and so forth to extract the associations among the connected entities. We employed text mining implemented through a vector space model to form a connected search space of the associations extracted from the crime reports. We used partition clustering to group the crime reports through dynamic i.e. user-defined features selection and validated the quality of the clusters through silhouette analysis. The proposed framework also enables the analysts to integrate domain knowledge with machine learning algorithms through a visualization feedback loop. It includes the setting of the parameters in the tokenization phase of the text mining process, user-defined features i.e. attributes selection for clustering crimes, choice of similarity functions for extraction of the criminal network and setting the weights to each of the crime components for ranking the extracted suspect list.

The contributions of this paper include 1) a detailed literature review showing data science contributions towards crime analysis, 2) an association discovery scheme incorporating a proposed multi-level associations model for identifying criminal linkages, 3) interactive clustering distinguishing the relevance or similarities among the criminal cases, 4) our approach to handling categorical data in the clustering of multidimensional associations of the crime entities, and 5) identifying and visualization of criminal groups through a hierarchical knowledge graph.

The rest of the paper is organized as follows. Section 2 presents related research. We unfold the proposed knowledge discovery scheme for crime matching in multiple sections, describing association miner, interactive clustering and visualization in sections 3, 4 and 5 respectively. A case study using the Anonymized data is also presented in section 6 and the conclusion is drawn in the last section of the paper.

2. Related Work

Knowledge discovery is an interactive and iterative process, that starts from acquiring domain knowledge, followed by selecting, preprocessing and cleaning the target dataset. The other stages of a knowledge discovery process, as described by (Fayyad, Piatetsky-Shapiro and Smyth, 1996), include dimensional reduction, data projection, and implementation of appropriate data mining algorithm for the required task. The knowledge discovery finally ends up with the interpretation of the mined pattern extracting knowledge from it. In recent years several researchers have proposed different theoretical variations of KDD models such as interaction model (Brehmer and Munzner, 2013), (Sacha *et al.,* 2017), and sense-making models (Brehmer and Munzner, 2013) in order to recognize and integrate human role with the analytic process. In crime analysis domain, (Jentner et al., 2016) demonstrated “analyst is in the loop” approach in an interactive visualization prototype to extract crime behavior of the offender from the modus operandi description.

Our proposed knowledge discovery scheme for criminal analysis, shown in Figure 1, incorporates human role with the machine learning algorithms for the elicitation of the associations between solved and unsolved crimes, offender network, and a list of potential suspects based on spatial, temporal and behavioral associations

from unstructured texts of crime reports. The spatial, temporal and behavioral features are extracted from the crime reports during the text tokenization step of the text mining. Text tokenization is an important step in text mining and is used to extract co-occurring "n" number of consecutive words called as N-grams from the sentences. Unigram (1-gram i.e. one word), bigram (2-gram i.e. two consecutive words) and 3-grams (i.e. three consecutive words) words models are commonly used for text tokenization. (Alruily, Ayesh and Zedan, 2014) in their proposed system for criminal profiling from Arabic text, employed N-gram model to extract crime-related information such as crime type, location, and nationality of a person from the crime reports (written in Arabic) and utilized these features in Self Organizing Map (SOM) to cluster the similar crime reports. However, their proposed system does not provide a human interaction for dynamic text tokenization. In another similar research (Jayaweera *et al.,* 2015) have used SVM to classify news articles of Sri Lankan English newspaper as crimes or no crimes articles.

Our proposed knowledge discovery scheme follows human in the loop approach and offers dynamic tokenization of the unstructured text, allowing choosing any of 1-gram, 2-gram or 3-gram model for extraction of temporal, spatial and behavioral features from the crime reports. These features are then used to extract associations among the crime entities to perform association analysis.

Association analysis utilizes data mining methods to extract the relationships, patterns, rules, criminal network from a large dataset of a specified domain. Researchers have employed association analysis in multiple domains using various kind of databases such as transactional, relational or unstructured databases. For example (Saeed Piri et al., 2018) performed an association analysis on an electronic medical database of diabetes patients and proposed a new assessment metric to identify rare items/patterns without over-generating association rules. (Wei Chen *et al.,* 2017) employed Apriori algorithm to extract association rules from categorical datasets. Other examples of association analysis include identifying topics in tweets (Zarrinkalam, Kahani and Bagheri, 2018), market analysis for extracting consumer purchase pattern (Valle, Ruz and Morrás, 2018) and recommendation systems(Liao and Chang, 2016),(Viktoratos, Tsadiras and Bassiliades, 2018).

In crime analysis domain, association analysis has been utilized for identifying criminal activity, linking burglaries with serial offenders, associating modus operandi with serial crimes, linking evidences with crimes, association extraction between criminals, predicting potential offenders for unsolved crimes and criminal networks/community detection etc. (Thongsatapornwatana *et al.,* 2017) used color, brand, and type of vehicles, employing journey path analysis techniques with the association rule mining, to detect potentially involved suspect/s in a criminal activity. (Borg, Boldt and Eliasson, 2017) developed an algorithm based on the modus operandi similarity of crime pattern using Jacquard coefficient for linking burglaries to a serial offender. Their research showed that crime series with the same offender on average had higher behavioral similarity than a random crime series. (Hong Chi *et al.,* 2017) developed a decision support system consisting of similarity algorithms, a classification model, a feature selection and parameter learning algorithm to link serial crimes through behavioral information. Other examples of association analysis in crimes include linking evidences with crimes through Naïve Bayes algorithm (de Zoete *et al.,* 2015), associating multiple offenders with separate offences (de Zoete, Sjerps and Meester, 2017) and predicting potential suspect for unsolved crime (Vural and Gök, 2017) etc.

In addition to the above, researchers also have utilized behavioral features of the crime pattern with machine learning algorithms including logistic regression (C. Bennell and D.V. Canter, 2002), (Tonkin *et al.,* 2012), probability inference (Wang and Lin, 2011), etc. for eliciting associations between crime and criminals.

(Bache *et al.,* 2010) applied unigram language model i.e. multinomial and multiple Bernoulli models over solved crimes dataset to link behavioral features with characteristics of offenders and found that Bernoulli models outperformed multinomial models.

(Al-Zaidy *et al.,* 2012) employed name entity recognition along with a modified Apriori algorithm to extract prominent criminal community from unstructured textual data of a chat log. Their method uses interaction frequency between two people to measure the strength of linkages. (Didimo, Liotta and Montecchiani, 2014) developed a visual analytics framework (VISFAN), to visualize financial activity networks. Their proposed framework extracts entities such as bank accounts, addresses, amount and types of the transactions, etc. from financial reports and visualizes it in the form of a network through graph drawing techniques and hierarchical clustering. (Isah, Neagu and Trundle, 2015) have demonstrated the use of the bipartite model over pharmaceutical dataset for extracting the hidden relationship between criminals. Some other earlier examples related to our work are commercial tools like COPLINK Explorer (Schroeder *et al.,* 2007), Dynalink (Park, Tsang and Brantingham, 2012), JIGSAW (Stasko, Görg and Liu, 2008). However, either most of these tools lack proper visualization or do not have the ability to extract criminal relationship from textual data.

Our work is different from the above mentioned related research, as we have incorporated associative questioning through a 5-WH associative search based model (Qazi *et al.,* 2016) for the elicitation of spatial, temporal, and behavioral associations of criminals from crime reports. In addition to this, we also have integrated interactive clustering to distinguish solved crimes with unsolved crimes on the basis of temporal-spatial and behavioral associations.

Clustering is a very commonly used unsupervised data mining algorithm that allows similar objects to be organized into groups. It has applied in a wide range of applications including sentiment classification (Onan, Korukoğlu and Bulut, 2017), electricity load management (Biscarri *et al.,* 2017), active learning (Min Wang *et al.,* 2017), tourism industry (Hu, Chen and Chou, 2017) etc. The examples from crime analysis literature include (Bsoul, Salim and Zakaria, 2013), who detected crime patterns in news articles through K-mean clustering over multiple crime types. They employed affinity propagation algorithm for determination of the number of clusters. In another research, (Thota *et al.,* 2017) constructed crime cluster zones of Indian crime dataset using the K-means method, however, they used numerical data.

The interactive clustering task performed in our work, however, deals with the high dimensional features of textual data, which due to the curse of dimensionality affects the performance of clustering. The reported solutions to improve cluster quality are automatic feature selection methods including filter (Alelyani, Tang and Liu, 2013), wrapper (Lin *et al.,* 2016), and hybrid (Bharti and Singh, 2014). Some researchers, however, have demonstrated the use of fixed features selection employing a vector space model for grouping similar items. For example (Dagher and Fung, 2013) have introduced subject-based semantic document clustering algorithm employing vector space model to groups documents into a set of overlapping clusters, each corresponding to one unique subject. (Reich and Porter, 2015) proposed a Bayesian model, utilizing crime locations and offender's modus operandi as fixed feature vector for burglary crime series identifications. (Borg *et al.,* 2014), demonstrated minimum cut based graph clustering to detect residential burglaries series. They used a fixed feature vector consisting of modus operandi, residential characteristics, stolen goods, spatial similarity, to group similar crimes.

In recent years interactive clustering emerges as a potential solution for the fixed feature vector problem. I-TWEC (Arın, Erpam and Saygın, 2018) is an interactive web-based clustering tool for twitter data that utilized

the suffix tree based algorithm to cluster user uploaded tweets using their semantic. Some other examples include iVisClustering tool (Lee *et al.,* 2012), Cluster Sculptor (Bruneau *et al.,* 2015), (Krause, Perer and Bertini, 2014) radial axes method for visual backward feature selection (Sanchez *et al.,* 2018), etc. These tools facilitate the analyst to steer the feature selection process according to their domain knowledge and specification.

The above literature review reveals that there is a need for a unified framework to integrate association extraction and interactive visualization under a single umbrella. Following this need, we in this work presented a unified framework that offers extraction of criminal associations, offender network and plausible suspect list from unstructured text and groups them through interactive 2D clustering with proper visualization under a single envelope. In our proposed framework the analyst is able to create a dynamic feature vector consisting of spatial, temporal and behavioral crime pattern attributes to group the crimes and associated offenders in the user-defined number of clusters. We also have used multidimensional scaling technique to visualize the hidden relationship between crime KPIs.

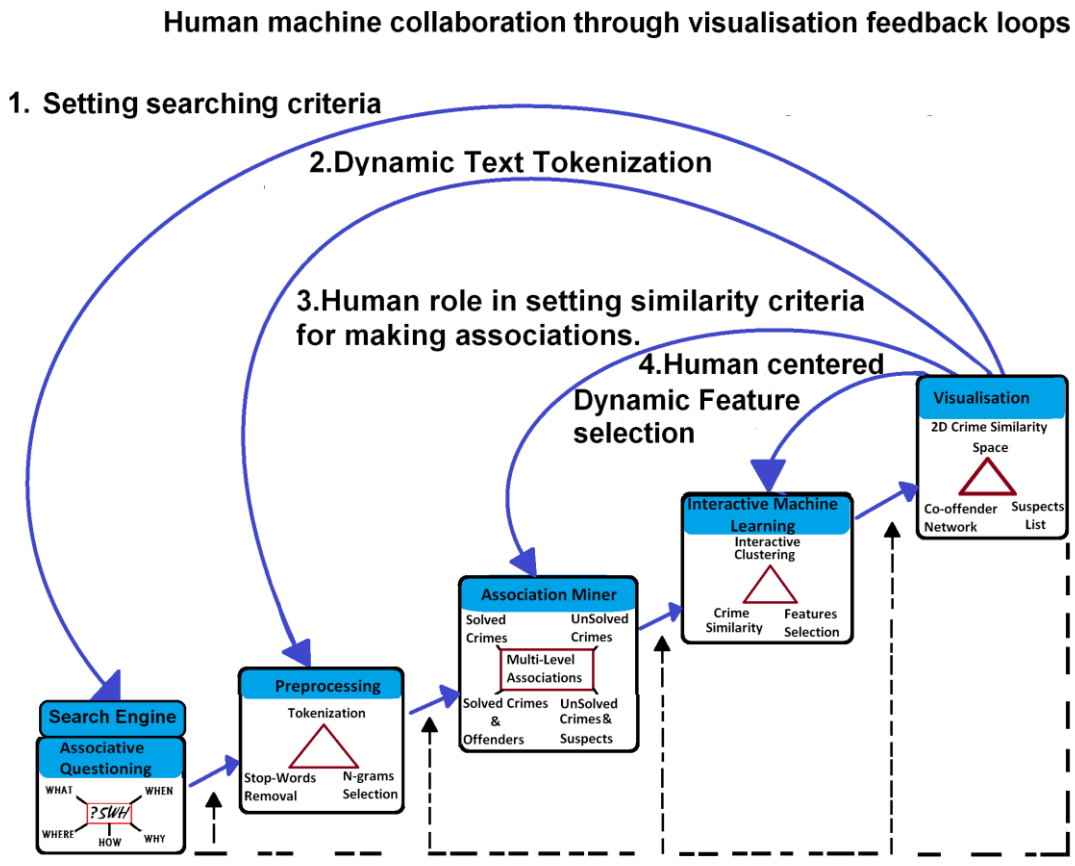

Figure 1: Knowledge Discovery Pipeline for Crime Analysis

3. Data Mining Framework for Crime matching

We now describe in detail, the pipeline of our proposed human-centered knowledge discovery and data mining (KDD) scheme for crime analysis. The architecture of the pipeline, shown in Figure 1 is composed of an associative search engine, an associations mining unit, an interactive clustering unit, and visualization unit. The visualization unit is connected to the remaining units through a feedback loop for human-machine collaboration. This feedback loop avoids the full automation of the discovery task and permits the user to steer the algorithm for optimal settings to solve a given problem. The functionality of this feedback loop is described for each unit later in the paper.

The framework takes a crime pattern as input and resolves it into its three components i.e. modus operandi, spatial and temporal components. It then elicits the multi-level associations on the basis of a temporal, spatial and behavioral characteristic, and group crime similarities in a 2D interactive clustering space. The associations in each crime cluster are then hierarchically visualized through a bipartite tree-based structure called as knowledge graphs in our framework, to depict the co-offender network and plausible suspect list using graph theory. We now describe each component of the pipeline in the following sections.

3.1 Associative Search Engine

The associative search engine unit, for the given crime pattern, generates the spatial, temporal and modus operandi based associative queries presented in Table 1. The relevant data, on the basis of these queries, is extracted from the knowledge base, and after pre-processing is fed into association miner, which elicits the multi-dimensional associations and is described below. The feedback loop from the visualization unit to search engine shown in Figure 1, allows the analyst to define the searching criteria through setting the crime pattern characteristics.

Table 1 Associative Queries

1. Who are the known offenders operating in an area and what is their modus operandi to commit crimes?
2. What are the additional details of the associated offenders/victims such as their past history etc.?
3. What are the geospatial profiles of the offender/s, including temporal, spatial and other similar criminal activities resembling with the given crime pattern?
4. How many times the offender has committed similar crimes and what are the temporal and spatial details?
5. What is his/her pattern of modus operandi?
6. Where an offender mostly likes committing an offense and who else has committed the same crime at this location?
7. What are the other offenses that have occurred with a similar given crime pattern?
8. How often have offenses like the given crime pattern occurred?

3.2 Association Miner

Associations miner unit of the KDD scheme, shown in Figure 1, elicits multi-level associations to unfold similar crimes, criminal network and plausible suspect list from a given crime dataset. It is accomplished through a multi-level association model (shown in Figure 2). This model is based on spatial-temporal characteristics (Ozgul *et al.,* 2012) and modus operandi behavior (Wang and Lin, 2011) of the given crime pattern. We compared the similarity of the given crime pattern with the other crime entities to establish these multi-level associations.

The proposed model through the rule-based heuristic and similarity matching extracts the associations in two levels as shown in Figure 2. Level 1 of the model elicits the relationships between solved and unsolved crimes. The heuristic rule employed to distinguish the crimes is based on the fact, that a solved crime is the one which has been solved and a perpetrator/offender has been identified or sentenced for this crime, and unsolved crime is the one, for which the goal is to identify potential/probable offender/s responsible for committing this crime. The level 2 of the association model calculates the temporal, spatial and modus operandi similarities between solved crimes, unsolved crimes, and offenders to associate solved crime with the offenders, offenders with each other, and unsolved crime with the suspects.

The feedback loop from visualization unit as shown in Figure 1, allows analyst through a user interface, to input queries to the model, such as give me a co-offender network of a given offender or what could be a possible suspect for a given unsolved crime. The feedback loop also leverages the analysts to choose the similarity functions to observe variation in the associations with the change in the similarity function.

The extracted associations are represented as an undirected heterogeneous graph (Sun, 2013). The root node of this graph is based on the user input question, which could be a criminal name or unsolved crime and based on this, the dynamically created children nodes may be the perpetrator/s, location/s, offense, time and modus operandi. The edges in the network are made of the associations connecting nodes on the basis of spatial-temporal and behavioral similarities. Thus for a given input node of a criminal name, it generates a graph for the co-offender network through comparing the crime pattern similarities of the root node with that of all the offenders/victims in other crime reports. On the other hand, if the given input node is an unsolved crime, then it compares similarities between offenders of similar solved crimes and the given unsolved crime, generating a list of possible suspects. We now describe these graphs separately in the following sections.

3.2.1 The Co-Offender Network

We modeled two types of co-offender network. The first model is based on crime associations i.e. when two or more offenders are reported together in a crime report for committing a crime. Our second network model is based on the spatial, temporal and modus operandi similarity and we have named it as (STM) model in our framework. These similarities are quantified through calculating three distance functions i.e. Cosine, Jacquard and Euclidean distances. However, following the co-reasoning approach between human and machine, the choice of calculating the similarity between the crimes is dynamically made by the user through selecting any of these distance functions. This thus implicitly steers the model each time an analyst chooses a different distance function to create similarity based associations.

In addition to this, the desired number of the retrieved offenders in the network graph is also set by the user and is implemented through following the K-Nearest Neighbor algorithm. We retrieved all the offenders that matched with the given similarity, ranked them in descending order according to the value of the chosen distance

function. However, instead of selecting all, we selected only the K number of offenders. This thus enables to weave the graph of crime objects having the largest similarity among them based on the chosen distance function.

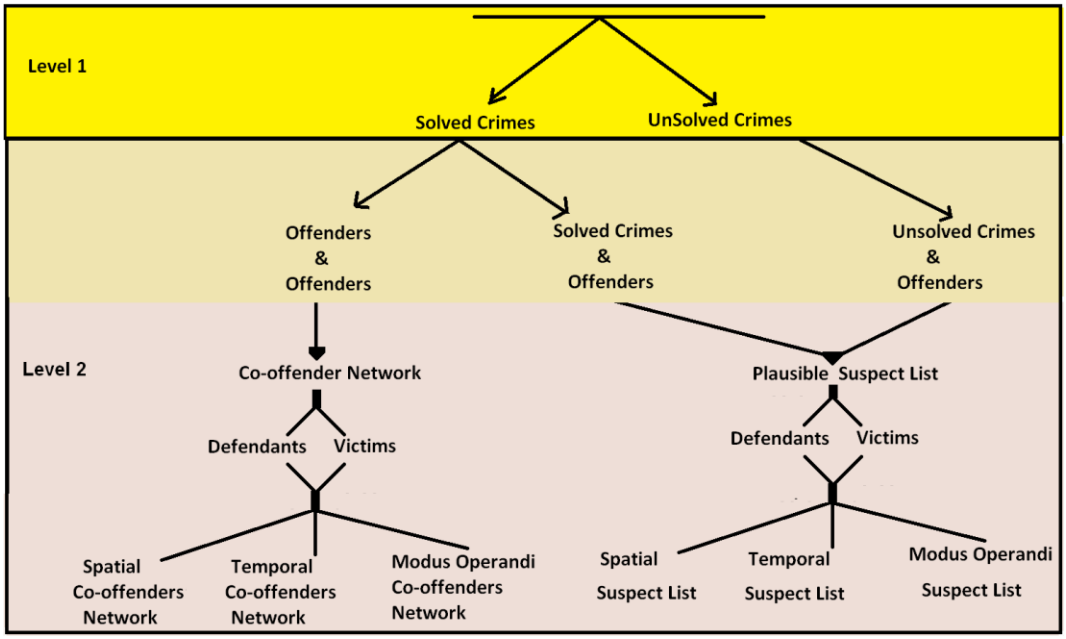

Figure 2 Spatial, Temporal, and Modus Operandi based Multi-level Associations model.

Mathematically Let A_t represents type of association where t ⊆ (C, L, T, MO) and P, C, L, T, MO represent the set of distinct values of the offenders, crimes, locations, time of the event, and modus operandi respectively. We define a criminal co-offender network model having nodes of offenders $P_s \subseteq P$ connected with the chosen association type $A_t \subseteq$ (C, L, T, MO). Any two offenders P_n and P_m in the network are said to be connected with each other if they share the same association type so that following is true.

$$A_k = \langle P_n, P_m \rangle$$

Where k ⊆(C, L, T, MO)

The offender may be associated with more than one offender based on the type of association. We represent the total number of common associations between the offenders through an adjacency matrix shown in Table 2. The first row and first column of this matrix contain the offender P_1 to P_n and the rest of the matrix elements bear a total number of common associations between the two offenders P_{ij}. For example Table 2 shows an adjacency matrix of associations for three offenders P_1, P_2, P_3. The similarity of an offender to him/herself is immaterial therefore it is shown with letter X in Table 2. It also shows that P_1 and P_3 have zero associations and hence is not connected with each other, whereas P_1 and P_2 are connected with each other with two associations and P_2 and P_3 are fully connected with each other sharing all four type of associations. Thus P_2 is connected with both P_1 and P_3. This adjacency matrix thus presents the similarity in the crime pattern of the multiple offenders, and thus facilitates an analyst to create a hypothesis to link two or more offenders based on the number of the associations between them.

Table 2: Adjacency matrix of associations between offenders

	P_1	P_2	P_3
P_1	X	2	0
P_2	2	X	4
P_3	0	4	X

3.2.2 Plausible Suspect List

In addition to the co-offender network, the proposed KDD also extracts a plausible suspect list for a given unsolved crime. For extracting plausible suspect list, we resolved the given pattern of unsolved crime into its modus operandi, temporal and spatial component, and compared each of this with that of the suspect's crime pattern. Each member of this list has at least one committed crime, exhibiting similarity with any or all the components of the given unsolved crime. The similarity of each component is measured through a cosine function. Each member is ranked based on the similarity score of his/her committed crimes with that of the given unsolved crime. However, the ranking is not fixed as the analyst defines the importance of the component of the crime pattern by setting its value between 0 and 1. The final similarity score S is thus the sum of the modus operandi, temporal and spatial component of the crime pattern. It is calculated as $S=W_T*S_T+W_L*S_L+W_M*S_M$; Where W_T, W_L, and W_M are the weights assigned to spatial, temporal, and modus operandi component of the crime respectively, by the analyst and S_T, S_L, S_M is the corresponding similarity value of each component with the given crime pattern.

Mathematically Let C_1, C_2, C_3,…., C_n be the solved crimes committed by perpetrators P_1, P_2, P_3,..., P_n having similarity with any or all the components of the given unsolved crime and hence may be considered as suspects for the given unsolved crime. Let S_1, S_2, and S_3… S_n are the corresponding similarity scores of unsolved crimes with these solved crimes. Let's also suppose P_1 has committed the crimes C_1, C2, while P_2 has committed the crime C_2, C_3 and P_3 have committed the crimes C_1, C_2, and C_3. The similarity ranking of each of these suspects then can be calculated using the following equations:

Rank of $P_1= S_1+S_2$ (As crime C_1 and C_2 has similarity score S_1 and S_2 respectively with given unsolved crime)

Rank of $P_2= S_1+S_3$ (As crime C_1 and C_3 has similarity score S_1 and S_3 respectively with given unsolved crime)

Rank of $P_3= S_1+S_2+S_3$ (crime C_1, C_2, and C_3 has similarity score S_1, S_2, and S_3 respectively with given unsolved crime)

Now suppose if $S_3 > S_2 > S_1$ which means that given unsolved crime is more similar to the C_3 and least similar to the C_1, then based on the similarity score, though P_1 has appeared in two crimes so is the P_2, but since solved crime C_3 is more similar to the given unsolved crime due to the high value of the S_3, therefore P_2 will rank high on plausible suspects list as compared to P_1.

Like the offender's network, the plausible suspect list also presents the similarity in the crime pattern of the suspect with the given crime pattern through an adjacency matrix of order nXm with n numbers of suspects and m numbers of the attributes of the crime pattern. The top row of this matrix contains all the attributes of a crime pattern i.e. spatial component including district and street, temporal component (time of the event) and lastly all the elements of modus operandi. The first column of this matrix contains the name of all the suspects in

the list. We tag each cell of the matrix S_{ij} corresponding to each suspect and the crime pattern component, either 0 or 1, to record the presence or absence of the similarity. This matrix thus represents a detailed picture of the offender similarity with the given crime pattern and hence facilitates the analyst in identifying a group of suspects for the unsolved crime. The analyst can then use this suspect list as the anchor to start the investigation process to solve an unsolved crime.

4. Interactive Clustering for 2-D Crime Space

The analysts, in addition to using the co-offender network and suspect list, also examine the spatial-temporal and behavioral similarities in crimes to make a hypothesis for solving an unsolved crime. Following this need, we integrated an interactive and unsupervised clustering algorithm to group the crimes into a 2D crime space, such that the similarity in a cluster is larger than among the clusters. We employed dynamic feature selection, using spatial, temporal and modus operandi attributes of the crime as an investigative lens, avoiding fully automated or manual system. The human-machine collaboration is integrated through a feedback loop from visualization module (as shown in Figure 1), that drives the analyst to redefine the feature vector selecting or de-selecting any or all of temporal, spatial or modus operandi variables to perform iterative clustering. In this way, the analyst is able to examine the spatial, temporal and behavioral similarities of the crimes and use it to reason unsolved crime for the crime matching.

4.1 Dynamic Feature Selection

For the spatial, temporal and behavioral attributes of the crime pattern, we represented spatial information through postcode, street, and town, temporal information through the month, day, and time of the offense occurred. For time of the offense, we, adopted the idea of conceptual scaling to transform 24 hours of the day into its symbolic value which resulted in four periods of the day: morning (from 6 am to 12 am), afternoon (from 12 pm to 6 pm), evening (from 6 am to 12 pm), and night (from 12 pm to 6 am). Finally, the modus operandi information of the committed crimes was represented through a set of twelve variables having a set of predefined values as shown in Table 3. Any or all of these spatial, temporal and modus operandi attributes may be selected or deselected by the analyst through the feature selection bar shown in Figure 4a, to form a dynamic feature vector for clustering the crimes.

4.2 Categorical Data Handling through VSM

The Feature vector described above involves categorical data. However most of the clustering algorithms work on numerical data, some researchers have demonstrated a solution to this critical challenge of clustering. (Keyvanpour, Javideh and Ebrahimi, 2011) have illustrated the use of categorical data into clustering algorithm by converting these variables into binary attributes and used 0 or 1 to indicate the categorical value either absent or present in a data record. This approach, however, is not suitable for high dimensional categorical data. Therefore in order to tackle this issue, we employed the Vector Space Model (VSM) and through the process of vectorization created a bag of words or (crime terms in this case) from the crime dataset.

The process of vectorization was accomplished following sequences of simple tasks including removing delimiters, converting all words to lower case, removing stop words and stemming words to their base. We also

made this preprocessing step interactive as shown in Figure 1 through the use of n-gram models which allow the user not to just use the unigram model, but also bi-gram and trigram model. The basic idea is to extract unique content-bearing words from the set of crime documents, assign weights to every term, based on the product of Term Frequency and Inverse Document Frequency (TF-IDF), and then treat these words as a numerical representation of the features to the clustering algorithm.

Mathematically, Let C= {C_1, C2, C3… Cn} be the crime space consisting of N crimes. Each crime C_i ;i=1,2...n is consisted of n numbers of terms t_1,t_2,t_3,…i_n, representing the spatial-temporal and modus operandi information of a crime.

Table 3 Modus Operandi Variables Details

Variable Name	Meaning
1. EntryPosition	How the offender/s entered into the premises such as rear or front.
2. EntryType 3. MO_Exit	It defines the method that an offender/s used to actually gain entry into the premise. It contains values such as Climbed, Cut, Forced, Removed Glass etc.
4. MOFixture: 5. MO_Exit_Fixture	It holds information regarding the feature at the point of the entry or exit, which may contain text like Door, Windows etc.
6. MOFixtureMaterial: 7. MO_Exit_FixtureMaterial	It contains information about the material of the point of the entry or exit such as wood, plastic, unknown etc.
8. MOFixtureType: 9. MO_Exit_Fixture_Type	It is the type of feature at the point of entry or exit which contains the values Casement, Fixed, Louvre, Patio, Sash, and Transom.
10. MOSearchLocation:	It identifies the rooms entered by the offender/s and contains the values All, Down (downstairs), Many, One and Up Stairs
11. MOSearch type: 12. MOSearchOther	This field determines the type of searches that the offender conducted. It contains the terms like Tidy and Un-Tidy or their synonyms.

We represented a crime C_i through the n-dimensional feature vector in the term space as C_i=W_1t_1, W_2t_2, W_3t_3,….. W_nt_n ; where W_n is the weight assigned to each term t_j in the crime document C_i through the following relationship.

$$W_n=Frq(t_j,C_i)*IDF$$

Where;

Frq (t_j,C_i) is the frequency of the term j in a crime document i and IDF is the inverse document term frequency calculated as:

$$\text{IDF}=1+\log\left(\frac{Total\ number\ of\ Crime\ Documents}{Number\ of\ Crime\ documents\ containing\ the\ term\ t_j}\right)$$

The numerical representation of the crime space C was thus represented through this weighted crime terms matrix consisting of rows as crime documents and columns containing weighted crime terms and was fed into a K-mean clustering algorithm. However, we used cosine similarity as the distance function in the K-mean algorithm rather than using Euclidean distance. Silhouette analysis was employed to calculate the optimal number of the clusters as required by the K-mean algorithm.

Lack of the ground truth is one of the important issues in grouping crimes, the challenge is how to validate and evaluate the clustering process. We used silhouette analysis by calculating the value of silhouette coefficients for each of the generated clusters. The value of silhouette coefficient ranges between 0 and 1, with a value near to 1 means the high value of similarity inside the cluster. Thus clusters that showed the high value of the silhouette coefficient were graded as high-quality clusters, while clusters having a coefficient value near to zero were graded low in cluster quality due to less similarity inside the cluster.

4.3 Dynamic Configuration of 2D Crime space

A significant feature of our adopted interactive clustering is the creation of a dynamic 2D cluster space having reconfigurable X and Y axes for visualizing the implicit relationship of KPIs with each other. It enables the analyst to observe the relationship between two KPIs with respect to each other revealing more insight into data. Thus for example, if the analyst wishes to examine how the crimes are spatially distributed over the streets of a town, s/he may choose to set these two KPI i.e. crimes and streets on either X or Y axis, to see their hidden relationship on a 2-dimensional crime space. Likewise, the similarities between any of the crime clusters can also be visualized by setting the similarity distance between them on either of X and Y axis of the crimes space.

We employed multi-dimensional scaling to map the similarity distances of clusters on either of X or Y axis. We first calculated an nXn distance matrix of centroids of each cluster and map each element of this matrix to the configuration points $x_1,x_2,x_3,\ldots x_n$ in such a way that the distance D_{ij} between any two clusters is well approximated by the distances $|x_i-x_j|$. Following this, either X or Y axis of the configuration space, when set to the cluster distance, would arrange clusters in a fashion, such that similar clusters will be placed near to each other and dissimilar clusters would appear far away from each other on the chosen axis. This enables the user to easily tag a cluster based on its crime pattern across the generated global similarity map.

5. Associations Visualization

The visualization unit of our KDD is connected to all other units through a feedback loop and presents the generated 2D crime space in aggregated and detailed views as shown in Figure 4a and Figure 4b respectively.

5.1 Aggregated View

The aggregated view presents the summary of the crime objects illustrating the associations between unsolved crimes and solved crimes along with their associated offenders. It is shown in Figure 4a. The clusters

are represented through visual doughnuts of varying arc length. The arc lengths of these doughnuts are kept proportional to the unsolved and solved crimes, whereas their associated offenders are represented in the center of the doughnuts. This size of the doughnut thus represents the heavy populated crime clusters. Hovering on each of these doughnuts (Figure 4a.) shows the statistic of the crime terms inside the selected crime clusters.

5.2 Detailed View

The detailed view of the crime space Figure 4b depicts how crimes are related to each other on the basis of the similarity inside the cluster. Each cluster is represented as a big gray circle, showing three types of the associations including the type of the crimes i.e. either solved or unsolved crime, the associated offenders and victims of the solved crime and lastly similarity of the crimes with each other. Hovering on each of these circles shows the information of the crime such as crime reference numbers as tool-tip. The association of an offender with crimes is visualized through the focus and context technique. When an offender is focused or hovered through a mouse, its association with all of its associated solved crimes in any cluster is highlighted through increasing the size of the related solved crime circles in the clusters, which goes back to normal when hover is off as shown in Figure 4b.

5.3 Offender Space: Knowledge Graph

According to visualization literature (Sallaberry *et al.,* 2016) nodes and links in a tree can signify relations among objects, consequently, we have employed the notion of a dynamic hierarchal bipartite tree called as a knowledge graph, to visualize the crime linkages, extracted in level 2 of the association miner. Each node represented through iconic graphic is collapsible and expandable, which means a user can click a node of interest to view its underlying children while closing any other node, so that only relevant/desired information is placed on the screen. The next section demonstrates a case study to show the working of the proposed scheme.

6. Case Study

We tested our proposed scheme over Anonymized burglary dataset, consisting of over 1.6 million crime reports along with associated offenders and victim's information, collected from UK Law Enforcement Agency. The process of anonymization through encryption techniques removed the individual's identifying information from the dataset so that the remaining data cannot be linked to that individual. This means the name and other related information such as crime reference number, location streets, town, and time, etc. are not real. A single crime report in the dataset contains details about the reported crime including crime reference number, offense category, spatial information such as street, district, town and postcode, temporal information such as time, date and day when the offense was first committed, and modus operandi description of the occurred crime. Twelve modus operandi variables with predefined values are presented in Table 3, which describe the modus operandi of the committed crime. The details of associated offender/s or victim/s include surname, forename, sex, date of birth, ethnicity, home address including postcode, street, town and district information, etc.

Several use cases were tested to demonstrate the performance of the scheme. However, here we present a use case where the objective was set to explore the clusters of similar crimes, hotspots, offender network, and plausible suspects list, for a given crime pattern. The chosen crime pattern for this use case was the modus operandi

used by the offender to enter the premise. The proposed framework was searched for the given modus operandi, where an offender entered the premises through “UPVC door or window”. This information, in our test dataset, is represented in the modus operandi field “MoFixtureMaterial” (Table 3) and contains the value “Plastic”. The search engine generated a list of solved and unsolved crimes having similarities with the given crime pattern. The next section shows the use of interactive clustering to group this result in the form of a 2D reconfigurable crime space and to observe any hidden crime pattern.

6.1 Effect of Crime Features on Clusters

For the resulted search data, to find an appropriate feature vector, we first examined the effect of feature vector on the number of clusters through silhouette analysis and measured average silhouette coefficients for each pair of the chosen number of clusters and feature vectors. Four sets of feature vectors were taken. The (Full FV) consisted of all features combining temporal, spatial and modus operandi attributes, while other three feature vectors were made of using spatial, temporal and modus operandi features separately. The result presented in Figure 3 shows that the silhouette coefficient value and hence the cluster quality decreases with the increasing number of clusters for all these three feature vectors used in making clusters. The highest value of the silhouette coefficient value also suggests that the feature vector consisting of only modus operandi information generated good quality clusters than the other two feature vectors.

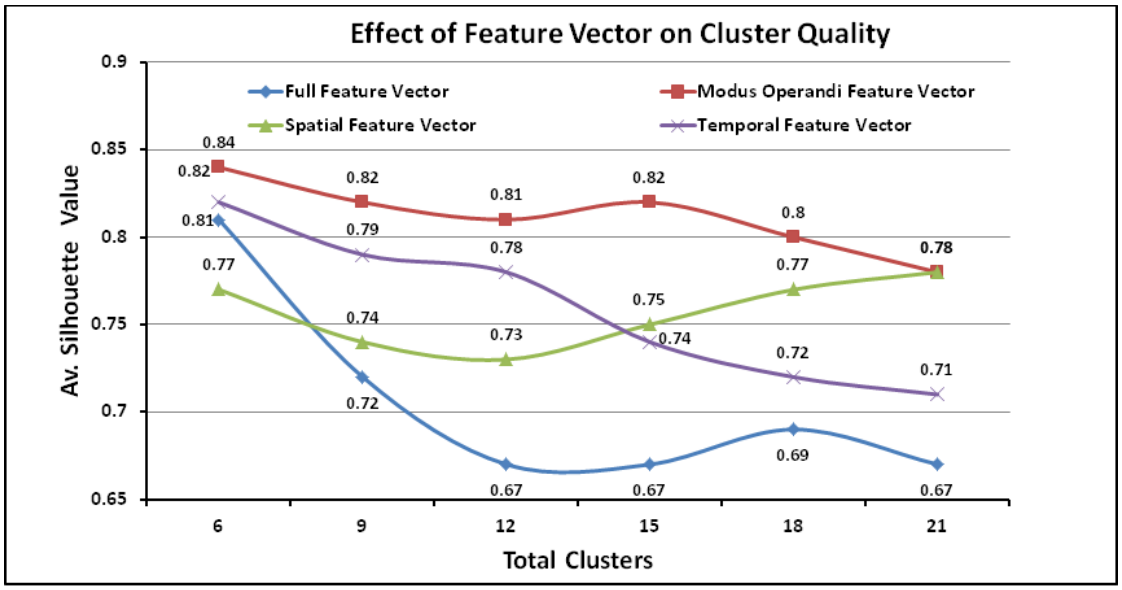

Figure 3 Effect of Feature Vector on Cluster Quality.

6.2 Crime Pattern

The 2D reconfigurable crime space is presented in aggregated and detailed views shown in Figure 4a and Figure 4b respectively. The reconfigurable X and Y axes of this 2D crime space, acting as analysts lens, allows him to steer the visualization process. The analyst can see several hidden relationships without actually performing clustering again, starting from a similarity-based grouping of the crimes to finding out temporal-spatial and behavioral crime hot spots in a particular area. For example, setting X-axis of the cluster space to the similarity

distance and Y axis to the proximity (Town), the 2D crime space projects the similarity of crime clusters with each other across the towns. It helps the analyst to examine where else similar crimes have occurred.

The similar clusters in Figure 4a and Figure 4b are more closed to each other indicating the higher similarities in the crimes in these clusters. In Figure 4a, it can be seen that in the town "DEWMAPLE" cluster 3 and cluster 4 are more similar (due to less distance between them) as compared to cluster 2. Hovering on cluster 4, the tooltip shows the centroid and statistic of the crimes in this cluster. It has six similar crimes, five of them are "BURGLARY DWELLING" committed in the town "DEWMAPLE" and "PLASTIC" was reported as fixture material to enter in the premise. These extracted patterns thus indicate that for most of the burglary dwelling crime in the town "DEWMAPLE" offenders have used "UPVC DOOR or WINDOWS" to enter in the premises. However one crime in this cluster is of "BURGLARY OTHER BUILDING", occurred in the town "YARNFORTH". The other groups of similar clusters (i.e. cluster 5 and cluster 1) are found in the town Yarnforth" and "Carsington". Other trends can also be seen in Figure 4a.

Steering again the axes lens of 2D crime space, and setting X and Y-axis to other crime variables shows another projection, answering the questions i.e. when, where and how these kinds of crimes occurred in these towns. These projections are shown in Figure 5a, Figure 5b, Figure 5c, Figure 5d, Figure 5e, and Figure 5f with Y-axis representing (Street, Day of period and modus operandi information including Exit from the premises i.e. MOExit, Fixture material used and search locations in the premises respectively), while for each of this Y-axis, the X-axis is set on proximity represented by Town. Figure 5a shows prominent streets, where similar crimes in the town "DEWMAPLE have occurred. The big orange arms of five unsolved crimes at "PAVEMENT ROW", "LINGSTON CLOSE" and "TEMPLEFIELD" streets indicate that given crime pattern is very common at these streets. Likewise "EASON CLOSE" street of the town "CARSINGTON and "OATLANDDRIVE" in town "YARNFORTH" are also the hot spots of similar crimes.

Another projection of the same clusters Figure 5b, rearranges the clusters to reveal temporal information of these crimes i.e. when did these events occur in these areas. It can be seen in the Figure 5b that all the three crimes clusters in town "DEWMAPLE" contain crimes that mostly occurred at midnight, while the two crime clusters in town "CARSINGTON" at "EASON CLOSE" contain crimes that were occurred in early mornings.

Additionally, The Figure 5c, Figure 5d, Figure 5e, and Figure 5f respectively, reveal days of the week and other modus operandi information (i.e. Exit from premises, fixture material, and search locations) respectively, of the committed crimes. The crime clusters in town "CARSINGTON" occurred on weekdays and the offenders in most of these cases used "PLASTIC" in committing the crimes and escaped from the premises through "REAR". Other patterns are also visible in Figure 5.

The important thing to note here is that analyst is steering the visualization wheel (in this case X-axis and Y-axis) without actually computing the algorithm again and more insight of the information is shown to the analyst. These crime patterns may be used as anchors in generating a hypothesis to facilitate reasoning process towards matching solved and unsolved crimes.

The detailed view Figure 4b reveals the similarity of the crimes with each other inside cluster along with associated offenders and victims. When an offender is focused or hovered its association with all of its associated solved crimes in any cluster is also highlighted by increasing the size of the related solved crime circles in the clusters, which goes back to normal when hover is moved off as shown in Figure 4b.

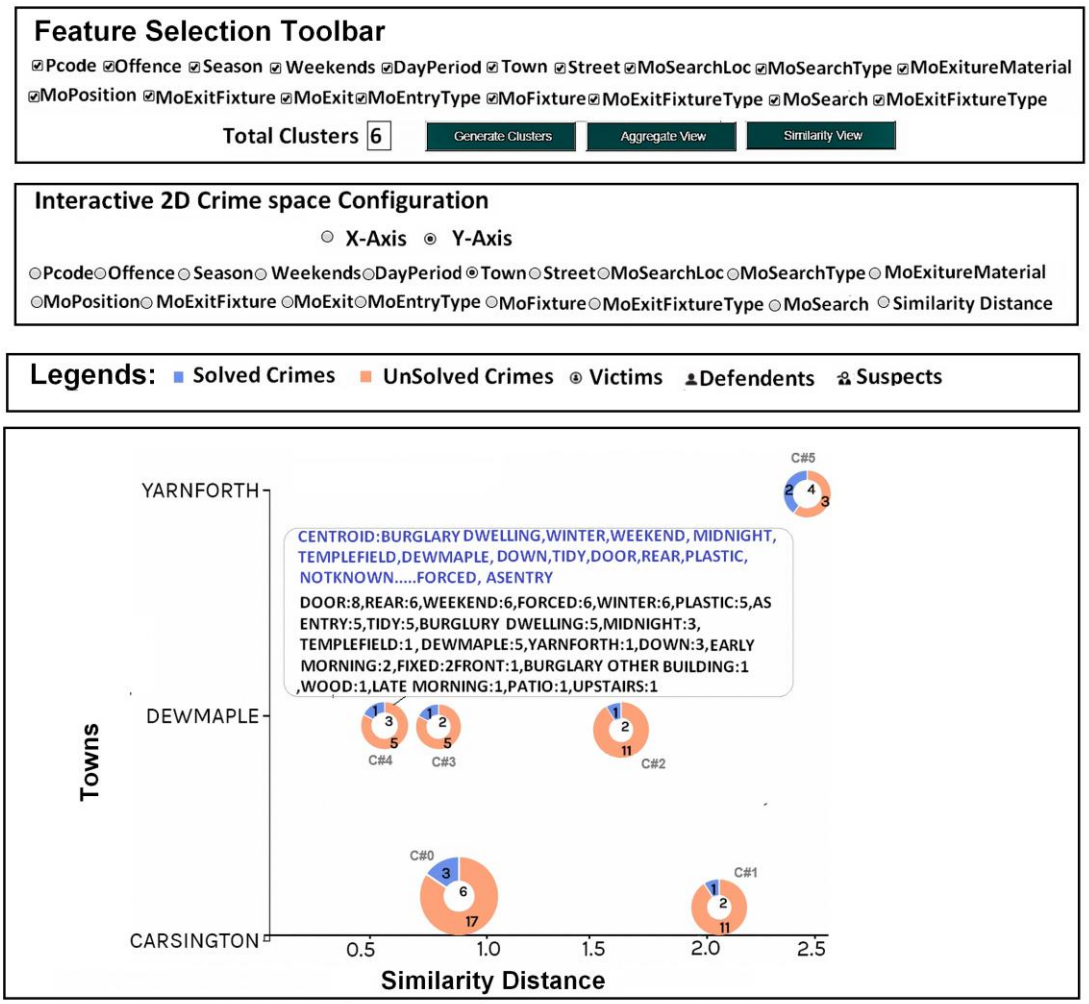

Figure 4 Crime cluster Space (Aggregated \& Detailed View)

Additionally, when any of this offender is clicked it generates a knowledge graph of all of his/her associated offender/s as shown in Figure 6. These kinds of similarities in the clusters may be useful to the analyst in hypothesis generation towards solving an unsolved crime or catching up a network.

6.3 Knowledge-Graph

After inspecting the crime pattern, the next anchor is to explore the details of associated offenders, which might be helpful to link unsolved crimes with a plausible suspect/s. It involves answering the questions such as who are the other offenders who have committed similar crimes showing similar characteristics in modus operandi, proximity and time. These associations are examined through a knowledge which is activated by clicking any offender as shown in Figure 4b. The generated knowledge graph of an associated offender is shown in Figure 6. Like clustering, the knowledge graph is also made interactive through a selection of a set of similarity attributes for modus operandi, proximity and temporal information allowing the analyst to set the associations criteria for the associated perpetrators of a given offender as shown in Figure 6.

The knowledge graph of the associated offender “TAJWAR GASKEL” is shown in Figure 6. It shows the group of offenders with whom “TAJWAR GASKEL” has committed crimes together i.e. they have reported in the same crime report. “TAJWAR GASKEL” is connected with the offender “IEZI SPURRLERR”, as a defendant through crime report “125628863”, however, he is nominated as a single defendant in the crime report “125642563”.

The Figure 6b highlights his spatial, temporal and modus operandi similarities with other defendants. The top section of Figure 6b represents the group of offenders having behavioral similarity i.e. modus operandi. For simplicity and sake of the space we have only expanded two nodes of modus operandi i.e. i) the (moFixtureMaterial) i.e. Fixture material used to enter into premises, which in this case is “PLASTIC”, and ii) the “moSearchLocation” i.e. where did the offender search in the premises, which in this case is “DOWN”. These two nodes are further expanded to reveal the location and type of these offenses. The leaf nodes answer the questions who are the other offenders, who have committed these crimes using same the modus operandi (MO). These leaf nodes thus link the offenders on the basis of modus operandi similarity of their committed crimes.

The “moFixtureMaterial” node shows a group of nine defendants consisting of eight men all of them are white skinned Europeans represented by red text and one woman of Asian origin represented by green text. The “moSearchLocation” node, however, shows six offenders who like “TAJWAR GASKEL”, while committing the crime have searched the “DOWN” portion of the premises. It can be seen that the three offenders i.e. “IEZI SPURRLERR”, “PRISCILE CHANG” and “TRAITH LASO” bear the similarities in their modus operandi signature, and hence may be thought to have more close associations with “TAJWAR GASKEL”.

The temporal node of the graph Figure 6b shows that “TAJWAR GASKEL” has committed in the month of the “MAR” and “FEB”. When the “MAR” node of the tree is further expanded, it shows that he has committed the offense “BURGLARY DWELLING” in the district “KNUTT COPSE”. The leaf nodes show the name of other offenders i.e. IEZI SPURRLERR", “PRISCILE CHANG”, and “TRAITH LASO”, who have committed “BURGLARY DWELLING” in the district “KNUTT COPSE” in the same month, and thus link the offenders on the basis of temporal similarity of their committed crimes.

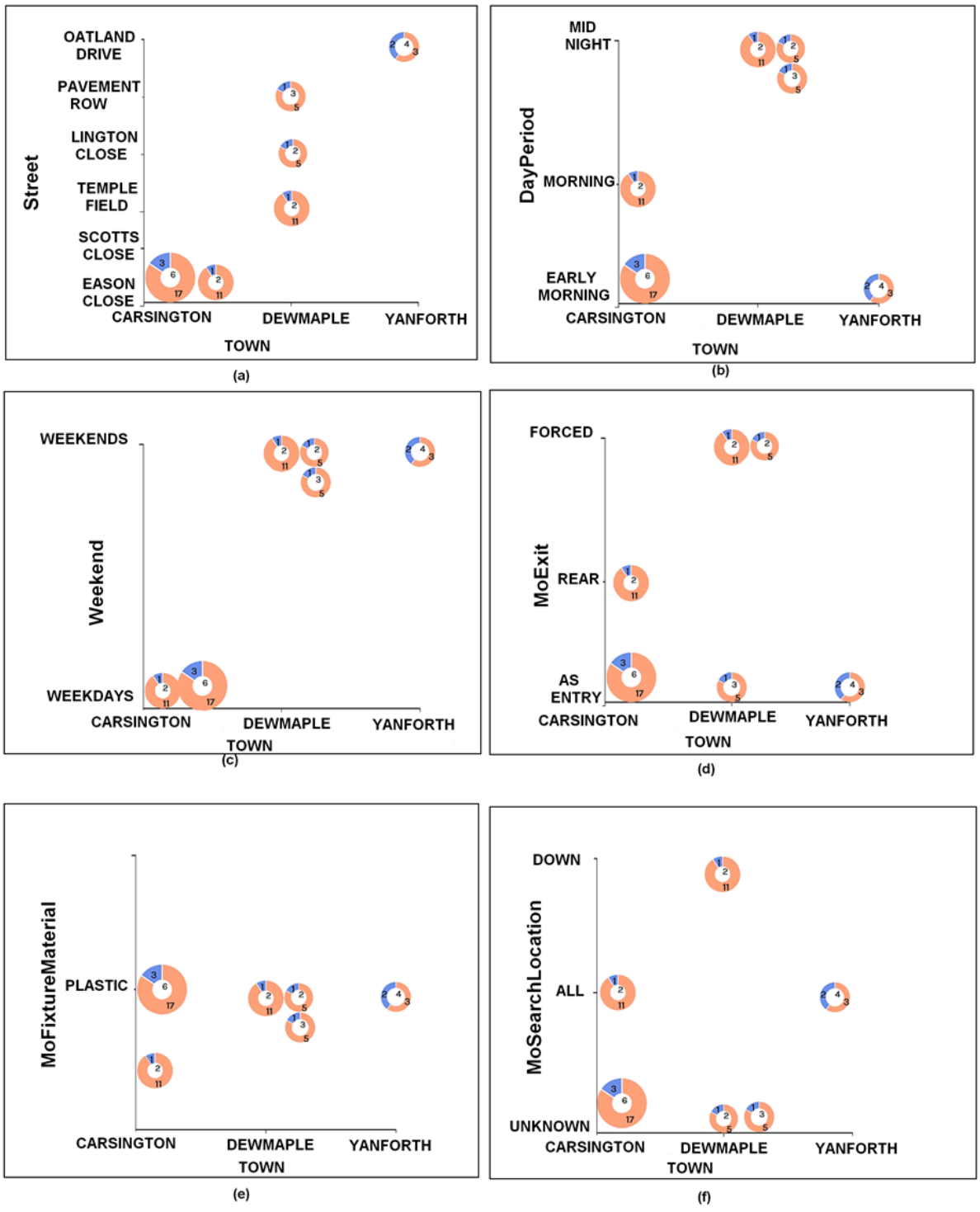

Figure 5 Detailed View of the Interactive Cluster space

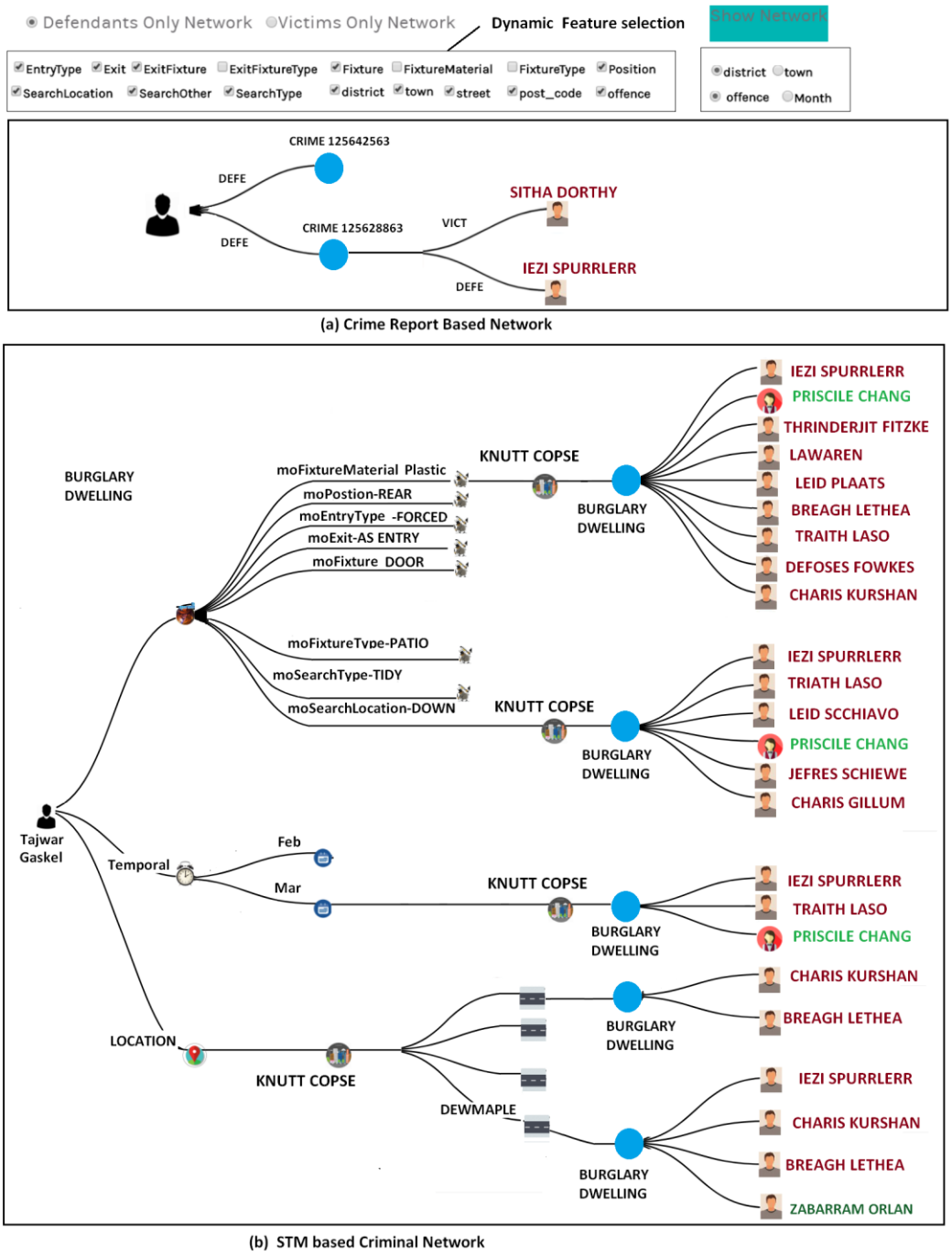

Figure 6 Knowledge graph of an offender: TAJWAR GASKEL.

The last node i.e. the spatial node of Figure 6b highlights the criminal groups with the spatial similarity. Two other offenders showed in leaf nodes, one Asian namely “PRISCILE CHANG” and a European namely “IEZI SPURRLERR” have committed similar crimes in the town “DEWMAPLE" of the “KNUTT COPSE” district. This graph thus shows that Both “IEZI SPURRLERR”, “PRISCILE CHANG” have exhibited very high similarity in crime pattern with that of the “TAJWAR GASKEL”. It is to remind once again that all this information are not real as the data has been anonymized.

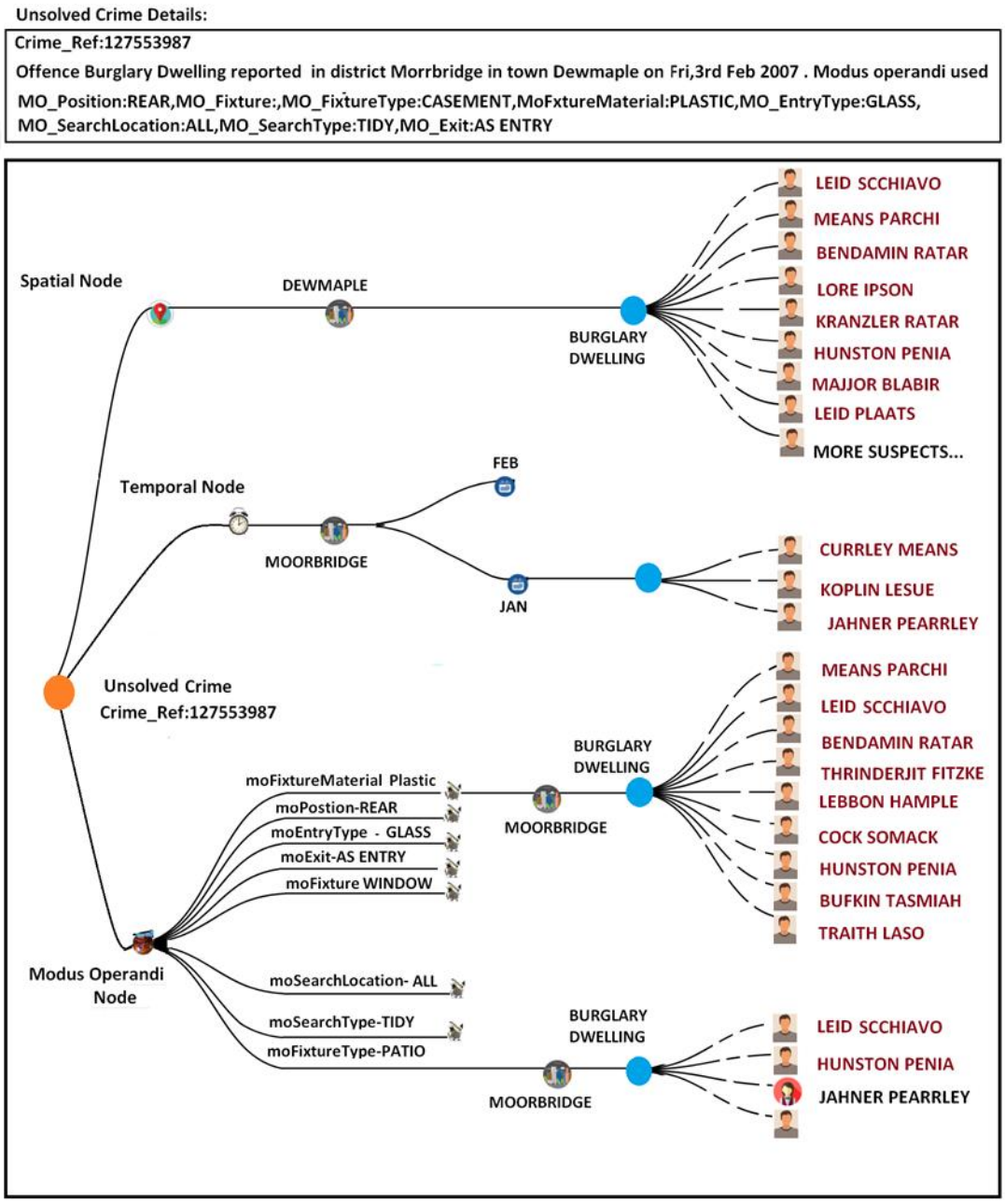

Figure 7 Knowledge Graph: Plausible Suspects List for Un-solved Crime

6.4 Plausible Suspect List Graph

During crime matching, an analyst may also be interested to see who could be possible suspect of an unsolved crime. It requires to link the plausible suspects to the unsolved crime based on the similarity in the given crime pattern. It is where our proposed plausible suspect list knowledge graph facilitates the analyst to compare the similarity of the crime pattern of the suspect/s with the unsolved crime and provides an opportunity to make a hypothesis based on the similarity in crime pattern. It is activated in our proposed framework, as shown in Figure 4b, when the unsolved crime of "Burglary Dwelling" having crime ref "127553987", is clicked, representing a plausible suspect list for the clicked unsolved crime and is shown in Figure 7. The root node for this suspect list

knowledge graph is the unsolved crime (which in this case is “Burglary Dwelling”) and is represented through the dark orange circle, showing a list of the offenders of the similar solved crimes. These offenders due to the similarity in their crime patterns with the unsolved crime may be thought as plausible suspects for the given unsolved crime.

It can be seen in Figure 7 that the root node is branched into three child nodes each for the spatial, temporal and modus operandi component. For example the spatial node in Figure 7 links more than eight suspects who have committed the “Burglary Dwelling” in the same Town (“DEWMAPLE”), where the unsolved crime of burglary dwelling was committed. The temporal node, on the other hand, connects the suspects who have committed the crimes in the same month when the unsolved crime was committed. For example, in Figure 7, three offenders ("CURRENTLY MEANS","KOPLIN LESUE", and “JAHEN PEARRLEY") have committed similar crimes in the month of “JAN” in the same district, where the unsolved crime was committed. The behavioral node i.e. modus operandi node connects suspects having similarities in various modus operandi component. For example, there are eight offenders who have committed the “Burglary Dwelling” using “Plastic” as a fixture material, as shown in Figure 7. When these suspects hover, it shows the detail of their similar committed crimes.

This widget thus connects the suspects to the given unsolved crime in three groups i.e. spatial, temporal and behavioral, each showing the possible suspects based on the respective similarity in the crime pattern. It thus provides an opportunity to link the unsolved crime with the suspects on the basis of the similarity in crime pattern with an unsolved crime.

The proposed knowledge graph also provides a reasoning inference to the analyst, for example, in Figure 7, the offender “LEID SCCHIAVO” has committed the similar crime of burglary dwelling in the same town “DEWMAPLE”, exhibiting same modus operandi (“Mofixturematerial”) as reported in the unsolved crime. In addition to this, “LEID SCCHIAVO” is also present in the co-offender group of the “TAJWAR GASKEL”, who has committed a similar “Burglary Dwelling” in the town “DEWMAPLE” as can be seen in Figure 6b. The suspect list widget thus along with Co-offender network knowledge graph gives the insight to facilitate the analyst in making hypotheses revealing the interesting hidden relationship. This hidden relationship could be helpful in relating an offender towards unsolved crime and thus help in crime matching.

7. Conclusion

In this work, we have presented a human-centered knowledge discovery scheme for the elicitation of tempo-spatial and modus operandi based criminal associations from the crime reports through text mining. We have handled the challenge of integrating human role with the machine learning algorithms during knowledge discovery through a feedback loop running, from the visualization unit to the searching, text mining and association extracting units of the knowledge discovery scheme. It thus avoids the automatic extraction of knowledge through human-machine collaboration, enabling the analyst to interact effectively with machine learning algorithms in making domain-specific semi-automatic models. In addition to this, we also have demonstrated the elicitation of multiple levels associations based on associative search mechanism through a multi-level association model.

The proposed KDD scheme is able to extract plausible associations identifying crime patterns, clusters of similar crimes, co-offender network and suspect list based on spatial-temporal and modus operandi similarity,

The analyst is able to create a 2D re-configurable crime space to see the hidden pattern in the crime reports, implemented through dynamic feature selection. We also have demonstrated the use of this scheme for a given crime pattern where "UPVC Door" or windows" is reported as modus operandi in committing a burglary crime. Our proposed approach of using temporal, spatial and behavioral pattern of a crime scene to find the associations among crime objects, identifying crime hot spots, extracting offender network and plausible suspect list along with the clustering of the similarities of the crimes with proper visualization under a single unified framework have not demonstrated before in the crime mining literature.

We also have demonstrated the use of silhouette analysis to tackle the absence of the ground truth, while grouping the crimes and have utilized the silhouette coefficient to examine the cluster quality of crimes clusters.

The key to this research is the belief, that there exist possible associations within the various dataset used by the analysts. Such associations can provide the basis for activating ideas/thoughts/tentative or plausible conclusions, that could trigger new lines of inquiry. We have shown that the simple visualization of these associations through human-machine collaboration can be helpful for analytical reasoning during a crime matching process. However, we do acknowledge that it does not capture all the problems. Our framework thus enables crime analysts to see the possibility of linkages between data and to make assessment rather than a recommendation.

Acknowledgments

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme through Project VALCRI, European Commission Grant Agreement N° FP7-IP-608142, awarded to Middlesex University and partners. The data used in the examples were anonymized from actual data produced to support the project**.**

References

Alelyani, S., Tang, J. and Liu, H. (2013) 'Feature Selection for Clustering: A Review.', *Data Clustering: Algorithms and Applications,* 29, pp. 110-121.

Alruily, M., Ayesh, A. and Zedan, H. (2014) 'Crime profiling for the Arabic language using computational linguistic techniques', *Information Processing & Management,* 50(2), pp. 315.

Al-Zaidy, R., Fung, B.C.M., Youssef, A.M. and Fortin, F. (2012) 'Mining criminal networks from unstructured text documents', *Digital Investigation,* 8(3â€“4), pp. 147.

Amershi, S., Fogarty, J. and Weld, D. (May 5, 2012) *Regroup: Interactive machine learning for on-demand group creation in social networks.* ACM, pp. 21.

Arın, İ, Erpam, M.K. and Saygın, Y. (2018) 'I-TWEC: Interactive clustering tool for Twitter', *Expert Systems With Applications,* 96, pp. 1-13. doi: 10.1016/j.eswa.2017.11.055.

Bache, R., Crestani, F., Canter, D. and Youngs, D. (2010) 'A Language Modelling approach to linking criminal styles with offender characteristic', *Data & Knowledge Engineering,* 69(3), pp. 303-315.

Bharti, K.K. and Singh, P.K. (2014) 'A three-stage unsupervised dimension reduction method for text clustering', *Journal of Computational Science,* 5(2), pp. 156-169. doi: 10.1016/j.jocs.2013.11.007.

Biscarri, F., Monedero, I., García, A., Guerrero, J.I. and León, C. (2017) 'Electricity clustering framework for automatic classification of customer loads', *Expert Systems With Applications,* 86, pp. 54-63. doi: 10.1016/j.eswa.2017.05.049.

Borg, A., Boldt, M. and Eliasson, J. (2017) *Detecting Crime Series Based on Route Estimation and Behavioral Similarity.* . Sept. pp. 1.

Borg, A., Boldt, M., Lavesson, N., Melander, U. and Boeva, V. (2014) 'Detecting serial residential burglaries using clustering', *Expert Systems with Applications,* 41(11), pp. 5252.

Brehmer, M. and Munzner, T. (2013) 'A Multi-Level Typology of Abstract Visualization Tasks', *IEEE Transactions on Visualization and Computer Graphics,* 19(12), pp. 2376-2385.

Bruneau, P., Pinheiro, P., Broeksema, B. and Otjacques, B. (2015) *Cluster Sculptor, an interactive visual clustering system*.

Bsoul, Q., Salim, J. and Zakaria, L.Q. (2013) 'An Intelligent Document Clustering Approach to Detect Crime Patterns', *Procedia Technology,* 11, pp. 1181.

C. Bennell and D.V. Canter, (2002) 'Linking commercial burglaries by modus operandi: tests using regression and ROC analysis', *Science & Justice,* 42(3), pp. 153.

Cao, L., Joachims, T., Wang, C., Gaussier, E., Li, J., Ou, Y., Luo, D., Zafarani, R., Liu, H., Xu, G., Wu, Z., Pasi, G., Zhang, Y., Yang, X., Zha, H., Serra, E. and Subrahmanian, V.S. (2014) 'Behavior Informatics: A New Perspective', *IEEE Intelligent Systems,* 29(4), pp. 62-80.

Cohen, L.E. and Felson, M. (1979) 'Social change and crime rate trends: a routine activity approach', *American Sociological Review,* , pp. 588-608.

Dagher, G.G. and Fung, B.C.M. (2013) 'Subject-based semantic document clustering for digital forensic investigations', *Data & Knowledge Engineering,* 86, pp. 224-241. doi: 10.1016/j.datak.2013.03.005.

de Zoete, J., Sjerps, M., Lagnado, D. and Fenton, N. (2015) 'Modelling crime linkage with Bayesian networks', *Science and justice,* 55(3), pp. 209-217.

de Zoete, J., Sjerps, M. and Meester, R. (2017) *Evaluating evidence in linked crimes with multiple offenders*.

Didimo, W., Liotta, G. and Montecchiani, F. (2014) 'Network Visualization for Financial Crime Detection', *J.Vis.Lang.Comput.,* 25(4), pp. 433-451.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'The KDD Process for Extracting Useful Knowledge from Volumes of Data', *Commun.ACM,* 39(11), pp. 27-34.

Gonçalves, F.M.F., Guilherme, I.R. and Pedronette, D.C.G. (2018) 'Semantic Guided Interactive Image Retrieval for plant identification', *Expert Systems With Applications,* 91, pp. 12-26. doi: 10.1016/j.eswa.2017.08.035.

Holzinger, A. (2016) *Machine Learning for Health Informatics.*

Holzinger, A. (2013) *Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together?* Springer, pp. 319.

Holzinger, A. and Jurisica, I. (2014) 'Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions'*Interactive knowledge discovery and data mining in biomedical informatics* Springer, pp. 1-18.

Hong Chi, Zhihong Lin, Huidong Jin, Baoguang Xu and Mingliang Qi, (2017) 'A decision support system for detecting serial crimes', *Knowledge-Based Systems,* 123, pp. 88.

Hu, Y., Chen, Y. and Chou, H. (2017) 'Opinion mining from online hotel reviews – A text summarization approach', *Information Processing and Management,* 53(2), pp. 436-449. doi: 10.1016/j.ipm.2016.12.002.

Isah, H., Neagu, D. and Trundle, P. (2015) *Bipartite network model for inferring hidden ties in crime data.* . Aug. pp. 994.

Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I. and Wijayasiri, A. (2015) *Crime analytics: Analysis of crimes through newspaper articles.* . April. pp. 277.

Keyvanpour, M.R., Javideh, M. and Ebrahimi, M.R. (2011) 'Detecting and investigating crime by means of data mining: a general crime matching framework', *Procedia Computer Science,* 3, pp. 872.

Krause, J., Perer, A. and Bertini, E. (2014) 'INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data', *IEEE Transactions on Visualization and Computer Graphics,* 20(12), pp. 1614-1623.

Kulesza, T., Amershi, S., Caruana, R., Fisher, D. and Charles, D. (2014) *Structured labeling for facilitating concept evolution in machine learning.* ACM, pp. 3075.

Lee, H., Kihm, J., Choo, J., Stasko, J. and Park, H. (2012) *iVisClustering: An interactive visual document clustering via topic modeling.* Wiley Online Library, pp. 1155.

Liao, S. and Chang, H. (2016) 'A rough set-based association rule approach for a recommendation system for online consumers', *Information Processing and Management,* 52(6), pp. 1142-1160. doi: 10.1016/j.ipm.2016.05.003.

Lin, K., Zhang, K., Huang, Y., Hung, J.C. and Yen, N. (2016) 'Feature selection based on an improved cat swarm optimization algorithm for big data classification', *The Journal of Supercomputing,* 72(8), pp. 3210-3221.

Min Wang, Fan Min, Zhi-Heng Zhang and Yan-Xue Wu, (2017) 'Active learning through density clustering', *Expert Systems with Applications,* 85, pp. 305.

Oatley, D.G.C., Zeleznikow, P.J. and Ewart, D.B.W. (2004) *Matching and predicting crimes.*

Onan, A., Korukoğlu, S. and Bulut, H. (2017) 'A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification', *Information Processing and Management,* 53(4), pp. 814-833. doi: 10.1016/j.ipm.2017.02.008.

Ozgul, F., Gok, M., Erdem, Z. and Ozal, Y. (2012) *Detecting criminal networks: SNA models are compared to proprietary models.* . June. pp. 156.

Park, A.J., Tsang, H.H. and Brantingham, P.L. (2012) *Dynalink: A Framework for Dynamic Criminal Network Visualization.* . Aug. pp. 217.

Qazi, N., Wong, B.L.W., Kodagoda, N. and Adderley, R. (2016) *Associative search through Formal Concept Analysis in Criminal Intelligence Analysis.* . Oct. pp. 1917.

Reich, B.J. and Porter, M.D. (2015) 'Partially supervised spatiotemporal clustering for burglary crime series identification', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 178(2), pp. 465-480.

Sacha, D., Zhang, L., Sedlmair, M., Lee, J.A., Peltonen, J., Weiskopf, D., North, S.C. and Keim, D.A. (2017) 'Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis', *IEEE Transactions on Visualization and Computer Graphics,* 23(1), pp. 241-250.

Sallaberry, A., Fu, Y., Ho, H. and Ma, K. (2016) 'Contact Trees: Network Visualization beyond Nodes and and Edges', *PLoS One,* 11(1). doi: 10.1371/journal.pone.0146368.

Sanchez, A., Soguero-Ruiz, C., Mora-Jiménez, I., Rivas-Flores, F.J., Lehmann, D.J. and Rubio-Sánchez, M. (2018) *Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions*.

Schroeder, J., Xu, J., Chen, H. and Chau, M. (2007) 'Automated Criminal Link Analysis Based on Domain Knowledge: Research Articles', *J.Am.Soc.Inf.Sci.Technol.,* 58(6), pp. 842-855.

Shixia Liu, Xiting Wang, Mengchen Liu and Jun Zhu, (2017) 'Towards better analysis of machine learning models: A visual analytics perspective', *Visual Informatics,* 1(1), pp. 48.

Stasko, J., Görg, C. and Liu, Z. (2008) 'Jigsaw: supporting investigative analysis through interactive visualization', *Information Visualization,* 7(2), pp. 118-132. doi: 10.1057/palgrave.ivs.9500180.

Sun, Y. (2013) 'MINING HETEROGENEOUS INFORMATION NETWORKS', .

Thongsatapornwatana, U., Lilakiatsakun, W., Kawbunjun, A. and Boongoen, T. (2017) *Analysis of criminal behaviors for suspect vehicle detection.* . Sept. pp. 15.

Thota, L.S., Alalyan, M., Khalid, A.O.A., Fathima, F., Changalasetty, S.B. and Shiblee, M. (2017) *Cluster based zoning of crime info.* . March. pp. 87.

Tonkin, M., Woodhams, J., Bull, R. and Bond, J.W. (2012) 'Behavioural case linkage with solved and unsolved crimes', *Forensic science international,* 222(1), pp. 146.

Valle, M.A., Ruz, G.A. and Morrás, R. (2018) 'Market basket analysis: Complementing association rules with minimum spanning trees', *Expert Systems With Applications,* 97, pp. 146-162. doi: 10.1016/j.eswa.2017.12.028.

Viktoratos, I., Tsadiras, A. and Bassiliades, N. (2018) 'Combining community-based knowledge with association rule mining to alleviate the cold start problem in context-aware recommender systems', *Expert Systems With Applications,* 101, pp. 78-90. doi: 10.1016/j.eswa.2018.01.044.

Vural, M. and Gök, M. (2017) 'Criminal prediction using Naive Bayes theory', *Neural Computing and Applications,* 28(9), pp. 2581-2592. doi: 10.1007/s00521-016-2205-z.

Wang, J. and Lin, C. (2011) *An Association Model Based on Modus Operandi Mining for Implicit Crime Link Construction.* Washington, DC, USA: IEEE Computer Society, pp. 548.

Wei Chen, Cong Xie, Pingping Shang and Qunsheng Peng, (2017) 'Visual analysis of user-driven association rule mining', *Journal of Visual Languages & Computing,* 42, pp. 76.

Wong, B.L.W. and Kodagoda, N. (2016) 'How Analysts Think', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 60(1), pp. 178-182.

Zarrinkalam, F., Kahani, M. and Bagheri, E. (2018) 'Mining user interests over active topics on social networks', *Information Processing and Management,* 54(2), pp. 339-357. doi: 10.1016/j.ipm.2017.12.003.