Contextual Visualization of Crime Matching through Interactive Clustering and Bayesian Theory

Nadeem Qazi[1], B.L. William Wong[1]
[1] Interaction Design Centre, Middlesex University
Burroughs, Hendon, NW4 4BT,London, United Kingdom
Email: n.qazi@mdx.ac.uk,w.wong@mdx.ac.uk

Abstract

One of the key challenges in crime matching is to identify the possible associations among crime entities. Aiming towards a solution of this challenge, in earlier research, we introduced associative search based on the 5WH questioning model to elicit associations among various crime entities. We demonstrated the use of three-dimensional, i.e. spatial, and temporal and modus operandi based similarity matching of crime pattern to establish hierarchical associations among the crime entities. We later employed it to develop a visualization tool that assists in discovering a possible suspect list for an unsolved crime. This answers the first question of the crime matching process: who might have committed the given unsolved crime?

In this paper, we extend our visualization framework of crime matching to visualize the spatial, temporal and modus operandi based multi-level associations in two-dimensional crime cluster space, projected through partition clustering. Each of these clusters groups the solved, unsolved crimes with associated perpetrator based on spatial, temporal and modus operandi similarity of the crimes. Additionally, we also attempt to answer the second question of the crime matching process: is it possible for a suspect to have committed a particular unsolved crime? We answer this question through identifying and extracting all the possible similarity linkages between a suspect and unsolved crimes. The plausible unsolved crimes are grouped and linked to the suspect corresponding to the behavioural and modus operandi pattern of the suspect. This is then depicted through acyclic graph weaving an association tree of plausible associations of the suspect with unsolved crimes.

We also evaluated this similarity based associations through Bayesian theory in order to examine how the crime pattern of the suspect in one crime can be used in another for making any type of association. This is implemented through measuring prior and posterior probabilities of suspect in committing a crime employing crime pattern characteristic as evidence.This thus facilitates crime matching process through visualizing all the plausible similarities and also assists in determining the likelihood of the given suspect in committing an unsolved crime.The proposed visualization aims to assist in hypothesis formulation reducing computational influence in the decision making of criminal matching process.

keywords— Data mining,exploratory data analysis,associative questioning,data visualization,knowledge graph,Naive Bayes Theory

I. Introduction

Analysts during crime matching assign criminals to the previously solved or unsolved crimes, establishing an association between them.They usually employ crime variables such as time, location, modus operandi of the committed crime, the profile of the offender to answer a question such as Given a Y crime identified by a modus operandi pattern, who could possibly be X offender.Alternatively in another approach of crime matching,they possibly try to answer the question for a given an X offender, what could be possible Y unsolved crimes, that could be potentially thought to be committed by him based on similarity pattern of the unsolved crimes pattern and the suspects.They possibly try to find out what are the chances that an offender to be responsible for an unsolved offence.

This process of crime matching is indeed a information-intensive querying process that requires establishing associations among crime entities to discover and reconstruct crimes through analysis of the evidence left at the crime scene.Analysts spend a large amount of time performing extensive database searches, reading crime reports looking clues for criminal associations among criminal entities such as criminals, vehicles, weapons, bank accounts, and organizations.The resultant data from these searches often requires a tedious job of grouping the crimes based on the crime pattern similarities.In addition to this, It also lacks appropriate visualization to help cognitive thinking process. It, therefore, causes analysts to 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.

In our earlier research towards the criminal analysis[1] ,we observed that police analysts follow a search based on associative questionings for establishing associations among the criminal to discover and reconstruct crimes scene. For example analyst for a given crime pattern,may be interested to generate contextual, cognitive and domain- specific queries such as Who else has used the same modus operandi in the similar crime, Who are the other offenders who have committed similar crimes, When was a suspect last stop and searched etc.This search however is also different than traditional keyword or semantic base search and requires a special search mechanism.Following this need in our earlier research we introduced associative search mechanism in digital criminal analysis[2].

Associative search, unlike keyword and semantic based search, searches along the networks of associations between objects such as people, places, other organizations, products, events, services, and so forth. we proposed a 5WH model made of five general but associated concepts, each representing a question and are linked with each other through a set of properties or attributes to reveal the relationship, association or relevancy among these crime concepts. These associated concepts are WHAT (what has happened i.e. type of offence), WHO (who has committed the crime identifying offender or group of offenders), WHEN (When an offence happened), WHERE (the Geo-spatial information about the offence) HOW (The modus operandi used in the committing a crime). We later employed it to develop a visualization tool that assists in discovering a possible suspect list for an unsolved crime. This answers the first question of the crime matching process: who might have committed the given unsolved crime?

In this paper, we extend our visualization framework of crime matching,employing data mining techniques for grouping non-structured crime reports forming two-dimensional crime cluster space,visualising multi-dimensional associations. Each of these clusters groups the solved, unsolved crimes with associated perpetrator based on spatial, temporal and modus operandi similarity of the crimes. Additionally, we also attempted to answer the second question of the crime matching process: is it possible for a suspect to have committed a particular unsolved crime? We answer this question through identifying and extracting all the possible linkages between a suspect and unsolved crimes based on spatial and modus operandi characteristic.In addition to this we also demonstrated the use of Naive Bayes theorem to determine the likelihood of an offender to be linked with the unsolved crime.

The contributions of this paper include 1) an association discovery scheme for multi-level associations for identifying criminal linkages with unsolved crimes,identifying and measuring association between suspects and unsolved crime through Naive Bayes theory. 2)Interactive clustering distinguishing the relevance or similarities among the cases and our approach to handle categorical data in the clustering and lastly 3)visualisation of hierarchical associations of crime entities through radial knowledge graph.We have employed vector space model, interactive partition clustering, and graph theory, to visualize these multi-level associations using temporal, spatial and behaviour characteristics of a given crime scene.The employed visualization project these computationally unconnected but operationally plausible associations in a same visual field of view to make sense out of this association.It enable analysts to see the possibility of linkages between data.It thus helps analyst to find out how crime entities that appear to be unrelated at the surface, are actually linked to each other thus facilitating hypothesis formulation encouraging questioning for sense-making and active learning. The developed framework thus provides a complete data analytic solution to expedite the criminal matching process and facilitates the analyst in identifying and understanding associations between criminal entities.

This rest of the paper is organized as follows. Section 2 presents related research followed by dataset description in section 3. We discuss the proposed knowledge discovery scheme for crime matching, describing each component including interactive clustering,association extractor and visualization in sections 4.The proposed visualization and its preliminary evaluation are described in section 5 and 6 respectively,while conclusion is made in the final section of the paper.

II. Related Work

The work presented in this paper integrates multiple domains from data science for criminal analysis under a single framework.It includes clustering for identifying similar crimes,criminal associations for finding relationship between crime entities,Naive Bayes algorithm for determining likelihood of criminal associations with the unsolved crime and lastly the network or tree visualization. Researchers have demonstrated the use of these domains separately to solve different criminal analysis problems. Clustering techniques either independently or in combination with other data mining techniques such as self-organizing map [3] has been utilized in criminal investigation for grouping similar crimes. For example [4] has utilized K-Means clustering algorithm on real time data acquired from sheriffs office to find the crime pattern.Researchers[5] have employed clustering technique to detect money laundering in banking industry, [6] performed crime data clustering for crime matching in two phases. The first phase uses a neural network to mine the attribute map and then K-Means algorithm groups the output. Recently, [7] has presented a two phase clustering algorithm called AK-modes to automatically find similar case subsets from large data sets.Reference [8] proposed a semi-supervised clustering approach based on Bayesian model,utilizing crime locations and offenders modus operandi for burglary crime series identifications.In an another approach researchers [9] have detected residential burglaries series through minimum cut based graph clustering.They used a feature vector consisting of modus operandi, residential characteristics, stolen goods, spatial similarity, to group similar crimes. Researchers [10] detected crime patterns on the news through document clustering using ten types of crime including traffic violation, theft, sex crime, murder,kidnap, fraud, drugs, cyber crime and arson gang articles.They employed affinity propagation algorithm for determination of the number of clusters to be fed in K-Means algorithm.VISFAN [11] employed k-cores hierarchical clustering to visualise financial activity networks connecting entities like bank accounts,addresses, amount and types of the transactions, motivations etc extracted from financial reports.

There is also a growing trend in criminal analysis towards the use of Naive Bayes to model crime linkage. Reference [12] proposed a crime linkage model based on Bayesian networks to show how evidences observed in one crime ,can be used in another and vice versa.Their proposed crime linkage model though simplifies the reality but does not capture all

the problems that play a role when linking crimes.Likewise working in the same direction [13] have employed Bayesian networks for modelling multiple offenders for two separate offences.They have discussed a mock case example to show that subtle differences between situations can lead to substantially different conclusions in terms of posterior probabilities of a certain suspect being one of the offenders in a particular crime. Reference [14] have constructed a Naive Bayesian model over synthetically generated incident-level crime data to express the probability of criminal for committing a crime.They used crime date,location,criminal name and criminals acquaintances as clues to predict the posterior probability of a criminal to be associated towards an unsolved crime.

Recently criminal network analysis has received great attention from researchers. Dynalink [15] is a framework for visualizing criminal network,uses the animation approach to visualize the changes of networks over time,thus assisting in discovering and analysing both relational and temporal patterns of criminal networks.Reference [16] have used bipartite network model for extracting hidden ties in both traditional and cyber crimes over pharmaceutical crime and underground forum data set respectively.Among other systems that use network visualization tools we mention,NETMAP,ANALYSTS NOTEBOOK, COPLINK by [17], [18], [19],Prep-Search [20],(RECAP)[21], and CrimeLink Explorer by [22].In addition to this JIGSAW by [23] is another interesting system that supports investigation processes.It extracts relevant entities (persons, addresses,dates, etc.) from a collection of documents and show correlation,connection between these entities. It uses nodelink diagram in a basic graph view to display connections between entities and documents thus allowing analysts to explore the documents. [24] have employed association analysis with reputation algorithm to identify the criminal vehicles that are potentially involved in criminal activities.They used criminal vehicle and criminal activity data set of Defence Technology Institute (DTI), Thailand.They employed matching of license plate, colour,brand, and type of target vehicle with that of one year checkpoint crossing data set to establish the criminal activity of the target vehicle.

The work cited above however, either lack appropriate visualisation for associations or lack knowledge discovery ability for association identification and require this information to be fed in by the user for its visualization. We in this work, bring the interactive dynamic visualisation along with the knowledge discovery ability of multi-level criminal associations in a single envelope over anonymized burglary data set.

III. Data-Set

Anonymized Burglary Dataset was collected from UK Law Enforcement agency for the experiment to evaluate the performance of the proposed associative search mechanism. The police utilize several kinds of updated information including a reported crime,modus operandi description, stop & search, Forensic and intelligence data gathered from different sources.

A crime report document is generally defined to be a logical unit of textual data consisting of the crime reference, offender information, modus operandi description, offence category, location, time and date of the crime occurred along with other related informations. Modus operandi is the method of the operation adopted to commit a crime, it can be preparation actions, crime methods, weapon used, position of entry etc. Stop & search information contains data of a person or vehicle whenever they are stopped for a specific reason, this information holds the name, address of the person along with car number plate, location and reason of the stopped and search and details of the duty officer.

For this work, we have used crime reports consisting of 164,800 record cases of five different types of burglary along with other crime entities such as nominals, modus operandi description and stopped and search from the provided dataset to generate the linked data.Twelve modus operandi variables including Entry position,Entry type,Fixture,Fixture type,Fixture material,Search location,Search type,Exit type, and Exit Fixture etc were used for each of the crime cases. Each of these modus operandi variable has a set of predefined values for selection in crime report filing.

IV. Association mining Scheme

The propose association discovery pipeline for crime analysis shown in Figure 1 is inspired by general process of KDD [25]. It takes a crime pattern as input and elicit the multi level associations and visualize it in 2D interactive clustering space along with criminal network and their hots spots in the form of a knowledge graph.This pipeline consist of associative query engine that during retrieval and integration phase generates the spatial,temporal and modus operandi based associative queries(presented below) for a given crime pattern, extracting data from the knowledge base for association extraction. The filtered data from the associative query engine is then fed into association miner which elicit multi-level associations to group the crime entities,which are then visualized through interactive clustering and knowledge graph component of the visualization module.

1) What other offences have occurred, for a given crime pattern?
2) Where else crimes like this have been committed?
3) Who are the known offenders operating in an area and what is their modus operandi to commit crimes?
4) Who else in past have committed crimes like this?
5) How often offences like the given crime pattern has occurred?
6) What are other modus operandi that has been used in committing crimes like this?
7) What are the additional details of the associated offenders/victims his/her past history etc.?

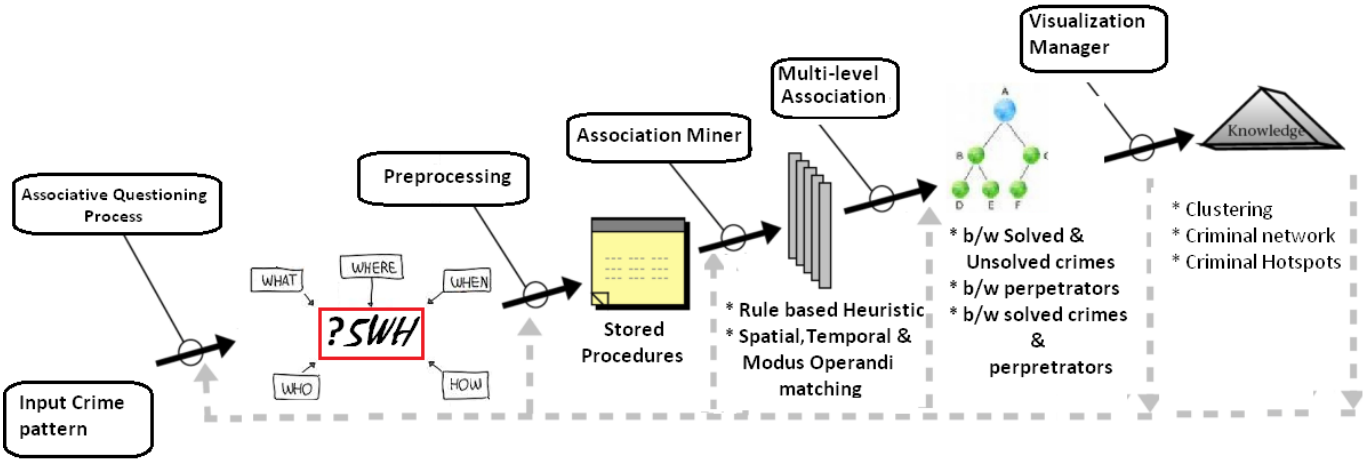

Fig. 1. Association Discovery Pipeline.

8) What are the Geo-spatial profiles of the offenders, including its temporal,spatial and other similar criminal activities resembling with the given crime pattern?
9) How many times the offender has committed the similar crimes and what are its temporal and spatial details?
10) What is his/her pattern of modus operandi?
11) Where an offender mostly likes committing an offence and who else has committed the same crime at this location?

A. THE ASSOCIATION Miner

The association miner unit of our KDD scheme elicit multi-level associations between crime entities as shown in the Figure 2. These associations elicited through rule-based heuristic and the similarity matching reveal entity associations and group identification in the linked crime dataset. The entity association relates solved,unsolved crimes and the associated offenders or victims as depicted in level 1 and level 2 of association hierarchy of Figure 2.The third level of associations in our proposed scheme assist in elicitation co-offender network,a plausible suspect list for an unsolved crime and plausible unsolved crimes list for a given offender. We propose to use these two lists in crime matching process and have represented these lists in the form of connected network showing spatial-temporal and modus operandi characteristics.The intuition behind using this approach is to find out how crime entities that appear to be unrelated at the surface,are actually linked to each other. The extraction rules for theses associations are described in detailed in the next section.

1) Crime and crime Associations: For distinguishing solved crime with un-solved crime we have followed a rule base heuristic approach. The implemented rule is that a solved crime is the one which has been solved and an perpetrator/offender has been identified/sentenced for a crime, and unsolved crimes is the one, which yet have to be solve without knowing who has committed this crime, and the goal is to identify potential/probable offender responsible for committing the crime. Mathematically, Let in the given Crime dataset $P = P_1, P_2, P_3...P_n$ be the set of the distinct full names of offenders/perpetrators; and $C = C_1, C_2, C_3, C_m$ be the set of the distinct Crime ids, then we tag a crime C_s from this subset as solved crime only if C_s contains at least one full name of the offenders/perpetrators taken from the set of offenders P in the knowledge-base.In other words a solved crime has one or more known offenders associated with it given by the following mathematical equation:

$$\exists C_s | \forall P_s \subseteq P, C_s \in P_s$$

All the crimes which do not satisfy the above relation are termed as unsolved crimes in this work.

2) Group Association: Researchers [26] and [27] have defined modus operandi as an important factor for establishing associations between crimes and criminals. In addition to this,Scottish philosopher David Hume has described contiguity in time and place, resemblance and causality as three main factors for determining the associations. We have merged the finding of these researchers and have used spatial,temporal and modus operandi to elicit plausible similarity based multi-level associations among crime entities. Later,inspired by the notion of heterogeneous representation of linkages[28],we have visualized these elicited multi-level associations in the form of a heterogeneous network through acyclic graph.

The generated associations network graph is consisted of multiple types of nodes and edges.These nodes may be perpetrator, location, offence, time and modus operandi. Each node of the perpetrator in turn may also have a different type that varies with their role such as defendant, victims etc.The edges in the network are associations that connect the nodes on basis of the

similarities in spatial, temporal and modus operandi characteristics, and for this reason we have named this network graph as spatio-temporal modus operandi(STM) network.

We propose to use this STM graph network to represent the Level 3 associations of our proposed framework Figure 2, a separate STM network for plausible suspect list, co-offender and plausible unsolved crime list, each having different root node and leaf nodes. For example the STM network graph for plausible suspect list, takes a given unsolved crime as root node and extract name of the suspects for unsolved crimes as leaf nodes of the network. Like wise, for co-offender network the root node is given offender and its co-offender are shown as leaf nodes of the network. Finally for the plausible unsolved crimes list which is described in this paper, the root node would be offender and leaf nodes would be unsolved crimes that may be committed by the offender. In our earlier study we have described plausible suspect list to answer first question of the crime matching process: who might have committed the given unsolved crime?.In this paper we focused towards the plausible unsolved crimes list to answer second questions of crime matching: is it possible for a suspect to have committed a particular unsolved crime?.

3) Plausible Unsolved crime list: We elicit the plausible crime list that may be committed by a suspect employing similarity matching of the crime component of the suspect with that of unsolved crime and later visualized in the heterogeneous network as mentioned earlier. We filtered the data based on the associative questioning given below and then compared the crime pattern of unsolved crime with that of suspect to elicit the node of associations.

1) What are the crimes that have been occurred in the criminal area of the suspect.
2) What are the similarity of unsolved crime reported in that area to that of the crime pattern of the suspect.
3) Who are the other offenders who have committed in the vicinity of the area and time, and what are their similarity with that of unsolved crime.

In addition to this we also have employed Naive Bayes theory to evaluate how likely it is that an unsolved crime may be thought to committed by the given offender. The Bayes theorem provides a way of calculating posterior probability given by the following equation:

$$p(H \mid E) = \frac{p(E \mid H) * p(H)}{p(E)} \qquad \text{Eq(1)}$$

where H is the hypothesis, in the legal context this is usually the statement for example "Defendant is innocent,which could be either true or false.The prior belief about the hypothesis is measured through prior probability of H and is written as $p(H)$,representing the probability of initial belief. The evidences such as crime pattern of the defendant may effect the initial hypothesis, the Bayes theorem answers the questions about the revised or posterior probability $p(H \mid E)$ of (belief), given the evidence. the term $p(E \mid H)$ is called likelihood and it answers the questions "what is the probability of seeing the evidence given our assumed hypothesis H. The term $p(E)$ in the denominator is the probability of the evidence irrespective of our knowledge about H.

We calculated the posterior probability of all the offenders that has shown the similarity to the unsolved crimes. We considered the spatial,temporal and modus operandi used in committing a crime as evidence and have set the hypothesis that a given suspect is committed the unsolved crime. The evidence vector in our case is consisted of Point of entry,Method of entry,Location of search,Type of search,Point of exit,Fixture,Fixture type,Fixture material, Day of week and Time of day as reported in a crime report. We however,made an assumption that all of these evidences are conditionally independent of each other to use the Bayesian rule for probability or in other words,we assume the presence or absence of a particular feature does not affect any other feature.

All of the evidence features in our case are categorical in nature, therefore we first created a frequency table for all of these evidences from the solved crime data and then calculated the prior probability of the evidences, their likelihood for a given offender for burglary crimes along with prior probability of the offender committing the burglary crimes to answer following questions:

1) What is the probability that a suspect may committed a burglary crime?
2) What is the probability that a suspect has committed a burglary crime in the same spatial location where this unsolved crime has occurred?
3) What is the probability of using a specific modus operandi by a suspect in committing a burglary?
4) How many similar crime, he/she has already committed?

Finally,we validated the prior probability of hypothesis through the posterior probability using equation 1 for the evidence mentioned in unsolved crime. Additionally,following the above mentioned method,we also have calculated posterior probabilities of all those suspects who have similarity in their crime pattern to that unsolved crime and present few of them here in Table I. However this model depends on the data and lack all the other relevant information that play a role when associating a crime with an offender. This posterior probabilities thus may be helpful to uncover the interesting aspects of the reasoning for linking unsolved crimes with a known offender.

The association scheme described above generates a heavy multi-dimensional information load consisting of key process indicators(KPI) such as total crimes, number of solved and unsolved crimes, associated offenders,proximity of the offence (represented through postcode), modus operandi used in committing an offence such as entry position, fixture type used, search

TABLE I
POSTERIOR PROBABILITY OF OFFENDER TO COMMIT AN UNSOLVED CRIME

	Unsolved CrimeID:191657520	Unsolved CrimeID:181908600	Unsolved CrimeID:251771339
Rator Gayor	0.698	0.217	0.901
JAYDA NAZHAT	0.757	0.657	0.815
JOHHAN TRINA	0.857	0.709	0.215

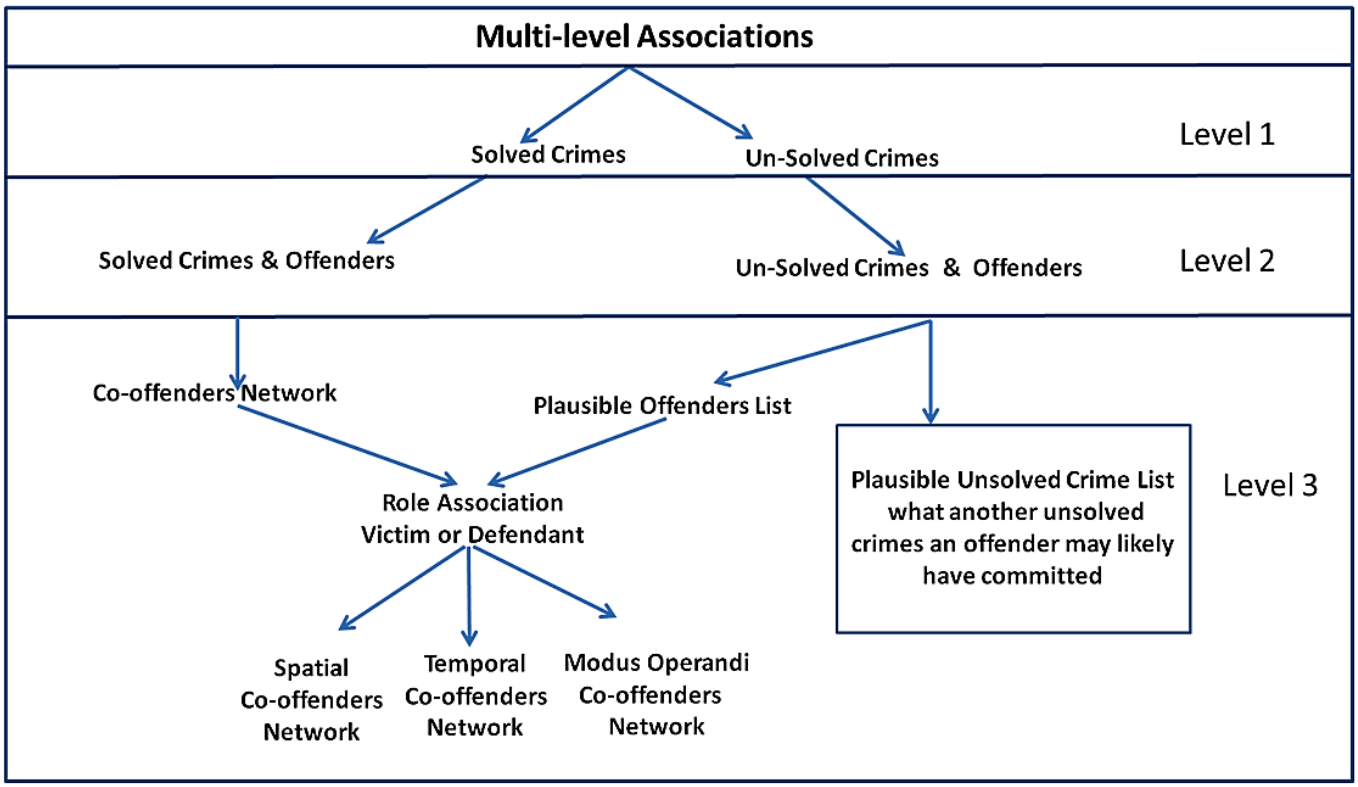

Fig. 2. Three Levels of Associations.

location, entry type etc.Aiming towards both explanatory and exploratory data criminal data analysis, we tackled this multi-dimensional information load through interactive clustering creating a user defined 2D dimensional crime space, grouping and projecting these association based on KPIs. We aim to project the association in every possible KPI dimension through grouping the similar pattern based on resemblance and contiguity in time, proximity and modus operandi and implemented it through interactive clustering which is defined in the next section.

B. SPATIAL-TEMPORAL AND BEHAVIOURAL SIMILARITY BASED CRIME CLUSTERING

Clustering is an unsupervised machine learning technique that groups data on the basis of similarity, however, it faces two important challenges including feature selection and categorical data handling. We selected spatial, temporal, and modus operandi component of the crime pattern as an investigative lens to form the cluster space aiming to group crimes into multiple clusters such that the similarity in a cluster is larger than among the clusters. We represented location through postcode, street, and town, temporal information through month day and time of the offence occurred. The time of the offence is a many-valued attribute having any value within 24 hours of the day. We, therefore, used the idea of conceptual scaling to transform this attribute into its symbolic value which resulted in four periods of the day: morning (from 6 a.m. to 12 a.m.), afternoon (from 12 a.m. to 6 p.m.), evening (from 6 p.m. to 12 p.m.), and night (from 12 p.m. to 6 a.m.), and lastly used 12 other variables described above for modus operandi. However using all these temporal, spatial and modus operandi attributes of the crime in a feature vector produces high dimensional feature vector, so unlike the existing trend in the crime clustering,the idea of dynamic feature selection was employed for interactive clustering.This enables a user to redefine the feature vector selecting or de-selecting any of temporal, spatial or modus operandi variables to re-cluster the crimes. This dynamic configuration of the feature vector for clustering also help in determining the effect of crime attributes in clustering.

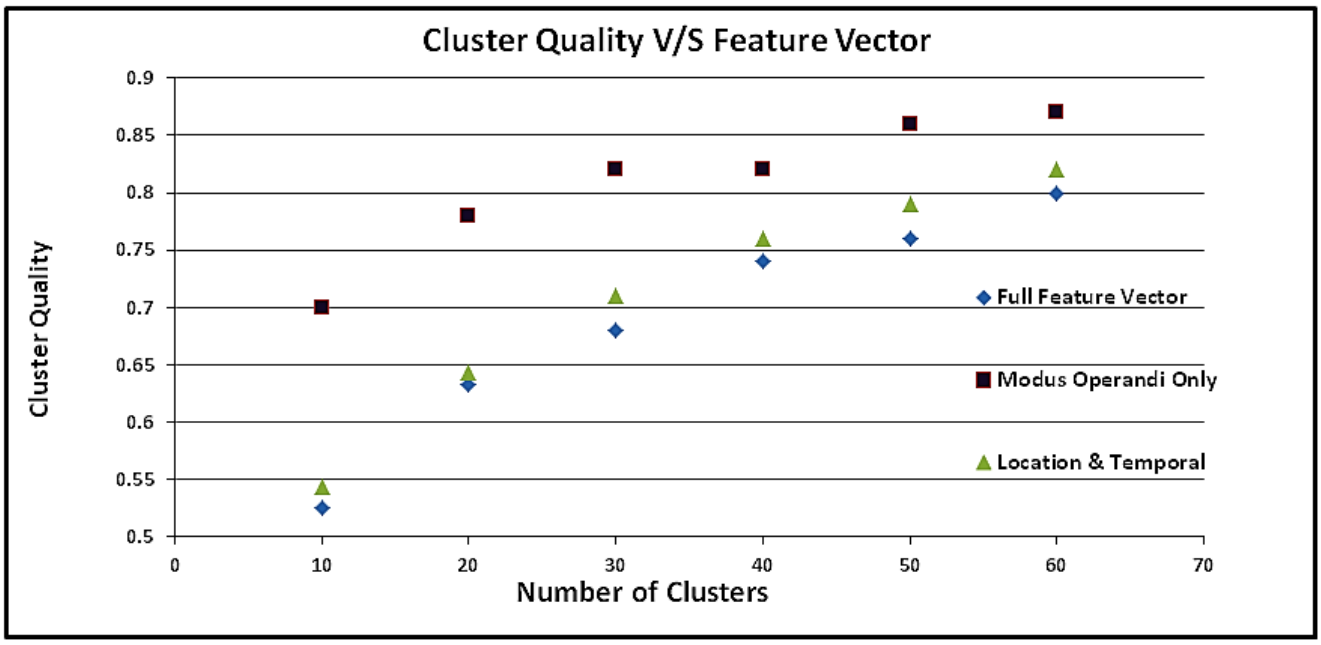

Fig. 3. Effect of Feature vector on Cluster Quality.

The data set used in this work mostly contains categorical variables, and compared to numeric data, due to different data type, cannot be used directly in the standard clustering. Therefore,it poses a unique challenge in clustering tasks of categorical variables. Some researchers [6] have demonstrated the use of categorical data into clustering algorithm by converting these variables into a binary attribute 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 Vector Space Model (VSM) and through the process of vectorization created a bag of words or (crime terms in this case) from the set of crime documents. Each of these words was weighted through the product of term frequency and inverse document frequency (TF-IDF). A numerical matrix having rows as crime documents and columns containing these weighted crime terms was constructed and used to represent the crime document. This term matrix was then fed into the K-Mean clustering algorithm. However knowing the fact that the mean is not very sensible for sparse data, and text vectors tend to be very sparse,we adopted cosine similarity as the distance function in the K-Mean algorithm rather than usual euclidean distance function.The K-mean algorithm, however also requires a prior number of cluster, therefore in addition to the user specified value for number of the cluster, we also used domain expert knowledge for calculation of the numbers of clusters given by the equation:

$$\frac{\text{Total crime documents in crime space}}{\text{User specified number of documents in a cluster}}$$

This process of clustering thus generated clusters based on similarity of the chosen feature vector, grouping solved and unsolved crime along with the associated offenders in each cluster.The quality of each cluster was also measured through a similarity score based on the sum of the similarity distances of each member of a cluster with its centroid.We also have measured the effect of the attributes in clustering and its cluster quality as shown in graph of Figure 3. We found cluster were more homogeneous when modus operandi alone was used as feature vector in clustering.

V. ASSOCIATIONS VISUALIZATION FOR CRIME MATCHING

We projected these clusters on a two-dimensional crime space having reconfigurable X and Y axis i.e. each of the KPIs can be set on either X or Y axis,thus enabling analyst to observe the relationship of the KPI with respect to each other.The adopted methodology for this reconfigurable crime cluster space is described in next section.

A. Reconfigurable Crime Cluster Space

It is implemented through mapping similarities of all the resulted clusters with each other in a two dimensional space.We first calculated a nxn distance matrix of centroids of each of the clusters and then converted it into a global similarity matrix through subtracting each element of the distance matrix from 1. Later multi-dimension scaling was employed to map this nxn

Fig. 4. Aggregated view of the Interactive Cluster space.

similarity matrix to the configuration points $x_1, x_2, x_3 x_n$ in such a way that the given similarities S_{ij} between any two clusters are well approximated by the distances $|x_i - x_j|$. Following this, either X or Y axis of the cluster space, when set to the cluster similarity, would arrange clusters in a fashion, such that clusters those perceived to be very similar to each other will be placed near to each other and those are perceived to be very different from each other 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. For example if a user sets cluster global similarity on X-axis and total number of the offenders in a cluster on Y-axis as shown in Figure 4, than the clusters would arrange themselves revealing how total number of offenders are distributed over cluster space as a function of clusters global similarity.In addition to this, a user can also set other KPIs such as proximity, total crime, and total offenders on any of these axis to see the hidden relationship on a 2-dimensional crime space. In this way any of KPIs can be set on either X or Y-axis to visualize crime clusters in user specified dimension revealing more insight of the data.Thus for example when proximity is chosen on X-axis and crime population on Y axis, a user can spot the associations of the proximity with the crime population on the 2D crime space.

B. Crime space: Aggregated View

We employed multi coordinated view technique to presents multidimensional associations in two separate views the aggregated and detailed view. The aggregated view, as shown in Figure 4, represents the summary of the crime KPIs inside a cluster. Each cluster in the Figure 4, is represented as doughnuts consisting of number of unsolved, solved crimes, and their associated offenders. The solved and unsolved crimes are represented by the two arcs of with arcs length showing their numbers whereas icon glyph is used to visualized offenders inside the doughnuts. The radius of the doughnuts is kept proportional to the sum of solved, unsolved and associated offenders, so that a big doughnuts represent a heavy population of the crime KPIs .

C. Crime space:Detailed View

The detailed view of the crime space as shown in Figure 5 depicts how crimes are related with each other on basis of the similarity of the given crime pattern. Each cluster is represented as big circle, showing three type of association including crime object i.e. crime and offenders, then type of the objects i.e. solved and unsolved crime ,associations of solved crimes with the offenders and lastly local similarity of crimes.Inside each cluster, each crime was represented as a turquoise and

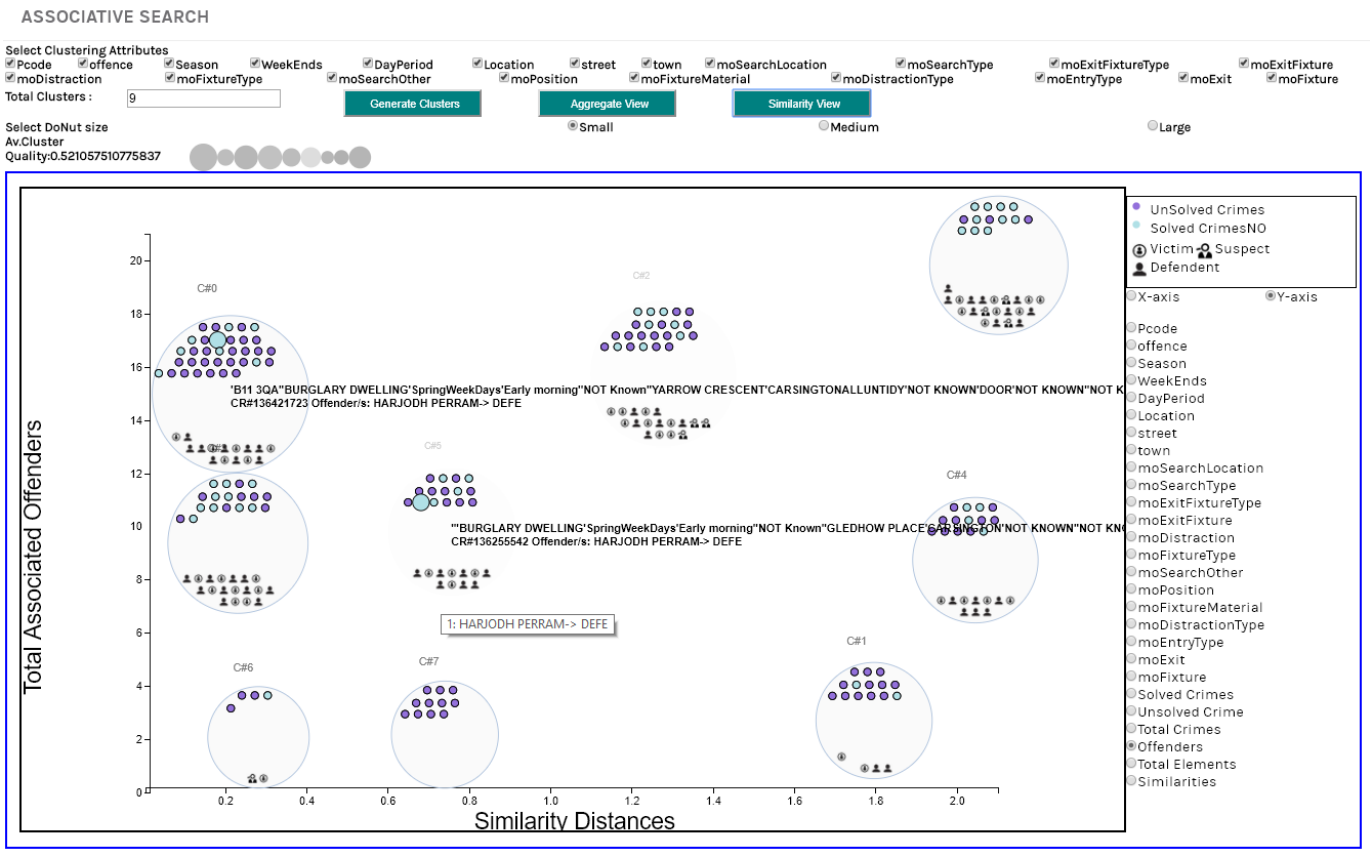

Fig. 5. Detailed view of the Interactive Cluster space.

purple circle for solved and unsolved crime respectively. Hovering on each of these circles shows the information of the crime such as crime reference numbers as tool-tip. An offender may be associated with more than one crime, it is visualized through focus and context technique. When an offender is focused or hovered through mouse its association with all of its associated solved crimes in any cluster is also highlighted through increasing the size of the related solved crime circles in the clusters, Figure 5, which goes back to normal when hover is off.

This detailed view thus encourages the analysts for further insight of the crime to visualised crime entities.Once an interesting solved crime is spotted to match with unsolved crime, the analyst may be curious to see the details of the associated offender such as criminal history, his criminal network,etc. It is where we have used the idea of Knowledge graph which is discussed in the next section.

D. Knowledge Graph

We present criminal associations in the form of a heterogeneous network called as knowledge graph in our framework. This graph presents the associations in a parent-child relationship weaving a heterogeneous radial tree graph structure as shown in the Figure 6. This tree structure clusters the perpetrators into three main groups corresponding to the spatial temporal and modus operandi similarity, showing linkages between a chosen root node and other linked entities.The root and child nodes are collapsible and are represented through iconic graphics while thick and broken lines are used to show the relationship.These collapsible nodes when are clicked expands to shows the level of similarity.The solid nodes means they have children and can be expanded,where the expanded nodes are shown hollow having no further children.

In the Figure 6 the root node for plausible crime list represents the suspect with a graphic icon. The root node(selected offender) is branched into child nodes each for the spatial and modus operandi component. The spatial component is further resolved up to two levels narrowing down the similarity of the spatial information over districts and streets. All unsolved/solved crimes that have occurred in these districts and streets are presented using purple and turquoise colour circles respectively.The temporal component of crime pattern of given suspect is used through a slider bar which is intervaled in number of weeks. The user can use this slider bar to filter the data before and after the date of crime committed by a suspect. The week zero(0W) means the actual crime committed date as mentioned in the crime report.Whenever a user moves this slider bar the data is filtered to the chosen week interval and similarity along with posterior probability is updated for the filtered data.Lastly the modus operandi node of the knowledge graph compares the behavioural component of the unsolved crime with that of suspect

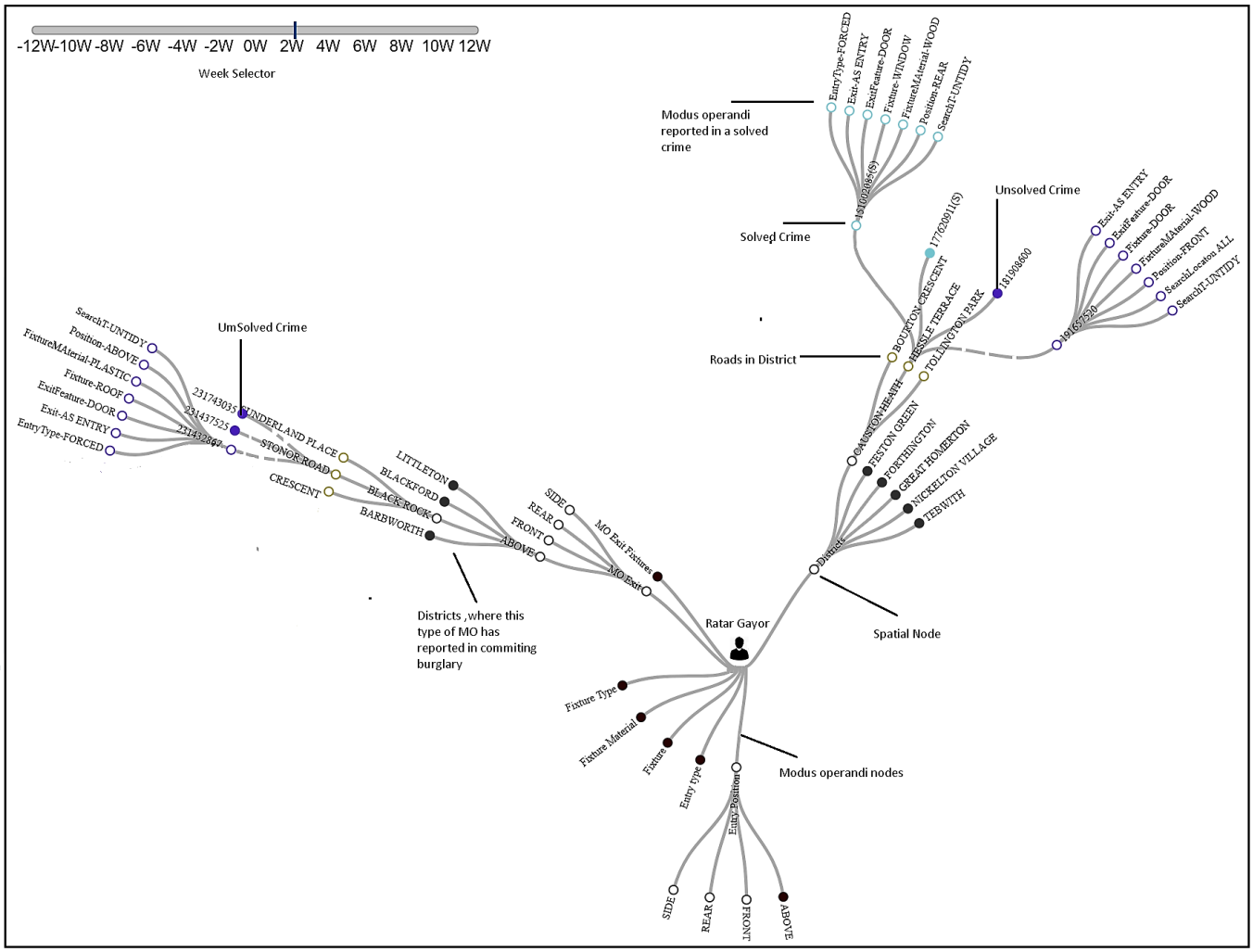

Fig. 6. Radial Knowledge graph for plausible unsolved Crime List.

pattern,starting from district and street, its underlying levels narrows down the crime pattern for each element of the modus operandi,showing the solved/unsolved crime having the same similarity.

The Figure 6 shows plausible unsolved crimes for the offender Rator Gayor. He has committed burglary crimes in the districts (Darkgray nodes) namely Causton Heath,Feston green,Forthington,Great Homerton,Nickelton Village and Tebwith.One of the expanded district node(showing hollow) shows three streets namely Bourton cresent,Hessle Terrace Tollington Park,where this offender has committed crime.On Hessle Terrace(hollow yellow cirlce), he has committed a solved crime (id,177620911(S),turquoise coloured circle), however there is an unsolved crime 191657520 reported on the same street which is shown purple colour and based on similarity, it may be thought to be committed by the same offender shown as broken lines.The modus operandi of the two crimes(solved and plausible unsolved) can also be seen and compared for similarity.The posterior probability of committing this crime based on modus operandi and spatial similarity is found to be 0.698 as shown in the Table I.These kinds of interesting associations may facilitate the cognitive thinking and reasoning strategies that analyst adopt during an investigation.

VI. Preliminary Evaluation

We conducted a preliminary evaluation with our police analyst end-users to elicit subjective feedback on our proposed associative search mechanism visualizations. The objective was to evaluate how the multi-level associations represented through clustering and knowledge graph support analysis and reasoning during a crime matching. The evaluation involved three qualitative focus groups of analysts, who participated in pairs. Each pair was from a different police organization. The prototype was demonstrated, illustrating the visualization for different association tasks to each of the focus group. Each group had 30 minutes for the demonstration and feedback. A separate group of observers recorder notes, ideas, and feedback from the end-users. We now report on the feedback as recorded by the observers, based on five questions:

- **Question 1:What is the purpose and value of the component?**

The end-users found that, the presented tool enables the discovery of relationships between crimes, offenders, etc. It helps the analysts to organize and, hence, to better understand complexity of a large amount of available data. For example, the solved and unsolved crimes can be clustered by use of varying criteria.

- **Question 2: Is the purpose and value clear to End Users?**
  The end-users understood the purpose of the prototype and it was found to add value to their current existing association elicitation processes.
- **Question 3: What do End Users like or dislike about the component?**
  The end users liked the visualization and commented that it has potential to increase the efficiency of the existing crime matching process. It allows the analyst to do something complicated with one click and avoid doing something that is currently very time consuming. The end user also liked the idea of using knowledge graph representation of the offender connections with other offenders and unsolved crimes. They found the tool is useful in observing the evolution of relationships. For example, it can be used to show the overlap between offenders and/or crimes, the evolution of the offender, or changing associations of a person (for example in the past he was associated with x offender, then became associated with y offender, etc.
- **Question 4: What features or functionalities would End Users like added, changed or removed?**
  The analyst should be able to easily import his own data into the tool. They wish to perform an effective compassion and identification of intersection within large number of available data points. They also wish to see visualization of the crime offender network, offender and offender network for a given time sliding window. They also want the colouring scheme for solved and unsolved crime should be user defined.
- **Question 5: Overall, is the End User groups assessment positive, negative or neutral?**
  The overall assessment was positive and end user enjoyed the demonstration of the tool. All the analysts found that the different aspect of the visualized associations could add value to what they are currently doing to make more effective decision making for criminal intelligence analysis.

VII. Conclusion

In this paper, we have shown how associative search can aid analysts in crime matching and presented a scheme for elicitation of multi-level association through 5WH questions over burglary dataset.Our proposed scheme elicits the associations between crimes at three level, distinguishing solved crime with unsolved crimes,identifying associations of the perpetrators with both solved and unsolved crimes and finally identifying plausible unsolved crimes list that could be associated to a known offender. We have merged temporal spatial and modus operandi characteristics in one envelope to identify these associations in a form of heterogeneous network, that could take weave associations for a given type of input node,highlighting either offender network, plausible suspect list or unsolved crime list depending upon the nature of given input node.We have project these associations through interactive clustering,visualizing networks of associations between criminal entities through graph theory.We have found that for our burglary dataset the feature vector consisting of only modus operand generated good quality clusters.

We also have evaluated this association of suspect with unsolved crime through Naive Bayes theory calculating posterior probability of an offender to be associated with an unsolved crime,using similarity of crime pattern as evidence.However we do acknowledge that it does not capture all the problems that play a role when associating a crime with an offender,though We do think it may be helpful to uncover the interesting aspects of the reasoning for crime matching. Our framework enables crime analysts to make assessment rather than recommendation and to act on the evidence only as appropriate.

The key to this research is the belief that there exists possible relationships within the various dataset used by the a police analyst, and simple visualization of these associations can be helpful for analytical reasoning during a crime matching process. Such associations can then provide the basis for activating ideas / thoughts / tentative or plausible conclusions, that could trigger new lines of inquiry.

The police analyst during a preliminary user feedback has given positive feedback indicating that this prototype have potential to improve the efficiency of the crime investigation process.In future we have plan to make the 5WH query model more intelligent. guessing itself what the user is intended to search and creating dynamic queries,extracting further deeper level of associations using video and social media data to make this tool more effective.

Acknowledgement

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.

References

[1] B. W. Wong and N. Kodagoda, "How analysts think," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 178–182, 2016. [Online]. Available: http://dx.doi.org/10.1177/1541931213601040

[2] N. Qazi, B. L. W. Wong, N. Kodagoda, and R. Adderley, "Associative search through formal concept analysis in criminal intelligence analysis," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2016, pp. 001 917–001 922.

[3] M. Alruily, A. Ayesh, and A. Al-Marghilani, "Using self organizing map to cluster arabic crime documents," in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*. IEEE, 2010, pp. 357–363.

[4] S. V. Nath, "Crime pattern detection using data mining," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 41–44.

[5] N. Le-Khac, S. Markos, and M. T. Kechadi, "A data mining-based solution for detecting suspicious money laundering cases in an investment bank," *CoRR*, vol. abs/1609.00990, 2016. [Online]. Available: http://arxiv.org/abs/1609.00990

[6] M. R. Keyvanpour, M. Javideh, and M. R. Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework," *Procedia Computer Science*, vol. 3, pp. 872 – 880, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050910005181

[7] L. Ma, Y. Chen, and H. Huang, "Ak-modes: A weighted clustering algorithm for finding similar case subsets," in *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on*. IEEE, 2010, pp. 218–223.

[8] B. J. Reich and M. D. Porter, "Partially supervised spatiotemporal clustering for burglary crime series identification," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 178, no. 2, pp. 465–480, 2015. [Online]. Available: http://dx.doi.org/10.1111/rssa.12076

[9] A. Borg, M. Boldt, N. Lavesson, U. Melander, and V. Boeva, "Detecting serial residential burglaries using clustering," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5252 – 5266, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414001110

[10] Q. Bsoul, J. Salim, and L. Q. Zakaria, "An intelligent document clustering approach to detect crime patterns," *Procedia Technology*, vol. 11, pp. 1181 – 1187, 2013, 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017313004659

[11] W. Didimo, G. Liotta, F. Montecchiani, and P. Palladino, "An advanced network visualization system for financial crime detection," in *2011 IEEE Pacific Visualization Symposium*, March 2011, pp. 203–210.

[12] J. de Zoete, M. Sjerps, D. Lagnado, and N. Fenton, "Modelling crime linkage with bayesian networks," *Science & justice*, vol. 55, no. 3, pp. 209–217, 2015.

[13] J. de Zoete, M. Sjerps, and R. Meester, "Evaluating evidence in linked crimes with multiple offenders," *Science & Justice*, vol. 57, no. 3, pp. 228 – 238, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1355030617300035

[14] M. S. Vural and M. Gök, "Criminal prediction using naive bayes theory," *Neural Computing and Applications*, vol. 28, no. 9, pp. 2581–2592, 2017.

[15] A. J. Park, H. H. Tsang, and P. L. Brantingham, "Dynalink: A framework for dynamic criminal network visualization," in *2012 European Intelligence and Security Informatics Conference*, Aug 2012, pp. 217–224.

[16] H. Isah, D. Neagu, and P. Trundle, "Bipartite network model for inferring hidden ties in crime data," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug 2015, pp. 994–1001.

[17] R. V. Hauck, H. Atabakhsh, P. Ongvasith, H. Gupta, and H. Chen, "Using coplink to analyze criminal-justice data," *Computer*, vol. 35, no. 3, pp. 30–37, Mar. 2002. [Online]. Available: http://dx.doi.org/10.1109/2.989927

[18] Y. Xiang, M. Chau, H. Atabakhsh, and H. Chen, "Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list," *Decision Support Systems*, vol. 41, no. 1, pp. 69 – 83, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923604001125

[19] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder, "Coplink: Managing law enforcement data and knowledge," *Commun. ACM*, vol. 46, no. 1, pp. 28–34, Jan. 2003. [Online]. Available: http://doi.acm.org/10.1145/602421.602441

[20] L. Ding, D. Steil, M. Hudnall, B. Dixon, R. Smith, D. Brown, and A. Parrish, "Perpsearch: An integrated crime detection system," in *Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics*, ser. ISI'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 161–163. [Online]. Available: http://dl.acm.org/citation.cfm?id=1706428.1706456

[21] D. E. Brown, "The regional crime analysis program (recap): a framework for mining data to catch criminals," in *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, vol. 3. IEEE, 1998, pp. 2848–2853.

[22] J. Schroeder, J. Xu, H. Chen, and M. Chau, "Automated criminal link analysis based on domain knowledge: Research articles," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 6, pp. 842–855, Apr. 2007. [Online]. Available: http://dx.doi.org/10.1002/asi.v58:6

[23] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, no. 2, pp. 118–132, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1145/1466620.1466622

[24] U. Thongsatapornwatana and C. Chuenmanus, "Suspect vehicle detection using vehicle reputation with association analysis concept," in *2014 IIAI 3rd International Conference on Advanced Applied Informatics*, Aug 2014, pp. 436–440.

[25] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996. [Online]. Available: http://doi.acm.org/10.1145/240455.240464

[26] J.-H. Wang, B. T. Lin, C.-C. Shieh, and P. S. Deng, *Criminal Record Matching Based on the Vector Space Model*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 386–386. [Online]. Available:

[27] J.-H. Wang and C.-L. Lin, "An association model based on modus operandi mining for implicit crime link construction," in *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 548–550. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2011.34

[28] Y. Sun, "Mining heterogeneous information networks," 2013.