Focus Article

Data mining and crime analysis

Giles Oatley[1]* and Brian Ewart[2]

An essential component of criminal investigation involves the interrogation of large databases of information held by police and other criminal justice agencies. Data mining and decision support systems have an important role to play in assisting human inference in this forensic domain that creates one of the most challenging decision-making environments. Technologies range widely and include social network analysis, geographical information systems, and data mining technologies for clustering crimes, finding links between crime and profiling offenders, identifying criminal networks, matching crimes, generating suspects, and predicting criminal activity. This paper does not intend to cover the gamut of techniques available to the investigator of crime as this has been presented elsewhere (Oatley GC, Ewart BW, Zeleznikow J. Decision support systems for police: lessons from the application of data mining techniques to 'soft' forensic evidence. *Artif Intell Law* 2006, 14:35–100). Rather, the objective is to highlight issues of implementation and interpretation of the techniques available to the crime analyst. To this end, the authors draw from their experiences of working with real-world crime databases (Oatley GC, Belem B, Fernandes K, Hoggarth E, Holland B, Lewis C, Meier P, Morgan K, Santhanam J, et al. The gang gun-crime problem—solutions from social network theory, epidemiology, cellular automata, Bayesian networks and spatial statistics. Accepted: book chapter for IEEE publication *Computational Forensics;* 2008; Oatley GC, McGarry K, Ewart BW. Offender network metrics. WSEAS Trans Inf Sci Appl 2006, 3:2440–2448; Oatley GC, Ewart BW. Crimes analysis software: pins in maps, clustering and Bayes net prediction. Expert Syst Appl 2003, 25:569–588), involving gun and gang crime, fraud, terrorism, burglary, and retail crime. © 2011 John Wiley & Sons, Inc. *WIREs Data Mining Knowl Discov* 2011 1 147–153 DOI: 10.1002/widm.6

THE CRIME INVESTIGATION CONTEXT

We are interested in supporting the investigation of crime with the tools available from computer science. For example, Canter[1] notes that a fundamental part of crime investigation is the generation of suspects. An important development would be the ability to link crimes, past and present, likely to have been committed by a particular offender. It is the essence of 'operational crime analysis,' according to Merry,[2] and involves the 'analysis of every crime with every other crime and with criminals to identify links that are not evident from routine police enquiries.'[2, p. 302] It has been shown that prolific offenders selected by police intelligence officers were seldom responsible for most of the crime in a specific area.[3] In this respect, human judgment and information processing are limited. The aim of this paper is not only to illustrate the role that information technology can play but also to highlight some important issues when implemented within the context of criminal investigation.

IMPLEMENTATION AND INTERPRETATION—EXAMPLES FROM REAL-WORLD APPLICATIONS

The first author was involved in a UK research study (EPSRC reference: EP/D078725/1) called 'modeling analysis of gun crime NETworks' (MAGNET) to tackle the increasing problem of gang-related gun crime. The study employed epidemiology, cellular automata, Bayesian networks, geographical information systems (GIS) and spatial statistics, and social network analysis (SNA). For instance, the GIS approach involved an analysis of the relationship between offender residence, victim residence, and place of offence. This was very different to the statistically

*Correspondence to: goatley@uwic.ac.uk

[1] Department of Information Systems, University of Wales Institute, Cardiff, UK

[2] Psychology Division, University of Sunderland, Sunderland, UK

DOI: 10.1002/widm.6

SIDEBAR: Core crime analysis technology: Social networks and link analysis

The use of offender network technologies has now become a routine part of crime prevention operations. The most well-known technology is social network analysis (SNA), supported by excellent (and freely available) software such as Pajek or R's SNA library (see below for links). Networks of offenders are developed on the basis of various links between offenders, such as co-conviction or co-arrest data. The final network can reveal to us much more than the original data, for instance, how offenders enter networks—how/what is their first contact; is there a periphery/core to the network; are there *holes* in the network, i.e., indicators of an offender not currently known; which crimes go together with which network contacts; what is the state of the network at the times of a crime; how behavior is transmitted through the network; and the effects on network dynamics by educating/reforming/removing certain nodes. The data mining community may be more familiar with the related subjects of link analysis, association rules, graph-based data mining, and relational data mining, which also have an important part to play in the analysis of offender networks.

Useful links:
http://pajek.imfm.si/doku.php?id=pajek
http://cran.r-project.org/web/packages/sna/index.html

SIDEBAR: Core crime analysis technology: Geographical information systems and spatial statistics

The spatial modeling of incidences of crime has its origins in the 'pins in maps' being translated to points (and polygons) in geographical information systems (GIS). This area is perhaps best known for the aggregated data representation of 'hotspots'; however, there are many more advanced technologies involving spatio-temporal and forecasting technologies. Excellent (and freely available) software includes CrimeStat and GeoBUGS.

Useful links:
http://www.icpsr.umich.ed/CRIMESTAT/
http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml

informed epidemiological model looking at the 'transmission' of gun using behavior, and similarly different to the network approach.

The objectives of using SNA were to model the dynamics of the gangs, their development and fragmentation into new gangs, and the roles that gun users occupy with the gang structure. In particular, we wanted to determine if any link types or network metrics, for instance, the frequency and type of person-to-person contact, could be used to reduce gun crime. We were specifically interested in using networks to target strategic interventions.

Figures 1 and 2 show networks of the rival gang members of which we studied (gangs A and D vs. gangs B and C).

SNA has been applied to a wide range of crime types, including terrorism,[4] cyber crime hate groups,[5] and sexual crime.[6] Supported by a Nuffield Foundation grant to the second author, SNA was applied to a database of over 30,000 known retail crime offenders. The aims were to identify unknown gangs specializing in retail crime and confirm the existence of gangs which the police and other retail security organizations had defined as being organized teams on the basis of intelligence and some instances of 'in store' detections. SNA is therefore implemented as an exploratory and confirmatory methodology.

Prior to broadening the discussion, these forms of criminality will be used to illustrate some initial comments concerning the implementation of technologies and the interpretation of their findings.

ISSUES OF IMPLEMENTATION

When SNA is used as an exploratory methodology, the completeness of the database and the variables need to be considered. Applied to a police database of detected crime, there is an assumption that all offences have an equal probability of being present. In the context of retail crime, this would be problematic for crimes such as using false credit cards and theft by concealment. The former network might be more complete as the likelihood of becoming known to the retailer, reported, investigated, and detected is greater. Links based on intelligence data such as sightings of suspects with other known offenders are less certain than those based on co-arrest. Obviously, co-conviction is the strongest link, but the completeness of any network is dependant on the level of both undetected and unreported offences. This increases the probability of absent links that become false negatives in any network produced. The criminological literature has an important role to play in that it informs the

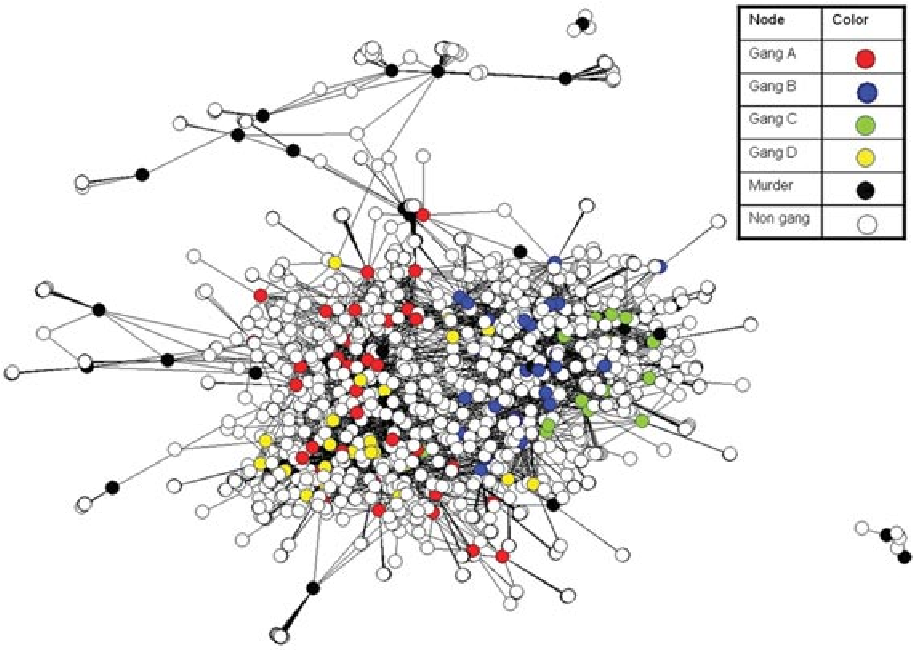

FIGURE 1 | Gangs A, B, C, D labeled, showing affinities between gangs A and D (red and yellow) and gangs B and C (blue and green).

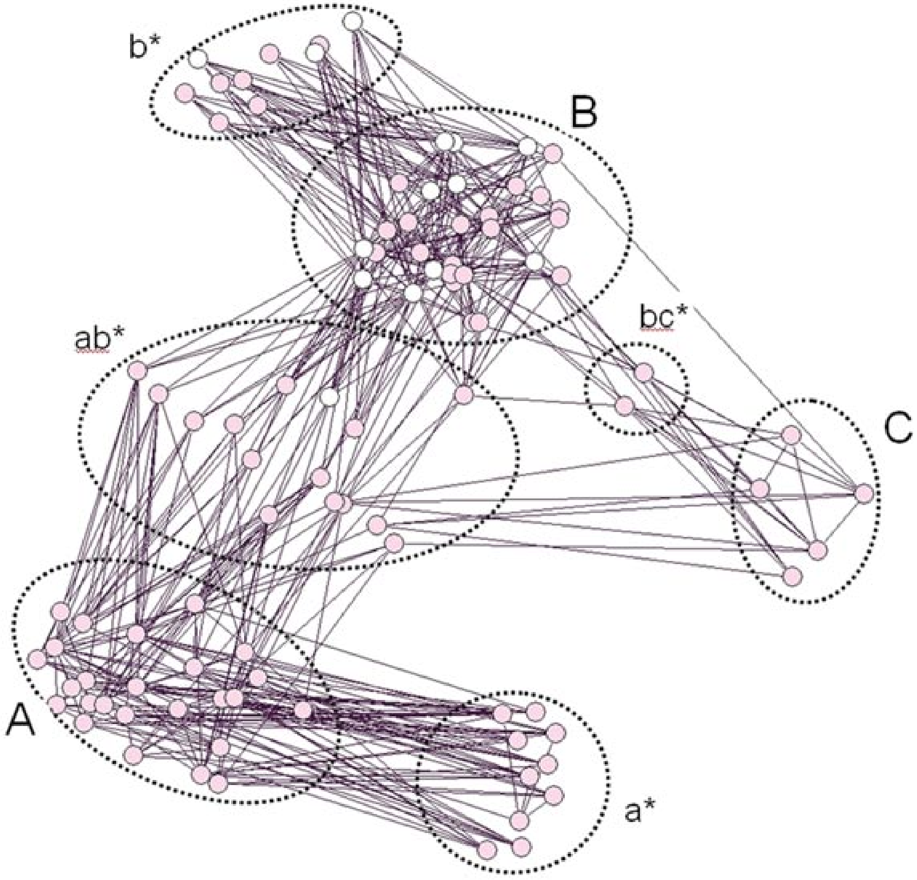

FIGURE 2 | Cores ($k = 5,6$) extracted showing gangs A and B and emergence of gang C. The figure also illustrates the large amount of nongang members who are associated with individual gangs (a*, b*) or who are intermediaries (ab*, bc*).

systems designer and end-user about the probability of crimes being reported, and the temporal and social factors that influence reporting and nonreporting.[7] This is essential to the evaluation of sociograms and their associated statistics, and provides an empirical foundation for the probabilistic linking available in SNA technologies.[8]

Using an SNA to discover networks involves identifying substructures that are locally dense, but separated to some degree from the rest of the graph. One technique is the *k-cores link reduction* method (see Figure 2). A *k*-core is a maximal group of actors, all of whom are connected to some number '*k*' of other members of the group. The problem is deciding the value of *k*. Robbery teams varied in size from three to seven people,[9] and organized drug networks between six and 23.[10] One definition of an organized crime is a collaboration of more than two people.[11] Such a small number applied to a large database of offenders would produce a meaningless number of networks, and some sizable and substantively important networks would not be evident. Too high a number, especially with a weak linkage variable, would generate a few networks with so many nodes that their operational value would be limited. So, prescribing a value of *k* when using SNA inductively is not without difficulties. This problem would apply to any clustering algorithm.

Taking a wider perspective, the representation of any crime in a database will affect the implementation of technologies. The lack of consistent representations and ways of classifying crime is a serious problem, especially for areas such as business crime[12] in which the value of technologies is increasingly recognized. For instance, the home office counting rules (HOCR)—by which crime is classified and therefore recorded or counted—will shortly be changing for the crime type of fraud. That is, the definition of what constitutes fraud is changing. There exist various fraud typologies: the previous HOCR,[13] the National Fraud Reporting Centre's emerging representation of fraud which is informing the new HOCR, Levi et al.'s[14] representation, and subsets of fraud covered in smaller typologies (e.g., FF-Poirrot[15]).

The ultimate implementation objective would be to prescribe technologies for particular investigative functions. As the examples above illustrate, research into just two forms of criminality (gun and retail gang theft) involves techniques ranging from social sciences (SNA) to physics (statistical complex graph literature). In the area of fraud, there are several excellent survey papers[16–19] providing a 'menu' of mutually exclusive techniques. Even then, a prerequisite of prescription is the ability to define the investigative problem. Provost (a prologue in[16]) comments in relation to fraud that . . .

> It would be useful to have a precise definition of a class (or of several classes) of fraud detection problems, which takes into account the variety of characteristics that make statistical fraud detection difficult. If such a characterisation exists already in statistics, the machine learning and data mining communities would benefit from its introduction.

Sadly, the characterization does not exist. Provost continues:

> Also, to succeed at detecting fraud, different sorts of modelling techniques must be composed, for example, temporal patterns may become features for a system for estimating class membership probabilities, and estimators of class membership probability could be used in temporal evidence gathering.

This structuring of the emerging body of knowledge needs to be developed. Provost emphasizes the need for progress to be measured against solution of the larger problem (of fraud in general), and this is indeed necessary for all crime types.

This is not to say that there has been no progress on formalizing criminal investigation technology. Since 1986, UK Police Forces employed the original home office large major enquiry system (HOLMES) in all major incidents, including serial murders, multimillion pound fraud cases, and major disasters. Its successor HOLMES2 builds upon operational practice. A code of practice that is the UK's National Intelligence Model[20] describes nine 'analytical techniques,' which include crime pattern analysis and network analysis. However, there is still very much a culture of reliance on products such as i2 and other 'intelligent network products' (e.g., Detica's NetReveal; Detica, Surrey, UK), geographical techniques including spatial clustering ('hotspot analysis'), all of which can lead to an impoverishment of critical thinking.

ISSUES OF INTERPRETATION

Utilizing our real-world example, we see the value of SNA is that it enables the criminal investigator to explore patterns of association (see Figure 1). However, Figure 2 illustrates that the large number of 'nongang' members may also feature in the network. This is a function of the linking variables employed, as strong links (such as co-conviction, co-arrest, related by family) were not distinguished from weaker ones (such as 'seen together').

Even when one variable is used for linking, interpretation must be cautious. The terms 'team,' 'group,' and 'network' are often used interchangeably in studies of organized crime and criminal groups.[10,21] If phone records are used, then network size and membership will vary much more than if co-conviction is used. The former may include the team who actually commit the offences, those used to 'fence,' insiders providing information, criminal associates, and innocents. Clusters linked by co-conviction only are likely to be the 'frontline team.' However, if a person's individual offending is more prolific, policing them as a gang member may be less appropriate. Data mining technologies could generate an individual/gang ratio of offences or contacts for each node, thus providing an important test of the validity of any network.

Implementing SNA as a confirmatory methodology is valuable. In the retail study above, gangs were sometimes defined by the police or retail security personnel on the basis of some incidents of co-arrest. A self-fulfilling prophecy is possible where offenders are subsequently perceived by investigators as a team even though they are more prolific as individuals. Again, technology can help interrogate a database to test the validity of these perceptions. In relation to SNA, Klerks[22] summarizes the challenge as the 'Next generation social network analysis must focus much more intensely on the content of the contacts, on the social context, and on the interpretation of such information.'

Taking this point more generally, technologies have the ability to generate new information based on aggregating crime data, but the investigator still has to be interpretative on this information. The problem is that human inference comprises some cognitive biases that may contaminate the process, and ultimately, the decision reached. For example, we exhibit a 'confirmatory bias'[23] in that we focus upon data that conform to the initial ideas formed (e.g., about who is the suspect) and so fail to test them by seeking evidence that disconfirms our notions. For example, if two offences at different locations involve a specific type of firearm, they may be readily attributed to the same offender. However, before this attribution is useful in linking the crimes to a suspect, the overall frequency, or base rate, of this feature across all offenders and firearms offences needs to be established, and here, technology has an essential role to play.

Decision-making needs to be informed by specialist knowledge from other domains. For example, criminal patterns such as 'near repeats'[24] and psychological concepts such as criminal range in relation to home location[25] are essential when interpreting the output of crime visualization technologies. Even then, the analyst must take account of issues. For example, the near repeats occur only if temporal and spatial data are sufficiently disaggregated[26] and algorithms prioritizing location and range criminal are not equally effective.[27] More generally, Snook et al.[28] note that constraint is needed by 'those intent on building confidence in computerised geographic profiling systems in the absence of strong empirical evidence in support of their use.'

Finally, the specialism of intelligence analysis has a potential contribution to make to this issue. The interpretative challenge of the intelligence analyst is that it involves different sources of information, such as human intelligence (HUMINT), imagery intelligence (IMINT), signal intelligence (SIGINT), measurement and signature intelligence (MASINT), technical intelligence (TECHINT), and open source intelligence (OSINT). The authors have recently developed a training course for intelligence analysts that develops these ideas of cognitive biases in decision-making and inference, based around literature such as[29] and[30]. The intelligence analyst has better practice in that approaches such as 'analysis of competing hypotheses' with supporting software[31] are employed to reason optimally with limited evidence. Surprisingly, this is not a common practice in crime analysis despite the obvious similarities between the disciplines.

CONCLUSION: EMERGING FRAMES OF REFERENCE

Despite the stated difficulties, we can see clear evidence that the area of crime data mining is developing. There *are* emerging crime typologies and methodology papers, and the use of statistical, data mining, mathematical technologies is more informed by forensic psychology and criminology.

Data standardization (collection and representation) is an obvious step required for collaborative solutions, and in the UK, since April 1, 2002, the National Crime Recording Standard was adopted, with a main aim to promote greater consistency between constabularies and the recording of crime. Overarching frameworks include[6] the global justice XML data model for representing crimes from the United States[32] and the National Policing Improvement Agency's 'corporate data model' from the United Kingdom.[33]

There are methodological frameworks from forensic psychology and criminology, for instance, 'become a problem solving *crime* analyst,'[34] and from forensic science, we have Ribaux and Margot's[35–37] attempts to develop a representational and inferential methodology—'organizing' data (adding new cases to a case-based reasoning system), 'scrutinizing' (inferencing over the case base) and 'exploiting' (developing police operations).

While we certainly agree with the development of such frameworks, and also Ribaux and Margot's development of systems based upon human inference strategies, we also emphasize the need to pay serious attention to the cognitive shortcomings of the crime bias introduced in criminal investigations and the importance of how an interdisciplinary approach can maximize the use and potential of these technologies and approaches.

REFERENCES

1. Canter D. Offender profiling and criminal differentiation. *Legal Criminol Psychol* 2000, 5:23–46.
2. Merry S. Crime analysis: principles for analysing everyday serial crime. In: Canter D, Alison L, eds. *Profiling Property Crimes*. Ashgate; 2000, 297–318. ISBN: 1-84014-785-7.
3. Townsley M, Pease K. How efficiently can we target prolific offenders? *Int J Police Sci Manage* 2002, 4:323–331.
4. Liu Q, Tang C, Qiao S, Liu Q, Wen F. Mining the core member of terrorist crime group based on social network analysis. Intelligence and security informatics. *Lect Notes Comput Sci* 2007, 4430/2007:311–313. ISBN: 978-3-540–71548-1.
5. Chau M, Xu J. Using web mining and social network analysis to study the emergence of cyber communities in blogs. In: Chen H, Reid E, Sinai J, Silke A, Ganor B, eds. *Terrorism Informatics, Integrated Series in Information Systems*, New York: Springer; 2008, 473–494. ISBN: 978–0-387–71612-1.
6. Chen H, Chung W, Xu JJ, Wang G, Qin Y, Cha M. *Crime Data Mining: A General Framework and Some Examples*. *IEEE Computer* 2004, 37:50–56.
7. MacDonald Z. Official crime statistics: their use and interpretation. *Economic J* 2002, 112:F85–F106.
8. Hanneman RA, Riddle M. *Introduction to Social Network Methods* (Electronic version). Riverside, CA: University of California; 2005.
9. McCluskey K, Wardle S. The social psychology of robbery. In: Canter D, Allison L,eds. *The Social Psychology of Crime. Groups, Teams and Networks*. Aldershot: Ashgate; 2000, 247–285.
10. Stovin G, Davies C. Beyond the network: a crime science approach to organized crime. *Policing* 2008, 2:497–505.
11. Van Der Heijden T. Measuring organized crime in Western Europe. In: Pagon M, ed. *Policing in Central and Eastern Europe: Comparing First Hand Knowledge with Experience from the West* (Electronic version). Slovenia: College of Police and Security Studies; 1996. Available at: http://www.ncjrs.gov/policine/mea313.htm. (Accessed August 12, 2010).
12. Ewart BW, Tate A. Policing retail crime: from minor offending to organised criminal networks. In: Froeling KT, ed. Criminology Research Focus. New York: Nova Science; 2007, 33–67.
13. HOCR. 2010. *Counting rules for recorded crime*. Available at: http://www.homeoffice.gov.uk/rds/countrules.html. (Accessed August 12, 2010).
14. Levi M, Burrows J, Fleming MH, Hopkins M, with the assistance of Kent Matthews. 2007. *The Nature, Extent and Economic Impact of Fraud in the UK*. Report for the Association of Chief Police Officers.' Economic Crime Portfolio; London: Association of Chief Police Officers. February 2007.
15. Kingston J, Schafer B, Vandenberghe W. Towards a financial fraud ontology: a legal modelling approach. *Artif Intell Law* 2004, 12:419–446.
16. Bolton RJ, Hand DJ. Statistical fraud detection: a review. *Statist Sci* 2002, 17:235–255.
17. Phua C, Lee V, Smith-Miles K, Gayler R. A comprehensive survey of data mining-based fraud detection research. Clayton, VIC: Clayton School of Information Technology, Monash University; 2005. Available at: http://clifton.phua.googlepages.com/. (Accessed August 12, 2010).
18. Kou Y, Lu CT, Sirwongwattana S, Huang YP. Survey of fraud detection techniques. In: *Proceedings of the 2004 International Conference on Networking, Sensing, and Control*. Taipei; 2004, 749–754.
19. Weatherford M. Mining for fraud. *IEEE Intell Syst* 2002, 17:4–6.
20. NIM. 2010. National Intelligence Model—analytical techniques and products. Available at: http://police.homeoffice.gov.uk/publications/operational-policing/nim-analytical/. (Accessed August 12, 2010).
21. Coles N. It's not what you know—it's who you know that counts. *Br J Criminol* 2001, 41:580–594.
22. Klerks P. The network paradigm applied to criminal organisations: theoretical nitpicking or a relevant

doctrine for investigators? Recent developments in the Netherlands. *Connections* 2001, 24:5365

23. Oskamp S. Overconfidence in case study judgments. *J Consult Psychol* 1965, 29:261–265.
24. Johnson SD, Bowers KJ. The stability of space-time clusters of burglary. *Br J Criminol* 2004, 44:55–65.
25. Barker M. The criminal range of small-town burglars. In: Canter D, Alison L, eds. *Profiling Property Crimes*. London: Ashgate; 2000, 57–75.
26. Ewart BW, Oatley GC. Patterns of victimisation and revictimisation: examining the communicability of burglary risk. Paper presented at the British Society of Criminology Annual Conference, Glasgow, 2006.
27. Canter D, Hammond L. Prioritizing burglars: comparing the effectiveness of geographical profiling methods. *Police Prac Res* 2007, 8:371–384.
28. Snook B, Taylor PJ, Bennell C. False confidence in computerised gepgraphical profiling [shortcuts to geographic profiling success: a reply to Rossmo (2005)]. *Appl Cognit Psychol* 2005, 19:655–661.
29. Lipton P. Testing hypotheses prediction and prejudice. *Science* 2005, 207:219–221.
30. Lipton P. *Précis of Inference to the Best Explanation* 2nd ed. Philosophy and Phenomenological Research 2007, 74:421–423.
31. PARC_ACH. 2010. PARC ACH 2.0.3 software. 'Analysis of Competing Hypotheses.' Available at: http://www2.parc.com/istl/projects/ach/ach.html. (Accessed August 12, 2010).
32. GJXDA. 2010. *Global Justice XML Data Model*. Available at: http://it.ojp.gov/jxdm/. (Accessed August 12, 2010).
33. CDM. *NPIA Corporate Data Model*2007. Available at: http://cordm.npia.police.uk/. (Accessed August 12, 2010).
34. Clarke RV, Eck J. *Become a Problem Solving Crime Analyst—In 55 Small Steps*. London: Jill Dando Institute of Crime Science, University College London; 2003. Available at: http://www.jdi.ucl.ac.uk/downloads/publications/other_publications/55steps/Prelims.pdf. (Accessed August 12, 2010).
35. Ribaux O, Margot P. Inference structures for crime analysis and intelligence: the example of burglary using forensic science data. *Forensic Sci Int* 1999, 100:193–210.
36. Ribaux O, Margot P. The analysis of serial crime through the use of different sources of data. *Probl Forensic Sci* 2001, 67:93–100
37. Ribaux O, Margot P. Case based reasoning in criminal intelligence using forensic case data. *Sci Justice* 2003, 43:135–143.