Prioritizing of Offenders in Networks

GILES OATLEY, KEN McGARRY & BRIAN EWART
School of Computing and Technology
University of Sunderland
St Peters Campus, Sunderland SR6 0DD
United Kingdom

Abstract: - This paper reports work that builds upon several years of experimentation using forensic psychology guided exploratory techniques from artificial intelligence, statistics and spatial statistics. Our central aim is the development of decision support systems for crime prevention and detection, and this paper presents a novel algorithm that incorporates geographical information, frequency and recency of criminal activity directly into the 'betweenness' metric of social network analysis. The algorithm is *ad hoc*, and design decisions are presented, alongside the operational use by police forces of such an algorithm, namely as a means for *prioritizing* of offenders in large networks. The data presented is from the crime of burglary from dwelling houses.

Key-Words: - Offender Networks, Spatial Statistics, Data Mining

1 Introduction

Technologies from spatial statistics, geographical information systems and link analysis (for instance from social network analysis) are routinely used by police forces in crime prevention and detection. While previous work by the author has presented an extensive survey of technologies useful for this area [1], it is clear that the current technologies used by a police forces' crimes analyst are either based on the geographic notion of crimes or finding links between offenders or crimes. This is evidenced by the available commercial software which targets these two 'disciplines', and part justified by the recognized need of expertise in representing domain knowledge and data mining technologies to create systems that can classify and predict.

Recent work by the author [2] investigated the addition of spatial, temporal and frequency data to offender networks, although only in a 'graphical' way, and this paper presents a computational means of combining this information.

The authors work in this area [3,4,5], in collaboration with West Midlands Police (WMP), is with the high volume crime of Burglary from Dwelling Houses (BDH). The focus of our research has been to develop techniques and a sound methodological framework for crime matching and prediction, and integration of evidence for such purposes. Innovative work includes the use of Bayesian belief networks for prediction of crimes, integrating as many evidence sources as determinable – concerning the offender, victim, time of crime, specific location of crime, general area of crime, and behavioural (modus operandi) features. This paper continues in this vein with the aim of combination of evidence, primarily from spatial statistics and the social network analysis metric of *betweenness*.

A common pitfall is the use of sophisticated social sciences methods without an understanding of the phenomena or meaning behind the links. 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 [6]. To this end Lyons and Tseytin [7] proposed an a priori expression of facts that may be used to infer *phenomena* from links, utilising situation calculus. We show how adding in a geographical component (not present in other approaches) and then a temporal and frequency component can add to the interpretation of the network and its key players

1.1 Link Discovery and Link Meaning

Within data mining, there has recently been great interest in the developing fields of g*raph-based data mining* [8] and *link mining* [9] - also known as link discovery or link analysis.

There has been a rapid increase in commercially available products that claim to perform link mining, and link mining encompasses a range of tasks including descriptive and predictive modelling [9]. In the study presented the links are already known,

being based on the co-defendant relationship, however the technique brings to bear additional information to better reveal the *nature of the link*.

The method used in this paper that operates over graphs (networks) created from links is the social network analysis (SNA) technology point centrality metric termed *betweenness*. This metric measures the extent to which a particular point lies 'between' the various other points in the graph: a point of relatively low degree may play an important 'intermediary' role and so be very central to the network. The betweenness of a point measures the extent to which an agent can play the part of a 'broker' or 'gatekeeper' with a potential for control over others.

Earlier work generated large networks of linked offenders, however it was not clear whether the link could be considered strong or weak, recent or old, and offender pairs committing many crimes together in the recent past would appear the same as those offenders whose activity together was a long time passed through only a single crime.

1.2 Structure of this paper

Section two presents the data used identifying the different forms of evidence that can be combined. The algorithm is presented in section three with section four containing initial results and a discussion of additional information not yet included. The paper concludes with section five.

2 Problem definition

2.1 Burglary Arrest Data Networks

The networks and geographical outputs presented are derived from 342 offenders who committed 1121 crimes, representing the time period 1997-2001. The network links are based upon whom the offender was arrested with for a particular crime and the geographical location of that offence. This represents a significant departure from previous methodologies in that links are on the basis of an established (albeit not proven in court) co-defendant relationship.

One reservation concerning the outputs presented is that links are on detected rather than unsolved crimes, which means the extent of the network and its range may an underestimate. However, the point is to illustrate the potential of such an approach and even these outputs provide important policing information on the offending range of the respective clusters. If date of the crime were added to this methodology, this temporal information (again routinely collected by police) would allow more substantive questions about the characteristics of networks to be explored.

2.2 Types of information

Figure 1 presents the information used in this approach. The square boxes represent offenders and the links are co-defendant. The number in the top-left of each box denote the ID of the offender (*URN*: 'unique reference number'). The bracketed information on the links are the dates of the crimes (in days) that form the co-defendant link, baselined from the beginning of the study period, i.e. larger values are more recent. Top-right is presented the amount of crimes the offender has committed. Bottom-left is the *betweenness* value of this node in the network. Each offender has a graphic displaying the geographic range of their activity.

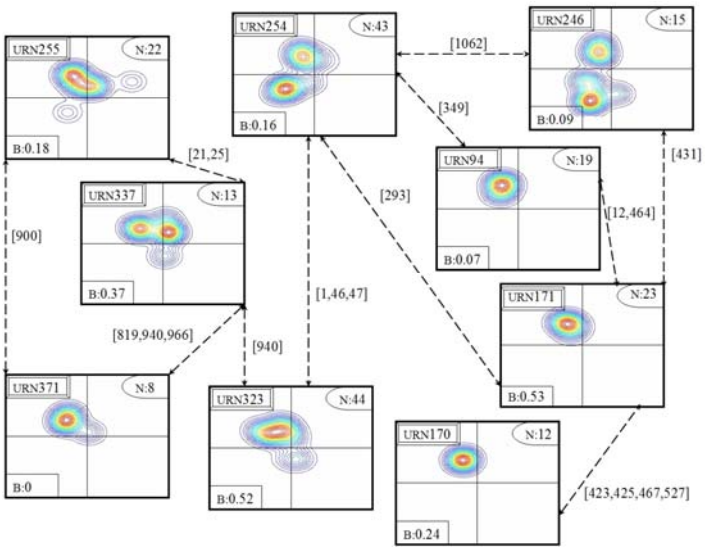

Figure 1. Network of offenders.

For example, offender *URN170* has committed 12 crimes in the 'northwest' (see spatial graphic), has a 'betweenness' value of 0.24, and has committed four crimes with offender *URN171* - 423, 425, 467 and 527 days after the start of the study period.

There is a lot of information displayed, and while it is very useful, it would be impossible to 'digest' if there were hundreds of offenders in the network (a typical situation), instead of this simple network of nine offenders. This paper now discusses initial thoughts on how to combine all of the information, to present on the same network structure a single value representing these diverse values of network position, offender range, amount

of crimes of an offender, and the strength of the links between offenders.

3 Problem Solution

The *betweenness* centrality measure is presented in equation 1.

$$C_B(n_i) = \frac{\sum_{j<k, i\neq j, i\neq k} w_i \cdot g_{jk}(n_i) \Big/ g_{jk}}{((n-2)\cdot(n-1))\Big/2} \quad (1)$$

$C_B(n_i)$ is the *weighted* betweenness centrality of node *i*, where: $g_{jk}(n_i)$ represents the number of geodesics linking *j* and *k* that contains *i* in between; and, g_{jk} represents the total number of geodesics linking *j* and *k*. This index is standardized by the value *((n-1)(n-2))/2* as the maximum value it can take is when node *i* is between all pairs of nodes, which is quantified as the total number of pairs in the network not including actor *i*.

It is the weighting factor, w_i, however, that is of interest to this paper. This is best illustrated initially by means of a diagram. The equations will then follow.

Consider the case from the network in figure 1 where we are calculating $C_B(n_{URN323})$ for offender *URN323* with *j* as *URN337* and *k* as *URN246*. One of several paths between *j* and *k* which pass through *i* – to calculate $g_{jk}(n_i)$ – is illustrated in figure 2.

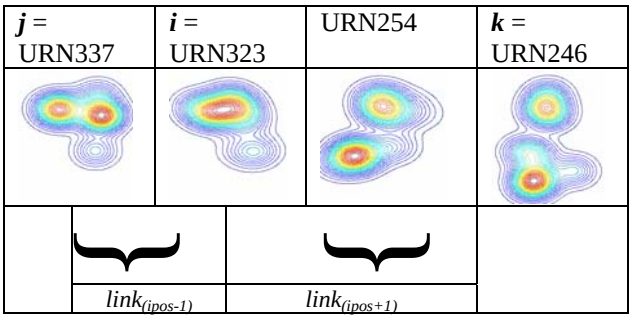

Figure 2. Illustration of weighting factor

The weighting factor for a node *i* considers the elements to which this node is connected. In the case of the example, *URN323* is connected to *URN337* and *URN254*. Hence equation 2, where in this example; *ipos-1=URN337*, and, *ipos+1=URN254*.

$$w_i = link_{(ipos-1)} + link_{(ipos+1)} \quad (2)$$

The value of *link*, shown in equation 3, is a combination of:

- *CR*: Offender range (geographical difference between linked offenders)
- *CD*: 'Danger' of offender (amount of crimes committed)
- *LS*: Strength of links (product of amount of co-crimes and recency of last crime)

$$link = CR \cdot CD \cdot LS \quad (3)$$

CR is the difference in geographical range to which the linked offender contributes. Consider figure 1, where the difference in criminal range is small between *URN337* and *URN323*, however is larger between URN323 and *URN254*, by virtue mainly because of *URN254's* activity in the 'southeast'. *CR* uses the *negative exponential* (or peaked) distribution, which falls off very rapidly with distance up to the circumscribed radius. Its functional form outside a specified radius, is shown by equation 4.

$$g(x_j) = 0 \quad (4)$$

And within a specified radius shown by equation 5, where d_{ij} is the distance between an incident location and any reference point in the region.

$$g(x_j) = \sum A \cdot e^{-K \cdot d_{ij}} \quad (5)$$

The absolute difference is calculated between the 2-dimensional matrices for each offender pair, and summed across columns and rows to a single value of dissimilarity.

The value *CD* is simply how many crimes the 'linked-to' offender has committed.

The value *LS* considers how much strength to give to the link (geographic range and prolificness of co-offending), and considers the recency of the last co-crime, and how many co-crimes constitute the link between these offenders.

There are clearly many ways that the information sources can be combined and calculated. For instance several versions of geographic range were created before the current one was chosen. It is also clear that perhaps in

some cases the number of co-crime links should outweigh the recency of the last co-crime, and so on. Various weightings can be applied to *CR*, *CD* and *LS*, however for these experiments all factors are set as equal.

4 Discussion

4.1 Initial results

The following results were obtained using the Matlab programming environment, with the 'betweenness' algorithm developed upon elements of Kevin Murphy's graph theory toolbox[1].

Rank	B		SB		WSB	
1	URN171	0.52	URN323	0.79	URN323	6.36
2	URN323	0.16	URN171	0.26	URN254	4.78
3	URN337	0.53	URN337	0.73	URN171	1.92
4	URN156	0.37	URN156	0.53	URN337	1.46
5	URN170	0.31	URN170	0.45	URN156	1.18
6	URN255	0.24	URN254	0.30	URN170	0.92
7	URN254	0.18	URN255	0.22	URN255	0.45
8	URN246	0.07	URN246	0.10	URN94	0.35
9	URN94	0.09	URN94	0.12	URN246	0.15
10	URN93	0.01	URN216	0.01	URN216	0.01
11	URN216	0.01	URN93	0.01	URN93	0.00

Table 1. Comparison of 'betweenness' calculations. B represents standard betweenness, SB incorporates the spatial value only, and WSB is weighted with offender range, 'danger' of offender, and strength of links.

The rank position of each offender is presented according to three algorithms, namely: standard 'betweenness' (B); 'betweenness' with spatial information (offender range - *CR* from equation 3) (SB); and, all of the factors presented in equation 3 (WSB).

Points of interest are the change of relative positions of *URN171* and *URN323* from B to SB and WSB. *URN171* drops in rank because the offenders they are connected are in the main operating in a similar geographical location. The exception is their link with *URN246*, however the link is not strong, by virtue of only one crime. *URN323*'s larger WSB value is strengthened through their strong link with the prolific offender *URN254* (and vice versa) by three co-crimes.

Other movements in rank are the change in positions of *URN246* and *URN94*, *URN93* and *URN216* etc. However, while it is interesting to see how the different algorithms change the positioning, the approach has been designed to illustrate that these diverse forms of information (network, spatial, temporal and other) *can* be combined, and to point out that over a large network the values of the offender nodes will be more *meaningful*, in an operational sense.

4.2 Additional information

Future work is to investigate the appropriate way of incorporating the following forms of additional information: modus operandi; repeat victimization; and, spatio-temporal.

The transmission of modus operandi across a network is an interesting and complex issue. Consider a new crime prevention technique that is slowly but surely compromised by criminal ingenuity, and passed through the network of offenders. This is a complex problem, and remains to be decided how best to calculate given the algorithm 'framework' presented here. It is possible that earlier work by the authors using a naïve Bayes approach with optimized modus operandi features could be useful. Indeed Adderley [10] uses clustering on modus operandi features to assign unsolved crimes to known criminals.

Repeat victimization, where a premise is burgled more than once, was the focus of earlier work by the authors [3-5] and this can be included.

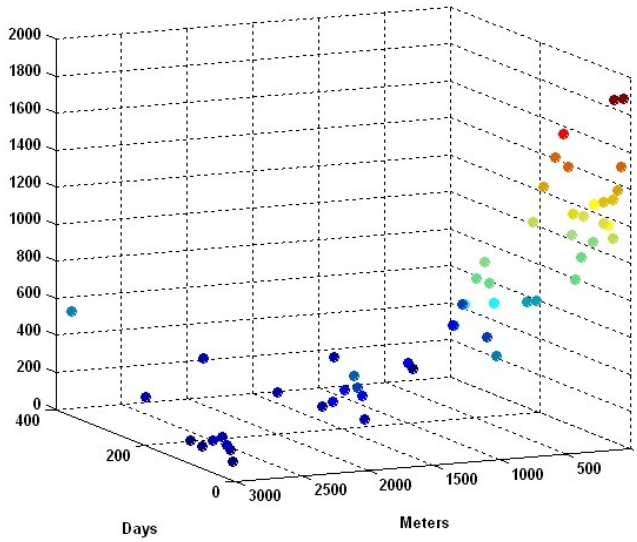

Figure 3. Knox index for spatio-temporal data.

While the reported algorithm includes both spatial and temporal information, there are known metrics from spatial statistics that combine these, for instance the Knox and Mantel indices. Figure 3 presents the same time period for all burglaries, and

[1] http://www.cs.ubc.ca/~murphyk/Software/

is crime centric instead of offender (-network) centric. Crimes are clustered most significantly within a few 10's of days and few 100's of meters of each other, and gradually tail off (the 'z' axis is the relative Knox value). The spatio-temporal indices provide a very useful information upon which decisions can be made [11,12], although much work remains as to how to combine this with the developed technique, and what benefits would be gained from it.

5 Conclusion

This paper has presented a novel algorithm that builds on earlier work and incorporates into the *betweenness* metric of social network analysis many forms of information from adjacent nodes.

The combination of information is ad hoc and results are presented of a small network for illustration of the method.

All information is considered equal, i.e. the weightings given to offender range, 'danger' of offender, and strength of links are all 1. Future work may look at different weightings, or even an optimization of weightings so that a 'desirable' ranking can be achieved (similar to earlier work [3]).

The *betweenness* metric is very useful, however it is clear that in the case of burglary it needs to be balanced with spatial data. Consideration of the temporal and frequency analysis of the crimes constituting the links will provide a better understanding of the nature of the links, and may highlight links that are not considered significant by the *betweenness* metric.

The approach is best seen as a means to use the 'betweenness' metric in a more meaningful way when there are large numbers of offenders in a network. The resultant values would be useful for knowledge sharing, and training new police officers to the criminal activities in an area. Offenders that may have been overlooked come to the fore as key players by virtue of linking diverse geographical regions, or because of recently committing offences.

Future work has been outlined, including the evaluation of this algorithm, namely the application of this approach over large networks to prioritise offender 'observation' and 'intervention'.

References:

[1] Oatley, G.C., Ewart, B.W., & Zeleznikow, J., 2006. Decision Support Systems For Police: Lessons From The Application of Data Mining Techniques To 'Soft' Forensic Evidence. To appear in: Journal of Artificial Intelligence and Law

[2] Oatley, G.C. & Ewart, B.W., 2005. The meaning of links. In: D. Nelson, S. Stirk, H. Edwards and K. McGarry (eds.), Data mining and knowledge discovery in databases workshop, 22nd British National Conference on Databases, Vol. 2, pp.68-76

[3] Ewart, B.W., Oatley, G.C., & Burn K., 2004. Matching Crimes Using Burglars' Modus Operandi: A Test of Three Models. International Journal of Police Science and Management 7(3), pp. 160-174.

[4] Oatley, G.C., & Ewart, B.W., 2003. Crimes Analysis Software: 'Pins in Maps', Clustering and Bayes Net Prediction. Expert Systems with Applications 25 (4) Nov 2003 569-588

[5] Oatley, G.C., Zeleznikow, J. and Ewart, B.W., 2004. Matching and Predicting Crimes. In: Macintosh, A., Ellis, R. and Allen, T. (eds.), Applications and Innovations in Intelligent Systems XII. Proceedings of AI2004, The Twenty-fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence, Springer: 19-32. ISBN 1-85233-908-X

[6] Washio, T., & Motoda, H., 2003. State of the Art of Graph-based Data Mining. In: S. Dzeroski and L. De Raedt (eds.), SIGKDD Explorations Special Issue on Multi-Relational Data Mining, July 2003, Volume 5, Issue 1, 2003, pp. 59-69

[7] Getoor, L., 2003. Link mining: a new data mining challenge. In: S. Dzeroski and L. De Raedt (eds.), SIGKDD Explorations Special Issue on Multi-Relational Data Mining, July 2003, Volume 5, Issue 1, 2003, pp. 84-90.

[8] Klerks, P., 2001. The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. Connections 24 (3): 53-65

[9] Lyons, D., and G. S. Tseytin. 1998. Phenomenal data mining and link analysis. In D. Jensen and C. Henry Goldberg (Eds.), *Artificial Intelligence and Link Analysis, 1998 Fall Symposium*, 68-75. AAAI, AAAI Press.

[10] Adderley R. & Musgrove, P., 2003. Modus operandi modelling of group offending: a data-

mining case study. International Journal of Police Science and Management 5 (4) 265-276

[11] Johnson, S.D. & Bowers, K.J., 2004. The burglary as a clue to the future: the beginnings of prospective hot-spotting, The European Journal of Criminology, 1 (2), 237-255

[12] Johnson, S.D. & Bowers, K.J., 2005. Domestic burglary repeats and space-time clusters. European Journal of Criminology, 2 (1), 67-92