Bayesian Networks for Sex–related Homicides: Structure Learning and Prediction

Stephan Stahlschmidt[*][†][‡], Helmut Tausendteufel[§] and Wolfgang K. Härdle[†]

[*] Corresponding author. Email: stahlschmidt@wiwi.hu-berlin.de

[†] School of Business and Economics, Humboldt–Universität zu Berlin, Germany

[‡] Centre for Social Investment, Heidelberg University, Germany

[§] Department of Police and Security Management, Berlin School of Economics and Law, Germany

Abstract

Sex-related homicides tend to arouse wide media coverage and thus raise the urgency to find the responsible offender. However, due to the low frequency of such crimes, domain knowledge lacks completeness. We have therefore accumulated a large data set and apply several structural learning algorithms to the data in order to combine their results into a single general graphic model.

The graphical model broadly presents a distinction between an offender and a situation driven crime. A situation driven crime may be characterised by, amongst others, an offender lacking preparation and typically attacking a known victim in familiar surroundings. On the other hand offender driven crimes may be identified by the high level of forensic awareness demonstrated by the offender and the sophisticated measures applied to control the victim.

The prediction performance of the graphical model is evaluated via a model averaging approach on the outcome variable offender's age. The combined graph undercuts the error rate of the single algorithms and an appropriate threshold results in an error rate of less than 10%, which describes a promising level for an actual implementation by the police.

Keywords: Bayesian Networks, Offender Profiling, Ensemble Learning, Model Averaging, Criminal Event Perspective

JEL classification: C49, C81, K42

AMS clasification: 62-09, 62P25

1 Introduction

Criminal profiling can be defined as the process of identifiying a suspect's behavioural characteristics and principal personality from a crime scene. Police profilers firstly analyse the crime scene carefully and infer the exact course of events. Based on this groundwork they try to discover why theses events occurred and finally what type of person could have committed these acts. The method thereby relies on certain assumptions, most notably the belief that the criminal's personality can be retrieved from the crime scene.

Gaining a psychological and social profile of the suspect has several advantages for the police. Known characteristics of the offender can narrow the number of potential suspects by excluding those not showing the specific traits. This hopefully leads to a faster arrest of the criminal, but also reduces costs for the police and society. Furthermore the knowledge may lead to certain investigative strategies and, as people show different reaction to police interrogation approaches, prove useful during questioning of suspects.

Offender profiling has been enhanced by scientific background knowledge. To this end statistical techniques, like multidimensional scaling, cluster analysis or logistic regression have been applied to data of resolved crimes. This data has been generated by collecting evidence on the crime scene and learning the characteristics of the convicted offender by an interview or the criminal's record. An overview of the applied techniques is given by Beauregard (2007). However most studies concentrate on a rather broad typology or predict only single variables, e.g. Davies (1997) and Salfati and Canter (1999). Therefore Aitken et al. (1996) propose the application of Bayesian Networks (abbr. BN) based on expert knowledge and we extent their idea by deriving the BN from data.

Although the use of BNs on data of crimes is a promising application, the technique itself has already been applied to several fields such as crop failure (Wright, 1921), medical diagonis (Heckerman, 1990) or biological networks (Friedman et al., 2000) among others. This broad scope of BNs may be explained by its unique characteristics. A BN consists of a directed acyclic graph (DAG) which mirrors a factorisation of a probability distribution over several variables by including a directed edge between two dependent variables. Conditional independence among certain variables leads to a sparse graph, in which the nodes, representing variables, can be endowed with conditional probabilities. This combination of (in-)dependence statements and conditional probabilities describes the possibly causal relations among all factors of a specific domain and facilitates thereby statistical inference (Pearl, 2000).

BNs offer several advantages for criminal profiling, as they describe the structure of a pre–specified domain. Hence the building of a BN mainly by data may be used for learning the structure of an unknown domain, e.g. certain types of homicides. Furthermore BNs may also be employed for prediction. Profilers could for example obtain a prediction of the offender's age and thereby reduce the number of suspects substantially by entering evidence found on the crime scene into an appropriate BN. At this, crime scenes often lack certain information or do not only render one course of events plausible, but several competing ones. By its very nature a BN can be exploited to order competing and mutually exclusive hypotheses according to their probability given the facts and allow for inclusion of soft evidence. These reasons make this statistical technique appealing to profilers and the purpose of this paper is to present a case study on sex–related homicides in Germany.

The restriction to sex–related homicides, resulting from requests by the German Federal Police Office, has several implications, which distinguishes our work from previous research. First, sex–related homicides occur infrequently and second the assumption of homology between the offender's characteristics and the crime scene actions lacks further verification (Alison et al., 2002). Therefore, experts may name several important factors which are to be included in any systematical approach to the domain, but refrain from giving a precise ordering of these factors or detailing the exact relationships between these factors. Hence building a BN solely from expert knowledge is unfeasible and we therefore rely on a data–driven approach to learn the BN. However, data on sex–related homicides is scarce and not collected routinely like for example data on car accidents. We therefore accumulated a sample of sex–related homicides which occurred in Germany between 1991 and 2006. This study leads to one of the biggest and most detailed databases on this specific type of crime in Germany and we sequentially base our computation of a graphical model on this data.

The number of potential edges in a BN grows exponentially in the number of variables (Robinson, 1977) and although we have more observations than variables, we have considerably fewer observations than potential edges with their corresponding parameters θ. This situation leads to the realm of "$\theta >> n$" and poses several challenges for structural learning and prediction. We address those by relying on a combination of several algorithms in structure learning to find edges persisting throughout the resulting graphs and implementing a model averaging approach for prediction.

In general, a notional scale with two oppositional prototypes of sex–related homicides and several increments in between can be deduced from the graphical model. On one hand several criminals show a high level of preparation and forensic awareness. They apply sophisticated measures to control the victim and are more likely to exhibit a sadistic or serial background. Furthermore, they often attack victims which are unknown to them and which they contact in unfamiliar surroundings. These crimes carry a rather long enquiry period. On the other hand, several criminals do not show high levels of preparation or forensic awareness. Instead alcohol often constitutes a vital part of the crime and the offenders are more likely to apply blunt force instead of more elaborated measures to control their victim. They are often known to the victim and act in familiar surroundings. Several crimes do not belong strictly to either one of these prototypes, but only exhibit some of the specified features or even show features of both prototypes. However, in any case, offenders have to interact with factors, which they cannot affect, like the victim's resistance or the characteristics of the contact location.

In order to test the graphical model we deploy it to predict the outcome variable offender's age. At this a model averaging approach is taken in which the BNs arising from the diverse algorithms are consulted separately and the learned posterior distributions of the outcome are weighted according to the BN's likelihood. This procedure reveals a lower error rate than the single BNs and a threshold of $\pi = 0.6$ for the minimal probability of the fitted value lowers the error rate to the promising level of less than 10%.

In section 2 we report on the data collecting process. Section 3 discusses the applied

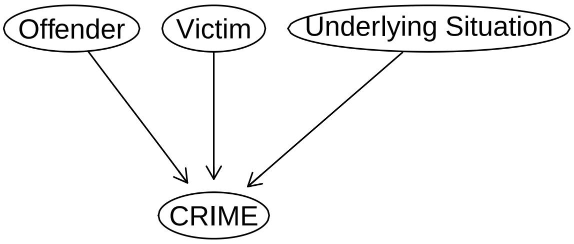

Figure 1: *Schematic overview on factors influencing a crime*

technique of BNs and section 4 describes the graphical model. Section 5 elaborates on its prediction performance, whereas section 6 concludes with a discussion of the results.

2 Data collection

The lack of data on sex–related homicides may be explained by the extremely tedious effort needed to collect the data and secondly by the restricted access to relevant information. Although court proceedings may be open to the public, the required time to attend several lawsuits would render this approach cumbersome. Furthermore information irrelevant to the judical conviction but essential from sociological or psychological perspectives may not be mentioned in detail. Hence, data collection relies on the assistance from authorities with access to adequate information.

The data presented in this paper is based upon support by the German police, which drew a sample of sex–related homicides from their internal documentation and provided access to the corresponding prosecutor's files. These files count between 1,500 and 10,000 pages, of which the crime scene report, the autopsy report, the psychiatric examination of the offender and the sentence contain almost all the essential information.

Transferring information from prosecutor's files into nominal variables requires comparable information throughout all cases. Therefore the prosecutor's files have to be scanned to determine which content would be available for an empirical analysis. Comparative text analysis (Strauss and Corbin, 1990) is a suitable technique to select the information satisfying this requirement and the variables presented in this study result from a comparative text analysis of 30 cases. However, not all available information is of use and the amount of information transferred into variables is restricted to a consistent set of important factors in the domain of sex–related homicides. The resulting variable selection is guided by sociological and psychological theory extended by the police's hands–on experience. These police experts provide information on relevant factors, but refrain from detailing the actual causal relations between these factors due to uncertain knowledge on this type of crime.

The information analysed in this study and transferred into variables focuses on four main elements: The offender, the victim, the underlying situation and the actual offence. A schematic overview is presented in Figure 1. The offender is described by his social, psychological and economic characteristics. Furthermore information regarding his medium–term and short–term disposition to commit a crime including his criminal record and any preparation to commit the crime are collected. Information on the victim is not widely available, however indicators on her social and economic status, as well on her prior relationship status with the offender is present throughout all cases. The underlying situation with its geographical and temporal information provides the general setting of the crime. The actual offence can be split up into several categories. First, any pre–attack events regarding the offender or shared activities between the offender and the victim before the attack are taken into account. Afterwards the actual crime begins with the offender's attack on the victim, which differs, for example, in the time needed, the victim's resistance or the level of applied violence. Resulting injuries including the fatal ones are recorded and sexual activities imposed on the victim are observed. Finally the offender's forensic awareness is measured and broadly divided into activities to hide his identity and activities to hide the crime.

The quality and quantity of the available information is highlighted by missing values and inter–rater reliability (Fleiss, 1971). Crimes resulting in limited traces entail a higher than average percentage of missing values. The same holds if the criminal refuses to testify, as several factors cannot be inferred from traces alone. Furthermore a high rate of missing values is accompanied by relatively low levels of inter–rater reliability. Raters seem to handle vague information differently. In general the data includes 6% missing values and Fleiss' meassure of inter–rater reliability between four raters amounts to $\kappa = 0.53$ with a percental match of 73%.

The actual collection of the data was carried out by reading all important documents of a particular crime and entering relevant information in a standardised form. This summary with all the essential information is thereafter deployed to assign all predefined variables. On average a rater completes two of the 252 cases on a typical working day. All together it took nearly a year to provide the data for the subsequent analysis.

3 Bayesian Networks and Structure Learning

A graph $\mathcal{G} = (\mathbf{V}, E)$ is defined by a set of nodes $\mathbf{V} = \{V_1, \ldots, V_p\}$ and a set of edges $E \subseteq \mathbf{V} \times \mathbf{V}$, which connect the nodes (Lauritzen, 1996). BNs form a particular subclass of graphical models and contain solely directed edges. Their set of edges E includes the entry (V_i, V_j), but not the entry (V_j, V_i) to denote a directed edge from node V_i to node V_j. Undirected edges are expressed as the entries (V_i, V_j) and (V_j, V_i) in E. In a directed edge (V_i, V_j) the node V_i is known as the parent of node V_j, and recursively the node V_j is said to be the child of V_i. The set of parents and children of a node V_i describe its adjacency. Extending the adjacency by all further parents of V_i's children, the Markov blanket of V_i is specified. The descendants $de(V_i)$ of any node V_i are defined by its children and any subsequent children. In order to distinguish clearly between descendants and non–descendants, we require the graph to omit circles. Consequentially the structure of a BN is known as a DAG (directed acyclic graph). A skeleton is a DAG without the arrow heads, such that all directed edges are converted into undirected edges. It includes several paths, describing a chain of nodes consecutively connected by edges. A chain of directed edges pointing all in the same direction is known as a directed path. If any two nodes point, via directed edges, at the same node without being adjacent, a collider arises.

A path π in a DAG $\mathcal{G} = (\mathbf{V}, E)$ is said to be blocked by a set $S \subseteq \mathbf{V}$ if node $V_w \in S$ on the path π is not a collider or some other collider $V_v \notin S$ on the path π exits and $V_w \notin de(V_v)$. Two disjoint subsets A and B of $\mathbf{V}$ are d–separated by S, if all paths between A and B are blocked by S (Jensen, 1996).

The probability distribution function of a random vector $\mathbf{X} = (X_1 \ldots X_p)^\top \in \mathbb{R}^p$ with an arbitrary ordering of the variables may be factorised as

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathrm{P}(X_i | X_1, \ldots, X_{i-1}). \tag{1}$$

Assuming that the conditional probability of some variable X_i is affected by only its Markovian parents $PA_i \subseteq \{X_1, \ldots, X_{i-1}\}$, which describe a subset of its predecessors, (1) can be shortened to

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathrm{P}(X_i | PA_i). \tag{2}$$

This assumption implies, conditional on the Markovian parents PA_i, independence between X_i and its non–Markovian parents predecessors $\overline{PA_i} = \{X_1, \ldots, X_{i-1}\} \backslash PA_i$, that is $\mathrm{P}(X_i | X_1, \ldots, X_{i-1}) = \mathrm{P}(X_i | PA_i, \overline{PA_i}) = \mathrm{P}(X_i | PA_i)$.

The probability distribution function (2) can be represented as a DAG establishing a tie between probability distribution functions and graphs. Variables X_i are displayed as nodes V_i and edges are drawn from the Markovian parents PA_i towards their child X_i. Such a DAG describes the corresponding probability distribution function graphically encoding dependencies in the distribution as edges. However, only if the probability function $\mathcal{P}$ allows for a factorisation according to (2) relative to a DAG $\mathcal{G}$, we may call $\mathcal{G}$ and $\mathcal{P}$ Markov compatible. As a consequence conditional independences in the probability function can be inferred from d–separations in the compatible graph (Lauritzen et al., 1990). A necessary and sufficient condition for this Markov compatibility is the so–called local Markov condition requiring that every variable in $\mathcal{P}$ may be independent of all its non–descendants conditional on its parents (Lauritzen, 1996).

3.1 Structure Learning

Structure learning refers to identifying the edges of a graphical model, where we assume that the i.i.d. data can be modeled as a sparse BN. The subsequent step of parameter learning endows the nodes with local probability functions or tables in order to transfer the DAG into a BN. As the space of DAGs grows exponentially in the number of variables, Chickering (1996) has shown that finding the correct structure of a BN is np–complete. Still several heuristic ideas exist to obtain the structure from observational data, which can be classified into constraint–based, score–based or hybrid approaches. Constraint–based approaches infer the existence of an edge by conditional independence tests and are vulnerable to errors in these tests. Furthermore the repeated application of independence tests inhibits any statement on the accuracy of the resulting graph, as the general significance level is unknown. Li and Wang (2009) have developed a constrained–based algorithm with a false discovery rate control which in comparison lacks power in disclosing existing edges. On the other hand score–based methods return a DAG, which possesses the highest score among all considered DAGs. Apart from choosing an appropriate score, these algorithms have to artificially narrow the search space in order to stay usable in large data sets. Finally, hybrid methods combine elements from constraint–based and score–based methods.

Although structure learning, defined as learning the existence of edges between nodes and consequently direct dependencies between the corresponding variables, is notoriously difficult, it constitutes our core interest. As a consequence the evaluation of different approaches by their error rate is ruled out, because an optimised prediction model may not resemble the existing dependencies and independencies in the data generating process (Meinshausen and Bühlmann, 2006). Furthermore the available data is limited in that there are much more potential edges than observations. The 53 variables in our analysis would lead to a complete graph of 1378 undirected edges, which existence we determine by analysing 252 observations.

We address these challenges by a combinatorial approach, which is loosely related to ensemble learning. In detail, we apply $J = 8$ different structure learning algorithms to the data, which return an indicator $ed_{ji} \in \{0, 1\}$ describing, if edge i has been included in the BN resulting from algorithm j. We combine these indicators ed_{ji} via the committee rule

$$ed_i^{Gen} = \mathbf{I}\left(\sum_{j=1}^{J} ed_{ji} > 0\right) ,$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. ed_i^{Gen} determines the inclusion of an edge in the combined graph shown in Figure 9 and consequently all edge included have been detected by at least one of the single algorithms. Obviously stricter committee rules lead to sparser combined graphs in which only edges found by several single algorithms persist. Meinshausen and Bühlmann (2010) propose a related approach for structure learning, which generates variation by sub–sampling and, via application of a single penalized structure learning technique to the sub–samples, allows for false discovery control in the final result.

Apart from the inclusion of an edge Figure 9 also reports on how often an edge has been detected across the algorithms. This frequency

$$ed_i^{Fre} = \sum_{j=1}^{J} ed_{ji} ,$$

determines the thickness of an edge i in the combined graph. Instead of deciding on a result via a committee rule, the graph offers, by the displayed frequencies ed_i^{Fre}, a degree of confidence in the existence of any edge which guides the resulting discussion of the graph.

We apply two score–based algorithms, five constraint–based algorithms and one hybrid algorithm. The plain Hill Climbing Greedy Search evaluates by which action the score improves most, conducts this step and reiterates until convergence. Feasible actions consist of adding or deleting an edge or changing an edge's direction. The Sparse Candidate algorithm (Friedman et al., 1999) also searches for the DAG with the highest score. Beforehand a set of potential parents are determined for every node and thereby the search space is reduced. The constrained–based algorithms Grow–Shrink Markov Blanket (Margaritis and Thrun, 1999) and Incremental Association Markov Blanket (Tsamardinos et al., 2003) identify a Markov blanket for every node and combine them to a BN. Whereas the Growth–Shrink algorithm adds any variable to the Markov blanket, as long as it exhibits some dependence given the current state of the Markov blanket, the Incremental Association algorithm adds the variable to the Markov blanket, which offers the maximal dependence given the current state of the Markov blanket. Finally in a backward phase both algorithms try to reduce the Markov blanket by rechecking the dependence. The constrained–based HITON algorithm (Aliferis et al., 2003) differs in executing the backward phase after each new inclusion of a variable to the Markov blanket instead of rechecking once at the end.

The Three Phase Dependence Analysis algorithm (Cheng et al., 2002) evaluates, via a statistical test, if a dependence between two variables can be explained by a path between them or if a separate edge connecting the two variables has to be included. As before a backward phase excludes any erroneously added edges. The well studied PC algorithm (Sprites et al., 2000) does not add edges but removes them immediately from a complete graph, if the corresponding variables exhibit independence given the neighbours of one of the two variables. The algorithm visits persisting edges multiple times, where the considered set of neighbours grows in cardinality. As an edge is instantly removed after recognising independence between the variables and consequently in sparse graphs many edges are only examined once or twice before discarding them, the PC algorithm provides a fast and consistent alternative in even high–dimensional settings (Kalisch and Bühlmann, 2007). The hybrid approach of the Max–Min hill climbing algorithm (Tsamardinos et al., 2006) combines the construction of a skeleton via independence tests with a score–based orientation of the edges.

4 Application

The application of the algorithms to our data yields several distinct graphs. We combine these graphs to a single one, in which the edge thickness is determined by how often an edge is found across the algorithms and indicates our confidence in an actual dependence between the corresponding variables in the data generating process. We omit the resulting edge direction and concentrate on the skeletons, as the algorithms do not agree uniformly on all edge directions. Figure 9 presents the resulting graph, which consists of 53 nodes and 83 edges. Beforehand any missing values were imputed five times and only edges persisting in all imputed data sets were included. Score–based algorithms were calculated via the Bayesian Information Criterion and the significance level in the constraint–based algorithms was set to 5%.

The single algorithms find between 20 and 68 edges. They agree completely on 4 edges and 1295 missing edges. A bar chart on the frequency of edges one or more algorithms, in changing combinations, agree upon is given in Figure 2. The maximal size of an adjacency in the final graph is 9, whereas the single algorithms provide adjacencies not larger than 8. The graph is considerably sparse taking into account the maximum of 1378 potential edges, which could arise from 53 variables.

The graph may be interpreted as showing the plain topology of an extensively organised offender, an offender lacking organisation and a mixture type (Ressler et al., 1988). However, this categorisation has been criticised for focusing solely on the offender and consequently has been enlarged to the Criminal Event Perspective (Miethe and Regoeczi, 2004). This theory stresses the influence of the victim and the underlying situation on the crime and thereby illustrates that for example, well prepared offenders may also show chaotic behaviour, if faced by unforeseen obstacles. The approach broadens the perspective to analyse a crime and we adapt it by including several variables describing the victim's behaviour and the underlying situation as illustrated in Figure 1.

Starting with the node "Preparation of offender", which is defined as the level of preparation to gain control over the victim, to hide the crime and to conduct the sexual assault, we observe 6 edges. The node may be located in the fourth row from below to the right in Figure 9 or in the lower centre of Figure 3. Of the emerging edges from the node "Preparation of offender", the edge towards the node "Sadistic Offender" sticks out by its thickness. The state of this node is defined via the psychiatric examination of the offender and is clearly connected to sadistic actions by the offender during the crime, included as the node "Sadism" in the graph. Examining the corresponding mosaic plots in Figure 6 it may be concluded that a sadistic offender is much more likely to behave sadistically and shows a higher level of preparation. Furthermore a sadistic offender conducts serial crimes more often than a non–sadistic offender. The node "Serial crimes", a dummy variable indicating if the specific crime is part of a wider series, exhibits profound edges to the offender's age and the enquiry period. Serial criminals usually belong to an age group of 24 to 33 years and obviously such a crime carries a longer enquiry period.

Apart from the sadistic offender, the serial criminal marks the second ideal example of an offender driven crime. On the other hand, there are situation driven crimes. These crimes show low levels of organisation by the offender and for the most part involve neither sadistic nor serial criminals. Rather they display a strong influence of the consumption of alcohol, which can be read in the graph by the edge between "Preparation by offender" and the node "Alcohol consumption by offender". This negative interaction is expanded by the node "Alcohol consumption by the victim" stating if the victim had consumed alcohol before the offender's attack. These also include cases in which the offender and victim voluntarily and

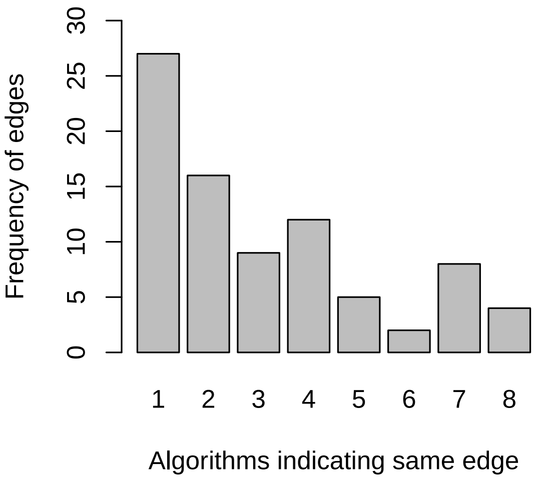

Figure 2: *Bar chart stating how many algorithms indicate the same edge and the frequency of such edges*

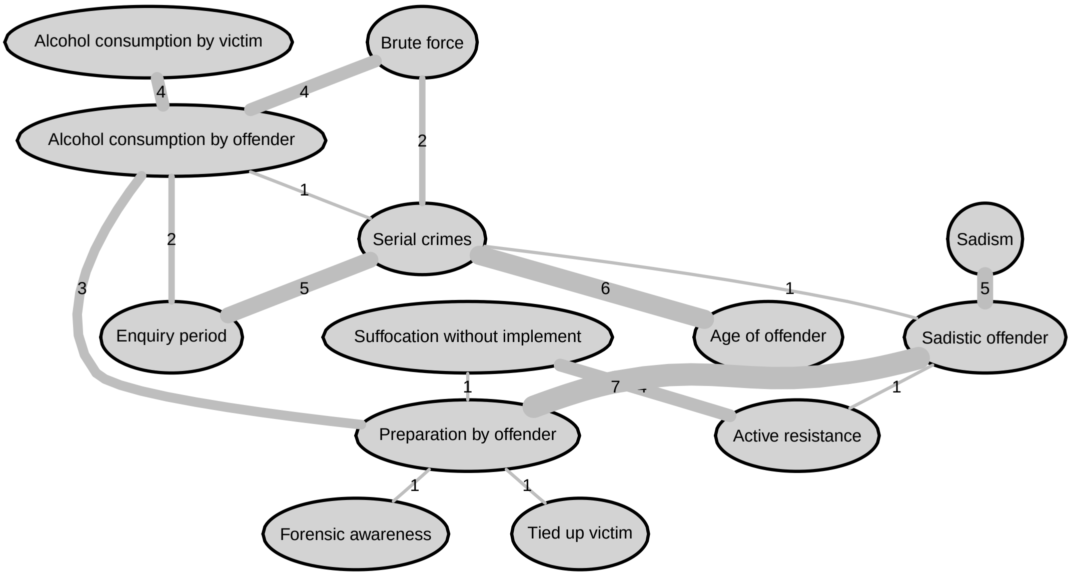

Figure 3: *Excerpt of Figure 9 showing variables which mark the difference between an offender and situation driven crime*

before the offender's attack engage in drinking. Most often either the victim and the offender have both consumed alcohol, which often leads to a situation driven crime, or neither the victim nor the offender have consumed alcohol, which characterises an offender driven crime. Apart from alcohol, the situation driven crimes are also marked by the use of brute force by the offender to gain and maintain control over the victim. The graphical model illustrates this interaction by the edge between "Alcohol consumption by offender" and the node "Brute force", which reflects any injuries of the victim due to the application of blunt force.

Serial criminals with their high level of preparation generally do not rely on blunt force, but apply more sophisticated measures to control the victim. This negative interaction can be read off the mosaic plot corresponding to the edge between "Brute force" and "Serial crimes" in Figure 7. One such measure to control the victim applied by offenders in a criminal driven crime is described by the edge between "Preparation by offender" and the node "Tied up victim". This node indicates if the victim is tied up by the offender and the corresponding cross–table reveals that offenders characterised by a high level of preparation are more likely to tie up their victim. Furthermore these offenders suffocate their victims less often with their hands, as highlighted by the cross–table corresponding to the edge between "Preparation by offender" and "Suffocation without implement". In general criminals with a high level of preparation apply a more instrumental mode to gain and maintain control, whereas a low level of preparation leads to a more expressive crime, where the offender likely applies blunt force. However, the likelihood of suffocation by the offender rises in both cases, whenever the victim strongly resists the attack. This general influence of the victim on the crime is specified by the edge between the nodes "Suffocation without implement" and "Active resistance", where active resistance is defined as resisting the assault physically, trying to escape or calling for help.

During the crime the level of planning carries over to the criminals' behaviour, as a high level of planning is accompanied by a high level of forensic awareness. Foren-

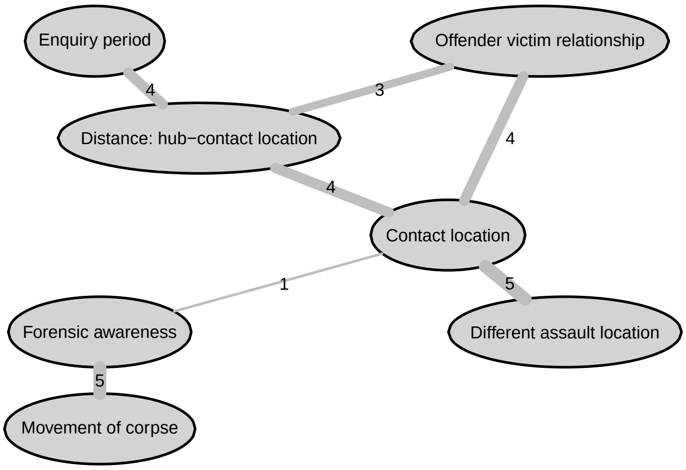
Enquiry period
Offender victim relationship
4
3
Distance: hub−contact location
4
4
Contact location
1
5
Forensic awareness
Different assault location
5
Movement of corpse

Figure 4: *Excerpt of Figure 9 showing geographical variables and their adjacency which mark the difference between an offender and situation driven crime*

sic awareness describes measures to hide the crime by for example using gloves or cleaning the crime scene afterwards. The corresponding node "Forensic awareness" is connected to the node "Preparation by offender" highlighting this positive interaction.

The node "Forensic awareness" links the degree of planning by the offender to certain geographical characteristics of the crime. The node may be found on the third row from below to the right in Figure 9 or to the left in Figure 4. Firstly, a criminal showing a high level of forensic awareness is more likely to hide the corpse at a separate location which serves solely for this purpose and complicates the prosecution. This interaction is reflected by the edge between "Forensic awareness" and "Movement of corpse".

Furthermore the node "Forensic awareness" is connected to the node "Contact location". This node describes the location of the first contact between the offender and the victim before the assault and distinguishes between location indoors, such as the victim's flat, the offender's flat or a shared flat, and locations outdoors. The corresponding mosaic plot reveals that offenders are less likely to show a high level of forensic awareness, if the contact takes place in their familiar surroundings, e.g. their own or a shared flat. On the contrary offenders meeting the victim in a rather unknown surrounding like the victim's flat or some location outdoors show a high level of forensic awareness and the corresponding crime is therefore most likely offender driven. The node "Contact location" exhibits a profound edge to the node "Offender victim relationship", which details the pre–attack relationship between the offender and the victim. Examining the corresponding cross–table reveals that the contact between the offender and an unknown victim is mostly established outdoors, whereas offenders meet any known victims rather indoors.

As an outdoor location is associated with a high level of forensic awareness, these outdoors contacts between the offender and the unknown victim may be attributed to the offender driven crime, whereas the indoor contact exhibits the characteristics of a situation driven crime and likely includes a victim known to the offender. An offender meeting the victim in his familiar surrounding obviously does not travel a great distance from his personal hub to the contact location, where a hub is defined as any location the offender is perfectly familiar with, e.g. his flat or work place. The graph therefore includes an edge between these two nodes and the corresponding mosaic plot is given in Figure 8. Furthermore the node “Distance: hub – contact location” is connected to the node “Enquiry period” and the corresponding mosaic plot details that a greater distance between the offender’s personal hub and the contact location complicates the prosecution, as the enquiry period rises.

Well organised offenders meet the victim in general not in their familiar surrounding, but have rather travelled a longer distance and hide the corpse at a separate location to impede the exposure of their crime. Less organised offenders meet the victim, which is most likely known to them, in rather familiar surroundings and do not travel a great distance. Furthermore they do not show a high level of forensic awareness or hide the corpse at a separate location. However, as before, the actual crime is not solely influenced by the criminal, as an examination of the edge between the nodes “Contact location” and “Different assault location” depicts. If the offender meets the victim in an outdoors location, in just over half of the crimes, the ensuing attack is conducted at a different location. The offender may not feel confident, that the contact location outdoors allows him to conduct the crime and is therefore forced to the change the location. This change of location occurs only in less of a quarter of all crimes, in which the contact location is indoors.

5 Prediction

The presented graphical model may be employed for prediction and information learned from a crime scene may be introduced in the BN to obtain a prediction for an unobserved outcome variable. The German police shows particular interest in learning the offender’s age from a crime scene, as such information often limits the number of potential suspects to a large extent and consequently may reduce the enquiry period to resolve the crime. Furthermore police profilers find it difficult to make a precise statement on the offener’s age or refrain from reporting any age classification.

In order to assess the prediction power of the combined graphical model we present a model averaging approach for the offender’s age. Due to the applied algorithms this outcome variable Y is classified in three age groups (younger than 24, older than 33 and in between), which account each for a third of the observations. To predict this discrete variable the eight candidate BNs $M_{j\in\{1,\dots,8\}}$ resulting from the diverse structure learning algorithms are deployed for predicting the outcome separately and the corresponding fitted values are pooled by a weighting schema. Beforehand the cross tables of the DAGs are parameterized via the expected value of the parameter’s posterior distribution originating from a uniform prior distribution. The actual prediction would start with introducing evidence X found on a crime scene in the BN by specifying the values of the corresponding variables. This information is transferred throughout the BN via a propagation algorithm altering the probability distribution of unobserved variables including offender’s age (Lauritzen and Spiegelhalter, 1988). The resulting fitted value $\widehat{Y}$ is chosen via a maximum a posteriori approach:

$$\widehat{Y} = \arg\max_{Y} \mathrm{P}(Y|Z)$$

In order to assess how well the the prediction fits the actual outcome value and due to the lack of data we apply a 10-fold cross validation procedure. The resulting training data Z, consisting of the training folds, is applied to build the eight diverse BNs and the remaining test fold X is deployed to calculate an error rate of the offender's age Y via the described model averaging approach

$$\mathrm{P}(Y|Z) = \mathrm{P}(f(X)|Z) = \sum_{j=1}^{8} \mathrm{P}(f(X)|M_j, Z)\,\mathrm{P}(M_j|Z).$$

At this $\mathrm{P}(f(X)|M_j, Z)$ denotes the posterior probability mass function obtained in the BN M_j learned via algorithm j on the training folds Z and deployed on the test fold X. This probability distributions is weighted by $\mathrm{P}(M_j|Z)$ which describes the likelihood of the BN M_j given the training folds Z (Hastie et al., 2009). Hence this pooled approach includes all algorithms of the combined graph in Figure 9, but adapts the weight of each algorithm by how well it accounts for the particularities of the training data.

The observed evidence to be entered in the BN may vary substantially from the data used to build the BN. Consequently the learned posterior distribution of the outcome variable may not differ to a large extent from the uniform prior distribution. In such circumstances it seems reasonable to refrain from any prediction statement in order to not mislead the police by doubtful indications. We therefore implement a threshold π which the posterior distribution needs to surpass in order to accept the corresponding fitted value. The threshold and corresponding error rate, calculated as 0–1 loss, of the single algorithms and the model averaging approach are depicted in Figure 5.

Whereas the model averaging approach belongs to the top three predictors if no threshold is imposed, this combinatorial approach surpasses its single components, if the threshold is set to $\pi \geq 0.45$. Taking into account the prior uniform probability of the three age categories and the need for a rather large jump to report trustworthy prediction statements to the police, this threshold value does not seem unreasonable. Hence the combination of algorithms, as highlighted by the combined graph in Figure 9, results in better prediction performance than any single algorithm. Incorporating a threshold of $\pi \geq 0.6$ lowers the error rate to less than 10%, which describes an appropriate level for a real-life implementation of the graphical model by the police.

6 Conclusion

This study demonstrates that learning a BN from data yields several insights into the domain of sex–related homicides. Hence we provide profilers with profound knowledge and extend previous statistical research in the realm of offender profiling. The combined skeleton of the obtained BNs shows the dependency structure in the domain and is calculated via various algorithms. The resulting single BNs of these algorithms are combined to a final structure by summing up on how often an edge is found by the diverse algorithms. The algorithms are applied on a new data set of 252 cases of sex–related homicides in Germany.

In general, a notional scale with two oppositional prototypes of sex–related homicides and several increments in between can be deduced from the graphical model.

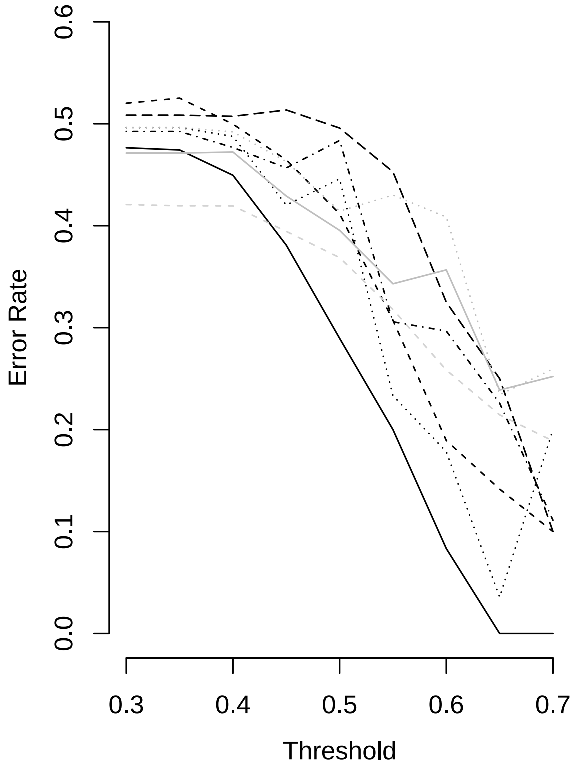

Figure 5: *Prediction performance via model averaging (black solid line), Growth–Shrink algorithm (black dashed line), Incremental Association Markov Blanket algorithm (black dotted line), Max–Min Hill Climbing algorithm (black dotted dashed line), Hill Climbing (black long dashed line), Three Phase Dependence Analysis algorithm (grey solid line), PC algorithm (grey dotted line) and HITON algorithm (grey dashed line).*

On one hand several criminals show a high level of preparation and forensic awareness. They apply sophisticated measures to control the victim and are more likely to exhibit a sadistic or serial background. Furthermore, they more often attack victims unknown to them, which they contact in unfamiliar surroundings. These crimes carry a rather long enquiry period. On the other hand, several criminals do not show high levels of preparation or forensic awareness. Instead alcohol often constitutes a vital part of the crime and the offenders are more likely to apply blunt force instead of more elaborated measures to control the victim. They are often known to the victim and act in familiar surroundings. Several investigated crimes do not belong strictly to one of these prototypes, but exhibit only some of the specified features or even show features of both prototypes. However, in any case offenders have to interact with factors, which they cannot affect, like the victim's resistance or the characteristics of the contact location.

The distinction between an offender and situation driven crime examines the graphical model in a certain perspective. Different views exist to analyse sex–related homicides and consequently different distinctions may be found. However, due to the low number and the heterogeneity of the analysed cases these points of view do not stick out as clearly as the described distinction, but have been noticed in smaller sub–samples (Safarik et al., 2002).

A model averaging approach based on the presented skeleton reveals promising prediction results. The combination of algorithms yields a lower error rate than the single algorithms if a reasonable threshold must be exceeded by the outcome variable's posterior distribution. In detail a threshold for the probability of the fitted value of 0.6 lowers the error rate to less than 10%, which denotes strong argument for a real-life implementation of the presented graphical model by the police.

Acknowledgements

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 ”Ökonomisches Risiko”, Humboldt-Universität zu Berlin, and the Bundeskriminalamt is gratefully acknowledged.

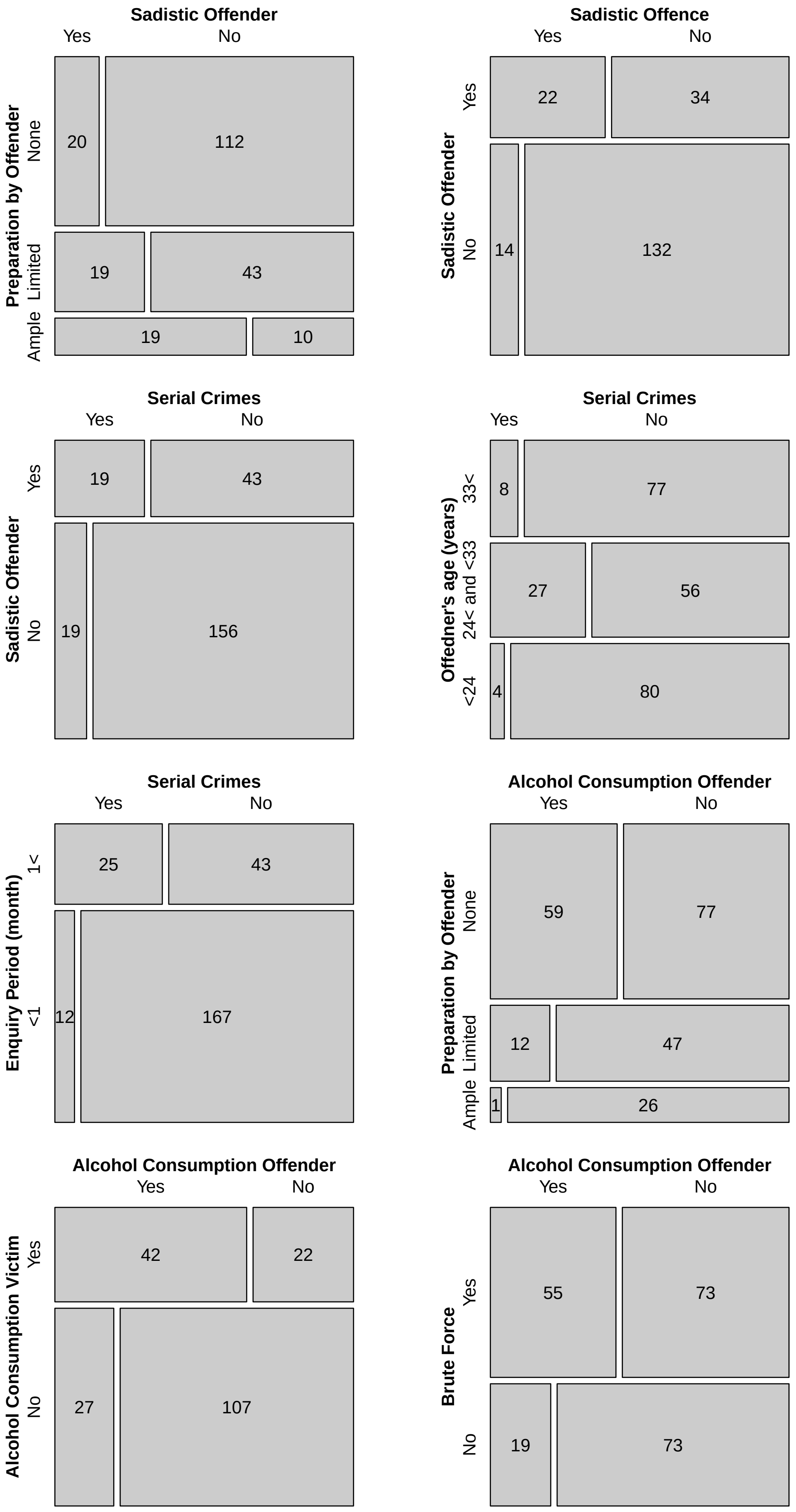

Figure 6: *Mosaic plots corresponding to discussed edges*

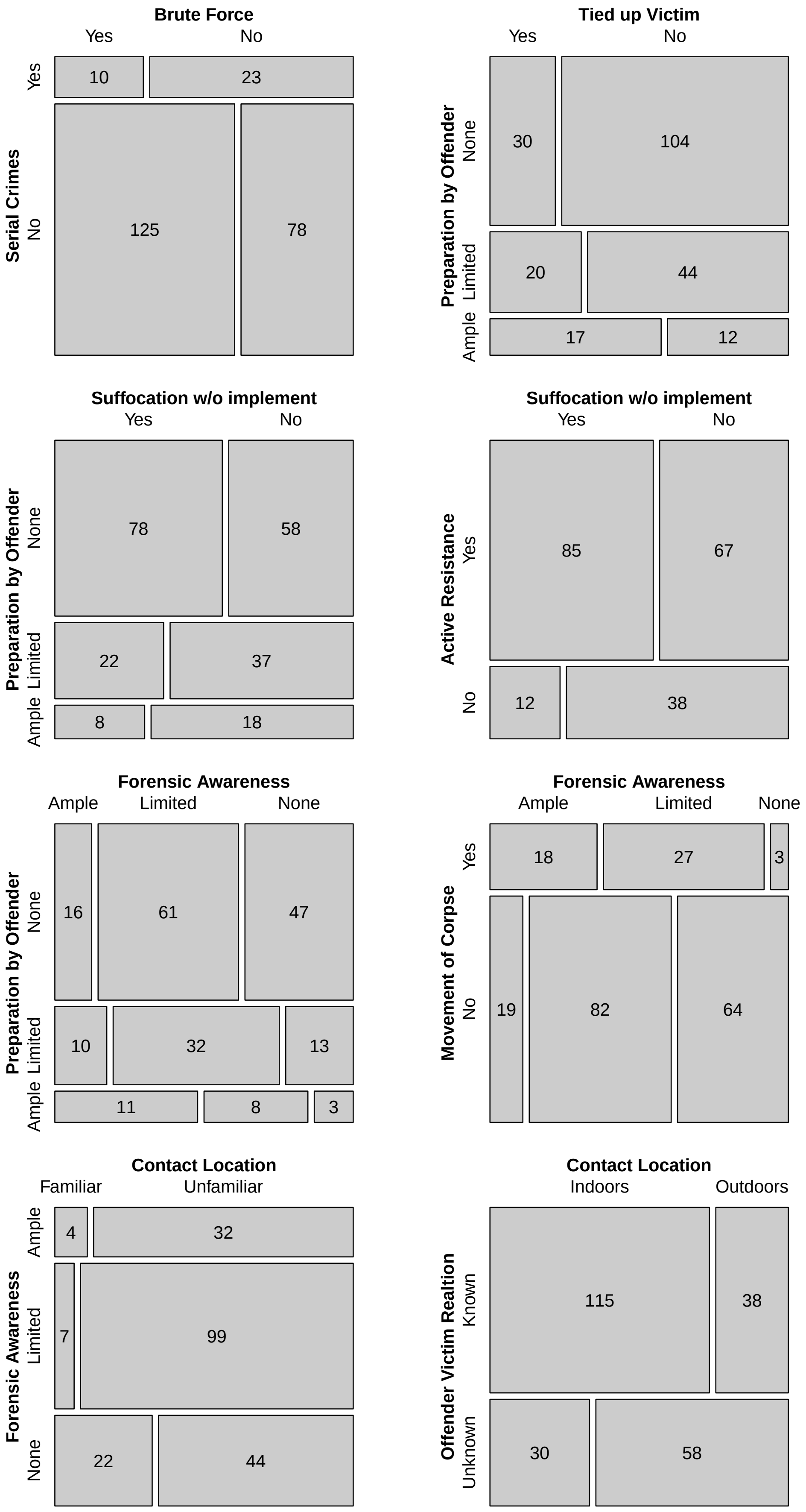

Figure 7: *Mosaic plots corresponding to discussed edges*

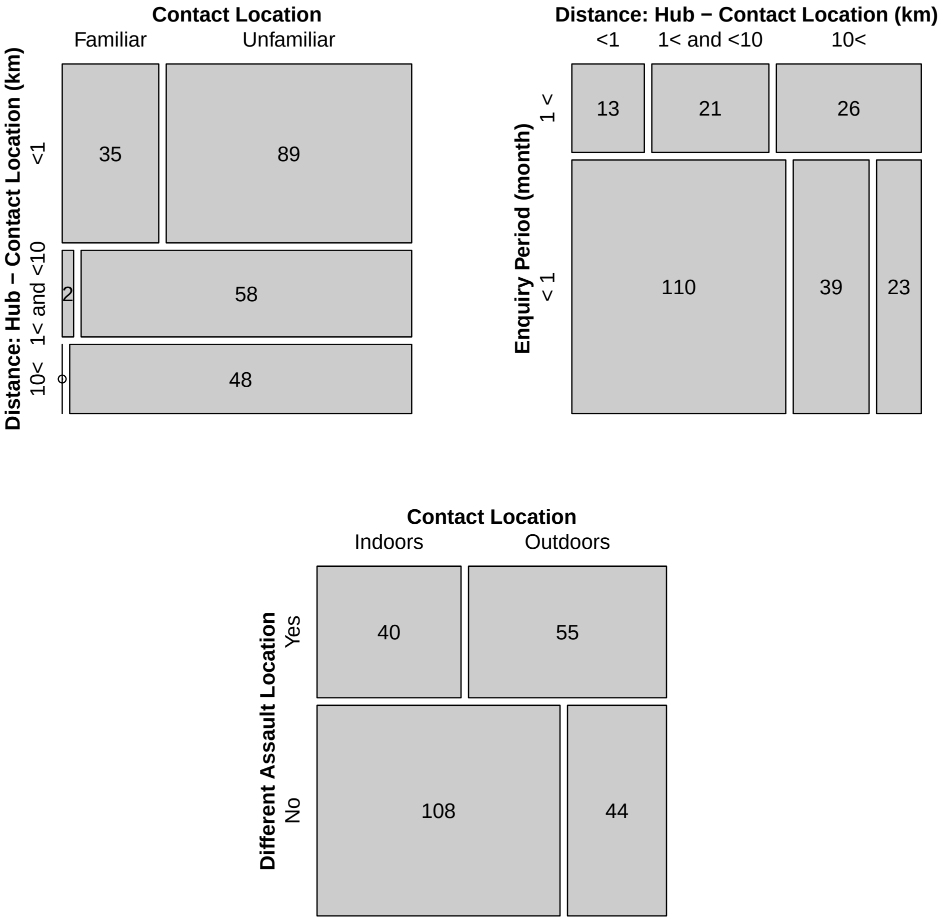

Figure 8: *Mosaic plots corresponding to discussed edges*

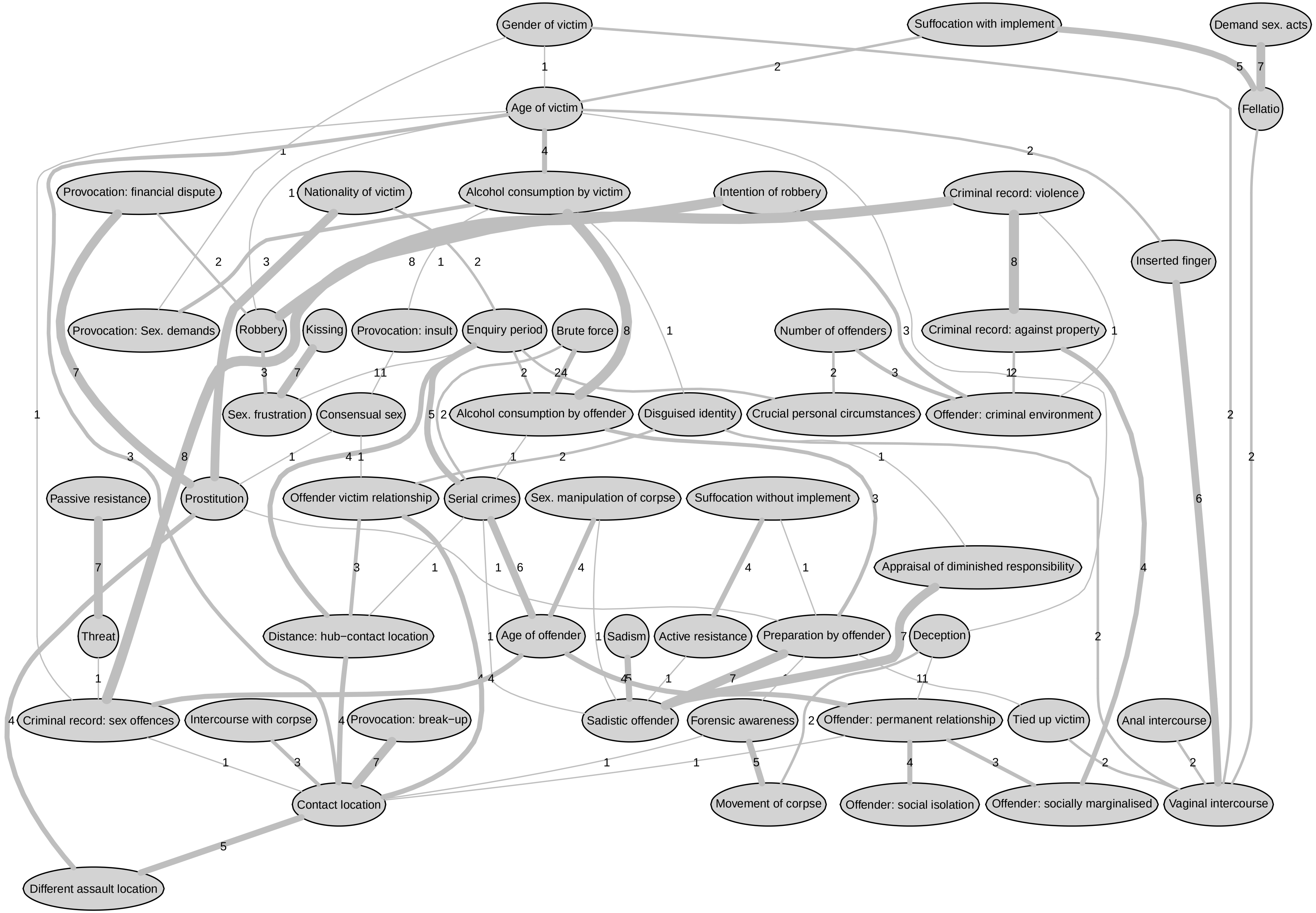
Gender of victim
Suffocation with implement
Demand sex. acts
Age of victim
Fellatio
Provocation: financial dispute
Nationality of victim
Alcohol consumption by victim
Intention of robbery
Criminal record: violence
Inserted finger
Provocation: Sex. demands
Robbery
Kissing
Provocation: insult
Enquiry period
Brute force
Number of offenders
Criminal record: against property
Sex. frustration
Consensual sex
Alcohol consumption by offender
Disguised identity
Crucial personal circumstances
Offender: criminal environment
Passive resistance
Prostitution
Offender victim relationship
Serial crimes
Sex. manipulation of corpse
Suffocation without implement
Appraisal of diminished responsibility
Threat
Distance: hub−contact location
Age of offender
Sadism
Active resistance
Preparation by offender
Deception
Criminal record: sex offences
Intercourse with corpse
Provocation: break−up
Sadistic offender
Forensic awareness
Offender: permanent relationship
Tied up victim
Anal intercourse
Contact location
Movement of corpse
Offender: social isolation
Offender: socially marginalised
Vaginal intercourse
Different assault location

Figure 9: *Graphical Model for sex–related homicides. Numbers on the edges denote the frequencies* ed_i^{Fre}.

References

Aitken, C. G. G., Gammerman, A., Zhang, G., Connolly, T., Bailey, D., Gordon, R. and Oldfield, R. (1996). Bayesian belief networks with an application in specific case analysis. In *Computational Learning and Probabilistic Reasoning* (ed A. Gammerman). Chichester: Wiley.

Aliferis, C. F., Tsamardinos, I. and Statnikov, A. (2003). HITON, a novel Markov blanket algorithm for optimal variable selection. In *Proceedings of the 2003 American medical informatics Association Annual Symposium*, 21–25.

Alison, L. , Bennell, C., Mokros, A. and Ormerod, D. (2002). The personality paradox in offender profiling *Psychology, Public Policy, and Law*, **8**, 115–135.

Beauregad, É. (2007). The Role of Profiling in the Investigation of Sexual Homicide. In *Sexual Murderers: A Comparative Analysis and New Perspectives* (eds J. Proulx, É. Beauregard, M. Cusson and A. Nicole). Chichester: Wiley.

Cheng, J., Greiner, R., Kelly, J., Bell, D. A. and Liu, W. (2002). Learning Bayesian Networks from data. *The Artificial Intelligence Journal*, **137**, 43–90.

Chickering, D. M. (1996). Learning Bayesian Networks is NP–complete. In *Learning from Data: Artificial Intelligence and Statistics V* (eds Fisher, D. and Lenz, H.–J.). New York: Springer.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, **93**, 935–948.

Davies, A. (1997). Specific profile analysis: a data–based approach to offender profiling. In *Offender profiling: Theory, Research and Practise* (eds Jackson, J. L. and Bekerian, D. A.). Chichester: Wiley.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378–382.

Friedman, N., Linial, M., Nachman, I. and Peer, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, **7**, 601–620.

Friedman, N., Nachman, I. and Peer, D. (1999). Learning Bayesian Network Structure from massive Datasets. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Inteligence*, 206–215.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of Statistical Learning.* New York: Springer.

Heckerman, D. (1990). Probabilistic similarity networks. *Networks*, **20**, 607–636.

Jensen, F. V. (1996). *Introduction to Bayesian Networks.* New York: Springer.

Kalisch, M. and Bühlmann, P. (2007). Estimating high–dimensional directed acyclic graphs with the PC–Algorithm. *Journal of Machine Learning Research*, **8**, 613–636.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H. G. (1990). Independence properties of directed markov fields. *Networks*, **20**, 491–505.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Clarendon Press.

Li, J. and Wang, Z. J. (2009). Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. *Journal of Machine Learning Research*, **10**, 475–514.

Margaritis D. and Thrun, S. (1999). Bayesian Network induction via local neighborhoods. In *Advances in Neural Information Processing Systems 12* (eds Solla, S. A., Leen, T. K. and Müller, K.–R.). Cambridge: MIT Press.

Meinshausen, N. and Bühlmann, P. (2006). High–Dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, **34**, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B*, **72**, 417–473.

Miethe, T. D. and Regoeczi, W. C. (2004). *Rethinking Homicide.* New York: Cambridge University Press.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press.

Ressler, R. K., Burgess, A. W. and Douglas, J. E. (1988). *Sexual homicide.* New York: Lexington Books.

Robinson, R. W. (1977). Counting unlabelled acyclic digraphs. In *Lecture Notes in Mathematics: Combinatorial Mathematics V.* Heidelberg: Springer.

Safarik, M. E., Jarvis, J. P. and Nussbaum, K. E. (2002). Sexual Homicide of Elderly Females. *Journal of Interpersonal Violence*, **17**, 500–525

Salfati, G. and Canter, D. V. (1999). Differentiating stranger murders: profiling offender characteristics from behavioral styles. *Behavioral Scinces and the Law*, **17**, 391–406.

Sprites, P., Glymour, C. and Scheines, R. (2000). *Causation, prediction and Search.* Cambridge: MIT Press.

Lauritzen, S.L. and Spiegelhalter, D. (1988). Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, **50**, 157–224.

Strauss, A. and Corbin, J. M. (1990). *Basics of qualitative research.* Thousand Oaks: Sage Publications.

Tausendteufel, H., Stahlschmidt, S. and Kühnel, W. (2011). *Bestimmung des Täteralters bei sexuell assoziierten Tötungsdelikten auf der Basis von Tatgeschehensmerkmalen.* Wiesbaden: Bundeskriminalamt.

Tsamardinos, I., Aliferis, C. F. and Statnikov, A. (2003). Algorithms for large scale markov blanket discovery. In *The 16th International FLAIRS Conference*, 376–381.

Tsamardinos, I., Brwon, L. E. and Aliferis, C. F. (2006). The max–min hill–climbing Bayesian Network structure learning algorithm. *Machine Learning*, **65**, 31–78.

Wright, S (1921). Correlation and Causation. *Journal of Agricultural Research*, **20**, 558–585.