Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Bayesian Network Modeling of Offender Behavior for Criminal Profiling

Kelli Crews Baumgartner, Silvia Ferrari, and C. Gabrielle Salfati

Abstract— **A Bayesian network (BN) model of criminal behavior is obtained linking the action of an offender on the scene of the crime to his or her psychological profile. Structural and parameter learning algorithms are employed to discover inherent relationships that are embedded in a database containing crime scene and offender characteristics from homicide cases solved by the British police from the 1970s to the early 1990s. A technique has been developed to reduce the search space of possible BN structures by modifying the greedy search K2 learning algorithm to include *a-priori* conditional independence relations among nodes. The new algorithm requires fewer training cases to build a satisfactory model that avoids zero-marginal-probability (ZMP) nodes. This can be of great benefit in applications where additional data may not be readily available, such as criminal profiling. Once the BN model is constructed, an inference algorithm is used to predict the offender profile from the behaviors observed on the crime scene. The overall model predictive accuracy of the model obtained by the modified K2 algorithm is found to be 79%, showing a 15% improvement with respect to a model obtained from the same data by the original K2 algorithm. This method quantifies the uncertainty associated with its predictions based on the evidence used for inference. In fact, the predictive accuracy is found to increase with the confidence level provided by the BN. Thus, the confidence level provides the user with a measure of reliability for each variable predicted in any given case. These results show that a BN model of criminal behavior could provide a valuable decision tool for reducing the number of suspects in a homicide case, based on the evidence at the crime scene.**

I. Introduction

The empirical research on offender profiling so far has been limited both in scope and impact due to the many variables describing the criminal act and corresponding investigation, as well as the high degree of uncertainty surrounding both. In this paper, a network modeling approach using Bayesian networks (BNs) is developed for modeling offender's behavior on the crime scene, with the purpose of predicting the offender's profile in unsolved cases. A valid network model of criminal behavior could potentially aide investigators by providing estimates of the biographical, motivational, and psychological profile of the unknown offender based on the analysis of past crimes, thereby reducing the number of potential suspects [1]. Current research using another network modeling approach investigates the usefulness of neural networks to criminal profiling [1]. It has been suggested (e.g. [2]) that offender profiling is not only possible, but that it is a psychologically straightforward process. However, it is much more complex than just a "multilevel series of attributions, correlations and predictions" [3]. Also, much of the psychological profiling used in investigations to date has been guesswork based on hunches and anecdotal information accumulated through years of experience, including error and misinterpretation [3]. Indeed, to date, much of the psychological profiling related to homicide investigations has been linked to particular individuals rather than to empirically tested and established scientific methods.

Recent developments in empirical crime scene analysis using statistical methods to understand the link between crime scene actions by an offender and his/her characteristics have shown promise [4]. Based on 82 British single offender-single victim solved homicides a statistical analysis was used to classify cases according to specific behavioral themes: the *expressive* theme, composed of behaviors that center on the victim as a certain person, and the *instrumental* theme, centered on the benefits they had for the offender (e.g., either sexual or material gain). The study that followed with a larger sample (247) of single offender-single victim solved homicide cases showed similar results [5].

The BN approach presented here seeks to discover these correlations in the offenders' behavior in order to obtain a useable criminal profile from the crime scene evidence. Building on the results from [1], this paper proposes a systematic approach for deriving a multidisciplinary behavioral model of criminal behavior. The proposed crime behavioral model is a mathematical representation of a system comprised of an offender's actions and decisions at a crime scene and the offender's personal characteristics. The influence of the offender traits and characteristics on the resulting crime scene behaviors is captured by a probabilistic graph or BN that maps cause-and-effect relationships between events, and lends itself to inductive logic for automated reasoning under uncertainty [6].

In order to overcome the challenges facing criminal profiling, probabilistic graphs are suitable modeling technique because they are inherently distributed and stochastic. In this work, the BN is initialized from expert knowledge, while the mathematical relationships naturally embedded in a set of crimes ([2], [7], [8]) are learned through training from a database containing solved criminal cases. The BN behavioral model enables the prediction of a criminal profile to produce a corresponding probabilistic confidence level or *likelihood*. Thus, it overcomes the critique that criminal profiling techniques lack substantiation, by offering the likelihood that a certain characteristic is present in an offender

K. Baumgartner is a graduate student of Mechanical Engineering at Duke University, Durham, NC 27707, USA kac20@duke.edu

S. Ferrari is with Faculty of Mechanical Engineering at Duke University, Durham, NC 27707, USA sferrari@duke.edu

C.G. Salfati is with Faculty of the Department of Psychology at John Jay College of Criminal Justice, New York, NY 10019, USA gsalfati@jjay.cuny.edu

0-7803-9568-9/05/$20.00 ©2005 IEEE

[9].

The BN approach to criminal profiling is demonstrated by learning a BN structure and parameters from a series of crime scene and offender behaviors designated by experts through their professional experience (expert knowledge). A modified K2 algorithm for structural learning is developed in order to reduce the computational complexity of learning a model with many variables when the number of training cases is fixed. This method is compared to the standard K2 algorithm. Both techniques are evaluated on a set of validation cases, not used for learning, by defining a prediction accuracy based on the most likely value of the output variables (offender profile) and its corresponding confidence level.

A. Bayesian Network Notation and Theory

In this paper, capital letters denote variables and lowercase letters denote the *states* or instantiations of the variables (i.e. X_i is said to be in its j^{th} instantiation when $X_i = x_{i,j}$). A variable or *node* in a BN corresponds to each item in a domain $\mathcal{X} = (X_1, ..., X_n)$ for $n > 1$ discrete variables in the probability space $\{\Omega, \mathcal{F}, \mathcal{P}\}$. The probability space of a BN refers to a structure or graph $\Omega = \{\mathcal{X}, \mathcal{S}\}$, where $\mathcal{S}$ is the set of directed arcs (denoted by arrows) between the variables $\mathcal{X} = (X_1, ..., X_n)$. The variables and *directed* edges of Ω together comprise a *graph*, referred to as a *directed acyclic graph* (DAG) [10]. The BN parameter $\mathcal{F}$ is the space of all possible instantiations of X_i, for $i = 1, ..., n$. $\mathcal{P}$ is the probability distribution for all X_i with respect to $\mathcal{S}$ and $\mathcal{F}$.

Let $\mathcal{B}$ be the set of all possible BNs, $\mathcal{B} = (\mathcal{S}, \Theta)$, where $\mathcal{S}$ is the DAG with parameters $\Theta = (\theta_1, ..., \theta_n)$ and $\Theta \in \mathcal{P}$. The parameter $\theta_i \in \Theta$ is the *conditional probability table* (CPT) attached to node X_i. A CPT lists in tabular form the conditional probabilities of each state of X_i with respect to each of its parents, $P(X_i|\pi_i)$, where π_i represents the parents of X_i. If a node has no parents, the CPT for θ_i is simply a prior probability distribution $P(X_i)$. Every X_i has a CPT that is either initialized by a user from prior knowledge or learned from the set of training cases, described in detail in Section III. A *sample* over $\mathcal{X}$ is an observation for every variable in $\mathcal{X}$. A database $\mathcal{D}$ is a compilation of d samples of $\mathcal{X}$, $\mathcal{D} = \{C_1, ..., \mathcal{C}_d\}$. $\mathcal{D}$ is said to have no *missing values* when all values of all variables are known. An assumption is made that each individual sample C_i is independent and identically sampled (i.i.d) with an underlying unknown distribution.

A BN is a mathematical model based on the acquired data and the implementation of Bayes' rule [10], [11]. Bayes' rule of dependence can be utilized to calculate the posterior probability distribution of X_i given the instantiations of X_i's children, represented as μ_i, as follows

$$P(X_i|\mu_i) = \frac{P(\mu_i|X_i)P(X_i)}{P(\mu_i)}. \tag{1}$$

The prior probability of X_i, $P(X_i)$, is the known probability distribution over the states of X_i, $(x_{i,1}, ..., x_{i,ri})$. The likelihood function, $P(\mu_i|X_i)$, contains the conditional probabilities of the instantiated children variables connected to X_i. This becomes the product of the likelihood probabilities of the instantiated variables $P(\mu_i|X_i) = \prod_{j=1}^{p} P(\mu_{i(j)}|X_i)$, where $\mu_{i(j)}$ is the instantiation of the j^{th} child of X_i. The marginalization of the observed variables, $P(\mu_i)$, accounts for the relationship between the instantiated variables and all possible states of X_i as follows:

$$P(\mu_i) = \sum_{k=1}^{ri} P(X_i = x_{i,k}) \prod_{j=1}^{p} P(\mu_{i(j)}|X_i), \tag{2}$$

where $\mu_{i(j)}$ is the j^{th} instantiated variable of X_i's p total children. The posterior probability of $X_i = x_{i,k}$, denoted by $P(X_i = x_{i,k}|\mu_i)$, is also known as the marginal probability of $x_{i,k}$ and represents its confidence as a probability for which to occur given the evidence. X_i is *inferred* from μ_i using (1).

The exact computation of the marginal probabilities of a BN is often too computationally expensive [12], [13]. Constructing an inference engine allows for a more tractable procedure to calculate the marginal probabilities in the BN [6]. Efficient inference engines identify the conditional independencies between the variables in a system in order to simplify computation. An important property addressing conditional independence is the *directed Markov property*, which states that a variable is conditionally independent of its non-descendants (i.e. $nd(\cdot)$) given its parents [6]: $X_i \perp nd(X_i)|\pi_i$. Typically, this property simplifies the inference procedure. Here, it also is exploited to simplify structural learning obtaining the so-called K2′ algorithm, discussed in Section III. See [6], [14] for a review on constructing an inference engine.

II. Criminal Profiling Modeling

A. Problem Formulation

Currently, a *criminal profile* (CP) is obtained from a psychological interpretation linking crime scene characteristics to the likely behaviors of the offender completed by an investigator or forensic psychologist. This research seeks an efficient and systematic discovery of non-obvious and valuable patterns from a large database of solved cases via a causal network (BN) modeling approach. The objective is to produce a more systematic and empirical approach to profiling, and to use the resulting BN model as a decision tool.

A criminal profile model is learned from a database of solved cases and it is tested by comparing its predictions to the actual offenders' profiles. The database $\mathcal{D}$ containing d solved cases $\{C_1, ..., C_d\}$, where C_i is an instantiation of $\mathcal{X}$, is randomly partitioned into two independent datasets: a training set $\mathcal{T}$ and a validation set $\mathcal{V}$, such that $\mathcal{D} = \mathcal{T} \cup \mathcal{V}$. The variables $\mathcal{X}$ are partitioned as follows: the *inputs* are the *crime scene* (CS) variables $(X_1^I, ..., X_k^I) \in \mathcal{X}$ (evidence), and the *outputs* are the *offender* (OFF) variables comprising the criminal profile $(X_1^O, ..., X_m^O) \in \mathcal{X}$.

The BN model is learned from $\mathcal{T}$, as explained in Section III-A, and it is tested by predicting the offender variables in the validation cases $\mathcal{V}$. Then, the BN is used to estimate the criminal profile, by designating as predictions the most likely values of offender variables. During the testing phase, the predicted value of X_i^O, denoted by $x_{i,a}^P$ where a=1,2, or r_i, is compared to the observed state $x_{i,b}^O$ obtained from the validation set $\mathcal{V}$, where $b = 1, 2$, or r_i. An example of an offender variable is "gender", with states "male" and "female". The overall performance of the BN model is evaluated by comparing the true (observed) states $x_{i,b}^O$ to the predicted values $x_{i,a}^P$ for all output variables inferred from the evidence over the $\mathcal{V}$ cases. This process tests the generalization properties of the model by evaluating its efficiency over $\mathcal{V}$.

B. Criminal Profiling Variables

The set of CP variables acquired from police reports of homicide crime scenes was defined by criminologists in [7], [15], [16], [5], [8]. The selection criteria for variable initialization [8] are: (*i*) behaviors are clearly observable and not easily misinterpreted, (*ii*) behaviors are reflected in the crime scene, e.g. type of wounding, and (*iii*) behaviors indicate how the offender acted toward and interacted with the victim, e.g. victim was bound/gagged, or tortured. 36 CS variables describing the observable crime scene and 21 OFF variables describing the actual offender were initialized based on the above selection criteria. Examples of the CS variables are multiple wounding to one area, drugging the victim, and sexual assault. Examples of the offender variables include prior offenses, relationship to the victim, prior arrests, etc. The variables all have binary values representing whether the event was present or absent.

C. Database of Solved Cases and Sample Demographics

A set of single offender/single victim homicides was collected by psychologists from solved homicide files of the British police forces around the UK spanning from the 1970s to the early 1990s. This same data was also used in criminal profiling research [5], [8].

In these 247 sample cases, the majority of the victims were female (56%) with a mean age of 41 years, ranging from 0 to 93. Male victims (44%) had a mean age of 39 years, ranging from 0 to 82. The offenders in this sample were predominantly male (89%) with a mean age 32 years ranging from 16 to 79. The female offenders (11%) had ages ranging from 17 to 70, with a mean age of 33 years. Only 15% of the cases were considered sex crimes and only 9% of the offenders had a prior sexual convictions. As for the victim/offender relationships, 10% of the victims were related to their offender (either by blood or otherwise) and 43% of the victims had a previous sexual relationship with the offender (excluding cases of prostitution). A total of 83% of the offenders knew the victim in any capacity at all prior to the offense.

III. Methods

A. Learning

Since the recent development of efficient inference algorithms [17], [13], BNs have become a common representation tool in computer science. They also are useful for control and decision making because they can model stochastic processes from data. A BN allows for causal interpretation of events in which predictions of intervention are made with some unknown information. A set of probabilistic Bayesian networks $\mathcal{B}$ can be constructed given a database containing the instantiation of a set of variables and an implicit assumption about the variables' characteristics and interactions with each other. A learning framework is used to obtain the network that "best" describes the database.

Ideally, if $\mathcal{B} = (\mathcal{S}, \Theta)$ denotes the set of all possible BNs with nodes $\mathcal{X}$ reflecting the variables in $\mathcal{D}$, then the compatibility of all DAGs with $\mathcal{T}$ would be compared pair-wise. The compatibility of each hypothesized structure, $\mathcal{S}^h \in \mathcal{S}$, with the training data is assessed by a so-called scoring metric that assigns a value, or *score*, to each $\mathcal{S}^h$ [12], [17], [11]. Thus, the optimal score is the maximum conditional probability of $\mathcal{S}^h$ given the training data $\mathcal{T}$, i.e.: $\max P(\mathcal{S}^h|\mathcal{T})$. Since the calculation $P(\mathcal{S}^h|\mathcal{T})$ is computationally infeasible, it is recognized that because $P(\mathcal{D})$ is independent of $\mathcal{S}^h$, a more feasible calculation is the joint probability $P(\mathcal{S}^h, \mathcal{T})$ [12]. Thus, the scoring metric becomes a joint probability calculation, where the joint probability distribution is given by

$$P(\mathcal{S}, \mathcal{T}) = \int_\Theta P(\mathcal{T}|\mathcal{S}, \Theta)P(\Theta|\mathcal{S})P(\mathcal{S})d\Theta. \qquad (3)$$

For the following assumptions, the computation of (3) becomes tractable: (*i*) all models are equally likely, $P(\mathcal{S}) \sim i.i.d.\ \ Uniform(\alpha)$; (*ii*) all cases in $\mathcal{T}$ occur independently given a BN model; (*iii*) all variables are discrete and known, making $P(\mathcal{T}|\mathcal{S}, \Theta)$ a probability mass function [12]. With assumptions (*i-iii*), the scoring metric becomes a joint-probability scoring metric [12] that can be simplified as follows

$$P(\mathcal{S}^h, \mathcal{T}) = P(\mathcal{S}^h) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(\bar{N}_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!, \quad (4)$$

where n discrete variables in $\mathcal{X}$ each have r_i possible states $(x_{i,1}, ..., x_{i,r_i})$, q_i is the number of unique instantiations for π_i, N_{ijk} is the number of cases in $\mathcal{T}$ where $X_i = x_{i,k}$, and $\bar{N}_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. $\mathcal{S}^h$ is encoded as a discrete variable whose state corresponds to the set of possible network structures in $\mathcal{B}$ and assesses the probabilities $P(\mathcal{S}^h)$. Since (4) depends on the relative compatibility of the hypothesized structure with the data and the goal is to find $\mathcal{S}^h$ with maximum score, the scoring metric is maximized with respect to $\mathcal{S}^h$.

Since the number of possible structures grows exponentially as a function of the number of nodes [18], a more feasible search algorithm is needed to systematically limit the search space in order to find a suitable local optimal structure, $\mathcal{S}^{opt}$, for a domain of variables $\mathcal{X}$. A search algorithm does not guarantee to find the structure with the highest probability, but it systematically reduces the computationally infeasible search space and, at the same time, maximizes the scoring function.

A greedy search algorithm [12], [19] referred to as the heuristic search K2 algorithm is one method explored in this research. The following simplifying assumptions are added to (*i-iii*): (*iv*) ordering of nodes, and (*v*) limited number of parents per node. The ordering of nodes in $\mathcal{X}$ refers to allowing only causal arcs in the forward path. In this algorithm, if it is assumed that X_1 precedes X_2, it excludes a causal arc from X_2 to X_1. These assumptions lead to a simplified score, from (4),

$$g = \log(\prod_{j=1}^{q_i} \frac{(r_i-1)!}{(\bar{N}_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!). \tag{5}$$

The log score is implemented because of its monotonically increasing characteristic that is computationally more efficient. Assumption (*i*) of marginal independence for $P(\mathcal{S}^h)$ holds for $P(\mathcal{S}^h) = P(\pi_{i,j} \rightarrow X_{i,j})$, which denotes the probability of the causal relationships of $\pi_{i,j}$ to $X_{i,j}$, when $i \neq j$ and π_i is independent of π_j. The complexity of the K2 algorithm is significantly less than the complexity of an exhaustive search. The function g in (5) is $\mathcal{O}(mur)$, where m is the maximum number of cases in $\mathcal{T}$, u is the maximum number of parents allowed per node, and $r = \max_{1 \leq i \leq n} r_i$. When this function is called at most $n-1$ times, it requires $\mathcal{O}(munr)$ computation time. Each of the total nodes n is limited to a maximum of u parents leading to a computation time of $\mathcal{O}(un)$. The resulting complexity of the K2 algorithm with a bound on the maximum number of parents is $\mathcal{O}(mu^2n^2r)$ [12].

The second learning method used in this research further reduces the computational complexity of (4) while still maintaining a suitable search space by introducing an additional assumption of input independence. The purpose of learning a BN is to infer variables that are non-observable from the values of the observable variables. If it is known prior to learning that a set of nodes *always* will be instantiated during the inference process, independence among these variables can be established. These conditional independence relationships are illustrated by the BN in Figure 1. Since X_4 has influence on X_1 which in turn has influence on X_2 and X_3, then evidence on X_2 and X_3 will effect the inference of both X_1 and X_4. However, if X_1 is known, this instantiation blocks communication to its parent and children respectively: X_4 is said to be *d-separated* from X_2 and X_3 [10]. Similarly, if it is known prior to learning that X_1 and X_4 are always instantiated and never inferred, then regardless of the connection between the X_1 and X_4, these variables are always independent of each other. This statement is derived from the property of admittance of *d*-separation in BNs, which states that if two variables X_4 and X_2 are *d*-separated in a BN with evidence e, then $P(X_1|X_4, e) = P(X_1|e)$ [10]. Inhibiting certain node connections prior to learning eliminates a subset of potential BNs and, thus, increases the efficiency of the greedy search algorithm. This independence assumption is insufficient if the data is incomplete. Hence, it should be used only for those nodes that will be instantiated by the observations.

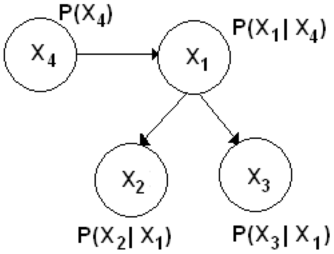


Fig. 1. Inserting variable X_4 as a parent to another input variable showing independence between X_1 and X_4 if they are both instantiated

In this paper, the modified K2 algorithm, where a particular set of arc is blocked *a priori*, is referred to as K2$'$. The complexity of the K2$'$ algorithm is significantly less than the complexity of the K2. The computational savings of K2$'$ over K2 comes with the additional conditional independence assumption among input variables. The computation time is reduced from $\mathcal{O}(un)$ to $\mathcal{O}(uk)$, where $n - d = k$, d is the number of variables that are independent of each other, and k is the total number of nodes with parents. Thus, the overall complexity of K2$'$, $\mathcal{O}(mu^2nkr)$ time, significantly decreases as the number of independent variables, d, increases.

B. Prediction

An initialized BN, either by the user or though learning, is used for probabilistic inference. Inference is the process of updating the probability distribution of a set of possible outcomes based upon the relationships represented by the model and the observations of one or more variables. With the updated probabilities, a prediction can be made from the most likely value of each inferred variable.

A BN mapping of n variables $(X_1, ..., X_n) \in \mathcal{X}$ represents a joint distribution over a discrete sample space. Thus, the joint probability of any particular instantiation of $X_{i \rightarrow n} \in \mathcal{X}$ can be calculated as,

$$P(X_1, ..., X_n) = \prod_i P(X_i|\pi_i), \tag{6}$$

where the variable X_i has n possibilities and π_i represents the instantiation of the parents of X_i. From the directed Markov property stated in Section I-A, the recursive factorization of (6) is simplified if the conditional independence relationships among the variables are identified, given the evidence. The inference engine is compiled through the steps of graphical manipulations described in Section I-A. In this research, the Matlab functions utilized are *jtree_inf_engine* to build the junction tree; *enter_evidence* to insert evidence; *marginal_nodes* to complete the inference on the specified nodes for the respective junction tree and evidence, and are found in *Bayes Net Toolbox* for Matlab [20].

The marginal probability is the probability that node X_i is in a particular state given the evidence and the usual property $\sum_{j=1}^{b} P(X_i = x_{i,j}|\pi_i) = 1$. The distribution of marginal probabilities for an inferred node is referred to as the predictive distribution. The state of a variable is predicted by choosing the state with the maximum marginal probability. In causal BNs generally the "causes" are the parent nodes and the "effects" are the children nodes. In this research, the offender profile is the cause for the resulting crime scene. Also, observations are made from the crime scene with the purpose of predicting the offender profile. Therefore, the inputs are the crime scene variables and the outputs are the offender variables (parent nodes), as is illustrated in Figure 2b for *m* outputs and *k* inputs.

IV. Results

A BN model of offenders behavior on the crime scene is learned and tested using murder cases solved by the British police forces from the 1970s to the early 1990s (Section II-B). The initial structure $\mathcal{S}_o$ is initialized as an empty set, assuming no prior knowledge about the node correlations, as seen in Figure 2a. The training data is used to build the BN by cycling through the set of possible BN, $\mathcal{B}^h \in \mathcal{B}$, using both the K2 and K2$'$ structural learning algorithms for comparison. The K2$'$ algorithm inhibits connections between the k input nodes X_i^I, for $i = (1, ..., k)$. After the structure is learned, the maximum likelihood parameter estimation (MLE) algorithm is used to find the corresponding parameters Θ^h based on $\mathcal{D}$.

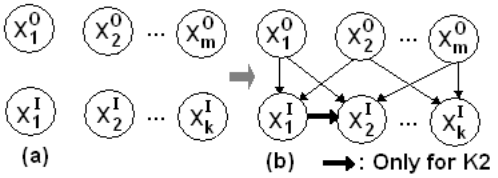


Fig. 2. The initial BN structure is an empty set with no connections (a) which is connected by the structural learning algorithms producing the final structure (b).

The database $\mathcal{D}$ of single offender/single victim homicides used in this research contains 247 cases and are divided into $\mathcal{T}$ (200 cases) and $\mathcal{V}$ (47 cases). The variables in $\mathcal{X}$ are partitioned into 36 *crime scene* input variables $(X_1^I, ..., X_{36}^I)$ (evidence), and into 21 *offender* output variables $(X_1^O, ..., X_{21}^O) \in \mathcal{X}$. The outputs comprise the criminal profile to be inferred from the evidence. All variables $X_i^{I,O} \in \mathcal{X}$ are binary ($r_i = 2$), with the value $x_{i,j}^{I,O}$ representing whether an event is either present ($x_{i,1}^{I,O}$) or absent ($x_{i,2}^{I,O}$). The optimal model of offender behavior, $\mathcal{B}^{opt}$, is learned from the training data $\mathcal{T}$. The maximum number of parents allowed per node (u) is set to 10. So, the complexity of K2$'$ is $\mathcal{O}(mu^2nkr) = \mathcal{O}(4.79 \times 10^7)$ and is reduced with respect to the K2 algorithm, with complexity $\mathcal{O}(mu^2n^2r) = \mathcal{O}(1.3 \times 10^8)$. Table I shows a comparison of the overall performance for the models obtained by the two algorithms. The improved accuracy brought about by the K2$'$ indicates that the conditional independence relations assumed between the crime scene variables correctly reflect the crime situation.

In every one of the 47 validations cases, 21 output variables are predicted, leading to a total of 987 predictions. Because the variables are all binary, a uniformly-random prediction procedure would produce ~50% predictive accuracy (PA). Where, the predictive accuracy is defined as the frequency at which output variables are inferred correctly over the 47 validation cases, $\mathcal{V}$. A predicted variable is said to be inferred correctly, or its prediction is said to be correct, when the true (observed) state $x_{i,b}^O$ is equal to the predicted value $x_{i,a}^P$. The overall model predictive accuracy (OPA) is the percentage of correct predictions over the total number of predictions (987). The predictive accuracy of an individual node (IPA) is computed by considering the correct predictions of that node value over the total number of validation cases (47). The results in Table I show that the predictive accuracy of the K2 and K2$'$ algorithms is better than 50%. This suggests that this BN method may have value in predicting offender profiles in unsolved cases. Also, the K2$'$ algorithm has a better predictive accuracy than the K2 algorithm.

TABLE I
Overall performance efficiency for K2 and K2$'$ algorithms for 987 total predictions.

Algorithm:	**K2**	**K2$'$**
Accuracy (%):	64.1%	79%
Correct Predictions (number):	633	780

Further comparison of the K2 and K2$'$ models involves the confidence levels of each prediction. When compared to other expert systems, such as Neural Networks, probabilistic networks have the added advantage that their predictions are based on posterior probability distributions for the states of each variable, also known as marginal probabilities. The marginal probability $P(x_{i,j}^P|e)$ is computed for each state of an inferred node X_i, and can be seen as the confidence level of a prediction stating that $X_i = x_{i,j}^P$. Table II shows that as the marginal probability for the predicted variable increases, so does the accuracy of the prediction. The accuracy of nodes predicted with a confidence level CL is denoted by CLA and is calculated by the following formula

$$CLA = \frac{K_{C,CL}}{K_{CL}} * 100, \tag{7}$$

where, $K_{C,CL}$ is the total number of correct predictions (subscript C) with a specified confidence level (subscript CL), and K_{CL} is the total number of nodes in the specified confidence level. For example, from Table II if the designated confidence level is $\geq 70\%$, K_{CL} is the number of nodes with a marginal probability $\geq 70\%$ ($K_{CL} = 573$ for K2 and $K_{CL} = 725$ for K2$'$), and $K_{C,CL}$ is the number of correctly predicted variables with the $\geq 70\%$ confidence level ($K_{C,CL} = 493$ for K2 and $K_{C,CL} = 618$ for K2$'$). Table II shows a comparison of the K2 and K2$'$ models with respect to the number of predictions with confidence levels of $\geq 50\%$, $\geq 70\%$, and $\geq 90\%$. It is apparent that the K2$'$ model has significantly more variables that are predicted with a higher confidence level, although the CLA for both methods are similar. For CL=$\geq 70\%$, $CLA = 86\%$ for K2 and $CLA = 85.2\%$ for K2$'$, but the number of variables predicted correctly is significantly higher for K2$'$

TABLE II

CONFIDENCE LEVEL OF PREDICTIONS FOR THE K2 AND K2′ ALGORITHM MODELS.

Algorithm:	**K2 ‖ *K2′***		
Confidence Level, CL (%):	$\geq 50\%$	$\geq 70\%$	$\geq 90\%$
K_{CL}:	798 ‖ *987*	573 ‖ *725*	168 ‖ *255*
$K_{C,CL}$:	633 ‖ *780*	493 ‖ *618*	159 ‖ *232*
Confidence Level Accuracy (CLA, %):	79.3% ‖ *79%*	86% ‖ *85.2%*	94.6% ‖ *91%*

($K_{C,CL} = 618$) compared to K2 ($K_{C,CL} = 493$).

Although the CLA of the models obtained by the K2 and K2′ algorithms are close, the number of inferred nodes with a marginal probability $\geq 50\%$ is consistently higher with the K2′ model. Only 798 out of 987 predictions had a predicted marginal probability ≥ 0.5 for the K2 model, because 189 variables had a predictive marginal probability of zero. In this research, the occurance $P(X_j = x_i|e) = 0$ for $i = 1, 2$, which results in $\sum_{i=1}^{r=2} P(X_j = x_i|e) \neq 1$, is referred to as "Zero Marginal Probability" (ZMP) node. ZMP nodes have been observed when inference is performed in a BN with an inadequate number of training cases. Where, the number of training cases required depends not only on the network size and database $\mathcal{D}$, but also on the number of variables that need to be inferred in an unsolved case. As can be deduced from Table II, the number of ZMP nodes is 189, or 19% of all predictions. This idea of ZMP leads to a more accurate calculation of the model's PA (recorded in Table I):

$$PA = \frac{K_t - (K_w + K_{ZMP})}{K_t} \cdot 100, \tag{8}$$

where K_w is the number of variables inferred incorrectly, K_{ZMP} is the number of ZMP variables, and K_t is the total number of predictions ($K_t = 987$ for OPA or $K_t = 47$ for IPA). $K_{C,CL}$ for CL=$\geq 50\%$ is related to K_w and K_{ZMP} by $K_{C,\geq 50\%} = K_w + K_{ZMP}$.

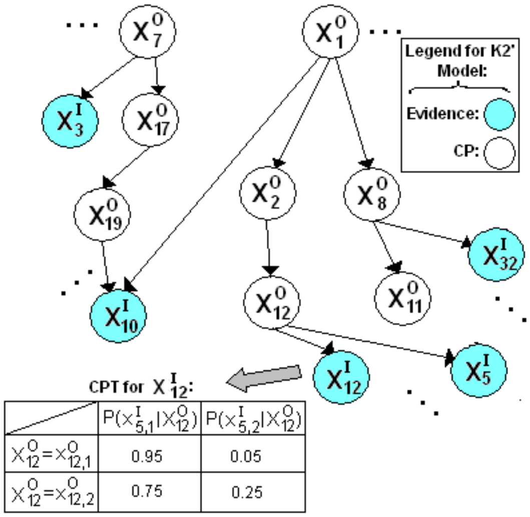
The decrease in prediction efficiency caused by ZMP is overcome by (*i*) using more training cases, (*ii*) decreasing the number of variables, or (*iii*) decreasing the number of variable relationships. However, the number of cases available is usually not up to the programmer, and it is not good practice to eliminate variables, since important relationships could be lost. The solution (*iii*) to decrease the search space through additional simplifying assumptions is typically the most useful. In this work, the search space is decreased through the K2′ algorithm, which reduces the number of possible variable relationships by exploiting conditional independence properties. Table II shows that the given training data (200 cases) is sufficient for learning the BN model through the K2′ algorithm but insufficient in learning for the K2 algorithm as seen by the presence of ZMP nodes. Although computational savings previously mentioned by the K2′ algorithm from the K2 algorithm may appear at first to be insignificant, it is enough to eliminate the ZMP nodes from the model for inference when the number of training cases is limited.

Another benefit of BN modeling is the graphical display of the relationships learned for a given system. A slice of the K2′ model is shown in Figure 3, to illustrate an example of relationships between 5 of the 36 crime scene input variables, (X_3^I, X_5^I, X_{10}^I, X_{12}^I, X_{32}^I), and 8 of the 21 output offender variables, (X_1^O, X_2^O, X_7^O, X_8^O, X_{11}^O, X_{12}^O, X_{17}^O, X_{19}^O). All of these variables are defined in Tables III-IV. Because the full DAG is too large to show here, the arcs are listed in the third column in Table IV. For example, X_8^O has three children variables listed in Table IV, with two of the three depicted in Figure 3. A total of 9 variables are disconnected from the graph (no arcs leading to or from the variable): X_{14}^O, X_7^I, X_{13}^I, X_{15}^I, X_{16}^I, X_{17}^I, X_{18}^I, X_{20}^I, X_{21}^I. This is due to either an insufficient number of training cases to recognize the relationships, or that relationships simply do not exist.

By inspecting the structure and CPTs of the BN learned from data, a pattern can be found linking the action of deliberately hiding the victim's face (X_5^I) to the offender's gender (X_{12}^O), as seen in Figure 3. The CPT acquired from the MLE parameter learning algorithm with respect to the structure for variable X_5^I is also shown in Figure 3. The values in the CPT are viewed as a probabilistic degree of influence supporting the state of the unknown variable based on the evidence. The influence between X_{12}^O and X_5^I is interpreted as strongly supporting a male offender ($X_{12}^O = x_{12,1}^O$) if the face is hidden (0.95 compared to 0.75). Instead, if the evidence shows that the victim's face is not hidden, the gender of the offender is more likely female ($X_{12}^O = x_{12,2}^O$, 0.25 compared to 0.05). However, the BN in Figure 3 also shows that when inferring the gender of the offender, X_{12}^O, the evidence on wounding from a blunt instrument, X_{12}^I, must also be taken into account. Of course, through inference in the BN, the influence of all observable crime scene variables on the offender profile is taken into account simultaneously. But this example shows how the learned BN structure also portrays the relationships discovered from the data, and thus can be easily utilized by a multidisciplinary team interested in understanding human behavior.

V. CONCLUSION

A Bayesian network modeling approach is developed to identify underlying patterns of criminal behavior from a database of solved cases. A well-known structural learning algorithm, known as K2, is implemented and compared to a modified version that exploits conditional independence



CPT for X_{12}^I:

	$P(x_{5,1}^I \mid X_{12}^O)$	$P(x_{5,2}^I \mid X_{12}^O)$
$X_{12}^O = x_{12,1}^O$	0.95	0.05
$X_{12}^O = x_{12,2}^O$	0.75	0.25

Fig. 3. A slice from the actual full BN structure that is learned from data by the K2′ algorithm (CPTs are not shown for simplicity).

relations among the input variables. The modified K2′ algorithm is faster, more effective, and requires fewer number of training cases for learning a BN from data for the purpose of predicting a criminal profile. This paper shows that additional conditional independence relationships can be effectively incorporated into the learning procedure to increase the final model performance. Inhibiting nodal connections systematically decreases the search space and is shown to improve the model performance considerably. Most importantly, the K2′ requires a smaller sample of training cases than the K2 algorithm, which may otherwise lead to ZMP predications. This attribute is particularly useful in applications where additional data is not easily acquired. These preliminary results support the idea that underlying patterns exist between offenders and their crime, and that they can be learned from a set of solved cases. Future research will expand upon this methodology to systematically evaluate and improve automated criminal profiling techniques.

VI. Appendix

The 21 offender and 36 crime scene variables are described in Tables III-IV. Table IV lists the children variables of each particular parent variable (i.e. the arcs of the DAG).

TABLE III

Definition of the crime scene variables.

Variable	**Definition**
X_1^I:	Vaginal penetration
X_2^I:	Anal penetration
X_3^I:	Foreign object penetration
X_4^I:	Victim found face up
X_5^I:	Victim's face not deliberately hidden
X_6^I:	Victim partially undressed
X_7^I:	Victim naked
X_8^I:	Clothing damage
X_9^I:	Bound (at one point)
X_{10}^I:	Blindfolded (at one point)
X_{11}^I:	Stabbed
X_{12}^I:	Blunt instrument
X_{13}^I:	Manual method (e.g. strangulation)
X_{14}^I:	Shot
X_{15}^I:	Wounds to head
X_{16}^I:	Wounds to face
X_{17}^I:	Wounds to neck
X_{18}^I:	Wounds to torso
X_{19}^I:	Wounds to limbs
X_{20}^I:	Multiple wounds to one area of body
X_{21}^I:	Multiple wounds distributed over body
X_{22}^I:	Weapon brought to scene
X_{23}^I:	Weapon from scene used
X_{24}^I:	Identity property taken
X_{25}^I:	Property taken beyond identity
X_{26}^I:	Property of value taken
X_{27}^I:	Body hidden
X_{28}^I:	Body transported
X_{29}^I:	Offender forensically aware
X_{30}^I:	Body not moved after death
X_{31}^I:	Sexual crime
X_{32}^I:	Suffocation (other than strangulation)
X_{33}^I:	Arson to crime scene or body
X_{34}^I:	Found in water
X_{35}^I:	Drugged or poisoned
X_{36}^I:	Victim found inside

TABLE IV

DEFINITION OF THE OFFENDER VARIABLES AND A LIST OF THE DIRECTED ARCS. NOTE THAT "PRIOR" REFERS TO A PRIOR CRIMINAL CONVICTION OF SAID OFFENSE.

Variable	**Definition**	**Arcs (children)**
X_1^O:	Prior theft	X_2^O, X_8^O, X_{21}^O, X_{10}^I, X_{24}^I, X_{25}^I
X_2^O:	Prior burglary	X_3^O, X_4^O, X_6^O, X_9^O, X_{12}^O, X_{26}^I
X_3^O:	Prior violence	X_4^O, X_8^I, X_{12}^I
X_4^O:	Prior damage	X_5^O, X_{20}^O
X_5^O:	Prior disorder	–
X_6^O:	Prison	X_9^O, X_{12}^O, X_3^I
X_7^O:	Young offender between 17-21 years	X_{10}^O, X_{17}^O, X_2^I, X_3^I, X_9^I
X_8^O:	Unemployed at the time of offense	X_{11}^O, X_{14}^I, X_{32}^I
X_9^O:	History of sex crime	X_{13}^O
X_{10}^O:	Armed service	–
X_{11}^O:	Familiar with area of offense occurrence	–
X_{12}^O:	Male	X_5^I, X_{12}^I, X_{31}^I, X_{35}^I, X_{36}^I
X_{13}^O:	Knew victim	X_{15}^O, X_{19}^O, X_{20}^O, X_{27}^I, X_{28}^I, X_{30}^I
X_{14}^O:	History of abuse	disonnected
X_{15}^O:	Suicide (attempted after crime)	X_{16}^O, X_{21}^I
X_{16}^O:	Psychiatric or social problems	X_{28}^I, X_{34}^I
X_{17}^O:	Prior fraud	X_{19}^O, X_{14}^I, X_{33}^I
X_{18}^O:	Related to victim	X_6^I, X_{11}^I, X_{19}^I, X_{21}^I, X_{24}^I
X_{19}^O:	Relationship with victim	X_{20}^O, X_{10}^I, X_{35}^O
X_{20}^O:	Blood related to victim	X_{32}^I
X_{21}^O:	Turned themselves in	X_1^I, X_4^I, X_6^I, X_{24}^I, X_{29}^I, X_{31}^I, X_{33}^I

REFERENCES

[1] M. Strano, "A neural network applied to criminal psychological profiling: An italian initiative," *International Journal of Offender Therapy and Comparative Criminology*, vol. 48, no. 4, pp. 495–503, 2004.

[2] R. Ressler, A. Burgess, and J. Douglas, *Sexual Homicide: Patterns and motives*. New York: Lexington Books, 1988.

[3] A. Pinizzotto and Finkel, "Criminal personality profiling: an outcome and process study," *Law and Human Behavior*, vol. 14, no. 3, pp. 215–233, June 1990.

[4] C. Salfati and D. Canter, "Differentiating stranger murders: Profiling offender characteristics from behavioral styles," *Behavioral Science and the Law*, vol. 17, pp. 391–406, 1999.

[5] C. Salfati, "Profiling homicide: A multidimensional approach," *Homicide Studies*, vol. 4, pp. 265–293, 2000.

[6] R. Cowell, "Introduction to inference for bayesian networks," in *Learning in Graphical Models*, M. Jordan, Ed., 1998, pp. 9–26.

[7] P. Santtila, H. Häkkänen, D. Canter, and T. Elfgren, "Classifying homocide offenders and predicting their characteristics from crime scene behavior," *Scandinavian Journal of Psychology*, vol. 44, pp. 107–118, 2003.

[8] C. Salfati, "Offender interaction with victims in homicide: A multidimensional analysis of crime scene behaviors," *Journal of Interpersonal Violence*, vol. 18, no. 5, pp. 490–512, 2003.

[9] C. Salfati and L. Kucharski, *The Psychology of Criminal Conduct. In J. Trevino and S. Fuarino (Eds.), The common subject of crime: A multi-disciplinary approach*. Anderson Publishing, In Press for 2005.

[10] F. Jensen, *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[11] D. Heckerman, "A bayesian approach to learning causal networks," *Technical Report MSR-TR-95-04*, pp. 1–23, May 1995.

[12] G. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.

[13] S. Ferrari and A. Vaghi, "Sensor modeling and feature-level fusion by bayesian networks," *Journal of smart Structures and Systems*, vol. 1, no. 1, pp. 1–9, November 2004.

[14] R. Cowell, "Advanced inference in bayesian networks," in *Learning in Graphical Models*, M. Jordan, Ed., 1998, pp. 27–50.

[15] C. Salfati, "Greek homocide, a behavioral examination of offender crime-scene actions," *Homicides Studies*, vol. 5, no. 4, pp. 335–362, November 2001.

[16] C. Salfati and F. Dupont, "Canadian homicide: An investigation of crime scene actions," *Homicide Studies*, In Press for 2005.

[17] D. Heckerman, D. Geiger, and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[18] R. Robinson, *Counting unlabeled acyclic digraphs. In C.H.C. Little(Ed.)* Lecture notes mathematics, 622: Combinatorial mathematics V. Springer-Verlag, 1977.

[19] D. Heckerman, "A tutorial on learning with bayesian networks," in *Learning in Graphical Models*, M. Jordan, Ed., 1998, pp. 301–354.

[20] K. Murphy, *How To Use Bayes Net Toolbox*. [Online]. Avaliable: http://www.ai.mit.edu/ murphyk/Software/BNT/bnt.html, 2004.