The Three Prisoners Problem Reviseted: issues in the use of Bayesian networks for security applications

J H Pickering (Presenter)

Department of Informatics
School of Computing & Engineering,
University of Huddersfield
Queensgate, Huddersfield, DH1 3DH
Email: j.h.pickering@hud.AC.UK

Z Xu

Department of Informatics
School of Computing & Engineering,
University of Huddersfield
Queensgate, Huddersfield, DH1 3DH
Email: z.xu@hud.AC.UK

***Abstract*—Bayesian networks are a type of causal network used for probabilistic reasoning, which have found wide application in biomedical environments and machine vision. We have considered their application in the realm of security, where behaviour that is deliberately intended to deceive has to be considered. As a first step to the the analysis of this behaviour we have analysed problems in which an one agent provides truthfull, but evasive, information to the other agents.**

The three prisoners problem, and its simpler relation the Monty Hall problem, are classic examples of statistical analysis giving rise to counter intuitive results. In this paper the source of the counter intuitive results is identified as an agent that only releases partial data about the true state of the system. Furthermore the data that is communicated is a function of the identity of the agent requesting the data. Under these circumstances two significant results are demonstrated; first different questioning agents, will arrive at different probability estimates for the same problem. Secondly, although if all the data is requested the estimated probability will converge, the convergence may be nonmonotonic. This means that some questions, truthfully answered, will lead to a less precise probability measurement.

Keywords — Bayesian reasoning; three prisoner problem;

I. Introduction

A reccuring problem in computer science and automation is reasoning about the truth or falsehood of propositions in the absence of conclusive evidence. The common example, from medicine, is to ask if a patient with a set of symptoms has a particular disease. Symptoms such as fever may occur with many diseases. Conversely for any disease few symptoms occur universally, most having some probability of occurring. As a result, for any set of symptoms a possible set of diseases can be found. In such cases a means of weighing the possibilities to indicate how likely they are relative to each other is required. Similar problems occur in image analysis, and interpretation, and sensor fusion. In all cases pieces of data can be assembled to indicate the truth of a propostion: does a picture contain the image of a cat; does the state of a set of engine sensors indicate imminent failure.

Problems of this nature appear to require a means of reasoning in which likelyhood is distributed within a set of possible propositions in accordance with the current evidence. Classical logic is clearly not suitable, as within it a statement must be true or false, and must not change. A host of techniques have been developed in responce to this problem — non-monotonic logic, fuzzy logic, neural networks, probabilistic logic and Bayesian networks — to list a few.

Choosing between a large set of methods is not easy, as there will be many relative advantages, often domain specific, associated with each technique. Our work has concentrated on Bayesian networks, in particular their application to image interpretation. We are interested in using Bayesian networks as a high level tool for interpreting images. By this we mean that the tool will take fragments, such a line segments or textures, generated by image analysis techniques and try to classify the objects or relations in the image. Bayesian networks are attractive in the domain for two main reasons.

1) They are based in probability theory which is a well established field of mathematics, so we do not have to worry about the foundations of the subject.
2) They allow, statistical knowledge about the objects being looked for, to be built in to the network. For example, Dick, Torr and Cipolla in their work on extracting building shapes from photographs [1], were able to include rules such as windows and doors usually being vertically aligned.

One major application of image interpretation is in the security field, where automating the identification of criminal activity in video surveillance material is desirable. One problem that would be encountered in this field is the deliberate obsfucation of material by criminals who, understandably, wish to avoid capture. Although the problem has arisen in the context of video surveillance and Bayesian networks, it is generic to any use of machine reasoning in the presence of hostile agents. Such applications might include attempts to detect money laundering by analysing banking transactions, or the detection of social security fraud by the analysis of claims

records.

As a result, we are interested in the behaviour of Bayesian networks in the case where there is a deliberate attempt to deceive them, by the supply of false evidence. This is a very complicated subject, so we have begun by considering the games in which an honest participant withholds some information from the other players. To this end we have carried out a detailed examination of the three prisoner problem introduced by Pearl [2], in which an honest but evasive jailor provides limited information to prisoners on their fate. The problem is analysed using Bayesian methods, and we demonstrate that despite the honesty of the jailor the convergence of the probabilities is not monotonic.

The next two sections describe the background to our work, covering Bayesian Networks and the Three Prisoner Problem. Our contribution occurs in section IV where we consider adding an observer to the problem, and analyse the development of the observer's estimates of the probabilities. It is from the evolution of the observer's estimates that we draw our conclusions.

II. Bayesian Networks

In this section the basic theory of Bayesian networks is outlined, together with some of their current applications. We concentrate on the basic theory as this will be used in the analysis of the three prisoner problem in the section III.

A. Bayes's Theorem

In his account of statistics [3] Thomas Bayes addressed the problem of the degree to which evidence can be used to support a hypothesis. Bayes assumed that before the evidence arrived there was some estimate of the probability of a hypothesis, called the prior probability. The discovery of evidence should lead to a new probability being calculated, called the posterior probability. Bayes related the posterior and prior probabilities in the equation that bears his name, equation 1.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{1}$$

where
H is the hypothosis E the evidence
$P(.)$ the probability of the parameter
$P(H|E)$ the probability of the hypothosis given the evidence (posterior)
$P(E|H)$ the probability of the evidence given the hypothosis is true

The equation is usefull because it is usually relatively easy to calculate the probability of the evidence on the assumption that the hypothosis is true. This may then be combined with the prior probability of the hypothosis and the probability of the evidence, to generate a new probability the posterior. The probability of the evidence can usually be found by combinatorial techniques, evaluating the number of ways the actual evidence occurred to the total number of ways evidence could occur.

If evidence is to be provided in discrete pieces over a period of time, it may be added to the existing evidence and equation 1 reevaluated. Computationally this can be an expensive procedure, if nothing else it requires that all the evidence be stored. Since Bayes's formula allows new probabilities to be calculated from old, it would appear that a reccursive version of the formula should exist that allows incremental updating. Providing the evidence is independent the following formula applies.

$$P(H|\mathbf{E_n} \wedge e) = P(A|\mathbf{E_n})\frac{P(e|\mathbf{E_n} \wedge H)}{P(e|\mathbf{E_n})} \tag{2}$$

where
$\mathbf{E_n}$ is a vector of n pieces of data
e is a new piece of data.
$\wedge$ is the logical 'and' operator

Equation 2 is central to the use of Bayesian statistics, as it allows new probabilities to be calculated from the current estimates, without reassessing all the evidence that has currently been received. Algorithms for Bayesian reasoning that use only equation 1 tend to be time and space intensive, unless they can find specific features of the domain that allow quick recalculation of the combined probabilities of the new and old evidence.

B. Bayesian Networks

In the previous section we considered Bayes's rule as a means of updating probabilities in the light of streams of evidence. It can also be used to compute probabilities of events that are related by cause and effect rules. Considering a number of discreet variables, statistical relationships between them may be noticed, which relate to the presence of cause and effect relationships between the variables. For example, the probability of a person on a street using an umbrella correlates to the truth of the variable "is raining". The natural way of handling the correlations it to build a joint distribution table in the form of an array, indexed by the states of the variables, containing the probability of each combination of values.

The problem with such an approach is that for most real problems the table becomes very large, also many of the probabilities are zero, as they are the probability of an effect without its cause. To handle these problems an alternative representation can be used, in which cause and effect relationships between variables are represented in a directed graph. The nodes of the graph are the variables, and the edges represent the cause and effect relationships between variables. Each node contains a conditional probability table, describing the probable outcomes of the variable in the node in terms of the values of its parents. We do not allow cyclic cause and effect relationships, and hence the graph must contain no directed cycles, resulting in a directed acyclic graph (DAG). Whithin the graph equations 1 or 2 may be used to update the the individual nodes.

Once set up a Bayesian network may be queried to find the probability of some outcome variables in terms of known

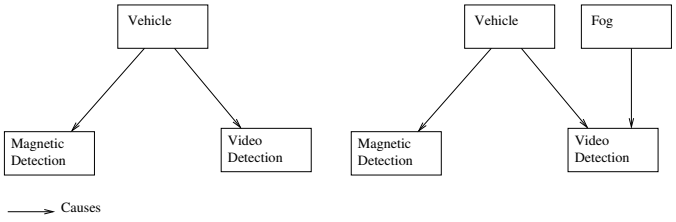

Fig. 1. Detecting a vehicle using a magnetic loop detector and a video camara, taken from [5]. On the left is a simple Bayesian sensor fusion network, in which the probabilities of a magnetic sensor and a video sensor for detecting a vehicle are known. Whenever one or both sensors give a signal we can compute the probability of a vehicle being present. On the right a fog sensor has been added, since fog can effect the performance of a video sensor but not a magnetic sensor, the node for fog is only linked to the video sensor node. The conditional probability associated with the video sensor now include a term for the presence of fog, the whole network will now adapt to the environment as a result.

query variables. There are two ways of doing this: exact inference in which all the probabilities are worked out starting from the query and leading to the outcome; and approximate inference in which the probabilities are approximated by sampling from the network. The details of these techniques are not important to the current paper. Some example of the applications of Bayesian networks follow.

a) Shape Reconstruction: Han [4] has developed a means of reconstructing three dimensional shape from a single image, using Bayesian networks containing prior knowledge about the shapes. Previous approaches to shape reconstruction used

1) Reconstruction 3D shapes from line drawings with human input to fill in the hidden lines
2) Construct a two and a half dimensional map using focus, texture and shading information, then parse this into a 3D model. This requires some prior model information, and is prone to lighting errors.
3) Computing the shape from multiple images, which requires multiple images be available.

Han first segmented a single image then extracted artifacts in the form of graphs of vertices, edges, planes, and curves. A Bayesian network was trained to detect known classes of objects from the graphs and produce a 3D reconstruction of them. Finally a scene graph showing the support relations between the object was developed.

b) Sensor Fusion: In sensor fusion the approach of machine vision is extended with several sensors — not necessarily cameras — being used to identify objects. The simplist case of sensor fusion is to combine two sensors, using the known probabilities of each to produce a correct result. Bayes's rule for conditional probabilities is an obvious way of achieving this. In this case a simple model of the probability of each sensor detecting an object can be built, and used to compute the probability of an object being present for any set of sensor readings. Such a network can also be extended to include background knowledge [5], an example is described in figure 1.

III. The Three Prisoners Problem

In this section we present our analysis of the three prisoner problem, in Bayesian terms. We start, however, with the simpler Monty Hall problem. While neither of these problems involves active deception, in both cases the agent providing the information is always truthfull, they do involve the concealment of vital data. It is our hope that analysis of these problems will provide a stepping stone handling deception by an overtly hostile agent.

A. The Monty Hall Problem

The name comes from the host of the television game show "Let's make a deal", widely syndicated in the USA. In the finale of the show a contestant was offered the chance to win a car. The vehicle was hidden behind one of three doors and the contestant has to guess which door hid the car. At this stage the host opened one of the two remaining doors, not concealing a car, and and then invited the contestant to choose between sticking with his or her original choice, or swapping for the remaining closed door. The problem is, should the contestant swap? The solution was published in [6].

The common and incorrect answer is as follows, there were three doors hence a one third probability of finding the car. Now one door has been removed the probability must be one half, so changing makes no difference. The problem with this analysis is that the show's host knows which door the contestant has chosen, which door hides the car, and must not end the show by revealing the car, when selecting the which door to open.

A look at possible alternative games is rewarding. Self evidently if one door was opened, revealing no car, before the contestant chose a door, the probability of the contestants choice being correct would be one half. Next consider the contestant choosing first and the host opens one of the two remaining doors at random. In one third of cases the car is revealed and the game ends, but in the remaining two thirds the the probability of the contestant being correct is one half.

The correct analysis of the problem is to observe that in the initial choice made by the contestant, the probability of selecting a door hiding the car is one third, hence the probability of the car being behind one of the other doors is two thirds. The crutial component of the analysis is the behaviour of the host. The host knows which door has been chosen by the contestant and which door hides the car. Two thirds of the time the contestant has chosen the wrong door, one of the remaining two doors hides the car and the host opens the other one. Alternatively if the contestant has chosen the door hiding the car, the host open one of the two remaining doors at random. Nothing the host has done alters the one third probability of the contestant's first choice being right, so the probabilities are now one third the contestant being correct, two thirds the remaining door hiding the car. Clearly the contestant should swap.

To express the problem in Bayesian form it must first be represented in the form of logical propositions.

$D =$ the door selected by the contestant hides the car

$I =$ an arbitrary door not selected by contestant is opened

Next the propostions are substituted Bayes's formula 1.

$$P(G|I) = \frac{P(I|G)P(G)}{P(I)} \tag{3}$$

Finally values must be found for the probabilities on the right hand side of equation 3. Clearly $P(G) = 1/3$ since this is the prior probability of the contestant selecting a door hiding a car, also $P(I|G) = 1/2$ since there are two doors and the choice was arbitrary. This leaves the value of $P(I)$ to be found. The host only makes the choice of which door to open after the contestant has selected a door, this leave the host with a choice of two doors. Since the door in the proposition D was arbitrary, the probability $P(I) = 1/2$. Inserting these values into equation 3, give the correct answer 1/3.

The error that leads to the answer 1/2, is assuming that $P(I) = 1/3$, in which case equation 3 yields $P(G|I) = 1/2$. This is wrong because it uses the prior probability of selecting a door as the probability for the host choosing any door, when one door has already been removed from the choice by the contestant. This highlights the fact that prior probabilities must be calculated relative to the current state of the system, not the initial state.

B. Three Prisoner Problem

The three prisoner problem is an extension of the Monty Hall, problem introduced by Pearl [2]. In the three prisoner problem a tyrant has imprisoned three people (Alice, Bob and Charlie), under the condition that, in the morning one will be executed, but none are to be told who is to be executed untill the execution itself. One prisoner, Alice, asks to see the jailor and, pointing out that at least one of the other two prisoners must be released, asks to be told the name of one fellow prisoner who will be released. The jailor consents and says Charlie.

The question is should Alice swap places with Bob? The answer, following the above analysis of the Monty Hall, problem is no. Forming the question in propositional form.

$I =$ Jailor tells Alice, Bob will be released

$A =$ Alice will be executed

Substitute into 1.

$$P(A|I) = \frac{P(I|A)P(A)}{P(I)} \tag{4}$$

The probability of the jailor saying Bob if Alice is to be executed $P(C|I)$ is 1/2, since, if Alice is to be executed, the jailor has a random choice, and we assume no bias between Bob and Charlie. The probability of Alice being executed $P(A)$ is 1/3, again assuming a random choice by the tyrant. This only leaves the probability of the jailor naming Bob $P(I)$, since Alice is asking the question, there are only two names that can be given, so the probability is 1/2. Inserting these numbers in equation 4 the probability of Alice being executed is 1/3, while Bob has a chance of 2/3.

One interesting corollary of the problem is that the probabilities are not independent of the person asking the question. If Bob asked the jailor the same question and received the answer Charlie, he would compute his probability of being executed at 1/3 and give Alice 2/3. Since the identity of the prisoner asking the question is a parameter the jailor has to take into account in answering this is inevitable.

C. Role of the Jailor

Given the importance of the jailor's behaviour, this role is now examined. The jailor has two directives.

1) The jailor must not tell a prisoner wheather they will be executed or released.
2) The jailor should be honest when answering the question, "Name one person who will be released?".

The algorithm is simple. If the person asking the question is to be released, then one of the other prisoners is to be executed, so the jailor names the other one. Alternatively if the person asking the question is the one to be executed, select the name of one of the other two at random. In effect the jailor is a communication channel, transmitting data on the current state of the jail to the prisoner asking the question. The identity of the prisoner asking the question is a parameter used in selecting the data.

Using this approach to the problem, we can give a frequentist account of the statistics, using the more humane, but less melodramatic, senario in which the selected prisoner has to do that day's cleaning. Now consider Alice asking who will not be cleaning the prison tomorrow. Approximately one in every three days she will get the answer Charlie, because Bob will be doing the cleaning. However, on approximately one in every six days she will get the answer Charlie, because she is doing the cleaning and the jailor chose to name Charlie, the other one in six she will be cleaning but the jailor names Bob. This gives the probability of 1/3 of doing the cleaning we found above in III-B. Symmetrical results apply if Bob or Charlie ask the question.

IV. An Observer of Three Prisoner Problem

In the preceeding sections it was assumed that the prisoners had no means of communicating with each other, and therefor could not compare the answers given by the jailor. Here the limitation is addressed, not by allowing the prisoners to communicate, but by adding an observer aware of all questions and answers to the problem. As before the analysis begins with a representation of the problem in propositional form.

$I_1 =$ Jailor tells Alice, Bob will be released

$I_2 =$ Jailor tells Bob, Alice will be released

$I_3 =$ Jailor tells Charlie, Bob will be released

$A =$ Alice will be executed

We will consider two scenarios, both beginning with Alice questioning the jailor and receiving the reply Bob. This allows the analysis presented in section III-B to be reused for the probabilities after the jailor has answered the first question.

Scenario One: In the first scenario, Bob also questions the jailor and is told Alice. Combinatorially it is trivial for the observer to determine that Charlie is the unlucky prisoner to be executed. However we wish to do the calculation using Bayesian methods. To do that we substitute the propostions into the Bayes's recursive updating formula, equation 2.

$$P(A|I_1 \wedge I_2) = P(A|I_1)\frac{P(I_2|I_1 \wedge A)}{P(I_2|I_1)} \tag{5}$$

Clearly if Alice is to be executed the answer I_2 is impossible, since the jailor does not lie. This makes the term $P(I_2|I_1 \wedge A)$ equal zero, and hence the overall probability is zero, which is in accordance with the combinatorial result.

Scenario Two: In the second scenario, the second prisoner to question the jailor is Charlie and he receives the answer, Bob. As before the probabilities are substituted in Bayes's law.

$$P(A|I_1 \wedge I_3) = P(A|I_1)\frac{P(I_3|I_1 \wedge A)}{P(I_3|I_1)} \tag{6}$$

The first term, the probability of Alice being executed given the jailor told her Bob would be released, was calculated to be 1/3 in section III-B. The probability of the jailor telling Charlie that Bob will be released, given Alice will be executed is one. Hence $P(I_3|I_1 \wedge A)$ is equal to one, the occurrence of proposition I_1 in the term making no difference. The remaining probability is that of propostion I_3 occurring given I_1 has occurred $P(I_3|I_1)$. It is important to note that this does not include the probability that Charlie will ask a question after Alice. The order in which the prisoners question the jailor is given as part of the problem, and is not involved in the statistics.

In calculating $P(I_3|I_1)$ it is important to observe the data that has already been obtained from I_1. Bob is not going to be executed and, as Alice asked the question she has a 1/3 chance of being executed, while Charlie the current questioner, has a 2/3 probability. There are two ways in which the answer Bob can occur, the probability is the sum of the individual probabilities for each option. First Charlie is to be executed, which occurs with a probability of 2/3. In this case the jailor has a random choice of Alice or Bob, so Bob is selected with overall probability $2/3 \times 1/2 = 1/3$. The other alternative is that Alice is to be executed and as a result the jailor must reply "Bob". The probability of this is clearly the probability of Alice being executed 1/3. The overall probability is the sum of the probabilities of these two alternative branches, 2/3. Inserting these figures into equation 6 gives the result that the probability of Alice being executed has risen to 1/2, while that of Charlie has dropped to 1/2.

In this scenario a third question is needed in which Bob asks the jailor for the name of one person who will be released. This question will act as a decider producing a probability of one, for Alice or Charlie. If Charlie is the unlucky victim, one important feature of the probabilities becomes clear, they do not monotonically converge. This is illustrated in table I, in which the probabilities of the three prisoners are listed, with an added final propostion $I_4 =$ the jailor tells Bob, Alice will be released. In this case the probability of Charlie being executed starts at 1/3, raises to 2/3, drops to 1/2, before finally rising to one.

Data	Alice	Bob	Charlie
	1/3	1/3	1/3
I_1	1/3	0	2/3
I_2	1/2	0	1/2
I_3	0	0	1

TABLE I
THE CHANGING PROBABILITIES IN REPONCE TO QUESTIONS ASKED OF THE JAILOR.

It is inevitable that when all three prisoners have questioned the jailor the identity of the victim will be found, the combinatoric nature of the problem guarantees that. The system of Bayesian estimators will converge to that answer, within the floating point accuracy of the computational machinery on which it is implemented. However, there is no guarantee that the convergence will be monotonic, for some steps in the process the probabilities may relax.

V. CONCLUSIONS

Most analysis of the behaviour of Bayesian networks has been based on the assumption of relatively 'well behaved' evidence. This is natural as most applications have been in the physical, biological or social sciences, in which cases the the evidence may be subject to errors, possibly systematic, but not deliberate evasion or deception. Issues relating to deliberate evasion or deception will have to be faced if Bayesian (or any other) methods are applied in the field of security. The analysis of simple games can be used to examine some of the problems encountered in this field.

The analysis of the three prisoner problem has been involved, and has given rise to several observations about the behaviour of a recursively updated Bayesian system.

1) In systems representing 'games' in which an agent provides information on the current state of the game (the host or jailor in the examples), it is best to interpret this agent as a communications channel reporting the current state.
2) If the identity of the agent requesting information from the communications agent is a parameter, used to decide what information is returned, then the probabilities derived by different agents may not be equivalent.
3) Although the Bayesian estimators will eventually converge to the correct answer, assuming an honest communications agent, the convergence will not be monotonic. At some stages the correct proposition may become less lightly.

The three prisoner problem may be a rather artificial example, but it is indicative of a range of problems relating to

games. It has some similarities to an investigation, in which witnesses may limit their answers to avoid incriminating or embarrassing themselves. It also provides a stepping stone to the more realistic and complex domains in which agents are actually deceitful. Finally, it illustrates the dangers of relying on assumptions about the behaviour of statistics derived from physical, biological or social science, in the fields of games and security.

REFERENCES

[1] Anthony R. Dick, Philip H. S. Torr, and Roberto Cipolla. A bayesian estimation of building shape using mcmc. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part II*, pages 852–866, London, UK, 2002. Springer-Verlag.

[2] J Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[3] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 1764.

[4] Feng Han and Song-Chun Zhu. Bayesian reconstruction of 3d shapes and scenes from a single image. In *HLK '03: Proceedings of the First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, page 12, Washington, DC, USA, 2003. IEEE Computer Society.

[5] M Junghans and H Jentschel. Qualification of traffic data by bayesian network data fusion. In *10th International Conference on Information Fusion, 2007*, pages 1–7, July 2007.

[6] Steve Selvin. On the monty hall problem (letter to the editor). *American Statistician*, 29, 1975.