UAI-02 Trip Report

Marco Valtorta

August 15, 2002

1 Introduction

The eighteenth conference on uncertainty in artificial intelligence (UAI-02) took place in at the University of Alberta in Edmonton, Alberta, Canada, from August 1 to August 4, 2002. This was the fifth time I attended UAI: I had attended 1993 (Washington, D.C.), 1995 (Montreal), 1996 (Portland), and 1997 (Providence). I had papers in 1993 (with Moninder Singh—student paper award winner), 1995 (with Young-Gyun Kim) and 1998 (with Mark Bloemeke), and served on the program committee from 1996-99 and in 2001.

Other notable conferences in the area are the biennial ECSQARU (European Conference on Symbolic and Quantitative Approaches to Reasoning under Uncertainty—the seventh one will be in July 2003 in Aalborg) and the biennial Artificial Intelligence and Statistics (AIS) workshops—the ninth one will be in January 2003 in Key West. Papers on this topic are also presented at more general conferences (typically, IJCAI, AAAI, ECAI, NIPS, IEA/AIE, FLAIRS, and some SPIE workshops). Over time, UAI has become more narrowly focused than either ECSQARU and AIS, concentrating almost exclusively on graphical probabilistic normative approaches to reasoning under uncertainty (Bayesian networks, Markov networks, chain graphs, influence diagrams, and related formalisms).

Moreover, there is no journal whose scope closely approximates that of UAI. The closest journal is the *International Journal of Approximate Reasoning*, which has a much wider scope, with a majority of papers on fuzzy and possibilistic approaches. and this makes the conference and its proceedings (published by Morgan Kaufmann since 1991) more valuable. Many papers from the UAI conferences are available on line, following links starting from the association for uncertainty in AI site (www.auai.org). I have all proceedings from 1991 on in my office. (I think that Juan has them too.)

2 Venue

Edmonton, Alberta, Canada, is a city of about 800,000 people (metropolitan area), built around the North Saskatchewan river. The river area is a very large park. The day I arrived, July 31, was the coldest on record for the time of year—I was told it snowed just outside the city. The accommodations in a dorm at the University of Alberta were good and unexpensive. The conference site itself, the Telus continuing education center at the university of Alberta, was modern and convenient, and just the right size for UAI. Several PCs and ethernet connections were provided. A subway system connects the University, just south of the river, to the city center, north of the river, where the conference banquet was held. An extensive system of covered or underground walkways criss-crosses the city. The conference chair noted that the facilities and food for breaks, conference banquet and conference reception were cheaper of comparable ones in the U.S. This allowed more student travel support than in the recent past.

The most popular tourist destination in town is West Edmonton Mall, billed as the largest multi-purpose mall in the world, which in addition to the usual shops, restaurants, and movie theaters, includes a skating rink, two large waterparks, and an amusement park with full-scale roller coaster, all under one roof.

3 General Data

There were 217 registered participants, approximately the same as in 2001. Of these, 90 were students, up from last year. Of the 192 papers submitted, 66 were accepted, for an acceptance rate of 34%. Papers were submitted from 25 countries (by indicated affiliation). By far, most of the papers came from the U.S. The UK had the second largest number of submitted papers, and Israel had the second largest number of accepted papers. The program co-chairs, Adnan Darwiche of UCLA and Nir Friedman of Hebrew University in Jerusalem showed charts indicating that the acceptance rate was slightly lower than last year, continuing a trend started (at least) in 1995, when it was 55%, with a roughly constant number of accepted papers. Twenty-six papers were presented in plenary sessions; the remaining forty were presented in two poster sessions, preceded by 2.5-minute "teaser" talks. As it is customary for UAI, there were no parallel sessions. Each paper was reviewed by three of the 78 program committee members (or, occasionally, auxiliary reviewers), with 14 area chairs responsible for the quality of the reviews. The program co-chairs made the final decision in a multi-day meeting in Los Angeles ("crusing coffee shops").

4 Tutorials

August 1 was devoted to four tutorials or, more precisely, a full-day course on advanced topics: the tutorials were clearly designed for workers in the field, rather than for newcomers. Titles, instructors, and highlights follow:

- Statistical Methods in Natural Language Processing, Michael Collins, AT&T.

- Markov Decision Processes, Craig Boutillier, University of Toronto. Highlight: results on how to exploit factorization of Conditional Probability Tables (CPTs) in Dynamic Bayesian Networks (DBNs) when performing Bellman backup, by approximation.

- Randomization and Rational Decision Making in Optimization, Carla Gomes and Bart Selman, Cornell. Highlight: Review of work in phase transition, "where the hard problems really are."

- Bayesian Networks for Genetic Linkage Analysis, Dan Geiger, Technion. Highlights: Transformation of pedigrees to Bayesian networks, in ways similar to that done by Robert Cowell and our own graduate student Xizhou Feng for forensic applications; the networks are used to estimate the recombination parameter; conditioning is done on top of the junction tree algorithm, in a classic space for time tradeoff; experiments show that implemented system (called "Superlink") is much faster than Fastlink, Vitesse, and Genehunter and can deal with much larger problems.

I have a copy of the tutorial slides as handed out. Updated slides, as actually used in the tutorials, were promised but I have not seen them posted yet.

5 Invited Talks

There were four invited talks and a banquet talk.

- "Particle Filters in Robotics," Sebastian Thrun, Carnegie-Mellon. This was an entertaining presentation, with plenty of engaging visual materials. The author started off by stating: "Judea Pearl's work has changed my life and changed robotics." He described various approaches to the localization problem: Kalman filters, multiple-hypothesis Kalman filters, Histograms, and Particles. He described the problem of simultaneous localization and mapping, which he called "the hardest perceptual problem in robotics," mentioned that the classical Kalman filter approach to this is improved by the insight that robot poses make landmark observations conditionally independent. For the initiated: robot positions ("poses") d-separate landmark observations. This insight allows a dramatic reduction in complexity and was implemented in the FastSLAM system.

- "The Problem of Thinking Too Much," by Persi Diaconis (Stanford). This was the conference banquet talk, by a famous mathematical statistician. It was entertaining. Diaconis used the process of building a progressively more complex model of coin flipping as his main example.

- "Using Statistical Models of Shape in Medical Image Analysis," by Eric Grimson (MIT). This talk was mainly on how to assist a surgeon by providing in real time realistic images of the patient. Grimson stated listed four sources of information for shapes: manual capture of shapes, statistical correction of variations in magnetic fields, atlases of anatomy (to compensate for local variations), specific shape models. The resulting system was evaluated in over 350 cases (of brain surgery). A main result of use of the image system was a 50% reduction in the time needed to perform the operation.

Grimson also described his current work on computational anatomy: use images to identify anatomical differences between normal and sick patients.

- "Decision Principles to Justify Carnap's Updating Method and to Suggest Corrections of Probability Judgements," by Peter P. Wakker (Maastricht University). This interesting talk by an economist included a new (at least for me!) technique for building utility functions that is tolerant of certain types of errors in assessing probabilities.

- "Advances in Boosting," by Robert E. Schapire (AT&T Bell Labs). Boosting has been around for more than ten years: I remember when Schapire (as a Ph.D. student of Ron Rivest at MIT) presented his initial results on boosting at COLT-89 at UC-Santa Clara. There was excitement about this, but it took a few more years to make the result practical. The idea is to take a weak classifier and improve its performance by training it (multiple times) on samples on which it fails, and taking a linear combination of the classifiers built each time as the final classifier. The most used variation of boosting is AdaBoost. The following long quote summarizes the algorithm and is from a paper in the proceedings (p.447). Similar papers are available at Schapire's web site (`www.research.att.com/~schapire`).

> AdaBoost calls a given *weak* or *base learning algorithm* repeatedly in a series of rounds $t = 1, \ldots, T$. One of the main ideas of the algorithm is to maintain a distribution or set of weights over the training set. The weight of this distribution on training example i on round t is denoted $D_t(i)$. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the base learner is forced to focus on the hard examples in the training set.
>
> The base learner's job is to find a *base classifier* $h_t : X \to \Re$ appropriate for the distribution D_t. In the simplest case, the range of each h_t is binary, i.e., restricted to $\{-1. +1\}$; the base learner's job is then to minimize the *error*
>
> $$\epsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i].$$
>
> Once the base classifier h_t [which only needs to be correct more than half of the time—it can be quite *weak*!] has been received, AdaBoost chooses a parameter $\alpha_t \in \Re$ that intuitively measures the importance that it assigns to h_t. ...

The distribution D_t is then updated using [the rule

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$

where Z_t is a normalization factor to insure that D_{t+1} is a distribution]. The *final* or *combined* classifier H is a weighted majority vote of the T base classifiers where α_t is the weight assigned to h_t, [i.e.:

$$H(x) = \mathrm{sign}(\sum_{t=1}^{T}(\alpha_t h_t(x)))$$

Schapire presented some results concerning training and generalization errors of AdaBoost. He concluded his talk by giving two applications that required extensions of the basic algorithm: trading agent for the 2001 Trading Agent Competition, and human-computer spoken dialogue system (`naturalvoices.att.com`). In the first case, AdaBoost had to be extended to provide distribution information (estimation of uncertainty) in addition to class information. In the second case, the challenge was to incorporate prior knowledge and active learning to compensate for the lack of data.

6 Papers

This is a highly subjective discussion of some papers.

6.1 Best Paper Award Winner

The best paper award winner was "A New Class of Upper Bounds on the Log Partition Function," by Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky (MIT). See `http://ssg.mit.edu/~mjwain/Talks/d_snow02_present.ps` for slides of a talk given in April that are very close to the ones Wainwright used at UAI. What is the partition function? It is a sum of cliques potentials over all configurations of a factored probability distribution, which gives the normalization constant needed to compute probabilities for a factored distribution. For a particular instantiation of variables, the partition function can be interpreted as the probability of the evidence. This is *not* a difficult problem in practice in Bayesian networks when the junction tree algorithm is used (because junction tree has almost invariably small cliques/treewidth), but it becomes difficult in applications of graphical models (e.g., grids with high connectivity) in statistical physics. Wainwright used the Ising model as an example.

Flame On Personally, I feel that the usefulness of this paper *in UAI* is limited. It is based on the assumption that the junction tree procedure has "limited practical use" [quote from p.47 of Wainwright's dissertation, `http://ssg.mit.edu/~mjwain/PhD_thesis/Final2_Phd_May30.ps`], which is empirically false. I would characterize this paper as a (partially successful at best) attempt by researchers in statistical physics to get leverage from techniques that are useful in their field to publish in the UAI area. Unfortunately, there were several papers of this kind at this year's UAI. *Flame Off*

6.2 Best Student Paper Award Winner

The best student paper award winner was "Generalized Instrumental Variables," by Carlos Brito (and Judea Pearl). The *identification problem* is to determine whether assumptions encoded in a Bayesian network (possibly with hidden variables) are sufficient to allow the assessment of causal effects from non-experimental data. In other words, we want to compute the probability that by *doing* X we see Y when we only have data of the form $P(X|Y)$ and a model of the conditional independencies governing the joint probability distribution $P(U)$, where U is a set of variables including X and Y.

There is no general solution of the identification problem, but many kinds of models are known to be identifiable (thus refuting the slogan that only controlled experimental studies allow the identification of causal effects). The main method for solving the identification problem is based on the notion of instrumental variables.

An *instrumental variable* Z relative to a cause X and an effect Y is, intuitively, a variable such that all correlation between Z and Y must be intermediated by X. Intuitively (again–sorry!), if an instrumental variable exists, we can use it to set up a virtual experiment. The paper describes a generalization of the method of instrumental variables for solving the identification problem.

6.3 Approximate Inference

- "Iterative Join Graph Tree Propagation," by Rina Dechter, Kalev Kask, and Robert Mateescu (presenter), U.C. Irvine. A theme that was explored in several papers was that of IBP (Iterative Belief Propagation). The idea is to apply Pearl's (1983!) belief propagation (BP) algorithm, which is exact for trees and polytrees to DAGs, *and iterate.* Quoting from the paper, p.128:

> [I]n recent years, [BP] was shown to work surprisingly well in many applications involving networks with [undirected] loops [i.e., DAGs], including turbo codes [McEliece and Cheng, "Turbo Decoding as an Instance of Pearl's Belief Propagation Algorithm," *IEEE Journal on Selected Areas in Communication*, 1997], when applied iteratively[1]. While there is still little understanding as to why and when Iterative Belief Propagation (IBP) will work well, some progress was made recently in understanding the algorithm's behavior, showing that it converges

[1] Also see: K.P. Murphy, Y. Weiss, and M.I. Jordan. "Loopy Belief Propagation for Approximate Inference: An Empirical Study. *UAI-99*, pp.467-475.

to a stationary point of Bethe energy, thus making connections to variational approaches to approximate inference [references given].

A serious drawback of IBP is that it does not have an anytime behavior, i.e., it does not give better results as it is allowed to run longer: once IBP converges to a bad approximation, it stops improving. Note that there are anytime algorithms for belief update: stochastic simulation methods (which have a long history) and abstraction-based methods, such as Dechter's own mini-bucket and mini-clustering algorithms. Dechter and al. propose an anytime variant of IBP called Iterative Join Graph Propagation (IJGP(i)), controlled by a parameter i, which "allows the user to control the tradeoff between time and accuracy." IJGP(i) is similar to an algorithm proposed by Yedidia et al [NIPS-01], called "Generalized Belief Propagation" (GBP). The experimental results indicate that IJGP(i) is superior to IBP and to mini-clustering for $i \geq 2$. Nobody knows why (yet).

- "Loopy Belief Propagation and Gibbs Measures," by Sekhar C. Tatikonda (presenter) and Michael I. Jordan (UC Berkeley). Loopy Belief Propagation (LBP) is an another name for IBP (see above). This paper presents a sufficient condition for convergence of IBP. I suggested to the author (indirectly) that he consider using the techniques used by Csizar (1975) to show the convergence of IPFP.

- "MAP Complexity Results and Approximation Methods," by James D. Park (UCLA). Consider a Bayesian network, evidence on some of its nodes, and a set of nodes of interest.

  - The MPE (Most Probable Explanation) problem consists of finding the most likely state for *all* of the nodes. This problem is NP-complete [Shimony, 1994].
  - The Pr problem consists of computing the probability of the evidence. This problem is PP-complete [Roth, 1996].
  - The MAP (Maximum A Posteriori assignment) problem consists of finding the configuration of the nodes of interest that maximizes the posterior probability of all nodes. This problem is in NP^{PP}. This was conjectured by Michael Littman and is proved in the paper.

  Intuitively, this is because MAP requires both belief updating and optimization, and one cannot interleave maximization and summation steps at will; typically, maximization is left for last; some interleaving is sometimes possible, but not in the worst case.

  The author notes that all previous algorithms for approximating MAP use an exact belief update engine and an approximate optimization engine. He proposes a new algorithm that uses IBP (What else?) for belief update and reports good experimental results.

In the Q&A session, Park noted that all his results hold even when all probabilities are positive.

6.4 Learning

- "A Bayesian Network Scoring Metric That Is Based on Globally Uniform Parameter Priors," by Mehmet Kayaalp (presenter) and Greg F. Cooper (University of Pittsburgh). The authors reported the results of Bayesian learning algorithms using the K2 and BDeu scores. The experiments involve generating cases by simulation from a Bayesian network (to be considered the gold standard). The cases are then input to a learning program that uses the K2 or BDeu scores, respectively. The main result is that BDeu in many cases produces networks that are much denser than the gold standard. The authors also propose a new score (Globally Uniform, "GU"). The problem with the new score is that it does not decompose on the structure of the network: if one wants to, say, compare two networks that differ only in the orientation of one edge, one must compute two scores from scratch. K2 and BDeu do not have this problem.

  These results provide a kind of vindication for the old K2 score. K2 may give different scores to networks that represent the same constraints on the conditional independence relation for the domain of interest. This is a problem, because data alone cannot (up to a point...) distinguish between networks that encode the same conditional independence relations ("equivalent" networks). BDeu was introduced precisely to overcome this problem. However, the authors' results show that BDeu has more serious problems than K2. The authors' interpretation (shared by other participants) is that this is a situation in which the uniform prior used by BDeu behaves unexpectedly. "The appropriateness of using uniform priors has been an area of discussion among Bayesian statisticians, ever since the concept was introduced by Bayes. ... A uniform prior is often found appropriate for statistical inference when it lets the data, rather than the priors, dominate posterior probabilities (Box and Tiao 1973, pp.21–23). The unexpected behavior of the BDeu scoring metric may be associated with the dominance of its prior over its likelihood. In such situations, when no prior knowledge is available, GU seems to offer a robust alternative to the BDeu metric." Several other participants felt that the way to go is to somehow "adjust" K2 and forget about BDeu and GU.

  On a personal note, the reimplementation of the Bayesian network learning algorithm CB [Singh and Valtorta, 1993 and 1995] by Bing Xia (Visual CB, M.S. thesis defended in April 2002), gives the user a choice of the K2 and BDeu scores. (The original CB used the K2 score only.) Bing reported much worse results using BDeu than using CB, and precisely because the reconstructed networks were too dense. I must confess that I was somewhat skeptical of these results, and even suspected an error in the implementation. Now I know better: it is comforting to see that this behavior is confirmed by the experience of others. One reason why I asked Bing to implement BDeu also, was a suggestion by Greg Cooper, coauthor of this paper and original coinventor (with Herskovits) of the K2 score. Sometimes we give up too easily on our own good ideas!

- "On the Construction of the Inclusion Boundary Neighborhood for Markov Equivalence Classes of Bayesian Network Structures," by Vincent Auvray (presenter) and Louis Wehenkel (University of Liège), and "Finding Optimal Bayesian Networks," by Max Chickering and Christopher Meek (Microsoft).

  An *essential graph* (also called a partially directed DAG or pdag) E is a way to represent a class of DAGs that encode the same conditional independence relation. The notion is based on a theorem of Verma and Pearl (1990) that states that two DAGs are equivalent iff they have the same skeleton (underlying undirected graph) and set of v-structures.

  Let I(E) be the independence relation encoded by the essential graph E. E' belongs to the inclusion boundary of E if $E' \neq E$ and either

  1. $I(E) \subset I(E')$ and there is no E'' verifying $I(E) \subset I(E'') \subset I(E')$, or
  2. $I(E') \subset I(E)$ and there is no E'' verifying $I(E') \subset I(E'') \subset I(E)$.

  The first paper shows that the inclusion boundary relation defines a connected space and gives an algorithm to construct the inclusion boundary of an essential graph incrementally. For scores that factor, such as K2 and BDeu, one can use the algorithm to evaluate the score of a neighbor in the inclusion boundary without actually constructing the inclusion boundary. This leads to an efficient greedy algorithm for finding Bayesian networks of highest score. Unfortunately, the boundary of an essential graph can be very large (exponential in the number of vertices of the essential graph).

  The second paper shows how to use similar results in an actual learning algorithm (called GES, for Greedy Equivalence Search), which uses a transformational characterization of the inclusion relation for Bayesian networks. The characterization was conjectured by Meek in his 1997 dissertation, proved correct in a special case by Kočka in a UAI-01 paper, and proved correct in the general case by Chickering in 2002. (The actual search is done in the space of Bayesian networks, not essential graphs.) The algorithm is proved correct under mild assumptions (mainly, compositionality), and (surprisingly) does not require the assumption of faithfulness, i.e., that the independence relation consistent with the distribution from which the data is generated is DAG-isomorph.

  Auvray told me that he did not implement an algorithm, because one of the reviewers communicated to him that Chickering and Meek had already implemented an algorithm based on similar results. Also, Tomas Kočka and several coauthors had related results in two UAI-01 papers.

- "Discriminative Probabilistic Models for Relational Data," by Ben Taskar (presenter?), Pieter Abbel, and Daphne Koller (Stanford). This was one of a couple of papers on learning and the web. The models learned are undirected graphs (Markov nets), rather than directed graphs (Bayesian nets). This eliminates the need for introducing hidden variables ("potential links," Getoor et al., IJCAI-2001), thus reducing the size of the model w.r.t. the directed model. However, learning is harder in principle, because there is no closed form solution for the maximum likelihood or MAP score of a Markov net. Another interesting aspect of this work is the extension of Markov network to a relational setting: "A relational Markov network (RMN) specifies a conditional distribution over all of the labels of all of the entities in an instantiation given the relational structure and the content attributes. ... Roughly speaking, [an RMN] specifies the cliques and potentials between attributes of related entities at a template level, so a single model provides a coherent distribution for any collection of instances from the schema."

6.5 Applications

- "Coordinate: Probabilistic Forecasting of Presence and Availability," by Eric Horvitz (presenter), Paul Koch, Carl M. Kadie, and Andy Jacobs (Microsoft Research, Seattle).

  This is yet another very good paper on applied of decision theory by Eric Horvitz, following earlier UAI papers on the VISTA system for spacecraft monitoring, the Lumière project (which became the "paper clip" in MS Office), etc. The focus of this paper is on how to model the probability of attending a meetings and the cost of interrupting a meetings. These models are used to help in automated scheduling and should lead to an improved version of the MS Priorities system for email message rating, which started shipping last year. Eric told me that Bill Gates himself is very supportive of this work.

- "Adaptive Foreground and Shadow Detection in Image Sequences," by Yang Wang (presenter) and Tele Tan (Laboratories for Information Technology, Singapore). A technique for distinguishing foreground moving objects in image sequences is presented. Applications are in surveillance, HCI, and content-based image coding. A Bayesian network is used to integrate edge, background, and intensity information. The technique is not real-time, and the authors want to speed it up.

- "Monitoring a Complex Physical System Using a Dynamic Bayesian Net," by Uri Lerner (presenter), Brooks Moses, Maricia Scott, Sheila McIlraith, and Daphne Koller (Stanford), and "Continuous Time Bayesian Networks," by Uri Nodelman, Christian R. Shelton (presenter), and Daphne Koller (Stanford).

The authors want to monitor the NASA Reversed Water Gas Shift Systems (RWGS), which produces oxigen from carbon dioxide. Since RWGS may be used on Mars, it needs to operate without human intervention. A major challenge is that some of the components of RWGS only have approximate and non-linear models. The paper reviews the classic *Extended Kalman Filter* method (linearize around the operating point, then use the Kalman filter) [Note for Juan: the reference given is *Estimation with Application to Tracking and Navigation*, by Bar-Shalom et al., Wiley 2001.]. They also describe Uhlmann and Juler's *Unscented Filter*, which uses a non-linear (second-order) approximation instead.

The authors use a *Dynamic Bayesian Network* with continuous and discrete nodes. (Discrete nodes do not have continuous nodes as parents, as usual.) The model is (first-order) Markovian and stationary, and it consists of nodes from two time slices with the associated conditional probability distribution. There are 228 continuous variables and 64 discrete ones. All the values of the discrete variables are known. This is a strong assumption: for example it means that it is known whether a sensor is broken, and the authors intend to relax the assumption in future work.

There are several challenging modeling problems. Perhaps the most interesting is due to the different time granularities involved: speed of sound, gas flows (1 Hz), and gas membrane for CO (hours). Rather than modeling at the smallest granularity, the time scale used is that of the observations. This leads to apparently instantaneous time changes: the conditional probability tables are recomputed by finding an equilibrium solution (fixpoint) whenever an observation is made. This part of the work is described in detail in "Continuous Time Bayesian Networks."

6.6 Temporal Models

- "Continuous Time Bayesian Networks," by Uri Nodelman, Christian R. Shelton (presenter), and Daphne Koller (Stanford). See above (in "Applications").

- "Decayed MCMC Filtering," by Bhaskara Marthi, Hanna Pasula, Stuart Russell, and Yuval Peres (UC Berkeley). MCMC stands for Markov Chain Monte Carlo. The key problem is that in many DBNs evidence makes the distant past d-connected to the present. This paper had a very good survey of approximate inference methods for Dynamic Bayesian Networks (DBNs) (p.320 in the proceedings). The title gives a good description of the method proposed by the authors.

6.7 Decision Making

- "Introducing Variable Importance Tradeoff into CP-Nets," by Ronen I. Brafman and Carmel Domshlak (presenter) (Ben-Gurion University). This is a paper on preference elicitation and building a utility model starting from a preference relation, which extends UAI-99 work by Boutilier, Brafman, Hoos, and Poole ("Reasoning with Conditional Ceteris Paribus Preference Statements") and UAI-01 work by Boutilier, Bacchus, and Brafman ("UCP Networks: A Directed Graphical Representation of Conditional Utilities"). The last paper has a discussion of how to obtain a utility table (to be used in an influence diagram) from a conditional preference network augmented with local utility information. [This may be especially of interest to John Rose.]

- "Unconstrained Influence Diagrams," by Finn V. Jensen (presenter) and Marta Vomlelova (Aalborg University). In (conventional) influence diagram, the (linear) order of decisions is fixed. In an unconstrained influence diagram, the order of decisions (and observations) is not determined. The paper presented an algorithm for determining strategies of maximal expected utility for unconstrained influence diagrams. The basic algorithm uses reverse-order variable elimination (as in the evaluation of conventional influence diagrams) in an auxiliary structure called a GS-DAG, with some tricks to deal with loops. The basic algorithm was implemented and performs much more efficiently than the naive approach of considering all possible sequences of decisions. The authors also note that irrelevance may be exploited to reduce the size of the GS-graph, using an algorithm by Nielsen and Jensen [UAI-99].

6.8 Latent Variables and Causality

- "On the Testable Implications of Causal Models with Hidden Variables," by Jin Tian and Judea Pearl (UCLA). Both authors were absent. The paper was presented very well by Blai Bonet (another student of Pearl's).

  Verma and Pearl (UAI-90) observed that for Bayesian networks with hidden variables the graphical structure may impose constraints on a probability distribution that are in addtion to conditional independence constraints. (This is not possible for Bayesian networks without hidden variables.) One could therefore sometimes distinguish two networks that have the same independence constraints by testing whether certain properties hold in the distribution that the networks are supposed to model. The paper presents a "recursive procedure for systematically finding [such] constraints."

- "Factorization of Discrete Probability Distributions," by Dan Geiger (Technion, presenter), Christopher Meek (Microsoft), and Bernd Sturmfelds (UC Berkeley). A joint probability distribution represented faithfully by a DAG (Bayesian network structure) always decompose multiplicatively. This is not true for undirected graphs (Markov networks), where the decomposition theorem (Hammersley and Clifford) holds only for strictly positive distributions. Geiger et al. give new "necessary and sufficient conditions for an arbitrary discrete probability distribution to factor according to an undirected graphical model," thereby generalizing the Hammersley-Clifford theorem. The constraints that the distribution must satisfy are *not* always conditional independence constraints.

6.9 Miscellaneous

- "Updating Probabilities," by Peter Grünwald (CWI, Amsterdam) and Joseph Halpern (presenter, Cornell). Joseph Halpern (editor in chief of the *Journal of the ACM* has never being quite convinced that updating is *always* the right way to In this paper, he considers several "difficult" puzzles (three prisoners, Monty Hall), for which conditioning in a "naive" space gives erroneous results. As Pearl argues in his 1988 book, it is necessary to model the protocol involved in the decision process: what the warden or Monty Hall are allowed to do, as well as what they actually do. The authors give technical conditions on which the "sophisticated space" can be coarsened to a "naive space," and show that they hold very rarely. In other words, it is difficult to overcome careful modeling. Good!

  A second part of the paper discusses similar results involving Jeffrey's rule and minimum relative entropy (MRE) updating. This may have an interesting relationship to the soft evidential update method of Valtorta, Kim, and Vomlel. Unfortunately, Halpern had no time to discuss these results in his talk!

- "Real-valued All-Dimension Search: Low-overhead Rapid Searching over Subsets of Attributes," by Andrew Moore (presenter) and Jeff Schneider (CMU). The authors present a new algorithm for computing high dimensional contingency tables iteratively, which is based on a recurrence relation that allows to compute a table of dimension $i + 1$ from a table of dimension i and (a modest amount of) additional statistics. The authors used their algorithm in data mining algorithms that find good decision rules, but it may be useful in (conditional-independence test-based) Bayesian network learning as well.

- "Exploiting Functional Dependence in Bayesian Network Inference," by Jirka Vomlel (Aalborg and University of Economics, Prague). There has been a steady amount of work on factorization of conditional probability tables (or potential tables), at least since Pearl's 1988 book. Here, (my former and hopefully future coauthor) Vomlel generalized the factorization method of Takikawa and D'Ambrosio (UAI-99) that applies to the Noisy-OR case to the general case. The method is based on introduction of hidden variables. Vomlel defines a rather natural looking problem on hyperrectangles called "minimal base of hyperrectangles" and proves that a CPT (or potential table) factorizes iff that problem has a solution.

  Vomlel also describes a nice application of Bayesian nets to computerized testing, related to work done at Educational Testing Services by Russell Almond, who proposed to use Bayesian nets to model student skills and performance on tasks. For models of this type, the average total clique size (for junction trees) is greatly reduced by using the factorization methods.

- "Robust Feature Selection by Mutual Information Distribution," by Marco Zaffalon (presenter) and Marcus Hutter (IDSIA). The paper presents a novel Bayesian approach (with priors) to estimating mutual information (a.k.a. cross entropy) and contrasts it to the use of the empirical cross-entropy in the context of feature selection for naive Bayesian classifiers. The paper also has a short overview of various approaches to feature selection.

7 Business Meeting

Continuing a longstanding tradition, the U.S.-based program co-chair (Adnan Darwiche) will become conference chair. The program co-chairs will be Chris Meek (Microsoft Research) and Uffe Kjaerulff (University of Aalborg). Meek will become conference chair for UAI-04, when the program chairs will be Max Chickering (Microsoft Research) and Joseph Halpern (Cornell University). This seemed to me to be a break with the tradition of having a non-U.S. program co-chair, but the conference chair (Daphne Koller of Stanford) noted that the same had happened in 2001, when she and Jack Breese (of Microsoft) were program co-chairs. There is some concern that the conference is becoming too U.S.-centered: the only time it was held outside North America was 1999, in Stockholm. It was supposed to be held in Europe in 2002, but the conference chair decided otherwise. It is unclear whether this is a reason for the declining number of European presenters: especially notable was the absence of Spanish researchers, who have been very active in the field in the past few years.

UAI-03 will most likely be held in Acapulco, Mexico (colocated with IJCAI) or Washington, D.C. (colocated with KDD and possibly ML). The majority of the participants in the business meeting seemed to favor Acapulco, although by a slim margin.

There was some discussion on bringing back panel discussion within the conference, re-introducing real tutorials for newcomers, and setting up an (industrial?) applications section with a separate review process.

8 Miscellaneous

- The institutions with most papers in the proceedings (excluding invited talks) were: Stanford (6), CMU (5.66), Microsoft (5.66), UCLA (4), MIT (3.75), Berkeley (3.66), Aalborg (3). Including invited talks, CMU had 6.66.

- Judea Pearl, who was coauthor of three papers, did not come. There was a moment of silence at the banquet in remembrance of his son.

- There were seven attendees with ties to Aalborg university, presenting four papers: Finn V. Jensen, Thomas Nielsen (who completed his Ph.D. last winter), Bo Thiesson (who is now at Microsoft research), Tomas Kocka (who is in the middle of a two-year postdoc), Jirka Vomlel (who just completed his postdoc and is now in Prague), Nevin Zhang (who recently completed his sabbatical and is now back in Hong Kong), and I (1999-2000 sabbatical).

- Stuart Russell had a draft copy of the second edition of his (and Norvig's) *Artificial Intelligence: A Modern Approach* on display. I do not know when the book will be published.

- The Italian language university of Switzerland ("Università della Svizzera Italiana"), which had opened in 1996 as a cantonal institution, has finally been recognized by the Swiss federal government as a University.

9 Closing Remarks

I enjoyed the papers on "how to do things right" and "how to do new things" (as in the papers by Grunwald & Halpern and Jensen & Vomlelova), on exact (structure-based) algorithms (as in the papers on learning), on complexity (as in the paper by Park), on causality (as in the student paper award winner). On the other hand, I did not particularly like the papers on approximate algorithms that were justified by clever technique rather than by real reductions in complexity (such as the paper award winner, at least as far as I understood it).

UAI has relations with well established areas: statistics, operations research, economics, epidemiology, and more. Members of these communities have contributed a lot to UAI: the most notable example is that of the distinguished mathematical statistician Steffen Lauritzen. However, researchers in these other fields often fail to share the concerns of algorithmic efficiency and automation of reasoning that are central to UAI.

The best way to increase the visibility and importance of UAI is to shore up its intellectual basis and to continue to demonstrate the relevance of the methods and techniques developed within it. The conference provided convincing evidence of the continued vitality and excitement of the field of Uncertainty in Artificial Intelligence.

10 Glossary

AAAI (Yearly conference of the) American Association for Artificial Intelligence, also known as National Conference on Artificial Intelligence (NCAI).
AIS (Biennial workshop on) Artificial Intelligence and Statistics.
ECAI (Annual) European Conference on Artificial Intelligence.
ECSQARU (Biennial) European Conference on Quantitative and Symbolic Approaches to Reasoning under Uncertainty.

FLAIRS (Annual conference of the) Florida Artificial Intelligence Research Society.
IEA/AIE (Conference on) Industrial and Engineering Applications of Artificial Intelligence and Expert Systems.
IJCAI (Biennial) International Joint Conference on Artificial Intelligence.
NIPS (Annual conference on) Neural Information Processing Systems.
SPIE International society for optical engineering.
UAI (Annual Conference on) Uncertainty in Artificial Intelligence.