

Bayesian Networks: an Exploratory Tool for Understanding ICT Adoption

Sergiu Nedeveschi[†], Jaspal S. Sandhu[‡], Joyojeet Pal[§], Rodrigo Fonseca[†], Kentaro Toyama⁺

[†]Department of Electrical Engineering and Computer Science,
University of California, Berkeley
Berkeley, CA 94720
{sergiu,rfonseca}@cs.berkeley.edu

[‡]Department of Mechanical Engineering,
University of California, Berkeley
Berkeley, CA 94720
jaspal@me.berkeley.edu

[§] Department of City and Regional Planning,
University of California, Berkeley
Berkeley, CA 94720
joyojeet@berkeley.edu

⁺ Microsoft Research India,
"Scientia"- 196/36 2nd Main,
Sadashivnagar, Bangalore 560 080, India
kentaro.toyama@microsoft.com

Abstract— Understanding technology adoption in emerging regions is challenging given the complex interrelations among socioeconomic factors that affect it directly and indirectly. The issue of impact assessment of technology adoption projects, especially the kind implemented in areas where prior technology has been very limited, is highly problematic and open to many methodological difficulties. Ethnographic evaluations have provided insight into the quality of interactions and into conceptions of technology and its adoption, whereas some quantitative analysis has been useful for high-level abstraction.

In this paper, we examine the use of Bayesian networks as tools that can be used in revealing the structure of the relationships between demographic, social, and economic factors, and penetration for various technologies. Our hypothesis is that technology adoption cases in emerging regions display unique aggregated characteristics that make Bayesian network-based analysis a useful starting point in defining relationships between variables in project analysis. We compare the usability of Bayesian networks in analyzing two data sets: (1) a detailed survey focusing on 500 respondents across 14 favelas in Rio de Janeiro; and (2) a comprehensive survey of 998 users of the Akshaya tele-kiosk initiative in Kerala, India. Our illustrations show how Bayesian networks can be useful as statistical analysis tools that reveal new hypotheses, suggest unintended correlations in data, and confirm standing hypotheses.

Index Terms— ICT4D, Statistical Methods, Bayesian Networks

I. INTRODUCTION

Deciphering the complexity of technology adoption in emerging regions using a quantitative methodology is difficult, partly because of the uncertainty of information, and partly because of the complex interaction of many

variables. We consider the use of *Bayesian networks* for exploring and analyzing such data, as a complement to standard regression models and statistical techniques.

A Bayesian network (“Bayes net”, “belief network”, and “probabilistic graphical models” are also used interchangeably) is a graph in which nodes represent random variables and directed edges between nodes encode information about variable dependence and independence. Based on self-consistent axioms of probability, the basic concepts behind Bayes nets are simple yet powerful. Bayesian analysis has the potential to provide richer models than those generated by, for example, traditional regression, because it explicitly models dependence relationships between variables and can handle complex relationships, not easily captured by simple functions. Bayesian techniques are, however, not without their drawbacks. As with any probabilistic technique, outcomes of analysis are highly dependent on data – both quantity and quality. In addition, there is a tradeoff between model accuracy and interpretability, especially when models are complex and variable relationships are represented by non-parametric means.

Extensive literature – formal and otherwise – exists on Bayesian models. In the context of this paper, it is not our purpose to extensively describe the underlying theory, as this would deviate too far from the main themes. However, we do present a few introductory concepts and a simple example in section 2. For the reader who requires a more comprehensive introduction to Bayesian networks, we offer the following resources which we have found to be most useful in our work: [5], [3], [8], and [11].

In the social sciences, [10] first introduced the concept of Bayesian model selection to social science research. [12] is perhaps most relevant to the presented work as it examines the relationships among country-level human development indicators using Bayesian methods. The presented work roughly shares a domain with [12], but is distinct in that it asks

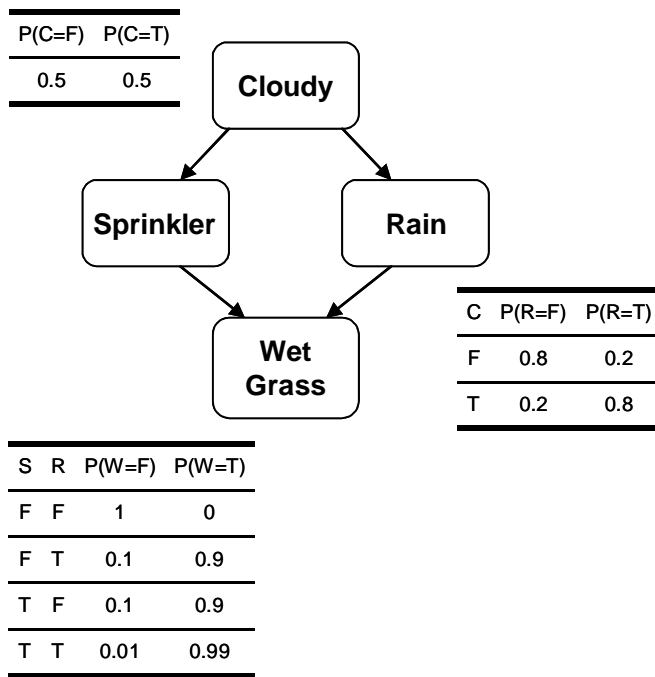


Figure 1. Example of a Bayesian network.

specific questions about ICT, it uses evidence at the individual level via survey data, and it uses Bayesian networks to propose hypotheses.

This paper illustrates the exploratory power of Bayesian networks on two data sets gathered to understand ICT adoption. Our work is meant to demonstrate the use of such methods in this domain, as a way to propose specific hypotheses regarding technology adoption. Formalizing the effect of personal factors on the potential success of technology adoption can be useful for targeted deployment of ICT (Information and Communication Technology) initiatives and for governmental policy with respect to ICT. In particular, we apply Bayesian analysis to provide hints to questions such as the following: What are the key social factors predicting computer usage and how are they interrelated? How do household technologies relate to computer usage? Is there any internal structure to ownership of household technologies, i.e. a “technological hierarchy”?

II. OVERVIEW OF BAYESIAN NETWORKS

“Graphical models are a marriage between probability theory and graph theory. [...] Fundamental to the idea of a graphical model is the notion of modularity -- a complex system is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. The graph theoretic side of graphical models provides an intuitively appealing interface by which humans can model highly-interacting sets of variables [...]” ---Michael Jordan, 1998.

Bayesian networks are used to represent the joint probability distribution of a set of random variables in a

compact manner. To illustrate, let us take the example¹ presented in Figure 1. In this example, all nodes are binary (either true or false), and thus their probability distribution is discrete, and is represented by conditional probability tables (CPTs).

The arrows in a Bayesian network indicate conditional dependency relationships. A causal relationship between two variables can be represented by a directed edge (represented as an arrow in the graph). Since we know that either the sprinkler or the rain will cause the grass to be wet, the variables are linked by the appropriate edges. The strength of these relationships is presented in the CPTs associated with node “Wet Grass”. For example, we see that the probability of the grass being wet, given that it is raining and the sprinkler is not on is: $P(W=T | S = T, R = F) = 0.9$.

It is critically important to note, however, that a *directed edge does not necessarily indicate causality*. The arrows only indicate probabilistic conditional relationships, and may not necessarily represent causal relationships. The simplest conditional independence relationship encoded in a Bayesian network can be stated as follows: a node is independent of its ancestors given its parents. In our example, it is enough to know the value of “Sprinkler” and “Rain” in predicting the outcome of “Wet Grass”, without having to inquire about the value of the variable “Cloudy”, which becomes irrelevant.

Using the chain rule of probability, the joint distribution of all the variables can be expressed as follows:

$$P(C, S, R, W) = P(C) * P(S | C) * P(R | C, S) * P(W | C, S, R)$$

However, by applying the conditional independence relationships encoded in the network, this can be simplified as follows:

$$P(C, S, R, W) = P(C) * P(S | C) * P(R | C) * P(W | S, R)$$

We were allowed to perform these simplifications because R is independent of S given its parent C, and W is independent of C, given its parents S and R. We can see that conditional independence allows us to represent the joint probability more compactly. While these savings seem insignificant in this constrained example, in the general case they become very important.

The joint probability distribution can be estimated given a large enough set of samples - tuples of possible values for the variables: (C,S,R,W). Using the compact and structured representation for the joint probability distribution provided by Bayesian networks is essential because it allows accurate estimation of the joint distribution using much fewer training samples. In our case this is especially important, because it allows us to perform accurate analysis given small data sets.

Once the network parameters (the probability distribution for each node) are learned from a set of training samples, the network can be used for inference, i.e., asking questions such as: “What is the probability of the sprinkler being turned on,

¹ The example is featured in [8]

knowing only that the grass is wet and it is raining? ($P(S=T|W=T,R=T)$)”.

III. SOFTWARE TOOLS

The machinery for computing on Bayesian networks is complex but repetitive and therefore particularly given to analysis by computer. We briefly discuss our choice of software. We used BNT (Bayes Net Toolbox) [7] as the basis for all analysis. This free, open-source MATLAB extension was appropriate to our application because of the low barrier to building models and extending the core functions (e.g., scripting for search over the structure space). BNT also supports multiple inference engines, both exact and approximate. We found that exact inference was possible for all of our experiments.

Many alternative software packages exist, but each has its own advantages and limitations. Still, there are two worth noting: BUGS/WinBUGS, perhaps the most widely used package (Gibbs Sampling for inference), and MSBNx (Microsoft Bayesian Network Editor and Tool Kit), which supports exact inference (Junction Tree), but cannot learn model parameters.

IV. DATA

The two data sets we explored provide usage-related data of technology projects being implemented for development. The first data set (henceforth, *CDI*) was a survey of 500 residents in neighborhoods (*favelas*) that had community computer access centers in Rio de Janeiro, Brazil [4]. Respondents were asked a number of questions relating to their use of computers and social dynamics in a structured multiple-choice survey. The respondents were selected randomly from within communities chosen through a stratified sample of computer-access points in the city, and approached at selected street corners of the locations surveyed.

The second data set (*Akshaya*) was of a survey of 1750 residents of randomly selected neighborhoods, also picked through a stratified selection method, in two districts of the southern Indian state of Kerala - one with computer kiosks for public use, and another comparison group without the same. The surveys were conducted in person via household visits. The questions asked related to respondents' use of services at public use kiosks, and their household decisions relating to technology use and purchase. For the purposes of this analysis, we used only the data set of the district with the public-use computer kiosks (998 respondents) and compared between the respondents who used the kiosks and those that did not. We used these comparisons to look at the relationship between computer kiosk use and intent to send children to computer-learning classes.

Both data sets come from comprehensive surveys of more than 100 questions, emphasizing personal and household demographics, along with details of computer & Internet usage. Overall, the surveys were well-designed, well-coded, and well-implemented. In spite of this, there were still some issues to address, as follow.

Data was missing for a variety of reasons (e.g., don't know, not applicable, choose not to respond), thus some questions could not be used. As this was survey data, there were also general self-reporting issues. In particular, income and computer usage, two variables that we were very interested in modeling, were very susceptible to self-reporting bias. By design, a large proportion of the respondents were from pre-selected groups, resulting in a non-random survey population. We explicitly considered this factor when developing our models since usage had a very strong impact on some variables (e.g., 100% of CDI users had used a computer) and would thus bias the models.

V. METHODOLOGY

As mentioned earlier, the focus in this work was on individual survey respondents rather than regional or domestic indicators. In this context, Bayesian networks have several key advantages. They can handle heterogeneous data, discrete and continuous, both of which appeared in this survey. They can incorporate incomplete data, which was necessary in this case.

Bayesian networks are characterized by two factors: the structure of the Bayesian network itself (the arcs in the graph), and the value of the parameters modeling the conditional probability distribution of each variable. The standard way to use Bayesian networks is the following: using a-priori knowledge, a network topology is constructed by hand, and the parameters modeling the probability distribution for each node are learned by examining a training set of samples. Once the model is trained, it can be used for inference.

The relationships among socioeconomic factors have not been adequately explored in the nascent ICT4D field, however, and consequently we could not assume any model structure. Instead, we used a technique to suggest model structure, by searching through a large space of possible network structures, by training each candidate model, and then comparing the performance of all these resulting trained models. By doing an exhaustive search, we can identify the model or models that best explain the given data. There is a fine balance between enumerating and exhaustively searching possible models on one hand, and using expert or *a priori* knowledge on the other. While exhaustive search is more thorough in some senses – it finds globally optimal solutions – it rapidly becomes computationally intractable over myriad indicators and graph permutations. The number of possible Bayesian network structures for 3 nodes is 25 and for 5 it is 29,281 – this explodes to 5.22×10^{26} for 12 variables.

Though this doesn't bode well for computational tractability, there are many methods to reduce the search space: using model equivalence as in [2], selecting better variables using other statistical techniques (e.g. Principal Components Analysis, PCA), or using local hill-climbing search approaches (e.g. flip/add/remove edge in the case of a fixed set of variables; fix a model, add a node, and determine connections to that node in the case of incremental variable addition). Later in the paper we present some of these techniques in detail. The flipside of the computational complexity is that it forces rigorous hypothesis formulation

and testing, and keeps practitioners from complacently using Bayesian networks as black boxes that can be used with little understanding.

We followed the following methodology for creating and evaluating each Bayes network; we performed this for each graph in the search space.

- a. Define a network topology
- b. Define statistical distributions on each node
 - i. continuous: (conditional) Gaussian
 - ii. discrete: CPT (conditional probability table), softmax function
- c. Learn parameters for distributions from (b) by feeding the model training data (some portion of the survey responses)
- d. Evaluate fitness of model (distributions + parameters) to “held out” data (remaining portion of survey responses) by maximizing likelihood.

The likelihood of a model is an unnormalized metric that conceptually represents the fitness of the model parameters to the data. Likelihoods can be used to compare graphs over a fixed set of variables with different topologies (connections between the nodes); however, it cannot be used to compare graphs over different variables (e.g. sex-age-computer_usage vs. race-religion-computer_usage). This makes sense from a non-technical standpoint as well, since the underlying questions are fundamentally different.

For performing inference we used the Junction Tree Algorithms, and we chose to estimate log-likelihoods rather than straight likelihoods. We used a mixture of experts approach, some discrete and some continuous, in conjunction with the expectation-maximization (EM) learning algorithm. For details please see [6].

As we ran exhaustive search for models as large as 4 nodes, over as many as 500 data points, we had to concurrently run simulations on multiple machines.

VI. ANALYSIS: CDI

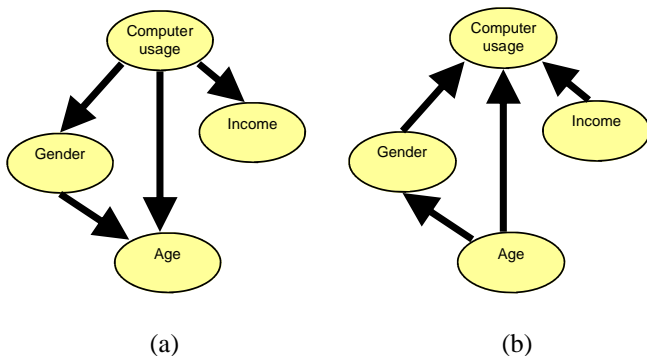


Figure 2. Best performing network topologies linking Computer Usage with Gender, Age and Income.

Social factors influencing computer usage. The first of our analyses involved the following variables from the CDI data set: age (continuous), gender (M/F), income (discrete, 6 categories), computer usage (Y/N). We only examined data

from non-computer-center users (251 respondents) as all CDI users have used computers and this would introduce undesired bias into the model. Through the process outlined above, we obtained the models presented in Figure 2. These two best models are Markov-equivalent, meaning that they encode equivalent conditional relationships, and hence have the same likelihood. Any of these models can then be used to perform predictive inference, i.e. to answer questions like: “Knowing

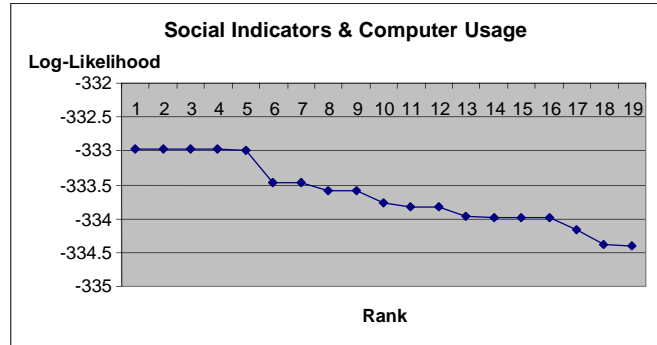


Figure 3. Best performing network topologies linking Computer Usage with Gender, Age and Income.

that a particular subject is a 38-year old female, what is the probability of the subject being a computer user?”. The network shown in Figure 2b, is perhaps more intuitive if the arrows are read as potential causal relationships.

The log-likelihoods of the 20 fittest models are plotted in

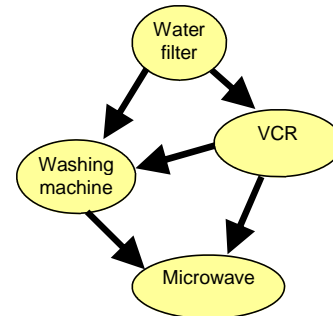


Figure 4. Best performing network topology exploring household technologies inter-relationships.

Figure 3; the clustering of log-likelihoods at a particular value indicates classes of Markov-equivalent models.

The analysis results in a hypothesis that age, gender, and income all affect computer usage, and that knowing any combination of these, we can still make a better prediction with information regarding the other variables. An unexpected part of the model is the link between age and gender. This prompts careful inspection of the original data, which, in fact, reveals that there is bias in the data set.

Hierarchy of household technologies. The next analysis involved ownership (Y/N) of the following household technologies: microwave, VCR, washing machine, and water filter. We used the CDI data with both computer-center users and non-users and obtained the model in Figure 4.

Here conditional independence becomes a good tool for intuiting the meaning of the model. Given that we know whether a person owns a VCR and a washing machine, we do

not need to know about microwave ownership to determine probability of water filter ownership (and vice versa). This is precisely what we mean by a hierarchy of technologies.

Influence of household ICTs on computer usage. This analysis examined the influence of ownership of household ICTs on computer usage; the variables (all Y/N) were: land phone, mobile phone, cable TV, and computer usage. As for

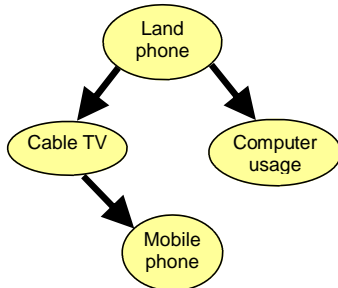


Figure 5. Best performing model linking Computer Usage with other technologies.

the first analysis, we used non-computer-center users and obtained the model depicted in Figure 5.

What does this model mean? Computer usage is independent of cable TV and mobile phones given information about land phone ownership. We can also view this as a technology hierarchy. Given information about a mobile phone, we can strongly predict whether they have cable TV or not (if they do have a mobile, they are likely to have cable TV; if they do not have a mobile phone, we know that there is a different known probability); similarly for cable TV and land phones. Also, given computer usage, we can predict land phone ownership.

VII. ANALYSIS: AKSHAYA

As mentioned earlier, for the Akshaya data, we only analyzed the portion of the data pertaining to the district with the public-use computer kiosks (n=998 respondents). We compared three groups in total in order to look at the relationship between computer-kiosk use and intent to send children to computer-learning classes: *Self-eliterate* (n=188), where the respondent had completed computer training; *Family-member-eliterate* (n=130), where the respondent had not completed computer training, but had a household family member who had; and *Chose-no-eliteracy* (n=133), where the household was knowledgeable about the existence of computer training, but did not send any household member to the training.

We analyzed these three groups independently using the same three variables: *occupation* (15 categories), *socioeconomicStatus* (4 levels), and *planToSendKid* (5 responses). The last of these indicated whether the parents intended to send a child to computer training in the future. We conducted exhaustive search over the set of all possible graphs (25 in the 3 node case) for these 3 groups. We again used a 2:1 ratio between training and evaluation data; these data were randomly assigned to the training and evaluation sets.

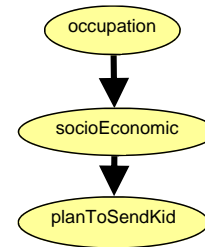


Figure 6. Akshaya model indicating conditional independence of occupation and plans to send a child to computer training given socioeconomic status.

An equivalent alternative to analyzing the three groups independently would be to perform the analysis on the entire data set, treating *eliteracyChoice* as an explicit, 3-valued variable in the model. We chose to split the data set in separate groups for clarity.

The search result for all three groups was identical in terms of graphical structure. The top graphs – 2 graphs that were Markov-equivalent – showed a relationship between *occupation* and *socioEconomicStatus*, as would be expected, but showed that *planToSendKid* was independent of both of the other variables. However, the “next-best” models (Figure 6) scoring very closely in log-likelihood, showed the same correlation between *occupation* and *socioEconomicStatus*, but additionally correlated *socioEconomicStatus* to *planToSendKids*. These models indicate an interesting conditional independence, namely that, knowing the socioeconomic status, the information about the subject’s occupation is superfluous in predicting whether the subject is planning to send her kids to school. A previous Akshaya study showed that computer application use indeed depended more on socio-economic factors than on the occupation of the persons living in a specific region [9]. These “next-best” graphs align more closely with the prior study as they show that socioeconomic status, not occupation, is the critical influencing factor for an outcome related to access.

The graphs resulting from the search are identified by their structure, identical in this case, and the learned parameters. The learned parameters distinguish the resulting

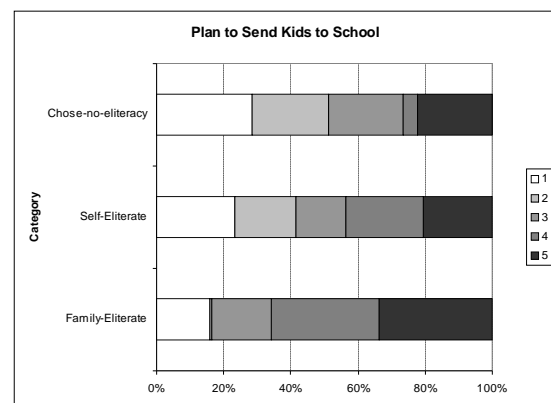


Figure 7. Comparing the conditional probability tables for variable *planToSendKid* for three groups within the Akshaya data set – *Self-eliterate*, *Family-member-eliterate*, and *Chose-no-eliteracy*. This is for a common model where the *planToSendKid* is an independent variable.

graphs from one another in these three cases. Since the variables are all multinomial (discrete), the parameters can be represented in CPTs (conditional probability tables), such as the ones from Figure 1. The difference here is that, given the large number of discrete values of each node, the resultant CPTs can be very large. For example, if *socioEconomicStatus* is conditioned just on *occupation*, the resulting CPT would have 60 entries (probability of each of 4 socioeconomic classifications, conditioned on each of 15 occupational categories).

In order to present the results in a meaningful manner, we have selected a key set of conditional probabilities to compare among the three groups. Figure 7 compares the *planToSendKid* CPT of each data set, for the models in which it is independent (top-scoring model). The categories in the bar graphs for both Figures 7 and 8 are as follows: (1) planning to send all the children to computer training; (2) planning to send some of the children to computer training based on intelligence/other quality (3) planning to send only

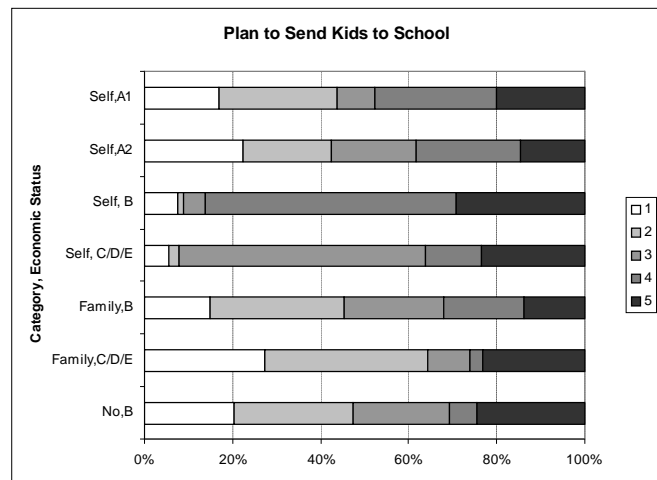


Figure 8. Comparing the conditional probability tables for variable *planToSendKid* for three groups within the Akshaya data set – *Self-eliterate*, *Family-member-eliterate*, and *Chose- no-eliteracy*. This is for a common model where the *planToSendKid* is conditioned on socioeconomic status (4 categories: A1, A2, B, C/D/E) as shown in Figure 6.

the boys and not girls; (4) planning to send only the girls and not boys; and (5) Not planning to send any of the children. This comparison yields some very interesting insights related to the perceived value of computer training, assuming that the intent to send children to computer classes is an indication of this perception; surprisingly, people who until now had no family member undergo a class are the ones most likely to send all their children to computer training, and the most likely to send a child to eliteracy courses. At the opposite end of the spectrum are the people who have a family member that already underwent training, but that haven't had training themselves. These are the people that seem to value the training the least. People who themselves underwent training are somewhat in the middle.

These findings seem to indicate that computer training didn't have an important perceived value for people taking it, and even less so for their family members. This is consistent with other findings from the same survey, which indicate that

most people found computer classes helpful in overcoming their fear of computers, but not helpful for other practical purposes.

Figure 8 again compares the *planToSendKid* CPT of each data set, but now for the 2nd best model (shown in Figure 6), conditioning on *socioEconomic*. As can be seen in both figures, the Bayesian networks encode complex and revealing probabilistic relationships among variables. As a single example, across the three data sets, conditioning on the socioeconomic group B, the probability of an individual sending only the girl(s) in the family is very high in the *Self-eliterate* group (57%), lower in the *Family-member-eliterate* group (18%), and extremely low in the *Chose-no-eliteracy* group (6%).

VIII. EXTENSIONS

As presented in Section IV, exhaustive search through the space of all possible Bayesian networks is not tractable for large numbers of variables, being already out of reach for more than 5 nodes. Several strategies to reduce the search space can be envisioned, such as hill-climbing, simulated annealing, genetic search, partitioning the problem space into additive sub-problems, etc. None of these techniques guarantees optimality, but models that are “good enough” can be found much faster.

Reducing problem dimensionality is another technique to cope with complexity, and mechanisms such as principal component analysis (PCA) or search through the space of equivalence classes of Bayesian networks rather than searching through individual networks [2] can be employed. We tested several extensions on the CDI data; descriptions of these experiments follow.

Genetic Algorithms: These algorithms are efficient tools for parallel search through fitness landscape by entertaining several solutions simultaneously. Any genetic algorithm follows the following general structure.

- 1) Generate an initial population
- 2) Select a “fit” subset of the organisms from the present population
- 3) Produce offspring by crossing different “fit” organisms
- 4) Allow mutations in the current population
- 5) Return to 2

In our implementation, the initial population of Bayesian networks was generated at random. We chose not to implement any cross-over operations, and our possible mutation operations were the following: a) *add a directed edge*; b) *remove an edge*, and c) *reverse the direction of a given edge*.

Social factors influencing computer usage (using genetic search). We revisit this problem by introducing an additional variable, namely whether the subject is a computer-center user. Having this variable explicitly in the model allows us to use the survey data for both computer-center users and non-users, and consequently model the social factors interrelationships more accurately. This approach was not possible

in the context of exhaustive search, since we could not handle 5 variable models. However, this became feasible by using the less computationally intensive genetic search. In the following, we present the conclusions of analyzing the top 4 performing models, which had very similar log-likelihoods, and very similar variable inter-dependencies.

All the top 4 performing models linked, as expected, the variable showing whether a computer course was taken to computer usage. A strong dependence between the same variable and age is observed as well, showing that people taking the course had a different age distribution than the rest of the population. Most of the 4 models also showed

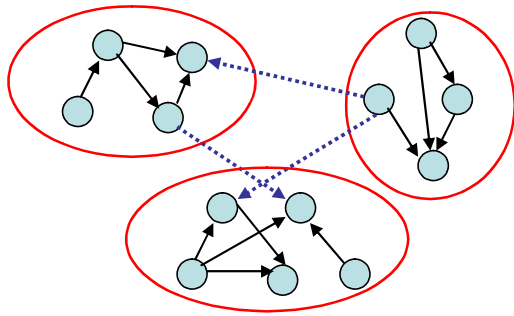


Figure 9. Partitioning the variable space.

dependency of the income to both age and gender, also to be expected. Other factors that influence computer usage in some cases are gender and income. To be noted is the fact that, if no constraint is imposed, a strong correlation between age and gender is observed. However, this fact is due entirely to the survey sampling methodology, and does not generally hold for the rest of the population. For this specific reason, a link between these variables should be explicitly forbidden.

Partitioning the Variable Space: In order to support larger models, with more relevant variables, we implemented a search method that partitions the variable space. The idea is to cluster together variables that are likely to influence each other. Each of these clusters contains less than 5 variables. In the first step, exhaustive search can be used to determine the network structures that best model interactions within each cluster. In the following step, these sub-networks can be assembled together, by searching through the whole variable space, but maintaining the links established in the first step as constraints. This second, more complex step will also require an approximate method such as the genetic search, without guaranteeing solution optimality. An example of such a partitioning is presented in Figure 9.

Predicting computer usage and other variables using generated models: Using the models generated and presented in Figures 2, 4, and 5, we examined whether these could be used to predict one variable (e.g. computer usage), by looking at other variables influencing it. After training the models with two thirds of our survey data, we evaluated how well the unknown variable predicted for the remaining one third of the data. Prediction accuracies for all our experiments varied between 70% and 85%. This is not very accurate, but it is reasonable, given that not all the relevant variables were present in the model. No matter how well the model is trained, age, gender and income alone are not enough to predict with

high accuracy whether the subject is a potential computer user. More complete data with other variables is probably required for better prediction.

IX. CONCLUSIONS

We have demonstrated the use of Bayesian networks as a method for analyzing ICT4D survey data. Bayesian network analysis shows promise in the ICT4D domain, but still requires significant technical expertise. It is certainly not necessary to do exhaustive search, or to apply “intelligent” search methods, but it is necessary to understand the fundamental principles of probabilistic graphical models, in particular conditional probabilities, inference, and statistical distributions. Consequently, there continues to be a barrier to accessing Bayesian network analysis and using it properly, but we hope that this paper goes some way toward demystifying the process.

In this case, we have demonstrated Bayesian networks to be valuable in at least three ways to the social scientist: First, Bayesian networks are useful in confirming existing hypotheses. In the CDI data set, the Bayesian network analysis confirms that several social factors affect computer usage in the general population, with age, gender, and income all directly influencing computer usage. This analysis also reveals a second value of Bayesian analysis – in detecting unintentional links in the data. We were also able to discover an unintended bias in the data set that linked age and gender. Finally, and perhaps the most valuable use of Bayesian networks is in generating hypotheses, given complex data. For example, in the analysis of household technologies, the analysis proposes “hierarchies” of technology adoption. It would be interesting to analyze this in conjunction with personal and/or household income since that is expected to strongly affect the ability to acquire some of these technologies. Such hierarchies can be used to develop proxy metrics in designing surveys or more complex models (e.g., What data can be obtained most easily and what variables become “redundant” knowing this data?). Perhaps most interestingly, household ICTs can serve as reasonable predictor of computer usage; within this analysis there exists another hierarchy of technologies.

These advantages were also applicable in the Akshaya data, where occupation and socioeconomic status were intuitively linked, but there is a more tenuous relation to the multinomial variable dictating whether a parent intends to send a child to computer training (*planToSendKid*). That the result was similar across the three subsets of Akshaya data indicates that there may be more meaningful predictors of this outcome. Alternatively, alternate analyses – such as distilling the 5-valued variable *planToSendKid* into a simple binary variable – might yield a stronger connection to socioeconomic status.

Ultimately, Bayesian networks are simply another tool in the toolbox, and as such, they have their strengths, as well as opportunities for misuse. We note that while the networks uncover interesting hypotheses from the data, they must be taken as hypotheses, which must be more carefully investigated before meaningful conclusions can be drawn.

ACKNOWLEDGMENTS

We would like to thank Nitin Sawhney for his recommendations during an informal conversation at a conference, and Kevin Murphy of the University of British Columbia for his leading role in the development of BNT, and equally as important for sharing this code publicly.

REFERENCES

- [1] AT&T Research. 2005. *Number of acyclic digraphs with n labeled nodes*, Online Encyclopedia of Integer Sequences, <http://www.research.att.com/cgi-bin/access.cgi/as/njas/sequences/eisA.cgi?Anum=A003024>.
- [2] Chickering D.M. 2002. *Learning equivalence classes of Bayesian-network structures*, J. Mach. Learn. Res., Vol. 2, pp445-498, MIT Press.
- [3] Cooper G.; Heckerman D.; and Meek C. 1997. *A Bayesian Approach to Causal Discovery*, Microsoft Research Technical Report, MSR-TR-97-05.
- [4] Ferraz C.; Fonseca R.; Pal J.; Shah M. 2005. *Computing for Social Inclusion in Brazil: A Study of the CDI and other initiatives*, 2005 UNIDO Bridging the Divide Conference, Berkeley.
- [5] Heckerman D. 1995. *A Tutorial on Learning with Bayesian Networks*, Microsoft Research, Technical Report, MSR-TR-95-06.
- [6] Jordan, M.I., 1994, Hierarchical Mixtures of Experts and the EM algorithm. *Neural Computation*, 6, pages 181-214, 1994.
- [7] Murphy K. *Bayes Net Toolbox for Matlab*. 2005. <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>.
- [8] Murphy K. 1998. *A Brief Introduction to Graphical Models and Bayesian Networks*, <http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html>.
- [9] Pal J.; Nedeveschi S.; Patra R.; Brewer E. 2005. *A Multi-Disciplinary Approach to Open Access Village Telecenter Initiatives: The Case of Akshaya*, Policy Options and Models for Bridging Digital Divides, Tampere, Finland.
- [10] Raftery A.E. 1995. *Model Selection in Social Research*, Social Methodology.
- [11] Russell S.J.; and P. Norvig. 2002. *Artificial Intelligence: A Modern Approach*, Prentice Hall.
- [12] Sawhney N. 2001. *Bayesian Model Selection for Human Development Indicators*, <http://web.media.mit.edu/~nitin/DevBayes/DevBayes.pdf>.