Same House + different occupants = different energy use.
Same occupants + different houses = different energy use.
Different occupants + different houses = radically different energy use
(EHCS ‘96, n = 3,676, Mean ~30,000 kWh)
One solution: model home energy use as joint distribution over a domain of variables.
Questions are then:

1. What knowledge domains are relevant?
2. What variables within these domains should we measure and model?
3. What are the relationships between these variables?
4. What probabilities describe these relationships?
Bayesian Belief Networks

'Graphical models are a marriage between probability theory and graph theory. ... 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 both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.' (Jordon 1999 p.1)
Methodological advantages

• Integration of qualitative and quantitative data from experts, case studies, data-sets and models;
• Integrate of new data as it becomes available;
• Highlight conflicts or synergies between variables.
• Intuitive display of relationships between variables;
• Straightforward sensitivity testing.
• ‘Subjective probability’ provides common epistemological ‘common ground’ between social and engineering approaches.
• Create consensus based decision support systems;
Growth of literature
TS="Bayesian Network*" ISI
Environmental applications

- Management of fisheries (Halls & Burn 2002);
- Management of wildlife (Cohen 1988);
- Management of forests (Crome et al 1996);
- Environmental management (Marcot et al 2002);
- Ecological decision making (Dixon and Ellison 1996);
- Decision support for land use change (Bacon et al 2002);
- Participatory resource management (Cain et al 1999);
- Integrated water resource management (Bromley et al 2004)
- Participatory agricultural land management (Cain et al 2003)
Graph theoretic definition of BBN

• Let $D = (V,E)$ be a Directed Acyclic Graph (DAG), where $V$ is a finite set of nodes and $E$ is a finite set of directed edges between the nodes. The DAG defines the structure of the Bayesian network.

• Each node $v \in V$ in the graph corresponds to a variable $X_v$. The set of variables associated with the graph $D$ is then $X = (X_v)_{v \in V}$.

  – Bottcher & Dethlefsen (2003 p.2)
Graph theoretic definition of BBN

- To each node $v$ with parents $\text{pa}(v)$ a local probability distribution, $p(x_v | x_{\text{pa}(v)})$, is attached. The set of local probability distributions for all variables in the network is $P$.

- A Bayesian network for a set of random variables $X$ is the pair $(D,P)$. The possible lack of directed edges in $D$ encodes conditional independencies between the random variables $X$ through the decomposition (factorization) of the joint probability distribution.

\[ p(X_1, X_2, \ldots, X_V) = \prod_{v \in V} p(x_v | x_{\text{pa}(v)}) \]
The medical ‘Alarm’ network

(Monitored variables of intensive-care patients)

- 37 2-state variables gives unstructured state-space of $2^{37}$ parameters
- Structuring this reduces it to 509 parameters
- The structure permits ‘factorization’ of the states-space and makes the problem tractable

Figure from N. Friedman via K. Murphy
BBN construction

- Identification of the domain variables;
- Identification of the relationships between these variables and;
- Identification of the probabilities describing these relationships
  - (Druzdzel & van der Gaag 2000).
1. What knowledge domains are relevant?

- Sociological theories
  - Socio-technical systems theory; Actor network theory
- Psychological theories
  - Attitude-behaviour models
- Economic theories
  - Rational action models
- Physical theories
  - Building thermal simulation
- Different sets of variables
- Different relationships between variables
Psychological theories:
Bagozzi’s Comprehensive Model of Consumer Action
(Ref: Jackson 2005, Figure 17)
Physical theories: The SAP (theory)

What variables within these domains should we measure and model?

- Reviews of:
  - Technical literature
  - Psychological literature
  - Sociological literature
  - Economic literature

- Look for variables which are:
  - Supported by empirical evidence
  - Supported by multiple authors
  - Supported by established theories
  - Likely to explain a significant means of energy use
  - Policy actionable or have good explanatory power
  - Allow replication of previous studies for longitudinal analysis
Some variables measured in CaRB DomNat survey

- Other Heating Controls & Usage
  - Additional Heating (Frequency of use)
  - Heating on if at home
  - Curtains use
- Ventilation
  - Windows & Doors Open
  - Extractor Fans / Cooker Hoods
- Occupancy Patterns
  - Weekly Occupancy Patterns
- Bathing Technology & Practices
  - Shower Technology
  - Bathing / Showering Practices
  - Pools, Sauna’s and Hot Tubs
- Built Form
  - Accommodation Type
  - Number of Storeys
  - Age of Building
  - Loft & Insulation
  - Walls & Insulation
  - Double-Glazing
  - Curtains – not in any so far
  - Draught-proofing
  - Number & Types of Rooms (in types of heating section)
  - Conservatory & Glazing
  - Internal Doors
Age of dwelling
Showers per person per week

![Histogram](image_url)

- **Mean = 4.97**
- **Std. Dev. = 3.382**
- **N = 353**
2. What are the relationships between these variables?

- In Bayesian Networks, the relationship between variables is called the ‘architecture’ or ‘structure’.
- Two main ways of determining structure:
  - Elicitation from domain experts
  - Learning from data
Elicitation from domain experts

- Excellent if:
  - The domain of knowledge are well defined
  - There is a consensus on main variables within that domain
  - There is separation between domains
  - There is a history of sound empirical statistical study in the domain
  - There is a consensus on relationship between variables within that domain.
  - There is a consensus on research approach
    - Theory vs Empirical
    - Qualitative vs Quantitative
    - Statistical vs Deterministic
Physical theories: The SAP (theory)

2. Structure learning from data

• The ‘model space’
  – Robinson (1977) showed that the size of the model space (number of different DAGs) grows super-exponentially with the number of nodes.
  – Thus: \( r(2) = 3; r(3) = 25; r(5) = 29,281; r(10) \approx 4.2 \times 10^{18} \)
    (Leray & Francois, 2004)

\[
r(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} r(n - i) = n^{2^{O(n)}}
\]
Searching model space for models which fit the data

- Model space is huge
- \( \therefore \) need a heuristic search strategy
- \( \therefore \) need a scoring system for DAGs
  - Need a decomposable and equivalent score
    - **Decomposable** if score is sum or product a function of a node and its parents
    - **Equivalent** if score is same for equivalent DAGs
- BIC (Schwartz 1978) is widely used

\[
BIC(\mathcal{B}, D) = \log \mathbb{P}(D | \mathcal{B}, \theta^{ML}) - \frac{1}{2} \text{Dim}(\mathcal{B}) \log N
\]
Markov Equivalence

- If two DAGs have the same joint probability distribution $P(X)$ (i.e. the structure creates the same set of conditional dependencies between variables) they are said to be ‘Markov equivalent’ and belong to the same ‘Markov equivalent class’.

- DAGs are Markov equivalent IFF they have the same edge support and the same set of ‘V’ structures. (Verma & Pearl, 1990)
Bayes’ rule and Markov Equivalent DAGs

- $P(A,B,C) =$
  $P(A)P(B|A)P(C|B) =$
  $P(A|B)P(B)P(C|B) =$
  $P(A|B)P(B|C)P(C) =$
  $\neq P(A)P(B|A,C)P(C) =$

\[ \begin{array}{c}
A \\
\rightarrow \\
B \\
\rightarrow \\
C \\
\end{array} \quad \begin{array}{c}
A \\
\leftarrow \\
B \\
\rightarrow \\
C \\
\end{array} \quad \begin{array}{c}
A \\
\leftarrow \\
B \\
\leftarrow \\
C \\
\end{array} \quad \begin{array}{c}
A \\
\rightarrow \\
B \\
\leftarrow \\
C \\
\end{array} \]
Searching model space for models which fit the data

- BIC measures fit of joint probability distribution (JPD) to data
- But JPD is unique only to Markov equivalence class level
- CPDAGs are unique to Markov equivalence classes
- Multiple DAGs per CPDAG
- Structure searches can’t distinguish DAGs within CPDAGs
Structure learning errors by number of cases for 'K2' algorithm on 37 node 'ALARM' network

Data from Cooper & Herskovits 1993
Search algorithm performance
(Leray & Francois, 2004)

- Test algorithms by dataset length, editing distance and BIC score ($n \approx 30$).

<table>
<thead>
<tr>
<th>INSURANCE</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
<th>15000</th>
</tr>
</thead>
<tbody>
<tr>
<td>MWST</td>
<td>37;-3373</td>
<td>34;-3369</td>
<td>36;-3371</td>
<td>35;-3369</td>
<td>34;-3369</td>
<td>34;-3369</td>
<td>34;-3369</td>
</tr>
<tr>
<td>K2</td>
<td>56;-3258</td>
<td>62;-3143</td>
<td>60;-3079</td>
<td>64;-3095</td>
<td>78;-3092</td>
<td>82;-3080</td>
<td>85;-3085</td>
</tr>
<tr>
<td>K2(2)</td>
<td>26;-3113</td>
<td>22;-2887</td>
<td>20;-2841</td>
<td>21;-2873</td>
<td>21;-2916</td>
<td>18;-2904</td>
<td>22;-2910</td>
</tr>
<tr>
<td>K2+T</td>
<td>42;-3207</td>
<td>40;-3009</td>
<td>42;-3089</td>
<td>44;-2980</td>
<td>47;-2987</td>
<td>51;-2986</td>
<td>54;-2996</td>
</tr>
<tr>
<td>K2-T</td>
<td>55;-3298</td>
<td>57;-3075</td>
<td>57;-3066</td>
<td>65;-3007</td>
<td>70;-2975</td>
<td>72;-2968</td>
<td>73;-2967</td>
</tr>
<tr>
<td>MCMC*</td>
<td>50;-3188</td>
<td>44;-2967</td>
<td>46;-2929</td>
<td>40;-2882</td>
<td>50;-2905</td>
<td>51;-2898</td>
<td>54;-2892</td>
</tr>
<tr>
<td>GS</td>
<td>37;-3228</td>
<td>39;-3108</td>
<td>30;-2944</td>
<td>33;-2888</td>
<td>29;-2859</td>
<td>25;-2837</td>
<td>28;-2825</td>
</tr>
<tr>
<td>GS+T</td>
<td>43;-3255</td>
<td>35;-3074</td>
<td>28;-2960</td>
<td>26;-2906</td>
<td>33;-2878</td>
<td>19;-2828</td>
<td>21;-2820</td>
</tr>
<tr>
<td>GES</td>
<td>43;-2910</td>
<td>41;-2891</td>
<td>39;-2955</td>
<td>41;-2898</td>
<td>38;-2761</td>
<td>38;-2761</td>
<td>38;-2752</td>
</tr>
<tr>
<td>SEM</td>
<td>50;-4431</td>
<td>57;-4262</td>
<td>61;-4396</td>
<td>61;-4092</td>
<td>69;-4173</td>
<td>63;-4105</td>
<td>63;-3978</td>
</tr>
</tbody>
</table>

Figure 3: Editing measures and BIC scores divided by 100 and rounded obtained with different methods (in row) for several dataset lengths (in column) (* As the method MCMC is not deterministic the results are a mean over five runs).
MATLAB: BNT: SLP: GES

ASIA original graph

GES CPDAG (cache)

GES DAG (cache)
MATLAB: BNT: SLP: PC

ASIA original graph

PC PDAG

PC DAG
3. What probabilities describe these relationships?

- **Parameter learning**
  - Quantitative data about each variable is gathered from the surveys for each household
  - This data is read into the BBN model in the form of a ‘case file’
Parameter learning

- Debate in literature on importance of sample size (N) vs. ‘subject to item’ ratio.
- Consensus that for other multivariate methods like BBNs (Principal Components Analysis; Exploratory Factor Analysis; etc) that ‘Subject to item’ ratio is more important.
- ‘Subject to Item’ ratio is the number of respondents (subjects) per line of the Conditional Probability Table in each node of the BBN.
Subject to Item ratios for parameter learning from literature

Statistical Confidence vs. Subject to Item ratio

Data from various papers
Interaction between network structure, variable states and case data

- In BNs, subject to item ratios depend on:
  - The number of states of each variable
  - The number of parents of each variable
- Limit BN variables to 3-states
- Limit number of parent variables per child variable to 3
  - $3 \times 3 \times 3 \times 10 = 270$ respondents for $\sim 70\%$ confidence
  - $3 \times 3 \times 3 \times 20 = 540$ respondents for $\sim 85\%$ confidence
Network development

1. Sensitivity analysis: Which variables are key?
2. Uncertainty analysis: Does the network as a whole remain within expected bounds.
3. Pruning: Delete insensitive variables and links.
4. Refining: Additional quantitative analysis to reassess key probabilities.
5. Extending:
   - Primary qualitative research to extend the network and refine contingencies
   - Primary quantitative research is conducted to populate new nodes with probability data.
6. Go to step 1 (repeat ‘till money or time runs out!)
Conclusions

- Provides policy focused decision support
- Supports evidence based policy making
- Transdisciplinary research epistemology
- Knowledge synthesis consistent with Realist Review method
- Models ‘learn’ through continual integration of data
- Provides ‘cross-fertilisation’ between fields
- Models specific ‘take-back’ effects
- Allows for identification of very specific ‘barriers’ and programme interventions to rectify them.
Top six sensitivities are:
SAP;
Living room temperature;
Outside temperature;
Head of Household income;
Dwelling type;
Employment status;
Head of Household SEC

Darkened variables all influence Total annual kWh
Acknowledgments:

CaRB is a consortium of five UK universities, supported by the Carbon Vision Initiative, which is funded by the Carbon Trust and the Engineering and Physical Sciences Research Council, with additional support from the Economic and Social Research Council and Natural Environment Research Council. The university partners are assisted by a steering panel drawn from UK industry and government.

Further information:

– website: http://www.carb.org.uk
– email: info@carb.org.uk