Probabilistic Graphical Models

Marek J. Drużdżel

University of Pittsburgh

School of Information Sciences and Intelligent Systems Program

marek@sis.pitt.edu
http://www.pitt.edu/~druzdzel

Politechnika Białostocka

Wydział Informatyki

m.druzdzel@pb.edu.pl
http://www.wi.pb.edu.pl/~druzdzel/

Overview

- Outline of the lectures
- (A preview of) the probabilistic graphical models
- A list of hard/interesting problems
- Tools and resources

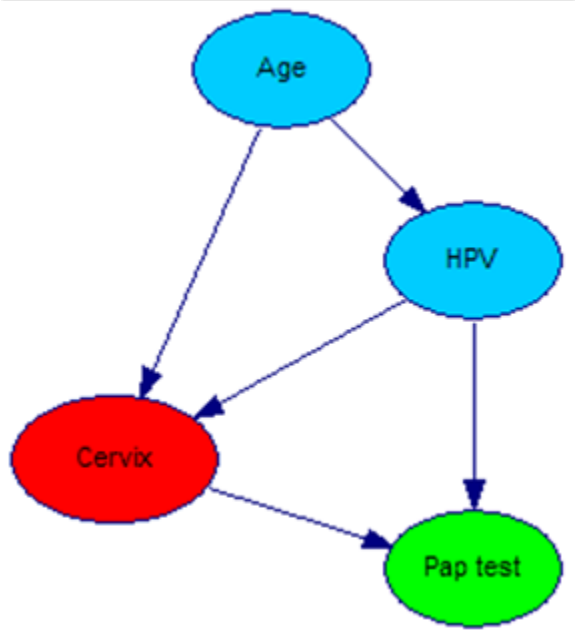
Essentially, a handful of slides enhanced by software.

Outline of the meetings

- Thursday (16:15 – 17:45) ☺
  - Introduction (basic tools and techniques)
- Friday (18:00 – 19:00)
  - Theoretical and practical techniques for decision support
- Saturday (14:00 – 16:45)
  - Learning/causal discovery (14:00 – 15:30)
  - Exercises (15:45 – 16:45)

Bayesian networks

A Bayesian network [Pearl 1988] is an acyclic directed graph consisting of:



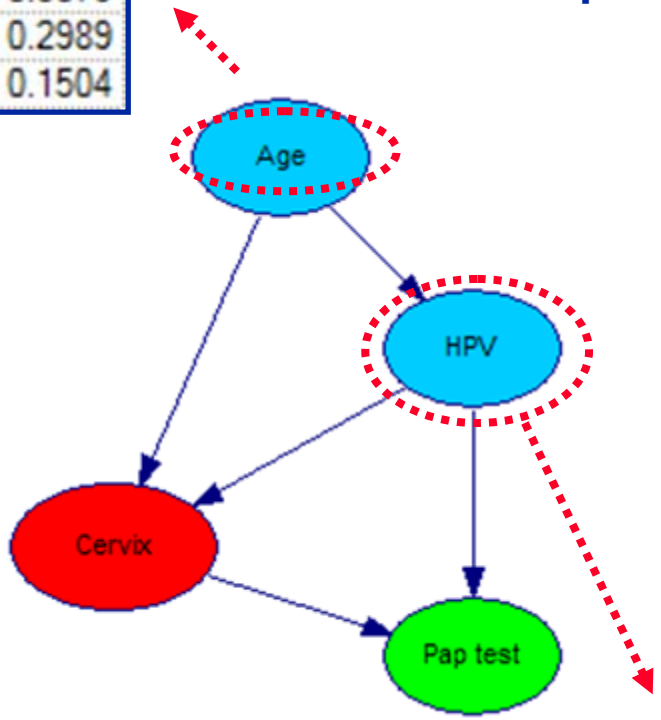
The qualitative part, encoding a domain's variables (nodes) and the probabilistic (usually causal) influences among them (arcs).

The quantitative part, encoding the joint probability distribution over these variables.

Bayesian networks: Numerical parameters

a1_below_20	0.0416
a2_20_29	0.2012
a3_29_45	0.3079
a4_45_60	0.2989
a5_60_up	0.1504

Prior probability distribution tables for nodes without predecessors (Age)

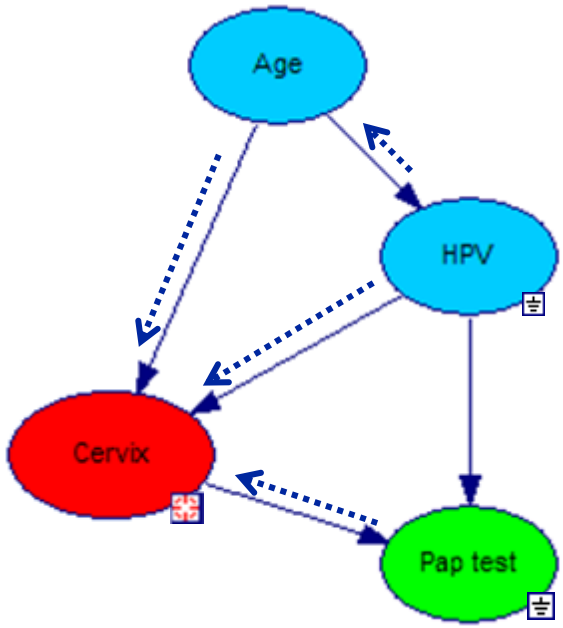


Conditional probability distributions tables for nodes with predecessors (HPV, Pap test, Cervix)

Age	a1_below_20	a2_20_29	a3_29_45	a4_45_60	a5_60_up
NA	0.8652	0.8387	0.7904	0.8055	0.8851
Negative	0.069	0.0901	0.1782	0.1765	0.1012
Positive	0.0613	0.0667	0.0282	0.0142	0.0082
Qns	0.0045	0.0045	0.0032	0.0038	0.0055

Reasoning in Bayesian networks

The most important type of reasoning in Bayesian networks is updating the probability of a hypothesis (e.g., a diagnosis) given new evidence (e.g., medical findings, test results).

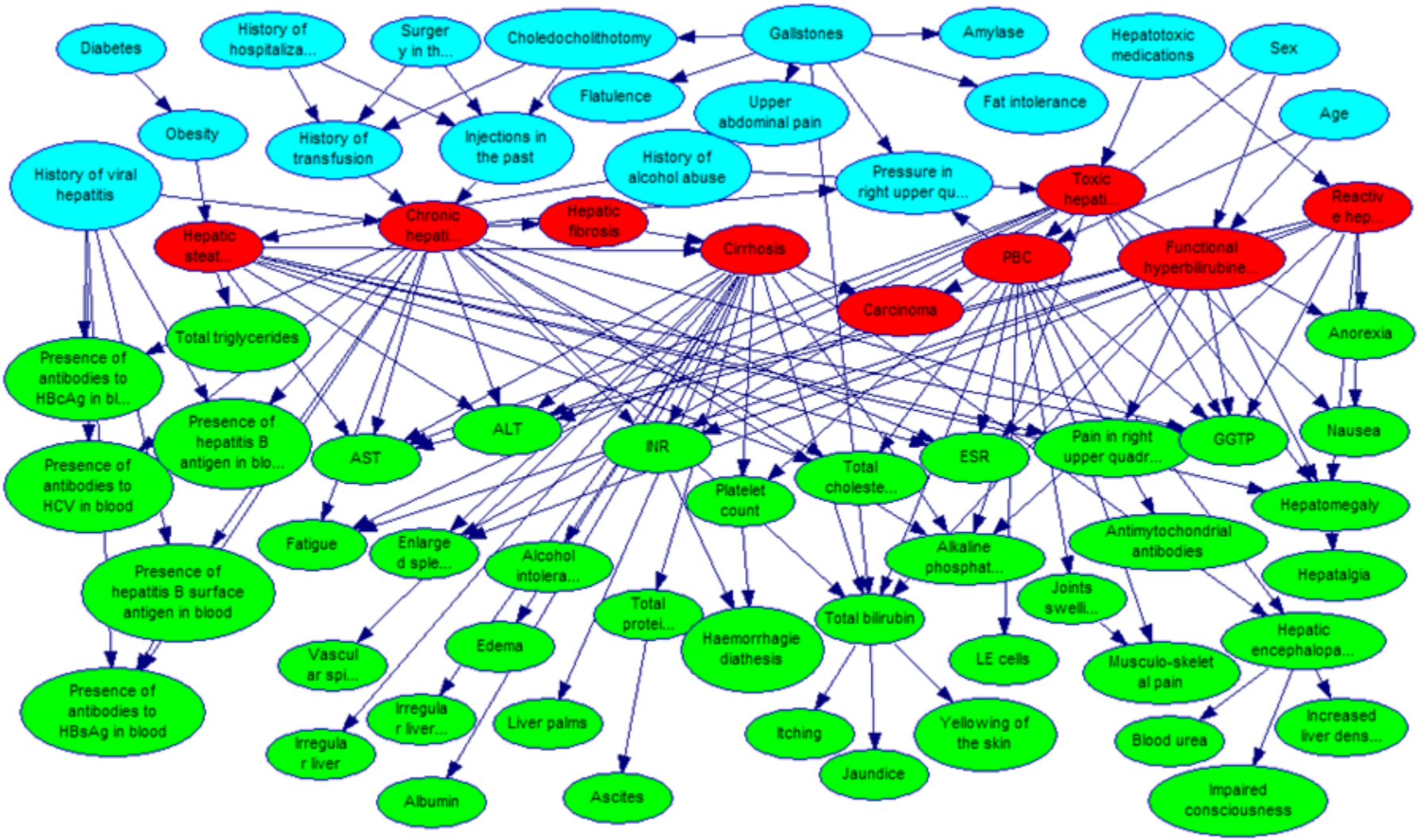


Example:

What is the probability of invasive cervical cancer in a (female) patient with high grade dysplasia with a history of HPV infection?

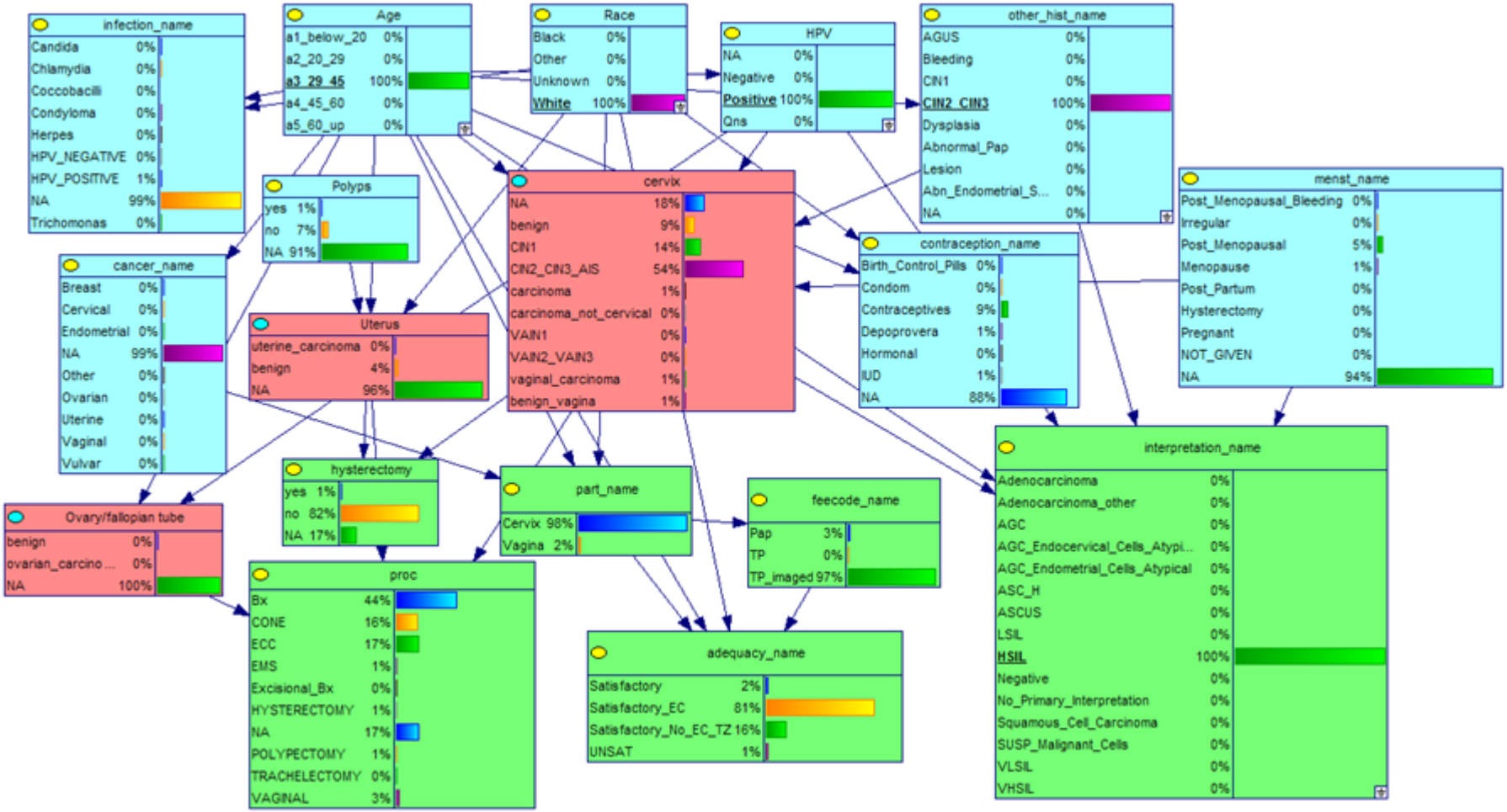
P(CxCa | HPV=positive, HSIL=yes)

HEPAR II Model



70 variables; 2,139 numerical parameters

Pittsburgh Cervical Cancer Screening Model



18 variables; 295,163 numerical parameters

Equation-based systems and graphical models

classsize = (*nstud* * *cload*) / (*nfac* * *tload*)

facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))

stratio = *nstud* / *nfac*

← Core equations

cload = 15

tload = 6

nstud = 22102

nfac = 3006

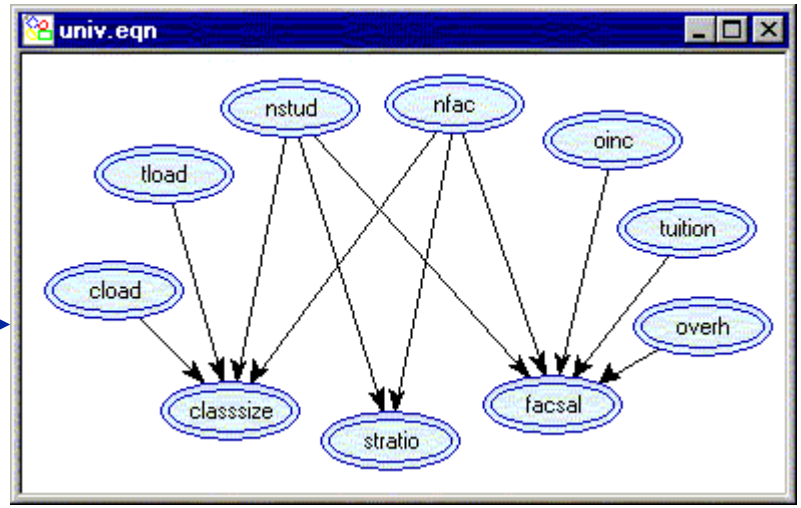
oinc = 30000000

tuition = 12000

overh = 0.48

← Equations for exogenous variables

Together they determine the structure of the model →



Equation-based systems: Reversibility of causal ordering

classsize = (*nstud* * *cload*) / (*nfac* * *tload*)

facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))

stratio = *nstud* / *nfac*

cload = 15

tload = 6

nstud = 22102

~~*nfac* = 3006~~

oinc = 30000000

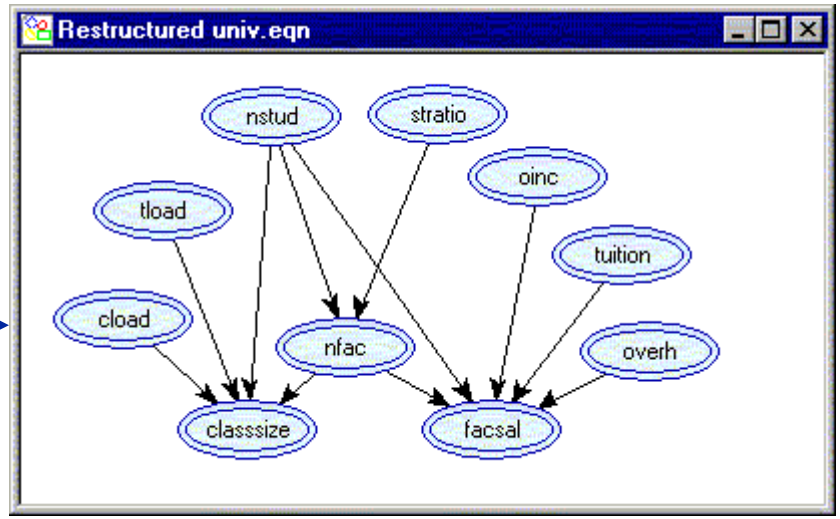
tuition = 12000

overh = 0.48

Setting *stratio* to be exogenous at the expense of *nfac*

stratio = 10

The new model structure



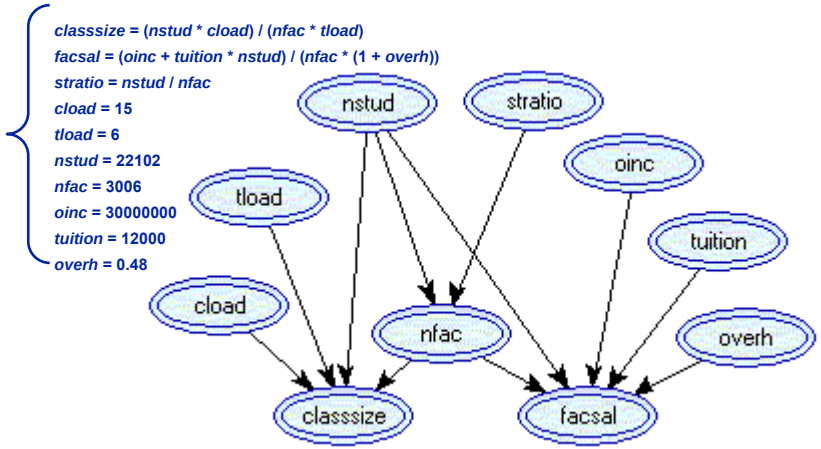
Explication of the asymmetries due to Herb Simon (early 1950s)

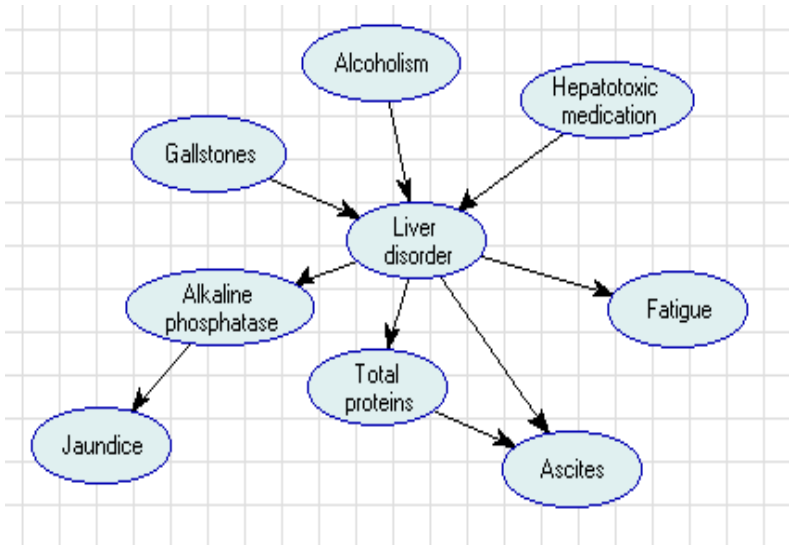
Advantages of directed graphs

- May be built to reflect the causal structure of a model (helps with obtaining insight into the problem)
- Can accommodate representation of uncertainty
- Can be reconfigured as needed
- Have sound theoretical foundations: We are dealing here with probability theory and decision theory
- We can talk (almost) the same language with statisticians, philosophers, and scientists

Family of directed graphs (a bigger picture)

(a.k.a. “influence nets,” “causal diagrams,” etc.)

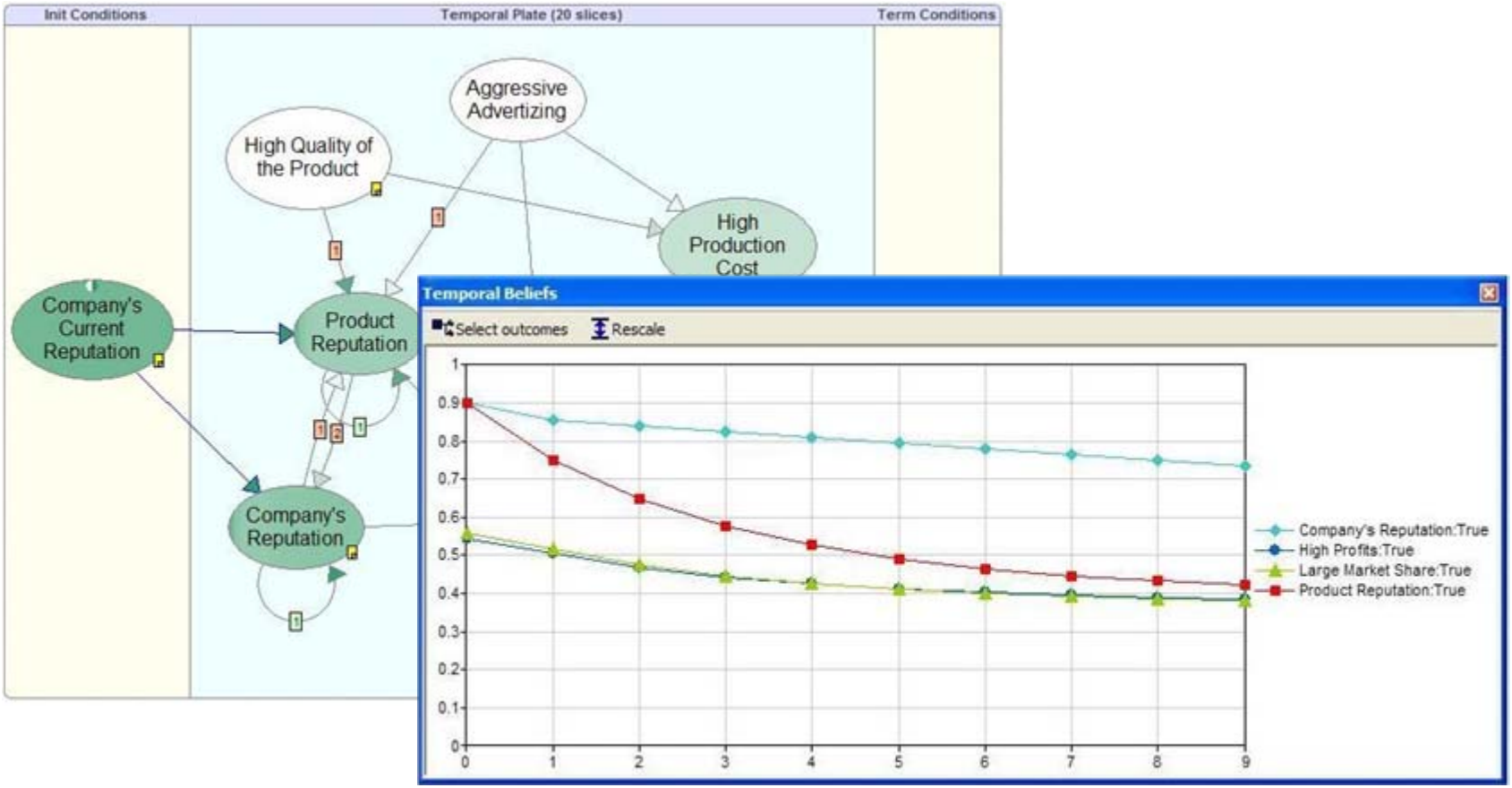




Both, systems of equations and joint probability distributions can be pictured by acyclic directed graphs.

Temporal reasoning

Temporal models allow for tracking development of a system over time and support decision making in complex environments, where not only the final effect counts.



Decision making

When a probabilistic graphical model is enhanced with an explicit representation of decision options and utilities, it implements the foundations of decision theory and allows for optimizing decisions.

Choice	Business Result	Final Cash Position ($)
Savings	Success	2200
	Flop	2200
Business	Success	5000
	Flop	0

Learning/Data Mining

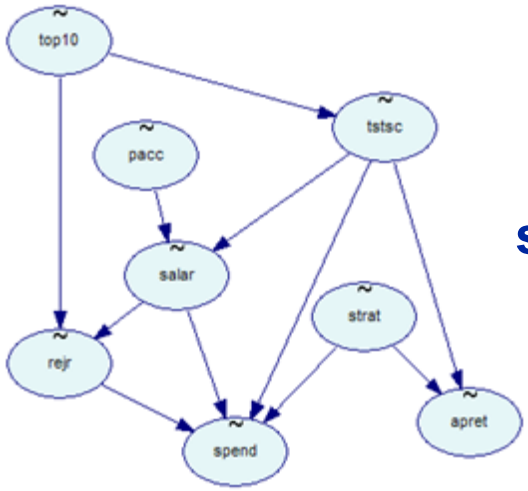
There exist algorithms with a capability to analyze data, discover causal patterns in them, and build models based on these data.

Retention.txt

spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	64.25	36	22.309	71.063	30.97	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289	17	51200
7251	62	17	22.635	56.25	46.78	18.1	48000
6967	66.75	40	9.718	65.625	53.103	18	57700
8489	70.333	20	15.444	59.875	50.46	13.5	44000
9554	85.25	79	44.225	74.688	40.137	17.1	70100
15287	65.25	42	26.913	70.75	28.276	14.4	71738
7057	55.25	17	24.379	59.063	44.251	21.2	58200
16848	77.75	48	26.69	75.938	27.187	9.2	63000
18211	91	87	76.681	80.625	51.164	12.8	74400
21561	69.25	58	44.702	76.25	26.689	9.2	75400
20667	65	68	22.995	75.625	28.038	11	66200
10684	61.75	26	8.774	66	33.99	9.5	52900
11738	74.25	32	25.449	66.875	27.701	12	63400
10107	74	43	11.315	71	29.096	16.2	66200
7817	65.75	36	33.709	64.25	52.548	17.7	54600
7050	26	11	0	55.313	55.651	18.8	59500
9082	83.5	73	64.668	77.375	43.185	13.6	66700
11706	60	56	16.937	73.75	39.479	12.7	62100
7643	49.25	23	36.635	62.813	39.302	18.7	57700
25734	90	77	67.758	80.938	44.133	10	80200
20155	86	84	69.31	79.688	48.766	17.6	74000
29852	94.5	84	75.009	81.313	51.363	10.6	74100
7980	68.5	34	9.122	63.875	35.294	16.3	53100

Row 1 of 170

data



structure

Success	0.2
Failure	0.8

Success	Success	Failure
Good	0.4	0.1
Moderate	0.4	0.3
Poor	0.2	0.6

numerical parameters

Hard problems

1. **Computation**: How do perform inference in general/flexible models?
2. **Modeling**: How to translate the complexity of a system into a manageable model?
3. **User interface**: How to show the results so that they are useful and make a difference?

Hard problems

- **Computation**
  - Exact inference in discrete BNs is worst-case NP-hard [Cooper 1990]
  - Exact inference in conditional linear Gaussian polytree is NP-hard [Lerner & Parr 2001]
  - Approximate inference in discrete BNs to any desired precision is also NP-hard [Dagum & Luby 1993]
  - Computation of MAP is NP-hard [Shimony 1994]
- **Modeling**
  - Practical models are reaching the size of thousands of variables: How do we build them and how do we reason with them?
  - Some real problems are more naturally represented by hybrid models (consisting of equations and discrete and continuous distributions)

Good dissertation problems

- Algorithms: Inference in hybrid models (mixtures of discrete and continuous variables and probability distributions).
  - Approximation of distributions (mixtures of normals, mixtures of truncated exponentials)
  - Stochastic sampling (most universal)
- Algorithms/philosophy: Causal graphs, causal discovery, causal reasoning.
  - Small data sets
  - Missing values
  - Mixtures of continuous and discrete data
  - Learning from temporal/time series data

Good dissertation problems

Effective user interfaces:

- **Knowledge engineering**
  - **support for interactive building of model structure**
  - **“canonical models” for parameter elicitation**
  - **computationally intensive techniques, such as sensitivity analysis**
  - **elicitation of continuous/parametric distributions**
- **Model exploration and explanation**
  - **model exploration (“instant gratification” interface)**
  - **graphical presentation of results**
  - **graphical and verbal explanation of results**

Good dissertation problems

Application of these methodologies to problems involving uncertainty.

For more: marek@sis.pitt.edu
http://www.pitt.edu/~druzdzel

Tools and resources

- All class materials will be available on my web site: http://www.pitt.edu/~druzdzel/phdopen/
- Software (GeNIe and SMILE©) is available at http://genie.sis.pitt.edu/

GeNIe and SMILE©

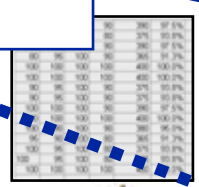
A developer's environment for graphical decision models (http://genie.sis.pitt.edu/).

Support for model building: ImaGeNIe

Diagnosis: Diagnosis

Qualitative interface: QGeNIe

Learning and discovery module: SMiner



Model developer module: GeNIe.

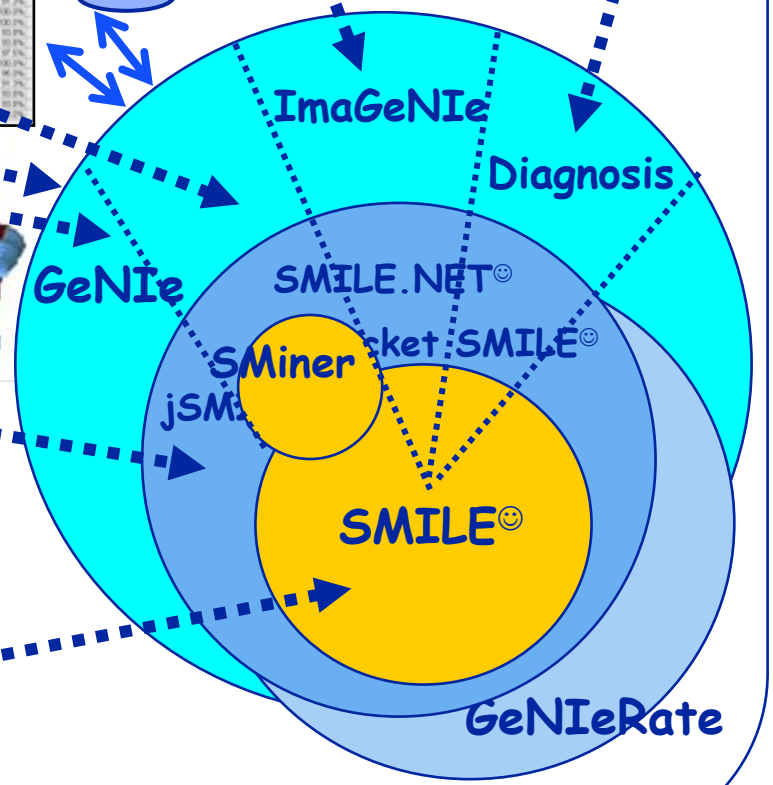
Implemented in Visual C++ in Windows environment.

Wrappers: SMILE.NET© jSMILE©, Pocket SMILE©

Allow SMILE© to be accessed from applications other than C++compiler

Reasoning engine: SMILE© (Structural Modeling, Inference, and Learning Engine).

A platform independent library of C++ classes for graphical models.



The remainder ☺