

Flight Data Analysis Using Bayesian Networks

Vasilis Megalooikonomou

(work with )

Introduction

- Objective: use data mining techniques to model the flight data and extract valuable information
 - What cause delay if the flights departure on time?
 - If large airplanes are delayed more than small ones?
- Dataset: 3 months flight data from Lockheed Martin contained in four tables (tracking, **flight**, plan, plan_point)
- Technique: Bayesian networks

What is a model?

- Probabilistic model of a restricted probabilistic domain

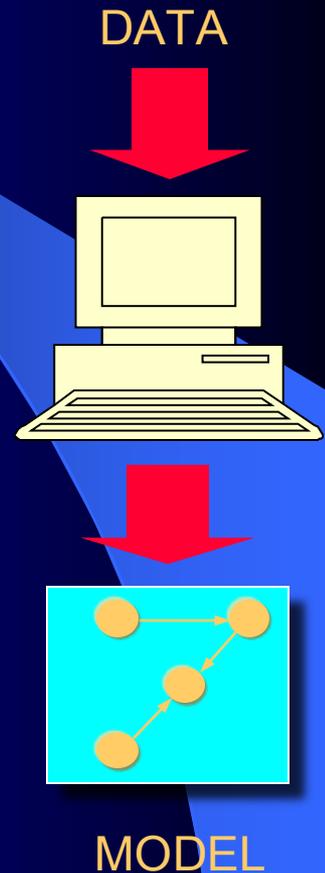
Input:

current knowledge (variable assignments)

Output:

$\Pr(\text{quantity of interest} \mid \text{current knowledge})$

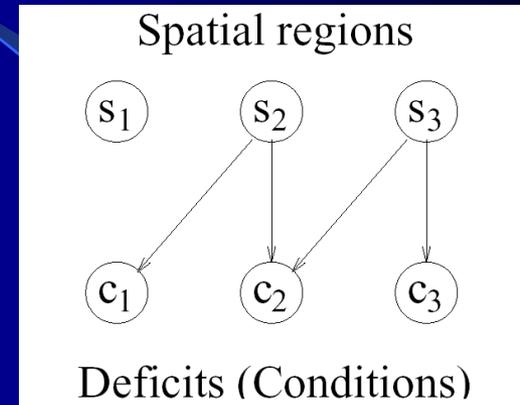
- Graphical models may show cause-effect or independence statements



Bayesian Networks

Bayesian Network Model:

- directed acyclic graph (DAG)
- represents the joint pdf of the domain, causal/independence information
- each node is a variable
- tables are conditional probabilities
- *conditional independence* principle

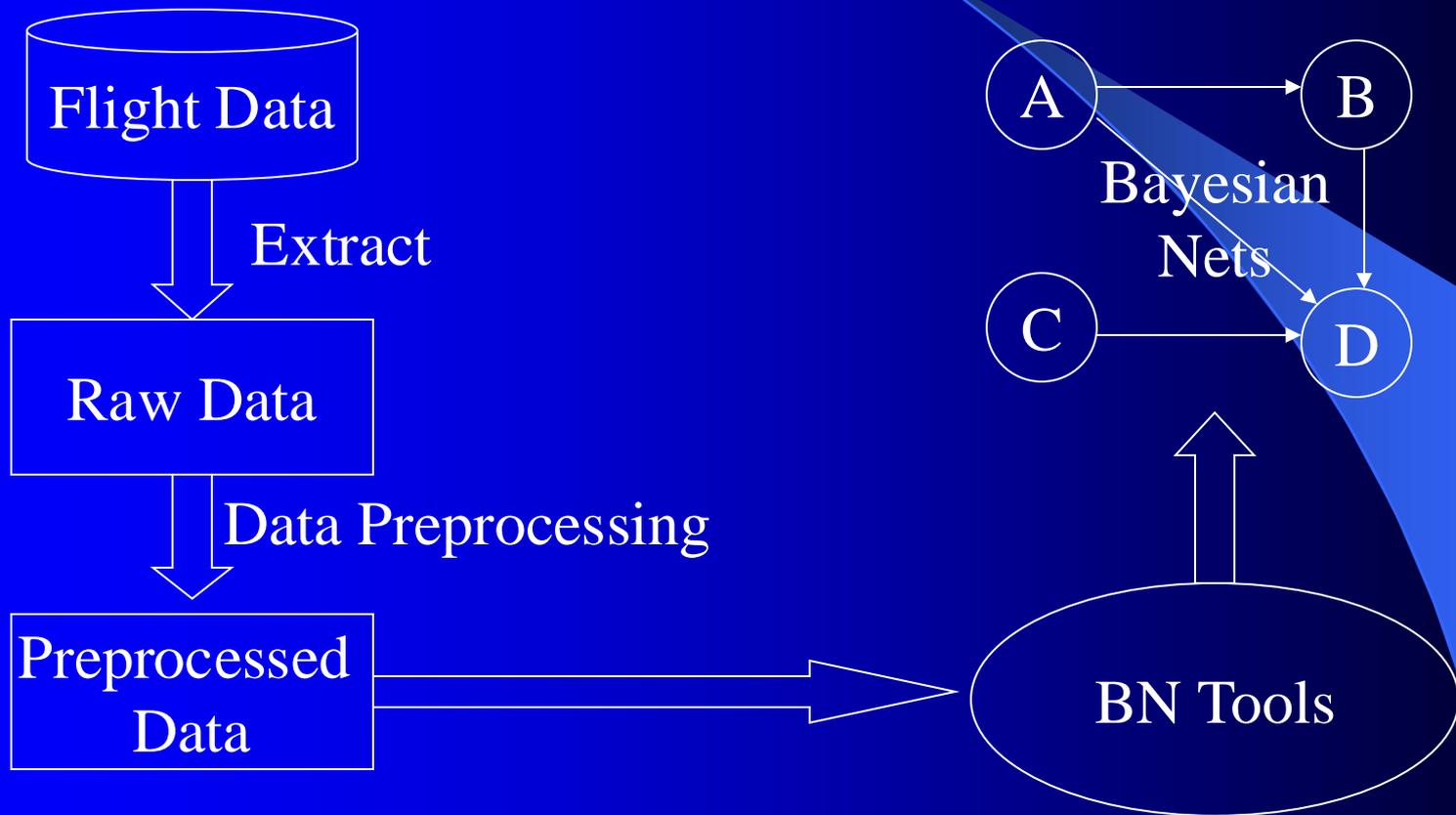


Advantages:

- mathematically sound (probability theory)
- visually expressive
- structurally meaningful (can represent generative processes)
- complete (any generative process)

node	conditional-probability table
s_1	$p(s_1) = 0.6$
s_2	$p(s_2) = 0.8$
s_3	$p(s_3) = 0.8$
c_1	$p(c_1 s_2) = 1.0, p(c_1 \overline{s_2}) = 0.6$
c_2	$p(c_2 s_2, s_3) = 0.4, p(c_2 s_2, \overline{s_3}) = 0.9,$ $p(c_2 \overline{s_2}, s_3) = 0.2, p(c_2 \overline{s_2}, \overline{s_3}) = 0.9$
c_3	$p(c_3 s_3) = 0.3, p(c_3 \overline{s_3}) = 0.9$

Overview



Data Preprocessing

- Eliminate unrelated attributes
 - attributes with only null values e.g. “a_out”
 - attributes with only one value e.g. “wx_alert”
- Incorporate experts suggestions
- Handle missing values
- Transform data
 - Convert Date type
 - Determine “DELAY” (assumption: flights depart on time)

Continuous:

Delay= arrival_delta – departure_delta

Discrete:

If arrival_delta – departure_delta > 5 then Delay=1

Else Delay=0

Data Preprocessing: Conversion

FLEET_ID – Airplane fleet identifier.

For each type of plane, decode as:

- numEngines an integer indicating the number of engines
- prop 1 if the enging type is prop, 0 otherwise
- turboProp 1 if the engine type is turbo prop, 0 otherwise
- jet 1 if the engine type is jet, 0 otherwise
- small 1 if the weight class is small, 0 otherwise
- large 1 if the weight class is large, 0 otherwise
- heavy 1 if the weight class is heavy, 0 otherwise
- climb the climb rate in feet per minute
- descend the descend rate in feet per minute
- catI 1 if the SRS category is I, 0 otherwise
- catII 1 if the SRS category is II, 0 otherwise
- catIII 1 if the SRS category is III, 0 otherwise

Discovering the Bayesian Network

- Split data into training set and testing set (70%, 30%)
- Edit a training plan
- Build the model
- Browse resulting Bayesian Net

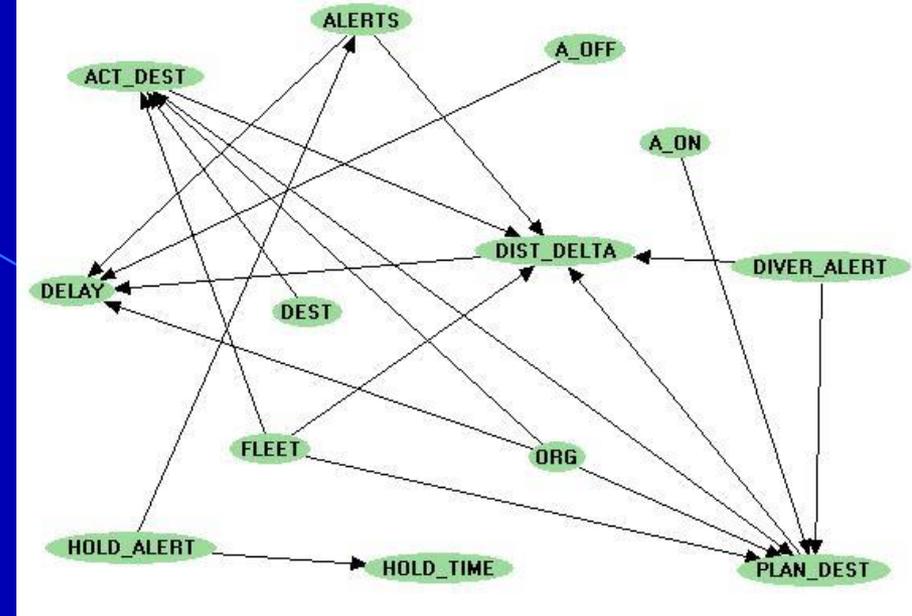
Experiment 1

- Dataset: First 200,000 records from table **Flight**.
- What causing delay if flight departs on time?
 - Distance_Delta
 - Plane.large
 - Hold_time
 - Dest_AP_ID
 - Plane.turboProp
- Whether the plane is large have more influence on delay than whether the plane is small.
- This model can be used for prediction

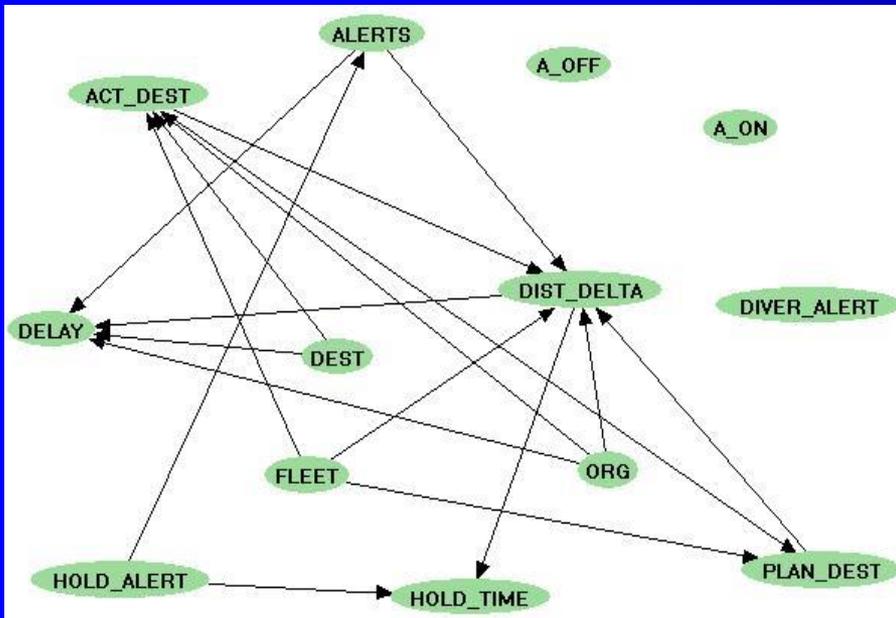
Experiment 2

- Analyze independently June, July, and August

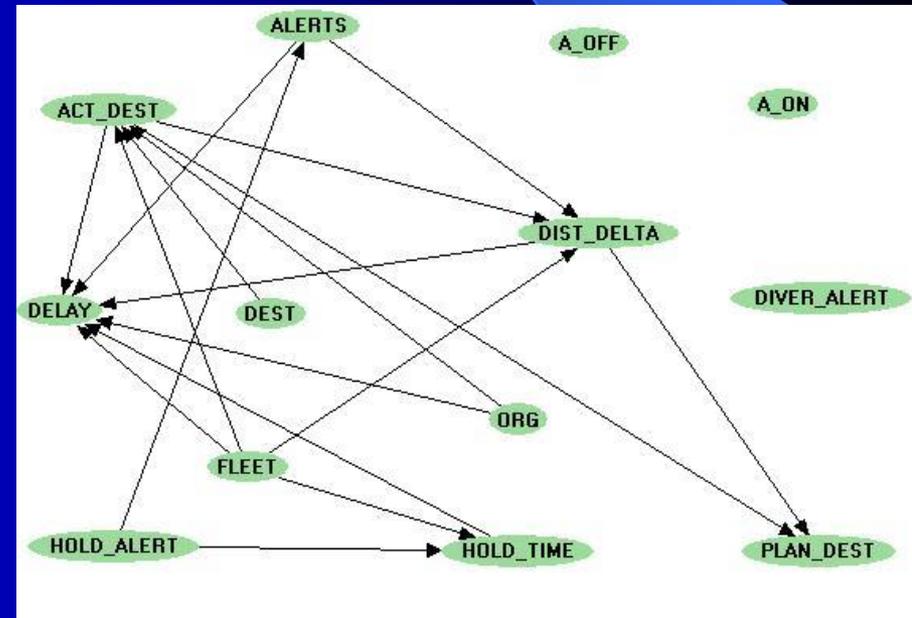
Dataset	Size
June	218,774
July	225,602
August	241,423



June



July



August

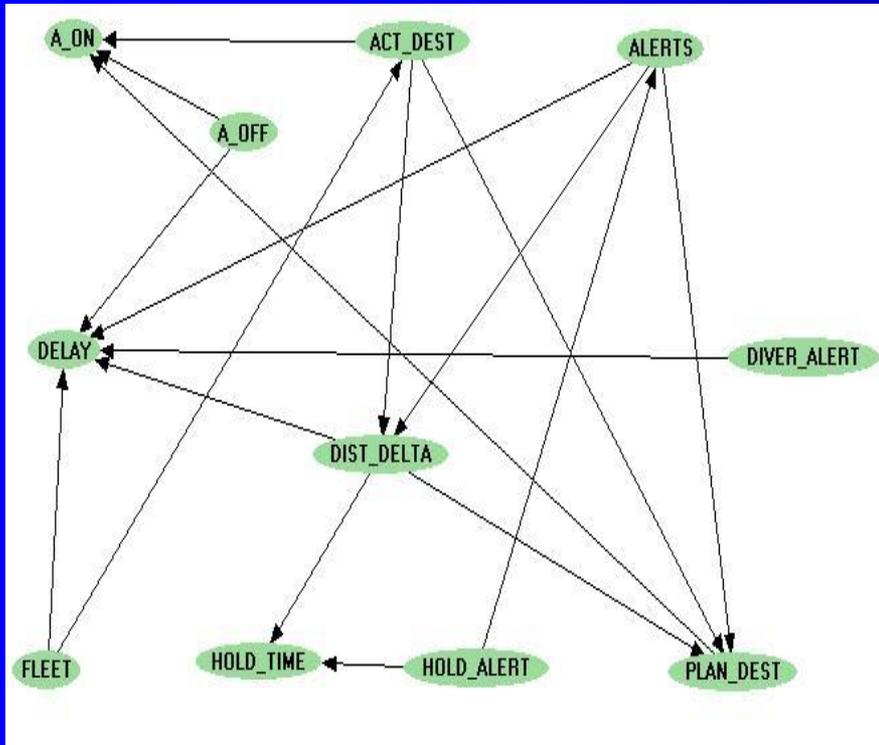
- Five strongest attributes causing DELAY

Strength	June	July	August
1	ALERTS	ALERTS	ALERTS
2	ORG	ORG	ORG
3	DIST_DELTA	DIST_DELTA	DIST_DELTA
4	A_OFF	FLEET	DEST
5	DEST	HOLD_TIME	HOLD_TIME

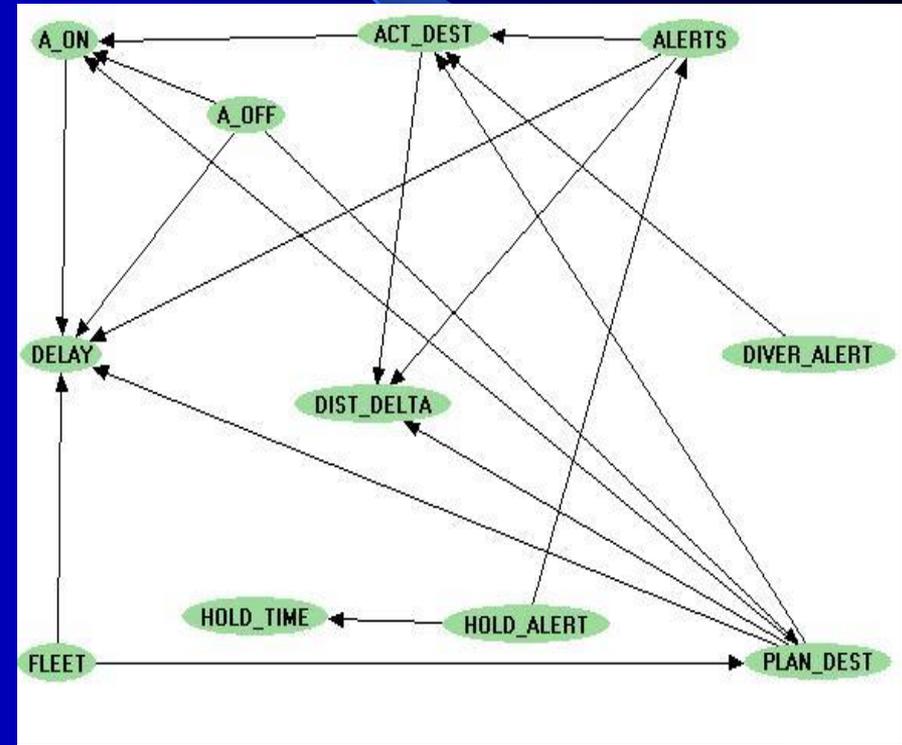
Experiment 3

- Compare resulting BNs from different data group by destination airport

Dataset	Size
To the most busy airport: DEN and ATL	103,456
To the airport with flights fewer than 2000	100,272



Busy Airports: DEN and ATL



Airports with few flights

Experiment 3

- First five strong attributes causing DELAY

Strength	Busy Airports	Airports with few flights
1	ALERTS	ALERTS
2	FLEET	A_ON
3	A_OFF	PLAN_DIST
4	DIST_DELTA	FLEET
5	DIVER_ALERT	A_OFF

Conclusion

- Bayesian Nets is a good technique to model the relationships among the attributes of flight data.
- Further work:
 - Include more attributes from other tables
 - Provide a comprehensive evaluation of this model
 - Integrate BN model with other data mining techniques (e.g., hypothesis testing, classification)
 - Incorporate spatial-temporal information (weather, etc)