In Pursuit of Patterns in Data

Reasoning from Data the Rough Set Way

Zdzisław Pawlak

Institute of Theoretical and Applied Informatics,
Polish Academy of Sciences,

zpw@ii.pw.edu.pl

MOTTO:

„It is a capital mistake to theorise before one has data"

Sherlock Holmes
In: A Scandal in Bohemia

AN EXAMPLE OF A DECISION TABLE

Fact no.	Driving conditions			Consequence	N
	weather	road	time	accident	
1	misty	icy	day	yes	80
2	foggy	icy	night	yes	140
3	misty	not icy	night	yes	40
4	sunny	icy	day	no	500
5	foggy	icy	night	no	20
6	misty	not icy	night	no	200

DECISION RULES

Let $S = (U, C, D)$ be a decision table. Every $x \in U$ determines a sequence $c_1(x), \ldots, c_n(x), d_1(x), \ldots, d_m(x)$ where $\{c_1, \ldots, c_n\} = C$ and $\{d_1, \ldots, d_m\} = D$

- The sequence will be called a *decision rule induced by* x (in S) and denoted by $c_1(x), \ldots, c_n(\mathrm{x}) \rightarrow d_1(x), \ldots, d_m(x)$ or in short $C \xrightarrow[x]{} D$
- The number $supp_x(C,D) = |C(x) \cap D(x)|$ will be called a *support* of the decision rule $C \xrightarrow[x]{} D$
- The number

$$\sigma_x(C,D) = \frac{supp_x(C,D)}{|U|}$$

will be referred to as the *strength* of the decision rule $C \xrightarrow[x]{} D$, where $|X|$ denotes the cardinality of X

CERTAINTY AND COVERAGE FACTORS

- **A *certainty factor* of the decision rule, denoted $cer_x(C, D)$ is defined as follows:**

$$cer_x(C,D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{\sigma_x(C,D)}{\pi(C(x))}$$

where $C(x) \neq \emptyset$ and $\pi(C(x)) = \frac{|C(x)|}{|U|}$

- **A *coverage factor* of the decision rule, denoted $cov_x(C, D)$ is defined as**

$$cov_x(C,D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{\sigma_x(C,D)}{\pi(D(x))}$$

where $D(x) \neq \emptyset$ and $\pi(D(x)) = \frac{|D(x)|}{|U|}$

INVERSE DECISION RULES

- **If $C \xrightarrow[x]{} D$ is a decision rule then $D \xrightarrow[x]{} C$ will be called an <u>inverse</u> <u>decision</u> <u>rule</u>**

- **The inverse decision rules can be used to give <u>explanation</u> (<u>reason</u>) for a decision**

CHARACTERIZATION OF DECISION RULES

Fact no.	*Strength*	*Certainty*	*Coverage*
1	0.082	1.000	0.308
2	0.143	0.877	0.538
3	0.041	1.167	0.154
4	0.510	1.000	0.695
5	0.020	0.123	0.027
6	0.204	0.833	0.278

PROPERTIES OF DECISION RULES

Let $C \xrightarrow[x]{} D$ be a decision rule. Then the following properties are valid:

$$\sum_{y \in C(x)} cer_y(C,D) = 1 \qquad (1)$$

$$\sum_{y \in D(x)} cov_y(C,D) = 1 \qquad (2)$$

$$\pi(D(x)) = \sum_{y \in C(x)} cer_y(C,D) \cdot \pi(C(y)) = \sum_{y \in C(x)} \sigma_y(C,D) \qquad (3)$$

$$\pi(C(x)) = \sum_{y \in D(x)} cov_y(C,D) \cdot \pi(D(y)) = \sum_{y \in D(x)} \sigma_y(C,D) \qquad (4)$$

$$cer_x(C,D) = \frac{cov_x(C,D) \cdot \pi(D(x))}{\pi(C(x))} = \frac{\sigma_x(C,D)}{\pi(C(x))} \qquad (5)$$

$$cov_x(C,D) = \frac{cer_x(C,D) \cdot \pi(C(x))}{\pi(D(y))} = \frac{\sigma_x(C,D)}{\pi(D(x))} \qquad (6)$$

GRANULARITY OF DATA AND FLOW GRAPHS

- **With every decision table we associate a <u>flow</u> <u>graph</u>**

- **To every decision rule $C \xrightarrow[x]{} D$ we assign a <u>directed</u> <u>branch</u> x connecting the <u>input</u> <u>node</u> $C(x)$ and the <u>output</u> <u>node</u> $D(x)$**

- **Strength of the decision rule represents a <u>throughflow</u> of the corresponding branch.**

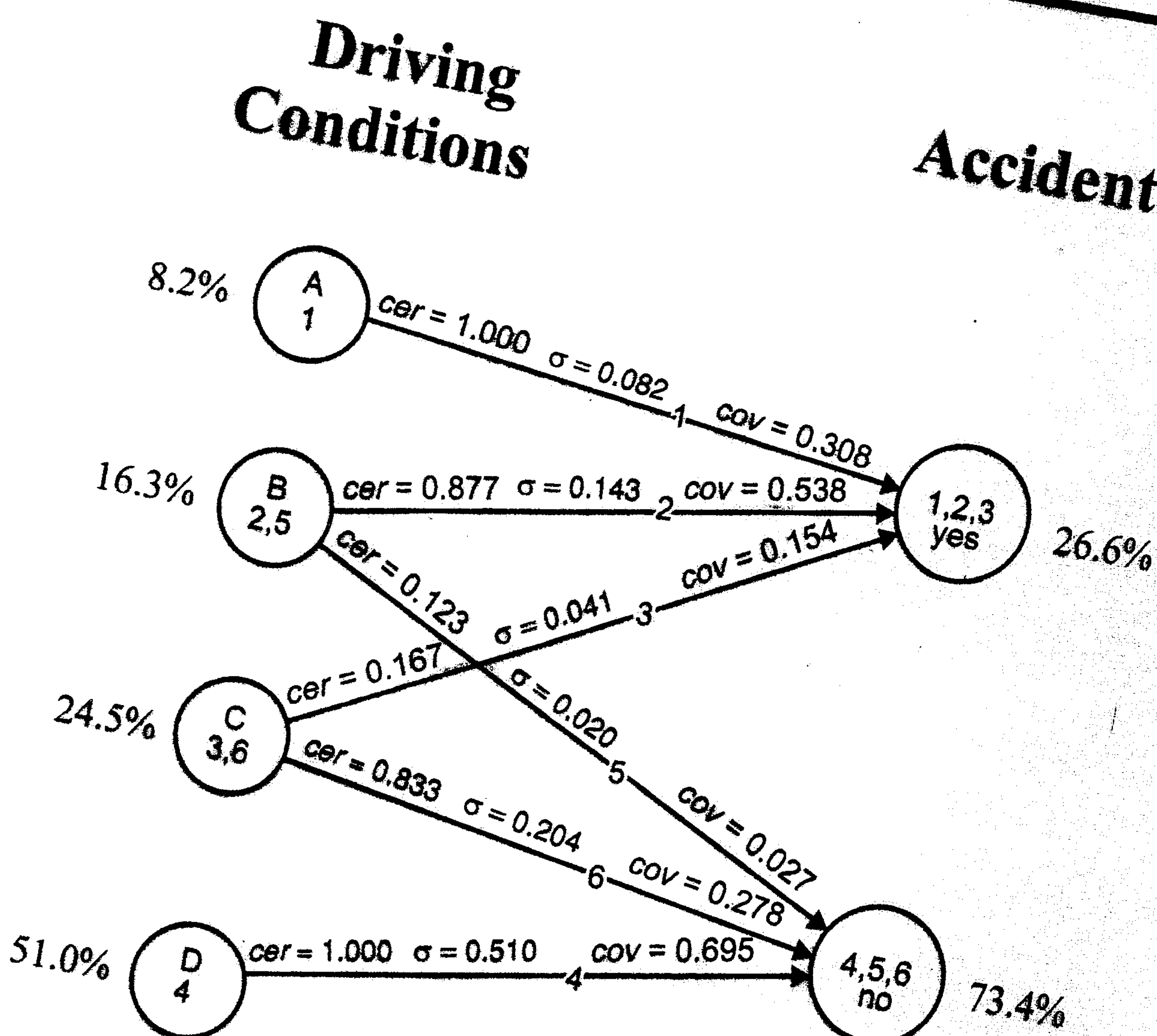
- **The throughflow of the graph is governed by formulas (1),....,(6)**

DECISIONS AND FLOW

- **Classification of objects boils down to finding the <u>maximal</u> <u>output</u> <u>flow</u> in the flow graph**

- **Explanation of decisions is connected with the <u>maximal</u> <u>input</u> <u>flow</u> associated with the given decision**

FLOW GRAPH


Driving Conditions
Accident
8.2%
A
1
16.3%
B
2,5
24.5%
C
3,6
51.0%
D
4
1,2,3
yes
26.6%
4,5,6
no
73.4%
cer = 1.000 σ = 0.082 1 cov = 0.308
cer = 0.877 σ = 0.143 2 cov = 0.538
cer = 0.167 σ = 0.041 3 cov = 0.154
cer = 0.123 σ = 0.020 5 cov = 0.027
cer = 0.833 σ = 0.204 6 cov = 0.278
cer = 1.000 σ = 0.510 4 cov = 0.695

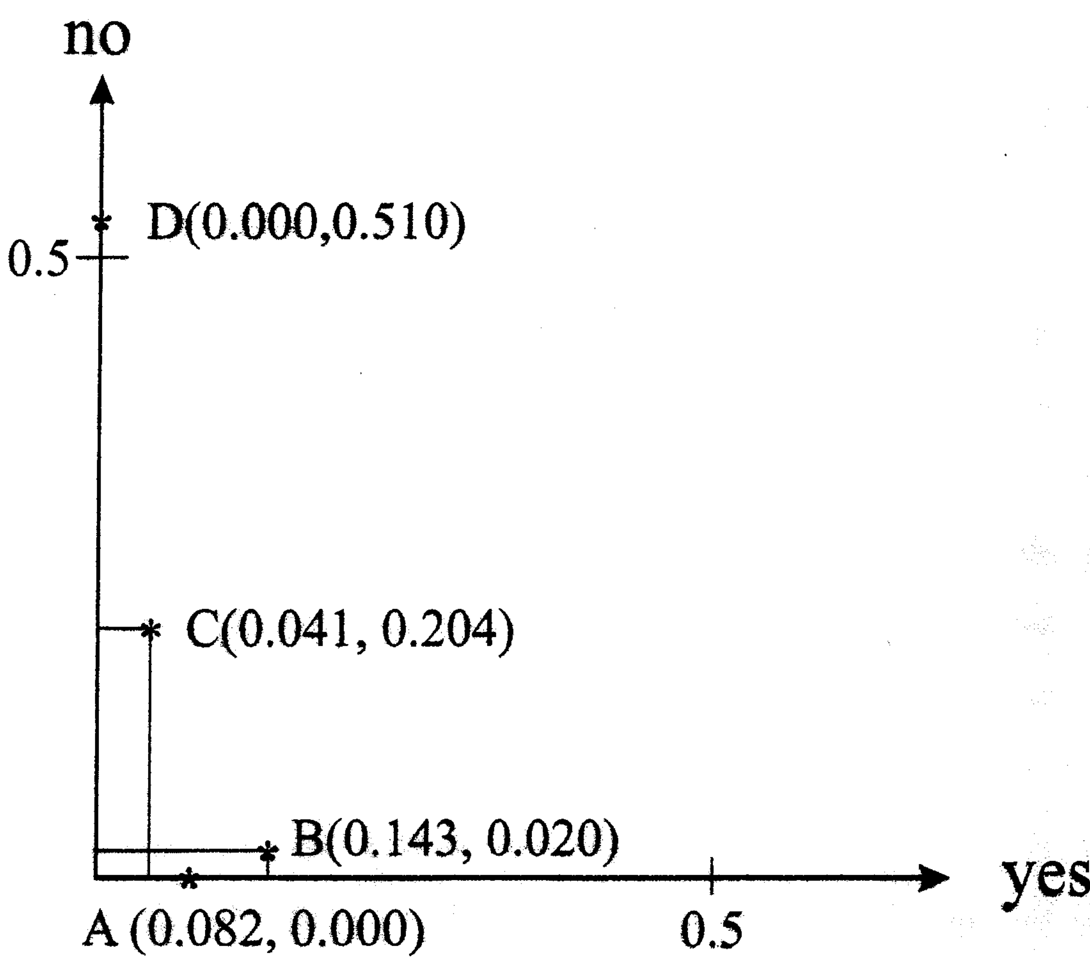
DECISION SPACE

- With every decision table with one n-valued decision attribute we associate n-dimensional Euclidean space
- Decision granules determine n axis of the space
- Condition granules determine points of the space
- Strengths of decision rules are coordinates of granules
- Distance $\delta(x, y)$ between granules x and y is defined as

$$\delta(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where $x = (x_1,\ldots,x_n)$ and $y = (y_1,\ldots,y_n)$

DECISION SPACE



DISTANCE MATRIX

	A	B	C	D
A				
B	0.064			
C	0.208	0.210		
D	0.517	0.510	0.309	

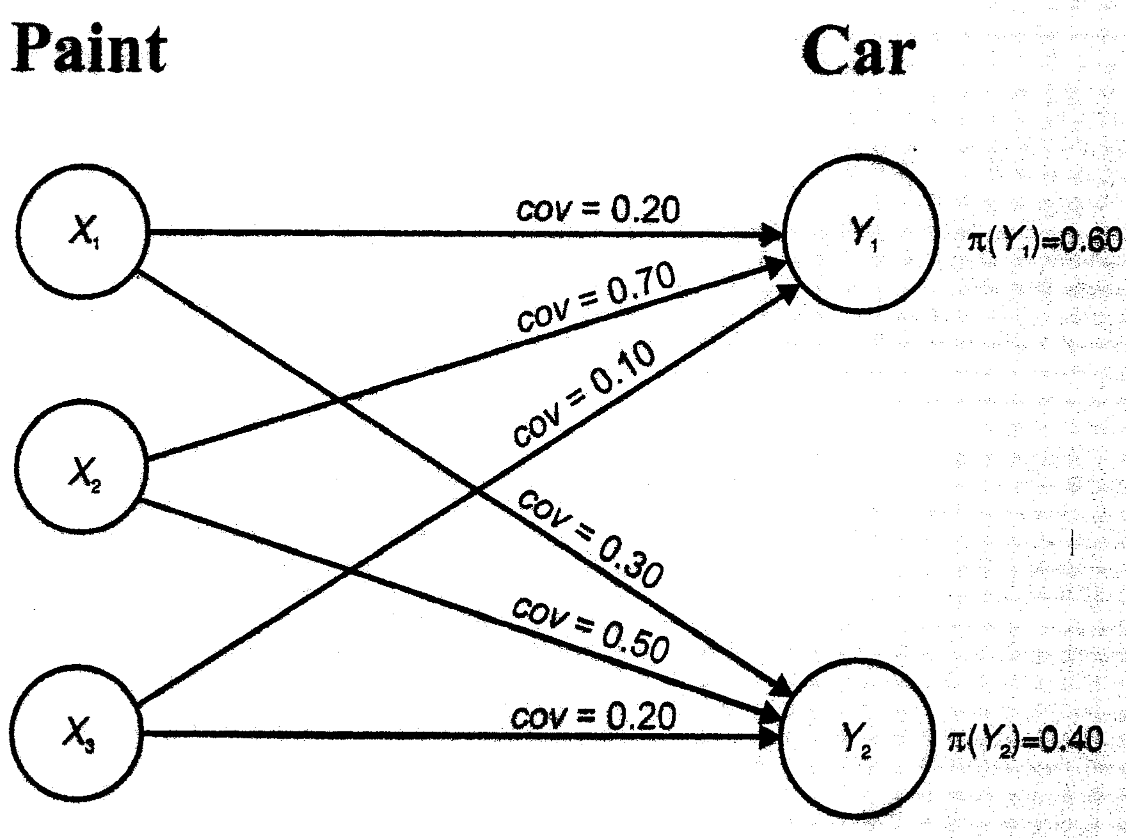
SUPPLY – DEMAND

Suppose that cars are painted into two colors Y_1 and Y_2 and that these colors can be obtained by mixing three paints X_1, X_2 and X_3 in the following proportions:

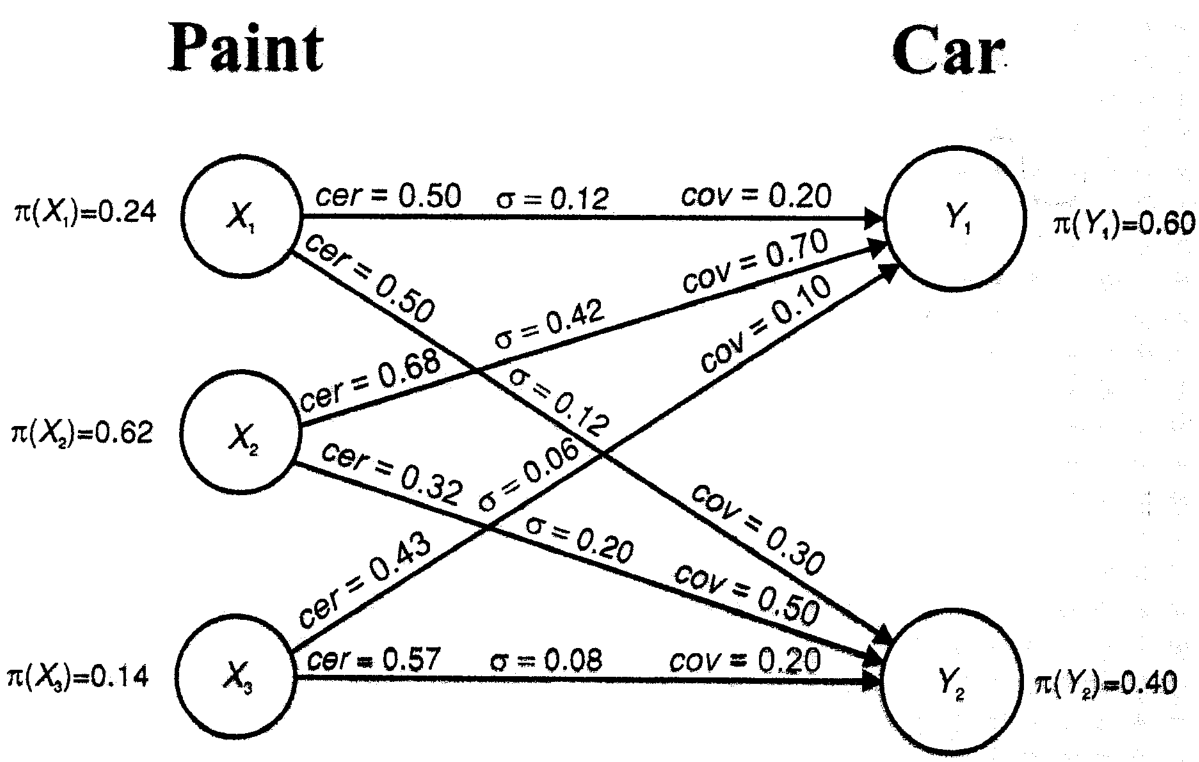
- **Y_1 contains 20% of X_1, 70% of X_2 and 10% of X_3**
- **Y_2 contains 30% of X_1, 50% of X_2 and 20% of X_3**

We have to find demand of each paint and their distribution among colors Y_1 and Y_2

SUPPLY – DEMAND



FINAL RESULTS



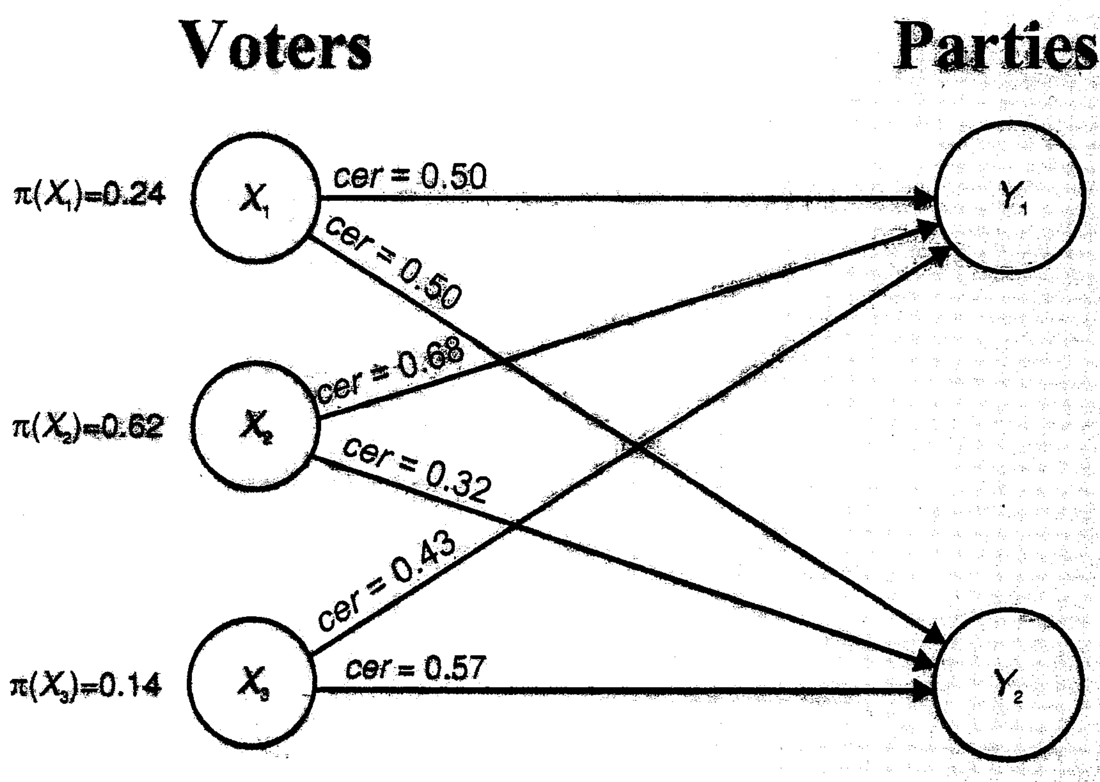
INVERSE PROBLEM

- **Distribution of votes of three disjoint group X_1, X_2 and X_3 of voters among two political parties Y_1 and Y_2**

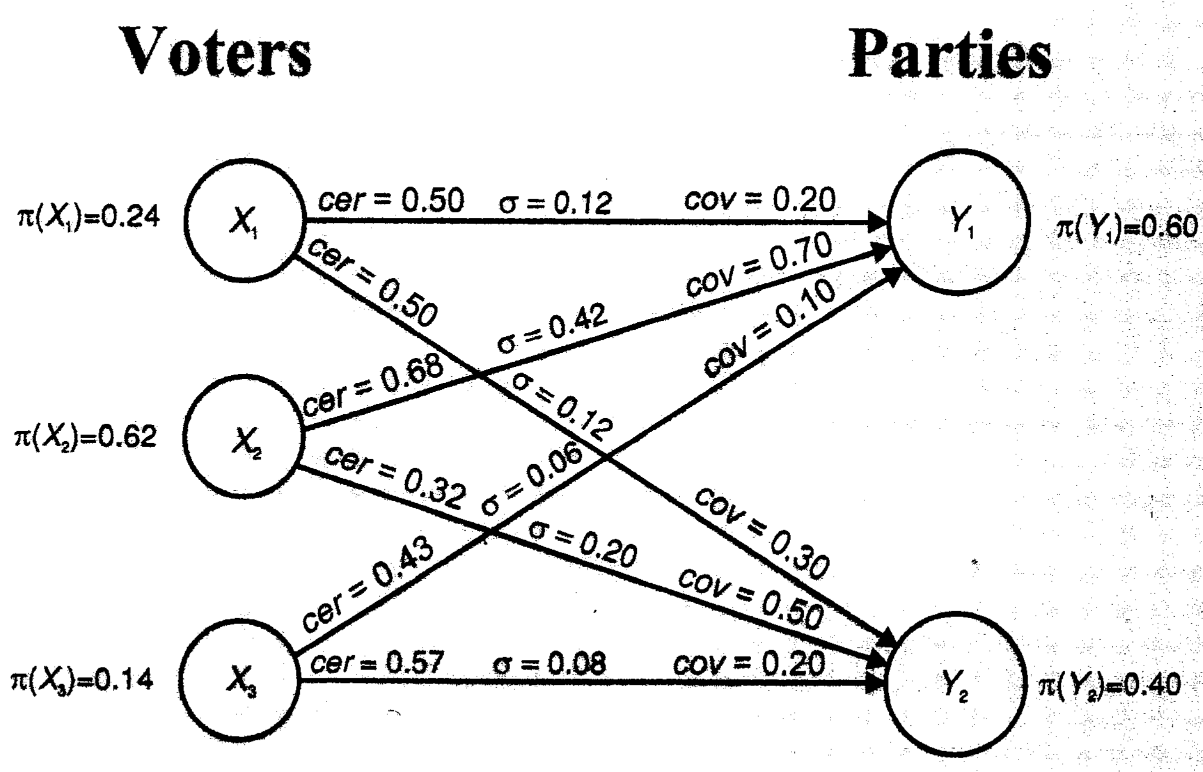
- **X_1 consists of 24% of voters, X_2 – 62% and X_3 – 14%**

- **Votes distribution among parties is as follows:**
  - **group X_1 gave 50% of its votes for each party**
  - **group X_2 gave 68% of votes for party Y_1 and 32% for party Y_2**
  - **group X_3 gave 43% votes for party Y_1 and 57% votes for party Y_2**

VOTING ANALYSIS



FINAL RESULTS



FINAL RESULTS

- Party Y_1 obtained 60% votes
- Party Y_2 obtained 40% votes
- Votes distribution for each party
  - Party Y_1 obtained
    - 20% votes from group X_1,
    - 70% from group X_2 and
    - 10% from group X_3
  - Party Y_2 obtained
    - 30% votes from group X_1,
    - 50% from group X_2 and
    - 20% from group X_3

FLOW GRAPH

A flow graph is a <u>directed</u>, <u>acyclic</u> finite graph

$$G = (N, \mathcal{B}, \varphi)$$

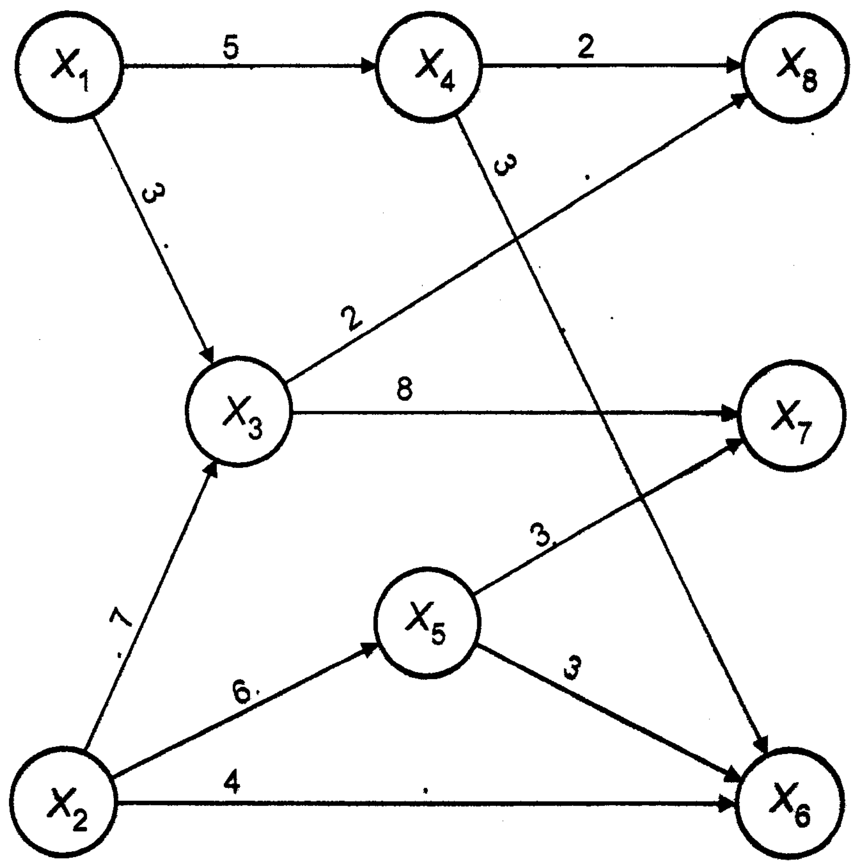
where

N **– set of nodes**

$\mathcal{B} \subseteq N \times N$ **set of <u>directed</u> <u>branches</u>**

$\varphi : \mathcal{B} \rightarrow R^{+}$ **– <u>flow</u> <u>function</u>**

A FLOW GRAPH



flow conservation
inflow = outflow

INPUTS AND OUTPUTS

- **Input** of $x \in N$

$$I(x)=\{y \in N : (y, x) \in \mathscr{B}\}$$

- **Output** of $x \in N$

$$O(x) = \{y \in N : (x, y) \in \mathscr{B}\}$$

- **Input** of G

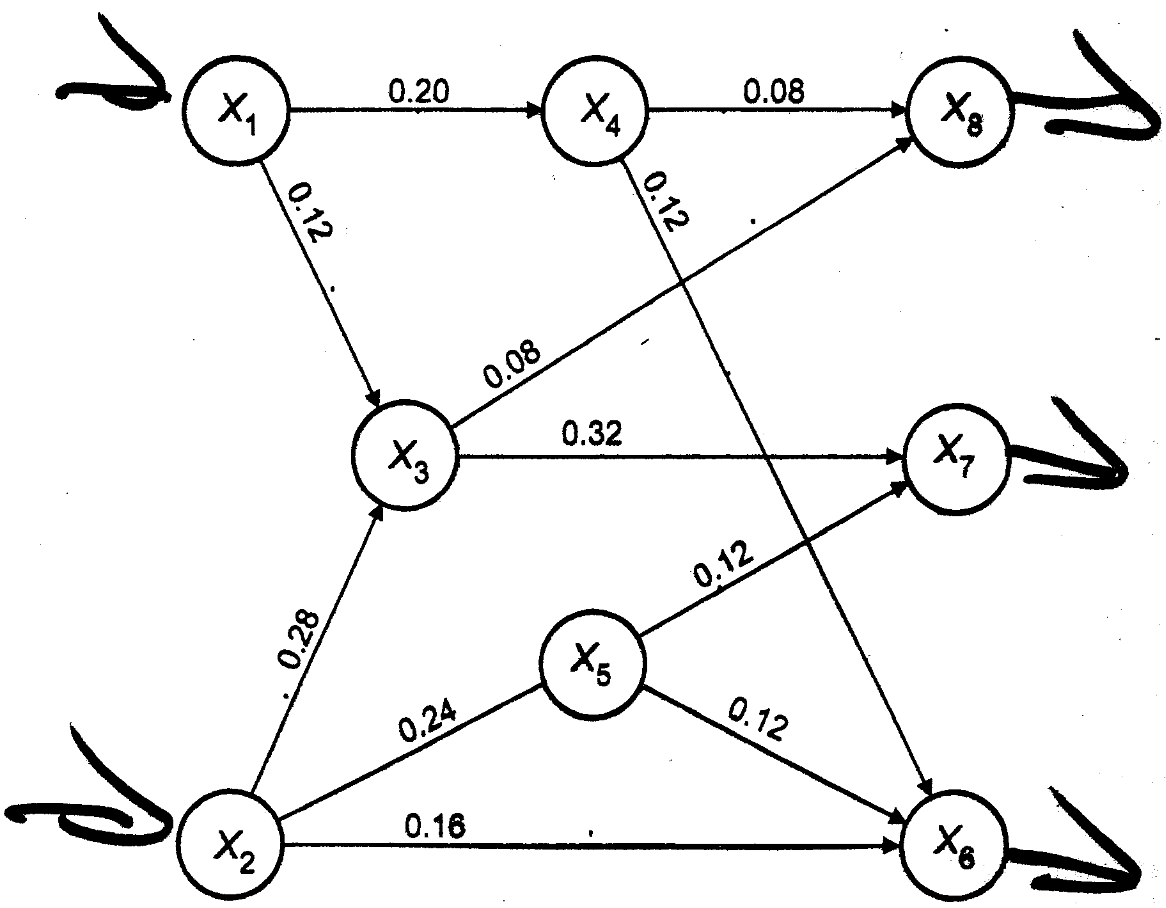
$$I(G) = \{x \in N : I(x) = \varnothing\}$$

- **Output** of G

$$O(G) = \{x \in N : O(x) = \varnothing\}$$

- **Inputs and outputs of G are external nodes of G; other nodes are internal nodes of G**

NORMALIZED FLOW GRAPH



$$\textbf{normalized flow} \equiv \frac{\textbf{flow}}{\textbf{total flow}}$$

FLOW

- If $(x, y) \in \mathscr{B}$ then $\varphi(x, y)$ is troughflow from x to y

- $\varphi_+(y) = \sum_{x \in I(y)} \varphi(x, y)$ is an inflow of y

- $\varphi_-(x) = \sum_{y \in O(x)} \varphi(x, y)$ is an outflow of x

- $\varphi_+(G) = \sum_{x \in I(G)} \varphi_-(x)$ is an inflow of G

- $\varphi_-(G) = \sum_{x \in O(G)} \varphi_+(x)$ is an outflow of G

FLOW CONSERVATION

- **We assume that for any internal node**

$$\boxed{\varphi_+(x) = \varphi_-(x) = \varphi(x)}$$

$\varphi(x)$ – **troughflow of** x

- **Consequently**

$$\boxed{\varphi_+(G) = \varphi_-(G) = \varphi(G)}$$

$\varphi(G)$– **troughflow of** G

STRENGTH, CERTAINTY AND COVERAGE OF FLOW

- The strength of (x, y)

$$\sigma(x, y) = \frac{\varphi(x, y)}{\varphi(G)}$$

- The certainty of (x, y)

$$cer(x, y) = \frac{\sigma(x, y)}{\sigma(x)}$$

- The coverage of (x, y)

$$cov(x, y) = \frac{\sigma(x, y)}{\sigma(y)}$$

- The normalized troughflow of x

$$\sigma(x) = \sum_{y \in O(x)} \sigma(x, y) = \sum_{y \in I(x)} \sigma(y, x)$$

PROPERTIES OF FLOW

- $\sum_{y \in O(x)} cer(x, y) = 1$ (1)

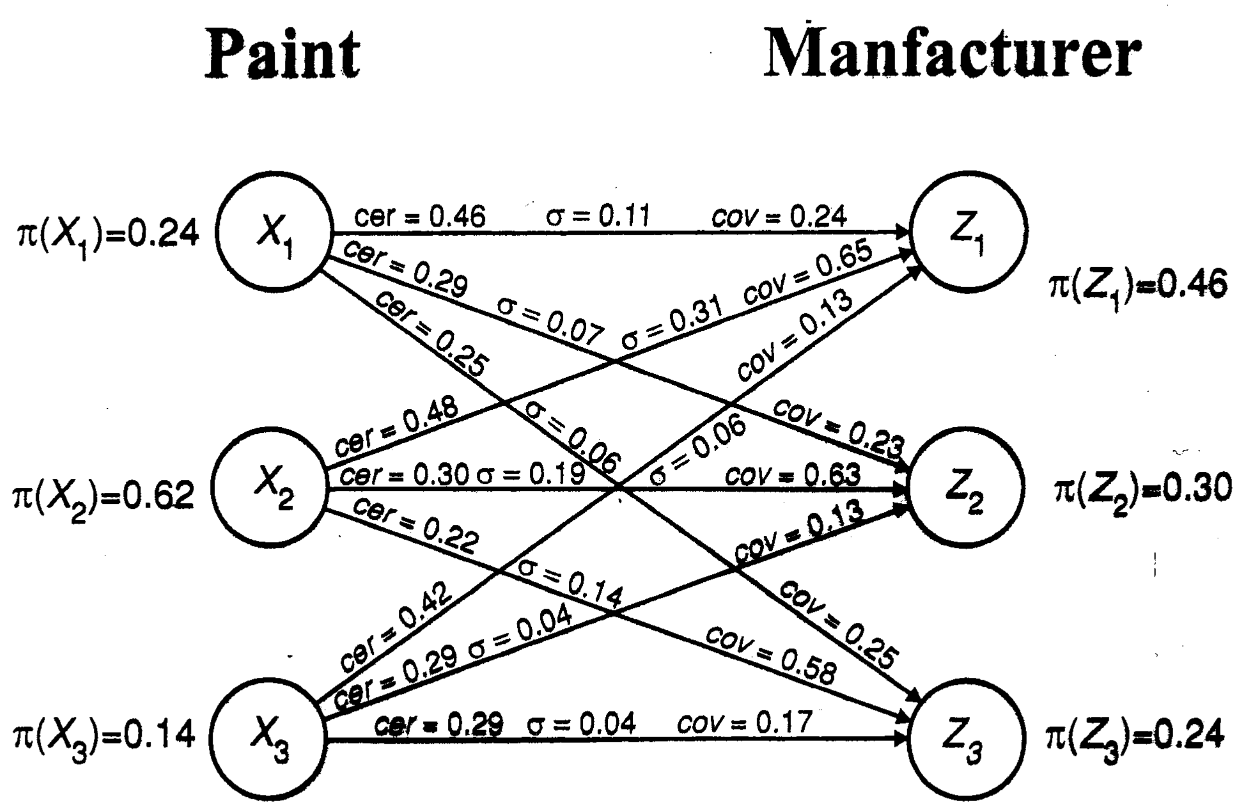
- $\sum_{x \in I(y)} cov(x, y) = 1$ (2)

- $cer(x, y) = \dfrac{cov(x, y)\sigma(y)}{\sigma(x)}$ (3)

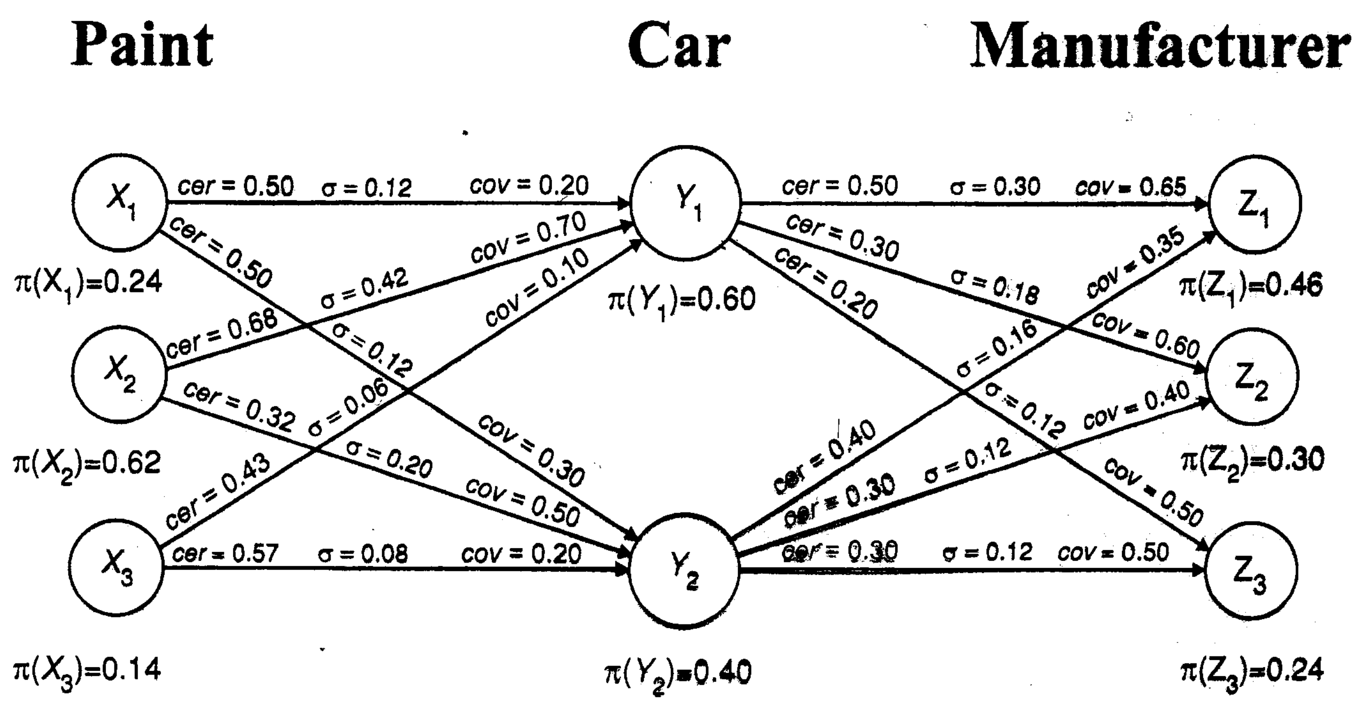
- $cov(x, y) = \dfrac{cer(x, y)\sigma(x)}{\sigma(y)}$ (4)

(3) and (4) are Bayes' formulas

SUPPLY – DEMAND SIMPLIFIED GRAPH



SUPPLY – DEMAND EXTENDED GRAPH



PATH

- A (<u>directed</u>) <u>path</u> from x to y, $x \neq y$ denoted $[x, y]$, is a sequence of nodes $x_1, \ldots, x_n$ such that $x_1 = x$, $x_n = y$ and $(x_i, x_{i+1}) \in \mathscr{B}$ for every i, $1 \leq i \leq n-1$

- The <u>certainty</u> of $[x_1, x_n]$

$$cer[x_1, x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1})$$

- The <u>coverage</u> of $[x_1, x_n]$

$$cov[x_1, x_n] = \prod_{i=1}^{n-1} cov(x_i, x_{i+1})$$

- The <u>strength</u> of $[x, y]$

$$\sigma[x, y] = \sigma(x)\, cer[x, y] = \sigma(y)\, cov[x, y]$$

CONNECTIONS

- **The set of all paths from x to y ($x \neq y$) denoted $< x, y >$, will be called a <u>connection</u> from x to y**

- **The <u>certainty</u> of $< x, y >$**

$$cer < x, y > = \sum_{[x,y] \in < x,y >} cer[x, y]$$

- **The <u>coverage</u> of $< x, y >$**

$$cov < x, y > = \sum_{[x,y] \in < x,y >} cov[x, y]$$

- **The <u>strength</u> of $< x, y >$**

$$\sigma < x, y > = \sum_{[x,y] \in < x,y >} \sigma[x, y]$$

THE RULE OF SUBSTITUTION

Let x, y $(x \neq y)$ be nodes of G. If we substitute the subgraph $< x, y >$ by a single branch (x, y) such that

$$\sigma(x, y) = \sigma < x, y >$$

then

$$cer(x, y) = cer < x, y >$$

$$cov(x, y) = cov < x, y >$$

and

$$\varphi(G) = \varphi(G')$$

where G' is the graph obtained from G by substituting $< x, y >$ by (x, y)

DECISION TABLES

	Paint	Car	Strength
1	X_1	Y_1	0.12
2	X_1	Y_2	0.12
3	X_2	Y_1	0.42
4	X_2	Y_2	0.20
5	X_3	Y_1	0.06
6	X_3	Y_2	0.08

	Car	Manu.	Strength
1	Y_1	Z_1	0.30
2	Y_1	Z_2	0.18
3	Y_1	Z_3	0.12
4	Y_2	Z_1	0.16
5	Y_2	Z_2	0.12
6	Y_2	Z_3	0.12

	Paint	Manu.	Strength
1	X_1	Z_1	0.11
2	X_1	Z_1	0.08
3	X_1	Z_3	0.06
4	X_2	Z_1	0.29
5	X_2	Z_2	0.18
6	X_2	Z_3	0.14
7	X_3	Z_1	0.06
8	X_3	Z_2	0.04
9	X_3	Z_3	0.04

SUMMARY

- Flow graphs can be used to decision analysis
-
- Flow in the graph represents strength of decisions
- Relation between decisions is expressed by Bayes' formula
- In this approach Bayes' formula has entirely deterministic character
- The presented approach leads to new computational algorithms and a new look on Bayesian methodology