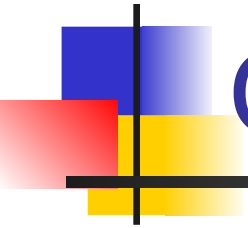
Capturing Independence Graphically; Undirected Graphs



COMPSCI 276, Spring 2014

Set 2: Rina Dechter

(Reading: Pearl chapters 3, Darwiche chapter 4)

Constraint Networks

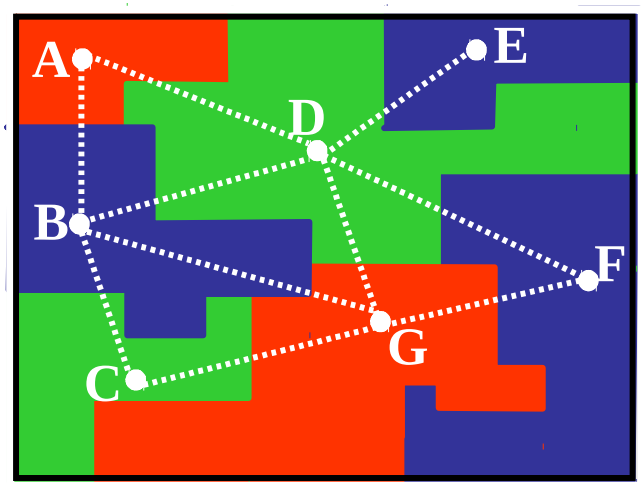
Example: map coloring

Variables - countries (A,B,C,etc.)

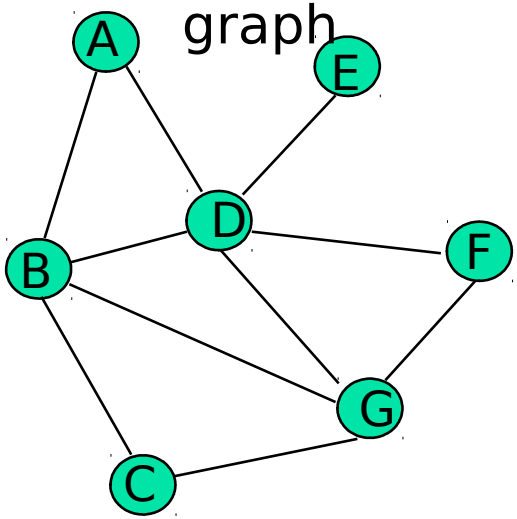
Values - colors (red, green, blue)

Constraints: $A \neq B,\ A \neq D,\ D \neq E,$ *etc.*

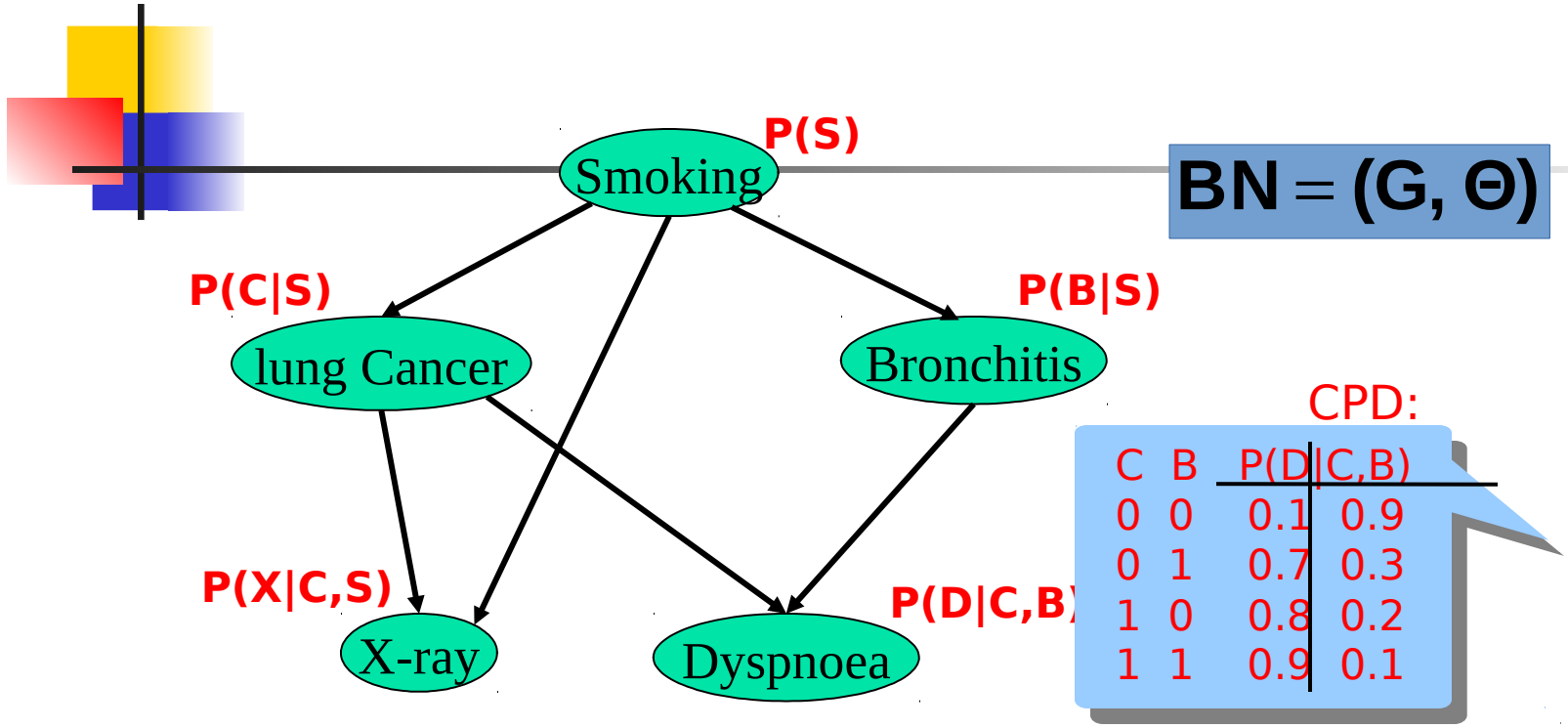
A	B
red	green
red	yellow
green	red
green	yellow
yellow	green
yellow	red



Constraint graph



Bayesian Networks (Pearl 1988)



P(S, C, B, X, D) = ***P(S) P(C|S) P(B|S) P(X|C,S) P(D|C,B)***

Combination: Product
Marginalization: sum/max

$$P(x_1 \ldots x_n) = \prod_i p(x_i \mid pa(x_i))$$

$$P(e) = \sum_{X-E} \prod_i p(x_i \mid pa(x_i))$$

$$mpe = \max_x P(x)$$

Sample Applications for Graphical Models

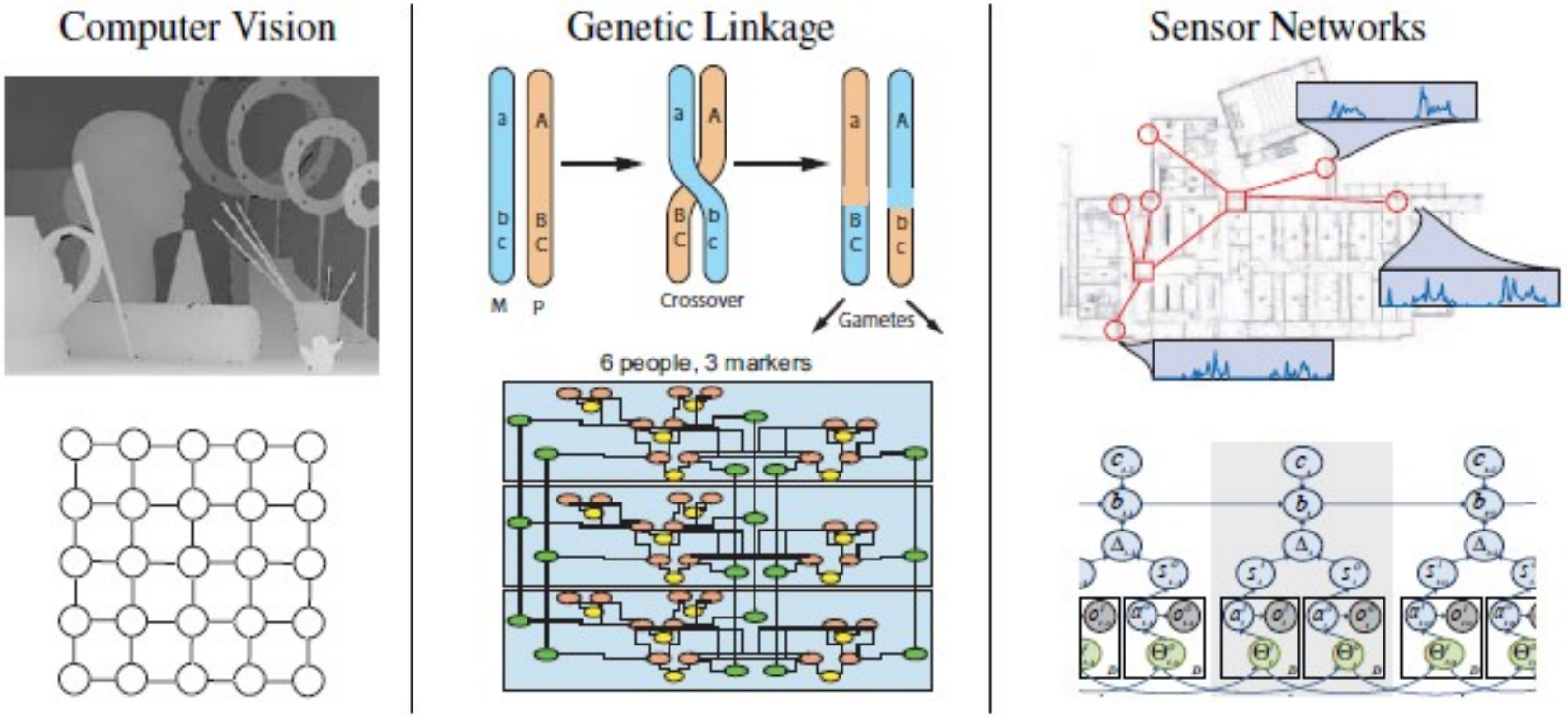


Figure 1: Application areas and graphical models used to represent their respective systems: (a) Finding correspondences between images, including depth estimation from stereo; (b) Genetic linkage analysis and pedigree data; (c) Understanding patterns of behavior in sensor measurements using spatio-temporal models.

The Qualitative Notion of Depedence

Motivations and issues

- Motivating example:
- What I eat for breakfast, what I eat for dinner?
- What I eat for breakfast, What I dress
- What I eat for breakfast today, the grade in 276
- The time I work on homework 1, my grade in 276
- Shoe size,reading ability
- Shoe-size, reading ability, if we know the age

The Qualitative Notion of Depedence

motivations and issues

- The traditional definition of independence uses equality of numerical quantities as in P(x,y)=P(x)P(y)
- People can easily and confidently detect dependencies, but not provide numbers
- The notion of relevance and dependence are far more basic to human reasoning than the numerical quantification.
- Assertions about dependency relationships should be expressed first.

Dependency graphs

- The nodes represent propositional variables and the arcs represent local dependencies among conceptually related propositions.
- Graph concepts are entrenched in our language (e.g., “thread of thoughts”, “lines of reasoning”, “connected ideas”). One wonders if people can reason any other way except by tracing links and arrows and paths in some mental representation of concepts and relations.
- What types of (in)dependencies are deducible from graphs?
- For a given probability distribution P and any three variables **X,Y,Z**,it is straightforward to verify whether knowing Z renders X independent of Y, but P does not dictates which variables should be regarded as neighbors.
- Some useful properties of dependencies and relevancies cannot be represented graphically.

Variable Independence

$\Pr$ finds $\mathbf{X}$ independent of $\mathbf{Y}$ given $\mathbf{Z}$, denoted $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, means that $\Pr$ finds $\mathbf{x}$ independent of $\mathbf{y}$ given $\mathbf{z}$ for all instantiations $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

Example

$\mathbf{X} = \{A, B\}$, $\mathbf{Y} = \{C\}$ and $\mathbf{Z} = \{D, E\}$, where A, B, C, D and E are all propositional variables. The statement $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is then a compact notation for a number of statements about independence:

$A \wedge B$ is independent of C given $D \wedge E$;

$A \wedge \neg B$ is independent of C given $D \wedge E$;

$\vdots$

$\neg A \wedge \neg B$ is independent of $\neg C$ given $\neg D \wedge \neg E$;

That is, $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is a compact notation for $4 \times 2 \times 4 = 32$ independence statements of the above form.

Properties of Probabilistic independence

THEOREM 1: Let X, Y, and Z be three disjoint subsets of variables from U. If $I(X, Z, Y)$ stands for the relation "X is independent of Y, given Z" in some probabilistic model P, then I must satisfy the following four independent conditions:

If Probabilistic independence is a good (intuitive to human reasoning) formalizm, then the axioms it obeys will be consistent with our intuition

Properties of Probabilistic independence

THEOREM 1: Let X, Y, and Z be three disjoint subsets of variables from U. If $I(X, Z, Y)$ stands for the relation "X is independent of Y, given Z" in some probabilistic model P, then I must satisfy the following four independent conditions:

- Symmetry:
  - **I(X,Z,Y) ⇔ I(Y,Z,X)**
- Decomposition:
  - **I(X,Z,YW)⇒ I(X,Z,Y) and I(X,Z,W)**
- Weak union:
  - **I(X,Z,YW)⇒I(X,ZW,Y)**
- Contraction:
  - **I(X,Z,Y) and I(X,ZY,W)⇒I(X,Z,YW)**
- Intersection:
  - **I(X,ZY,W) and I(X,ZW,Y) ⇒ I(X,Z,YW)**

Decomposition

If some information is irrelevant, then any part of it is also irrelevant.

$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$ only if $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{W})$.

If learning **yw** does not influence our belief in **x**, then learning **y** alone, or learning **w** alone, will not influence our belief in **x** either.

Pearl language:
If two pieces of information are irrelevant to X then each one is irrelevant to X

Decomposition

The opposite of Decomposition, called Composition:

$$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{W}) \Longrightarrow I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

does not hold in general.

Two pieces of information may each be irrelevant on their own, yet their combination may be relevant.

Example: Two coins and a bell

Contraction

$$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W}) \Rightarrow I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

If after learning the irrelevant information **y**, the information **w** is found to be irrelevant to our belief in **x**, then the combined information **yw** must have been irrelevant from the beginning.

Compare Contraction with Composition:

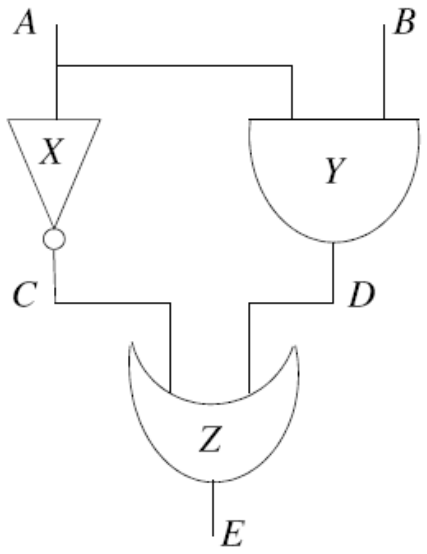
$$I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{W}) \Rightarrow I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

One can view Contraction as a weaker version of Composition.
Recall that Composition does not hold for probability distributions.

Strictly Positive Distributions

Definition

A strictly positive distribution assign a non-zero probability to every consistent event.



Example

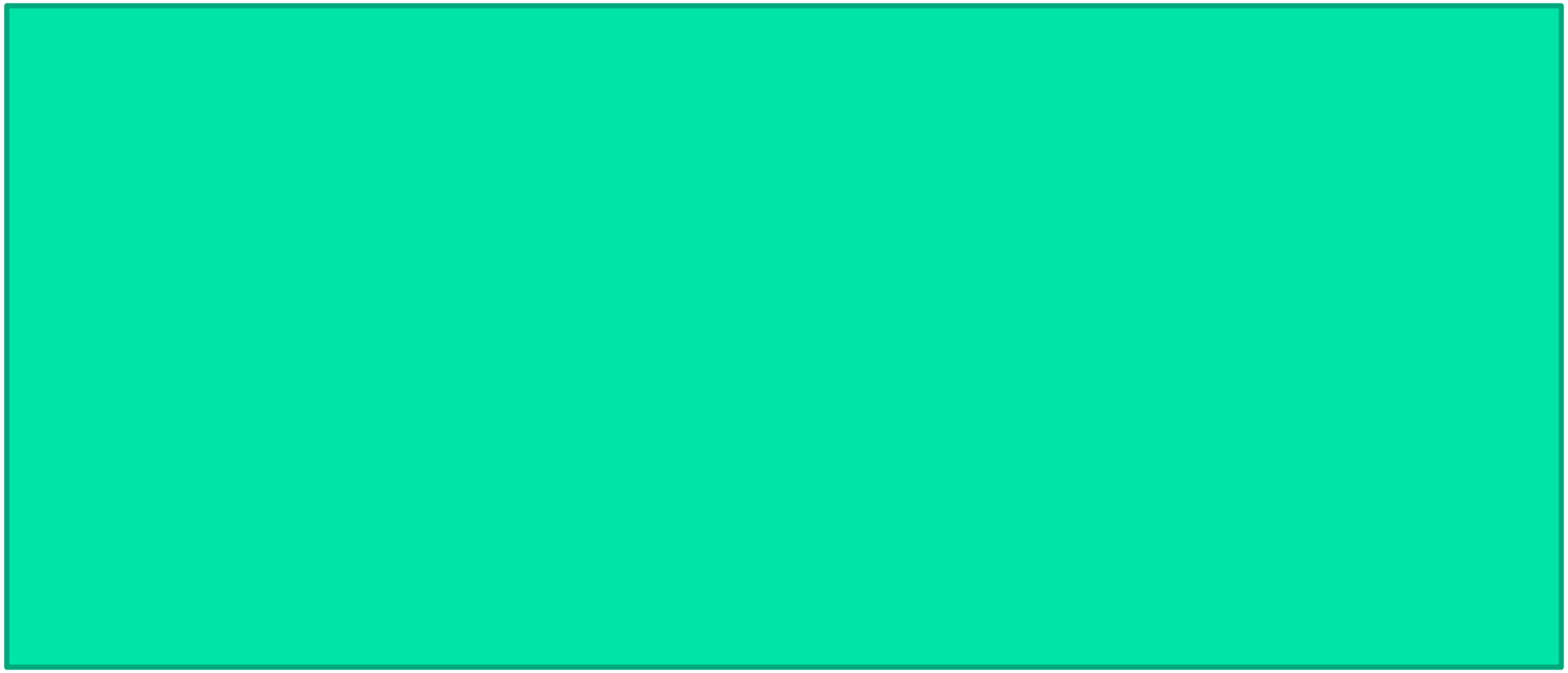
A strictly positive distribution cannot represent the behavior of Inverter X as it will have to assign the probability zero to the event $A = \text{true}$, $C = \text{true}$.

A strictly positive distribution cannot capture logical constraints.

Intersection

Holds only for strictly positive distributions

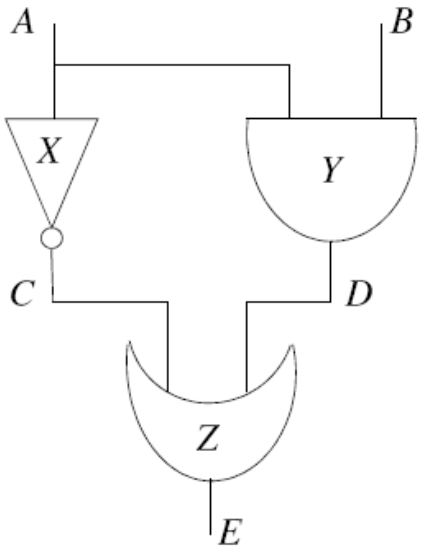
$I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y})$ and $I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$ only if $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$
If information **w** is irrelevant given **y**, and **y** is irrelevant given **w**, then combined information **yw** is irrelevant to start with.



Intersection

Holds only for strictly positive distributions

$I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y})$ and $I_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$ only if $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$
If information **w** is irrelevant given **y**, and **y** is irrelevant given **w**, then combined information **yw** is irrelevant to start with.



- If we know the input A of inverter X, its output C becomes irrelevant to our belief in the circuit output E.
- If we know the output C of inverter X, its input A becomes irrelevant to this belief.
- Yet, variables A and C are not irrelevant to our belief in the circuit output E.

THE INTERSECTION AXIOM AND STRICTLY POSITIVE DISTRIBUTIONS

- The intersection axiom requires $P(x) > 0$ for all x
- Will not hold if the variables in U are constrained by logical dependencies.
- Y: "The water temperature is above freezing"
- W: "The water temperature is above 32°F"
- Knowing the truth of either proposition renders the other superfluous.
- However, Y and W might still be relevant to a third proposition X = "We will enjoy swimming in that water," for example.
- If two properties exert influence on X, then it is impossible that each of the two properties will render the other irrelevant.
- Such symmetrical exclusion is possible only with analytical or definitional properties (e.g., Y = "The water temperature is above 32°F," W = "The water temperature is not equal to or lower than 32°F") and not with properties defined by independent empirical tests.

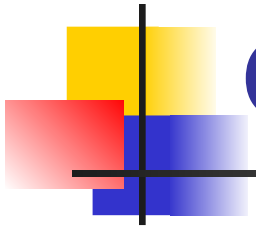
Graphs vs Graphoids

- Symmetry:
  - **I(X,Z,Y) ⇔ I(Y,Z,X)**
- Decomposition:
  - **I(X,Z,YW)⇒ I(X,Z,Y) and I(X,Z,W)**
- Weak union:
  - **I(X,Z,YW)⇒I(X,ZW,Y)**
- Contraction:
  - **I(X,Z,Y) and I(X,ZY,W)⇒I(X,Z,YW)**
- Intersection:
  - **I(X,ZY,W) and I(X,ZW,Y) ⇒ I(X,Z,YW)**

- **Graphoid**: satisfy all 5 axioms
- **Semi-graphoid**: satisfies the first 4.

- Decomposition is only one way while in graphs it is iff.
- Weak union states that w should be chosen from a set that, like **Y** should already be separated from **X** by **Z**

Why Axiomatic Characterization?



- Allow deriving conjectures about independencies that are clearer
- Axioms serve as inference rules
- Can capture the principal differences between various notions of relevance or independence

Dependency Models and Dependency Maps

- A dependency model is a set of independence statements I(X,Y,Z) that are either true or false.
- An undirected graph with node separation is a dependency model
- We say $< X, Z, Y >_G$ iff once you remove Z from the graph X and Y are not connected
- Can we completely capture probabilistic independencies by the notion of separation in a graph?
- Example: 2 coins and a bell.

Independency-map (i-map) and Dependency-maps (d-maps)

- A graph G is an independency map (i-map) of a probability distribution iff

  $< X, Z, Y >_G$ implies I_P(X,Z,Y)

- A graph G is a Dependency map (d-map) of a probability distribution P iff

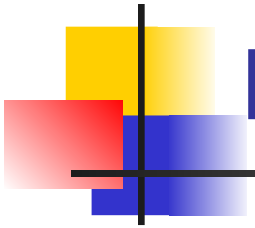
  $not\ < X, Z, Y >_G$ implies $not\ I_P$(X,Z,Y)

A model with induced dependencies cannot have a graph which is a perfect map
Example: two coins and a bell… try it
How we then represent two causes leading to a common consequence

Axiomatic Characterization of Graphs

- **Definition:** A model M is graph-isomorph if there exists a graph which is a perfect map of M.
- **Theorem (Pearl and Paz 1985)**: A necessary and sufficient condition for a dependency model to be graph–isomorph is that it satisfies
  - Symmetry: **I(X,Z,Y)** ⇔ **I(Y,Z,X**)
  - Decomposition: **I(X,Z,YW)**⇒ **I(X,Z,Y) and I(X,Z,Y**)
  - Intersection: **I(X,ZW,Y)** and **I(X,ZY,W)**⇒**I(X,Z,YW)**
  - Strong union: I(**X,Z,Y)** ⇒ I**(X,ZW, Y)**
  - Transitivity: I(**X,Z,Y**) ⇒ exists t s.t. I(**X,Z**,t) or I(t,**Z,Y**)

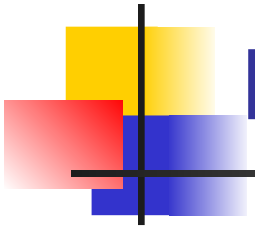
- This properties are satisfied by graph separation



Markov Networks

- **Graphs and probabilities:**
  - Given P, can we construct a graph I-map with minimal edges?
  - Given (G,P) can we test if G is an I-map? a perfect map?

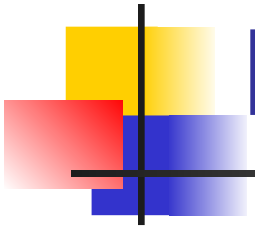
- **Markov Network Definition:** A graph G which is a minimal I-map of a probability distribution P, namely deleting any edge destroys its i-mappness, is called a **Markov network of P**.



Markov Networks

- **Theorem (Pearl and Paz 1985):** A dependency model satisfying symmetry decomposition and intersection has a unique minimal graph as an i-map, produced by deleting every edge (a,b) for which I(a**,U**-a-b,b) is true.
- The theorem defines an edge-deletion method for constructing G_0
- **Markov blanket** of a is a set S for which I(a,S,U-S-a).
- **Markov Boundary**: a minimal Markov blanket.

- **Theorem (Pearl and Paz 1985):** if symmetry, decomposition, weak union and intersection are satisfied by P, the Markov boundary is unique and it is the neighborhood in the Markov network of P



Markov Networks

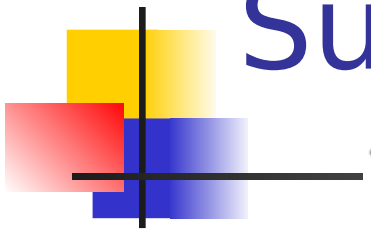
- **Corollary:** the Markov network G of any strictly positive distribution P can be obtained by connecting every node to its Markov boundary.
- The following 2 interpretations of direct neighbors are identical:
  - Neighbors as blanket that shields a variable from the influence of all others
  - Neighborhood as a tight influence between variables that cannot be weakened by other elements in the system
- So, given P (positive) how can we construct G?
- Given (G,P) how do we test the G is an I-map of P?
- Given G, can we construct P which is a perfect i-map? (Geiger and Pearl 1988)

Testing I-mapness

- Theorem 5 (Pearl): Given a positive P and a graph G the following are equivalent:
  - G is an I-map of P iff G is a super-graph of the Markov network of P
  - G is locally Markov w.r.t. P (the neighbors of a in G is a Markov blanket.) iff G is a super-graph of the Markov network of P

- There appear to be **no** test for I-mappness of undirected graph that works for extreme distributions without testing every cutset in G (ex: x=y=z=t)
- Representations of probabilistic independence using undirected graphs rest heavily on the intersection and weak union axioms.
- In contrast, we will see that directed graph representations rely on the contraction and weak union axiom, with intersection playing a minor role.

Markov Networks: Summary



- The essential qualities of conditional independence are captured by five logical axioms: (a) symmetry, (b) decomposition, (c) weak union, (d) contraction and (e) intersection.

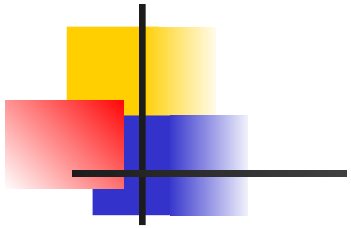
- Intersection holds only for strictly positive distributions (i.e., reflecting no functional or definitional constraints) and is essential to the construction of undirected graphs.

- Symmetry, decomposition, and intersection enable us to construct a minimal graph G_0 (Markov network), in which every cutset corresponds to a genuine independence condition.

- The weak union axiom is needed to guarantee that the set of neighbors that G_0 assigns to each variable α is the smallest set required to shield α from the effects of all other variables.

- If we identify the Markov boundaries associated with each proposition in the system and treat them as neighborhood relations defining a graph G_0, then we can correctly identify independence relationships by testing whether the set of known propositions constitutes a cutset in G_0.

- Not all probabilistic dependencies can be captured by undirected graphs because graph separation is strictly normal and transitive.



CONCEPTUAL DEPENDENCIES AND THEIR MARKOV NETWORKS

- An agent identifies the following variables as having influence on the main question of being late to a meeting:

  1. The time shown on the watch of Passerby 1.
  2. The time shown on the watch of Passerby 2.
  3. The correct time.
  4. The time it takes to travel to the meeting place.
  5. The arrival time at the meeting place.

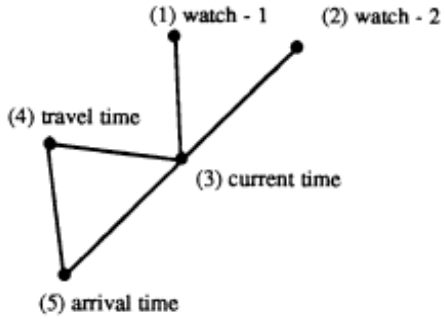
- The construction of G_0 can proceed by one of two methods:
  - The *edge-deletion* method.
  - The *Markov boundary* method.

- The first method requires that for every pair of variables (α, β) we determine whether fixing the values of all other variables in the system will render our belief in α sensitive to β.

- For example, the reading on Passerby 1's watch (1) will vary with the actual time (3) even if all other variables are known, so connect node 1 to node 3

- The Markov boundary method requires that for every variable α in the system, we identify a minimal set of variables sufficient to render the belief in α insensitive to all other variables in the system.

- For instance, once we know the current time (3), no other variable can affect what we expect to read on passerby 1's watch (1).



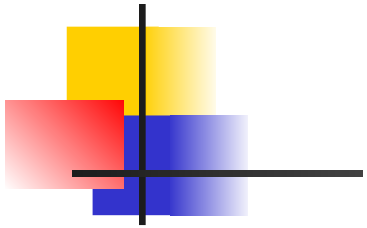
The unusual edge (3,4) reflects the reasoning that if we fix the arrival time (5) the travel time (4) must depends on current time (3)

Figure 3.6. The Markov network representing the prediction of *A* 's arrival time.

- G_0 can be used as an inference instrument.

  - For example, knowing the current time (3) renders the time on Passerby 1's watch (1) irrelevant for estimating the travel time (4) (i.e., $I(1,3,4)$); 3 is a cutset in G_0, separating 1 from 4.

Summary



- The essential qualities of conditional independence are captured by five logical axioms: (a) symmetry, (b) decomposition, (c) weak union, (d) contraction and (e) intersection.
- Intersection holds only for strictly positive distributions (i.e., reflecting no functional or definitional constraints) and is essential to the construction of undirected graphs.
- Symmetry, decomposition, and intersection enable us to construct a minimal graph G_0 (Markov network), in which every cutset corresponds to a genuine independence condition.
- The weak union axiom is needed to guarantee that the set of neighbors that G_0 assigns to each variable α is the smallest set required to shield α from the effects of all other variables.
- If we identify the Markov boundaries associated with each proposition in the system and treat them as neighborhood relations defining a graph G_0, then we can correctly identify independence relationships by testing whether the set of known propositions constitutes a cutset in G_0.
- Not all probabilistic dependencies can be captured by undirected graphs because graph separation is strictly normal and transitive.

MARKOV NETWORK AS A KNOWLEDGE BASE

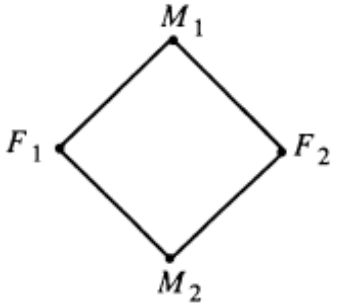
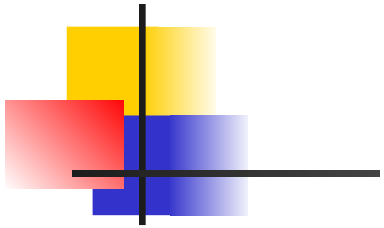


Figure 3.2. An undirected graph representing interactions among four individuals.

How can we construct a probability Distribution that will have all these independencies?

QUANTIFYING THE LINKS

- If couple (M_1, F_2) meet less frequently than the couple (M_1, F_1), then the first link should be weaker than the second
- The model must be consistent, complete and a Markov field of G.
- Arbitrary specification of $P(M_1, F_1)$, $P(F_1, M_2)$, $P(M_2, F_2)$, and $P(F_2, M_1)$ might lead to inconsistencies.
- If we specify the pairwise probabilities of only three pairs, incompleteness will result.



- A safe method (called *Gibbs' potential*) for constructing a complete and consistent quantitative model while preserving the dependency structure of an arbitrary graph G.

1. Identify the cliques† of G, namely, the largest subgraphs whose nodes are all adjacent to each other.
2. For each clique C_i, assign a nonnegative compatibility function $g_i(c_i)$, which measures the relative degree of compatibility associated with the value assignment c_i to the variables included in C_i.
3. Form the product $\prod_i g_i(c_i)$ of the compatibility functions over all the cliques.
4. Normalize the product over all possible value combinations of the variables in the system

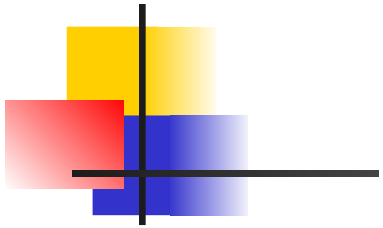
$$P(x_1,\ldots,x_n) = K \prod_i g_i(c_i), \tag{3.13}$$

where

$$K = [\sum_{x_1,\ldots,x_n} \prod_i g_i(c_i)]^{-1}.$$

† We use the term *clique* for the more common term *maximal clique*.

So, How do we learn Markov networks From data?



G is locally markov
If neighbors make every
Variable independent
From the rest.

THEOREM 6 [Hammersley and Clifford 1971]: A probability function P formed by a normalized product of nonnegative functions on the cliques of G is a Markov field relative to G, i.e., G is an I-map of P.

Proof: G is guaranteed to be an I-map if P is locally Markov relative to G (Theorem 5). It is sufficient, therefore, to show that the neighbors in G of each variable α constitute a Markov blanket of α relative to P, i.e., that $I(\alpha, \boldsymbol{B}_G(\alpha), \boldsymbol{U} - \alpha - \boldsymbol{B}_G(\alpha))$ or (using Eq. (3.5c)) that

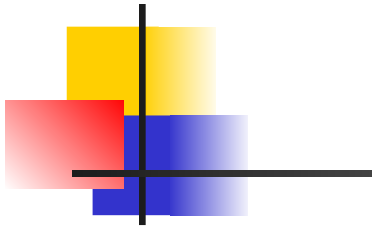
$$P(\alpha, \boldsymbol{B}_G(\alpha), \boldsymbol{U} - \alpha - \boldsymbol{B}_G(\alpha)) = f_1(\alpha, \boldsymbol{B}_G(\alpha))\, f_2(U - \alpha). \tag{3.17}$$

- Let J_α stand for the set of indices marking all cliques in G that include α, $J_\alpha = \{j : \alpha \in \boldsymbol{C}_j\}$. Since P is in product form, we can write

$$P(\alpha, \beta, \ldots) = K \prod_j g_j(\boldsymbol{c}_j) = K \prod_{j \in J_\alpha} g_j(\boldsymbol{c}_j) \prod_{j \notin J_\alpha} g_j(\boldsymbol{c}_j). \tag{3.18}$$

- The first product in Eq. (3.18) contains only variables that are adjacent to α in G; otherwise, $\boldsymbol{C}_j$ would not be a clique. According to the definition of J_α, the second product does not involve α. Thus, Eq. (3.17) is established. Q.E.D.

INTERPRETING THE LINK PARAMETERS



- It is difficult to assign meanings to the parameters of the compatibility functions.
- Given the joint probability $P(M_1, F_1, F_2, M_2)$, to infer the compatibility functions g_i we must solve a set of simultaneous nonlinear equations for g_i
- The solution for g_i will not be applicable to new situations.
- For a parameter to be meaningful, it must be an abstraction of some invariant property of one's experience.
- The quantities $P(f_1 | m_1, \neg m_2)$ and $P(f_1 | \neg m_1, \neg m_2)$ and their relations to the frequency of interaction of couple $\{M_1, F_1\}$ are perceived as invariant characteristics of the disease.
- The Markov network formulation does not allow the direct specification of such judgmental input.
- Judgments about low-order conditional probabilities (e.g., $P(m_1 | f_1, \neg m_2)$) can be taken only as constraints that the joint probability distribution (Eq. (3.13)) must satisfy; from them, we might be able to calculate the actual values of the compatibility parameters.