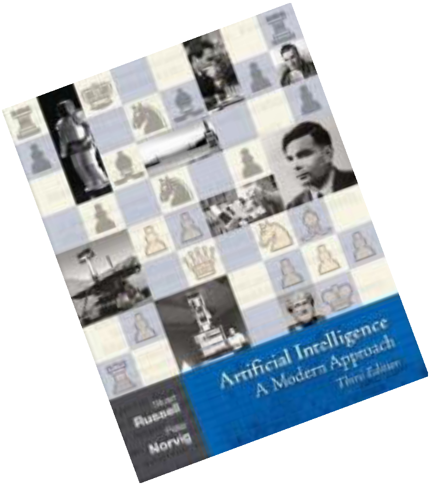


Introduction to Artificial Intelligence

כללי

מפגש שישי

- פרק 1 - מבוא
- פרק 2 - סוכנים אינטיליגנטיים
- פרק 3 - פתרון בעיות באמצעות חיפוש
- פרק 4 - אלגוריתמים לחיפוש מקומי ובעיות אופטימיזציה (רק סעיף 4.1)
- פרק 5 - חיפוש בתנאי יריבות (סעיפים 5.1-5.4)
- פרק 6 - בעיות סיפוק אילוצים (סעיפים 6.1-6.4)
- פרק 7 - סוכנים לוגיים
- פרק 8 - לוגיקה מסדר ראשון (סעיפים 8.1-8.3 בלי 8.3.3)
- פרק 9 - היסק בלוגיקה מסדר ראשון (בלי 9.3.3, בסעיף 9.4 רק את 9.4.1)
- פרק 10- תכנון (10.1-10.3, 10.4.1, 10.4.4)
- פרק 13 - אי-וודאות
- פרק 14 - הנמקה הסתברותית (14.1-14.2)
- פרק 17 - קבלת החלטות מורכבות (בלי 17.4)
- פרק 18- למידה מדוגמאות (18.1-18.5)

אי-וודאות

Outline

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

אי ודאות (Uncertainty)

- אי ודאות (חוסר ודאות-Uncertainty) היא מצב שבו אינה קיימת האינפורמציה הנחוצה בכדי לתאר באופן מדויק מצב קיים או מצב עתידי. כאשר קיימת חוסר ודאות, מספר המצבים האפשריים בהווה או בעתיד (מצבי הטבע) גדול מ-1.
  - לדוגמה – האם ב-1 בינואר בשנה הקלנדארית הבאה ירד גשם? שני מצבי טבע אפשריים – 'גשם' ו-'אין גשם' ואין ביכולתנו לתאר באופן מדויק מה צפוי להיות בתאריך זה.

UNCERTAINTY

Let action A_t = leave for airport $_t$ minutes before flight
Will A_t get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2.
2. noisy sensors (traffic reports)
3.
3. uncertainty in action outcomes (flat tire, etc.)
4.
4. immense complexity of modeling and predicting traffic
5.

Hence a purely logical approach either

1. risks falsehood: "A_{25} will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

אי ודאות

- תורת ההסתברות היא הכלי המתמטי העיקרי בו משתמשים כיום סוכנים הפועלים בתנאי אי וודאות. הסיבה לשימוש בהסתברות היא שהתורה מבוססת יחסית, תורה שנחקרה היטב וכבר קיימות במסגרתה שיטות מוכנות לייצוג ולטיפול במצבי אי וודאות.
-
- למרות זאת, קיימים כלים אלטרנטיביים לטיפול באי וודאות, אחד מהם הוא לוגיקה עמומה. הרעיון בכלי זה הוא לשייך מידה של "אמיתות" לכל עובדה בעולם. לכל עובדה תשוייך הסתברות המציינת את השכיחות היחסית של התקיימות העובדה. ניתן להתייחס להסתברות זו גם כמידת האמונה (degree of belief) שלנו שהעובדה מתקיימת.

דוגמה: סוכן נהג מונית

- תהי הפעולה A_t: צא לדרך לכיוון שדה התעופה t דקות לפני זמן ההמראה.
- האם A_t תביאני לשדה בזמן?
-
- בעיות אפשריות:

1.נראות חלקית (מצב הכביש וכו')

2.המידע שהסוכן מקבל מהסביבה באמצעות חיישניו הוא "רועש" כלומר אינו מדוייק

3.אי וודאות לגבי תוצאת פעולה (תקר בגלגל וכו')

4.מורכבות ניבוי התנועה

דוגמה: סוכן נהג מונית

- כדי לקבל החלטות בתנאי אי-וודאות, צריך שיהיו לסוכן העדפות בין תוצאות שונות אפשריות של תכניות שונות.
- בהתייחס לדוגמת נהג מונית הצריך להביא את נוסעו בזמן לשדה התעופה, נניח שהסוכן מאמין כי:

- $P(A_{25}$ gets me there on time$|\ldots) = 0.04$
- $P(A_{90}$ gets me there on time$|\ldots) = 0.70$
- $P(A_{120}$ gets me there on time$|\ldots) = 0.95$
- $P(A_{1440}$ gets me there on time$|\ldots) = 0.9994$

איזו פעולה לבחור?

דוגמה: סוכן נהג מונית

- הבחירה תלויה בהעדפותיו, עד כמה הוא מעדיף לפספס את הטיסה לעומת האפשרות להגיע הרבה זמן לפני ולהמתין זמן רב בשדה התעופה וכד'.
- התועלת של מצב עבור (ביחס ל) סוכן היא מדד המשקף עד כמה מצב הוא "טוב" עבורו או מדד המשקף את "מידת השגת המטרה". הסוכן יעדיף מצבים עם תועלת גבוהה יותר.
- נשתמש בתורת התועלת (utility theory) כדי לייצג ולהסיק העדפות.

Methods for handling uncertainty

- Default or nonmonotonic logic:
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?
- Rules with fudge factors:
  - $A_{25} \mid\rightarrow_{0.3}$ get there on time
  - $Sprinkler \mid\rightarrow_{0.99} WetGrass$
  - $WetGrass \mid\rightarrow_{0.7} Rain$
- Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??
- Probability
  - Model agent's degree of belief
  - Given the available evidence,
  - A_{25} will get me there on time with probability 0.04

סיכון (Risk)

- בכלכלה נהוג להגדיר סיכון (Risk) באופן שונה משפת היומיום ומתחומים מדעיים אחרים.
- ביומיום אנו משתמשים במונח סיכון לתאר מצב שבו חוסר הודאות עלולה להסתיים בתוצאה שלילית יחסית למצב המוצא (הפסד או אובדן).
- אנו מתייחסים לסיכון כמצב של חוסר ודאות בו אנו מסוגלים לייחס הסתברות (מידת סבירות) למצבי הטבע האפשריים השונים.
  - עירית עובדת בחברת נסיעות והיא יודעת שמעבידתה תעניק לה ביום הולדתה אחד מהשניים: (1) נופש משפחתי באיי סיישל; או (2) טיול משפחתי בניו-זילנד. (הניחו כי שתי האלטרנטיביות הן משאת נפשה של עירית).

  על-פי ההגדרה הכלכלית, אם עירית איננה יודעת את ההסתברות לכל תרחיש, נאמר כי עירית נמצאת בחוסר ודאות.

  ואם היא מייחסת ההסתברויות לכל תרחיש (למשל 0.5-0.5) אזי נאמר שעירית נמצאת ב**סיכון**.

  מה עירית תספר למשפחתה?

סיכון – דוגמה נוספת

- הצעה A איננה מסוכנת – היא ודאית.
- הצעה B מסוכנת למרות שבכל תרחיש אפשרי נזכה בפרס חיובי.

הצעה	הסתברות	פרס
A	1	- 100
B	0.25	100
	0.75	101

מערכת לאיבחון תקלות רכב

- **משתנים** (. לכל גורם רלוונטי בעולם הבעיה נתאים משתנה מקרי):
  - Radio – הרדיו פועל?
  - Battery – המצבר תקין?
  - Ignition – הצתה?
  - Starts – המנוע מניע?
  - Gas – יש דלק?
  - Moves – הרכב זז?

מה ההסתברות שהמנוע יפעל
בהינתן שהרדיו פועל

מערכת לאיבחון תקלות רכב

- ברור כי זה מודל פשטני של עולם הבעיה שלנו, ברכב אמיתי יש עוד הרבה גורמים המשפיעים על התהליכים המענינים אותנו. העובדה שאנו עובדים עם מודל הסתברותי מאפשרת לנו את הפשטנות הזאת – כל הגורמים מהם התעלמנו ישפיעו על ההסתברויות של המשתנים המקריים.
- אנו רוצים להיות מסוגלים לשאול שאלות כמו : "מה ההסתברות שהמצבר לא תקין בהנתן שהרכב לא זז ?", "מה ההסתברות שהמנוע יפעל בהינתן שהרדיו פועל" וכיו"ב.
- אם סוכן יודע לשערך נכון הסתברויות מותנות, הוא יכול לבחור את הפעולה הבאה שלו כדי שתביא למקסום תוחלת התועלת.
- בדוגמה שלנו, בהנחה שהרכב לא זז, הסוכן צריך להחליט אם ללכת לתחנת דלק או לנסות לעצור מכונית לעזרה בהנעה.
- הסוכן יחשב לכל פעולה את תוחלת התועלת ויבחר בפעולה הנותנת את תוחלת התועלת הגדולה ביותר.

- הרעיון הבסיסי של תורת ההחלטות (decision theory) הוא שסוכן הוא רציונלי אם"ם הוא בוחר את הפעולה המניבה את התועלת הגבוהה ביותר, כאשר התועלת מחושבת כאמור כממוצע התועלות של כל התוצאות האפשריות של הפעולה.
- זהו עקרון **מקסימום תוחלת התועלת** – **MEU**

הרעיון - MEU

- קבלת החלטות על סמך הסתברות.
- מה צריך לפני שמתחילים ליישם ?
- מי יכול להכריע?
- איפה אנחנו עומדים לפני הפיתוח אילו עבודות הכנה צריכים לעשות?
- הקשר בין זה לבין יורסטיקה?

MEU

הרעיון של קבלת החלטות על סמך הסתברויות ותועלת הוא העקרון המרכזי בתורת ההחלטות. אנחנו נתרכז בשיערוך ההסתברויות המותנות.

(שילוב בין ההסברות לבין התועלת)

Decision Theory = Probability Theory + Utility Theory

Syntax

- רכיב בסיסי: משתנה אקראי
- דומה להיגיון פסוקים: עולמות אפשריים מוגדרים על ידי הקצאה של ערכים למשתנים אקראיים.
- משתנים אקראיים בוליאנית
- למשל: *Cavity* (do I have a cavity?)
-
- משתנים אקראיים בדידים
- למשל:, *Weather* is one of <*sunny,rainy,cloudy,snow*>
- התחום חייב להיות מוגדר
- פיתרון צריך להיות הקצאה של ערך
-
- random variable: e.g., $Weather = sunny$, $Cavity = false$
- (abbreviated as $\neg cavity$)
- טענות מרוכבות מורכבות מהנחות יסוד לוגיות למשל
- $Weather = sunny \vee Cavity = false$
-

Syntax

- אירוע אטומי: תיאור מלא של מצבו של העולם על אף שהסוכן אינו ודאי.

לדוגמה, אם העולם מורכב רק משני משתנים בוליאנים $Cavity$ ו-$Toothache$, אז יש 4 אירועים נפרדים אטומיים:

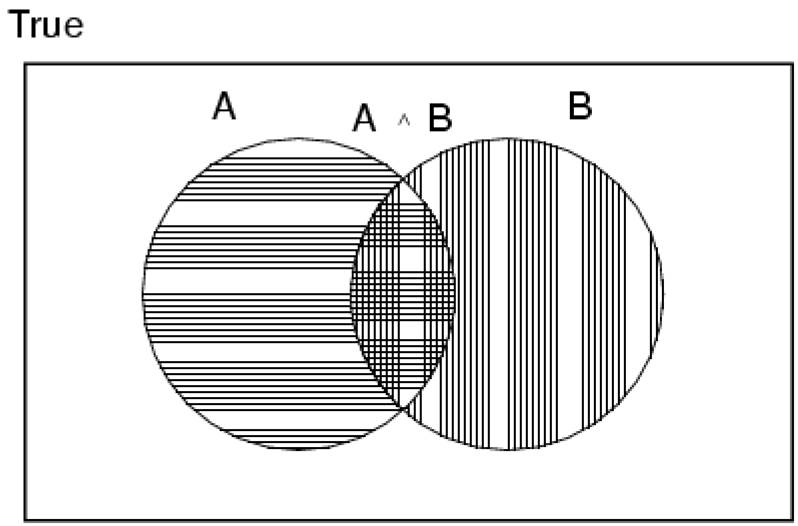
$$
\begin{aligned}
&Cavity = false \wedge Toothache = false\\
&Cavity = false \wedge Toothache = true\\
&Cavity = true \wedge Toothache = false\\
&Cavity = true \wedge Toothache = true
\end{aligned}
$$

- אירועים אטומים הם יחודיים וממצים!!!
-

אקסיומות של הסתברות

- עבור כל שני משתנים A, B מתקיים

- $0 \le P(A) \le 1$, $0 \le P(B) \le 1$
- $P(true) = 1$ and $P(false) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
-



הסתברויות קודמות

- הסתברויות קודמות או ללא תנאי של טענות
- למשל
- P(Cavity = true) = 0.1 and P(Weather = sunny) = 0.72
- מתאים לאמת (לא סותר שום טענה) לפני ההגעה של כל ראיה (חדשה)

- התפלגות הסתברות נותנת ערכים עבור כל המשימות אפשריות:

- P(Weather) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)

- התפלגות הסתברות משותפת עבור קבוצה של משתנים אקראיים נותנת את ההסתברות של כל אירוע האטום על משתנים אקראיים אלה
-

$\mathbf{P}(Weather, Cavity)$ = a 4 × 2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

- בכל שאלה לגבי תחום ניתן לענות על ידי צירוף הסברויות

הסתברות מותנית

- הסתברויות מותנות או מוקדמות
- למשל $P(cavity \mid toothache) = 0.8$
-

משמע בהנתן שכל מה שאני יודע הוא יש לי כאב שיניים אז הסיכוי ל .... הוא

- ובצורה אחרת באופן מותנה
$\mathbf{P}(Cavity \mid Toothache)$ = 2-element vector of 2-element vectors)

אם אנחנו יודעים ששתי הפעילויות נכונות
$P(cavity \mid toothache, cavity) = 1$

- ראיות חדשות עשויות להיות לא רלוונטיות, אבל מאפשרות פישוט,
- $P(cavity \mid toothache, sunny) = P(cavity \mid toothache) = 0.8$
- סוג זה של היסק, המאושרר על ידי ידע מושלם, הוא חיוני
-

הסתברות מותנית

- הגדרה של הסתברות מותנית:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
-
- Product rule מוביל לנוסחה חלופית
-

$P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$

- לדוגמא

$\mathbf{P}(\mathit{Weather},\mathit{Cavity}) = \mathbf{P}(\mathit{Weather} \mid \mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})$

- (View as a set of 4 × 2 equations, not matrix mult.)
-
- Chain rule נגזר על ידי יישום רצוף :
-

$$\begin{aligned}\mathbf{P}(X_1, \ldots,X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \mathbf{P}(X_1,\ldots,X_{n-2})\ \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \ldots \\ &= \pi_{i= 1}\text{^}n\ \mathbf{P}(X_i \mid X_1, \ldots ,X_{i-1})\end{aligned}$$

הסקה על ידי מנייה

- נתחיל עם ההסתברות המשותפת:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- לכל רכיב φ, סכום האפשרויות בהם הוא true הוא

$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

הסקה על ידי מנייה

- נתחיל עם ההסתברות המשותפת:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- לכל רכיב φ, סכום האפשרויות בהם הוא true

$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

- $P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$
-

הסקה על ידי מנייה

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- לכל רכיב φ, סכום האפשרויות בהם הוא true הוא $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

- $P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$
-

הסקה על ידי מנייה

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- ניתן גם לחשב הסתברויות מותנות:
-

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

נירמול

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- מכנה יכול לראות כ מנורמל
-

$$
\begin{aligned}
\mathbf{P}(Cavity \mid toothache) &= \alpha,\ \mathbf{P}(Cavity, toothache) \\
&= \alpha,\ [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha,\ [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&= \alpha,\ \langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}
$$

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

הסקה על ידי מנייה המשך

Typically, we are interested in
the posterior joint distribution of the query variables **Y**
given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

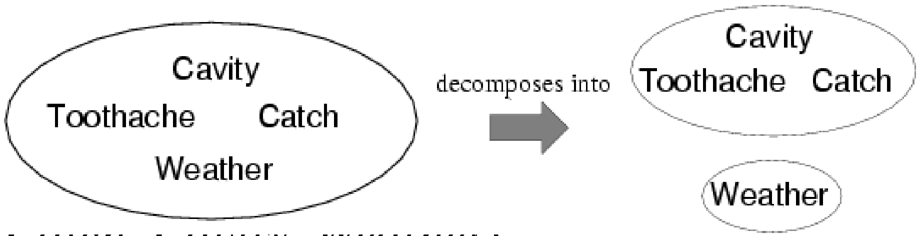
Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_h\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
-
- Obvious problems:
-
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2.
  2. Space complexity $O(d^n)$ to store the joint distribution
  3.
  3. How to find the numbers for $O(d^n)$ entries?

INDEPENDENCE

- A and B are independent iff
  $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ or $\mathbf{P}(B \mid A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$



$\mathbf{P}(Toothache, Catch, Cavity, Weather)$
$= \mathbf{P}(Toothache, Catch, Cavity)\,\mathbf{P}(Weather)$

- 32 entries reduced to 12; for n independent biased coins, $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- $\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries
-
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
- (1) $\mathbf{P}(catch \mid toothache, cavity) = \mathbf{P}(catch \mid cavity)$
- The same independence holds if I haven't got a cavity:
- (2) $\mathbf{P}(catch \mid toothache, \neg cavity) = \mathbf{P}(catch \mid \neg cavity)$
- *Catch* is conditionally independent of *Toothache* given *Cavity*:
- $\mathbf{P}(Catch \mid Toothache, Cavity) = \mathbf{P}(Catch \mid Cavity)$
- Equivalent statements:
  $\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$

  $\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\, \mathbf{P}(Catch \mid Cavity)$

Conditional independence contd.

- Write out full joint distribution using chain rule:
- $\mathbf{P}(Toothache, Catch, Cavity)$
  $= \mathbf{P}(Toothache \mid Catch, Cavity)\,\mathbf{P}(Catch, Cavity)$
  $= \mathbf{P}(Toothache \mid Catch, Cavity)\,\mathbf{P}(Catch \mid Cavity)\,\mathbf{P}(Cavity)$
  $= \mathbf{P}(Toothache \mid Cavity)\,\mathbf{P}(Catch \mid Cavity)\,\mathbf{P}(\text{Cavity})$

  I.e., 2 + 2 + 1 = 5 independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*.
-
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.
-

Bayes' Rule

- Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
- $\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) / P(b)$

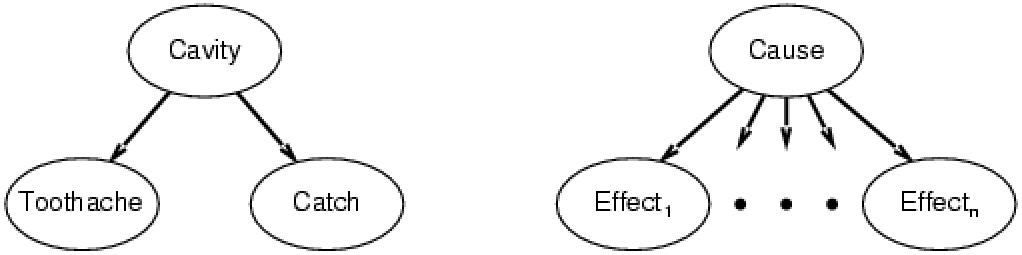
- or in distribution form
- $$\mathbf{P}(Y \mid X) = \mathbf{P}(X \mid Y)\, \mathbf{P}(Y) / \mathbf{P}(X) = \alpha \mathbf{P}(X \mid Y)\, \mathbf{P}(Y)$$
- Useful for assessing diagnostic probability from causal probability:
  - $P(Cause \mid Effect) = P(Effect \mid Cause)\, P(Cause) / P(Effect)$
  - E.g., let *M* be meningitis, *S* be stiff neck:
  - $P(m \mid s) = P(s \mid m)\, P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$
  - Note: posterior probability of meningitis still very small!

Bayes' Rule and conditional independence

$\mathbf{P}(Cavity \mid toothache \wedge catch)$
$= \alpha\mathbf{P}(toothache \wedge catch \mid Cavity)\ \mathbf{P}(Cavity)$
$= \alpha\mathbf{P}(toothache \mid Cavity)\ \mathbf{P}(catch \mid Cavity)\ \mathbf{P}(Cavity)$

- This is an example of a naïve Bayes model:
-

$$\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause})\ \pi_i \mathbf{P}(\text{Effect}_i \mid \text{Cause})$$



- Total number of parameters is linear in n
-

Summary

- Probability is a rigorous formalism for uncertain knowledge
-
- Joint probability distribution specifies probability of every atomic event
- Queries can be answered by summing over atomic events
-
- For nontrivial domains, we must find a way to reduce the joint size
-
- Independence and conditional independence provide the tools
-

Probabilistic Inference

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

$$P(Cavity \vee Toothache) = 0.108 + 0.012 + \ldots$$
$$= 0.28$$

Probabilistic Inference

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

$$P(\text{Cavity}) = 0.108 + 0.012 + 0.072 + 0.008$$
$$= 0.2$$

Probabilistic Inference

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

Marginalization: $P(c) = \Sigma_t \Sigma_{pc} P(c \wedge t \wedge pc)$
using the conventions that c = Cavity or ¬Cavity and that Σ_t is the sum over t = {Toothache, ¬Toothache}

Conditional Probability

- $P(A \wedge B) = P(A|B)\,P(B)$
  $= P(B|A)\,P(A)$

  $P(A|B)$ is the posterior probability of A given B

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

$$P(\text{Cavity}|\text{Toothache}) = P(\text{Cavity} \wedge \text{Toothache})/P(\text{Toothache})$$
$$= (0.108+0.012)/(0.108+0.012+0.016+0.064) = 0.6$$

Interpretation: After observing Toothache, the patient is no longer an "average" one, and the prior probabilities of Cavity is no longer valid

P(Cavity|Toothache) is calculated by keeping the ratios of the probabilities of the 4 cases unchanged, and normalizing their sum to 1

	Toothache		$\neg$Toothache	
	PCatch	$\neg$PCatch	PCatch	$\neg$PCatch
Cavity	0.108	0.012	0.072	0.008
$\neg$Cavity	0.016	0.064	0.144	0.576

P(Cavity|Toothache) = P(Cavity$\wedge$Toothache)/P(Toothache)

= (0.108+0.012)/(0.108+0.012+0.016+0.064) = 0.6

P($\neg$Cavity|Toothache)=P($\neg$Cavity$\wedge$Toothache)/P(Toothache)

= (0.016+0.064)/(0.108+0.012+0.016+0.064) = 0.4

$$
\begin{aligned}
P(c|\text{Toochache}) &= \alpha\, P(c \wedge \text{Toothache}) \\
&= \alpha\, \Sigma_{pc}\, P(c \wedge \text{Toothache} \wedge pc) \\
&= \alpha\, [(0.108, 0.016) + (0.012, 0.064)] \\
&= \alpha\, (0.12, 0.08) = (0.6, 0.4)
\end{aligned}
$$

normalization constant

Conditional Probability

- $P(A \wedge B) = P(A|B)\, P(B)$
  $= P(B|A)\, P(A)$
- $P(A \wedge B \wedge C) = P(A|B,C)\, P(B \wedge C)$
  $= P(A|B,C)\, P(B|C)\, P(C)$
- $P(Cavity) = \sum_t \sum_{pc} P(Cavity \wedge t \wedge pc)$
  $= \sum_t \sum_{pc} P(Cavity|t,pc)\, P(t \wedge pc)$
- $P(c) = \sum_t \sum_{pc} P(c \wedge t \wedge pc)$
  $= \sum_t \sum_{pc} P(c|t,pc) P(t \wedge pc)$

INDEPENDENCE

- Two random variables A and B are independent if
$$P(A \wedge B) = P(A)\, P(B)$$
hence if $P(A|B) = P(A)$

- Two random variables A and B are independent given C, if
$$P(A \wedge B|C) = P(A|C)\, P(B|C)$$
hence if $P(A|B,C) = P(A|C)$

Updating the Belief State

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

- Let D now observe Toothache with probability 0.8 (e.g., “the patient says so”)
- How should D update its belief state?

Updating the Belief State

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

- Let E be the evidence such that $P(Toothache|E) = 0.8$
- We want to compute $P(c \wedge t \wedge pc|E) = P(c \wedge pc|t,E)\, P(t|E)$
- Since E is not directly related to the cavity or the probe catch, we consider that c and pc are independent of E given t, hence: $P(c \wedge pc|t,E) = P(c \wedge pc|t)$

Updating the Belief State

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	~~0.108~~ 0.432	~~0.012~~ 0.048	~~0.072~~ 0.018	~~0.008~~ 0.002
¬Cavity	~~0.016~~ 0.064	~~0.064~~ 0.256	~~0.144~~ 0.036	~~0.576~~ 0.144

- Let E be the evidence such that P(Toothache|E) = 0.8
- t∧pc|E) = P(c∧pc|t,E) P(t|E)
- Since E is not directly related to the cavity or the probe catch, we e independent of E given t, he t)

To get these 4 probabilities we normalize their sum to 0.8

To get these 4 probabilities we normalize their sum to 0.2

Issues

- If a state is described by n propositions, then a belief state contains 2^n states (possibly, some have probability 0)
- → Modeling difficulty: many numbers must be entered in the first place
- → Computational issue: memory size and time

HISTORY

‘60s The first expert systems

1968 Attempts to use probabi
systems (Gorry & Barne

1973 Gave up - to heavy calcu

1976 MYCIN: Medical predicate logic expert
system with <u>certainty factors</u> (Shortliffe).

1976 PROSPECTOR: Predicts the likely
location of mineral deposits. Uses Bayes’
rule. (Duda

Certainty Factor (MYCIN):

A <u>real</u> value (-1,+1):

-1:	expression is known to be <u>false</u>.
0:	<u>no belief</u> either way.
+1:	expression is known to be <u>true</u>.

Example:

rule 34:	$a \wedge b \wedge c \rightarrow q$	(+0.7)
rule 35:	$d \wedge q \rightarrow r \vee s$	(-0.9)

Example (PROSPECTOR):

$$P(d \mid a) = P(d \mid a \wedge b) \cdot P(b \mid a) + P(d \mid a \wedge \neg b) \cdot P(\neg b \mid a)$$

Summary of the time up until mid ‘80s:

- “Pure logic <u>will</u> solve the AI problems!”
- “Probability theory is <u>intractable</u> to use and too <u>complicated</u> for complex models.”

פרק 14
הנמקה הסתברותית

BUT...

1986 Bayesian networks were revived and reintroduced to expert systems (Pearl).

1988 Breakthrough for efficient calculation algorithms (Lauritzen & Spiegelhalter) $\Rightarrow$ <u>tractable calculations</u> on BNs.

1995 In Windows95™ for printer-trouble shooting and Office assistance (“the paper clip”).

1999 BN is getting more and more used. Ex. Gene expression analysis, Business strategy etc.

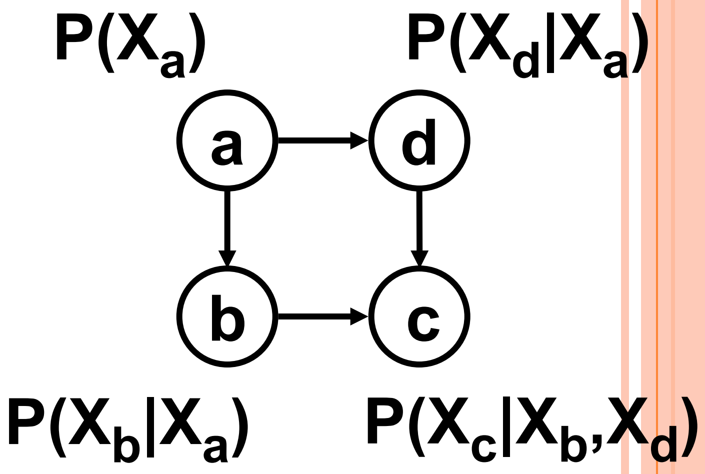
2000 Widely used - a BN tool will be shipped with every Windows™ Commercial Server.

Bayesian Networks is the future!

What is a Bayesian Network?

A Bayesian Network has two parts:

1) <u>qualitative part</u>
   - the structure
   - directed acyclic graph (DAG)
   - vertices represent variables
   - edges represent relations between variables

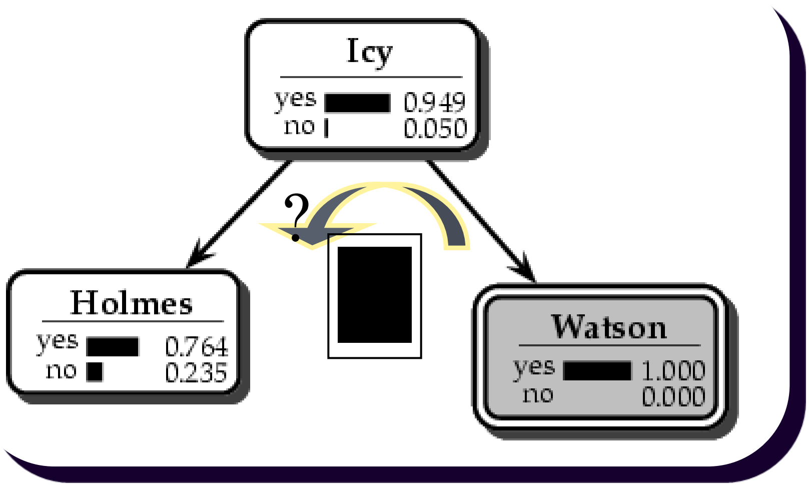


2) <u>quantitative part</u>
   - the strength of relationship between variables
   - conditional probability function

WHY BAYESIAN NETWORKS?

- Models probabilities
- Gives posterior beliefs given some observations
- Can be used as a classifier
- Can explain why
- Can find the variables with the most impact
- Algorithms exists
- Can model expert (subjective) knowledge
- Can automatically be learned from raw data
- Simple - my grandmother can use it!

"Watson has had an accident!" $\Rightarrow P(X_{Watson}=yes)=1$



Bayes' Rule $\Rightarrow$

$\Rightarrow P(X_{Icy} \mid X_{Watson}=yes) = (0.95,0.05)$

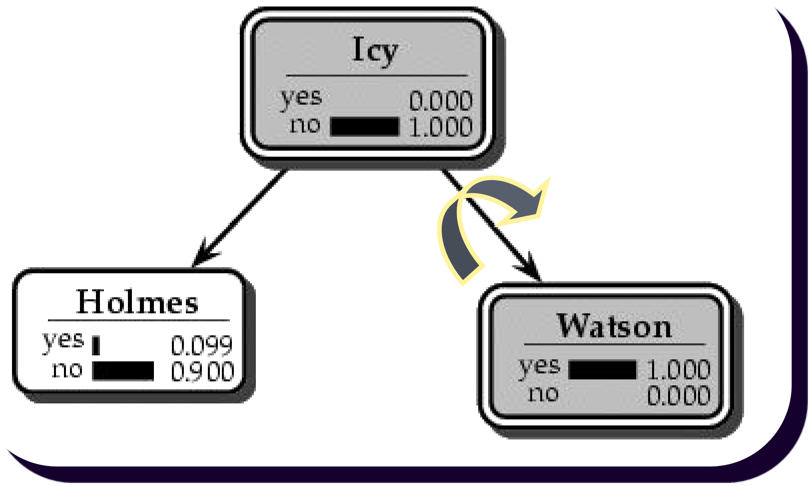
$(0.70,0.30)_{\text{a priori}}$

Joint Probability + Marginalization $\Rightarrow$

$\Rightarrow P(X_{Holmes} \mid X_{Watson}=yes) = (0.76,0.24)$

$(0.59,0.41)_{\text{a priori}}$

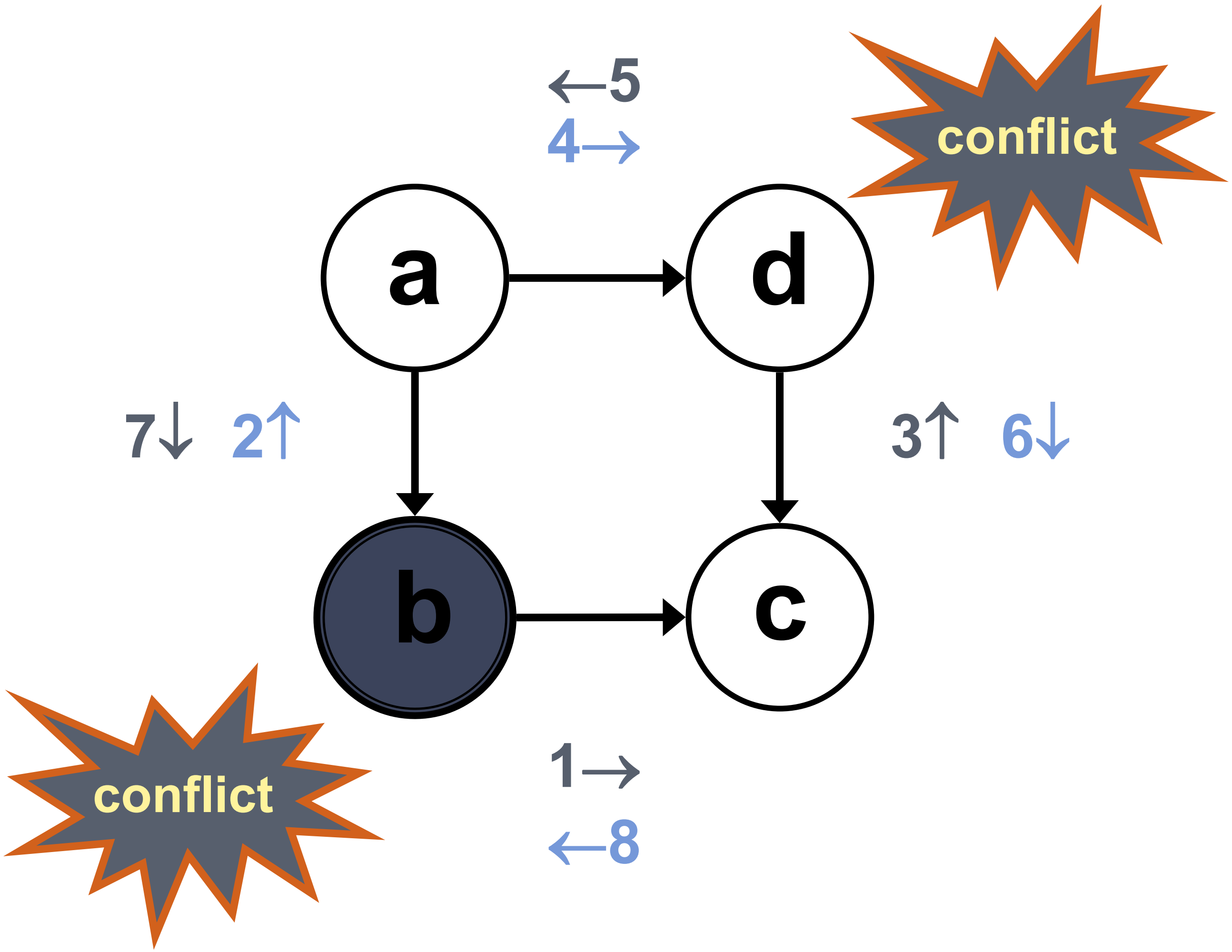
“No, the roads are not icy.” $\Rightarrow$ $P(X_{Icy}=no)=1$



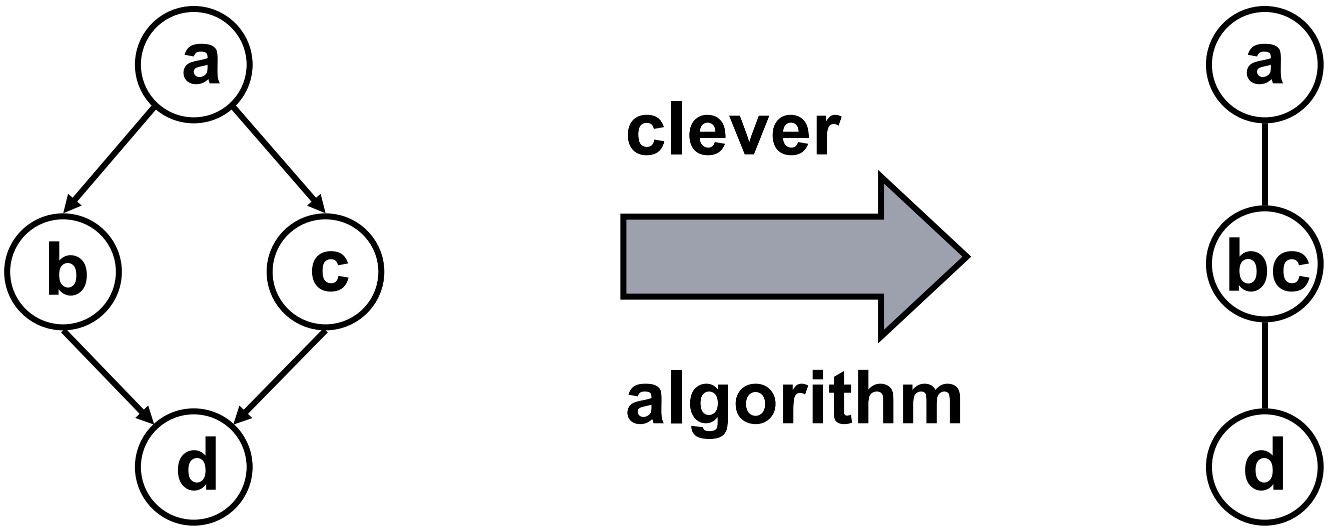
When initiating X_{Icy}
X_{Holmes} becomes independent of X_{Watson} ;

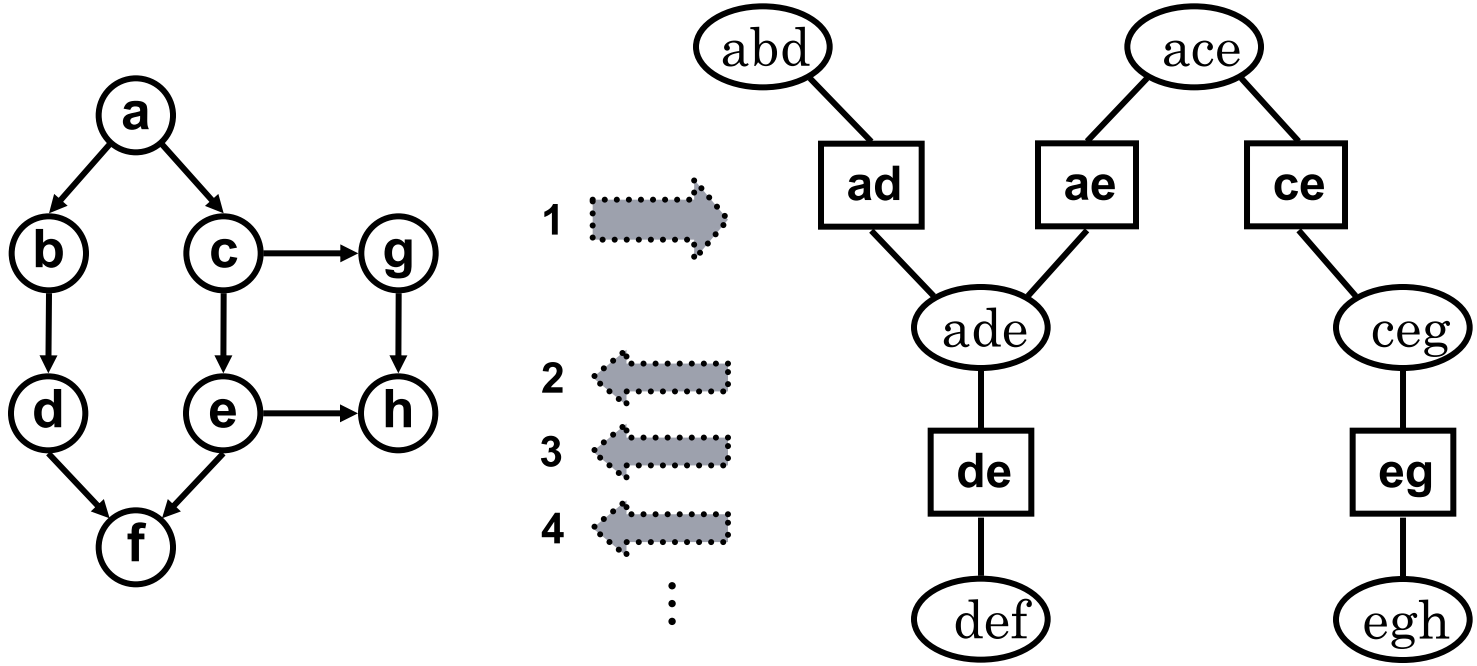
X_{Holmes} $\quad$ X_{Watson} $\perp\!\!\!\perp X_{Icy}$

This naive approach of updating the network inherits oscillation problems!



Idea behind the Junction Tree Algorithm





58 (47)

Bayesian Network

- one-dim. random variables
- conditional probabilities

Secondary Structure/ Junction Tree

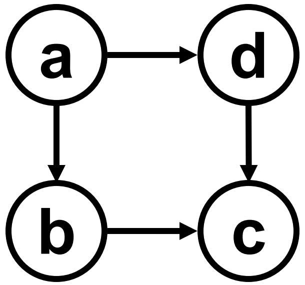
- multi-dim. random variables
- joint probabilities (potentials)

Graph Theory

A **graph** $G = (V, E)$ consists of:

- a set of vertices V
- a set of edges E

Ex. V = {a,b,c,d}, E = {a⌫b, a⌫d, b⌫c, d⌫c}



e=u⌫v : u is a **parent** of v, and v a **child** of u.

pa(u) = the parent set of u, *ch*(u) = the child set of u

Ex. *pa*(b) = *pa*(d) = {a}, *ch*(a) = {b,d}, *ch*(d) = ∅

the **family** of u: *fa*(u) = {u} ∪ *pa*(u)

Ex. *fa*(a) = {a}, *fa*(b) = {a,b}, *fa*(c) = {a,c}, *fa*(d) = {b,c,d}

the **neighbors** of u: *ne*(u) = {u} ∪ *pa*(u) ∪ *ch*(u)

Ex. *ne*(a) = {a,b,d}, *ne*(b) = {a,b,c}, *ne*(c) = {b,c,d}, *ne*(d) = {a,c,d}

Bayesian networks

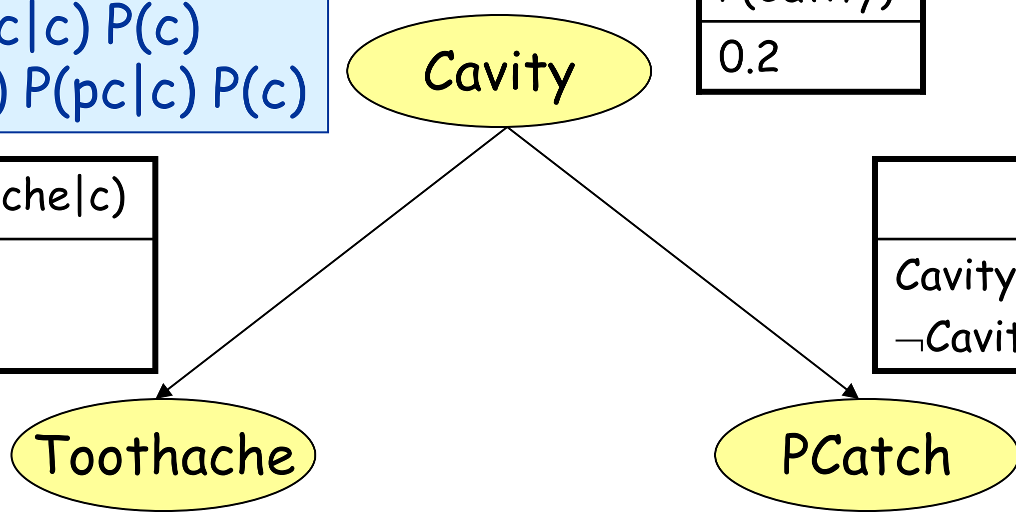
- רשת בייסיאנית מייצגת בגרף את יחסי התלות המותנה בין משתנים ומהווה דרך קומפקטית ואינטואיטיבית להצגת התפלגות ההסתברות המשותפת המלאה (full joint probability distribution) של המשתנים השונים.
- רשתות בייסיאניות משמשות באוסף רחב של אפליקציות: זיהוי ספאם, זיהוי טקסט, רובוטיקה ומערכות דיאגנוסטיקה.

Bayesian networks

- Notice that Cavity is the "cause" of both Toothache and PCatch, and represent the causality links explicitly
- Give the prior probability distribution of Cavity
- Give the conditional probability tables of Toothache and PCatch

$$
\begin{aligned}
P(c \wedge t \wedge pc) &= P(t \wedge pc|c)\,P(c) \\
&= P(t|c)\,P(pc|c)\,P(c)
\end{aligned}
$$

P(Cavity)
0.2



	P(Toothache|c)
Cavity	0.6
¬Cavity	0.1

	P(PClass|c)
Cavity	0.9
¬Cavity	0.02

5 probabilities, instead of 7

Bayesian networks

- רשת בייסיאנית היא גרף מכוון חסר מעגלים עם טבלאות הסתברויות מותנות המשוייכות לצמתיו.
- כל צומת בגרף מייצג משתנה מקרי בעולם וכל קשת מכוונת מצומת X לצומת Y בגרף פירושה: ל-X יש "השפעה ישירה" על Y.
- X (הסיבה) הוא ההורה של Y (התוצאה). כלומר, הרשת מייצגת את כל היחסים ה"סיבתיים" הישירים בין משתנים.

Bayesian networks

- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

- Toothache and PCatch are independent given Cavity (or ¬Cavity), but this relation is hidden in the numbers ! [Verify this]
- Bayesian networks explicitly represent independence among propositions to reduce the number of probabilities defining a belief state

Bayesian networks

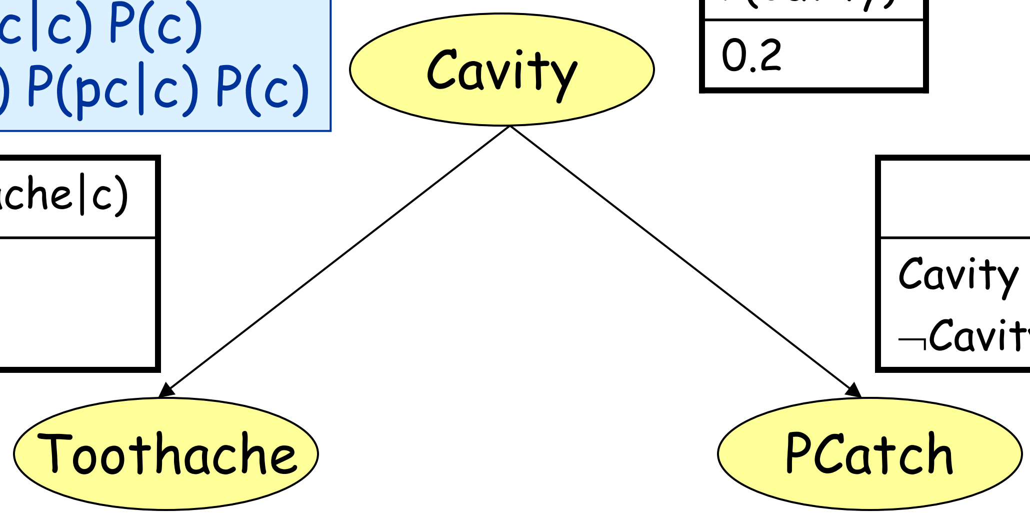
- לכל צומת X_i בגרף משוייכת טבלה המתארת את התפלגות ההסתברות המותנה (conditional probability distribution) של המשתנה המקרי X_i ביחס למשתנים המקריים המשמשים כהוריו בגרף.
- טבלה זו נקראת **טבלת ההסתברות המותנה** (CPT)והיא מכילה את התפלגות ההסתברות המותנה של המשתנה X_i בהינתן כל צירופי הערכים האפשריים של משתני ההורים.

Bayesian networks

- Notice that Cavity is the "cause" of both Toothache and PCatch, and represent the causality links explicitly
- Give the prior probability distribution of Cavity
- Give the conditional probability tables of Toothache and PCatch

$$
\begin{aligned}
P(c \wedge t \wedge pc) &= P(t \wedge pc|c)\,P(c) \\
&= P(t|c)\,P(pc|c)\,P(c)
\end{aligned}
$$

P(Cavity)
0.2



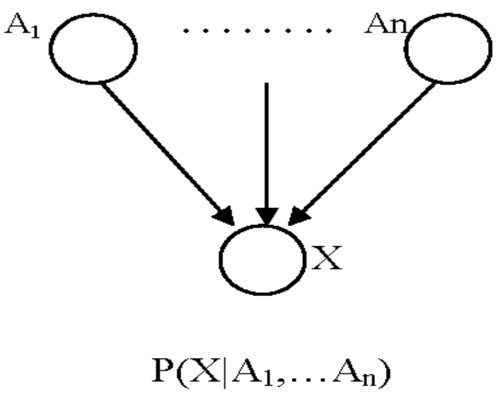
	P(Toothache|c)
Cavity	0.6
¬Cavity	0.1

	P(PClass|c)
Cavity	0.9
¬Cavity	0.02

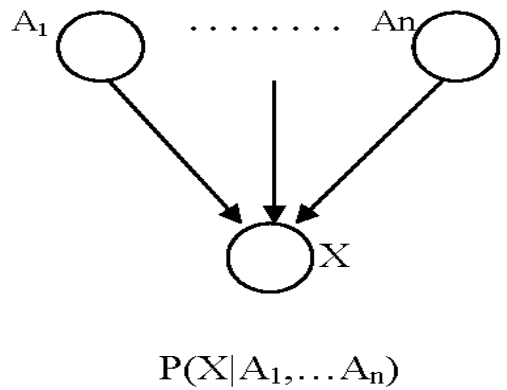
5 probabilities, instead of 7

Bayesian networks

- ההתפלגות המותנה עבור צומת X_i בהינתן הוריו היא:
- $P(X_i \mid Parents(X_i))$
- והיא מכמתת את השפעת ההורים על הצומת.



Bayesian networks



- טבלת ההסתברות המותנה מופיעה להלן:

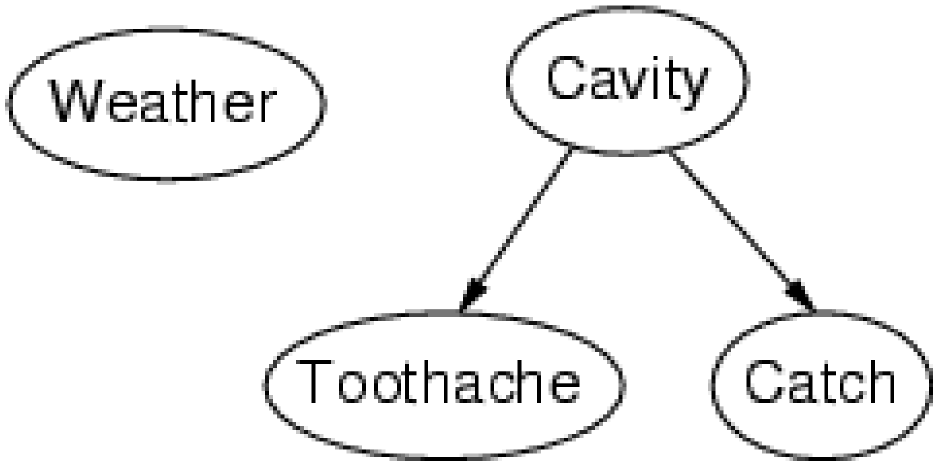
$A_1 \ldots A_n$	P(X)
t,…,t	0.5
t,…,f	0.01
. . .	. . .
f,…,f	0.3

Bayesian networks

- עבור משתנה בוליאני X_i בעל k הורים, יש 2^k שורות בטבלה, שורה לכל צירוף אפשרי של ערכי הוריו.
- משתנה ללא הורים יכיל רק שורה אחת שתייצג את ההסתברות הא-פריורית שלו.
- ההסתברויות בכל שורה בטבלה צריכות להסתכם ל-1.

Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

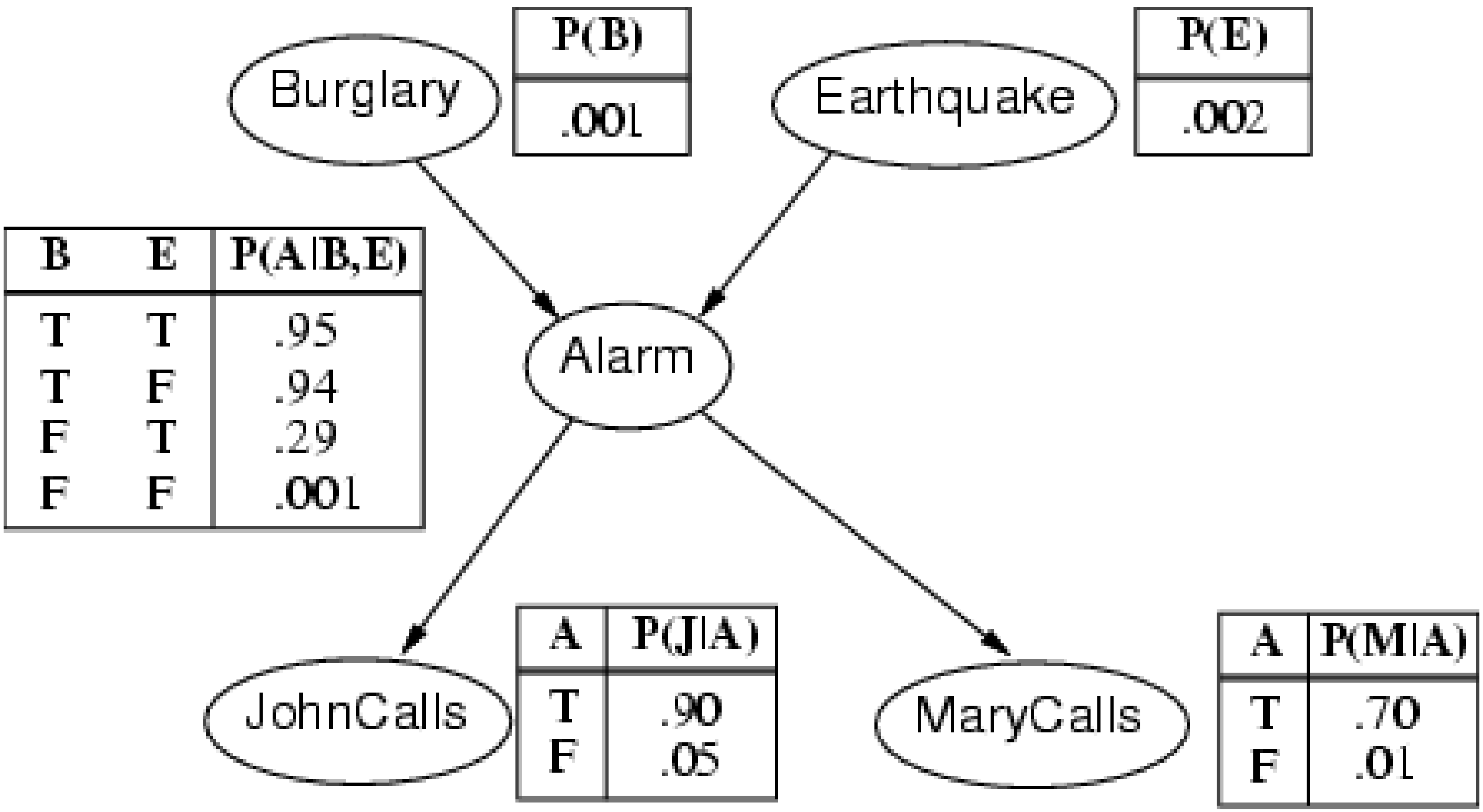
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Example contd.

- ניתן לראות על-פי הקשרים ברשת כי פורץ ורעידת אדמה משפיעים ישירות על ההסתברות שהאזעקה תפעל ואילו התשובה לשאלה האם John ו-Mary מתקשרים תלוייה רק באזעקה.
- נשים לב כי ברשת אין צמתים המתאימים ל"Mary מאזינה כעת למוסיקה רועשת" או "הטלפון מצלצל" ועלול לבלבל את John. גורמים אלה מתומצתים באי-הוודאות המשוייכת לקשרים שבין Alarm לבין JohnCalls ו-MaryCalls.

- בדרך זו הסוכן יכול להתמודד עם כמות גדולה של מידע ("עולם גדול"), לפחות בקירוב. ניתן לשפר את דרגת הקירוב (הדיוק) אם נציג מידע רלבנטי נוסף.

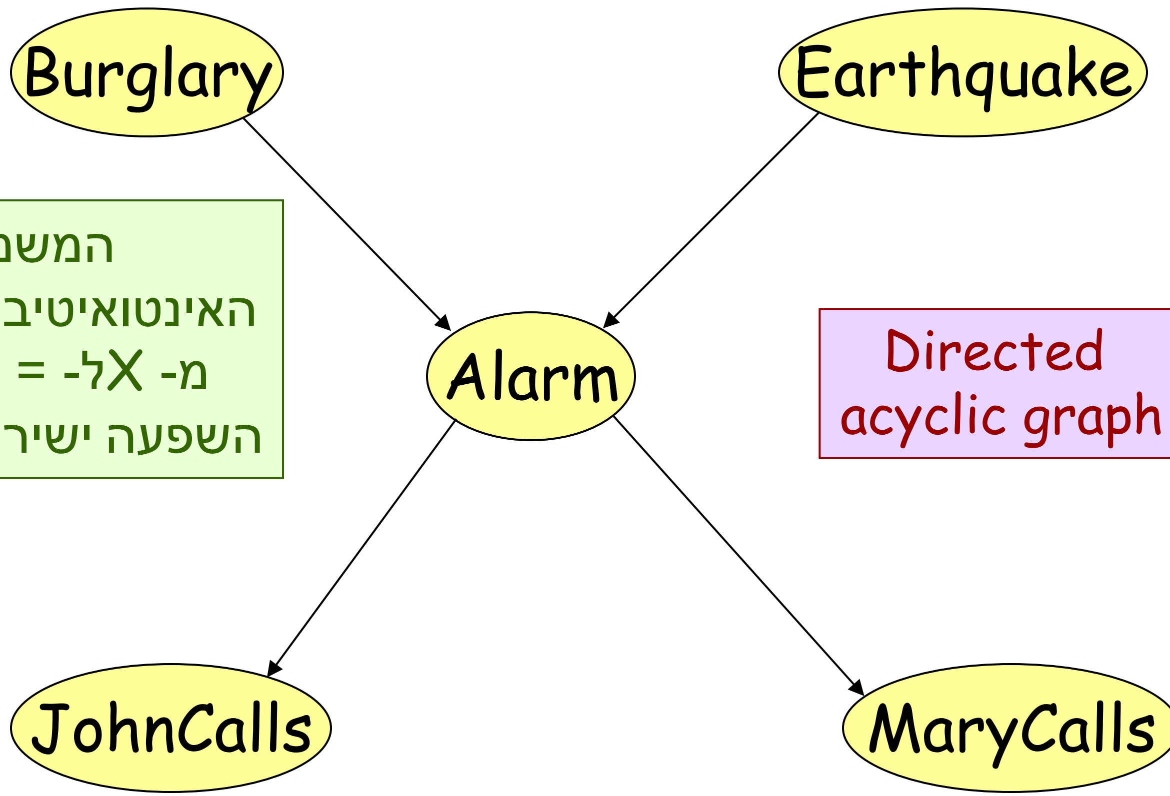
Semantics

- רשתות בייסיאניות הן דרך לתיאור ההסתברויות המשותפות תוך שימוש בהתפלגויות מקומיות (הסתברויות מותנות).
- מתוך הרשת ניתן לחשב כל ערך של ההסתברות המשותפת של קבוצת המשתנים המקריים.
-

SEMANTICS

- נגדיר את משמעות (סמנטיקה) הרשת הבייסיאנית בעזרת האופן בו היא מייצגת התפלגות משותפת מעל כל המשתנים:
- כל ערך בהתפלגות המשותפת של המשתנים X_{i-1}, ..., X_1 ניתן לחישוב כמכפלה של האיברים המתאימים של ה-CPTs ברשת בעזרת הנוסחה (14.2)שבסוף עמ' 513 בספר.

A More Complex BN



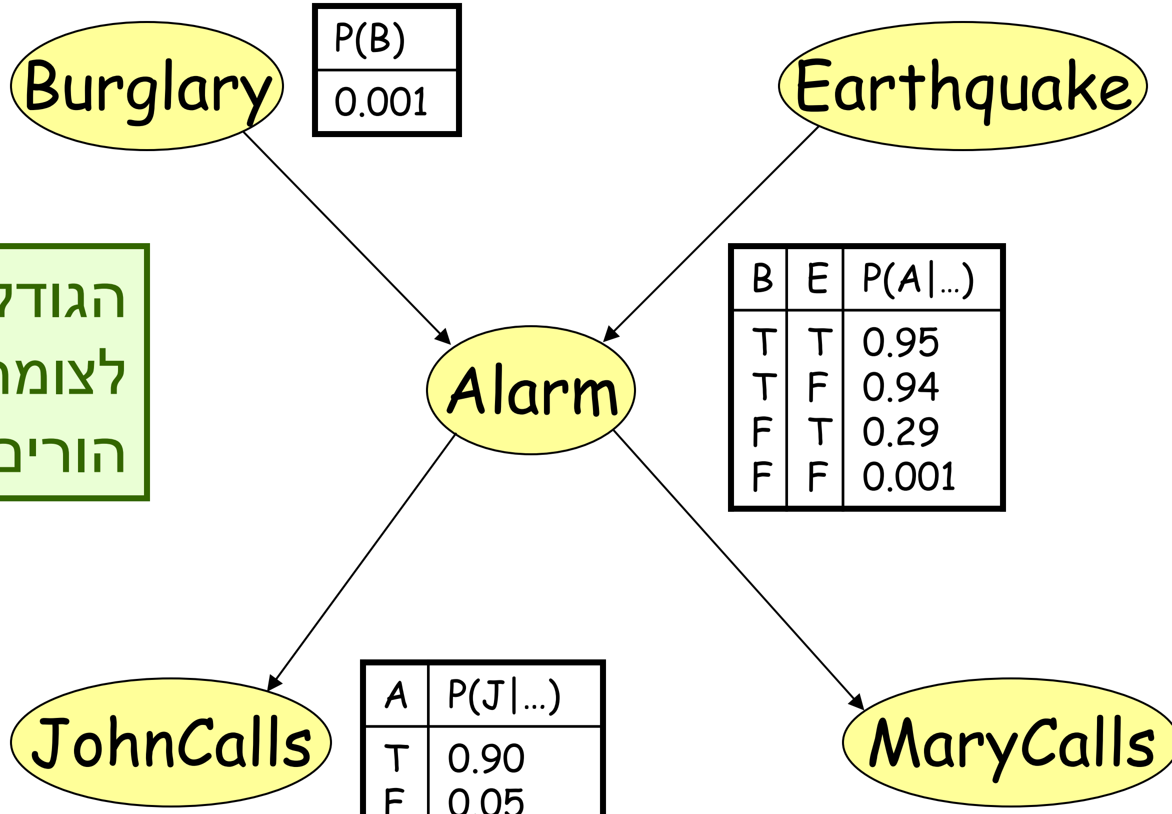
המשמעות
האינטואיטיבית של קשת
מ- X ל- Y = ← יש
השפעה ישירה ל X על Y

Directed acyclic graph

causes

effects

A More Complex BN



P(B)
0.001

P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

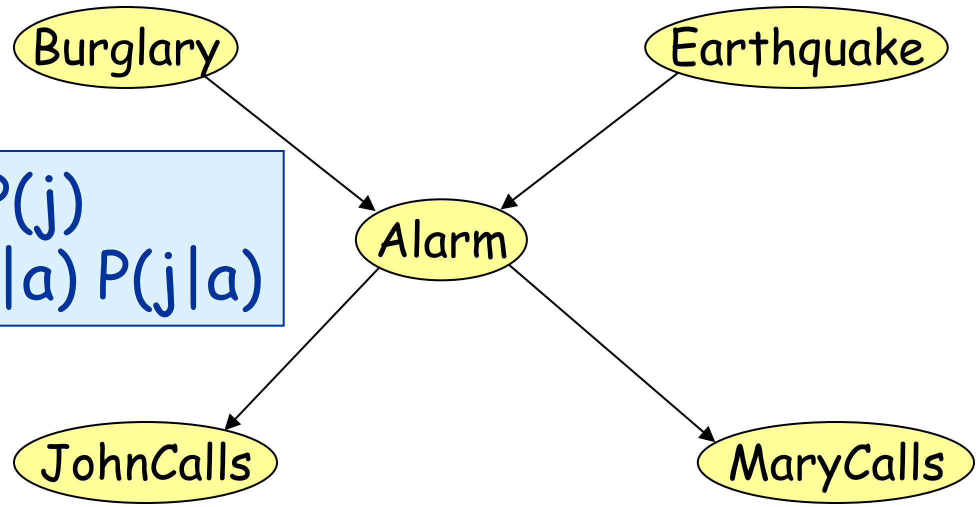
A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

הגודל של CPT
לצומת עם K
הורים הוא?

10 probabilities, instead of 31

What does the BN encode?

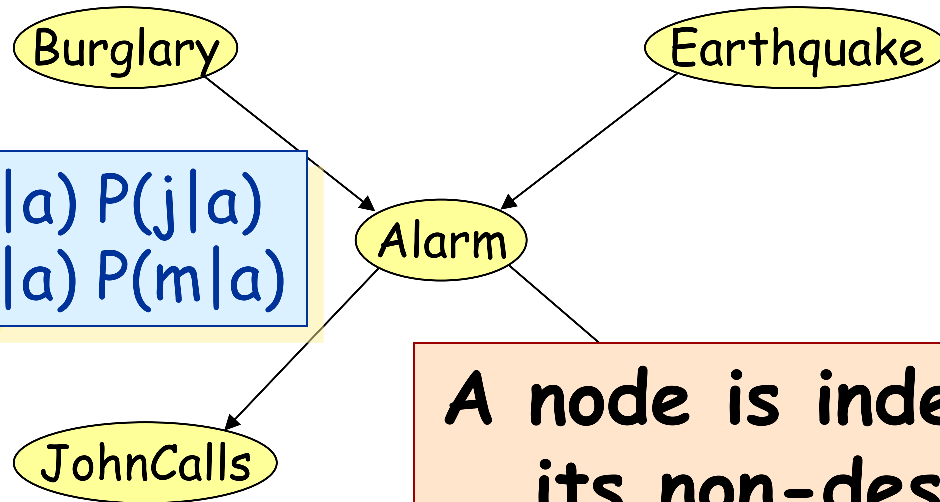


$P(b \wedge j) \neq P(b)\,P(j)$
$P(b \wedge j|a) = P(b|a)\,P(j|a)$

Each of the beliefs JohnCalls and MaryCalls is independent of Burglary and Earthquake given Alarm or ¬Alarm

For example, John does not observe any burglaries directly

What does the BN encode?



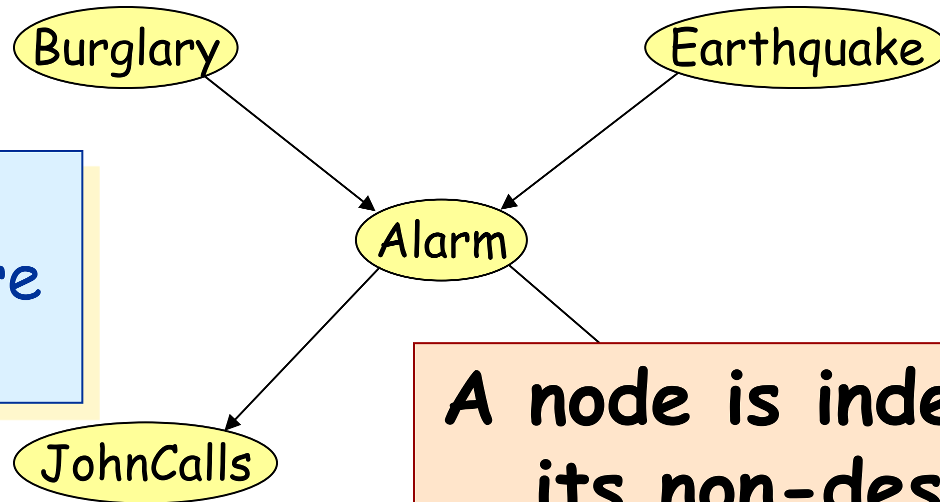
$P(b \wedge j \mid a) = P(b \mid a)\, P(j \mid a)$
$P(j \wedge m \mid a) = P(j \mid a)\, P(m \mid a)$

A node is independent of its non-descendants given its parents

The beliefs JohnCalls and MaryCalls are independent given Alarm or $\neg$Alarm

For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

WHAT DOES THE BN ENCODE?



Burglary and Earthquake are independent

A node is independent of its non-descendants given its parents

The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

Locally Structured World

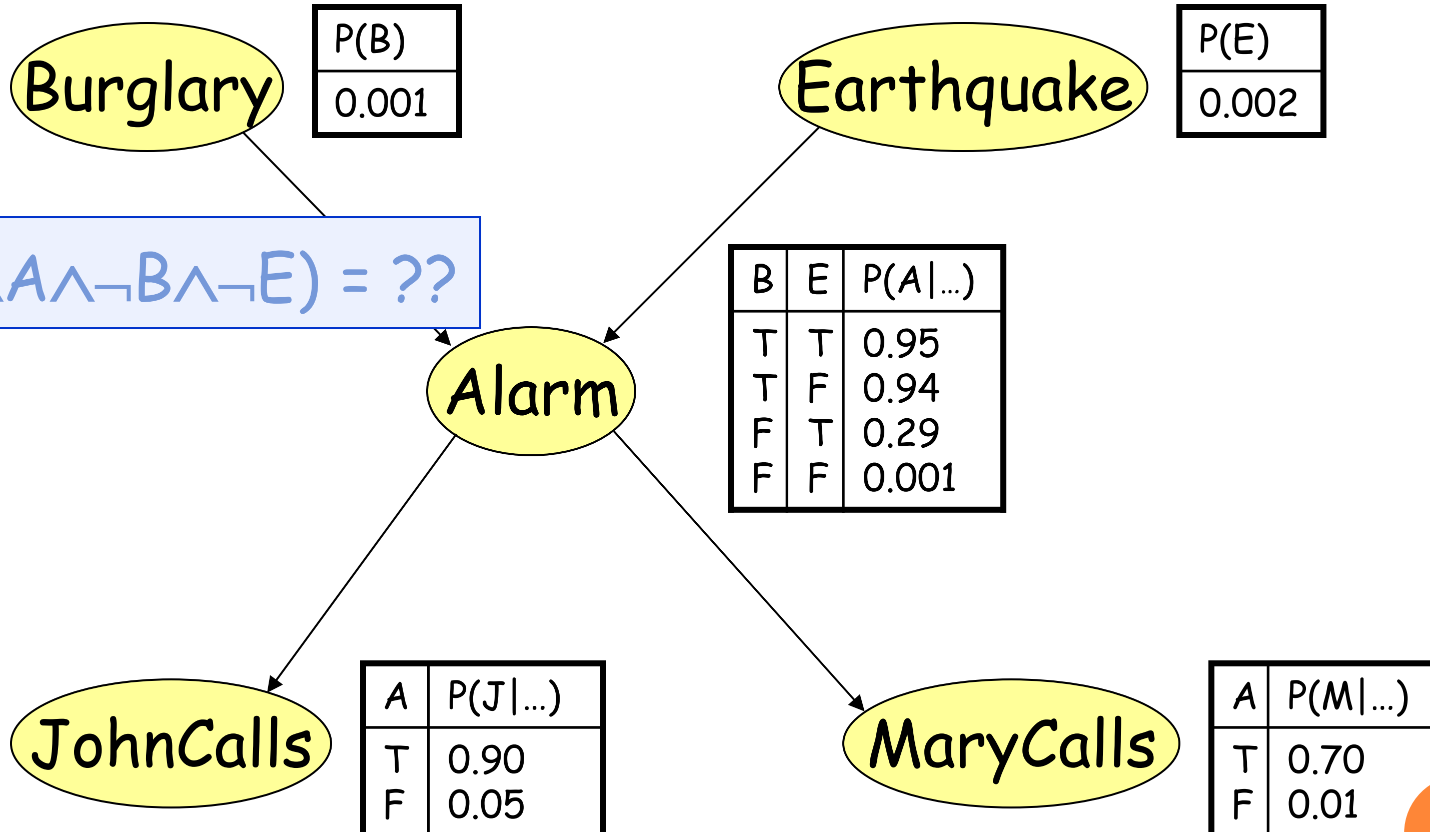
- A world is locally structured (or sparse) if each of its components interacts directly with relatively few other components
- In a sparse world, the CPTs are small and the BN contains much fewer probabilities than the full joint distribution
- If the # of entries in each CPT is bounded by a constant, i.e., O(1), then the # of probabilities in a BN is linear in n – the # of propositions – instead of 2^n for the joint distribution

But does a BN represent a belief state?

In other words, can we compute the full joint distribution of the propositions from it?

Calculation of Joint Probability

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = ??$



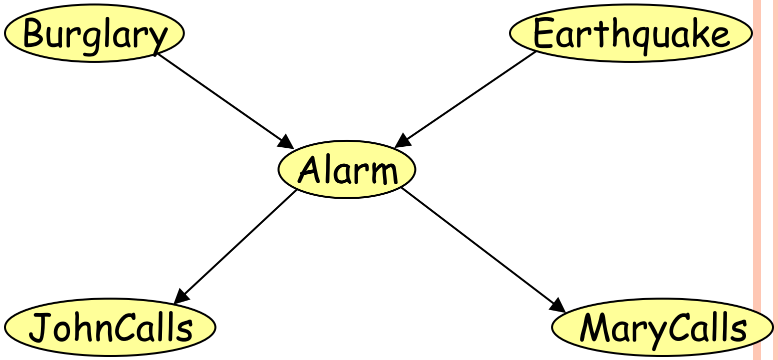
P(B)
0.001

P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

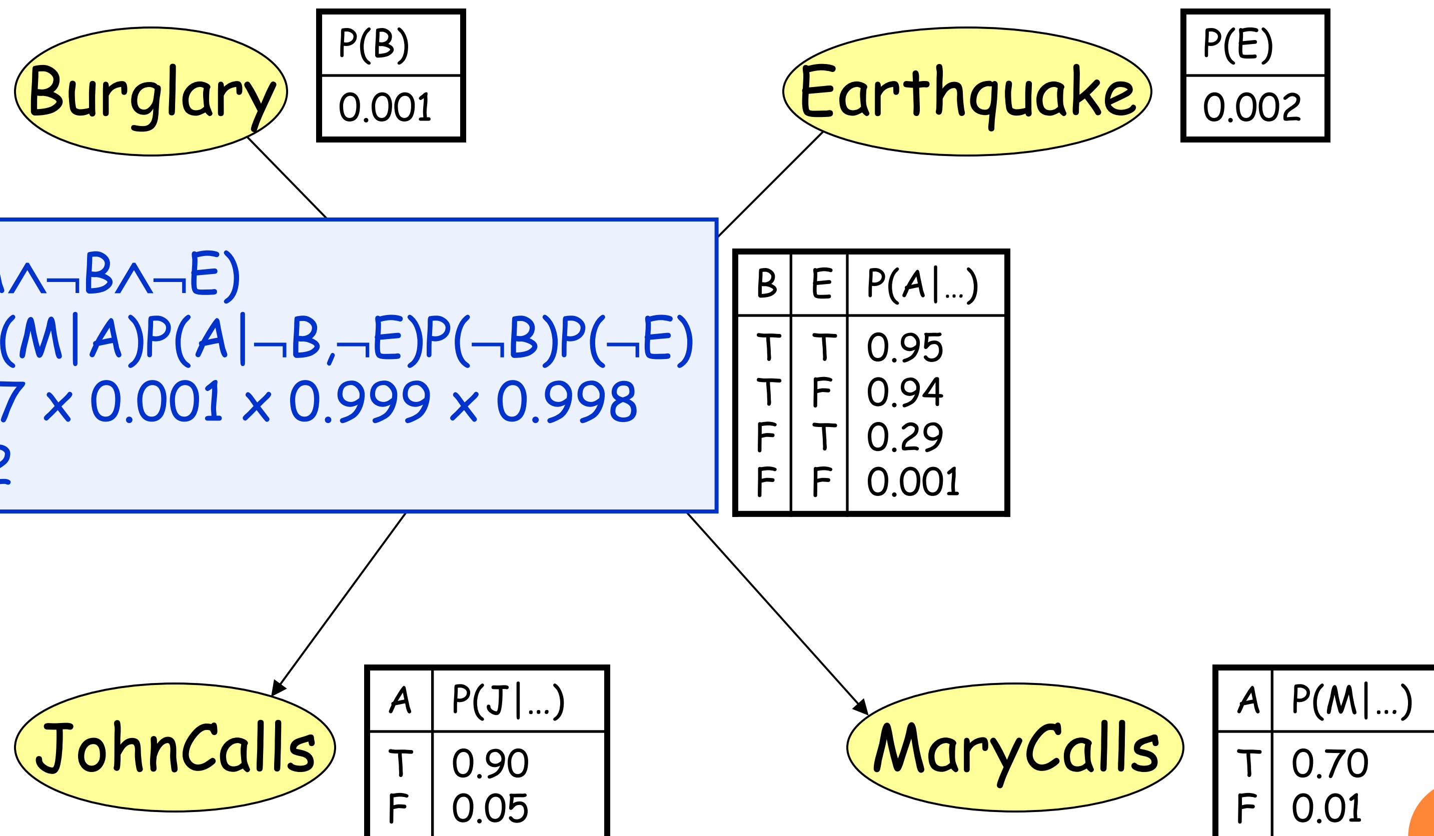
A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01



- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
  $= P(J \wedge M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  $= P(J | A, \neg B, \neg E) \times P(M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  (J and M are independent given A)
- $P(J | A, \neg B, \neg E) = P(J | A)$
  (J and $\neg B \wedge \neg E$ are independent given A)
- $P(M | A, \neg B, \neg E) = P(M | A)$
- $P(A \wedge \neg B \wedge \neg E) = P(A | \neg B, \neg E) \times P(\neg B | \neg E) \times P(\neg E)$
  $= P(A | \neg B, \neg E) \times P(\neg B) \times P(\neg E)$
  ($\neg B$ and $\neg E$ are independent)
- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J | A) P(M | A) P(A | \neg B, \neg E) P(\neg B) P(\neg E)$

Calculation of Joint Probability



P(B)
0.001

P(E)
0.002

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B, \neg E)P(\neg B)P(\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

Calculation of Joint Probability

Burglary

P(B)
0.001

Earthquake

P(E)
0.002

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)P(\neg B)P(\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

$$P(x_1 \wedge x_2 \wedge \ldots \wedge x_n) = \Pi_{i=1,\ldots,n} P(x_i|\text{parents}(X_i))$$

→ full joint distribution table

Calculation of Joint Probability

Burglary

P(B)
0.001

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

Since a BN defines the full joint distribution of a set of propositions, it represents a belief state

T	F	0.94
F	T	0.29
F	F	0.001

$$P(x_1 \wedge x_2 \wedge \ldots \wedge x_n) = \Pi_{i=1,\ldots,n} P(x_i|\text{parents}(X_i))$$

→ full joint distribution table

- כאשר אני מגיע הביתה, אני רוצה לדעת האם מישהו מבני משפחתי נמצא בבית לפני שאני נכנס.
- נתון המידע שלהלן:

1. כאשר אשתי עוזבת את הבית, היא מדליקה את האור מחוץ לבית לעתים קרובות (אך לא תמיד).(כמו כן, היא מדליקה לעתים את האור כשאמור להגיע אורח).
2. כאשר אין איש בבית, משאירים את הכלב לעתים קרובות בחוץ.
3. גם כאשר לכלב יש בעיות מעיים, משאירים אותו לעתים קרובות בחוץ.
4. אם הכלב בחוץ, יתכן ואשמע אותו נובח (למרות שיתכן שאינו נובח, או שאני עשוי לשמוע נביחות של כלב אחר ואחשוב שזהו הכלב שלי).

שלב ראשון נגדיר משתנים

- נגדיר 5 משתנים מקריים (בוליאניים):
  1. O: כולם מחוץ לבית
  2. L: האור דולק
  3. D: הכלב בחוץ
  4. B: לכלב יש בעיות מעיים
  5. H: אני יכול לשמוע את הכלב נובח

שלב שני : השפעות

1. H מושפע ישירות רק מ-D. לכן בהינתן D, יש אי תלות מותנה בין H לבין L,O, ו-B.
2. D מושפע ישירות רק מ-O ומ-B. לכן בהינתן O ו-B, יש אי תלות מותנה בין D לבין L.
3. L מושפע ישירות רק מ-O. לכן בהינתן O, יש אי תלות מותנה בין L לבין H,D, ו-B.
4. O ו-B הם בלתי תלויים.

שלב שלישי – רשת בייסיאנית

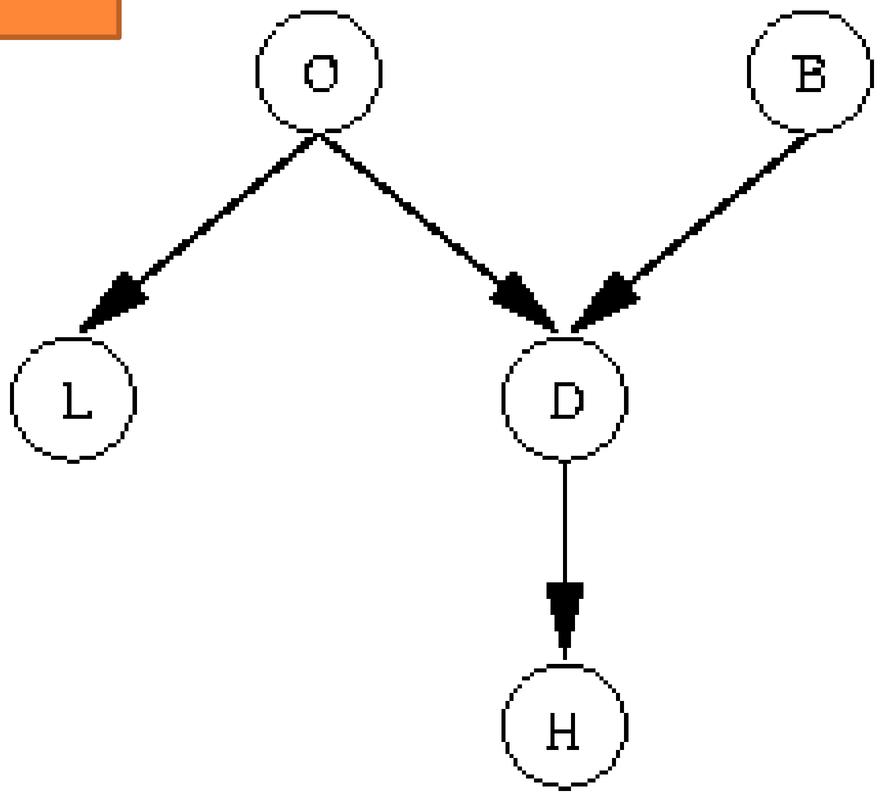
O: כולם מחוץ לבית

L: האור דולק

D: הכלב בחוץ

B: לכלב יש בעיות מעיים

H: אני יכול לשמוע את הכלב נובח



שלב רביעי : נוסיף את הטבלאות (CPT)

- עבור שורש (צומת ללא הורים), מוגדרת ההסתברות האפריורית של המשתנה המקרי המתאים.
- לכל צומת שאינו שורש, מוגדרת טבלת ה-CPT (ההסתברויות המותנות של המשתנה המתאים לצומת בהינתן כל הצירופים האפשריים של ערכי הוריו).

P(B)
0.3

P(O)
0.6

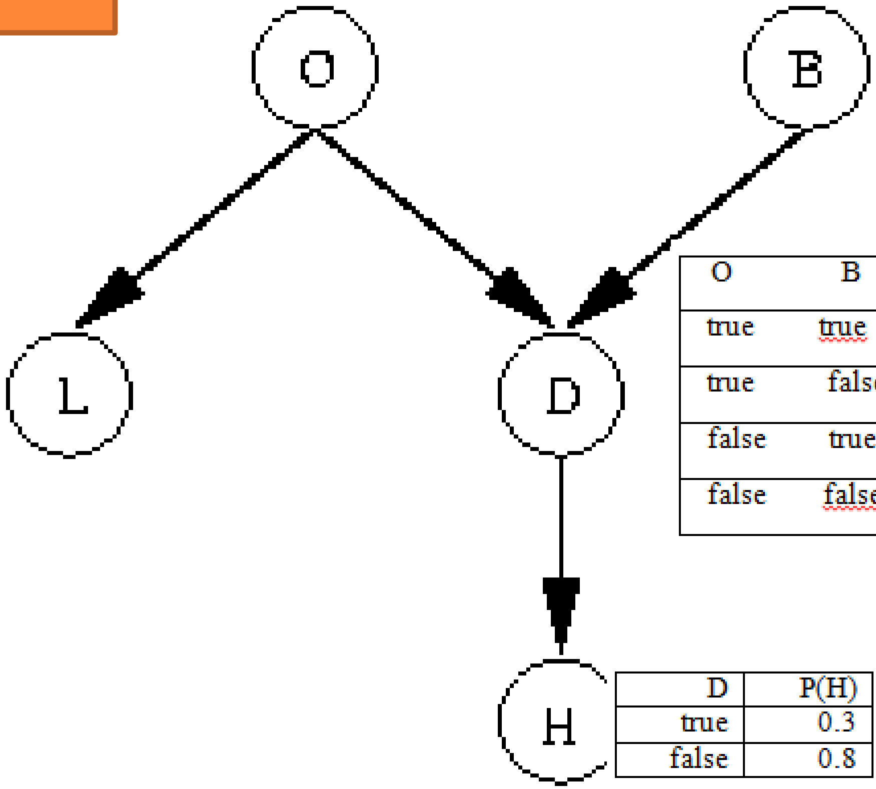
D	P(H)
true	0.3
false	0.8

O	P(L)
true	0.3
false	0.6

O	B	P(D)
true	true	0.5
true	false	0.1
false	true	0.1
false	false	0.2

שלב רביעי : נוסיף את הטבלאות (CPT)

O: כולם מחוץ לבית
L: האור דולק
D: הכלב בחוץ
B: לכלב יש בעיות מעיים
H: אני יכול לשמוע את הכלב נובח



P(O)
0.6

P(B)
0.3

O	P(L)
true	0.3
false	0.6

O	B	P(D)
true	true	0.5
true	false	0.1
false	true	0.1
false	false	0.2

D	P(H)
true	0.3
false	0.8

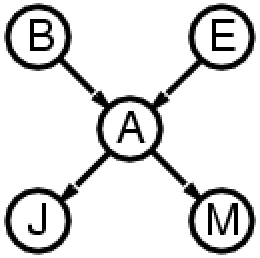
נשים לב כי...

- בדוגמה זו אוחסנו ברשת הבייסיאנית 10 הסתברויות, בעוד שהתפלגות ההסתברות המשותפת המלאה דורשת טבלה המכילה $2^5 = 32$ הסתברויות.
- האי תלות המותנה של חלק גדול מהמשתנים מאפשרת להפחית את מספר ההסתברויות.
-

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$



e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

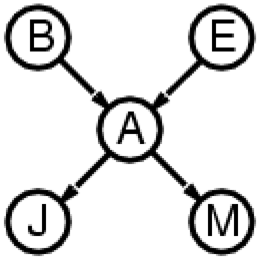
$$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$$

Compactness

- כאמור רשת בייסיאנית היא ייצוג שלם ולא יתיר (nonredundant) של התחום ולעתים קרובות ייצוג זה הינו קטן משמעותית מייצוג ההתפלגות המשותפת המלאה.
- לכן ניתן להשתמש בו בתחומים שבהם יש הרבה משתנים.
- רשת בייסיאנית אינה מכילה ערכים הסתברותיים מיותרים ולכן היא תמיד עקבית

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1-p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)

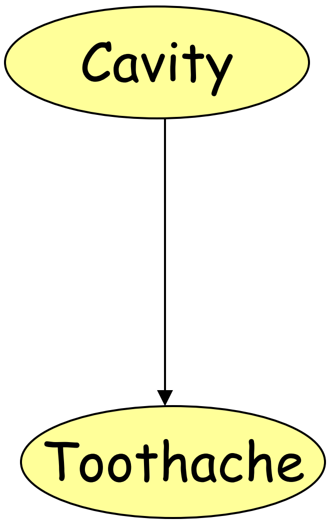
Compactness

- החסכון במקום (קומפקטיות) מתאפשר ברשת בייסיאנית הודות לניצול העובדה שבתחומי בעיות רבים בעולם האמיתי, התלות בין משתנים היא בדרך כלל מקומית, כך שיש הרבה משתנים שהאי תלות ביניהם מותנה.
- לכן בדרך כלל אין צורך במקום אחסון בגודל אקספוננציאלי כדי לשמור את כל המידע שבטבלת התפלגות ההסתברות המשותפת. השימוש באי תלות מותנה מפחית ברוב המקרים את גודל הייצוג של ההתפלגות המשותפת מגודל אקספוננציאלי במספר המשתנים (הצמתים) לגודל לינארי במספר המשתנים.
- (חישוב ההסתברות של כל איבר בטבלת ה-CPT הינו מהיר – הוא לינארי במספר הצמתים).

Constructing Bayesian networks

- באופן אינטואיטיבי, כדי לבנות רשת בייסיאנית עבור קבוצה נתונה של משתנים, יש לשרטט קשתות ממשתנים המהווים סיבה למשתנים שהם התוצאה הישירה. ההורים של צומת X_i צריכים להכיל את כל הצמתים $X_1, \ldots, X_{i-1}$ המשפיעים ישירות על X_i.
- שני משתנים שאינם מחוברים ישירות על ידי קשת בגרף (הרשת הבייסיאנית) יכולים עדיין להשפיע זה על זה.
-

Querying the BN



P(C)
0.1

C	P(T|c)
T	0.4
F	0.01111

- The BN gives P(t|c)
- What about P(c|t)?
- P(Cavity|t)
  = P(Cavity $\wedge$ t)/P(t)
  = P(t|Cavity) P(Cavity) / P(t)
  [Bayes' rule]

- P(c|t) = α P(t|c) P(c)

- Querying a BN is just applying the trivial Bayes' rule on a larger scale

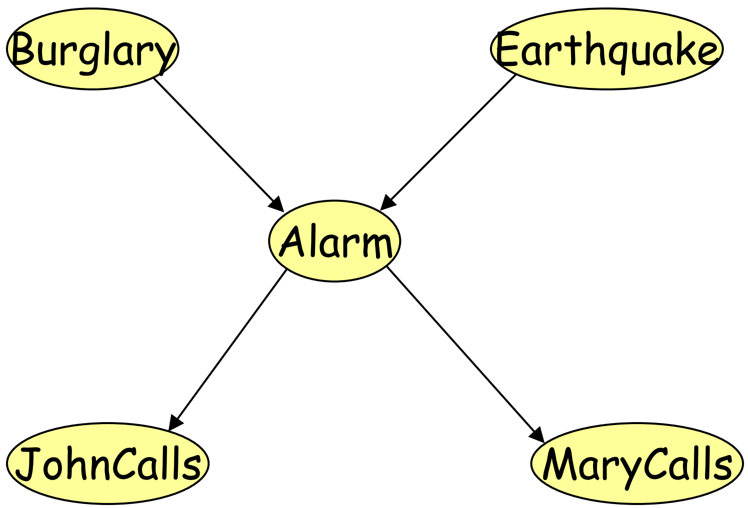
Querying the BN

- New evidence E indicates that JohnCalls with some probability p
- We would like to know the posterior probability of the other beliefs, e.g. P(Burglary|E)
- $$\begin{aligned} P(B|E) &= P(B \wedge J|E) + P(B \wedge \neg J|E) \\ &= P(B|J,E)\, P(J|E) + P(B\,|\neg J,E)\, P(\neg J|E) \\ &= P(B|J)\, P(J|E) + P(B|\neg J)\, P(\neg J|E) \\ &= p\, P(B|J) + (1-p)\, P(B|\neg J) \end{aligned}$$
- We need to compute $P(B|J)$ and $P(B|\neg J)$

Querying the BN

- $P(b|J) = \alpha\, P(b \wedge J)$
  $= \alpha\, \Sigma_m \Sigma_a \Sigma_e P(b \wedge J \wedge m \wedge a \wedge e)$ [marginalization]
  $= \alpha\, \Sigma_m \Sigma_a \Sigma_e P(b)P(e)P(a|b,e)P(J|a)P(m|a)$ [BN]
  $= \alpha\, P(b)\Sigma_e P(e)\Sigma_a P(a|b,e)P(J|a)\Sigma_m P(m|a)$ [re-ordering]
- Depth-first evaluation of $P(b|J)$ leads to computing each of the 4 following products twice:
  $P(J|A)\,P(M|A)$, $P(J|A)\,P(\neg M|A)$, $P(J|\neg A)\,P(M|\neg A)$, $P(J|\neg A)\,P(\neg M|\neg A)$
- Bottom-up (right-to-left) computation + caching – e.g., variable elimination algorithm (see R&N) – avoids such repetition
- For singly connected BN, the computation takes time linear in the total number of CPT entries ($\rightarrow$ time linear in the # propositions if CPT's size is bounded)

Comparison to Classical Logic

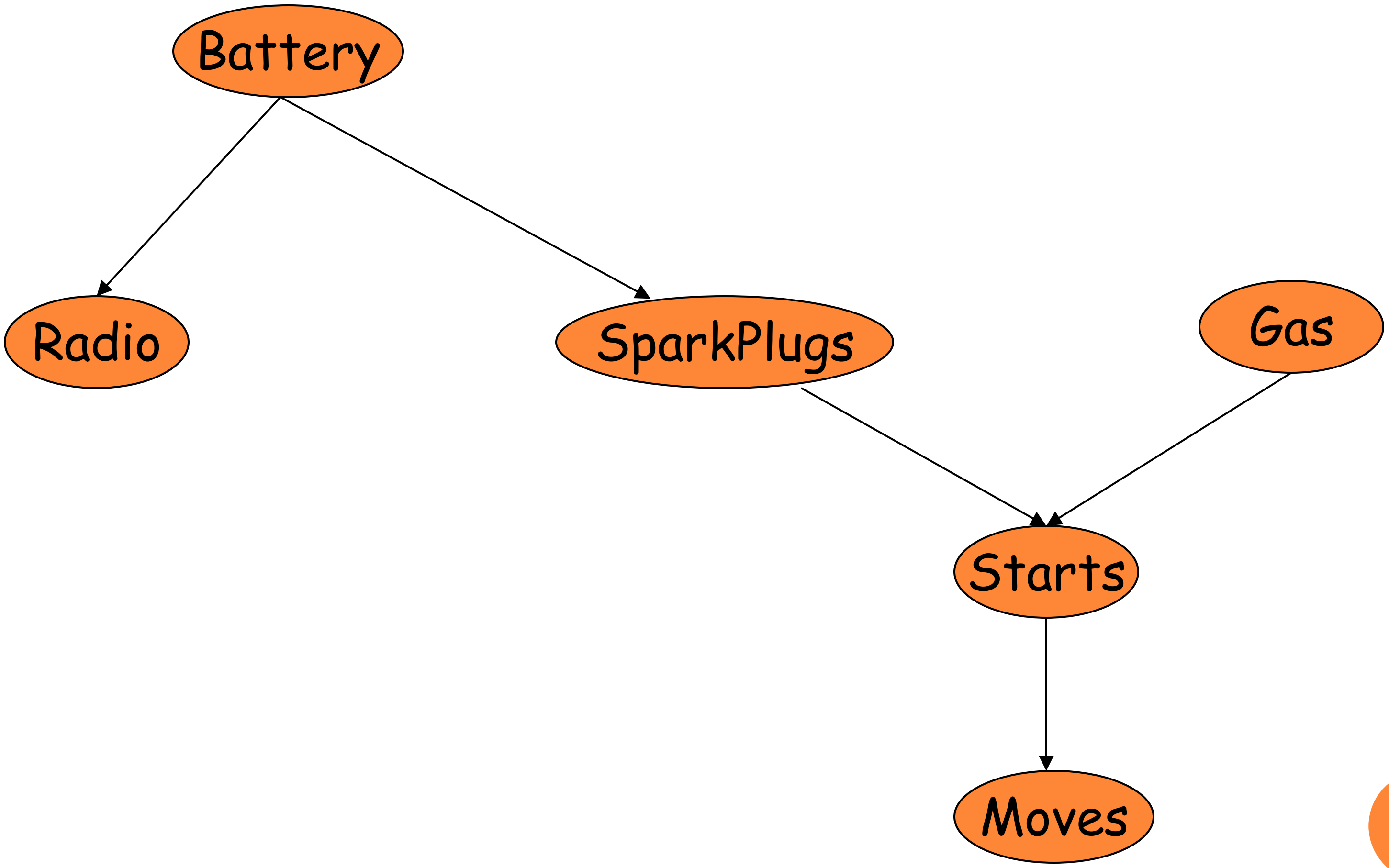


Burglary → Alarm
Earthquake → Alarm
Alarm → JohnCalls
Alarm → MaryCalls

If the agent observes ¬JohnCalls, it infers ¬Alarm, ¬MaryCalls, ¬Burglary, and ¬Earthquake

If it observes JohnCalls, then it infers nothing

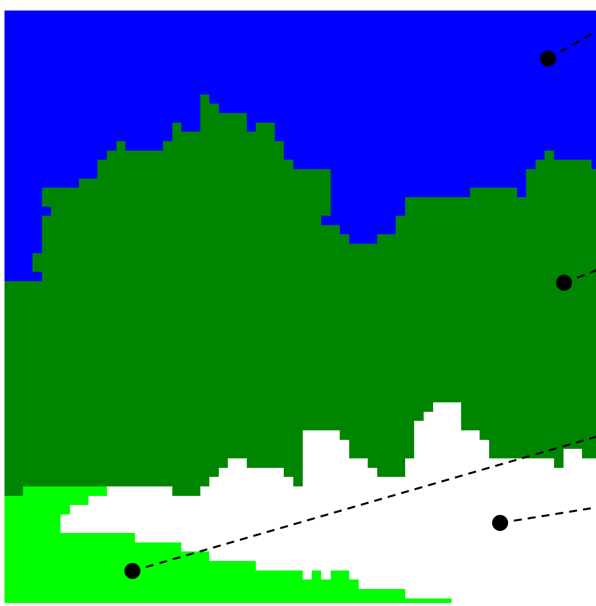
More Complicated Singly-Connected Belief Net

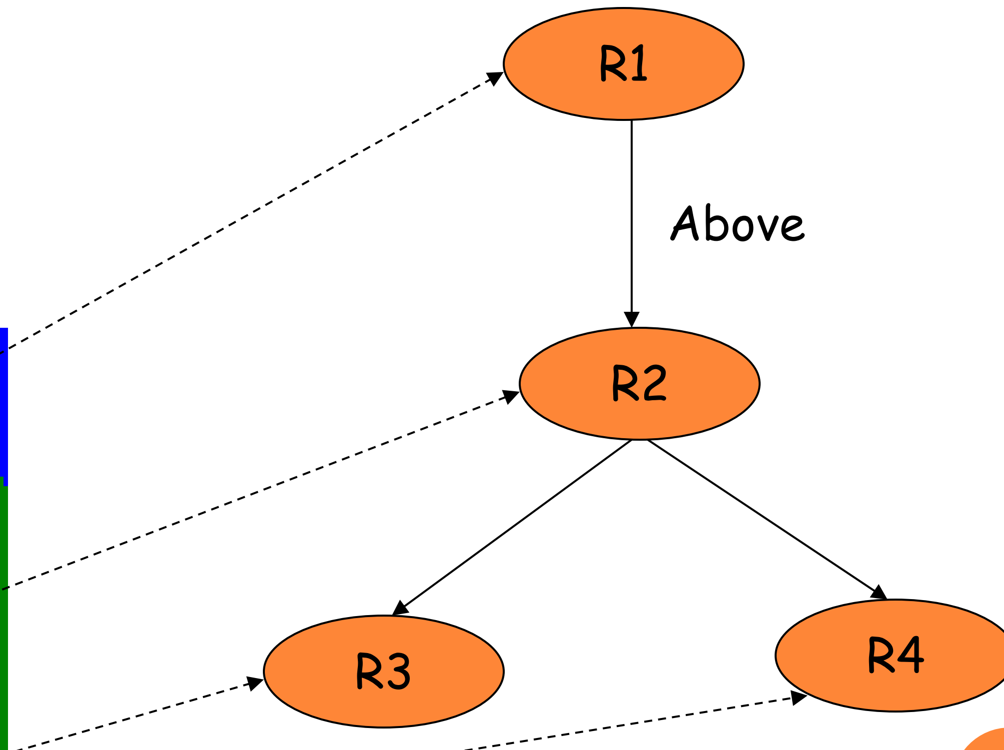


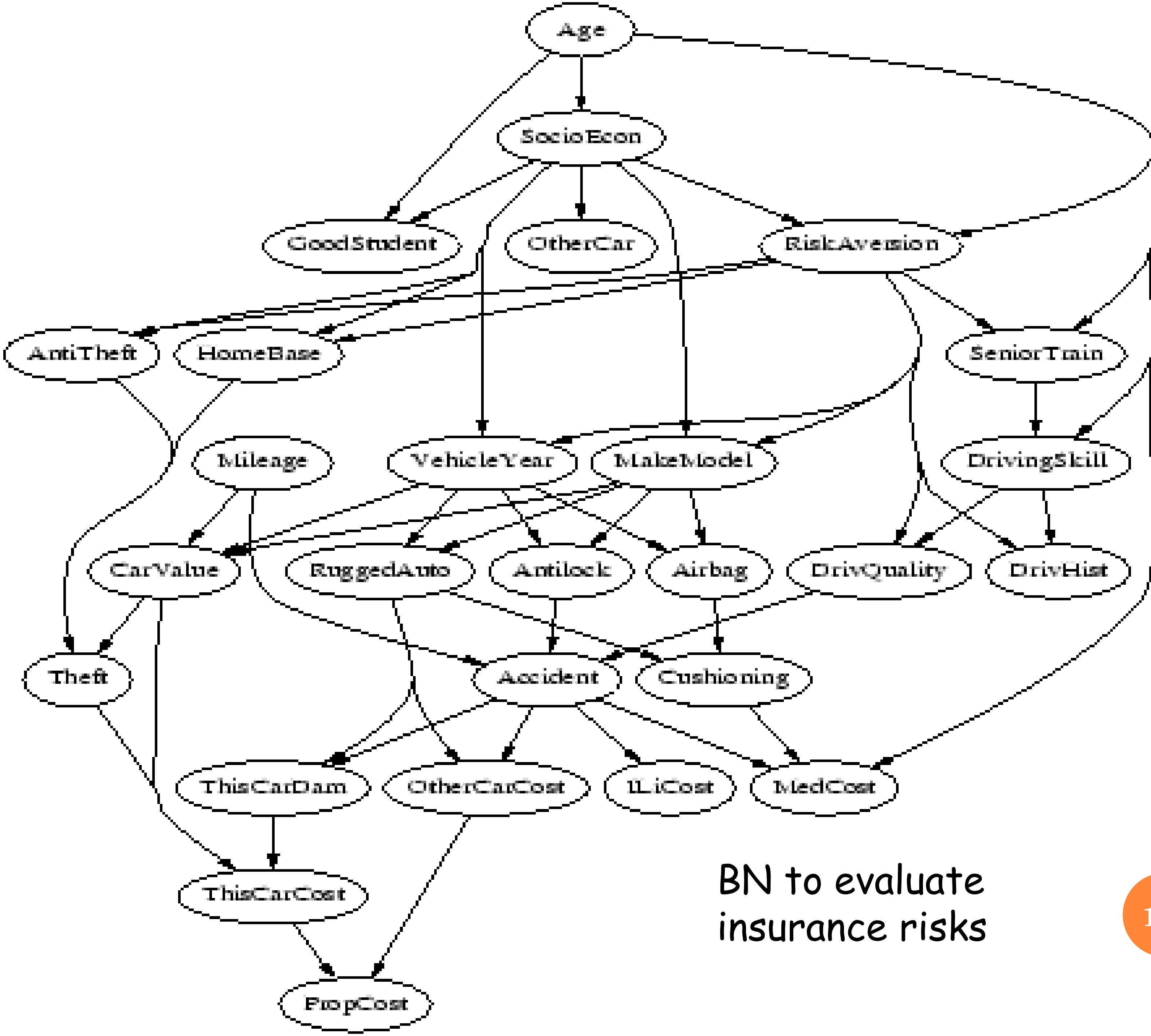
Some Applications of BN

- Medical diagnosis, e.g., lymph-node diseases
- Troubleshooting of hardware/software systems
- Fraud/uncollectible debt detection
- Data mining
- Analysis of genetic sequences
- Data interpretation, computer vision, image understanding

Region = {Sky, Tree, Grass, Rock}







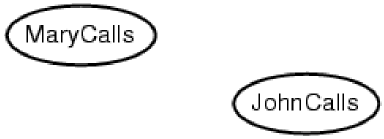
BN to evaluate
insurance risks

Constructing Bayesian networks

- עבור קבוצה נתונה של משתנים מקריים, הצגת ההסתברות המשותפת כרשת בייסיאנית איננה יחידה. למעשה, לכל סדר שנבחר על המשתנים המקריים אנו עשויים לקבל רשת אחרת. בדרך כלל ננסה לסדר את המשתנים לפי סדר סיבתי – מהסיבות הראשוניות לכיוון התוצאות. סדר זה יתן לנו לרוב רשת דלילה המנצלת אי תלויות מותנות רבות.
- ואולם, כל הרשתות המתקבלות מייצגות את אותו מידע. במלים אחרות, מכל רשת שנבנה ניתן לחשב כל ערך בהתפלגות ההסתברות המשותפת.
-
- הרשת "הטובה ביותר" מתקבלת כאשר בצעד 1 של האלגוריתם, המשתנים מסודרים כך שכל משתנה בא לפני כל ילדיו (הסיבות לפני התוצאות). באופן זה הצמתים הראשונים צריכים להיות השורשים, לאחריהם הצמתים המושפעים ישירות וכן הלאה.
-
- האלגוריתם לא יבנה רשת שאינה חוקית במובן של הפרת חוקי ההסתברות.

EXAMPLE

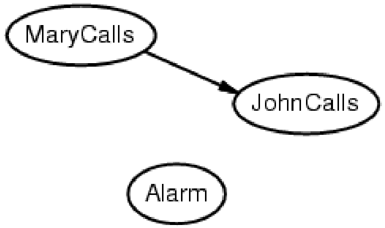
- Suppose we choose the ordering *M, J, A, B, E*
-



$$\mathbf{P}(J \mid M) = \mathbf{P}(J)?$$

EXAMPLE

- Suppose we choose the ordering *M, J, A, B, E*
-



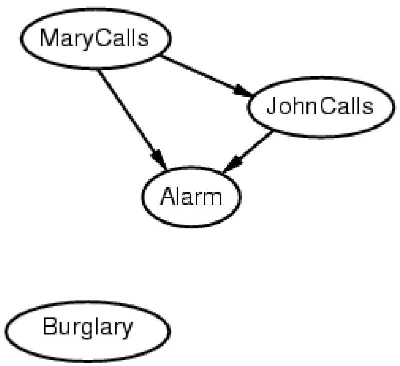
$\boldsymbol{P}(J \mid M) = \boldsymbol{P}(J)$?

No

$\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A \mid J)$? $\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A)$?

EXAMPLE

- Suppose we choose the ordering *M, J, A, B, E*
-



$\boldsymbol{P}(J \mid M) = \boldsymbol{P}(J)$?

No

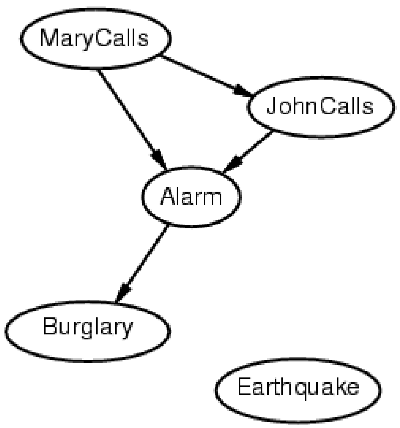
$\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A \mid J)$? $\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A)$? **No**

$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$?

$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B)$?

EXAMPLE

- Suppose we choose the ordering M, J, A, B, E
-



P*(J | M)* = ***P****(J)?*

No

P*(A | J, M)* = ***P****(A | J)*? ***P****(A | J, M)* = ***P****(A)*? **No**

P*(B | A, J, M)* = ***P****(B | A)*? **Yes**

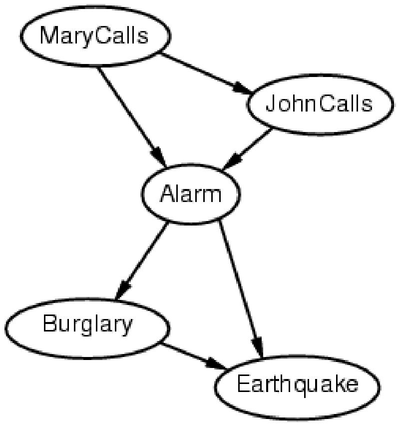
P*(B | A, J, M)* = ***P****(B)*? **No**

P*(E | B, A ,J, M)* = ***P****(E | A)*?

P*(E | B, A, J, M)* = ***P****(E | A, B)*?

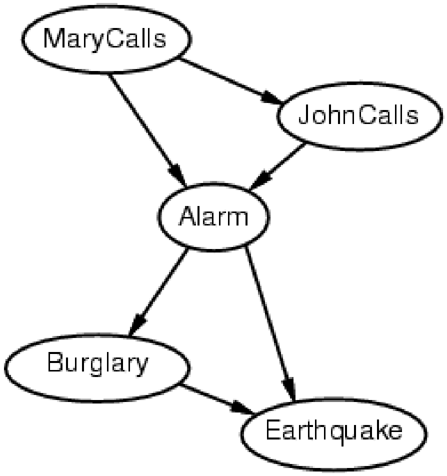
EXAMPLE

- Suppose we choose the ordering M, J, A, B, E
-



$\boldsymbol{P}(J \mid M) = \boldsymbol{P}(J)$?
No
$\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A \mid J)$? $\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A)$? **No**
$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$? **Yes**
$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B)$? **No**
$\boldsymbol{P}(E \mid B, A ,J, M) = \boldsymbol{P}(E \mid A)$? **No**
$\boldsymbol{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A, B)$? **Yes**

EXAMPLE CONTD.



- Deciding conditional independence is hard in noncausal directions
-
- (Causal models and conditional independence seem hardwired for humans!)
-
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed
-

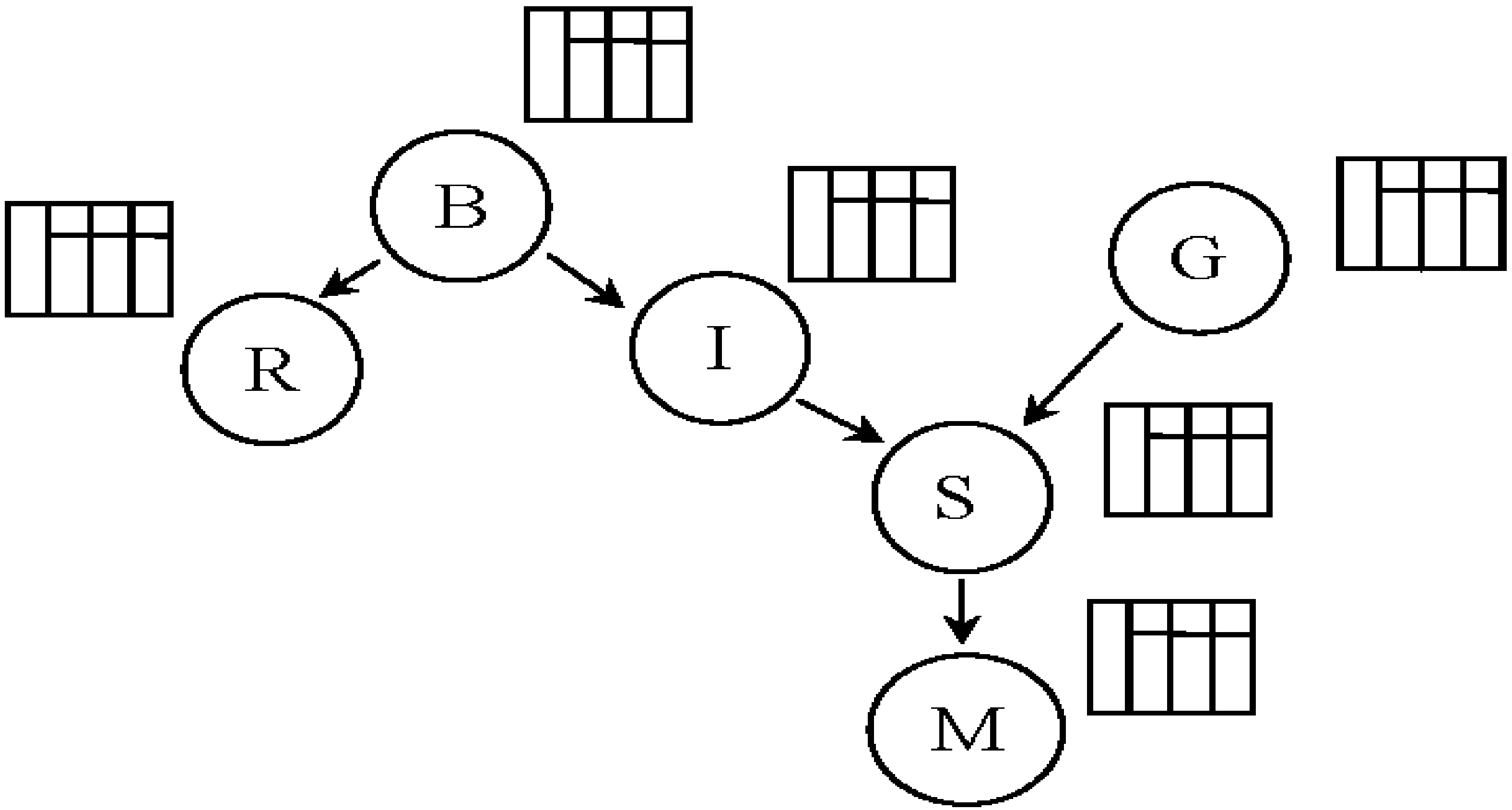
תזכורת

- בהינתן רשת בייסיאנית, ניתן בקלות לדעת את יחסי האי תלות המותנה המוצגים בה.
-
- כל משתנה מקרי (צומת) Xבגרף (ברשת) איננו תלוי באופן מותנה בכל המשתנים שאינם צאצאיו של Xבהינתן הוריו של X.
- למשל, בדוגמה 3 יש אי-תלות מותנה בין H לבין B,O ו-L בהינתן D
- ולכן $P(H \mid B,D,O,L) = P(H \mid D)$

דוגמת אבחון תקלות רכב:

- משתנים:
- Radio – הרדיו פועל?
- Battery – המצבר תקין?
- Ignition – הצתה?
- Starts – המנוע מניע?
- Gas– יש דלק?
- Moves – הרכב זז?

דוגמת אבחון תקלות רכב:



דוגמאות לטבלאות CPT:

P(G)
0.9

I	G	P(S)
false	true	0
true	false	0
true	true	0.95
false	false	0

דוגמת – מישהו בבית?

- כאשר אני מגיע הביתה, אני רוצה לדעת האם מישהו מבני משפחתי נמצא בבית לפני שאני נכנס.
- נתון המידע שלהלן:

1. כאשר אשתי עוזבת את הבית, היא מדליקה את האור מחוץ לבית לעתים קרובות (אך לא תמיד).(כמו כן, היא מדליקה לעתים את האור כשאמור להגיע אורח).
2. כאשר אין איש בבית, משאירים את הכלב לעתים קרובות בחוץ.
3. גם כאשר לכלב יש בעיות מעיים, משאירים אותו לעתים קרובות בחוץ.
4. אם הכלב בחוץ, יתכן ואשמע אותו נובח (למרות שיתכן שאינו נובח, או שאני עשוי לשמוע נביחות של כלב אחר ואחשוב שזהו הכלב שלי).

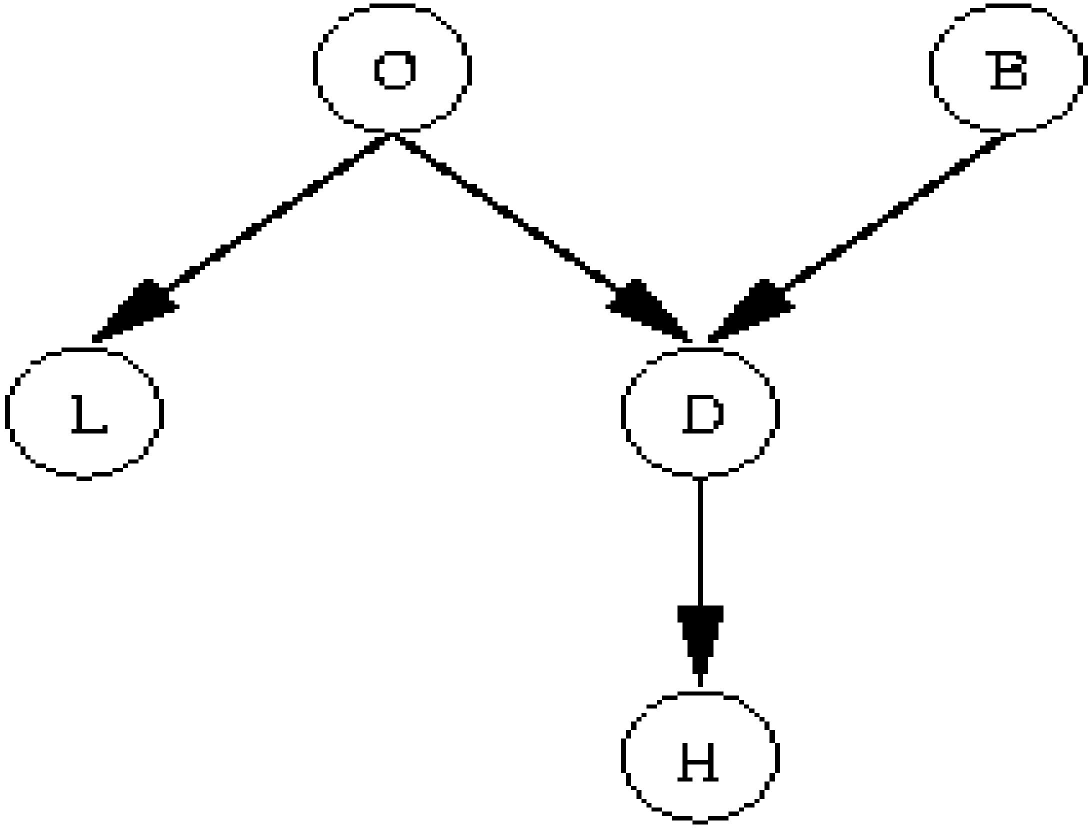
שלב 1 – המשתנים

- נגדיר 5 משתנים מקריים (בוליאניים):
  1. O: כולם מחוץ לבית
  2. L: האור דולק
  3. D: הכלב בחוץ
  4. B: לכלב יש בעיות מעיים
  5. H: אני יכול לשמוע את הכלב נובח

שלב 2: ההשפעות

- ננסח כעת את ההשפעות הסיבתיות הישירות:
  1. H מושפע ישירות רק מ-D. לכן בהינתן D, יש אי תלות מותנה בין H לבין O,L, ו-B.
  2. D מושפע ישירות רק מ-O ומ-B. לכן בהינתן O ו-B, יש אי תלות מותנה בין D לבין L.
  3. L מושפע ישירות רק מ-O. לכן בהינתן O, יש אי תלות מותנה בין L לבין H,D, ו-B.
  4. O ו-B הם בלתי תלויים.

שלב 3: הרשת



CPT - שלב 4: הטבלאות

P(B)
0.3

P(O)
0.6

D	P(H)
true	0.3
false	0.8

O	P(L)
true	0.3
false	0.6

O	B	P(D)
true	true	0.5
true	false	0.1
false	true	0.1
false	false	0.2

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct