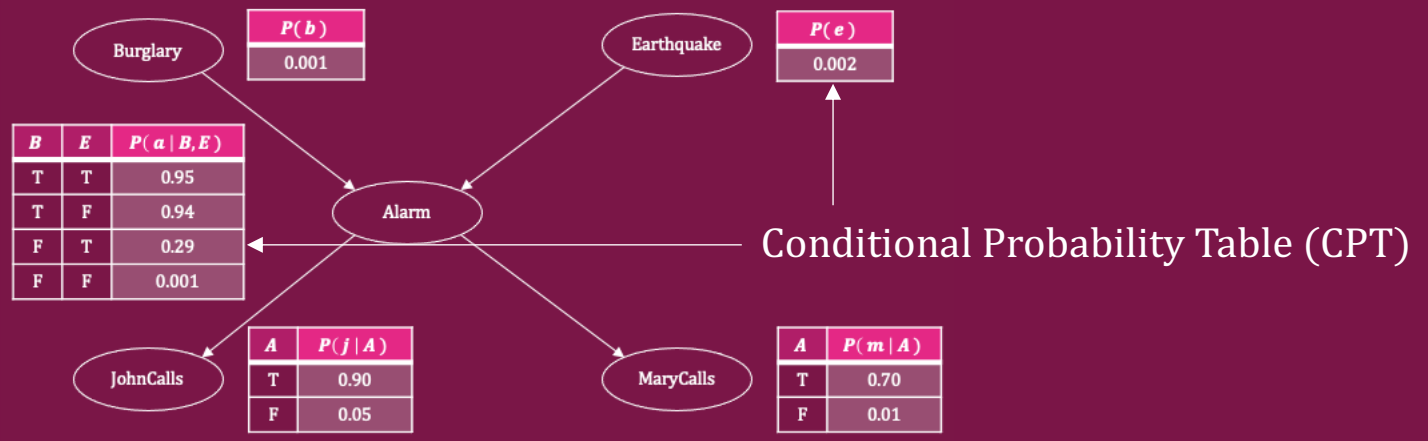
Bayes Nets

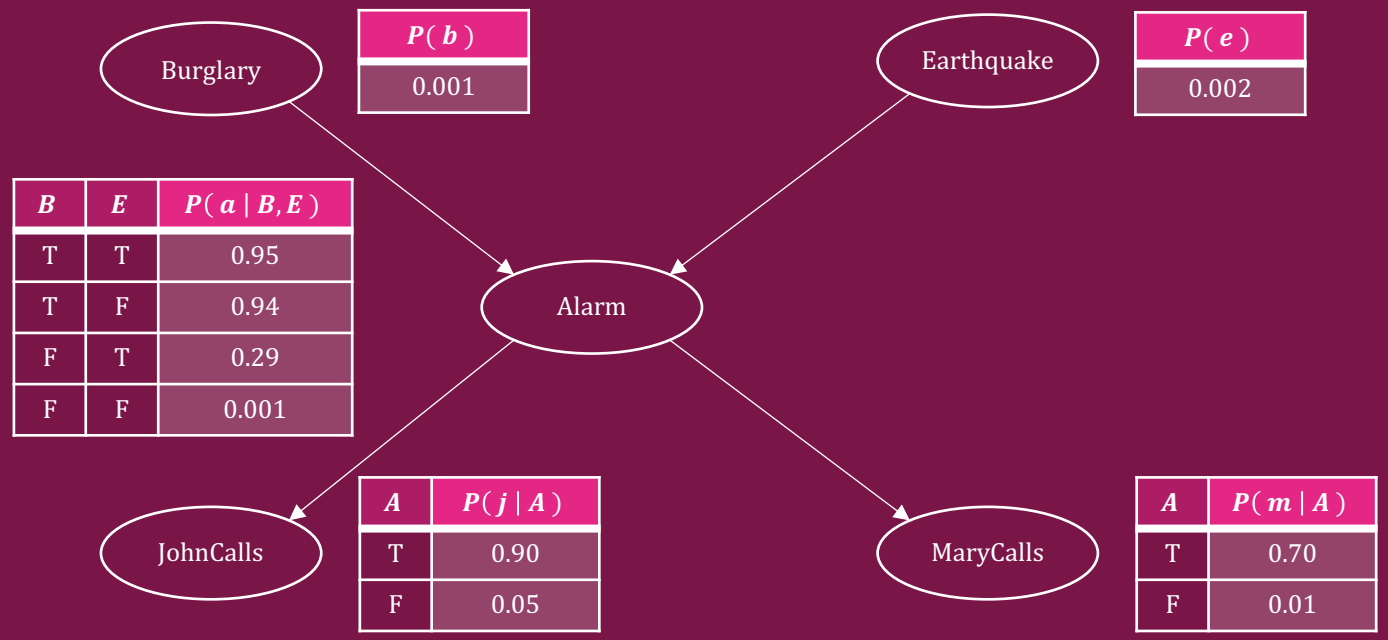
MSAI 348: Intro to Artificial Intelligence
Instructor: Mohammed A. Alam

Northwestern University

A Bayes Net (BN), short for "Bayesian Network" (coined by Judea Pearl, who researches causality), is a Directed Acyclic Graph (DAG) that uses the Chain Rule to represent the Joint Probability Distribution (JPD) over some random variables compactly—and thus efficiently—enabling easy inference.



Instead of writing the entire Joint Probability Table (JPT), we write the Conditional Probability Table (CPT) for every variable given only its ancestors – *variables it depends on*.

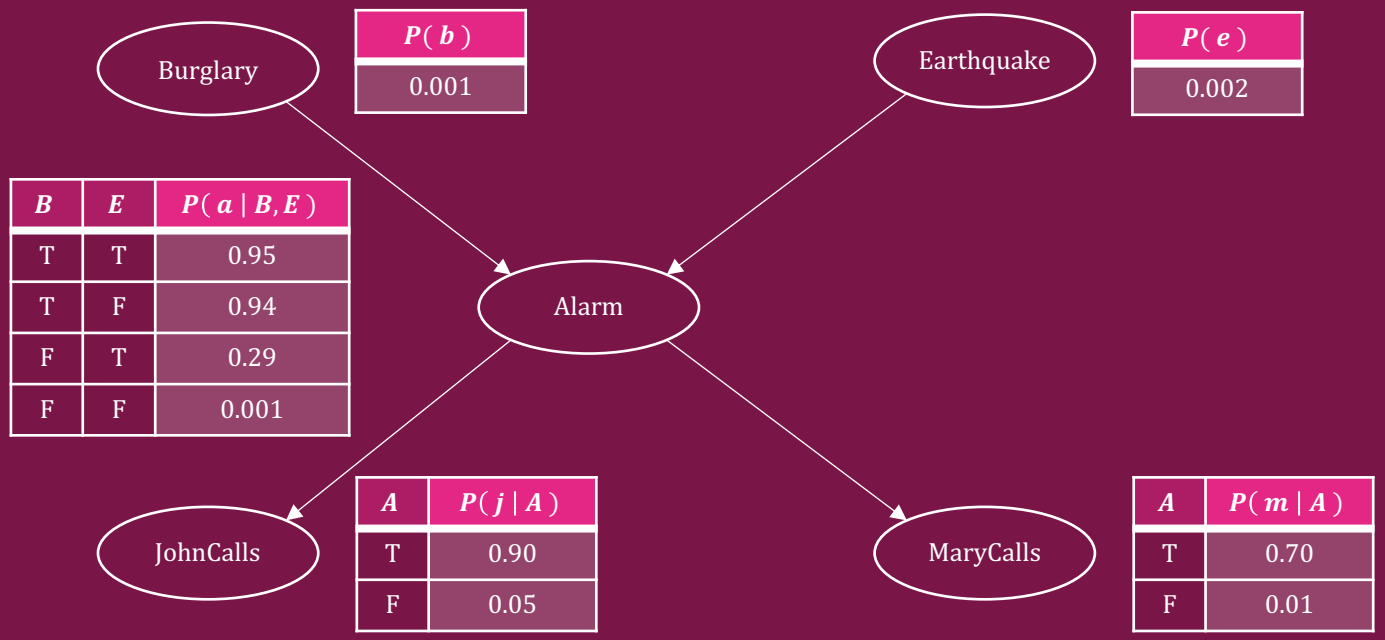


The BN structure must reflect appropriate causal/semantic intuitions as relationships.

In other words, a BN must be built, often manually, in a reasonable way.

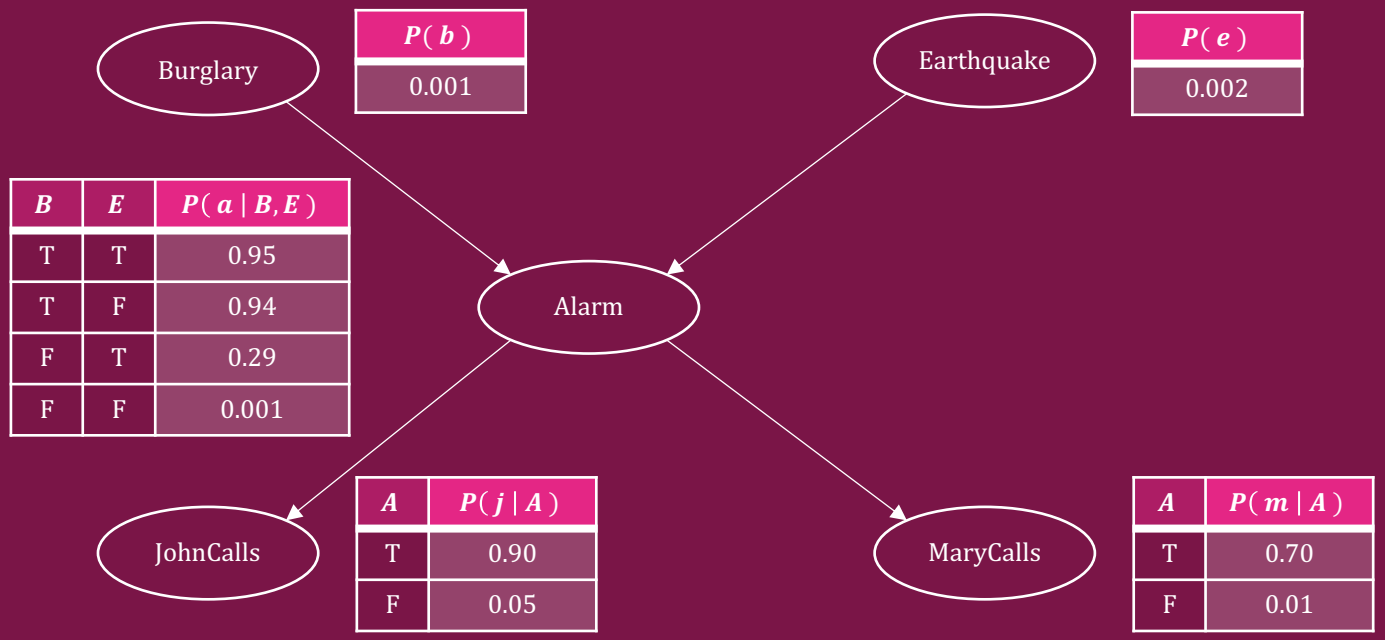
BNs can also be figured out from data, but here, we'll only use pre-built ones to calculate probabilities.

Shown above is the classic "burglary example."



Every random variable is a node in the DAG.

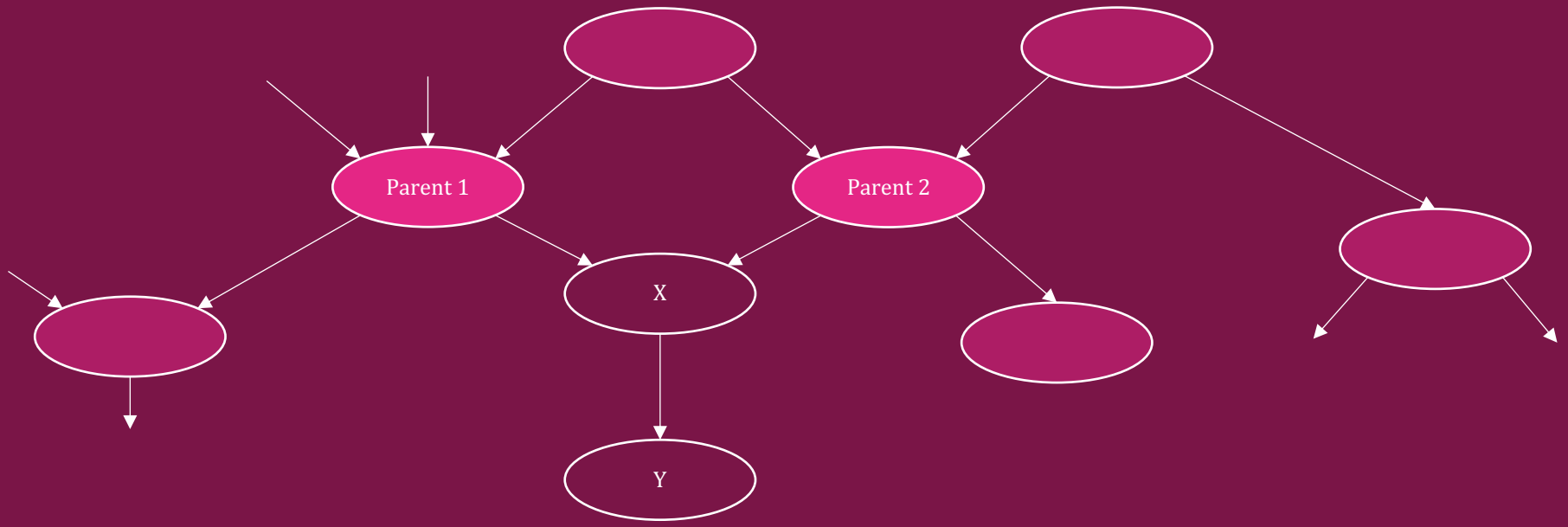
Causal relationship is represented as dependency, and every dependency is an edge.



Each node is conditionally independent of its non-descendants given its parents.

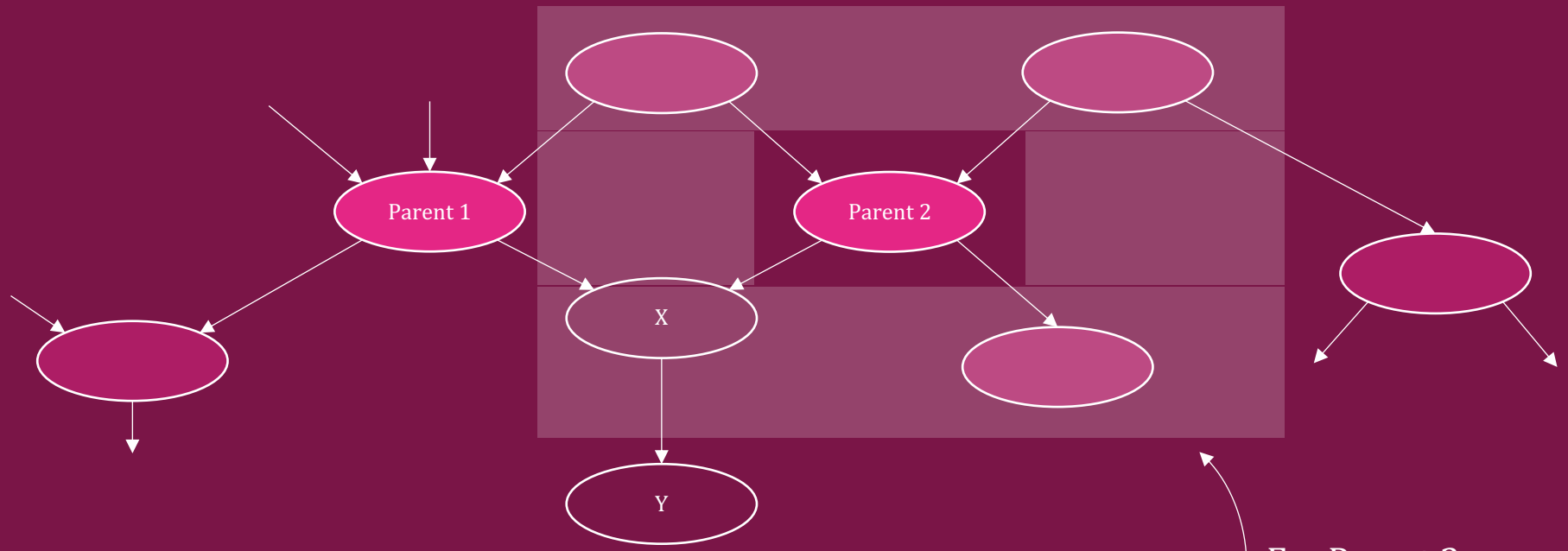
Here, JohnCalls is independent of Burglary, Earthquake, and MaryCalls given Alarm.

This constitutes local semantics for BNs.



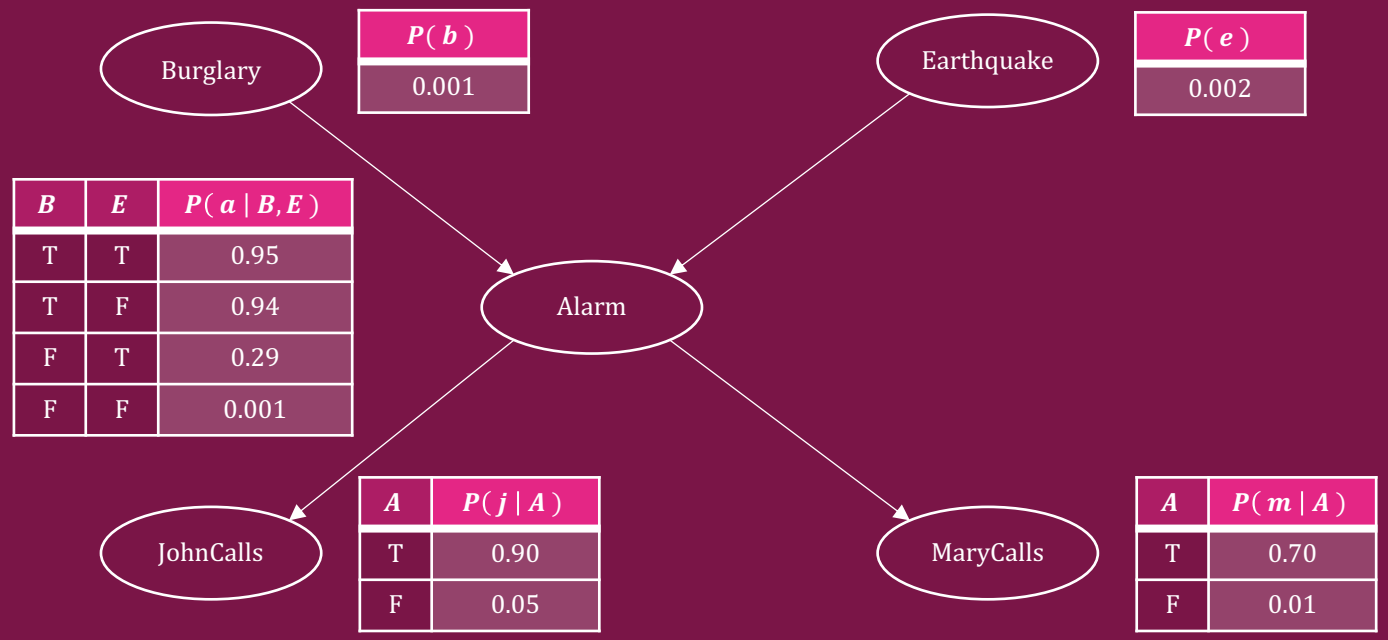
Each node is conditionally independent of its non-descendants given its parents.

Here, given Parent 1 and Parent 2, X is independent of any node that is not Y.



Each node is conditionally independent of all others given its Markov Blanket.

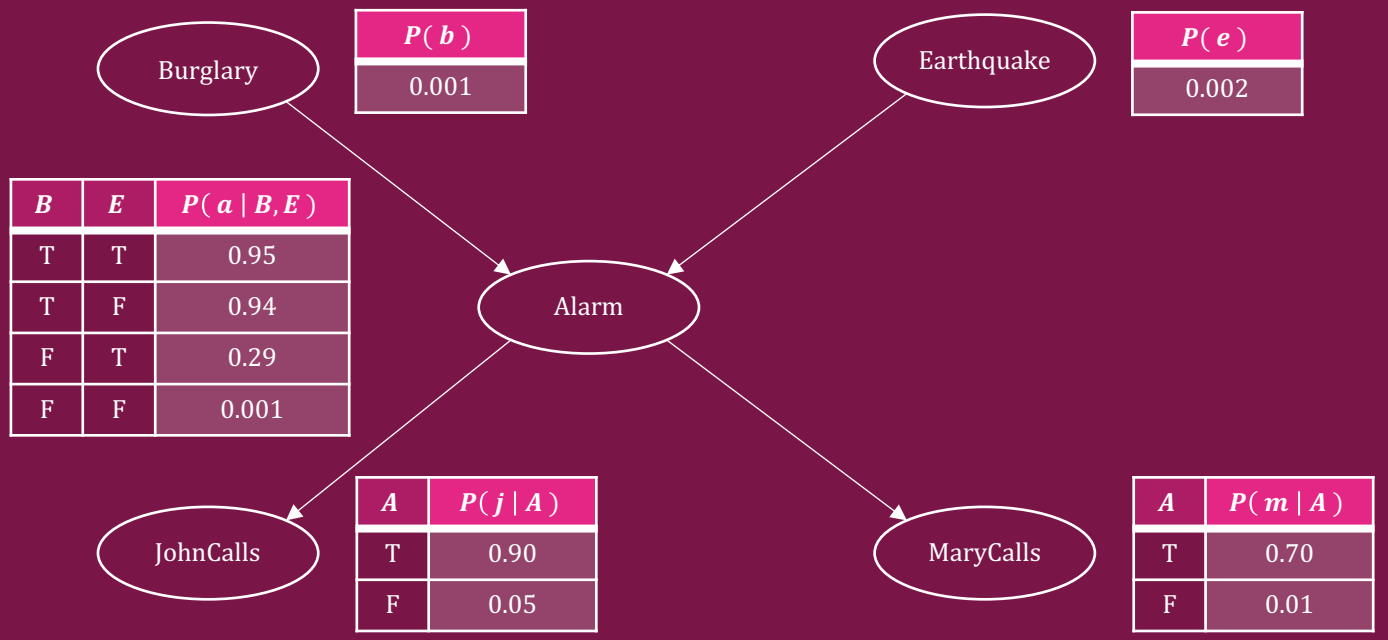
A node's Markov Blanket: The node's parents, children, and children's parents.



An event's probability can be computed by multiplying relevant probabilities in the CPTs.

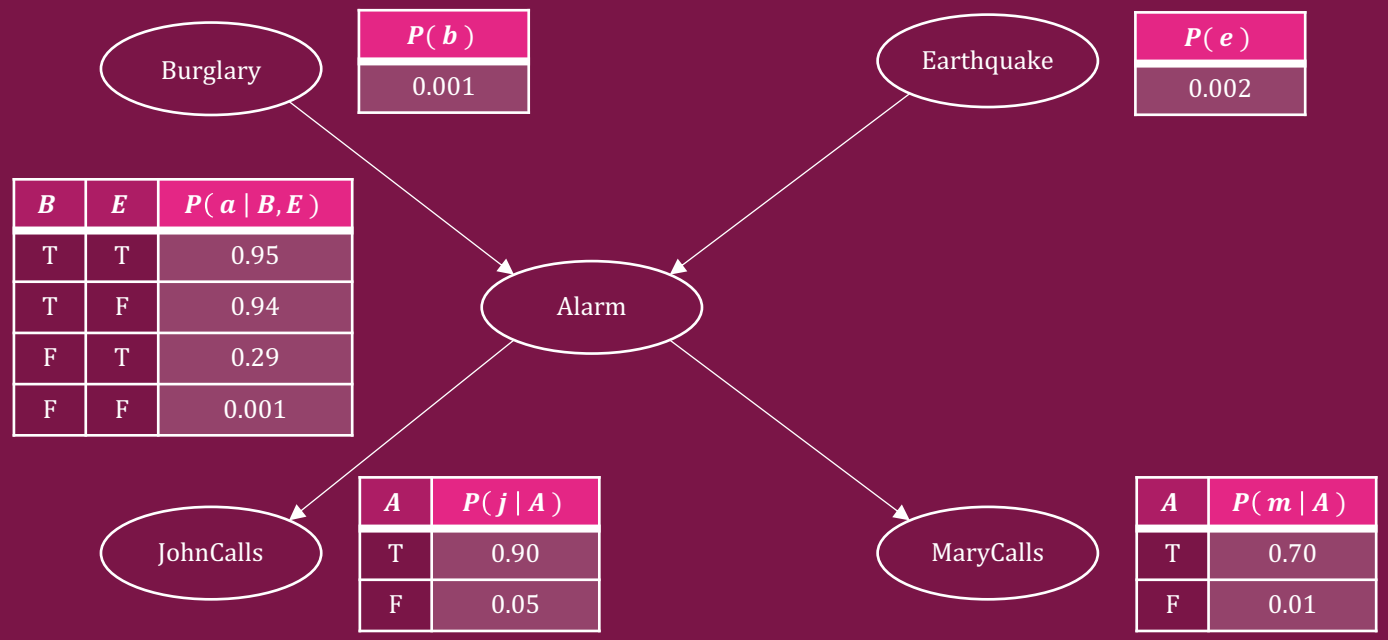
And conditional independencies help simplify the probability tables needed.

This constitutes global semantics for BNs: $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(x_i))$



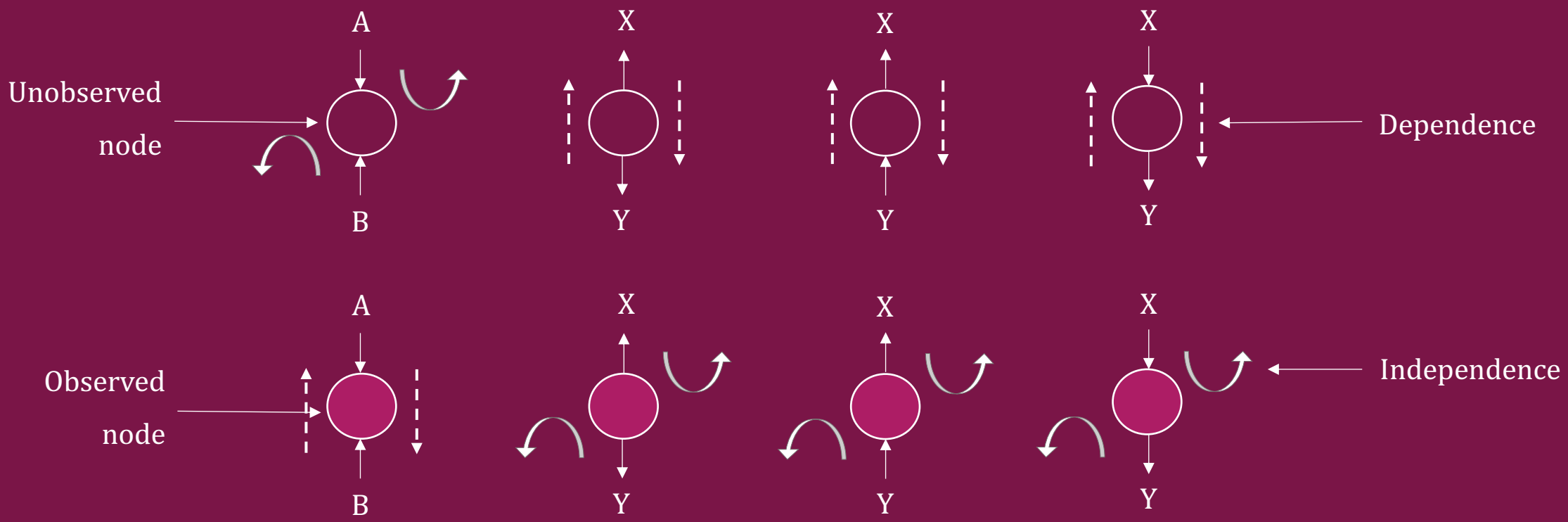
To create a Bayes Net:

1. Create a node for each important variable in domain.
2. Create causal edges using domain knowledge or by learning from data.
3. Obtain CPTs using domain knowledge or from data.



Note something implicit: We know more than the CPTs tell us explicitly.

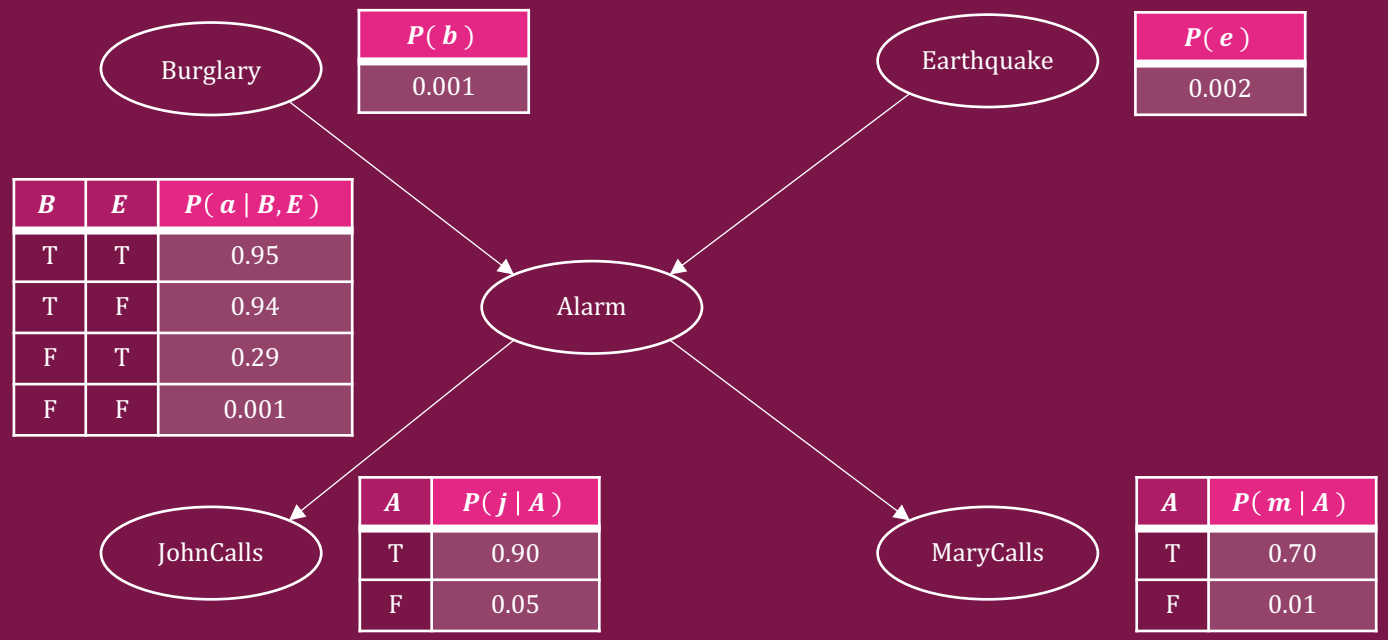
E.g., we know that $P(\neg a \mid B = T, E = T) = 0.05$ and $P(\neg m \mid A = F) = 0.99$.



The unobserved node between A and B makes them independent due to D-separation.

Observing the node between A and B makes them dependent.

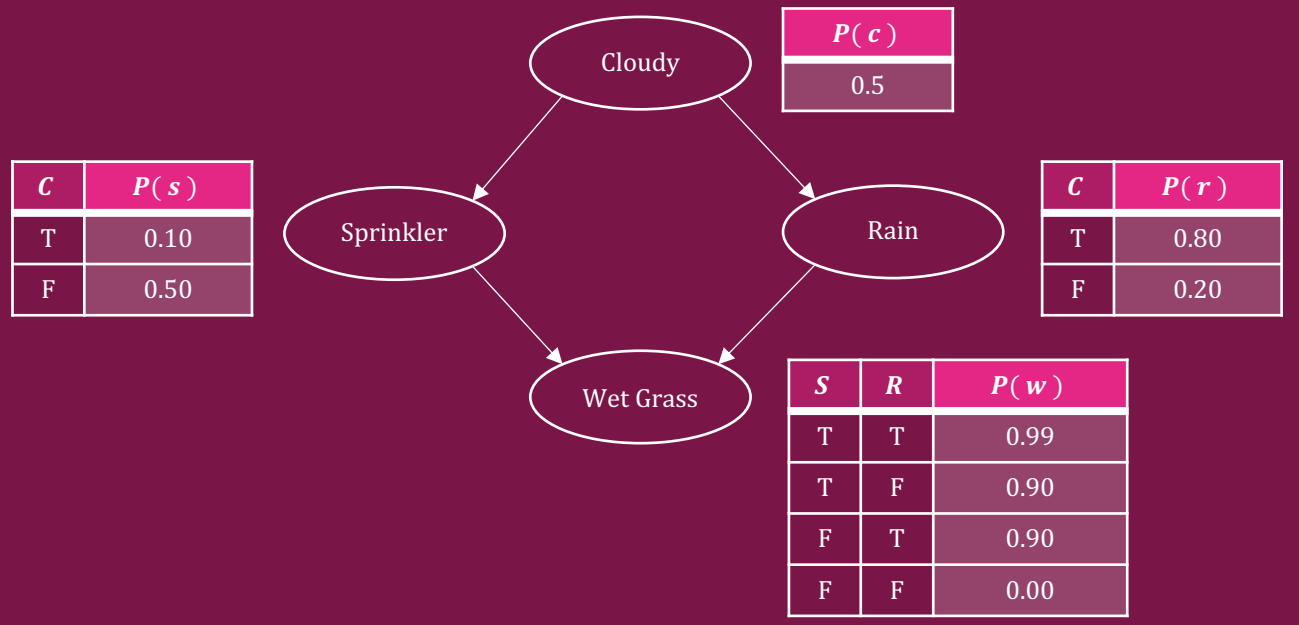
The opposite is true for X and Y. And this diagram is known as Bayes Ball, developed by Ross Shachter.



Alarm (A) is explained by both Burglary (B) and Earthquake (E).

If E is true, we don't need B to explain A, so our belief in B decreases.

This is called "explaining away." Knowing E explains B away.

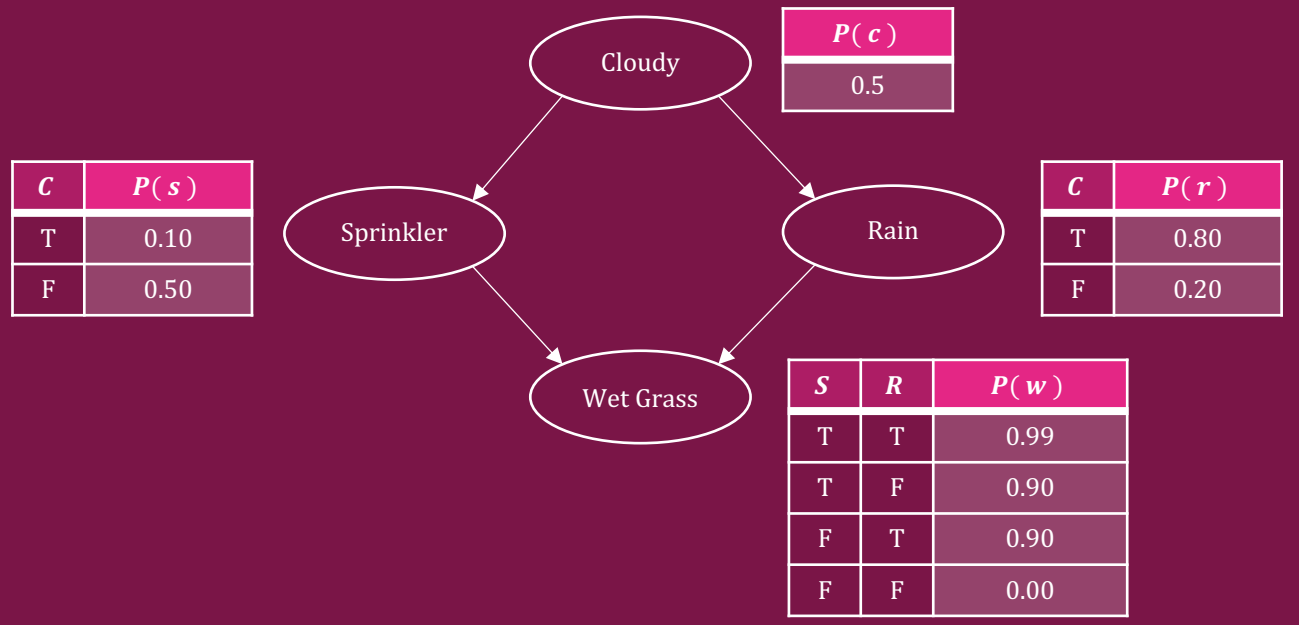


Using the Chain Rule, the JPD of all nodes in the "sprinkler example" above is:

$$P(C,S,R,W) = P(C)\,P(S|C)\,P(R|C,S)\,P(W|C,S,R)$$

$$= P(C)\,P(S|C)\,P(R|C)\,P(W|S,R) \longleftarrow \text{using conditional independence (Bayes Ball)}$$

Notice the reduction in parameter count.

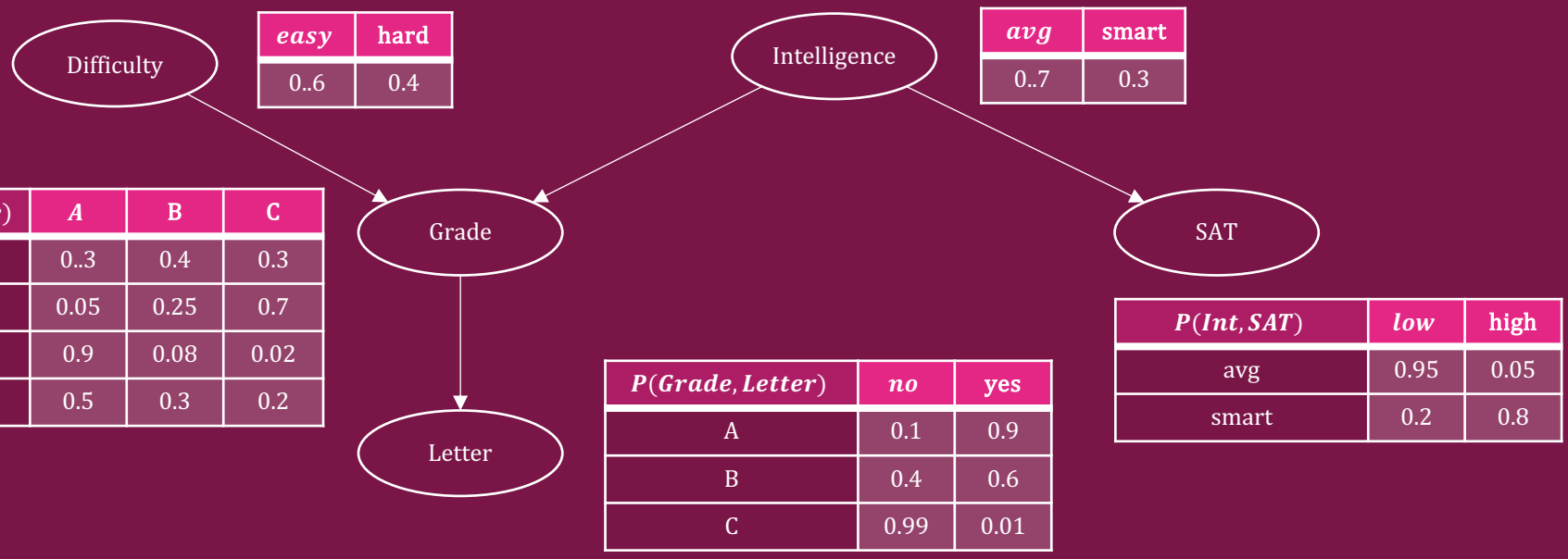


Notice also how we are left with only the probability of every variable given its parents.

$$P(C,S,R,W) = P(C)\,P(S|C)\,P(R|C,S)\,P(W|C,S,R)$$

$$= P(C)\,P(S|C)\,P(R|C)\,P(W|S,R) \longleftarrow \text{using conditional independence (Bayes Ball)}$$

So, the breakdown using Chain Rule is easy: Write the probability of every variable given its parents.



easy	hard
0..6	0.4

avg	smart
0..7	0.3

$P(Int, Diff, Grade)$	*A*	B	C
avg., easy	0..3	0.4	0.3
avg., hard	0.05	0.25	0.7
smart, easy	0.9	0.08	0.02
smart, hard	0.5	0.3	0.2

$P(Grade, Letter)$	*no*	yes
A	0.1	0.9
B	0.4	0.6
C	0.99	0.01

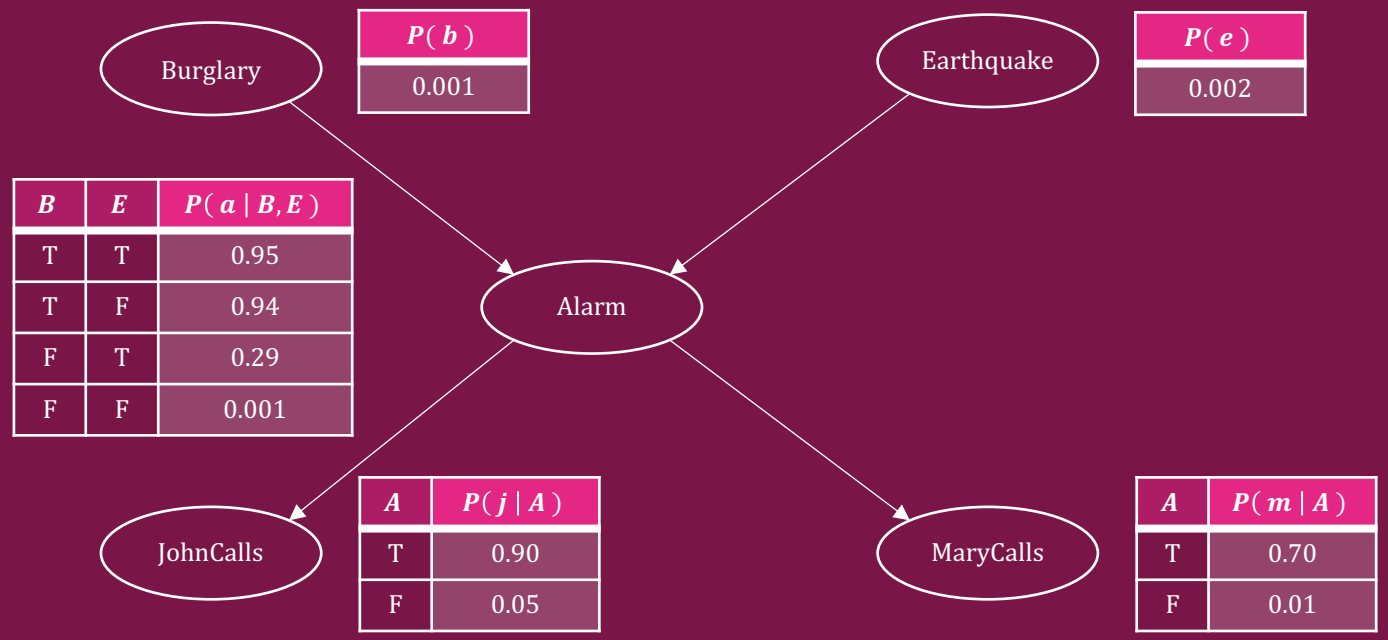
$P(Int, SAT)$	*low*	high
avg	0.95	0.05
smart	0.2	0.8

Ques 1: *What's the P(someone smart scoring a B on an easy exam and high on the SAT gets no letter)?*

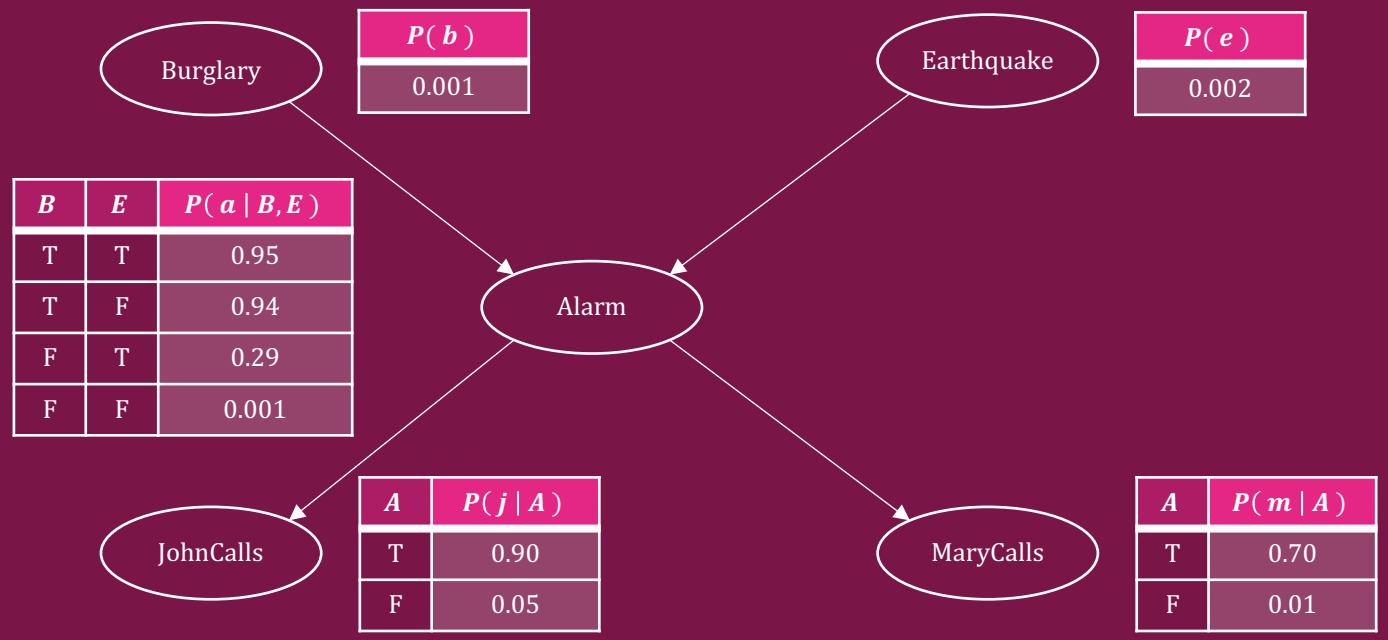
$$P(smart, easy, B, high, no)$$

$$= P(smart)\ P(easy)\ P(B \mid smart, easy)\ P(high \mid smart)\ P(no \mid B)$$

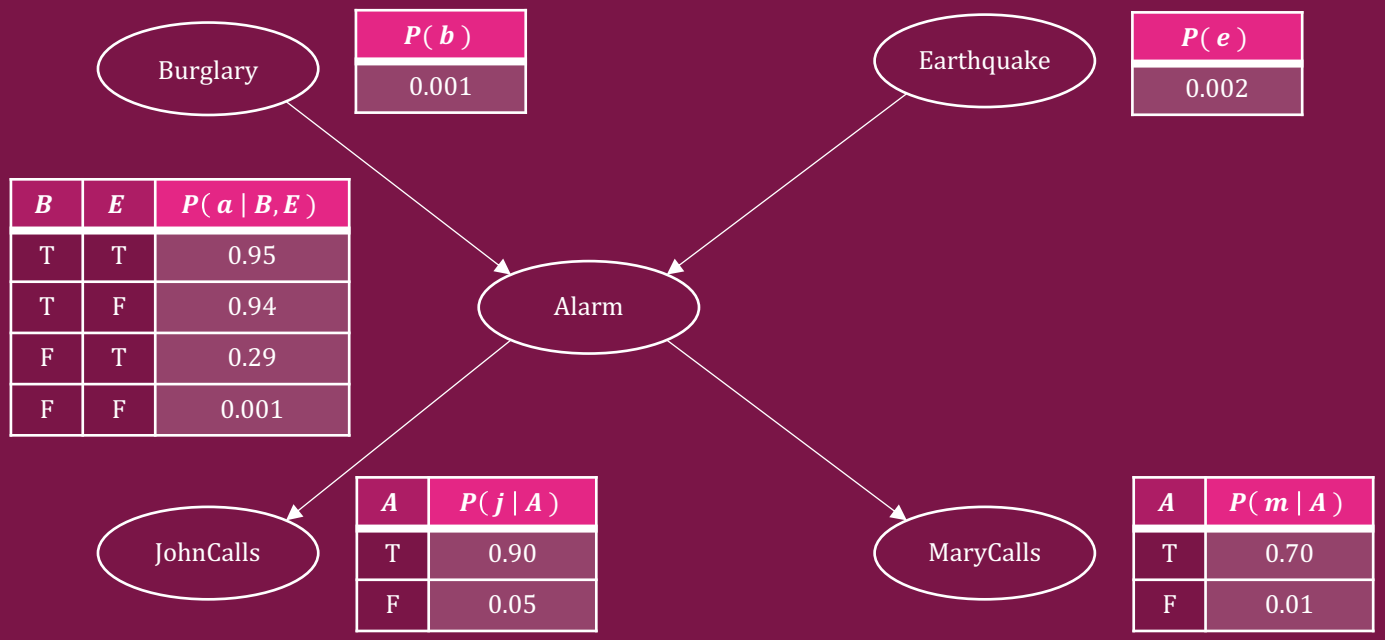
$$= 0.3 \times 0.6 \times 0.08 \times 0.8 \times 0.4 = 0.004608$$



Ques 2: *What's the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, yet both John and Mary have called?*



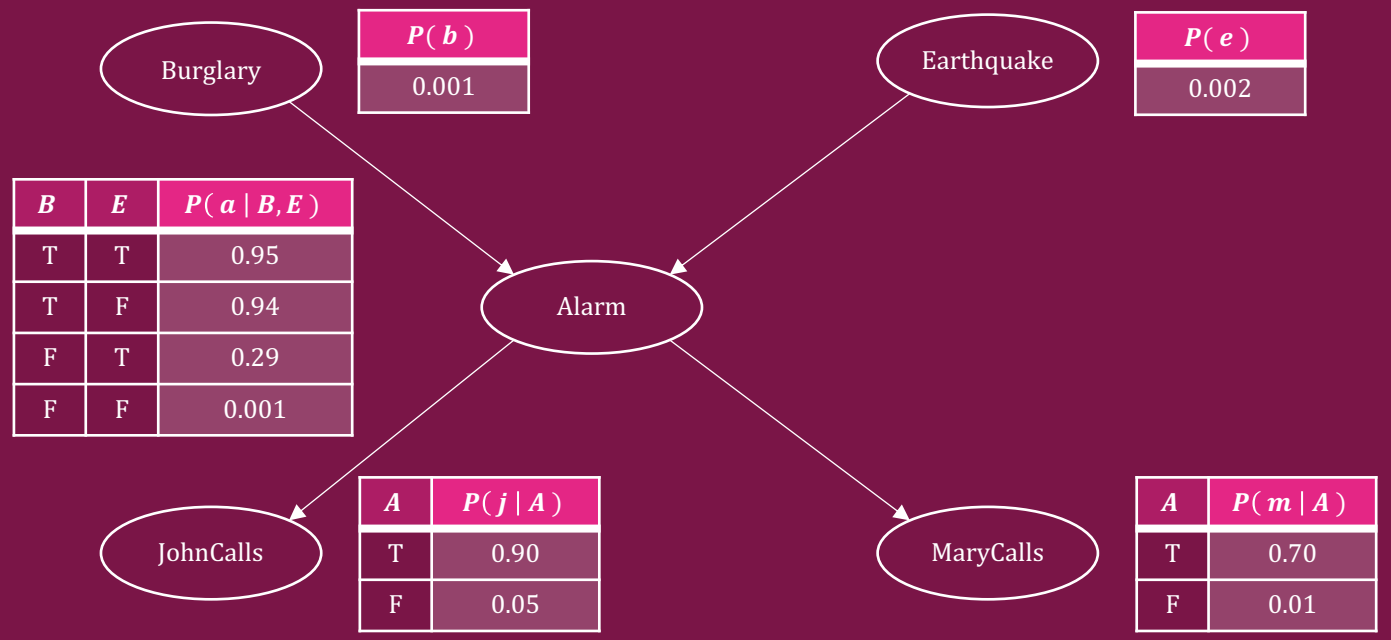
Let's denote the event that Mary calls despite the alarm not having sounded as $m \mid \neg a$, which means the probability of that event is $P(m \mid \neg a) = 0.01$.



$P(a, \neg b, \neg e, j, m)$ ⟵ what we want to answer

$= P(a \mid \neg b, \neg e)\, P(\neg b)\, P(\neg e)\, P(j \mid a)\, P(m \mid a)$ ⟵ the probability of every variable given its parents

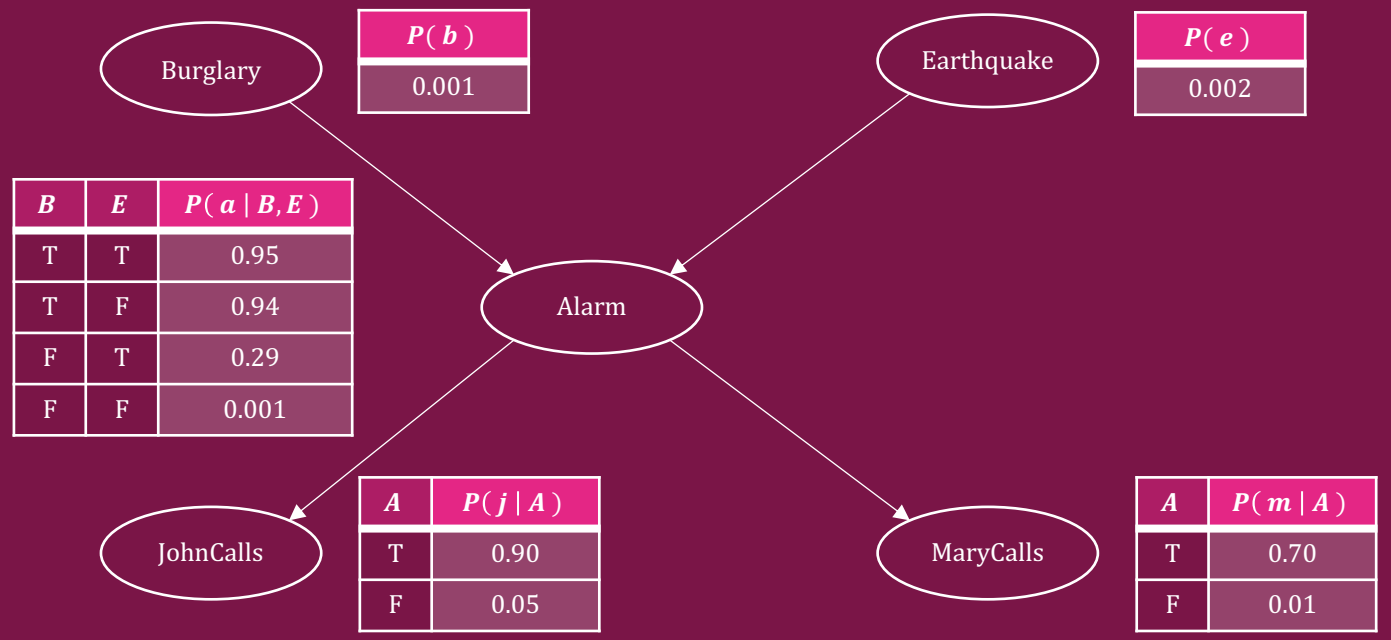
$= 0.001 \times 0.999 \times 0.998 \times 0.9 \times 0.7 = 0.00062$



Ques 3: *What's the probability that John will call?*

Note that we know nothing about burglary, earthquake, alarm, or a call from Mary.

Also note that this is the marginal probability of the event that John calls.

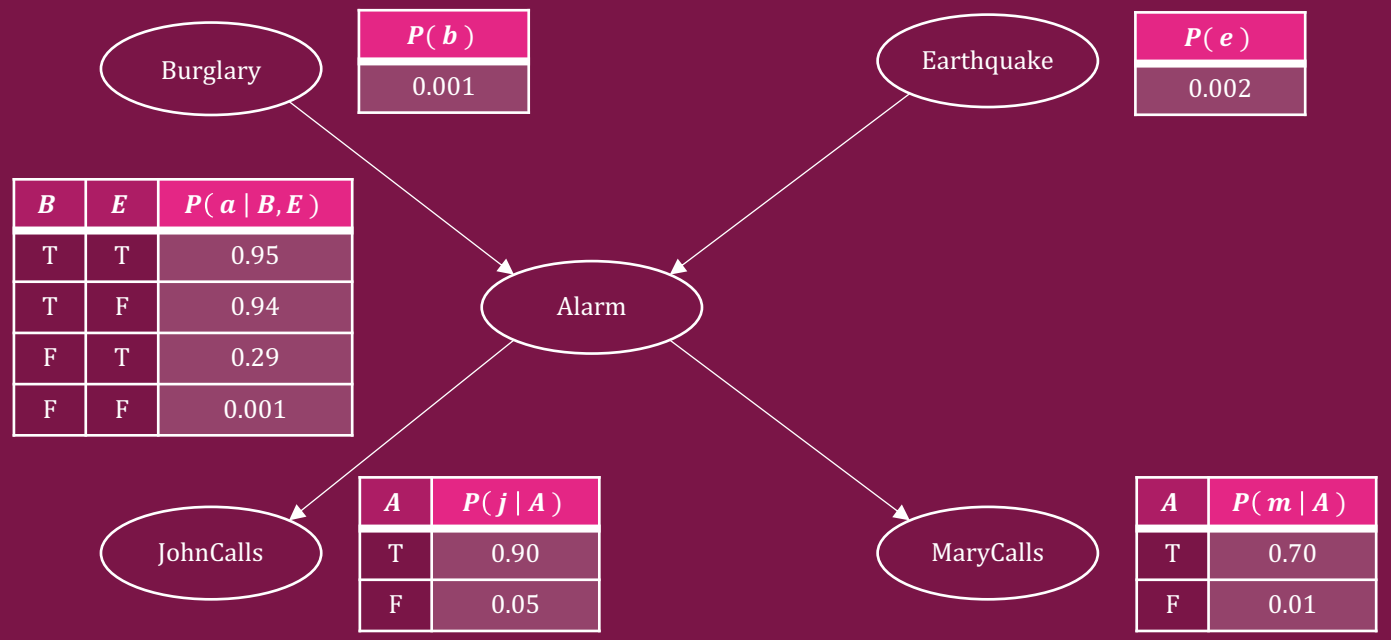


What we want to know: $P(j)$.

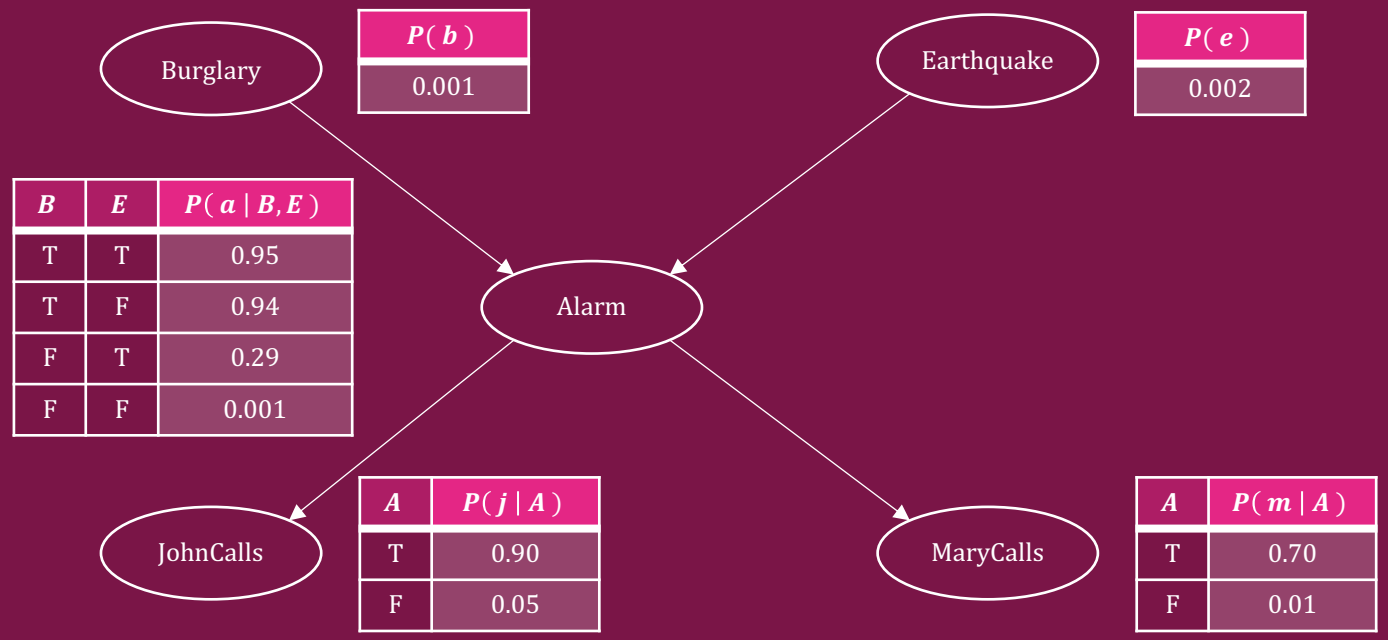
We don't have much to go off of. So, we bring in j's parent, A:

$$P(j) = \sum_A P(j, A)$$

The right-hand side in the above equation is expressing the left-hand side summed over A.



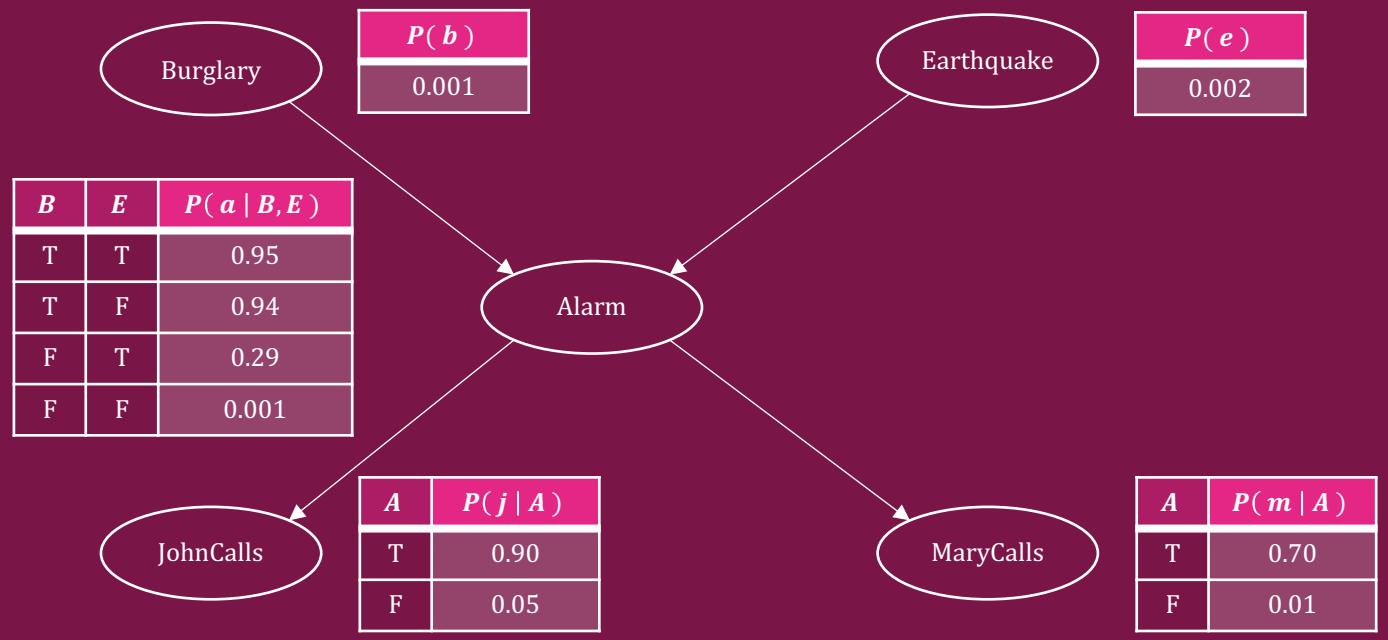
In other words, we wrote the joint probability of the event that John calls and the event that the alarm sounds or doesn't. We considered both possibilities for the alarm, hence marginalization. Why do that? Because that might let us use the CPTs that we have, so we can use the probabilities in them.



$$P(j) = \sum_A P(j, A) = \sum_A P(j \mid A)\, P(A) = P(j \mid a)\, P(a) + P(j \mid \neg a)\, P(\neg a)$$

We have a problem, though: We don't know $P(a)$ and $P(\neg a)$, both marginal probabilities.

We need to go one step further and marginalize A over B and E.



$$P(j) = \sum_A P(j, A) = \sum_A P(j \mid A)\, P(A) = \cancel{P(j \mid a)\, P(a) + P(j \mid \neg a)\, P(\neg a)}$$

$$= \sum_A P(j \mid A)\, P(A) = \sum_A [P(j \mid A) \sum_{B,E} P(A, B, E)]$$

$$P(j) = \sum_A [P(j \mid A) \sum_{B,E} P(A,B,E)]$$

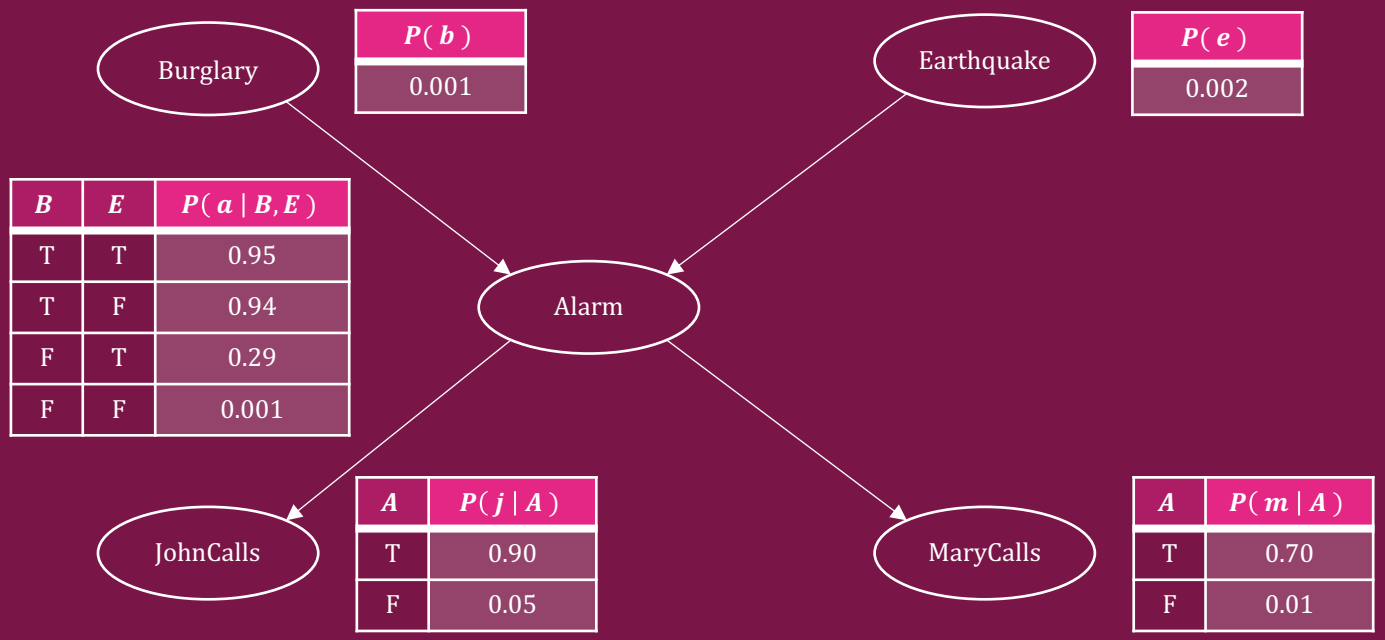
$$= P(j \mid a) \sum_{B,E} P(a,B,E) + P(j \mid \neg a) \sum_{B,E} P(\neg a,B,E)$$

$$= P(j \mid a) \sum_{B,E} P(a \mid B,E)\, P(B,E) + P(j \mid \neg a) \sum_{B,E} P(\neg a \mid B,E)\, P(B,E)$$

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

$$+ P(j \mid \neg a)\, \{P(\neg a \mid b,e)\, P(b,e) + P(\neg a \mid \neg b,e)\, P(\neg b,e) + P(\neg a \mid b,\neg e)\, P(b,\neg e) + P(\neg a \mid \neg b,\neg e)\, P(\neg b,\neg e)\}$$

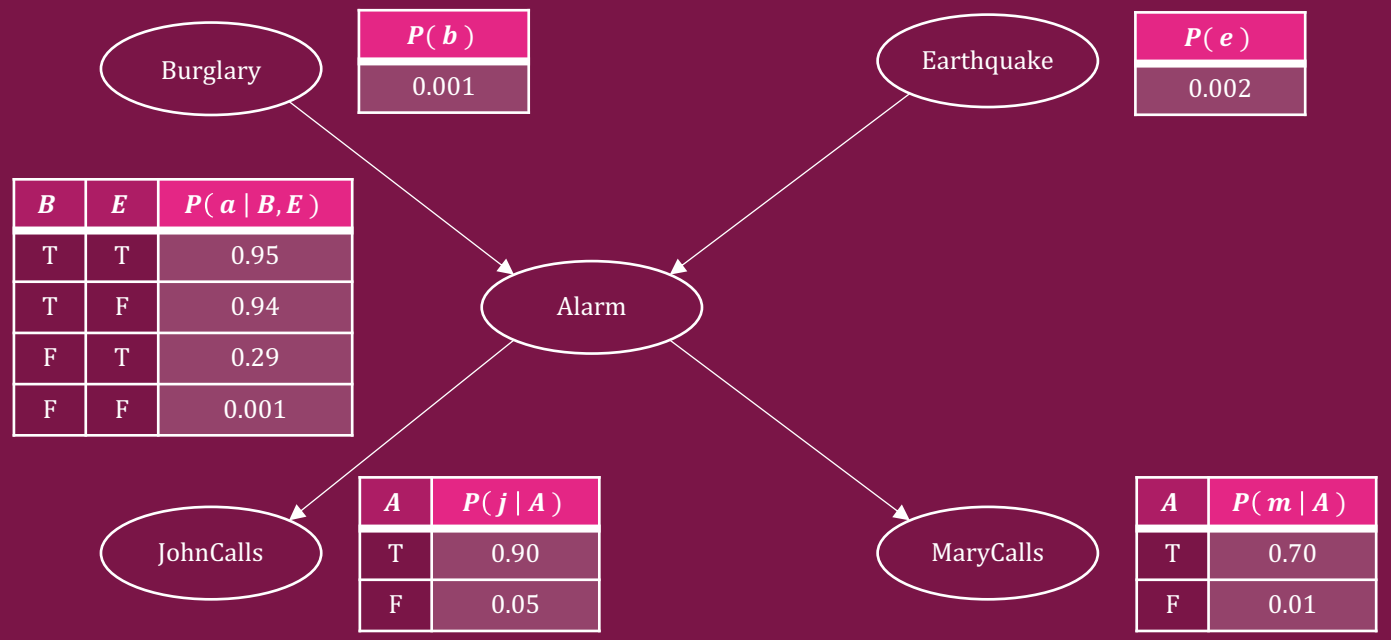
$$= 0.9 \times 0.00252 + 0.05 \times 0.9974 = 0.0521$$



Now for a really useful question:

Ques 4: *What's the probability that there's a burglary given that both John and Mary have called?*

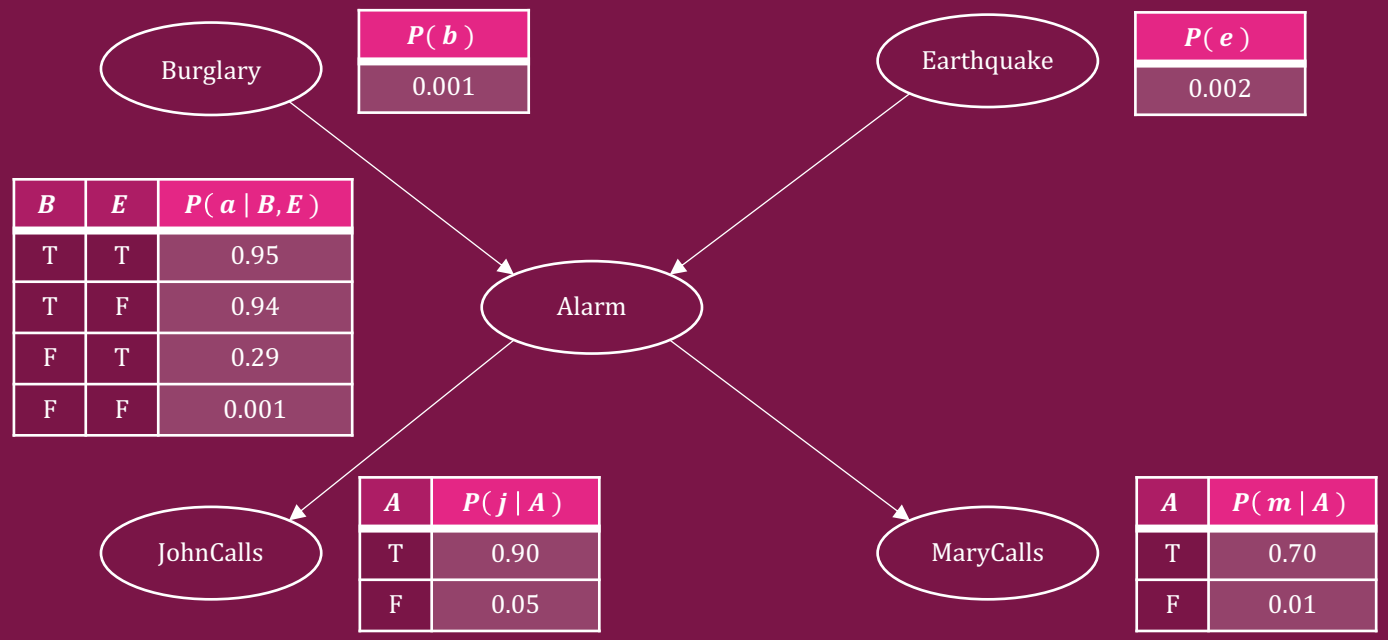
So, $P(b \mid j, m)$?



In $P(b \mid j, m)$, b is the model and j, m are the data.

Using Bayes' Theorem, $P(b \mid j, m) = \frac{P(j, m \mid b)\, P(b)}{P(j, m)} = \frac{P(b, j, m)}{P(j, m)}$

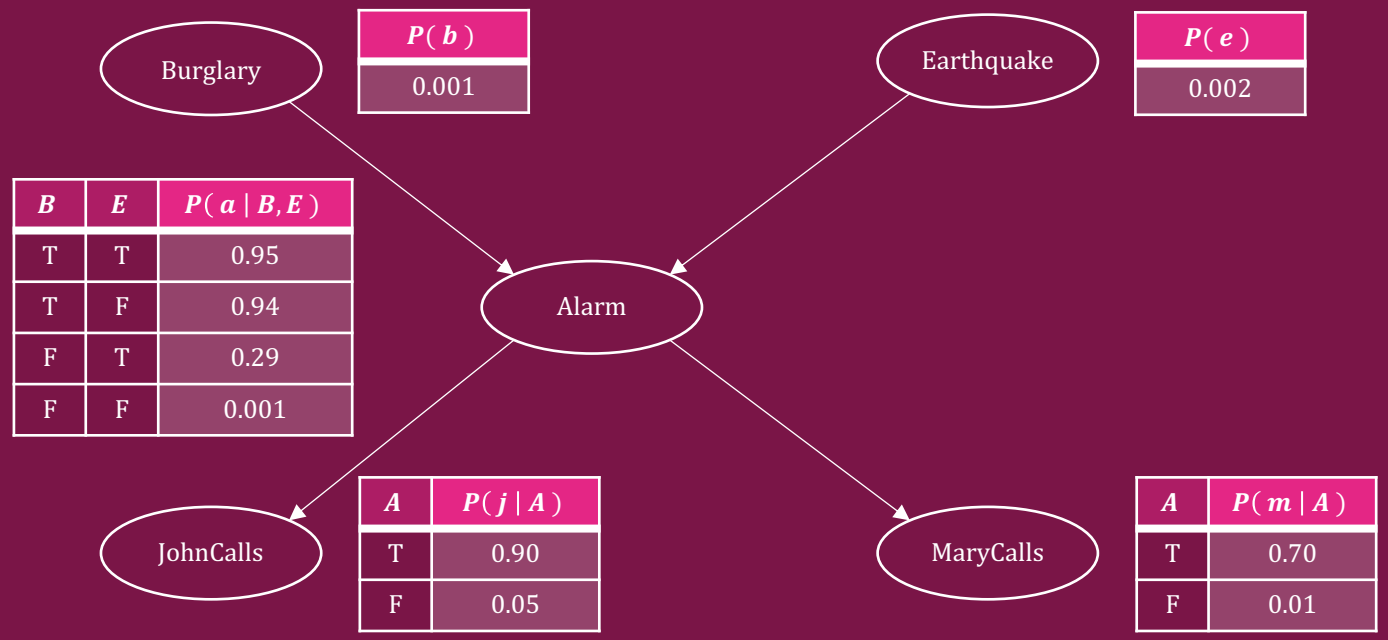
We'll compute $P(b, j, m)$ and leave $P(j, m)$ (the normalization constant) for later.



$P(b, j, m) = P(b)\, P(j \mid A)\, P(m \mid A)$.

However, we don't know which value of A to consider: a or $\neg a$.

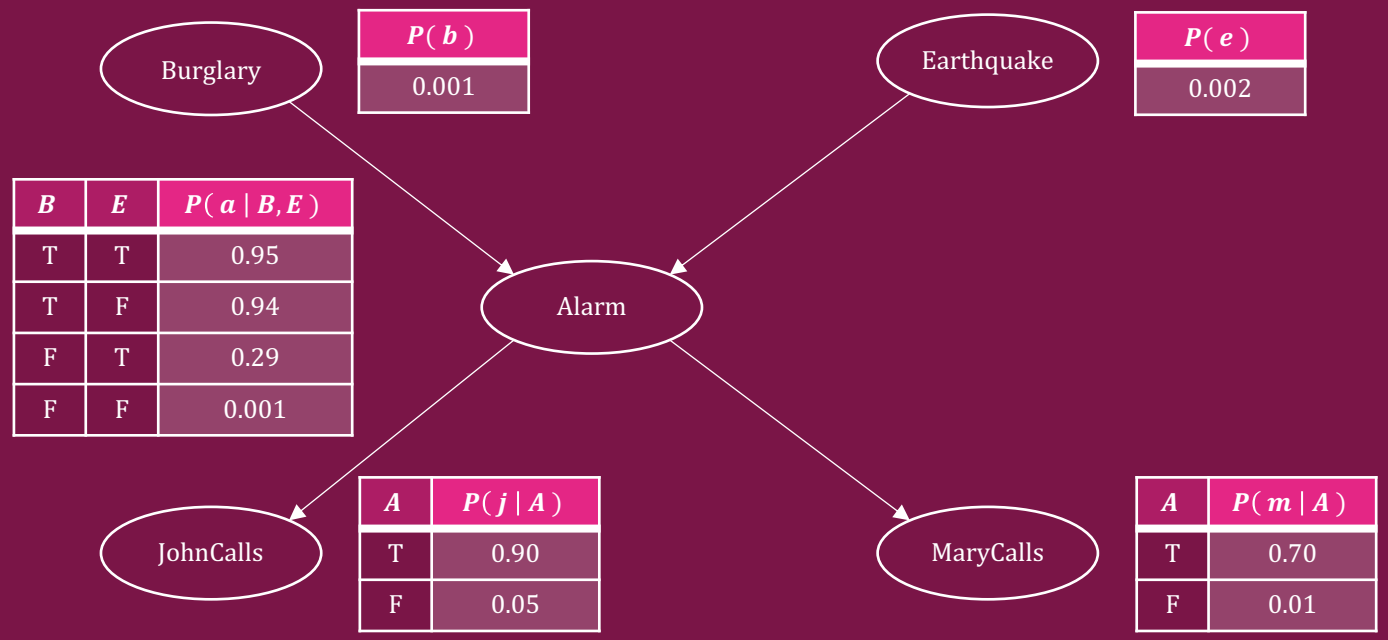
Let's introduce A in hopes of getting more information.



$$\sum_A P(b, j, m, A) = \sum_A [P(b)\ P(j \mid A)\ P(m \mid A)\ P(A \mid b, E)]$$

However, we don't know which value of E to consider: e or $\neg e$ either.

Let's introduce E in hopes of getting more information.



$$\sum_A \sum_E P(b, j, m, A, E) = \sum_A \sum_E [P(b)\, P(j \mid A)\, P(m \mid A)\, P(A \mid b, E)\, P(E)]$$

$$= P(b) \sum_A \sum_E [P(j \mid A)\, P(m \mid A)\, P(A \mid b, E)\, P(E)]$$

$$\sum_A \sum_E P(b, j, m, A, E) = P(b) \sum_A \sum_E [P(j \mid A)\, P(m \mid A)\, P(A \mid b, E)\, P(E)]$$

$$= P(b) \sum_A [P(j \mid A)\, P(m \mid A) \sum_E P(A \mid b, E)\, P(E)]$$

$$= P(b) \sum_A [P(j \mid A)\, P(m \mid A)\, \{P(A \mid b, e)\, P(e) + P(A \mid b, \neg e)\, P(\neg e)\}]$$

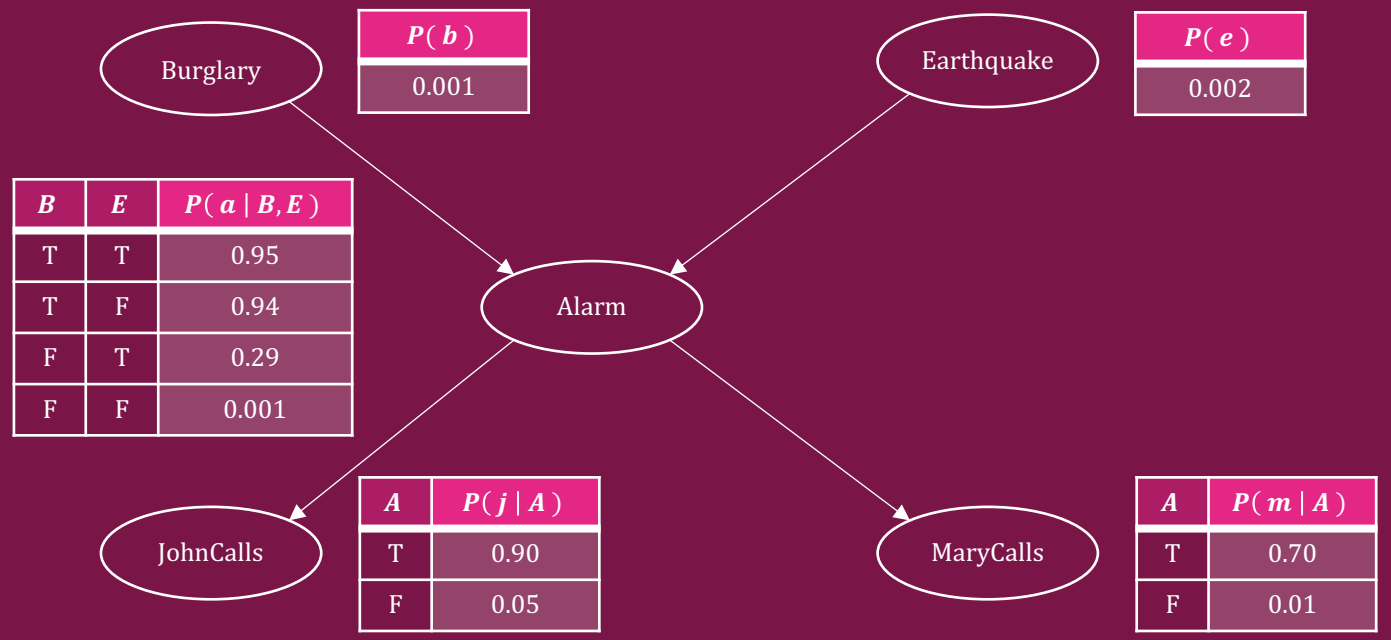
$$= P(b)\, [P(j \mid a)\, P(m \mid a)\, \{P(A \mid b, e)\, P(e) + P(A \mid b, \neg e)\, P(\neg e)\} +$$

$$P(j \mid \neg a)\, P(m \mid \neg a) \{P(\neg a \mid b, e)\, P(e) + P(\neg a \mid b, \neg e)\, P(\neg e)\}]$$

$$= 0.001\, [0.90 \times 0.70\, \{0.95 \times 0.002 + 0.94 \times 0.998\} +$$

$$0.05 \times 0.01\, \{0.05 \times 0.002 + 0.06 \times 0.998\}]$$

$$= 0.00059$$

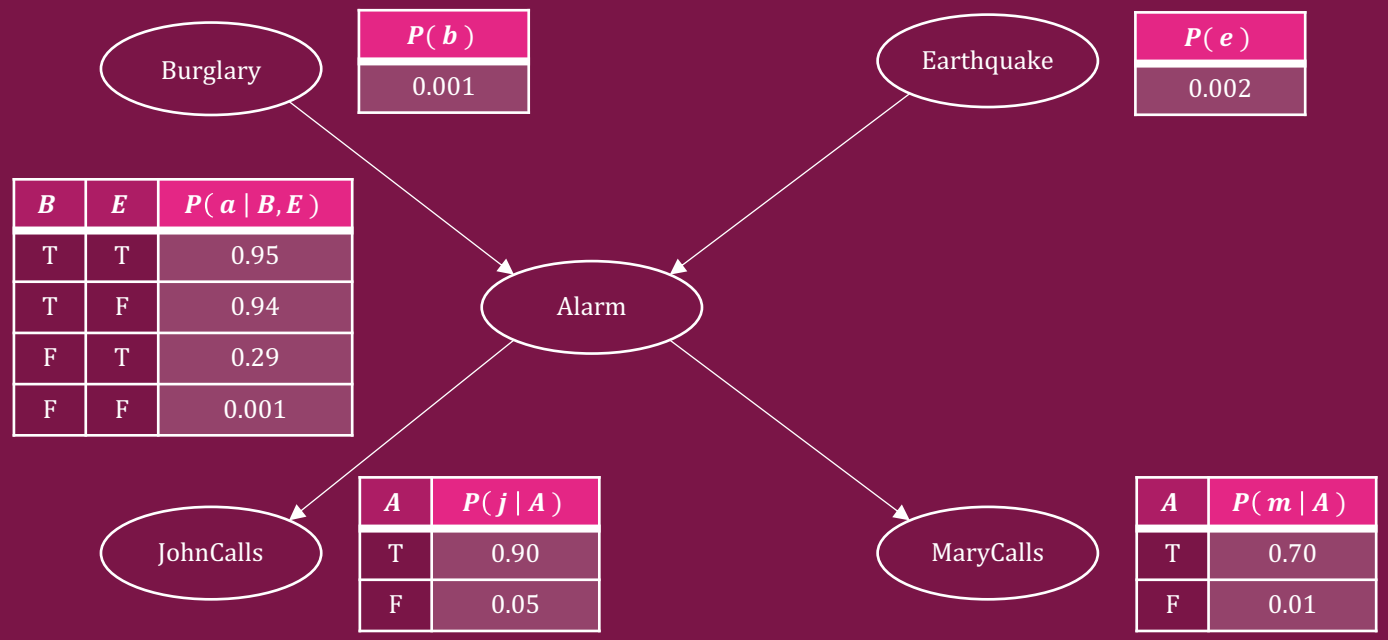


But wait! Remember the normalization constant, $P(j, m)$? We need to calculate that.

Recall that $P(j, m)$ is $P(data)$, so that's all the ways the data can be observed:

$$P(j, m) = P(j, m \mid b)\, P(b) + P(j, m \mid \neg b)\, P(\neg b) = P(b, j, m) + P(\neg b, j, m)$$

The first colored term is exactly what we just computed. We need to compute the other.



After computing $P(\neg b, j, m)$ in the same way in which we computed $P(b, j, m)$, we have:

$P(\neg b, j, m) = 0.0015$ and $P(b, j, m) = 0.00059$

So, $P(j, m) = P(b, j, m) + P(\neg b, j, m) = 0.00059 + 0.0015 = 0.00209$

$$\therefore P(b \mid j, m) = \frac{P(j, m \mid b)\, P(b)}{P(j, m)} = \frac{P(b, j, m)}{P(j, m)} = 0.28$$