

Machine Learning for Data Science (CS4786)

Lecture 22

Inference in Bayesian Networks

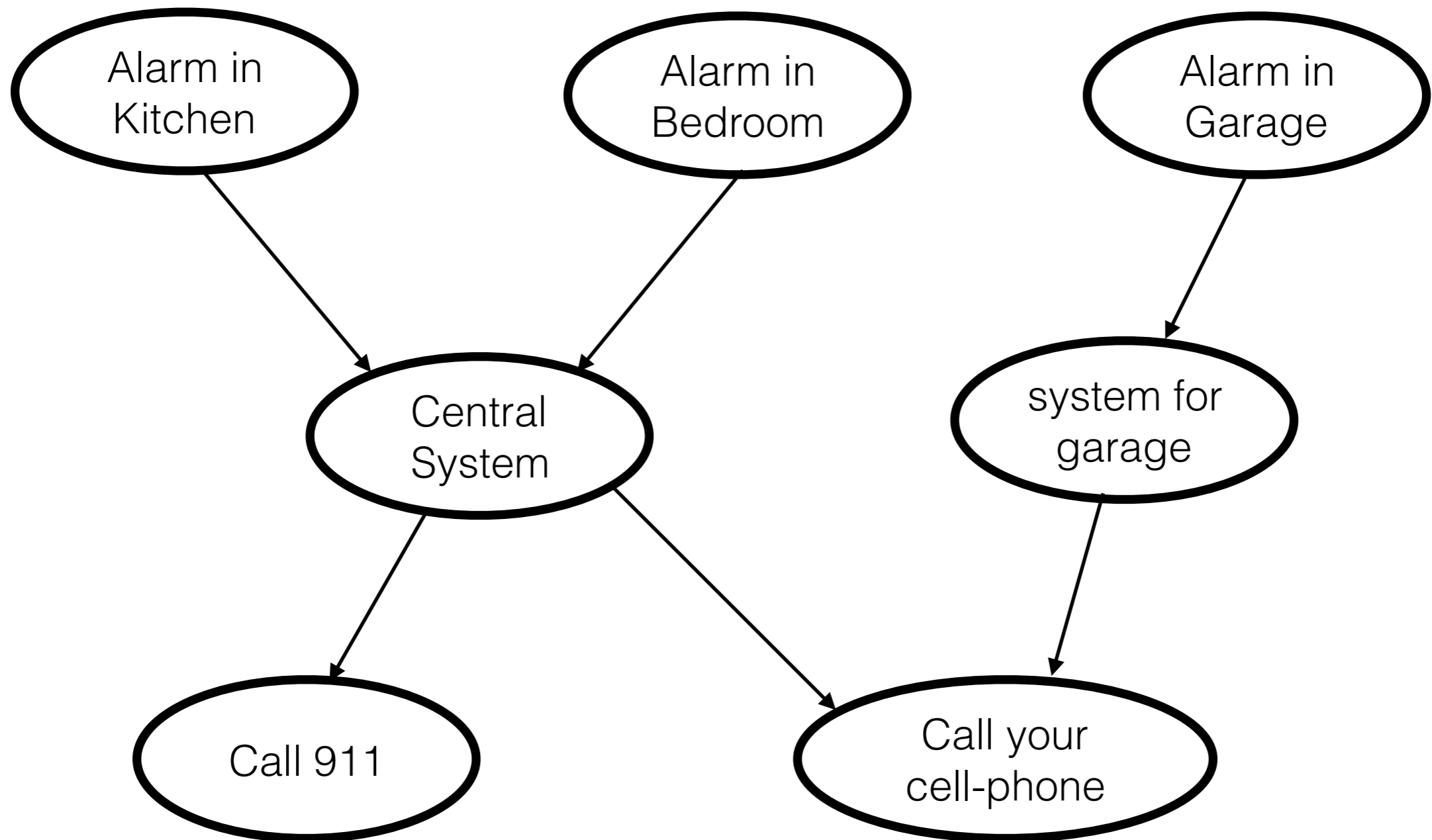
Course Webpage :

<http://www.cs.cornell.edu/Courses/cs4786/2017fa/>

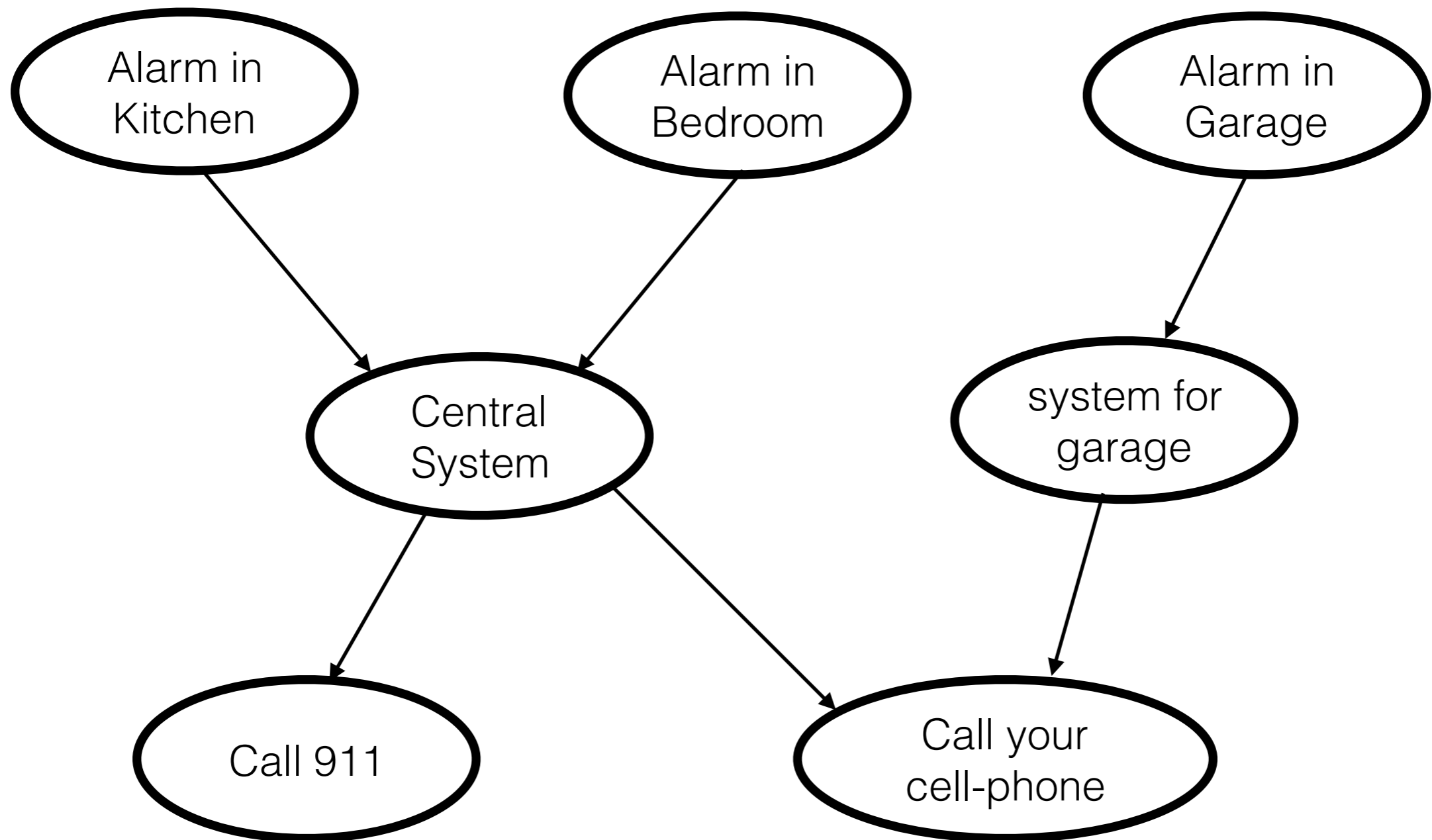
Announcement

- Competition 1: Overall excellent performance, reports look good, great work!
- Competition 2 will be posted in a couple days, its focused on HMM

BELIEF PROPAGATION

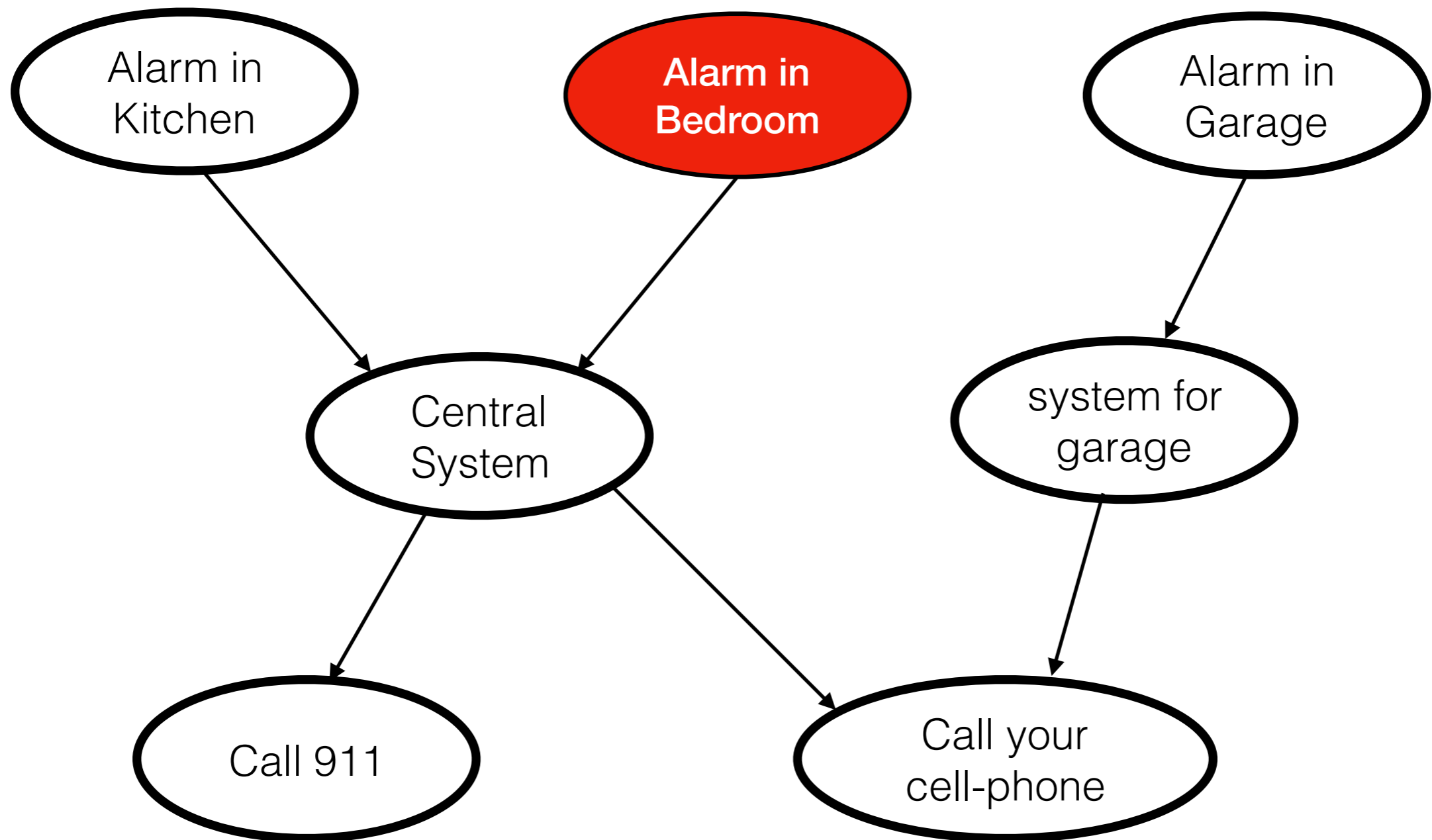


BELIEF PROPAGATION



You receives phone call

BELIEF PROPAGATION



You receives phone call

BELIEF PROPAGATION

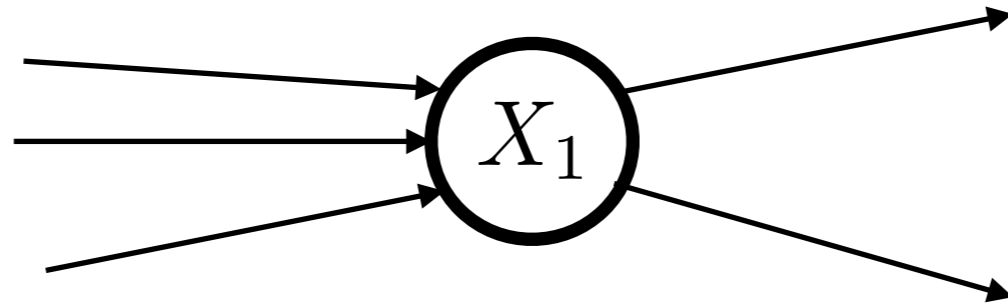
- Compute probability of fire in kitchen using messages received in last round

BELIEF PROPAGATION

- On each round: Receive messages from previous round

BELIEF PROPAGATION

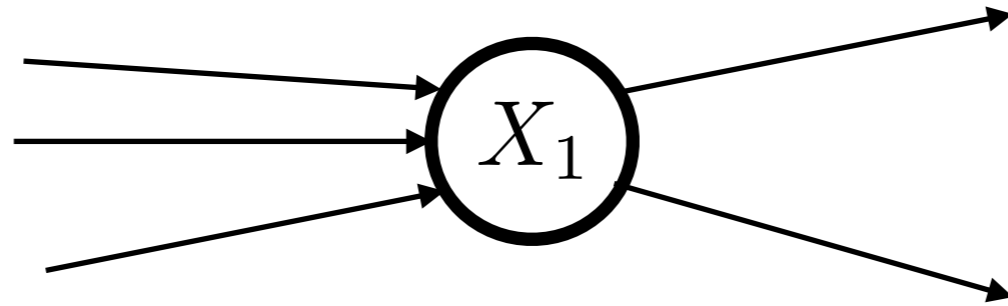
- On each round: Receive messages from previous round



BELIEF PROPAGATION

- On each round: Receive messages from previous round

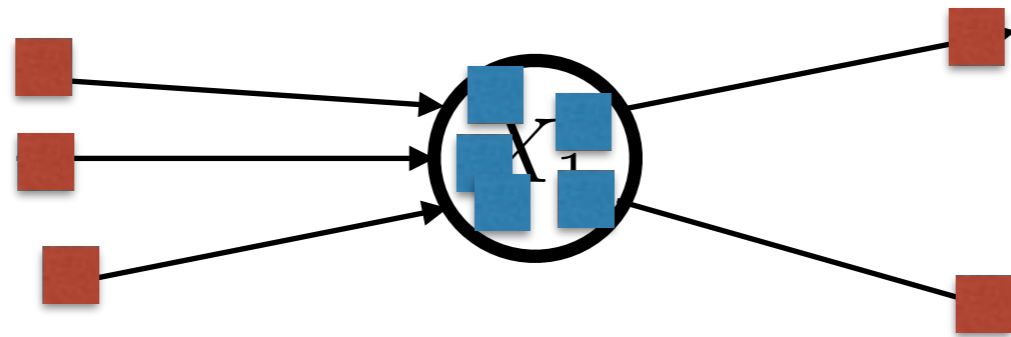
Round $t-1$



BELIEF PROPAGATION

- On each round: Receive messages from previous round

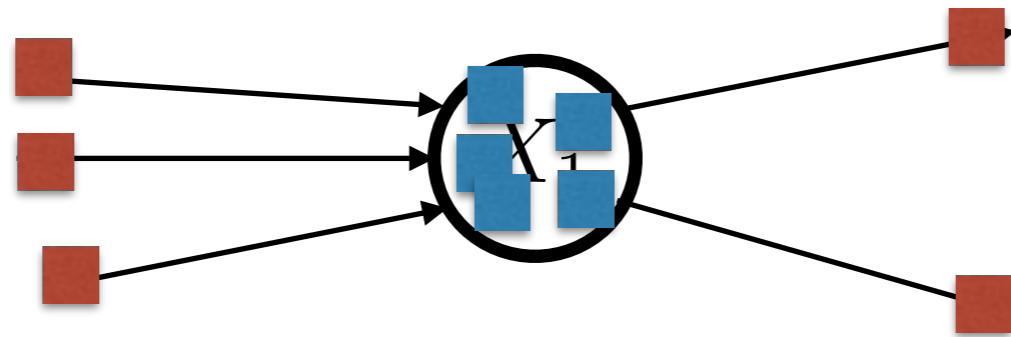
Round t



BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node X_i to **Child** X_k on round t

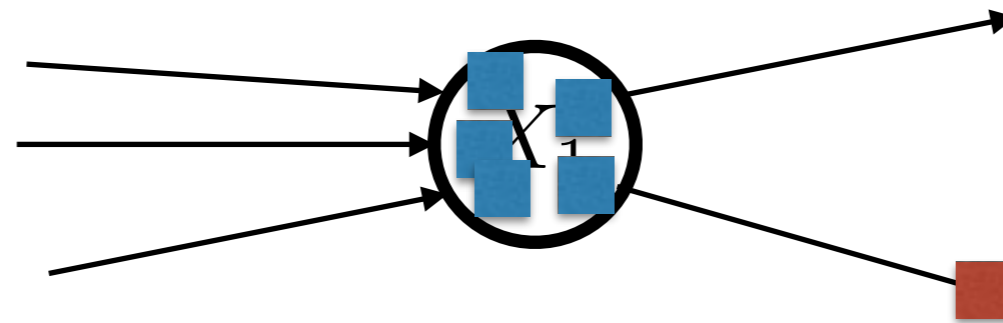
$$M_{i \rightarrow k}^t(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i)) \text{ (product of all messages but one from } X_j \text{)}$$

from previous round (t-1)

BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node X_i to **Child** X_k on round t

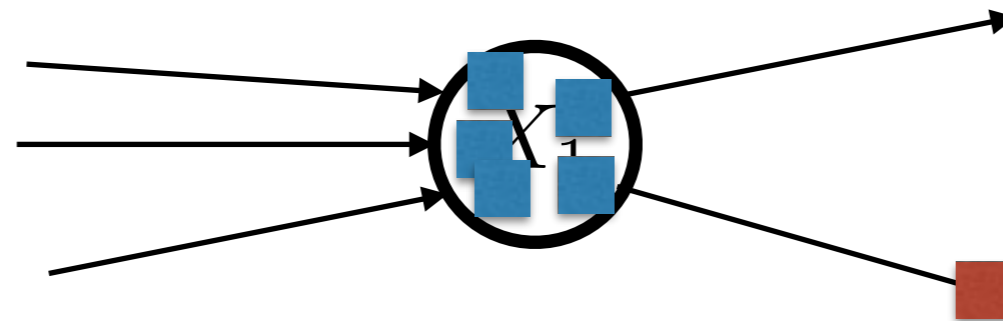
$$M_{i \rightarrow k}^t(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i)) \text{ (product of all messages but one from } X_j \text{)}$$

from previous round (t-1)

BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node X_i to **Child** X_k on round t

$$M_{i \rightarrow k}^t(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i))$$

(product of all messages but one from X_j)

from previous round (t-1)

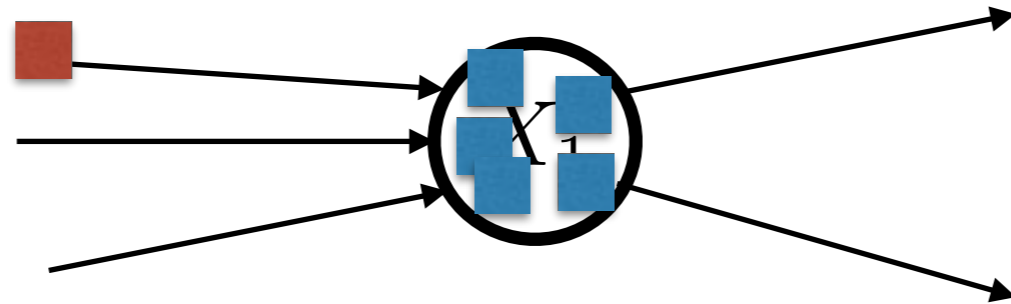
BELIEF PROPAGATION

- On each round: Receive messages from previous round

BELIEF PROPAGATION

- On each round: Receive messages from previous round

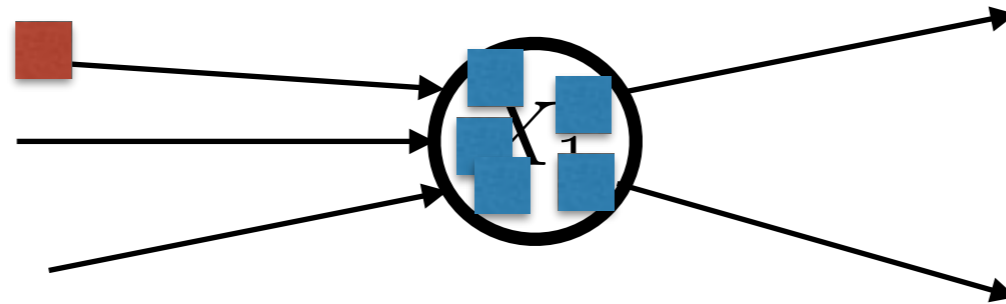
Round t



BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



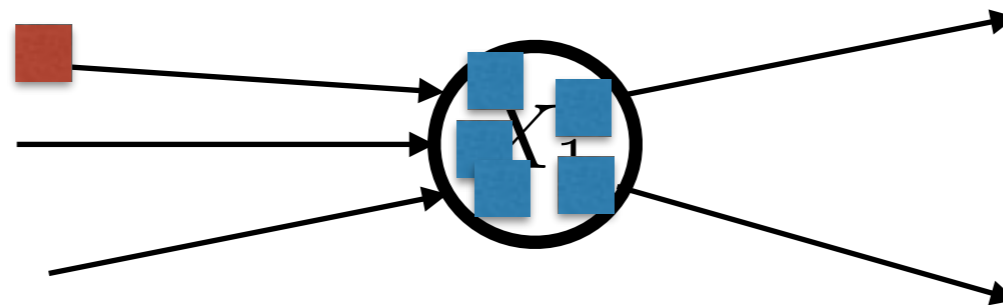
Message from node X_i to **Parent** X_j on round t

$$M_{i \rightarrow j}(x_j) = \sum_{x_i, \text{Parents}(X_i) \setminus X_j} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i)) \text{ (product of all messages but one from } X_j \text{) from previous round (t-1)}$$

BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node X_i to **Parent** X_j on round t

$$M_{i \rightarrow j}(x_j) = \sum_{x_i, \text{Parents}(X_i) \setminus X_j} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i))$$

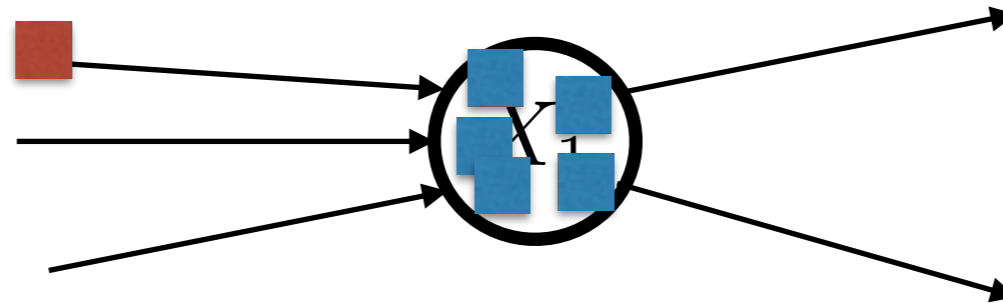
(product of all messages but one from X_j)
 from previous round (t-1)



BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node X_i to **Parent** X_j on round t

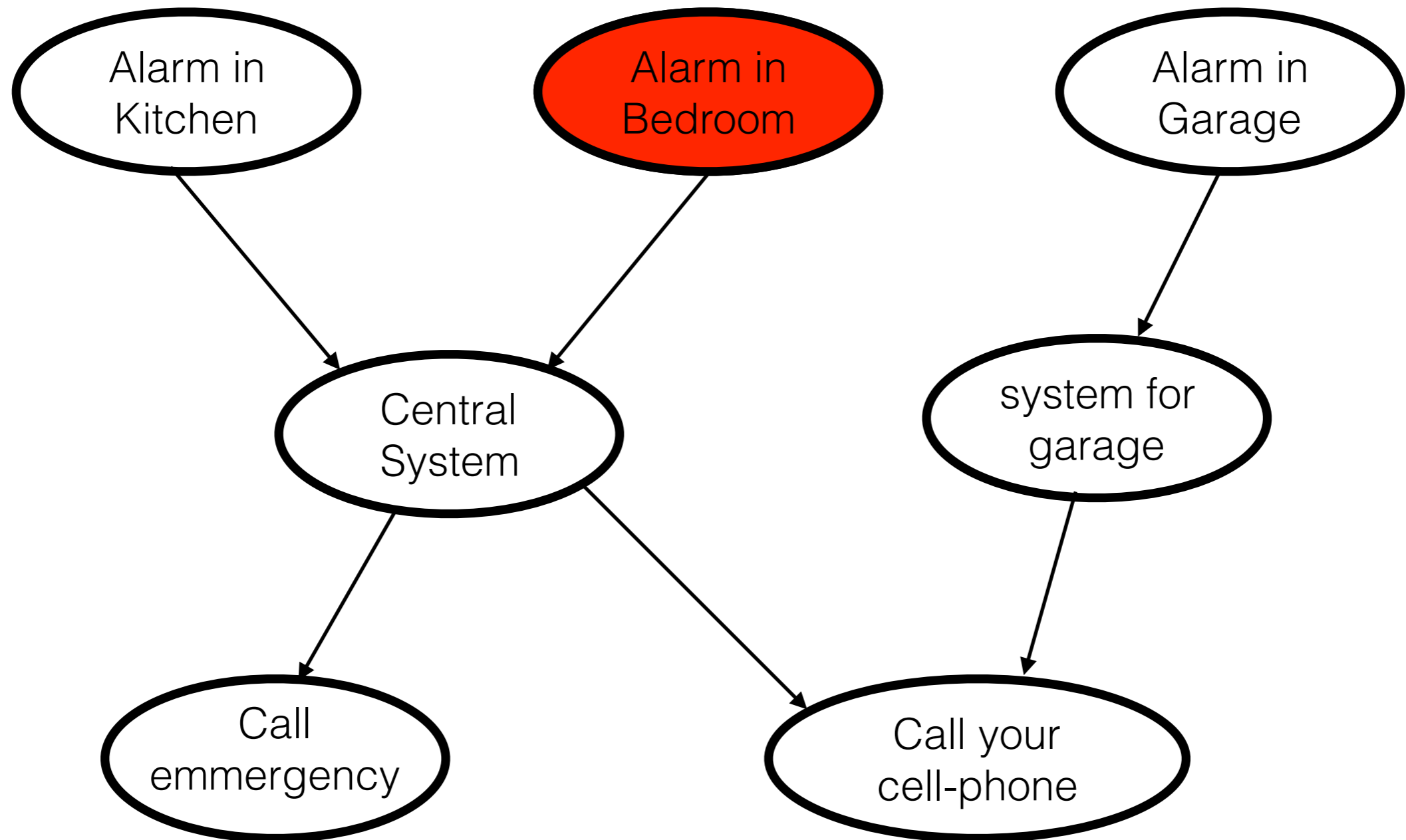
Only these vary over iterations

$$M_{i \rightarrow j}(x_j) = \sum_{x_i, \text{Parents}(X_i) \setminus X_j} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i))$$

(product of all messages but one from X_j)

from previous round (t-1)

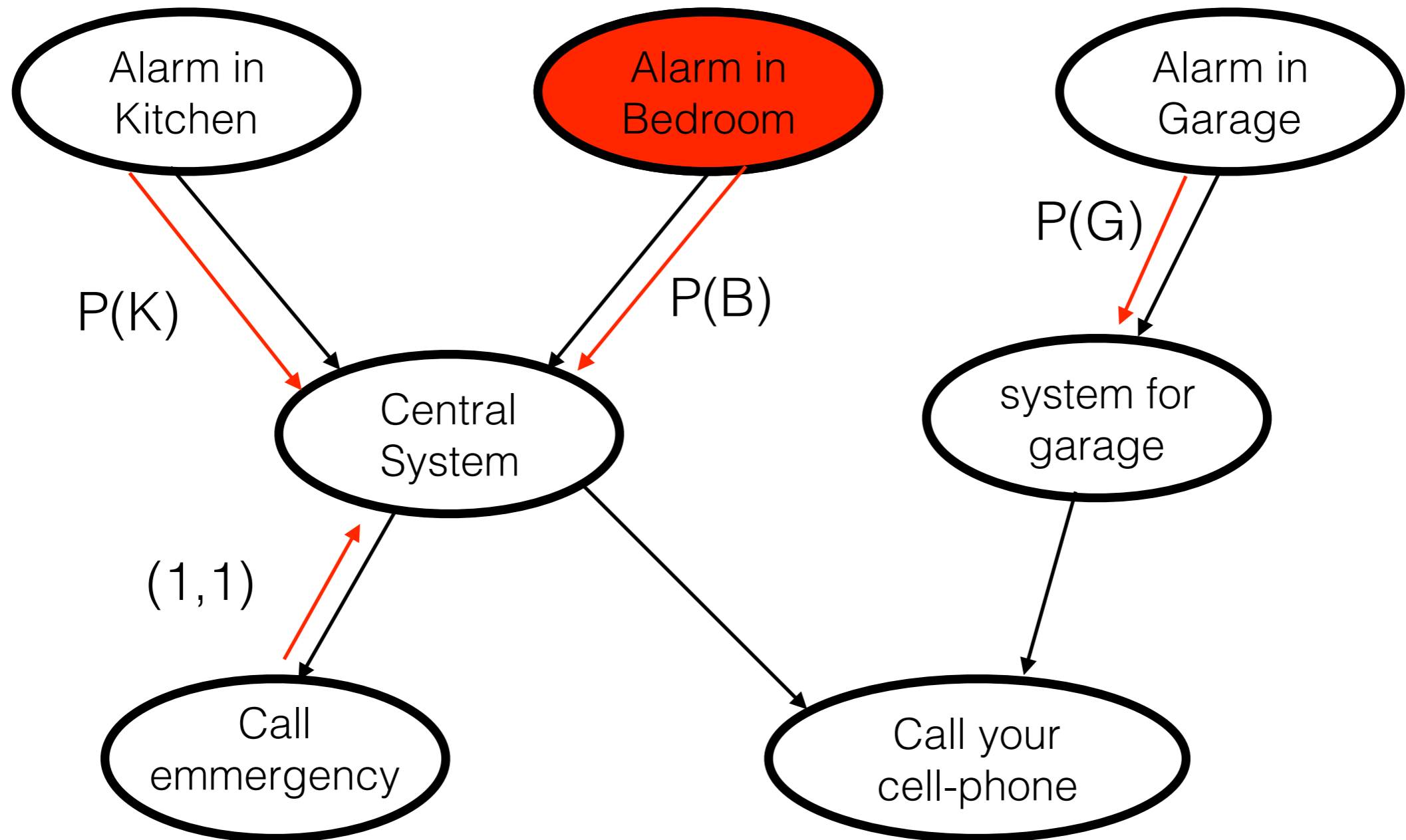
BELIEF PROPAGATION



You receive phone call

BELIEF PROPAGATION

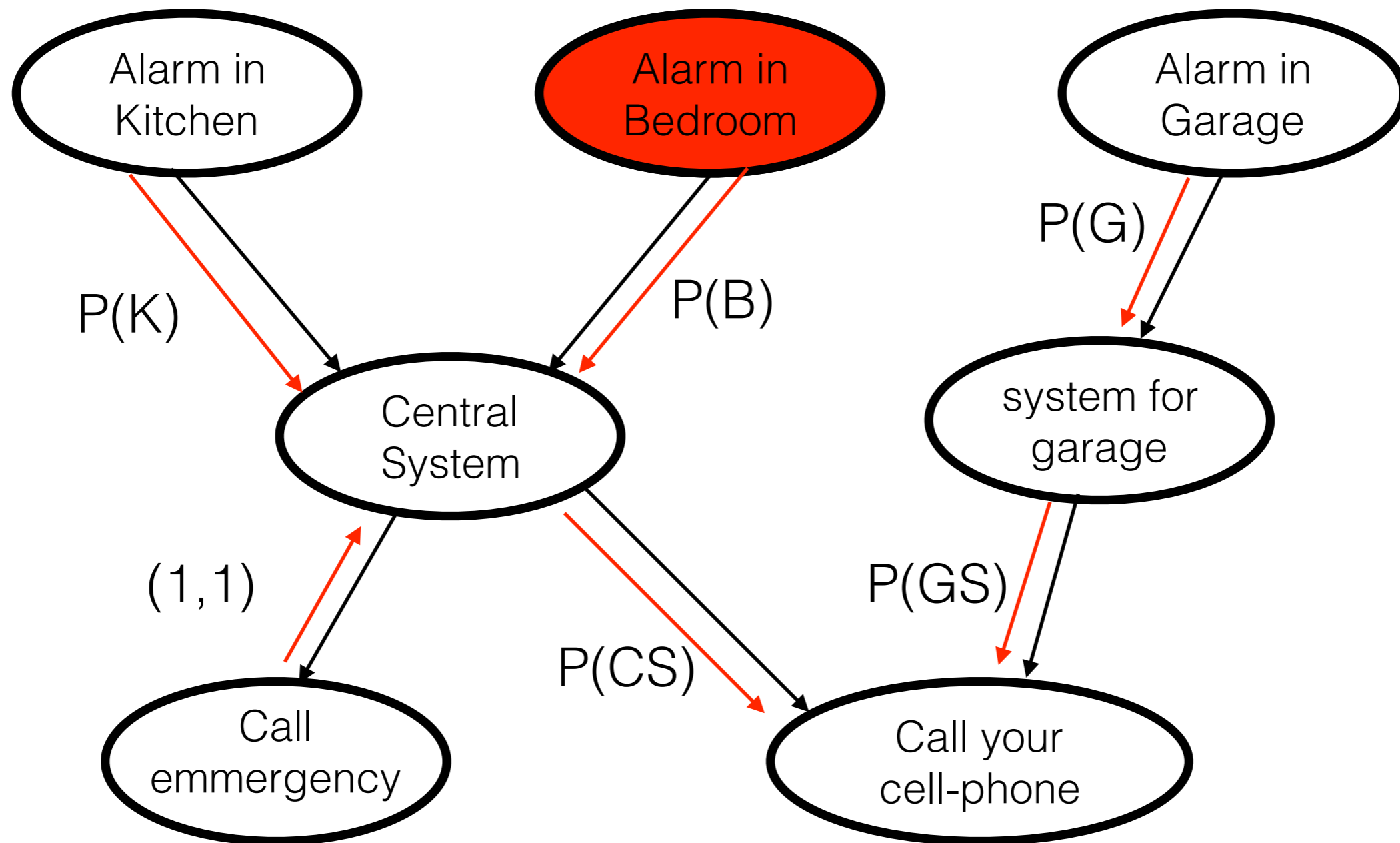
$i=1$



You receive phone call

BELIEF PROPAGATION

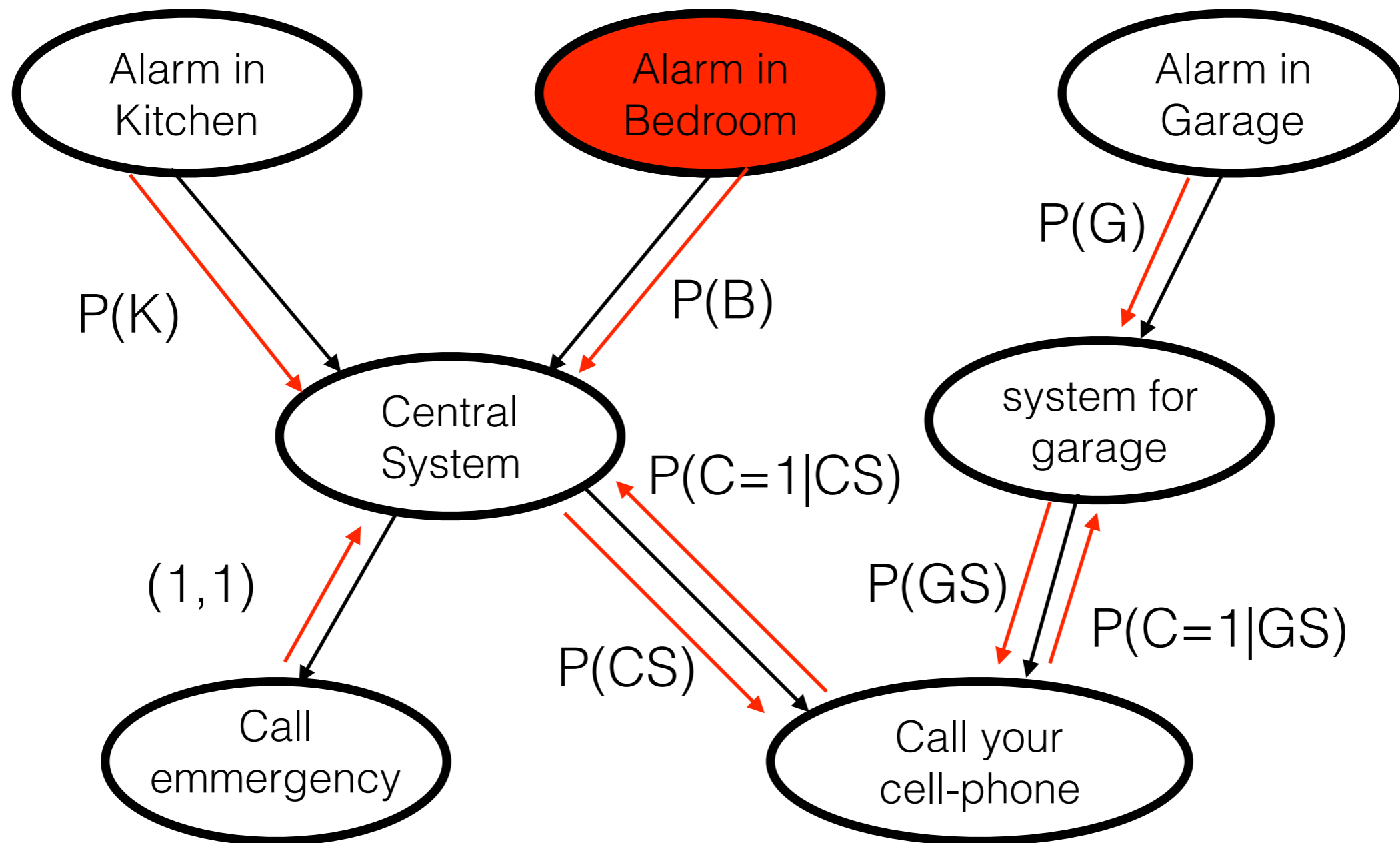
$i=1,2$



You receive phone call

BELIEF PROPAGATION

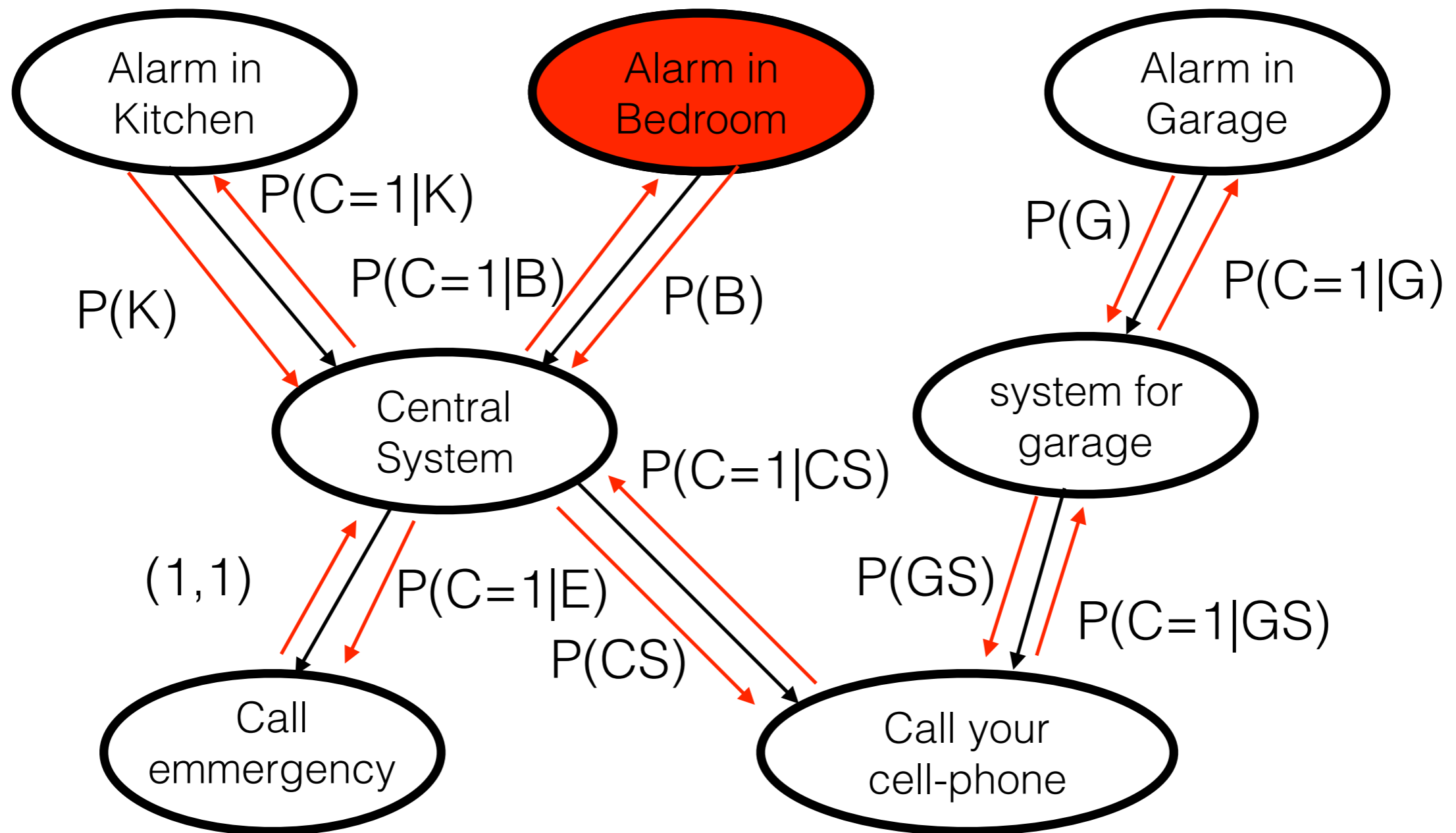
$i=1,2,3$



You receive phone call

BELIEF PROPAGATION

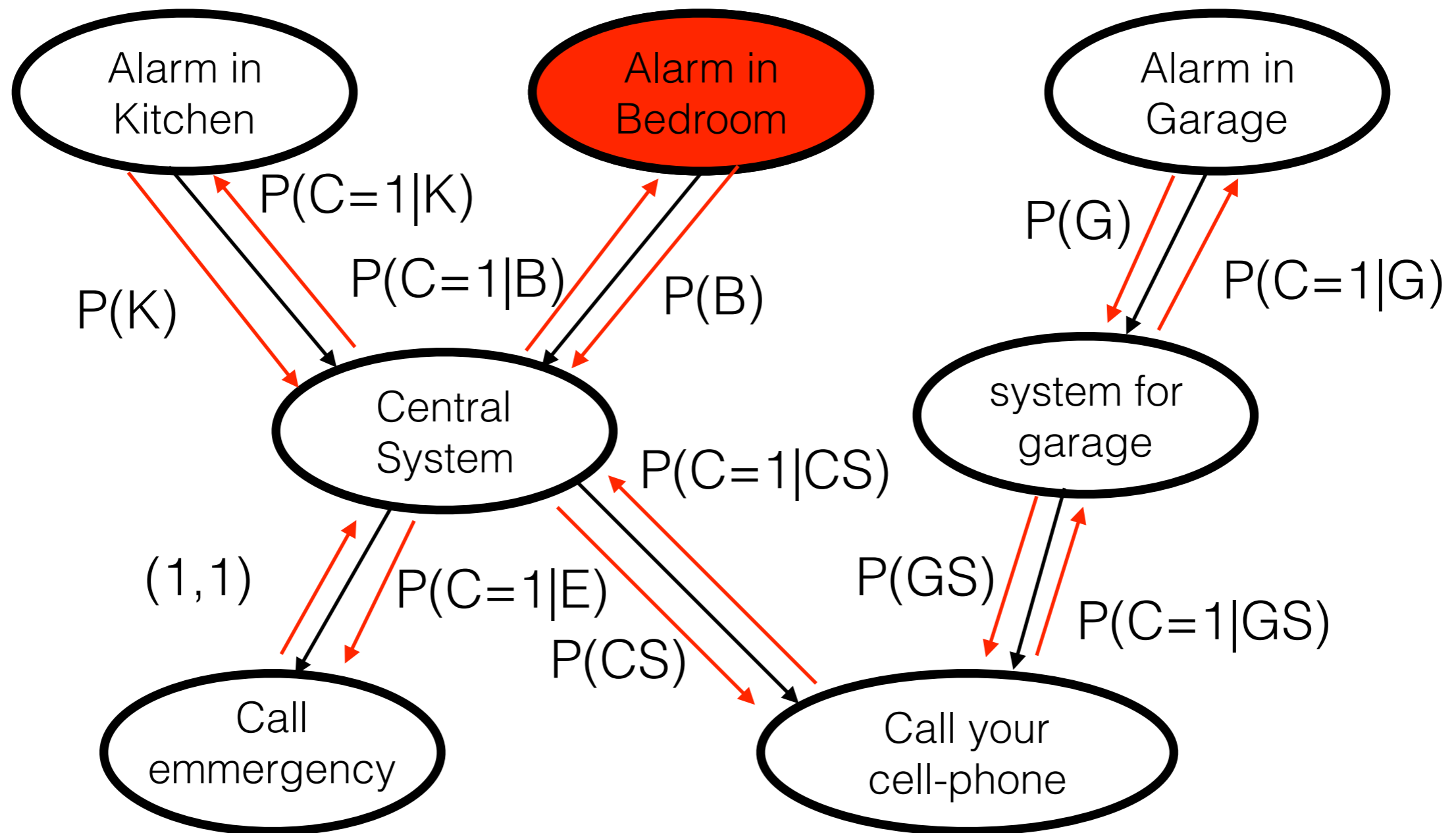
$i=1,2,3$



You receive phone call

BELIEF PROPAGATION

$i=1,2,3$



You receive phone call

BELIEF PROPAGATION

BELIEF PROPAGATION

After convergence:

$$P(X_i = x_i | \text{Observation}) \propto \sum_{\text{values of Parent}(X_i)} E_{X_i}(x_i) \times P(X_i = x_i | \text{Parent}(X_i)) \times \text{Product of all messages}$$

We have inference, what about learning parameters for the model from data?

PARAMETER ESTIMATION (LEARNING)

- What are the parameters for a Bayesian Network?

PARAMETER ESTIMATION (LEARNING)

- What are the parameters for a Bayesian Network?
 - The conditional probability distributions/tables/density functions

PARAMETER ESTIMATION (LEARNING)

- MLE: n independent samples $(X_1^1, \dots, X_N^1), \dots, (X_1^n, \dots, X_N^n)$ where each (X_1^t, \dots, X_N^t) is drawn from the Bayesian network

PARAMETER ESTIMATION (LEARNING)

- MLE: n independent samples $(X_1^1, \dots, X_N^1), \dots, (X_1^n, \dots, X_N^n)$ where each (X_1^t, \dots, X_N^t) is drawn from the Bayesian network

In this scenario, how do we learn conditional probability tables?

PARAMETER ESTIMATION (LEARNING)

- Simple case of finite outcomes

θ_i^{MLE} = empirical conditional probability table

PARAMETER ESTIMATION (LEARNING)

- MLE: n independent samples $(X_1^1, \dots, X_N^1), \dots, (X_1^n, \dots, X_N^n)$ where each (X_1^t, \dots, X_N^t) is drawn from the Bayesian network

$$\begin{aligned} & \arg \max_{\theta} \sum_{t=1}^n \log(P_{\theta}(X_1^t, \dots, X_N^t)) \\ & = \arg \max_{\theta} \sum_{t=1}^n \sum_{i=1}^N \log(P_{\theta}(X_i^t | \text{Parent}(X_i^t))) \end{aligned}$$

If θ_i is the parameter only involving $P_{\theta}(X_i^t | \text{Parent}(X_i^t))$ then

$$\theta_i^{MLE} = \arg \max_{\theta_i} \sum_{t=1}^n \log(P_{\theta_i}(X_i^t | \text{Parent}(X_i^t)))$$

PARAMETER ESTIMATION (LEARNING)

- MLE: n independent samples $(X_1^1, \dots, X_N^1), \dots, (X_1^n, \dots, X_N^n)$ where each (X_1^t, \dots, X_N^t) is drawn from the Bayesian network

PARAMETER ESTIMATION (LEARNING)

- MLE: n independent samples $(X_1^1, \dots, X_N^1), \dots, (X_1^n, \dots, X_N^n)$ where each (X_1^t, \dots, X_N^t) is drawn from the Bayesian network

What is the problem?

Hint: think of the HMM example

PARAMETER ESTIMATION: LATENT VARIABLES

- EM Algorithm: Initialize parameters randomly
- For $j = 1$ to convergence
 - E-step: For each of the Latent variable X_i , perform inference to compute

$$Q^{(j)}(\text{Latent variables}) = P_{\theta^{(j-1)}}(\text{Latent variables}|\text{Observation})$$

- M-step:

$$\theta^{(j)} = \arg \max_{\theta} \sum_{\text{Latent variables}} Q^{(j)}(\text{Latent variables}) \sum_{t=1}^n \log P_{\theta}(X_1^t, \dots, X_N^t)$$

which can be simplified to:

$$\theta_i^{(j)} = \arg \max_{\theta_i} \sum_{\text{Latent}} Q^{(j)}(\text{Latent}) \sum_{t=1}^n \log P_{\theta_i}(X_i^t | \text{Parent}(X_i^t))$$

PARAMETER ESTIMATION: LATENT VARIABLES

- EM Algorithm: Initialize parameters randomly
- For $j = 1$ to convergence
 - E-step: For each of the Latent variable X_i , perform inference to compute

$$Q^{(j)}(\text{Latent variables}) = P_{\theta^{(j-1)}}(\text{Latent variables}|\text{Observation})$$

- M-step:

$$\theta^{(j)} = \arg \max_{\theta} \sum_{\text{Latent variables}} Q^{(j)}(\text{Latent variables}) \sum_{t=1}^n \log P_{\theta}(X_1^t, \dots, X_N^t)$$

$$\theta_i^{(j)} = \arg \max_{\theta_i} \sum_{t=1}^n \sum_{X_i^t, \text{Parent}(X_i^t)} P_{\theta^{(j-1)}}(X_i^t, \text{Parent}(X_i^t)|\text{Observation}) \log P_{\theta_i}(X_i^t|\text{Parent}(X_i^t))$$

PARAMETER ESTIMATION: LATENT VARIABLES

M-step for simple case of finite outcomes

$\theta_i^{(j)}$ = empirical conditional probability table weighted by $Q^{(j)}$

For HMM this is called the Baum Welch algorithm

PARAMETER ESTIMATION: LATENT VARIABLES

- So if we had inference, learning follows easily via EM algorithm
- M-step is simply computing weighted MLE

PARAMETER ESTIMATION: LATENT VARIABLES

E-step computed using inference
How?

- So if we had inference, learning follows easily via EM algorithm
- M-step is simply computing weighted MLE

INFERENCE IS COMPUTATIONALLY HARD!

- Belief propagation is exact on trees
- For general graphs, belief propagation need not work
- Inference for general graphs can be computationally hard

Can we perform inference approximately?

WHAT IS APPROXIMATE INFERENCE?

- Obtain $\hat{P}(X_v|\text{Observation})$ that is close to $P(X_v|\text{Observation})$
 - Additive approximation:

$$|\hat{P}(X_v|\text{Observation}) - P(X_v|\text{Observation})| \leq \epsilon$$

- Multiplicative approximation:

$$(1 - \epsilon) \leq \frac{\hat{P}(X_v|\text{Observation})}{P(X_v|\text{Observation})} \leq (1 + \epsilon)$$

APPROXIMATE INFERENCE

Two approaches:

- Inference via sampling:
generate instances from the model, compute marginals
- Use exact inference but move to a close enough simplified model