Machine Learning for Data Science (CS4786)
Lecture 23

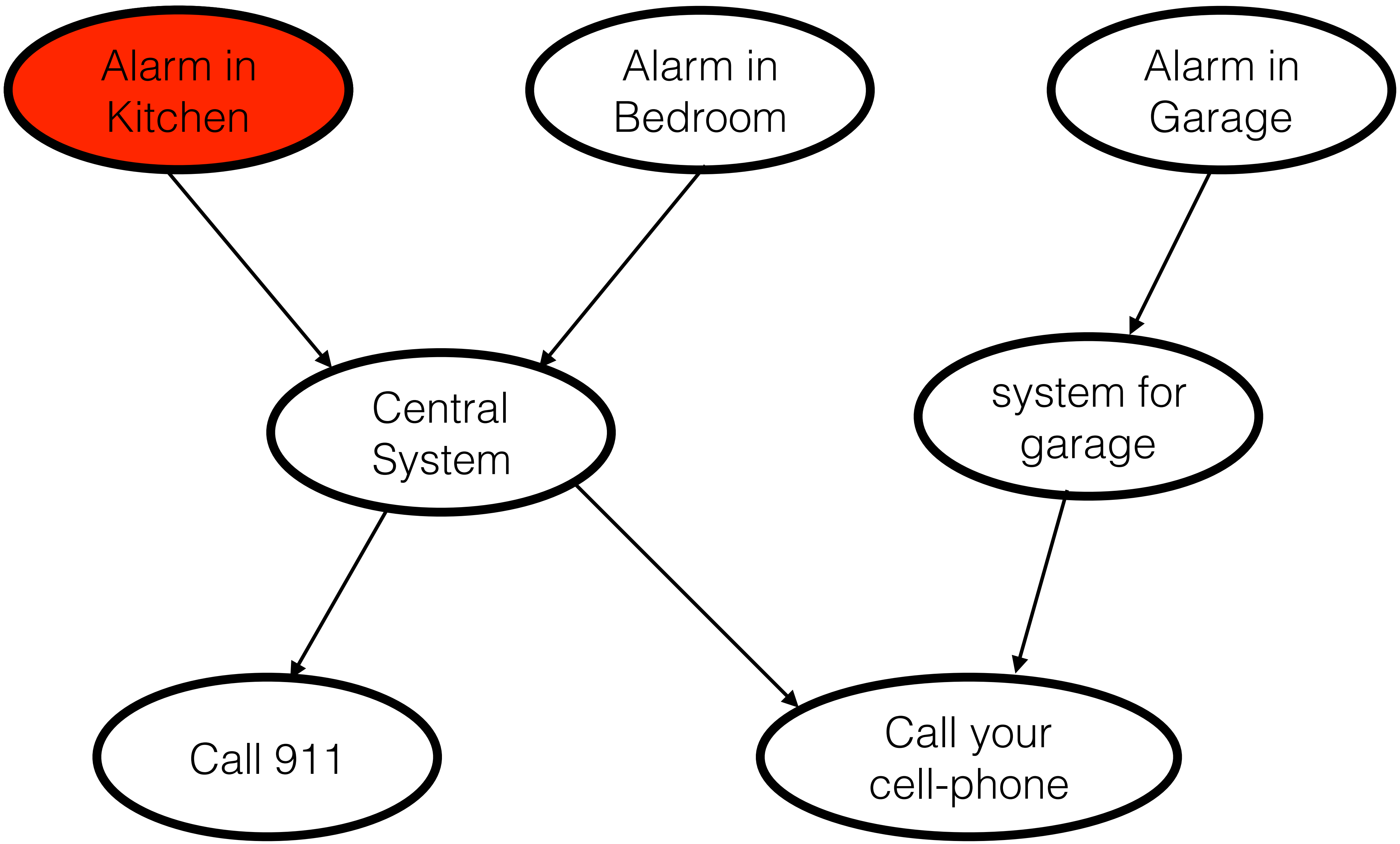
Message Passing and Learning in Graphical Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016fa/

Announcement

- Competition 1 feedback will be posted by end of this weekend.
    - Overall excellent performance, reports look good, great work!
- Competition 2 will be posted tonight, its a focused one based on HMM example in class

Belief Propagation



Neighbor receives phone call

Belief Propagation

- Nodes in the Bayesian network propagate beliefs or messages to their neighbors over multiple iterations
- Belief’s are vectors
  - Messages to children, belief about own value
  - Messages to parents, beliefs about parents’ value
- Evidence for each node takes into account observations

Belief Propagation

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

Belief Propagation

$$M_{i \mapsto j} \quad = \quad \text{Evidence-for-}X_i$$

$$\times \; P(X_i | \text{Parent}(X_i))$$

$$\times \; (\text{Product-of-all-messages-but-one-from-}X_j)$$

$$\text{(from previous round)}$$

Belief Propagation

If X_j is parent of X_i :

$M_{i \mapsto j}(x_j) =$

$$\sum_{\substack{\text{all other parents} \\ \text{and value of self}}} \left(\begin{array}{l} \text{Evidence-for-}X_i \\ \\ \times \ P(X_i | \text{Parent}(X_i)) \\ \\ \times \ (\text{Product-of-all-messages-but-one-from-}X_j) \\ \qquad\qquad\qquad (\text{from previous round}) \end{array} \right.$$

Belief Propagation

$$M_{i \mapsto j} = $$

$$\text{Evidence-for-}X_i$$

$$\times \; P(X_i | \text{Parent}(X_i))$$

$$\times \; (\text{Product-of-all-messages-but-one-from-}X_j)$$

$$\text{(from previous round)}$$

Belief Propagation

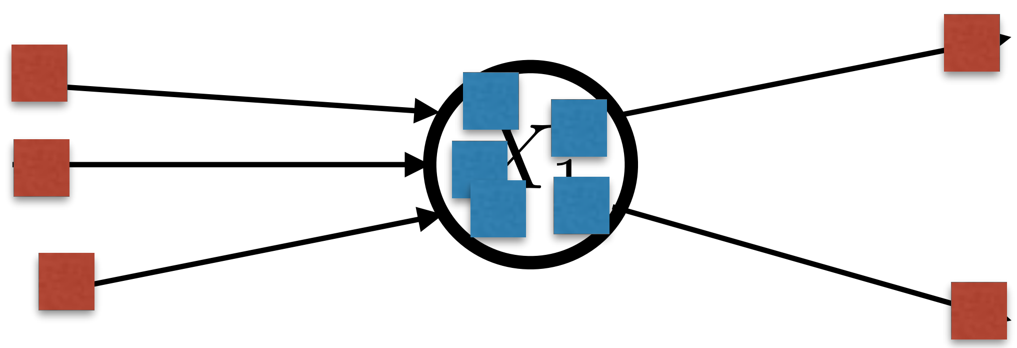
If X_j is the child of X_i :

$$M_{i \longmapsto j}(x_i) =$$

$$\sum_{\text{all parents' values}} \left(\begin{array}{l} \text{Evidence-for-}X_i \\ \\ \times\ P(X_i|\text{Parent}(X_i)) \\ \\ \times\ (\text{Product-of-all-messages-but-one-from-}X_j) \\ \qquad\qquad\qquad\qquad (\text{from previous round}) \end{array} \right.$$

- On each round: Receive messages from previous round

Round t



Message from node X_i to Child X_k on round t

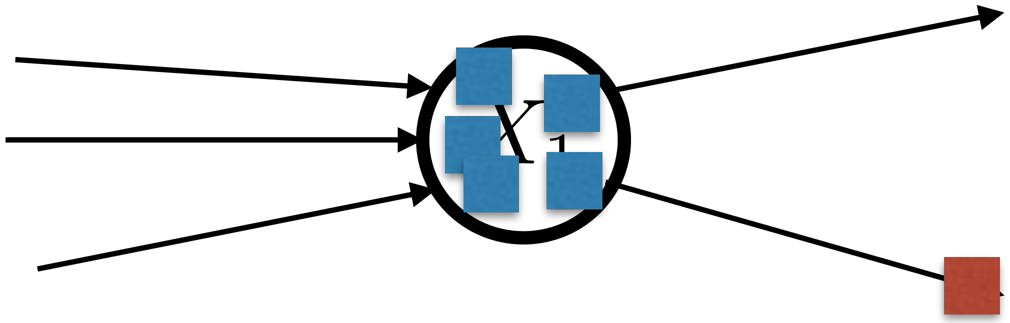
$$M^t_{i\mapsto k}(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i|\text{Parents}(X_i))\,(\text{product of all messages but one from } X_j)$$

from previous round (t-1)

Belief Propagation

- On each round: Receive messages from previous round

Round t



Message from node Xi to Child Xk on round t

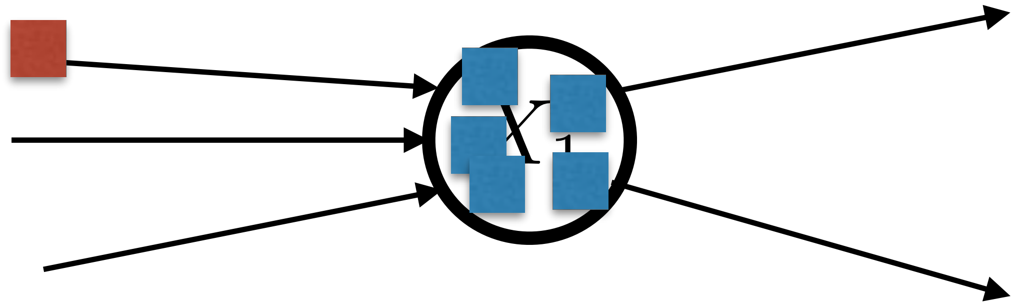
$$M^{t}_{i\mapsto k}(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i|\text{Parents}(X_i))\,(\text{product of all messages but one from } X_j)$$

from previous round (t-1)

Belief Propagation

- On each round: Receive messages from previous round

Round t



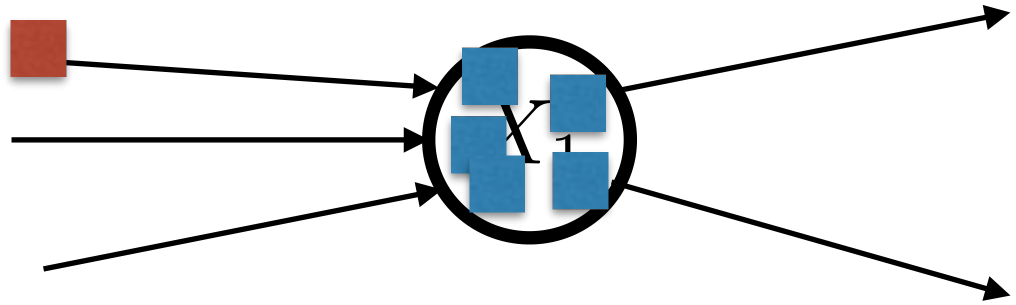
Message from node X_i to Parent X_j on round t

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

from previous round (t-1)

- On each round: Receive messages from previous round

Round t



Message from node Xi to Parent Xj on round t

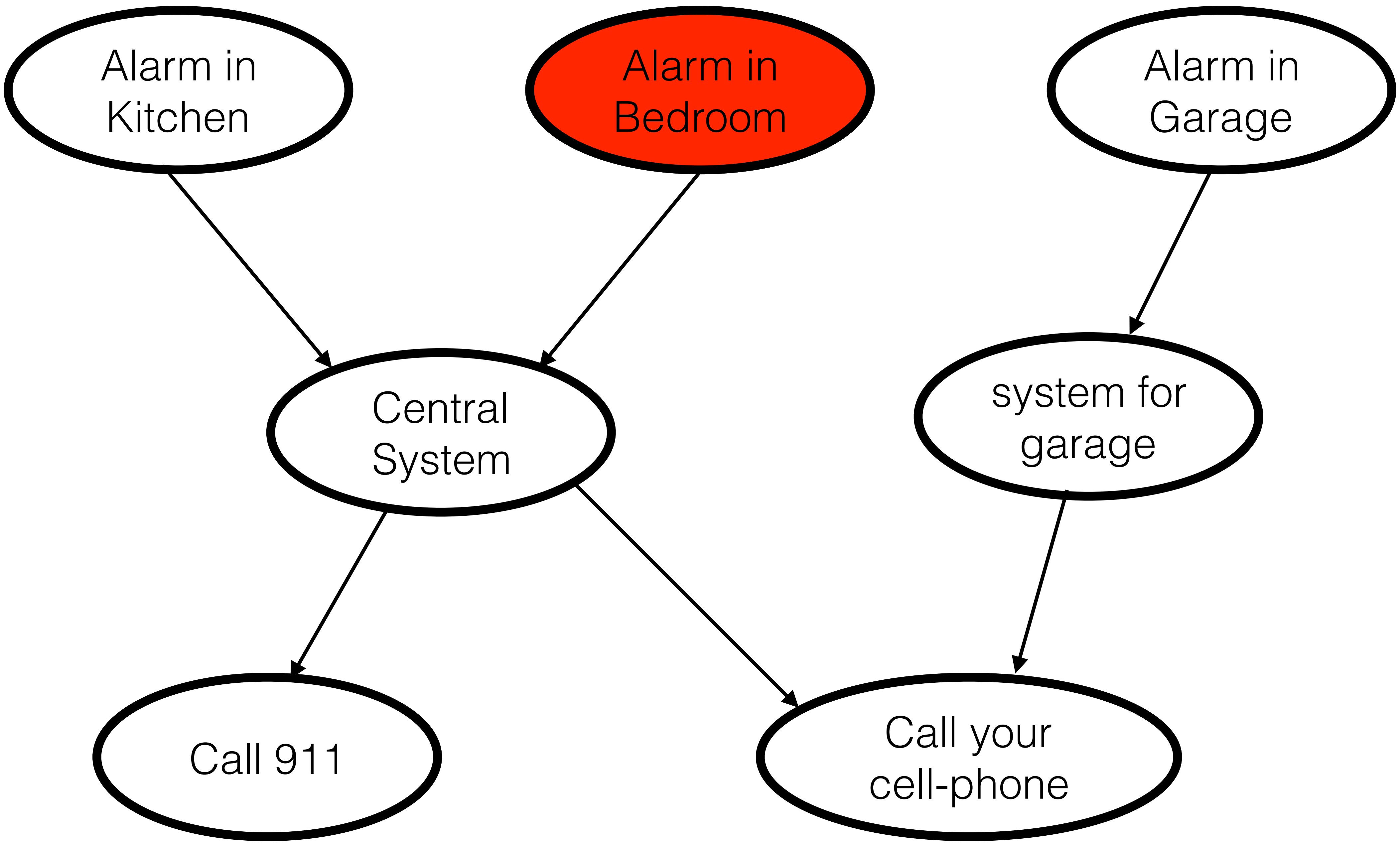
Only these vary over iterations

X X X

$$M_{i \mapsto j}(x_j) = \sum_{x_i, \mathrm{Parents}(X_i) \setminus X_j} E_{X_i}(x_i) P(X_i = x_i | \mathrm{Parents}(X_i)) \boxed{(\text{product of all messages but one from } X_j)}$$

from previous round (t-1)

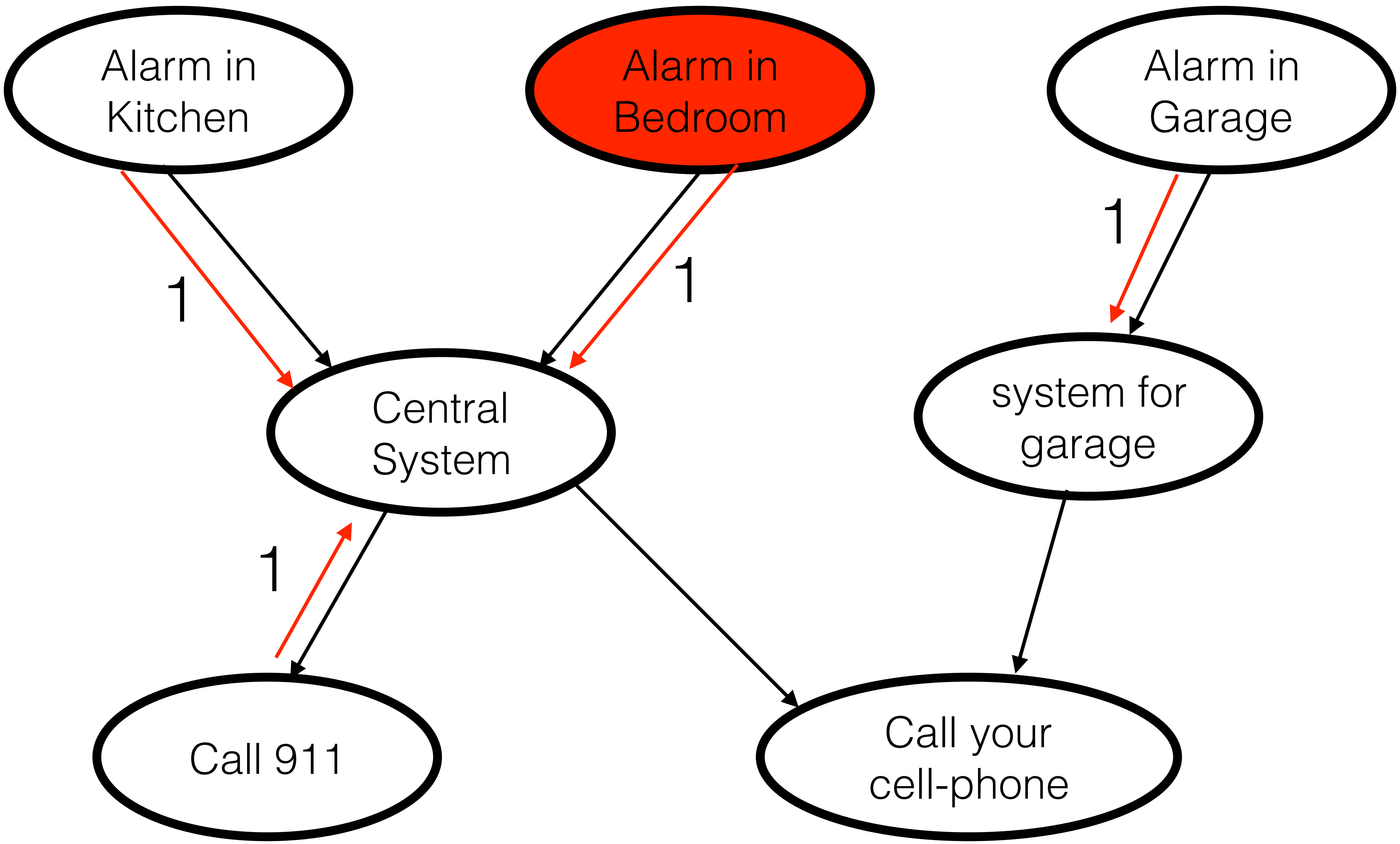
Belief Propagation



You receive phone call

Belief Propagation

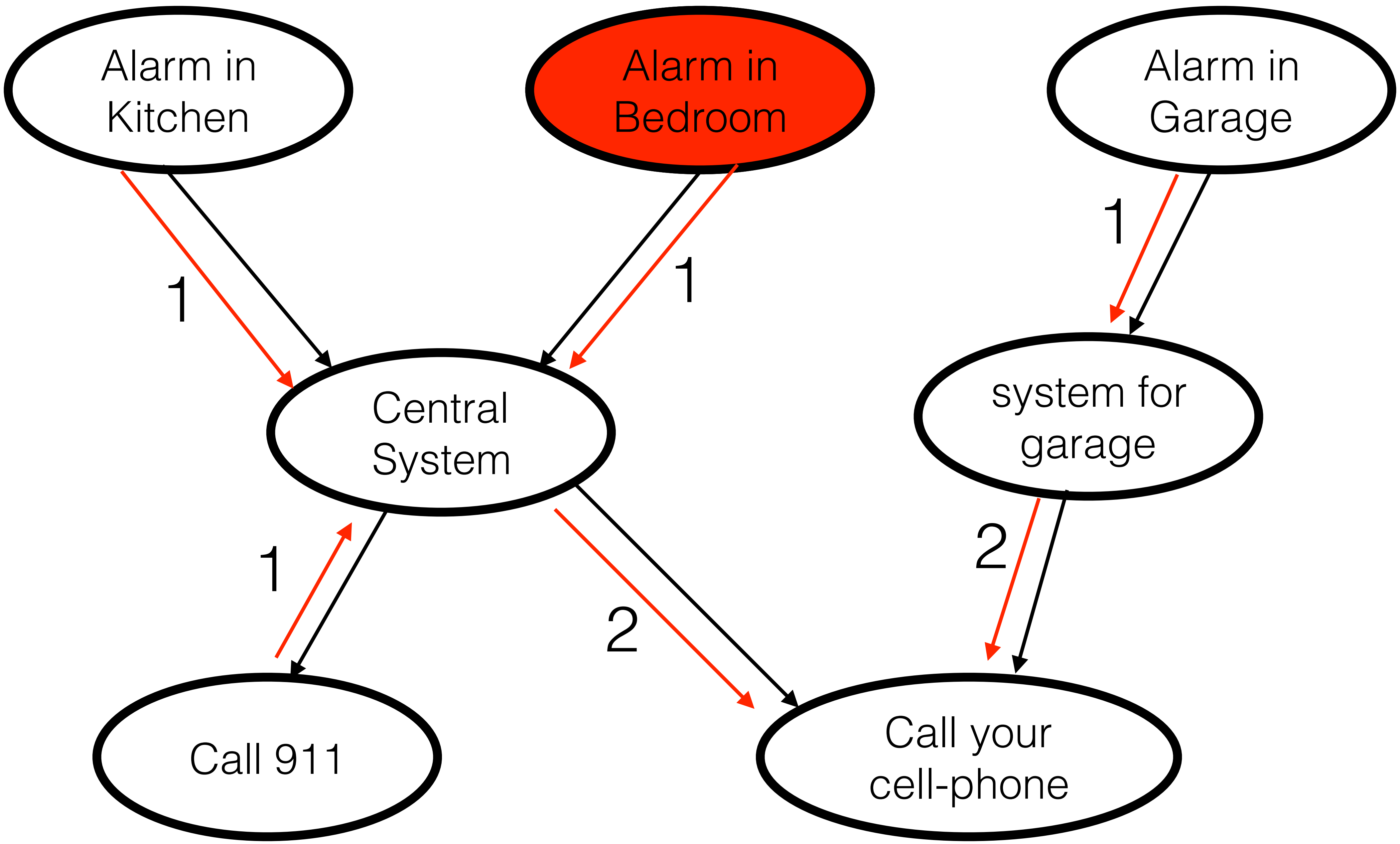
i=1



You receive phone call

Belief Propagation

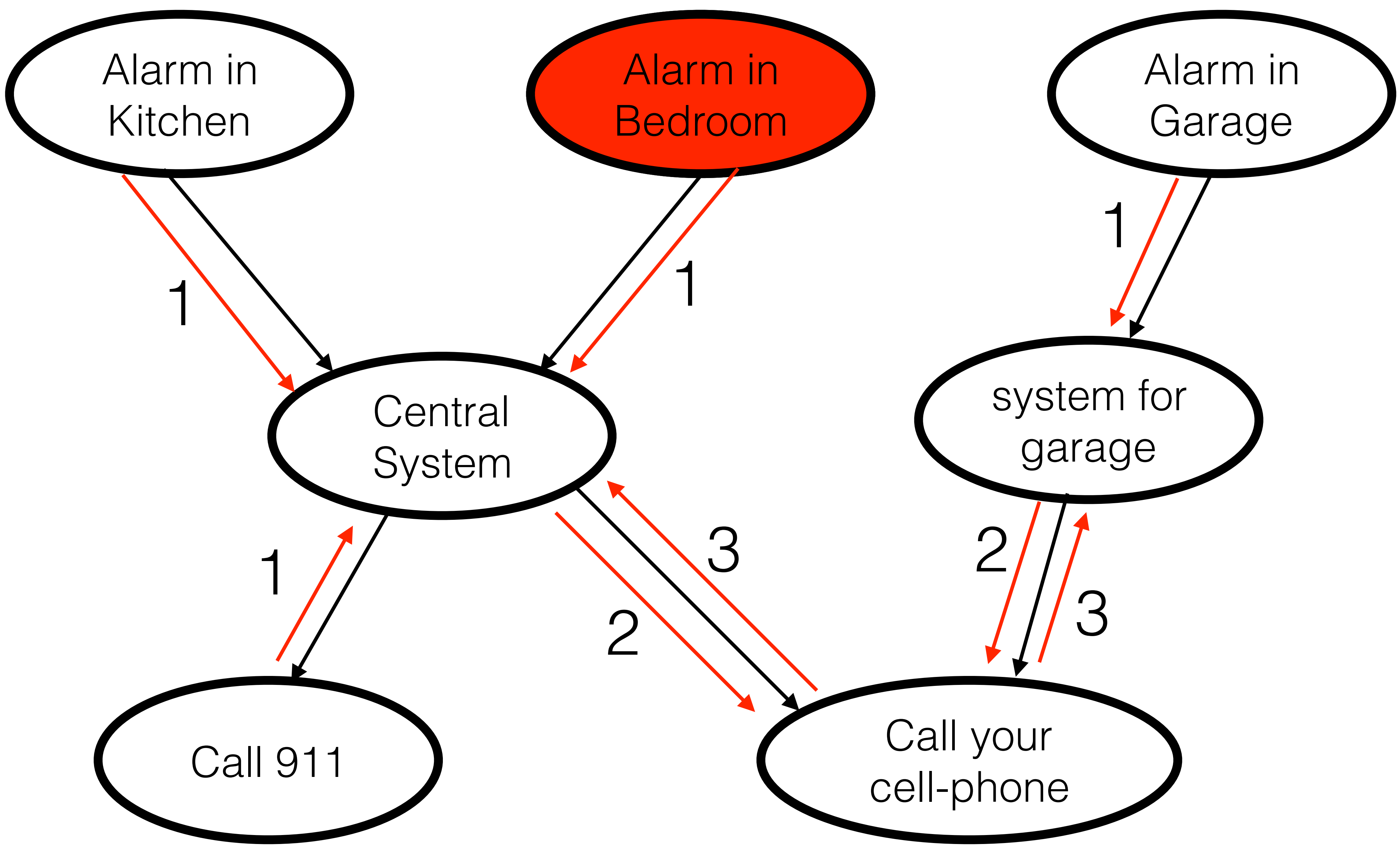
i=1,2



You receive phone call

Belief Propagation

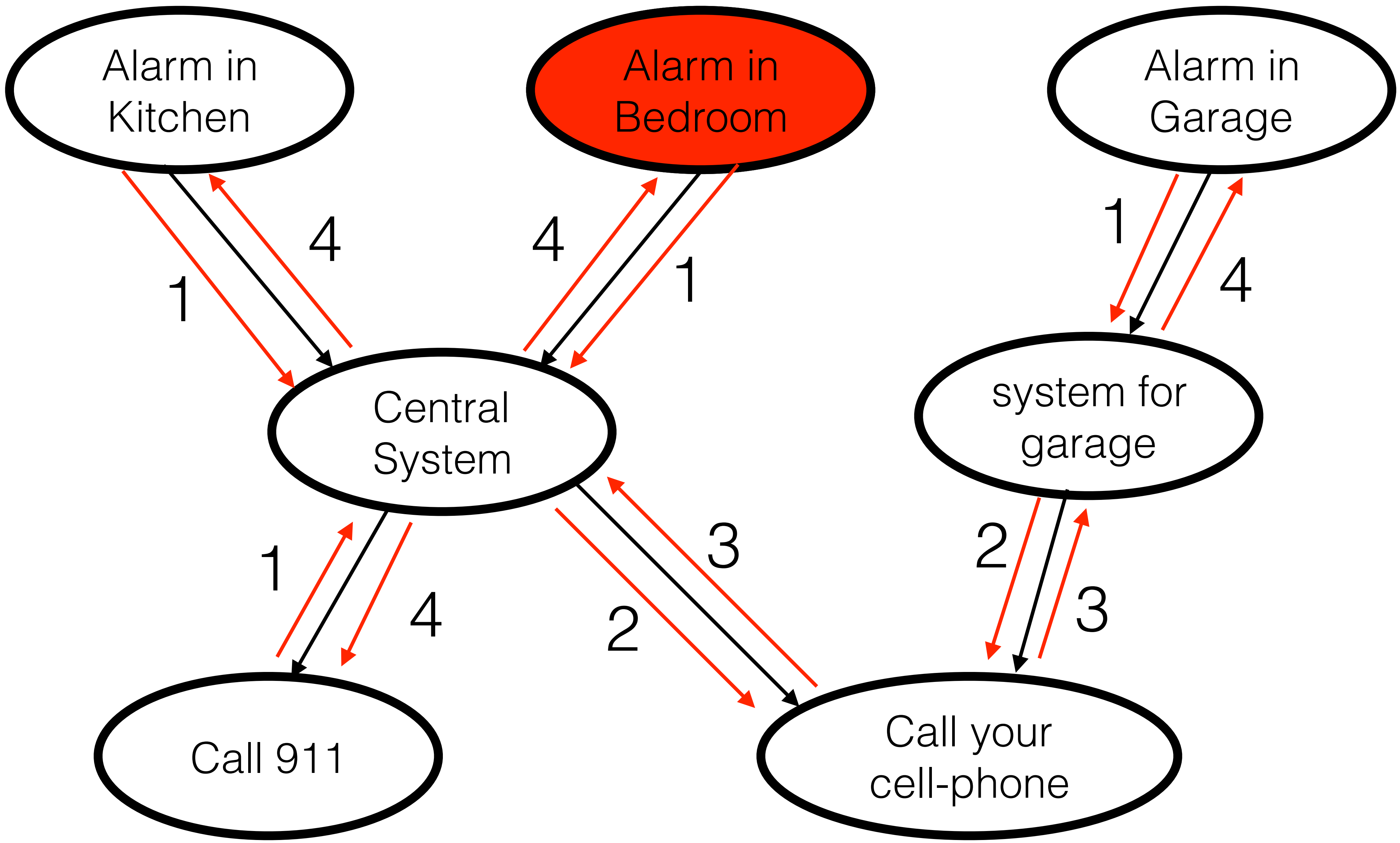
i=1,2 ,3



You receive phone call

Belief Propagation

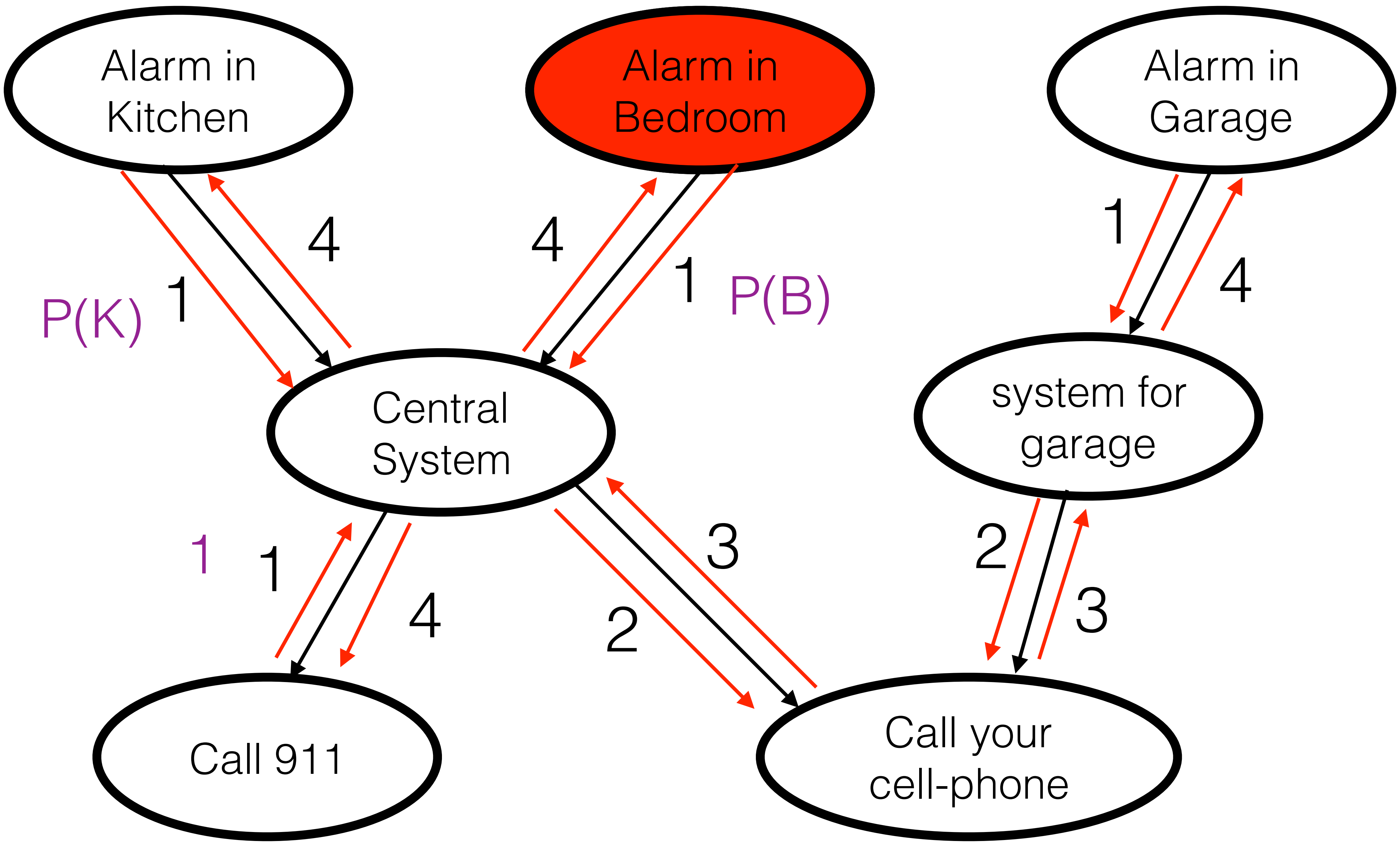
i=1,2 ,3



You receive phone call

Belief Propagation

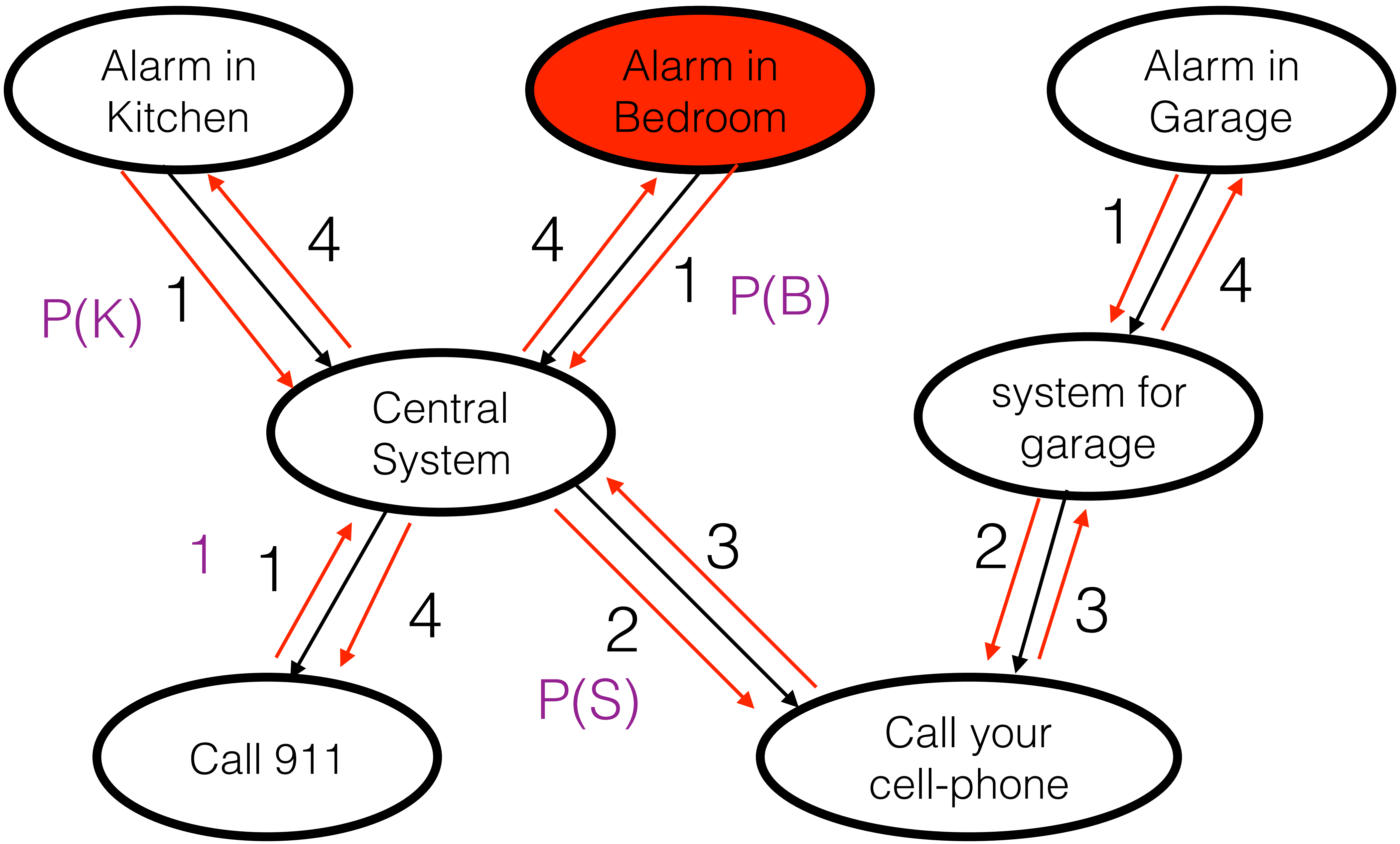
i=1,2 ,3



You receive phone call

BELIEF PROPAGATION

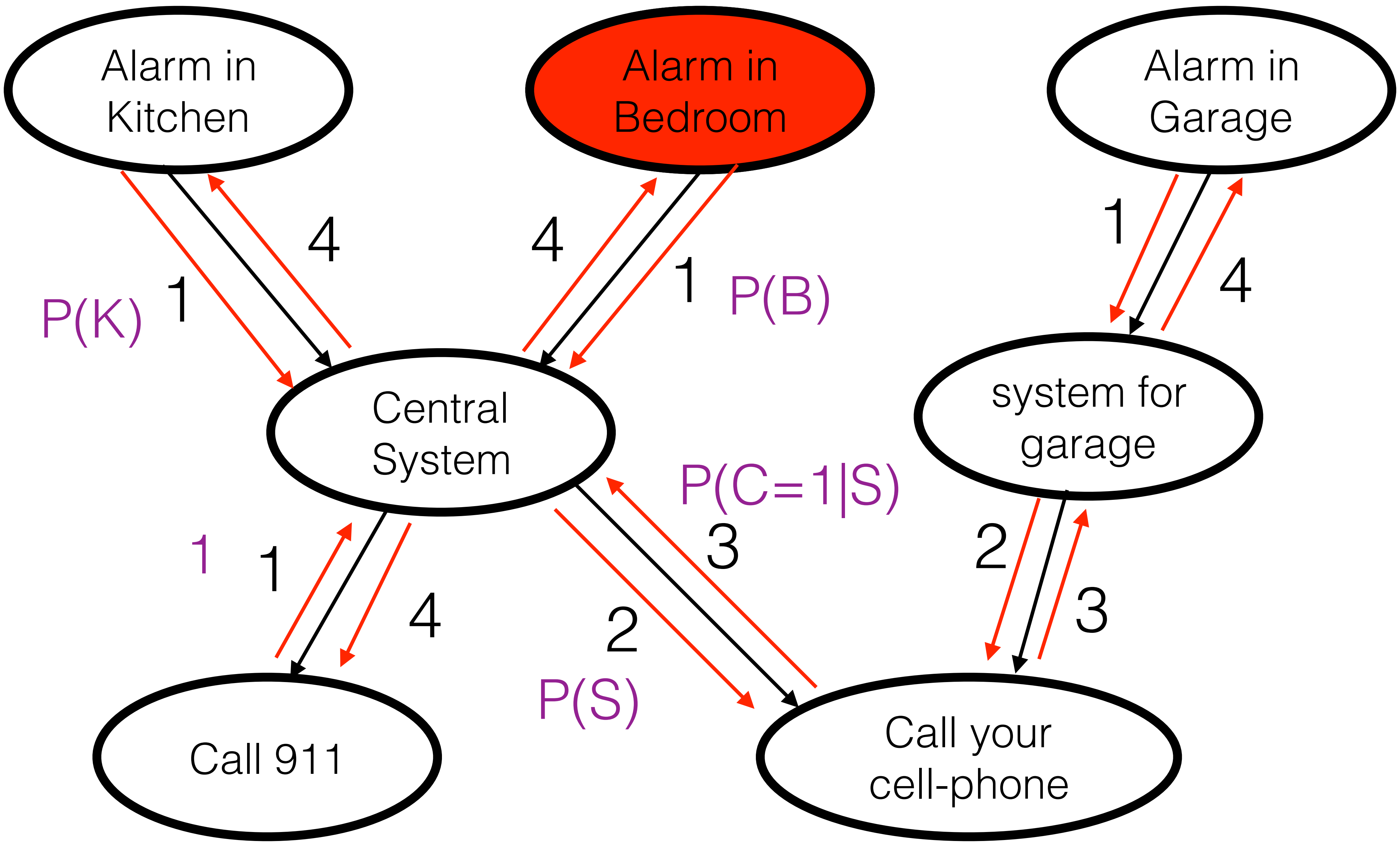
i=1,2 ,3



You receive phone call

Belief Propagation

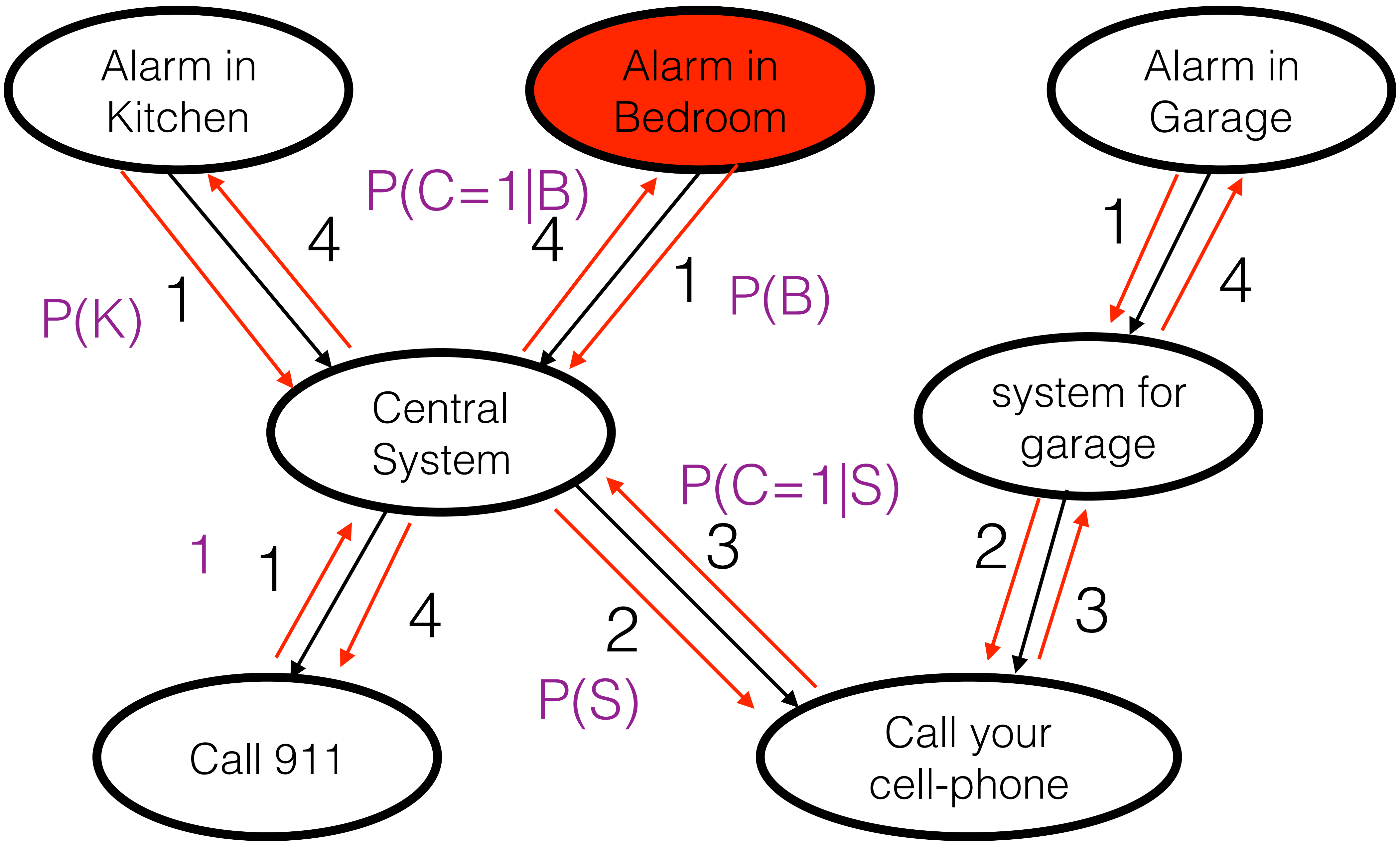
i=1,2 ,3



You receive phone call

Belief Propagation

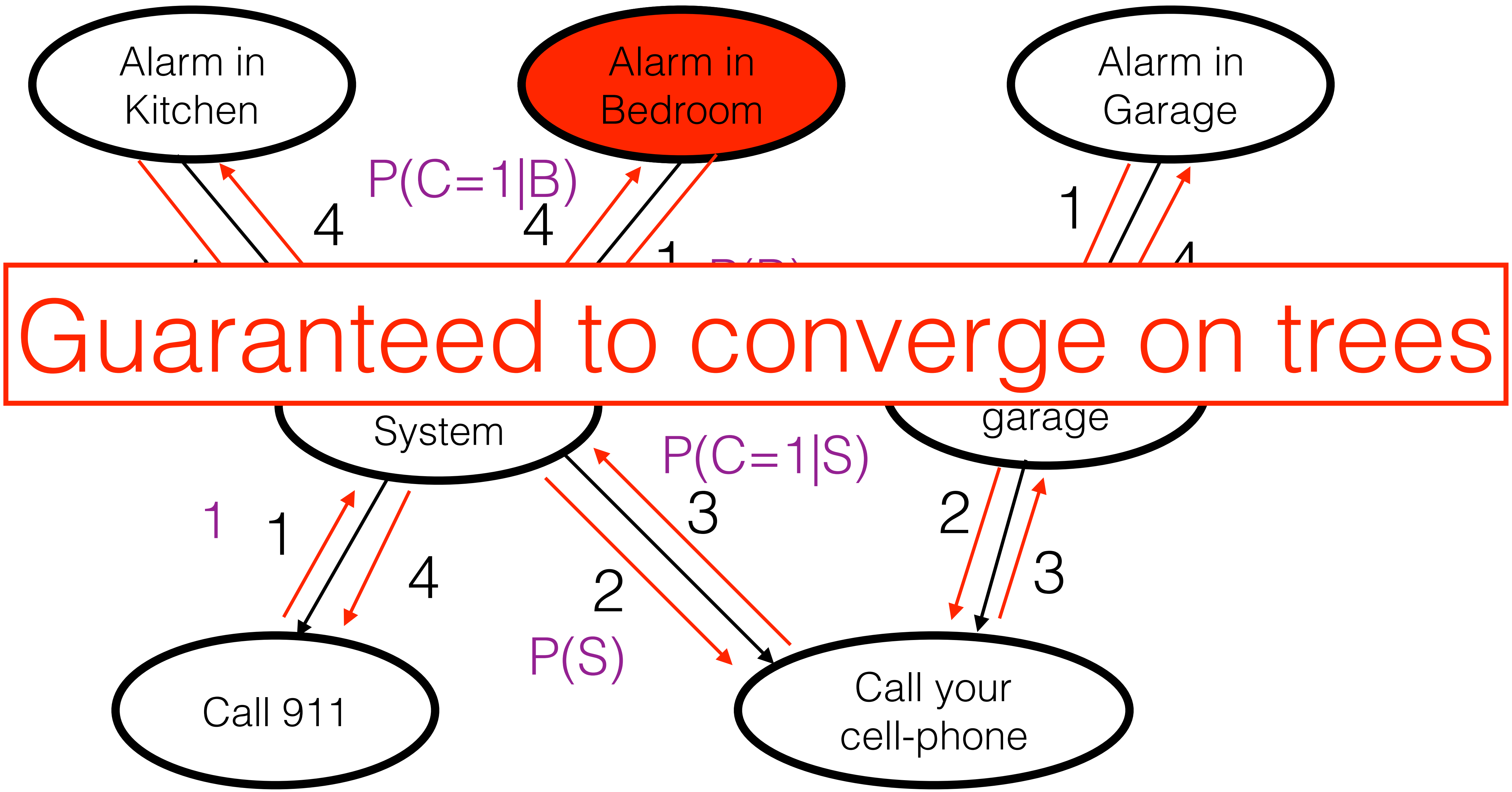
i=1,2 ,3



You receive phone call

Belief Propagation

i=1,2 ,3



You receive phone call

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 Baysian Network?

Parameter Estimation (Learning)

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

Parameter Estimation (Learning)

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

PARAMETER ESTIMATION (LEARNING)

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

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

Parameter Estimation (Learning)

- Simple case of finite outcomes

$$\theta_i^{MLE} = \text{empirical conditional probability table}$$

Parameter Estimation (Learning)

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

$$\arg\max_\theta \sum_{t=1}^{n} \log(P_\theta(X_1^t, \ldots, X_N^t))$$

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

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 $(X_1^1, \ldots, X_N^1), \ldots, (X_1^n, \ldots, X_N^n)$ where each $(X_1^t, \ldots, X_N^t)$ is drawn from the Bayesian network

Parameter Estimation (Learning)

- MLE: n independent samples $(X_1^1, \ldots, X_N^1), \ldots, (X_1^n, \ldots, X_N^n)$ where each $(X_1^t, \ldots, X_N^t)$ 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:

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

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:

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

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

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:

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

  - 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