Machine Learning for Data Science (CS4786)
Lecture 21

Inference in Bayesian Networks

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2017fa/

Announcement

- November 14th: Guest lecture by Prof. Kilian Weinberger

Bayesian Networks

- Directed acyclic graph (DAG): $G = (V, E)$
- Joint distribution P_θ over $X_1, \ldots, X_n$ that factorizes over G:

$$P_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{N} P_\theta(X_i | \text{Parent}(X_i))$$

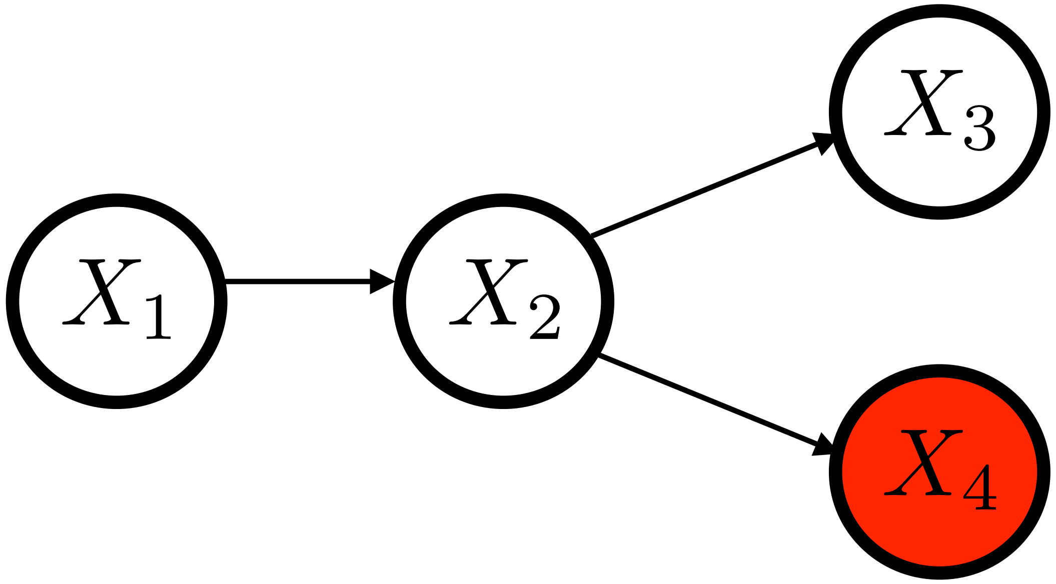
- Hence Bayesian Networks are specified by G along with CPD's over the variables (given their parents)

VARIABLE ELIMINATION: EXAMPLES

- Marginals are enough:

$$P(X_j = x_j, X_k = x_k | X_i = x_i, X_h = x_h) = \frac{P(X_j = x_j, X_k = x_k, X_i = x_i, X_h = x_h)}{P(X_i = x_i, X_h = x_h)}$$

Variable Elimination: Examples



P(Given variables) = Sum over all other variables (P(All variables))
= Sum over all other variables (Product P(Xi|Parents(Xi)))

$$P(X_4) = \sum_{x_1}\sum_{x_2}\sum_{x_3} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4)$$

$$= \sum_{x_1}\sum_{x_2}\sum_{x_3} \left(P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_2 = x_2) \cdot P(X_4 | X_2 = x_2)\right)$$

$$= \sum_{x_1} \left(P(X_1 = x_1) \sum_{x_2} \left(P(X_2 = x_2 | X_1 = x_1) P(X_4 | X_2 = x_2) \left(\sum_{x_3} P(X_3 = x_3 | X_2 = x_2) \right) \right) \right)$$

Variable Elimination: Bayesian Network

Initialize List with conditional probability distributions

Pick an order of elimination I for remaining variables

For each $X_i \in I$

Find distributions in List containing variable X_i and remove them

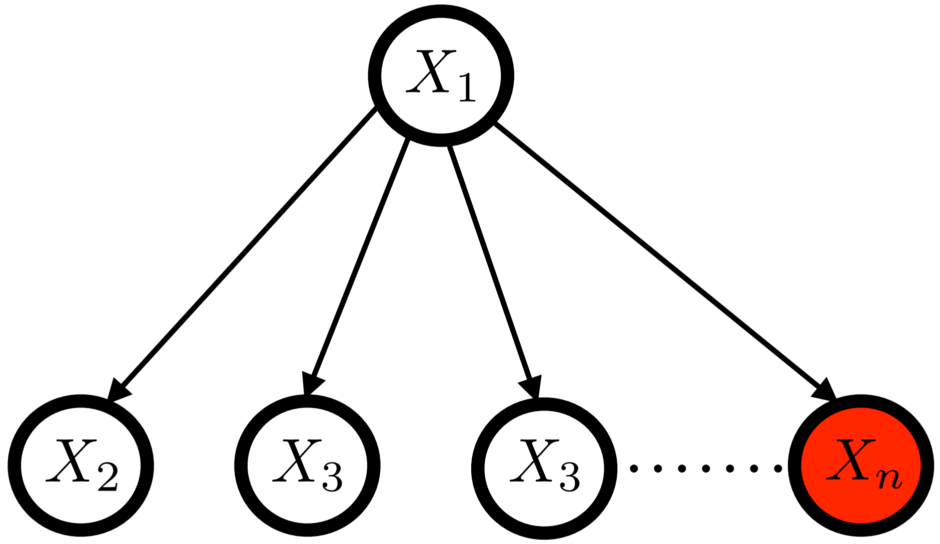
Define new distribution as the sum (over values of X_i) of the product of these distributions

Place the new distribution on List

End

Return List

Variable Elimination: Order Matters



List initialized to : $\{P(X_1), P(X_2|X_1), P(X_3|X_1), \ldots, P(X_n|X_1)\}$

Say $\mathbf{I} = (1,2,3,\ldots,n-1)$

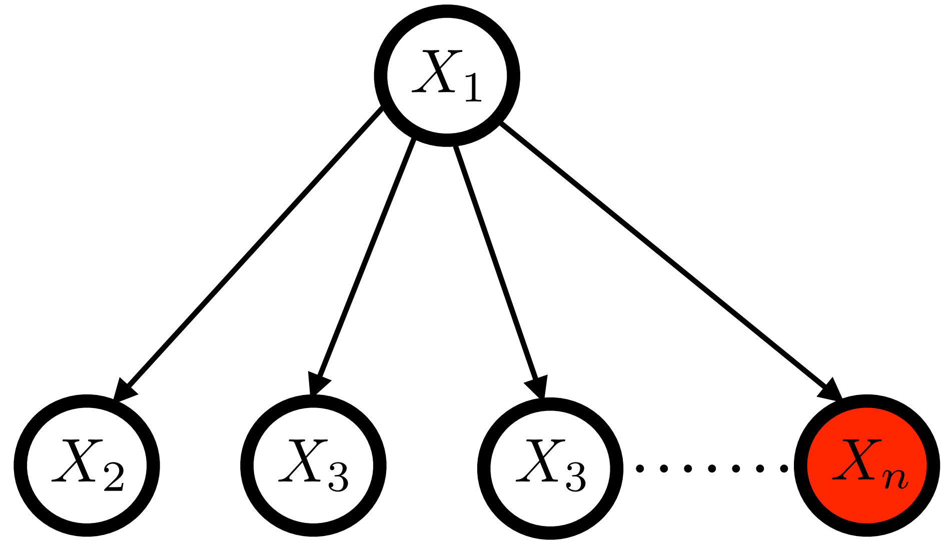
Iteration 1: Eliminate X_1

All terms in list involve X_1 so remove all of them

Replace them by table:

$$L_2(x_1,\ldots,x_n) = \sum_{x_1}\left(P(X_1 = x_1)\prod_{t=2}^{n} P(X_t = x_t|X_1 = x_1)\right)$$

Variable Elimination: Order Matters



List … X_1))}

Say

What is the size of table L1?

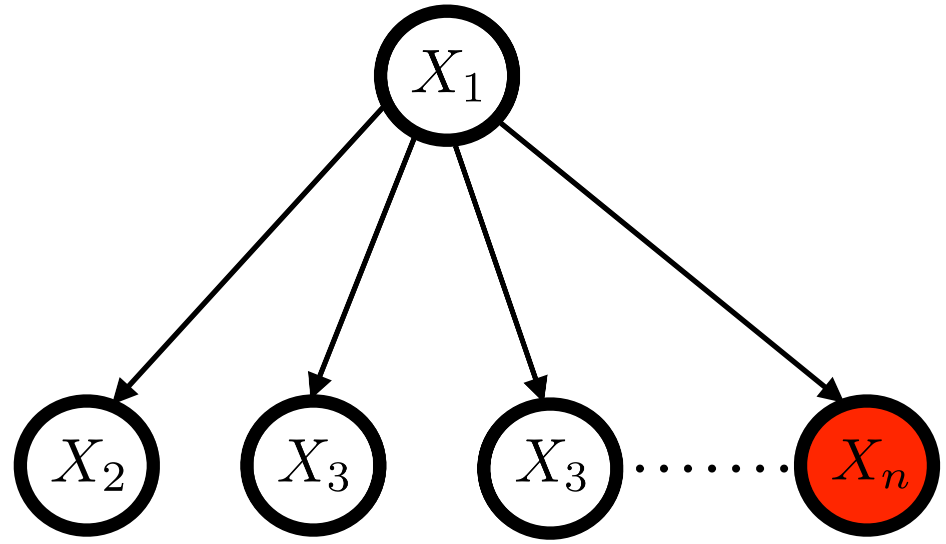
Iteration 1: Eliminate X_1

All terms in list involve X_1 so remove all of them

Replace them by table:

$$L_2(x_1, \ldots, x_n) = \sum_{x_1} \left(P(X_1 = x_1) \prod_{t=2}^{n} P(X_t = x_t | X_1 = x_1) \right)$$

VARIABLE ELIMINATION: ORDER MATTERS



List initialized to : $\{P(X_1), P(X_2|X_1), P(X_3|X_1),\ldots,P(X_n|X_1)\}$

Say $\mathbf{I}$ = (n-1,n-2,…,1)

Iteration 1: Eliminate X_{n-1}

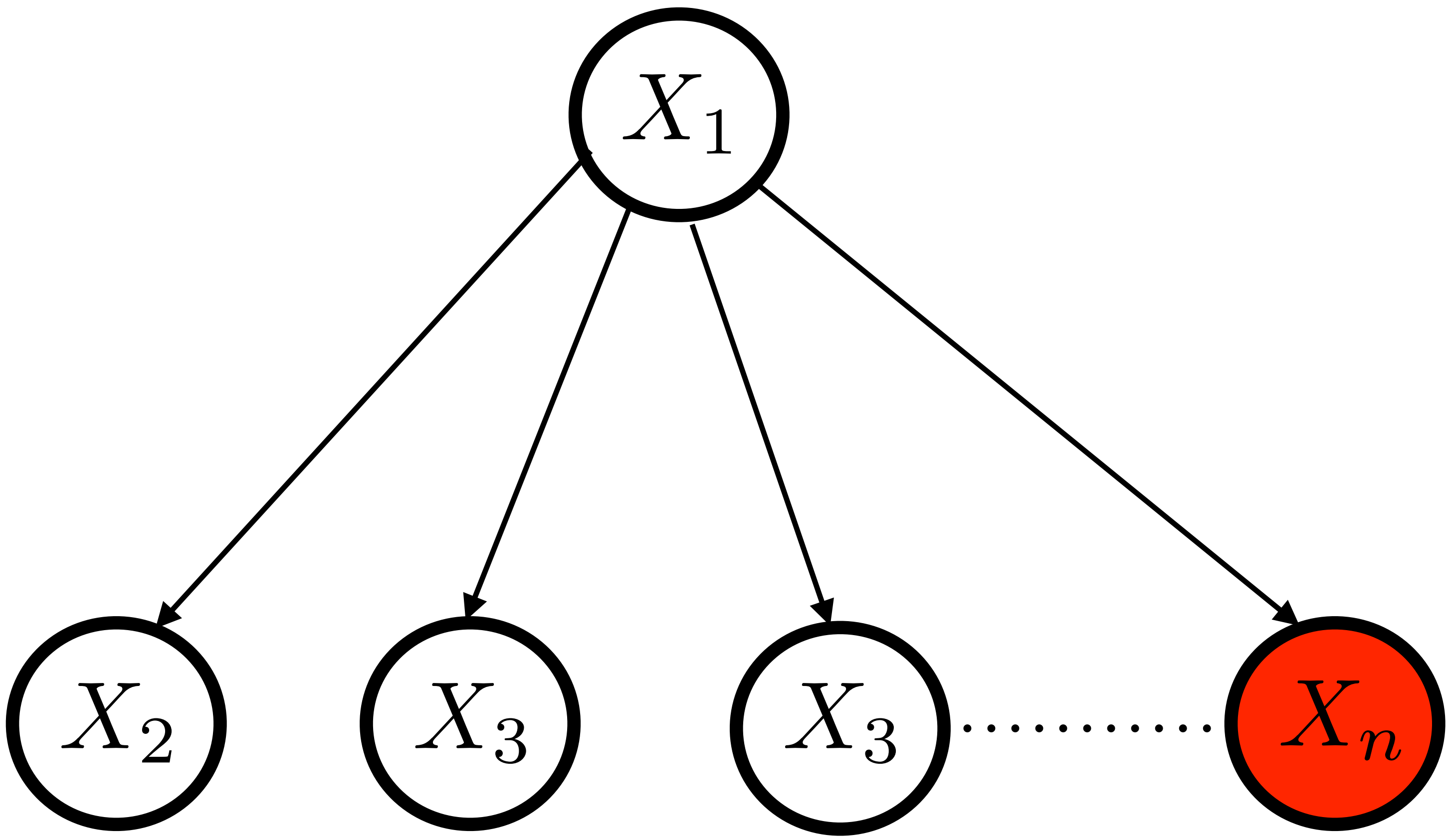
Remove $P(X_n|X_1)$ from List and replace by

$$L_{n-1}(x_1) = \sum_{x_{n-1}} P(X_{n-1}|X_1 = x_1) = 1$$

All the way up to X_2 we replace by all ones message

In then end we only have $P(X_1),P(X_n|X_1)$

Variable Elimination: Order Matters



Right order: O(n)

Wrong order: $O(2^n)$

Message Passing

- Often we need more than one marginal computation

- Can we exploit structure and compute these intermediate terms that can be reused?

Eg. forward backward algorithm for HMM

Belief Propagation

- This is specific to Bayesian networks
  - Messages to parents: belief about parent's value
  - Messages to Children: belief about self

Belief Propagation

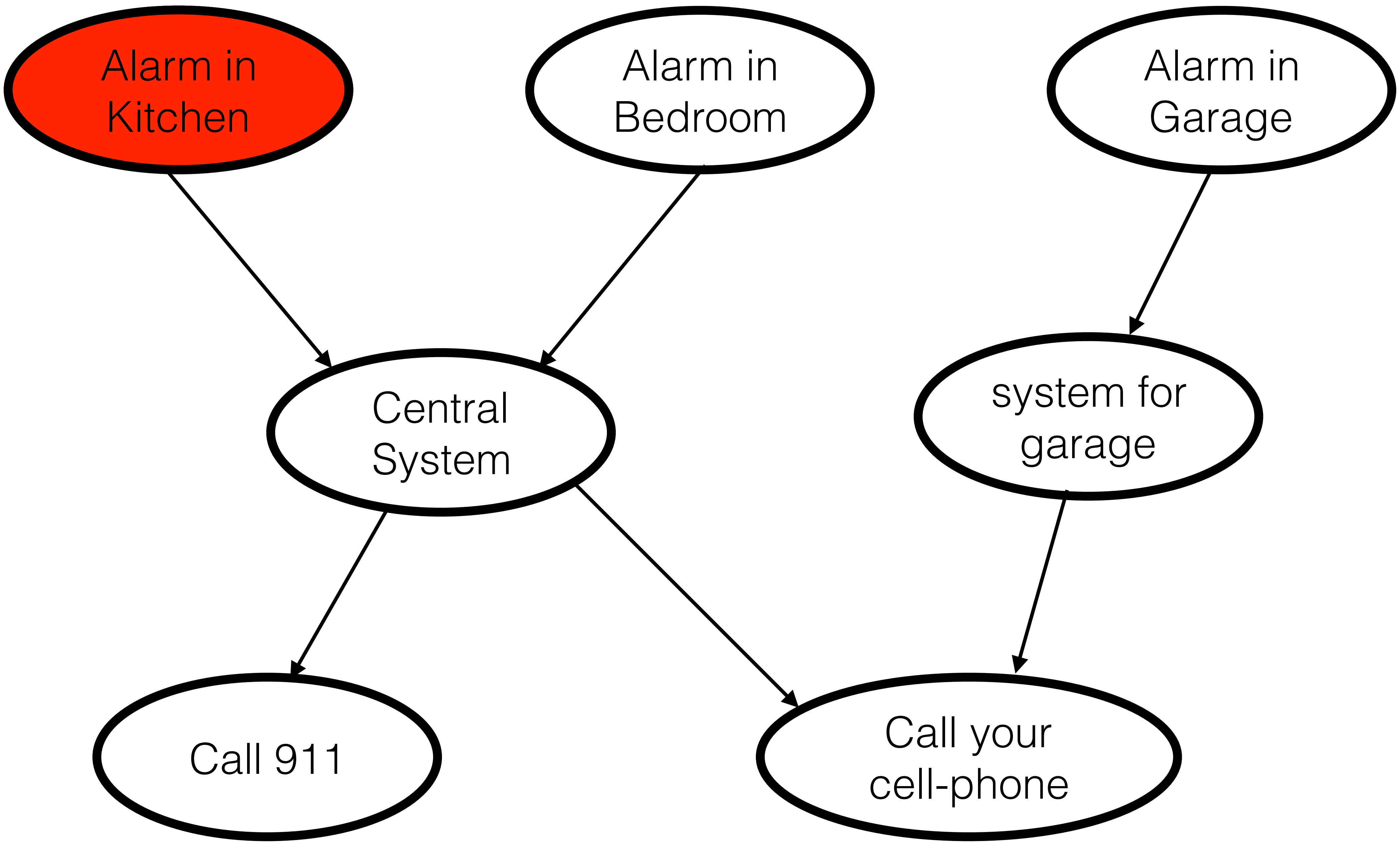
- Evidence E_i for node X_i can be seen as a priori belief of each node about itself.
  - if unobserved, set it to uniform distribution
  - if $X_i = x_i$ is observed,

    set $E_i(x_i) = 1$ and $E_i(x) = 0$ for any other x

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



You receives phone call

Belief Propagation

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

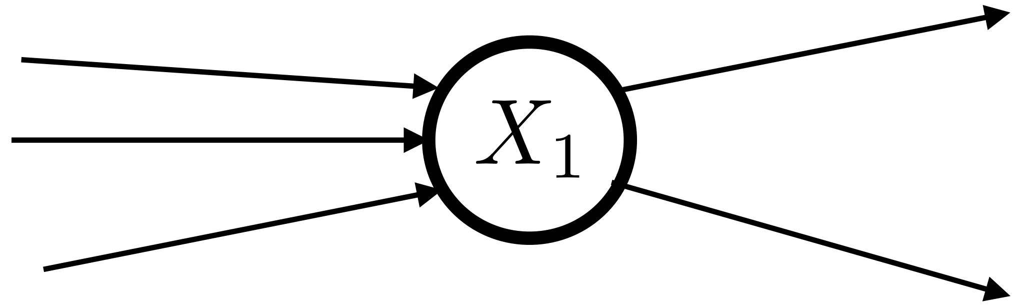
If X_j is the child of X_i :

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

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

Belief Propagation

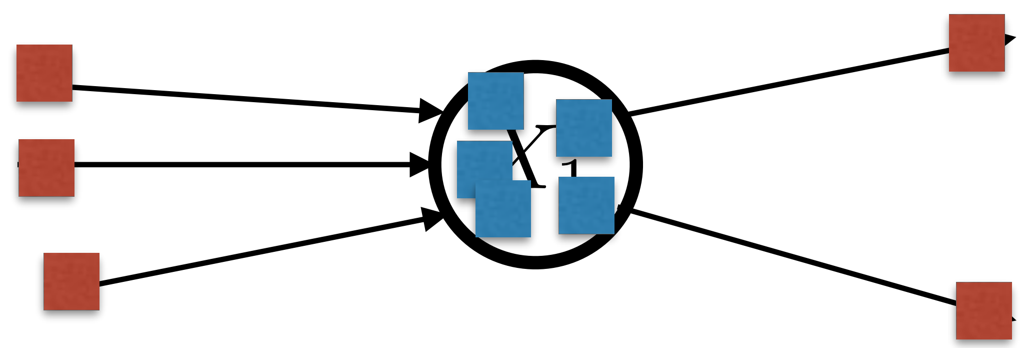
- On each round: Receive messages from previous round



Belief Propagation

- On each round: Receive messages from previous round

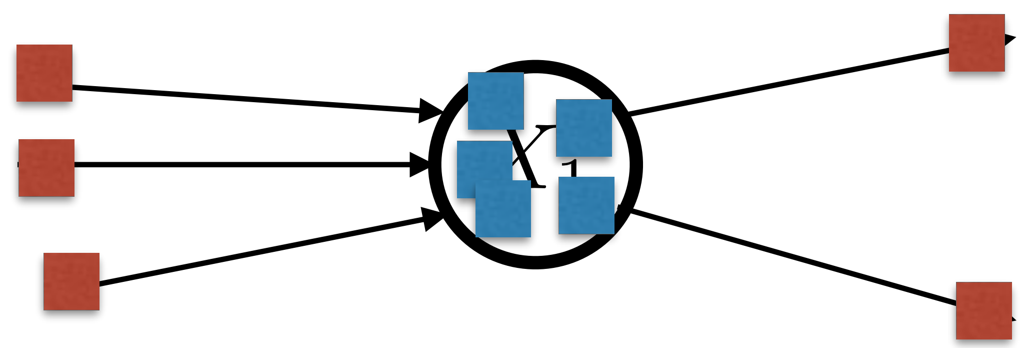
Round t



BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node X_i to Child X_k on round t

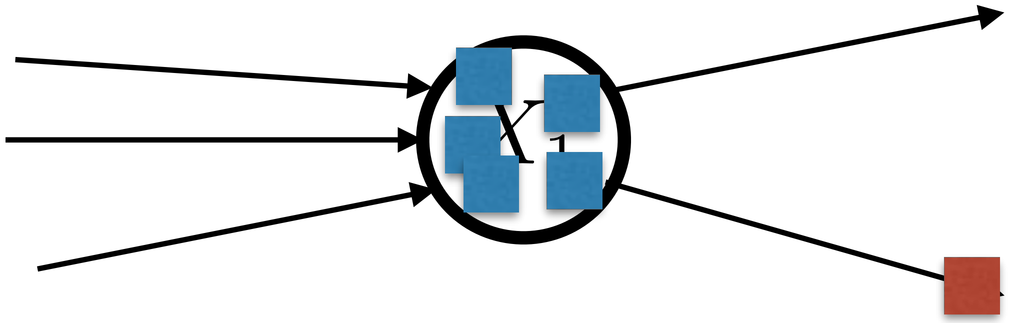
$$M^t_{i \mapsto k}(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i)) \text{ (product of all messages but one from } X_j)$$

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

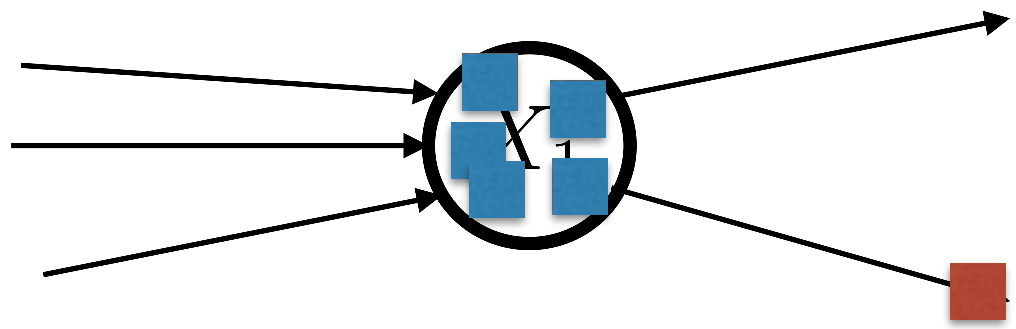
$$M^t_{i \mapsto k}(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i)) \; (\text{product of all messages but one from } X_j)$$

from previous round (t-1)

BELIEF PROPAGATION

- On each round: Receive messages from previous round

Round t



Message from node Xi to Child Xk on round t

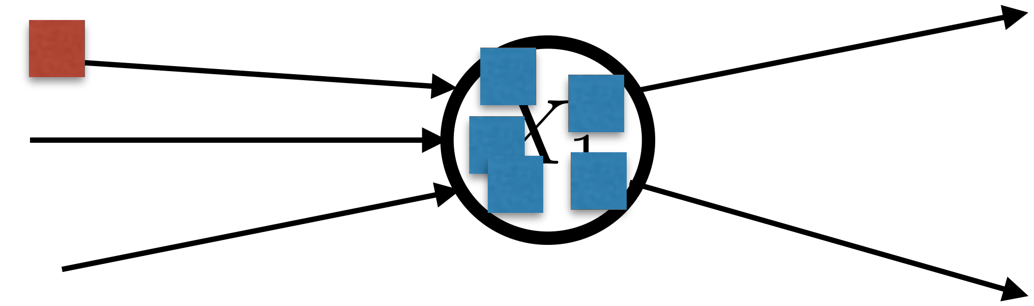
$$M^{t}_{i\mapsto k}(x_i) = \sum_{\text{Parents}(X_i)} E_{X_i}(x_i) P(X_i = x_i | \text{Parents}(X_i)) \text{ (product of all messages but one from } X_j)$$

from previous round (t-1)

Belief Propagation

- On each round: Receive messages from previous round

Round t



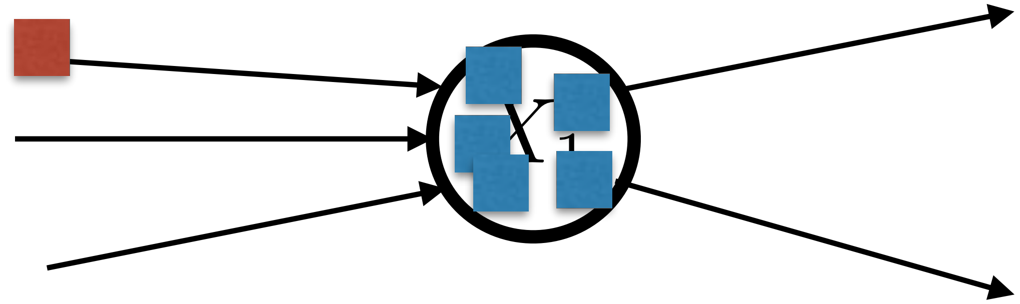
Message from node Xi to Parent Xj on round t

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

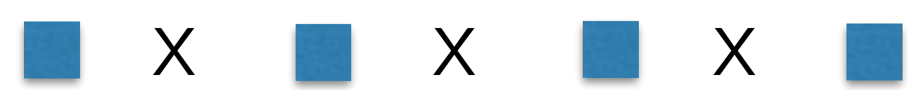
from previous round (t-1)

- On each round: Receive messages from previous round

Round t



Message from node Xi to Parent Xj on round t

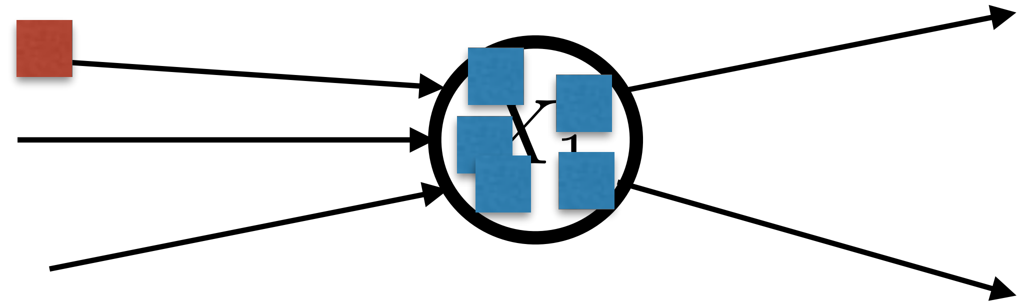


$$M_{i\mapsto j}(x_j) = \sum_{x_i,\text{Parents}(X_i)\setminus X_j} E_{X_i}(x_i)P(X_i = x_i|\text{Parents}(X_i))\ (\text{product of all messages but one from } X_j)$$

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

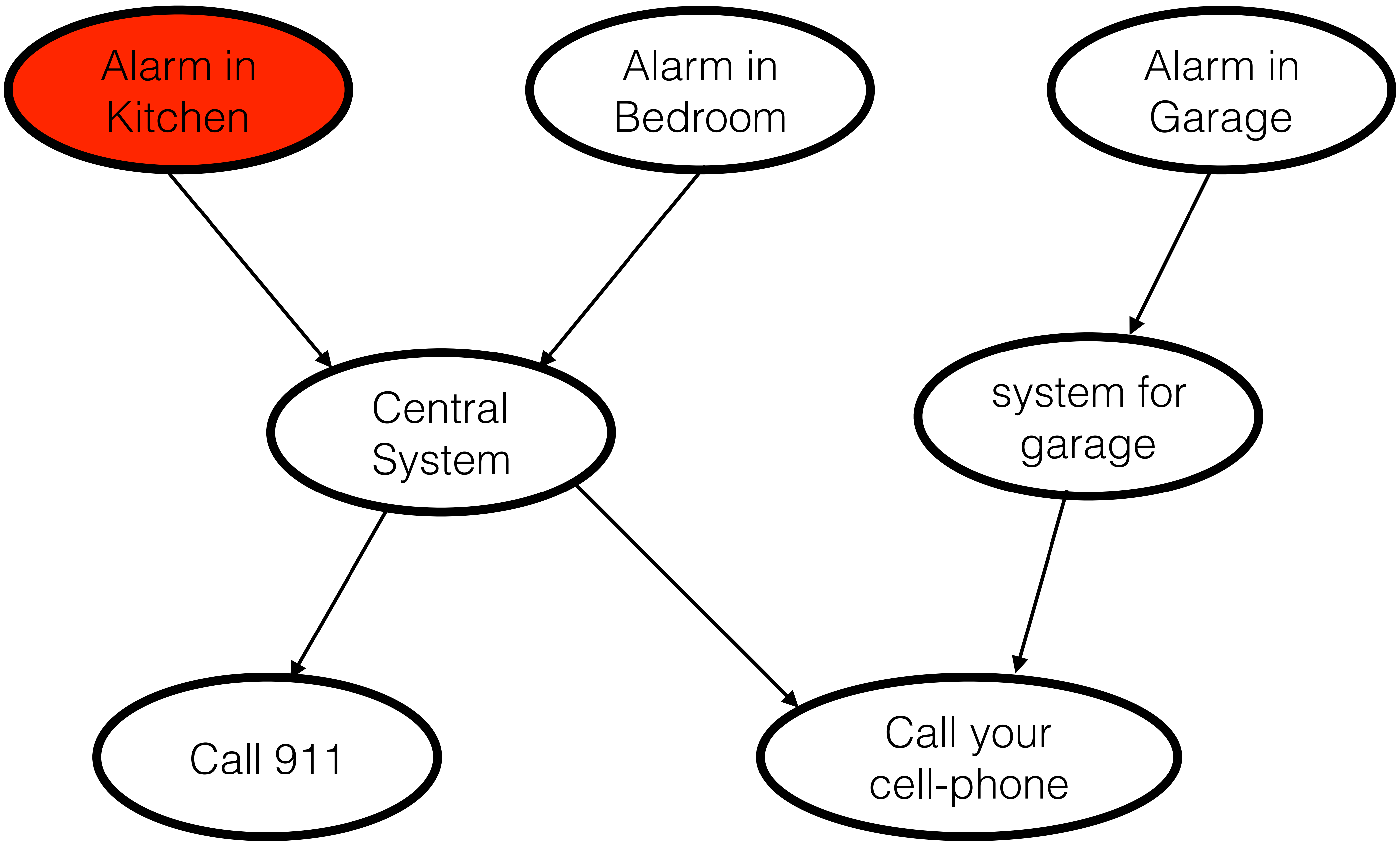
Only these vary over iterations

X X X

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

from previous round (t-1)

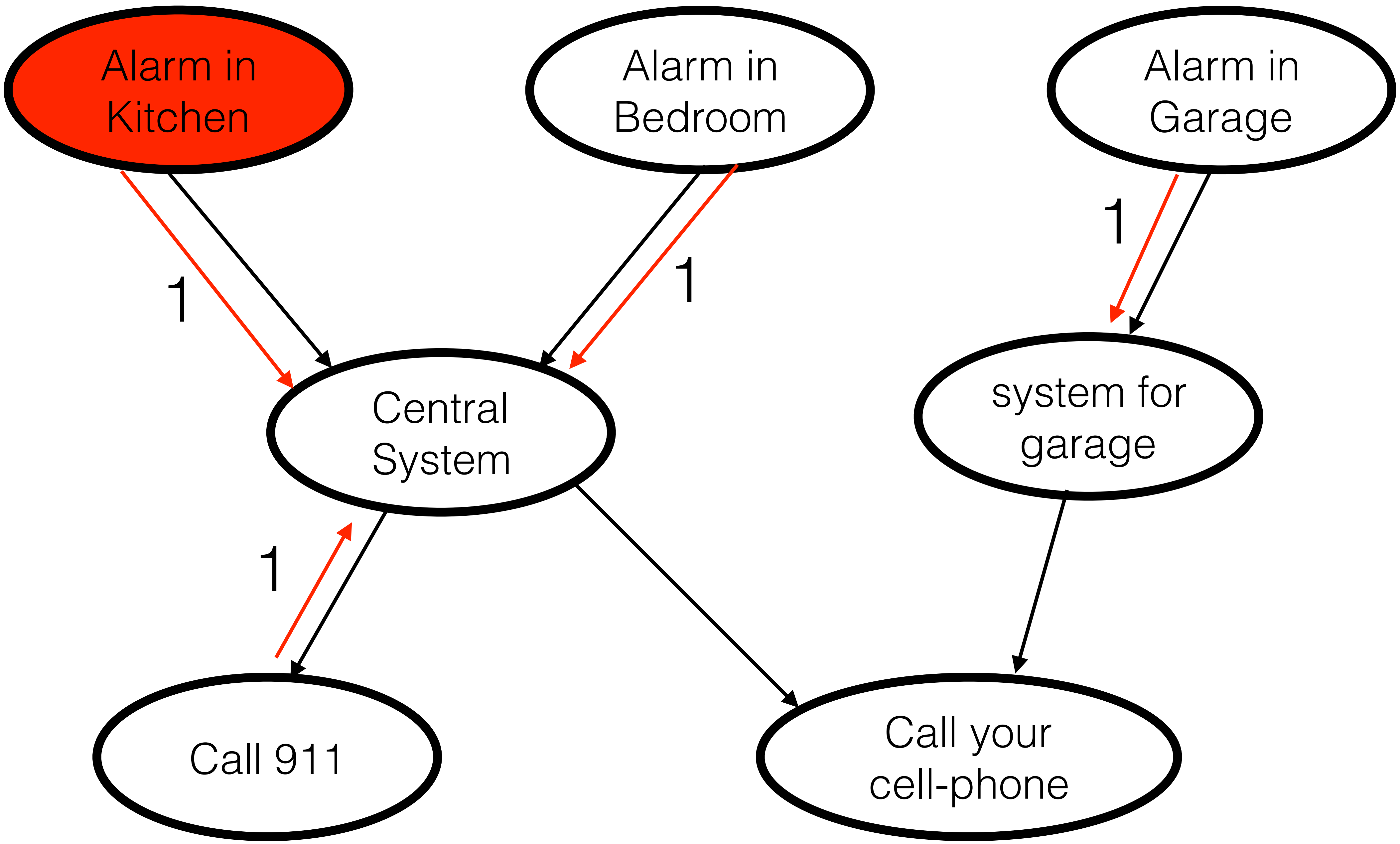
BELIEF PROPAGATION



You receive phone call

Belief Propagation

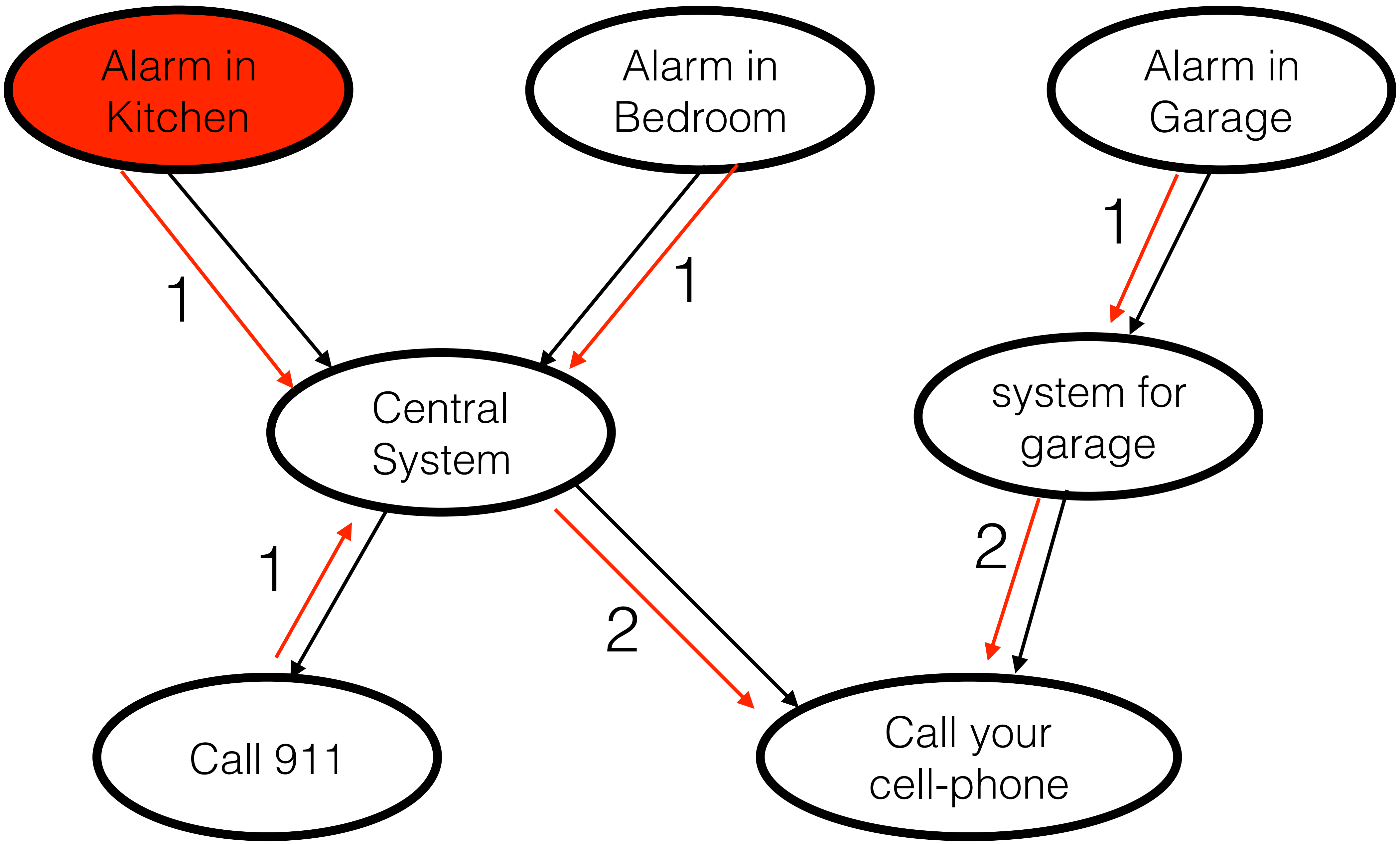
i=1



You receive phone call

Belief Propagation

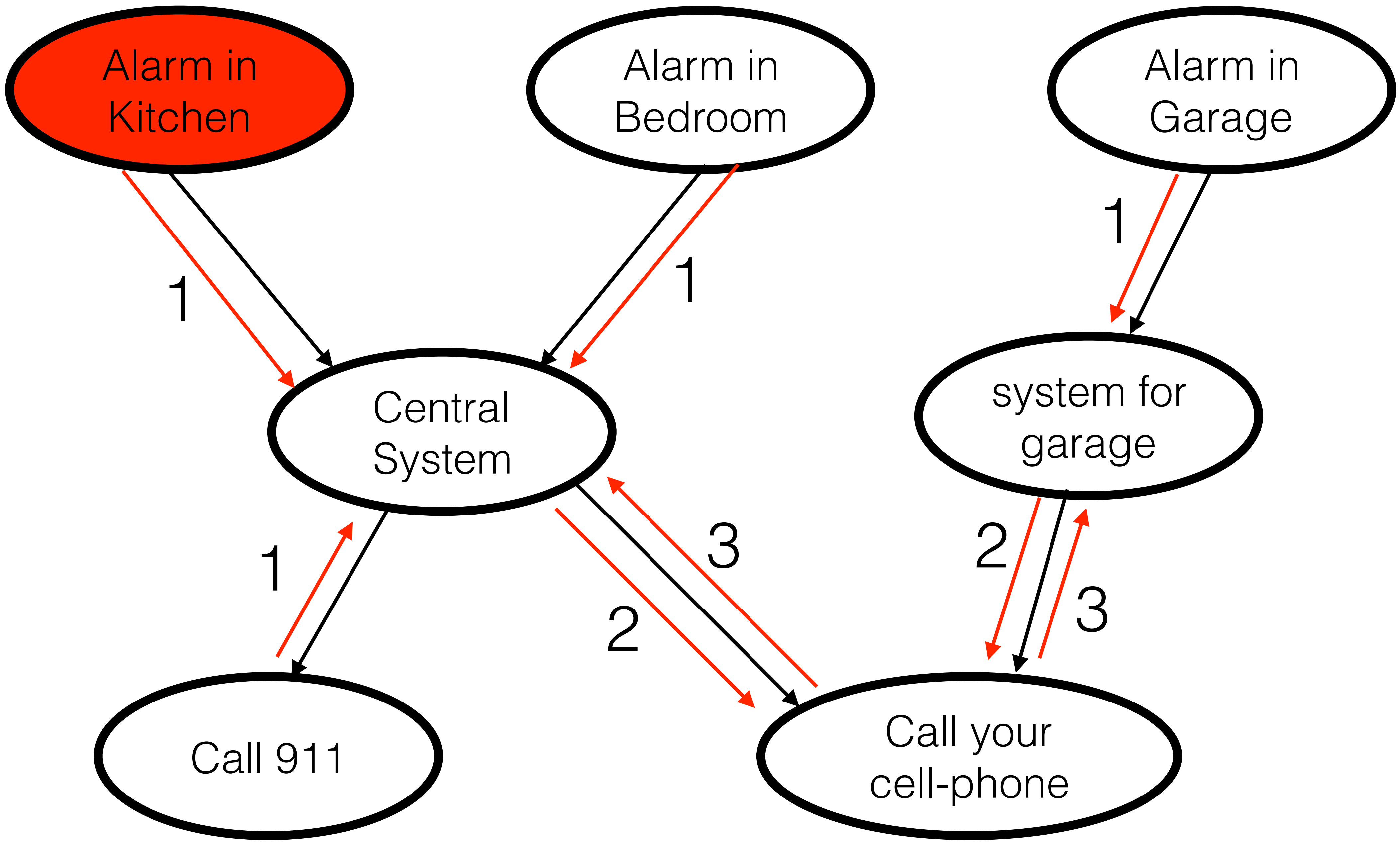
i=1,2



You receive phone call

Belief Propagation

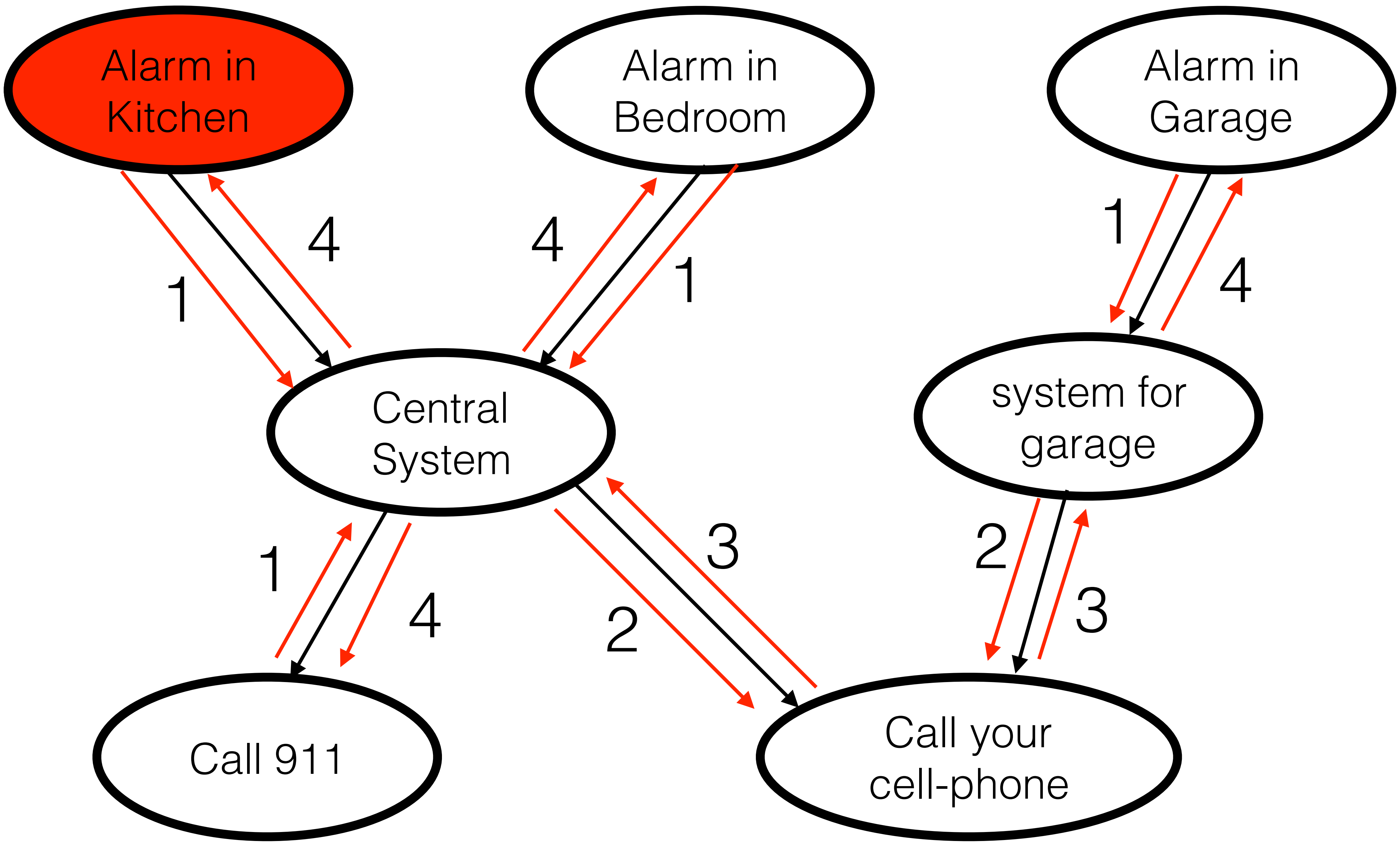
i=1,2 ,3



You receive phone call

Belief Propagation

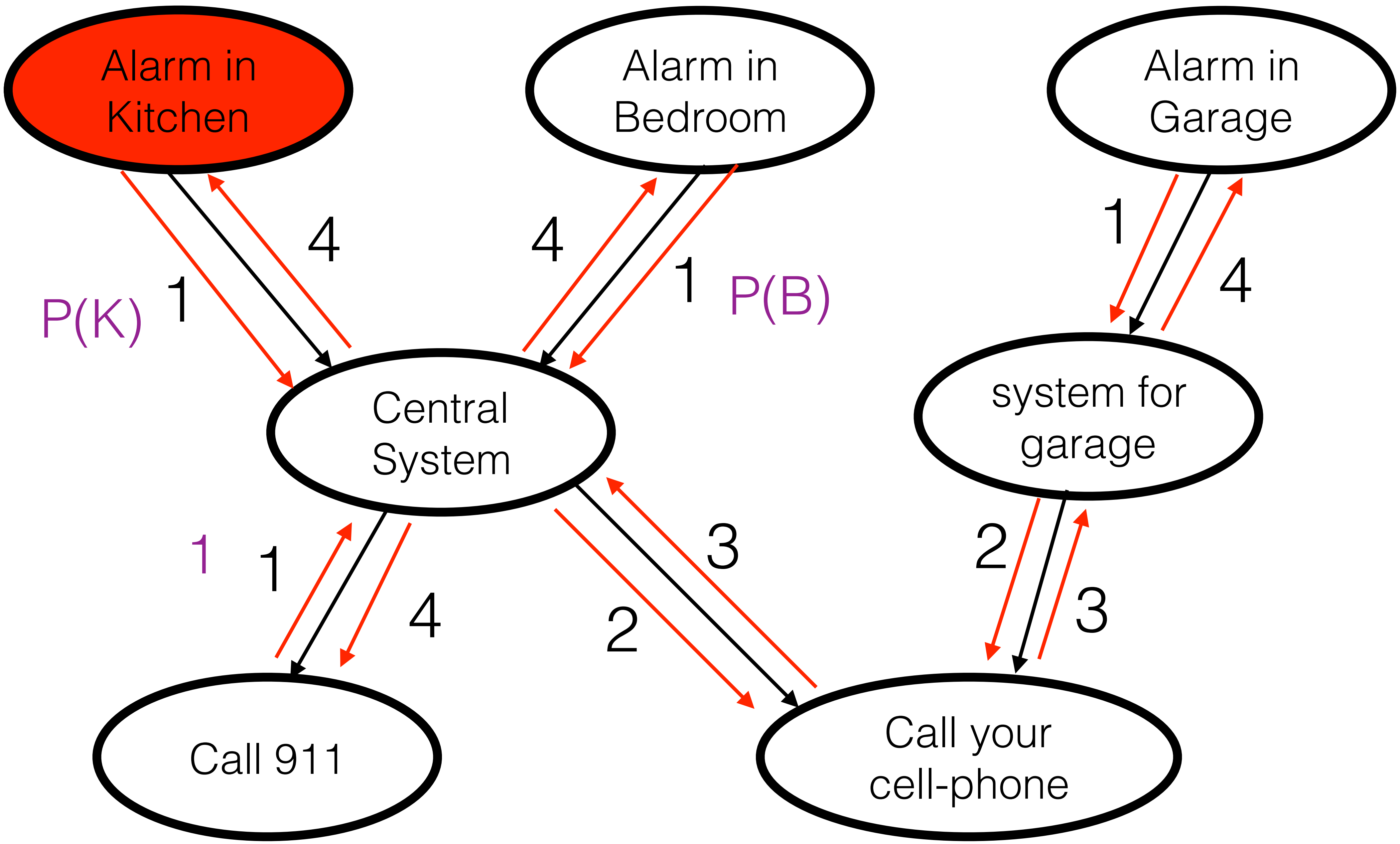
i=1,2 ,3



You receive phone call

BELIEF PROPAGATION

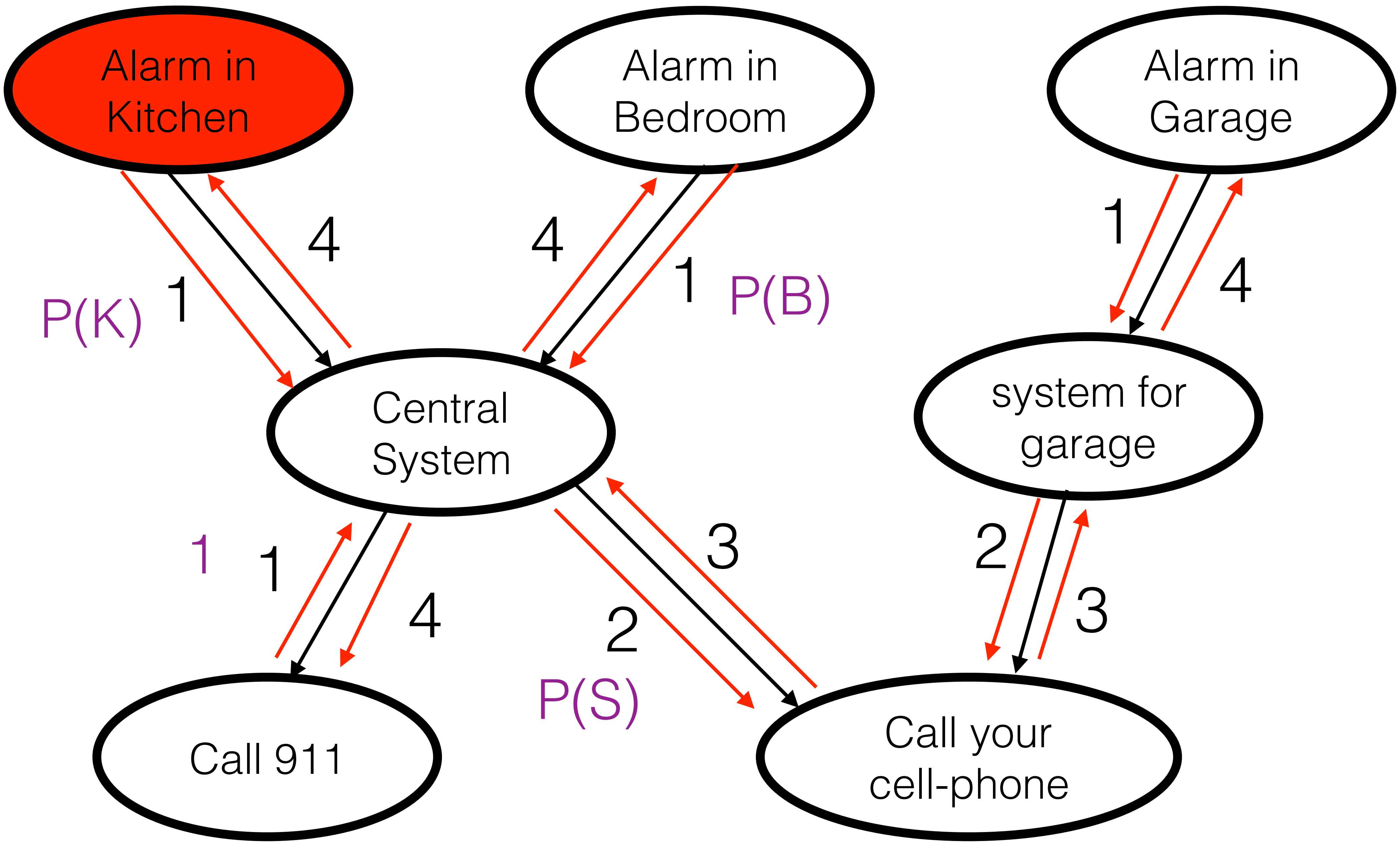
i=1,2 ,3



You receive phone call

Belief Propagation

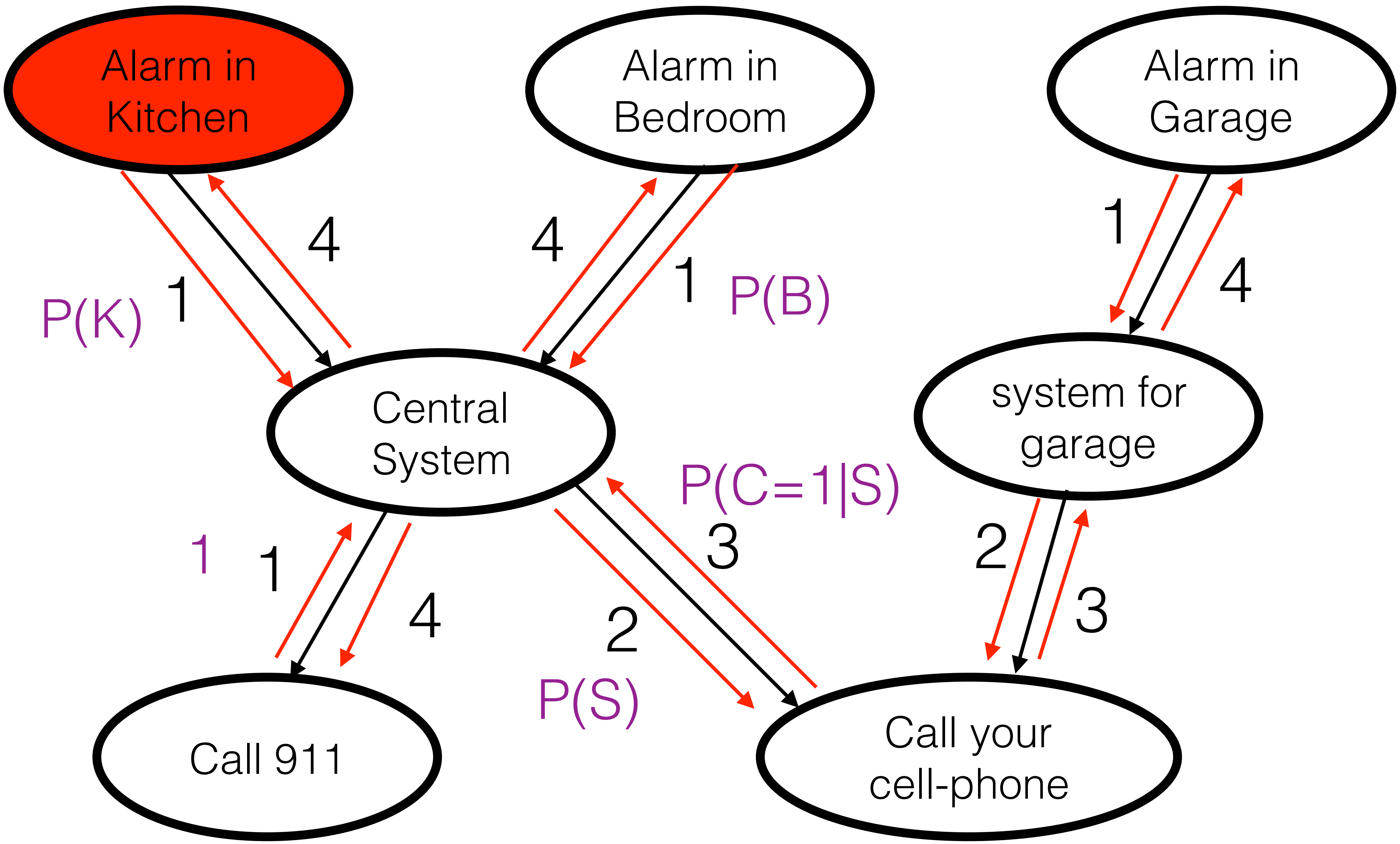
i=1,2 ,3



You receive phone call

BELIEF PROPAGATION

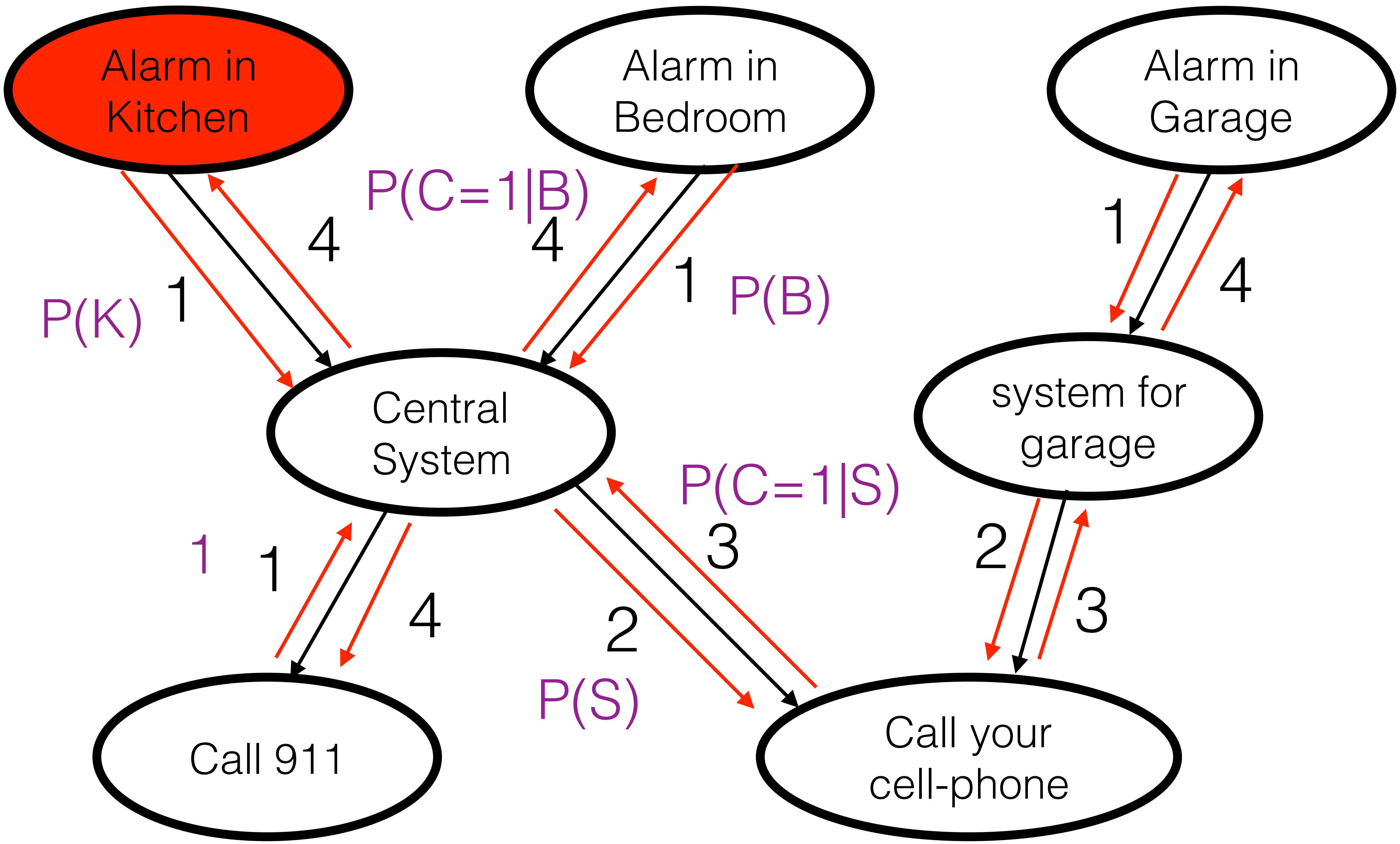
i=1,2 ,3



You receive phone call

Belief Propagation

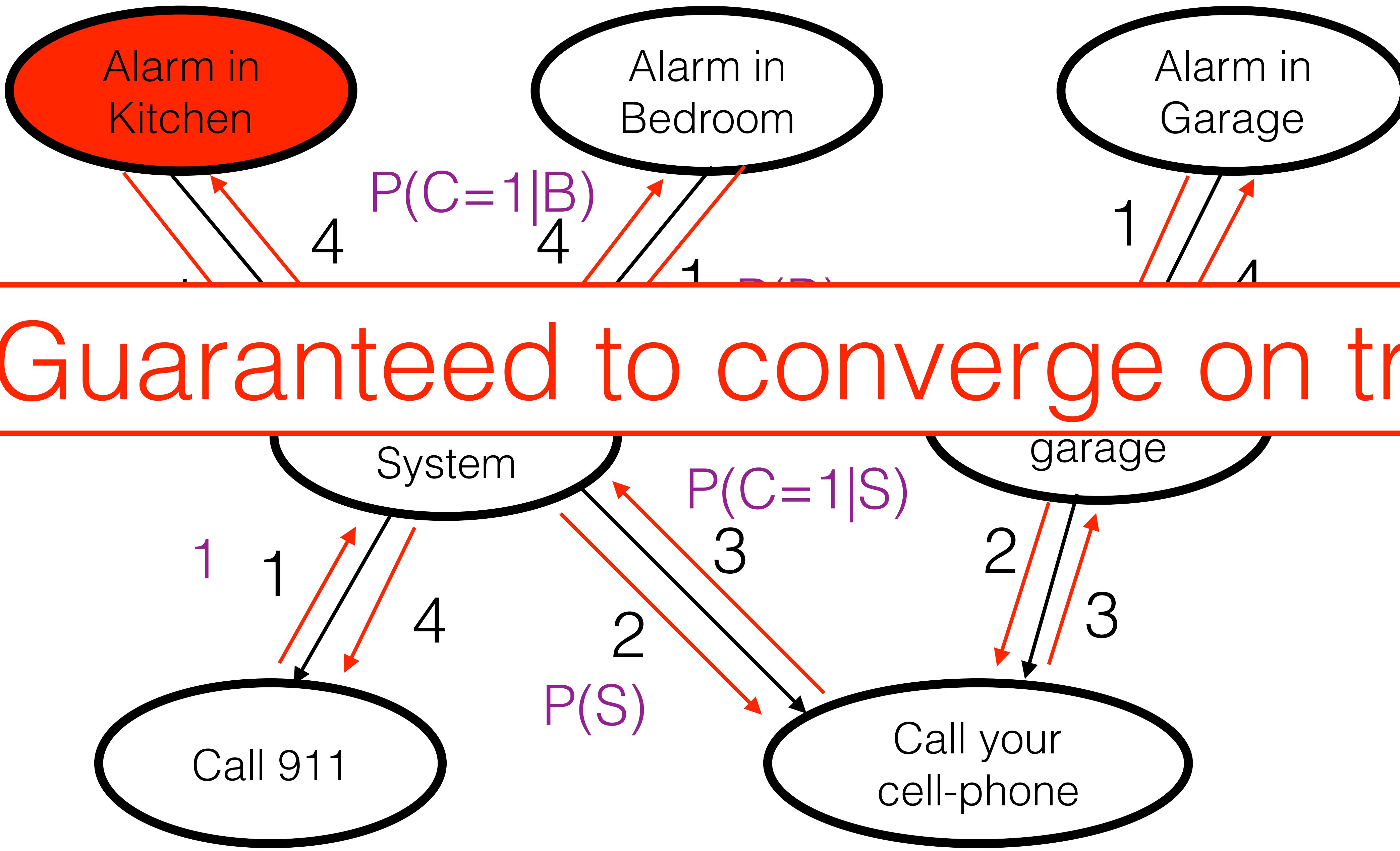
i=1,2 ,3



You receive phone call

Belief Propagation

i=1,2 ,3



Guaranteed to converge on trees

You receive phone call

Belief Propagation

After convergence:

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