Advanced Algorithms part 2

December 18, 2018

BROADCAST

Definition: Consider a distributed computing system where only one entity x, knows some important information; this entity would like to share this information with all the other entitites in the system. This problem is called broadcasting.

To solve this problem meand to design a set of rules that, when executed by the entities, will lead to a configuration where all entitieswill know the information; the solution must work without regardless of which entity has the information at the beginning.

Hence, for its solution, broadcasting requires connecctivity restriction (every entity has to be reachable from every other entity).

Cost: We will denote by M(Bcast) and T(Bcast) the message and time complexity.

The first message is sent by the initiator and has to reach all the entities.

Given d(a,b) the distance, i.e., the minimal number of edges between a and b, the ideal time depends on the time spent to go from the initiator to any node, thus the eccentrity or radius of G.

$$r(initiator) = Max\{d(x, y) : x, y \in V\}$$

Given that any node can be the initiator we have to compute the diameter of the graph

$$T(Bcast) = diameter(G) \le n-1 = O(n)$$

We consider m as the number of links. We could say that there will be ≤ 2 messages on each link. Let s be our initiator:

$$M(Bcast) = |N(s)| + \sum_{x \neq s} (|N(x)| - 1)$$
$$= \sum_{x} |N(x)| - \sum_{x \neq s} 1$$
$$= 2m - (n - 1)$$

Code: We have to consider three different states, INITIATOR, SLEEPING, DONE and two possible event, spuntaneous impulse event and receving the message.

```
if INITIATOR
        spontaneous impulse event
                send(I) to N(X)
                become (DONE)
        receving(I)
                do-nothing
if SLEEPING
        receving(I)
                send(I) to N(x) - \{sender\}
                become (DONE)
        spontaneous impulse event
                do-nothing
if DONE
        receving(I)
                do-nothing
        spontaneous impulse event
                do-nothing
```

SPANNING TREE

A spanning tree T of a graph G = (V, E) is an acyclic subgraph of G such that T' = (V, E') and $E' \subset E$.

PROTOCOL SHOUT

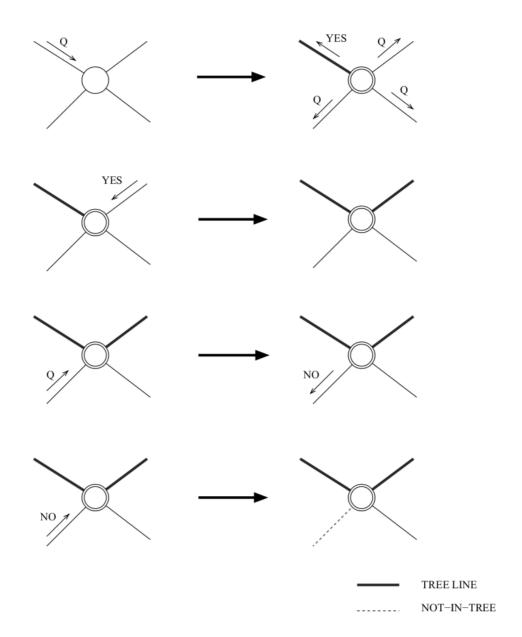
Assumption: There is a single initiator, there are bidirectional links, total reliability and G is connected.

Definition: We want to build a spanning tree in a distributed system, but we have to consider that the entities do not know G, not even its size. The only things an entity is aware of are the labels on the ports leading to its neighbors and the fact that, if it sends a message to a neighbor, the message will eventually be received.

Solution: Ask-Your-Neighbors

• The initiator s will ask its neighbors; it will send a message Q = ('Are you my neighbor in the spanning tree?') to all its neighbors.

- An entity $x \neq s$ will reply 'Yes' only the first time it is asked and, in this occasion, it wull ask all its other neighbors; otherwise it will reply 'No'.
- Each entity terminates when it has received a reply from all neighbors to which it asked the question.



Example of SHOUT

Code: We have four states, INITIATOR, IDLE, ACTIVE, DONE; the terminal state is DONE.

```
if INITIATOR
        spontaneous event
                root = true
                tree-neighbours = {}
                send(Q) to N(x)
                counter = 0
                become ACTIVE
if IDLE
        receving(Q)
                root = false
                parent = sender
                tree-neighbours = {sender}
                send('Yes') to sender
                counter = 1
                //managing leaves and intermediate node cases
                if counter = |N(x)| then
                         become (DONE)
                else
                         send (Q) to N(x) - \{sender\}
                         become ACTIVE
if ACTIVE
        receving(Q)
                send ('No') to sender
        receving('Yes')
                tree-neighbours = tree-neighbours ∪ sender
                counter += 1
                if counter = |N(x)|
                         become (DONE)
        receving('No')
                counter += 1
                if counter = |N(x)|
                         become (DONE)
```

Complexity: To compute the total amount of messages sent in this protocol we have to consider the messages sent by the initiator or the other sent to the neioghbours, the amount of 'Yes' and 'No' answer.

$$M(SHOUT) = Q + NO + YES$$
$$= (n-1) + 2(m - (n-1)) + 2(m - (n-1)) + (n-1)$$
$$= 4m - 2n + 2 = 2(2m - n + 1)$$

We could improve this solution but how? Do we really need the 'No' message?

SHOUT+

The code is almost the same, but we do not need to send the 'No' message. But how can we manage that?

What we have to change is the following:

```
 \begin{array}{c} \text{if ACTIVE} \\ \text{receving}\left(Q\right) \\ \text{counter += 1} \\ \text{if counter = } \left|N(x)\right| \\ \text{become}\left(DONE\right) \end{array}
```

Complexity: The messages are much less.

```
M(SHOUT+) = 2m
```

when

$$M(SHOUT) = 2(2m-n+1)$$

LEADER ELECTION

Definition: Moving the system from an initial configuration where all entities are in the same state (called avaiable), into a final configuration where all entities are in the same state (called follower), except one which is in a different state (called leader).

No restriction, we assume a bidirectional graph, connected and no failures.

SATURATION TECHNIQUE

This technique is used to solve many other problems, independently from the number and positions of the initiators (multiple initiators).

The states are avaiable, awake and processing. the algorithm can be divided in three parts:

- Activation phase: started by all iterators, wake-up, all nodes are activated
- Saturation phase: started by the leaves, a unique pair of neighbours is identified(saturated node), the other become processing.
- Resolution phase: started by the saturated nodes

Code: We have four states AVAILABLE, ACTIVE, PROCESSING, SATURATED. The initial state is AVAILABLE.

```
if \ AVAILABLE \ (I \ haven't \ been \ activated \ yet) spontaneously send('Activate') \ to \ N(x) neighbours = N(x) if \ |neighbours| = 1 \ //leaf \ case
```

```
parent = neighbours
                         send('Saturation') to parent
                         become (PROCESSING)
                 else
                         become (ACTIVE)
                receving('Activate')
                         send('Activate') to N(x) - \{sender\}
                         neighbours = N(x)
                         if |neighbours| = 1
                                 parent = neighbours
                                 send('Saturation') to parent
                                 become (PROCESSING)
                         else
                                 become (ACTIVE)
if ACTIVE
        receving (M)
        neighbours = neighbours - {sender}
        if |neighbours| = 1
                         parent = neighbours
                         send('Saturation') to parent
                         become (PROCESSING)
if PROCESSING
        receving (M)
                become (SATURATED)
```

Complexity:

$$M(Saturation) = 2n - 2 + n + n - 2 = 4(n-1)$$

But how this technique can be useful for our election problem?

SOLUTION TO THE ELECTION PROBLEM

Let us assume we assign a label v(x) to each node, then it's possible to:

- Run the saturation technique.
- The two saturated nodes exchange the value and if it is different they can elect as a leader the one with smaller value.

Observations:

- If nodes do not have distinct labels the leader cannot be elected.
- Each saturated node knows where the other one is, since it knows from which edge it has received the last message.

MINIMUM FINDING WITH SATURATION TECHNIQUE

Each entity x has in input its value (value(x)). At the end each entity should know whether it is the minimum or not.

The states are AVAILABLE, ACTIVE, PROCESSING, MINIMUM, LARGE. The starting state is AVAILABLE.

Code: The states are AVAILABLE, ACTIVE, PROCESSING, MINIMUM, LARGE. The starting state is AVAILABLE.

```
if AVAILABLE
        spontaneously
                send('Activate') to N(x)
                min = v(x)
                neighbours = N(x)
                if |neigbours| = 1 //leaf
                        M = ('Saturation', min)
                         parent = neighbours
                         send (M) to parent
                         become (PROCESSING)
                else
                         become (ACTIVE)
        receving('Activate')
                send('Activate') to N(x) - \{sender\}
                min = v(x)
                neighbours = N(x)
                if |neighbours| = 1
                                         //leaf
                        M = ('Saturation', min)
                         parent = neighbours
                         send (M) to parent
                         become (PROCESSING)
                else
                         become (ACTIVE)
if ACTIVE
        receving (M)
                min = min\{min, M\}
                neighbours = neighbours - {sender}
                if |neighbours| = 1 //leaf
                        M = ('Saturation', min)
                         parent = neighbours
                         send(M) to parent
                         become (PROCESSING)
```

if PROCESSING

Complexity: M(minimum) = 2(n-1) + n + n - 2 where 2(n-1) is the cost for the activation; n the cost for saturation and n - 2 the cost for notification.

$$M(minimum) = 4n - 4$$

RANKING PROBLEM IN ARBITRARY NETWORKS

Definition: Given a graph G(V, E) rank it from the lowest value to the largest. **Solution:** In an arbitrary network, this problem could be solved in the following way:

- Find a spanning tree
- Use saturation and the minimum finding to find a starting node
- Do ranking (Centralized or De-centralized)

CENTRALIZED RANKING

This problem could be solved in the following way:

- Build a spanning tree using SHOUT algorithm
- We elect a leader(minimum) using saturation technique
- The leader knows the minimum, it sends in that direction a ranking message.
- Every node knows the minimum in its subtrees, it can then forward the ranking message (ranking, minimum) in the right direction.
- When the node to be ranked receives the message it sends up a notification and update message(new-minimum) that will travel up to the leader.

Complexity:

$$M(rank_centr) = 2(n-1) + 2(n-2) + ... + 2(1) = n(n-1)$$

DE-CENTRALIZED RANKING

Solution: The starter node sends a ranking message of the form: (first, second, rank) in the direction of the first. The first has the smallest value, the second the second smallest known so far. If no value is indicated (or the value is ∞) it means that the smallest in the corresponding subtree is unknown(for the moment).

The ranked node attempts to send a ranking message to the next node to be ranked. The second variable of the rank message is updated during its travel and the minimum values on the links of the tree are also updated.

Complexity:

$$M(rank_decentr) = 2(n-1) + (n-2) + (n-3) + ... + 1 = (n-1)n/2 + (n-1) = (n-1)(n/2+1)$$

Code:

- Build a spanning tree using SHOUT
- We elect a leader (e.i. minimum using saturation technique)
- Jump from the min to the second one and so one. After having elected the minimum, we know the second smallest element and so we send a message with (secondValue, ∞ , 2nd), this message travel across the spanning tree to find the third element. After the second recives the message put the value of the subtree as ∞ because there is nothing to rank in that subtree and so on.

ELECTION PROBLEM

The election problem cannot be generally solved if the entities do not have different identities.

Consider the system where:

- Unique entities
- · Same state
- Anonymous
- Synchronous

At each moment, they are doing the same thing, Receive the same messages, move to the same state.

Election in trees

To each node x is associated a distinct identifier v(x). A simple algorithm:

- Execute the saturation technique
- · Choose the saturated node holding the minimum value

Election in rings A ring is a sparse network (few arcs) such as trres but it's not symmetric. **Solution:** Each id is fully circulated in the ring, in few entity sees all identities. Assumptions:

- Two versions: unidirectional and bidirectionals
- Local orientation
- Distinct identities

All the entities can be the initiator. The bidirectional version assumes that it is oriented. The protocol works even without knowing the dimension of the ring and having messages that arrive in a non-FIFO order. Each node sends it value, if it receives a message it forwards it and it keeps track of the minimum value seen so far.

Code: The states are ASLEEP, AWAKE, FOLLOWER, LEADER; the starting state is ASLEEP and the terminal states are FOLLOWER and LEADER.

```
INITIALIZE
        count = 0
        size = 1
        known = false
        send('Election', id(x), size) to right
        min = id(x)
if ASLEEP
        Spontaneously
                INITIALIZE
                become (AWAKE)
        receving ('Election', value, counter)
                INITIALIZE
                send('Election', value, counter+1) to other
                min = min {min, value}
                count = count + 1
                become (AWAKE)
if AWAKE
        receving ('Election', value, counter)
                 if value \neq id(x)
                         send('Election', velue, counter+1) to other
                         min = min{min, value}
                         count+=1
                         if known
                                 become (CHECK)
                         else
                                  ringsize = counter
                                 known = true
                                 become (CHECK)
if CHECK
        if count = ringsize
                 if min = id(x)
                         become (LEADER)
                 else
```

become (FOLLOWER)

Complexity: Each entity crosses each link, so the complexity is n^2 the size of each message is log(id).

 $M(Alltheway) = O(n^2)$ messages $T(Alltheway) \le 2n-1$ Observations:

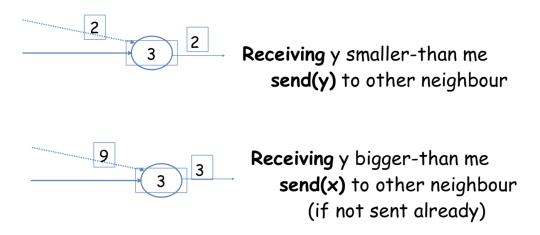
- The algorithm also solves the data collection problem
- It also works for unidirectional/bidirectional

POSSIBLE SOLUTION: ASFAR

Idea: It is not necessary to send and recive messages with larger id's than the id's that have already seen.

Assumptions:

- Unidirectional/bidirectional ring
- Different id's
- · Local oriented



AsFar example

Code: The states are ASLEEP, AWAKE, FOLLOWER, LEADER; the starting state is ASLEEP, the terminal states are FOLLOWER and LEADER.

Unidirectional version:

```
if ASLEEP
     spontaneously
     send('Election', id(x)) to right
     min = id(x)
```

```
become (AWAKE)
        receving ('Election', value)
                 send('Election', id(x)) to right
                 min = id(x)
                 if value < min
                         send('Election', value) to right
                         min = value
                 become (AWAKE)
if AWAKE
        receving('Election', value)
                 if value < min
                         send('Election', value) to right
                         min = value
                 else
                         if value = min
                                 NOTIFY
        receving (NOTIFY)
                 send(NOTIFY) to other
                 become (FOLLOWER)
NOTIFY
        send(NOTIFY) to right
        become (LEADER)
```

SYNCHRONOUS SYSTEMS

We consider as synchronous system all systems in which all local clocks are incremented by one unit symultaneously; in other words all local clocks 'tick' in the same moment. Notice that this assumption does not mean that all the clocks have the same value, but just that their values are incremented at the same time.

The first part of this section is about calculating how much time will take a message to be send in a synchronous system. (not going to treat this part)

MAN-FINDING AND ELECTION

General idea: messages travel at different speed.

- Knowledge of n is not necessary
- · Unidirectional version
- Synchronous

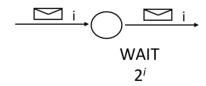
There are two ways of eliminating ids, one is like in AsFar, large ids are stopped by smaller ids; the other is that small ids travel faster so to catch up with larger ids and eliminate them.

Important: Each message travels at a speed which depends on the identity it contains.

Speed is assumed to be unitary, the same for every message. How can we change it? Introducing appropriate delays.

When a node receives a message containing i, it waits f(i) ticks. When a node receives its own id, it becomes the leader and sends a notification message around the ring. This message will not be delayed.

Another example could be done with $f(i) = 2^i$. When a node recives a message containing i, it waits 2^i ticks.



Graphical example of 2^i delay

When a node receives its own id, it becomes the leader and sends a notification message around the ring. This message will not be delayed.

In time 2^i n + n the smallest id i traverses the ring. Let the second smallest be i+1 (waiting time 2^{i+1}). How many links does it have the time to traverse while the smallest id goes around?

$$(2^i+n)/2^{i+1}\approx n/2$$

We can generalize that by thinking about the jth id: in time: $2^i n + n$, has traversed $(n/2^j)$ links. While:

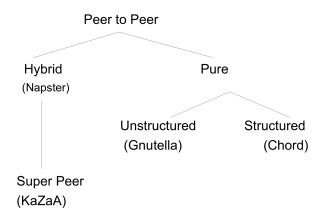
totalMessages =
$$\sum_{j=1}^{n-1} = n \sum_{j=1}^{n-1} 1/2^j = O(n)$$

So we can say that the total time is:

$$totalComplexity = O(2^{i}n)$$

PEER-TO-PEER

ARCHITECTURE

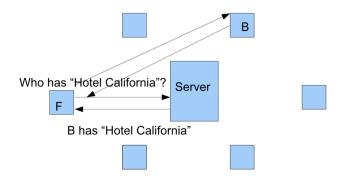


p2p architecture

SEARCH

HYBRID P2P

Napster presents a centralized lookup. The peers send metadata to the look-up server. When a resource request arrives, the server returns the list of the peers that store the resource. Data are then exchanged among peers.



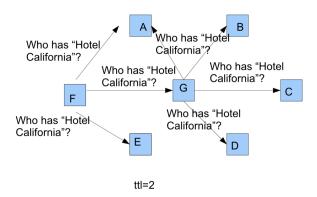
Search in hybrid p2p systems

Unstructured pure P2P

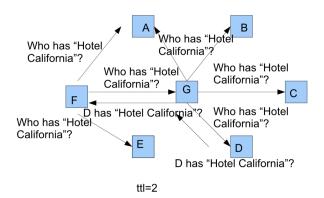
Gnutella is a pure p2p without a central server. Peers have an initial set of addresses known for the first connection; peers are equally treated, no metter which bandwidth they have and how many files they share. Each peer provides both the files and sends/replies to routing requests. Each peer is both client and server.

Each peer knows a subset of neighbouring peers, there are some cache servers that maintains as many peer addresses as possible; when the application starts, it contacts one of these cache servers that will add the p2p network to the new peer.

When the search starts, requests from a peer R are sent from neighbour to neighbour (PING). The message stops either when the resource is found or after a limited and predefined number of steps(TTL = time to live). If the resource has been found, the address of the peer P that stores it, sends to R(PONG). R will directly contact P and will download the resource. **Problem:** hard to control/regulate



PING with Gnutella



PONG with nutella

ITERATIVE DEEPENING

- The system is flooded with a limited TTL
- If the resource is not found, we start with a bigger TTL
- We repeat up to when the resource is found or when a boundary TTL is found Many variations: blind, informed...

SUPER PEER (ZAZAA)

It behaves like a pure and a hybrid p2p system.

- There are different servers, the super peers supervives the subnetworks (subtrees)
- The super peers execute the queries for the leaves (e.i. they work as servers)
- The super peers are also operate as normal peers
- The peers exchange information directly
- Combine the characteristics of hybrid and pure systems

Search? We have to do some questions:

- Good numbers of leaves for each super peer?
- How should super peers connect together?
- The system is made more reliable k redundant

BIT TORRENT

Many users can simultaneously download the same file without delay too much each other. Files are not downloaded from the same server but they are divided into different portions. Each peer that makes a request, offers already downloaded portions to the other peers, thus contributing to the downloading of the other peers.

We concentrate on the efficient fetching and not on the search. There is a single publisher and many downloaders. The same file is distributed among many peers.

INFORMATION STORAGE

The main questions are: where to store information in a p2p system? How to find it? Parameters:

- System scalability: limit the communication overhead and the memory used by the nodes with respect to the number of nodes in the system.
- Robustness and adaptability: in the presence of failures and changes

There could be many approach in which we can store our inforation.

- Centralized approach: we send a request to the server. We use O(N) memory on the server, search O(1) step to reach the server
- Decentralized approach: we send a request to all our neighbours (different optimizations, e.i. TTL). We use O(1) memory and we use $O(n^2)$ steps for the search.

Another approach is to have **distributed hash tables** in the system. It's a good compromise for what regard the complexity of the operations. We use $O(\log(N))$ steps to find the informations and $O(\log(N))$ entries in the routing table of each node. Technique:

- Adaptability: it is simple to insert (assing the information) and remove nodes (reassignment to the neighbouring nodes)
- Balances information on the nodes (makes routing efficient)

Peers and data are mapped in the same address space throungh hash functions. For nodes the hash of the IP; for data hash of the content (title, ...)

CHORD

Data distribution: data are distributed among peers using a precise algorithm. Data are replicated to improve availability.

Assignment:

- Each peer has an ID (hash of the IP)
- Each resource has a key (hash of the title, ...)
- We use long keys to avoid collision
- The peers stores resources with keys similar to one of the peers
- Given a key of a resource the peer sends the request to the peer with the most similar key

Using Chrod we have to use a consistent hashing:

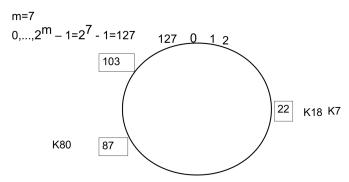
- To assign the keys to the peers
- There should be load balancing so that with very high probability each peer receives the same number of resources/keys
- We ask for key updates when a peer connects/disconnects from the network
- \bullet The insertion of the N-th key with hight probability will require the movement of 1/N other keys

Advantages:

- We maintain consisting hashing although the information is not stored in all the nodes
- The system is simple, correctness and performance are easy to prove
- It takes at most $O(\log(N))$ to reach the destination

- A peer requires a table of $O(\log(N))$ bits for an efficient routing, but the performance decrease when the tables are not updated
- The insertion/removal of a node generates $O(\log^2(N))$ messages

Idea: The logical space is circular O,..., 2^m -1, ring (mod 2^m). A resource with key K is stored in the node which is the closest successor from K (successor(K)).



A key is stored at its successor: node with next higher ID

Example of how Chord assign ID

Successor: the successor of peer 22 is 87, of 87 is 103, of 103 is 22.

Routing: each node stores only its successor. If the resource is not on this node, then the query is sent to the successor.

It requires O(1) of memory, O(N) for search: in the worst case we have to visit the whole ring. The failure of a node blocks the procedure.

This algorithm in the worst case it requires too many operations (we maintain other routing informations).

Note that this extra information is not required for the correctness of the procedure but it helps speeding up the search. This protocol works given that the calue of the next successor is correctly maintained/updated.

The best solution is that each node stores m values. The search of k is sent to the furthest known predecessor of k. We store more values of the close nodes and less of more distant nodes, routing is more precise close to a node. It requires $O(\log(N))$ of memory and $O(\log(N))$ for the search.

Finger table

Each node maintains a finger table (the i-th entry of the finger table of x is the first node that succeeds equals $(x + 2^i) \mod 2^m$) and the predecessor node. An item is identified by id is stored on the successor node of id.

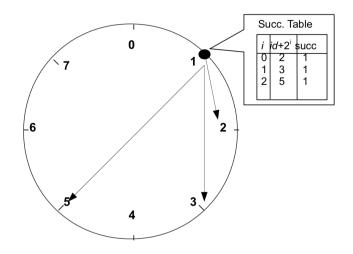
EXAMPLES

Here below it will be provided some examples of how this algorithm works.

Assume m=3, 2^m=8

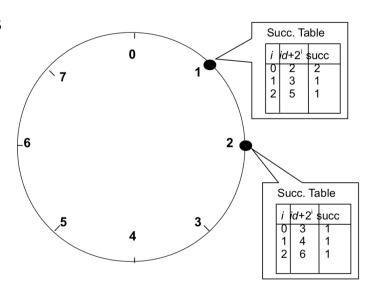
Node n1:(1)
joins-> all entries in
its finger table are
initialized to itself
(no other peer is
there)

$$1+2^0=2$$
, $1+2^1=3$, $1+2^2=5$

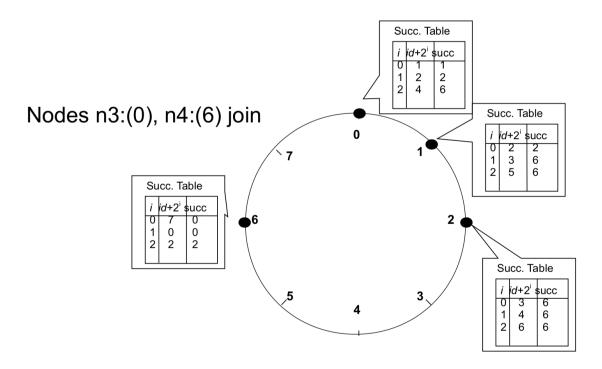


First assignment with Chord

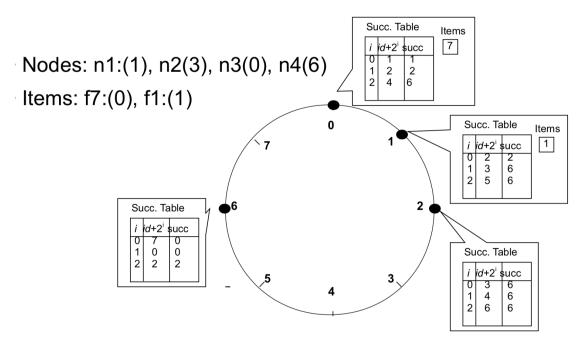
Node n2:(2) joins



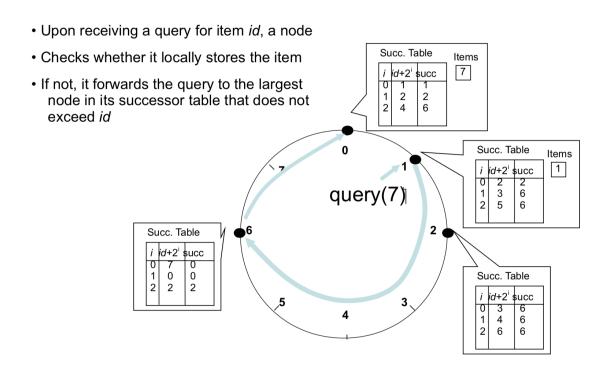
Second node joins



Third and forth joins



Item assignment



Routing after having received a query

MOBILE AGENTS AND THE BLACK HOLE PROBLEM

Definition: Consider that we have some mobile agents that can travel across the graph and could communicate together with a whiteboard and they also have some storage, the problem is to find a malicious node (black hole) that makes disappear our mobile agents. Each agent knows the number of nodes.

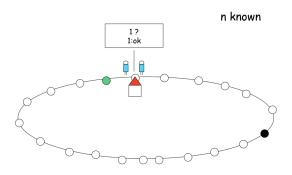
CAUTIOUS WALK

We can consider our graph as a set of ports and in every port there is a traffic light. If the next port has not been controlled, the traffic light is red, going to the next node we activate it (yellow light), and when we come back from that node, that edge (and the next node) is safe (not the black hole).

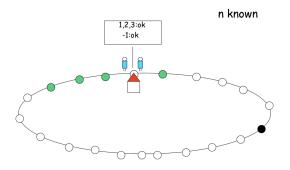
What about if the configuration of the graph is a ring?

We need two agents!

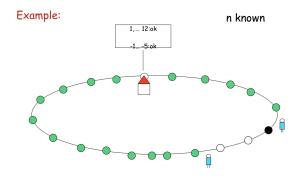
The agents, starting from the home base, must explore using cautious walk. The agents must explore disjoint areas otherwise they could both disappear. So they finally explore step by step all the ring until one of the two agents disappear of they finish the graph and the last node that has not been marked is the black hole.



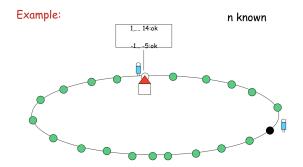
First step of the left agent



First step of the right agent



Black hole for the right agent



Left agent complete the ring and knows where is the black hole (for the whiteboard)

But what if also the intruder is mobile?

THE INTRUDER CAPTURE PROBLEM

Let's put some rules, the intruder:

- can moves from a node to a neighbouring one
- can moves arbitrarily fast
- · cannot cross a node guarded by an agent
- is invisible to the agents
- can permanently see the position of the other agents

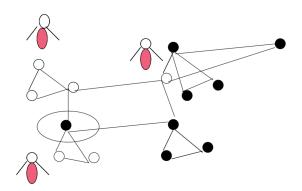
Initially the agents are located at the homebase and form a team while in the end the agents capture (surround) the intruder.

The intruder capture problem is equivalent to the decontamination problem.

As said initially the agents are located at the homebase and form a team. The whole network

is contaminated (except the homebase). An agent cleans a node when it enters in it. At the end the whole network must be clean.

Contamination rule: a node is contaminated if it is not protected by an agent and at least one of its neighbours is contaminated (remember that the intruder is fast).

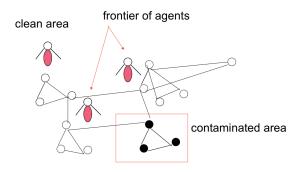


Contaminated node

Decontamination process: As said several times, initially the agents are located at the homebase and form a team. The whole network is contaminated (except the homebase). An agent cleans a node or an edge when it enters (or traverse the edge) in it, a node should not be recontaminated. A node becomes contaminated if it is not protected by an agent and at least one of its neighbours is contaminated.

CONTIGUOUS MONOTONE STRATEGIES

We consider the situration where agent move only to neighbouring nodes (contiguous) and no recontamination can occur(monotone).



Contaminated node

Complexity issues:

· Number of agents

- · Number of moves
- Time (sync or async)
- Memory (agents, nodes(whiteboard))

Strategies are not necessarily connected and monotone (i.e. agents can jump. We have to minimize the number of searchers).

There are two model of visibility, one with visibility of the other neighbours (clean, contaminated, guarded) and the other in which agents have only local knowledge.

DECONTAMINATING A MESH

We have to be in this situation:

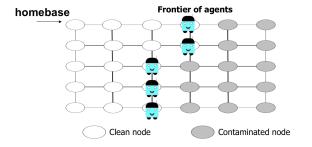
- Asynchronous system
- Node + agent storafe $O(\log(n))$ bits
- Models (visibility, local)

We consider a m x n mesh topology. We have two possible strategies:

- With synchronizer: Searching agents do not have visibility power. The synchronizer is an agent that coordinates the moves.
- Agents with visibility: Visibility power means agent can see their neighbouring nodes, agents can move independently, the synchronizer is not required.

The main idea behind the strategy is to:

- Start from the homebase
- Contiguously clean the contaminated network by maintaining a vertical barrier of agents (to avoid recomtamination, works asynchronously)
- Move one column at the time



Contaminated node

We can perform the search of a node or of an edge moving the barrier. (see slides)