CONTAMINATION SOURCE INFERENCE IN WATER DISTRIBUTION NETWORKS

INFERENCIA DEL ORIGEN DE CONTAMINACIÓN EN REDES HÍDRICAS

A. Braunstein[a], A. Lage-Castellanos[b†] and E. Ortega[b]

a) DISAT, Politecnico di Torino, Corso duca degli Abruzzi 24, Turin, Italy HUGEF, Via Nizza 52, Turin, Italy, Collegio Carlo Alberto, Moncalieri, Italy
b) Facultad de Física, Universidad de la Habana. ale.lage@gmail.com
† corresponding author

We study the inference of the origin and the pattern of contamination in water distribution networks after the observation of contaminants in few nodes of the network.

Se estudia el problema de la inferencia del origen y del patrón de contaminación en una red de distribución a partir de la observación del contaminante en pocos nodos de la red.

I. INTRODUCTION

Fluid networks are ubiquitous and fundamental. They come in a variety of sizes and shapes, from circulatory and lymphatic systems in vertebrates, to ocean conveyor belts and sea currents, passing by industrial fluid networks, water distribution systems in cities and hydrographic basins. In all of these cases, the spread of contamination, defined as the presence in the fluid of non desired material, is problematic. We study how to infer the position in the network at which a contaminant first appeared after the observation of its presence in some distant nodes of the network. This is a relevant question at least in industries and cities, where pipe networks are not susceptible of being explored or sensed at every node, and where future or further damage can be prevented from the knowledge of the contamination source.

The study of contamination origin has received a lot of attention in the near past. Several approaches have been developed [1–7] with the goal of infering the patient zero after observing an epidemic outbreak. In [8] there is a state of the art of studies made in related topics. However, epidemic models in networks differ from fluid networks in that they are essentially stochastic. Pipe systems can have some source of stochasticity, like turbulence, but to the scope of the present study the movement of the fluid is considered deterministic. The simplicity of fluid networks is compensated by the fact that we will try to infer the origin with very little information, namely the observation of contamination in one or two nodes of the system.

There is a variety of questions related to the inference of contamination in water distribution systems, and a variety of methods and approaches. Some researchers [9] have focused on the inference of the pattern at the origin, after observation of a contamination pattern at an observation node, but under the assumption that both the origin and the observation points are known. In such cases, the target is to reproduce the real-valued time dependent concentration of contaminants.

Instead, we will pay more attention to discovering the unknown source of contamination.

Within the scope of detecting the source of contamination, many researchers make use of extensive forward in time simulations (usually resorting to EPANET) [10–16] to test different possible origins, and compare their predicted patterns with the observed one, or to create a rule that allows them to make inference later on. Some consider that there is a source of stochasticity coming from the uncertainties in the demands on the network at any given time [13, 17] or on the reliability of the sensors [18], and try to infer in such a context, sometimes resorting to Monte Carlo simulations [17] or Bayesian methods [13, 17, 18]. Some have attempted the detection of the origin by directly reversing the flow dynamics in the network [19].

Our approach has some contact points with some of these. As in [13, 17, 18] we formalize our method in Bayesian terms, although we do not consider stochasticity in the demand or on the sensors. We assume a simplified version of the problem, with known deterministic dynamics, where time is discretized and contamination is reduced to be binary (contaminated or not) as in [13]. Our inference method will rely on a neat mathematical approach using linear programming to find the most effective explanation for a contamination event. The linear programming optimization will arise from an Occam's Razor-like argument by the assumption that contamination is a rare event.

Section II establishes the simplifications assumed throughout this work in the modeling of the water distribution networks, and the treatment of their dynamics. A precise statement of our mathematical problem appears in Section III, where it is also transformed into an optimization problem. The results obtained in random models of cities appear in Section V, and everything is summarized at the end.

II. ASSUMPTIONS AND SIMPLIFICATIONS

Fluid networks can be grouped in two big classes, non cyclic and cyclic regarding on whether any piece of fluid is allowed to pass by a given position in the network more than once. Circulatory systems and the ocean conveyor belt are examples of cyclic networks. On the other hand, many industrial and city networks are non cyclic. The procedure presented by us will focus on non cyclic networks. Extensions to cyclic networks might be simple, but have not been explored so far.

Furthermore, we will focus on water distribution networks (WDN), this is, the system of pipes, pumps, elevated tanks, junctures and consuming points that characterize a standard city clean water network. The behavior of such networks are described by the Todini-Pilati equations [20], in which two fundamental constraints, the conservation of mass and the conservation of energy, are grouped in matricial form in the following way

$$
\begin{bmatrix} A_{pp} & A_{pn} \\ A_{np} & 0 \end{bmatrix} \begin{bmatrix} Q \\ H \end{bmatrix} = \begin{bmatrix} -A_{p0}H_0 \\ q \end{bmatrix} \tag{1}
$$

On right-hand side, the flow of demand q at the consumption nodes is given, as well as the head pressure H_0 at the pumps and elevated tanks of the city. On the left-hand side, the flow in each pipe Q and the pressure in internal junctures H are to be found in order to satisfy the energy and continuity equations. Detailed explanations of each term and a discussion on the meaning and solution of the equations can be found in [21]. Despite their apparent simplicity, these equations are non-linear, since the elements of the matrix $A_{pp}(i,i) = R_i|Q_i|^{n-1}$ depend on the flows Q we are looking for.

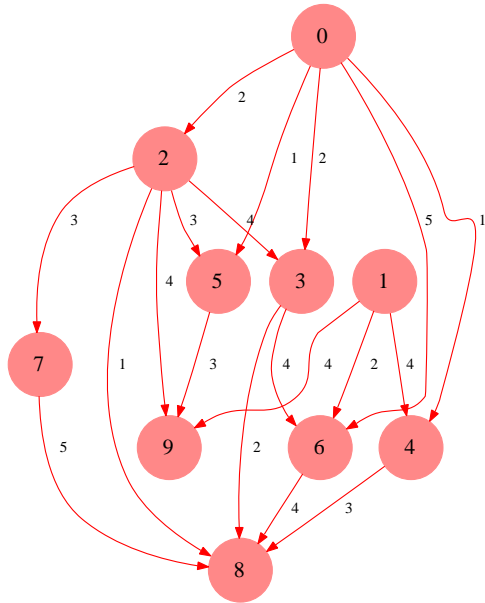
Figure 1. Schematic representation of an abstract water distribution network. Numbers inside nodes are labels. Numbers besides edges are time delays Δ_{ij}.

Inside the matrices A_{pp}, A_{pn}, A_{np} and A_{p0} resides the information about the topology of the network, the radii

of the pipes as well as the drag coefficients inside the pipes. In principle, all these parameters are known (at least approximately) and the flows Q and the fluid velocities v inside the pipes, can be found using free and public modeling software like EPANET [22]. This software allows also the study of the diffusion of contaminants in the network.

However, we will concentrate our effort on studying the direct (forward in time) problem and inverse (inference) problem in a simplified version of the water dynamics. In order to do this we will assume the following simplifications:

- **discrete time:** once the fluid velocities v are known, we obtain the pipe time $\Delta = \text{Round}(\frac{L}{v})$ from the pipe lengths L. We will consider this time Δ an integer, using a suitable discretization. This discretization is used for describing the states s_i^t of each site i, at time t, in the network.

- **binary contamination:** at each time the state of node i can be either clean $s_i^t = 0$ or contaminated $s_i^t = 1$. We shall make the distinction between the state $s_i^t \in \{0, 1\}$ of a node, and the variable $v_i^t \in \{0, 1\}$ signaling whether node n_i is being actively contaminated by an external source in time t. In other words $v_i^t = 1$ marks node n_i at time t as origin of the contamination.

- **deterministic:** we will suppose all real parameters of the system (head pressures, drag coefficients, pipe lengths, demand flow, etc) as known in detail and fixed in time, such that the system operates in stationary regime. Stationarity was assumed for simplicity, but is not essential to our method, and can be readily lifted.

- **WDN is a graph:** any water distribution network is representable as a directed, non cyclic graph with timed edges $G(V, E, \Delta)$:

  - $V = \{n_i | i \in [0, 1, \dots N]\}$ is the set of all nodes (vertexes) in the graph,
  - $E = \{(n_i, n_j)\}$ is the set of directed $n_i \to n_j$ edges with time delay Δ_{ij}
  - Δ_{ij} is the time delay of each pipe. When $(n_i, n_j) \in E$, $\Delta_{ij} > 0$ and $\Delta_{ji} = -\Delta_{ij}$. If $(n_j, n_i) \notin E$ then $\Delta_{ij} = \infty$.

In practice the graph $G(V, E, \Delta)$ is constructed using the topological network data (which nodes are joined by pipes). But this is not enough. We also need the stationary solutions of the Todini-Pilati equations, since we need the time Δ_{ij} the fluid takes to travel through pipe (n_i, n_j) that depends on the velocities of the fluid. In this sense, $G(V, E, \Delta_{ij})$ summarizes topological as well as dynamical information of our system.

II.1. Simplified dynamics for direct problem

The forward in time evolution of contaminants in the network is ruled by the following consistency equation between the state of node n_i and the sates of it's predecessors $\partial^+ j = \{n_j | (n_j, n_i) \in E\}$:

$$
s_i^t = m(s_{\partial^+ i}^{t-\Delta}, v_i^t) \equiv v_i^t \bigvee_{j \in \partial^+ i} s_j^{t-\Delta_{ji}} \tag{2}
$$

The meaning of this equation is quite obvious. The node n_i at time t can be contaminated ($s_i^t = 1$) if either it is the origin of the contamination ($v_i^t = 1$) or it is receiving contaminants from nodes connected to it $\exists_{j\in\partial^+i} s_j^{t-\Delta_{ji}} = 1$.

II.2. *Graph-time expansion*

Equation (2) suggests that the dynamics of contamination can be described in a time-extended graph, where nodes s_i^t are not only defined by their spatial location i in the WDN but also by a (discrete) time index t, where we define directed edges to be existent between nodes ($s_j^{t-\Delta_{ji}} \to s_i^t$) if a pipe connects node j with node i with a time delay of Δ_{ji}. In Figure 2 we show the first three time slices of the network of Figure 1. To avoid overloading the Figure, we only represented three edges, two of them connecting physical node 0 to physical node 5, in a $\Delta = 1$ time delay, and one connecting node 0 to node 2 in $\Delta = 2$ time steps.
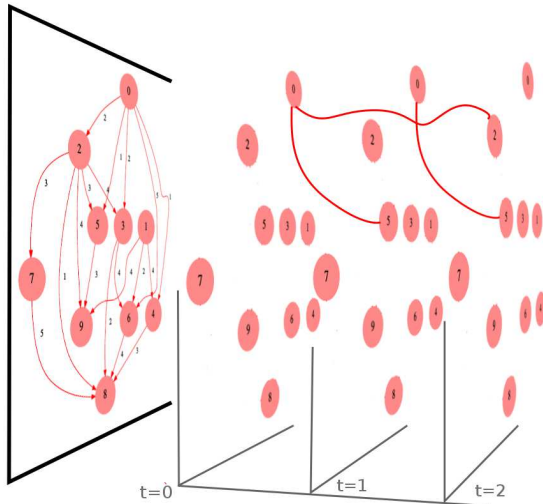

Figure 2. Representation of the time-extended graph derived from the graph in figure 1.

The time-extended graph will be the starting point of our construction, allowing us to simulate simplified dynamics through eq. (2). To test our inference procedure, we generate random graphs with the desired properties (directed, non cyclic and timed edges) pick up a random source node and a random pattern $v_i^t = (0,0,0,1,1,0,1,1,0,0,\dots)$, and use equation (2) to track the evolution throughout the network. We will also pick, among the contaminated nodes, one or two that will be used as sensor, and whose evolution in time $s_o^t = (0,0,0,0,0,0,1,1,0,1,1,\dots)$ will be accessible to the inference procedure that we explain next.

III.   INFERENCE AS OPTIMIZATION PROBLEM

The inverse problem we want to solve can be stated as follows. Given a water distribution network, characterized by a graph $G(V, E, \Delta)$, and a set $O = \{o_1, o_2, \dots\}$ of observed nodes with their patterns $s_{o_i}^t = (0,0,0,1,1,\dots)$, find the most

probable set $P = \{p_1, p_2, \dots\}$ of nodes and their patterns $v_p^t = (0,0,1,1,\dots)$ for the origin of the contaminants in the network. When convenient we will denote both, the observed nodes and the contamination patterns by O and P in each case.

In the estimation theory language, and using the Bayes formula, we want to maximize, over the set of origins P, the probability

$$Prob(P|O) = \frac{Prob(O|P)Prob(P)}{Prob(O)}.$$

To the scope of maximization, the denominator $P(O)$ is unknown but irrelevant. On the other hand, the probability of observing contamination O given a set of sources P

$$Prob(O|P) = \sum_{s_i^t \notin O} \prod_i \prod_t \delta_{s_i^t, m(s_{\partial^+i}^{t-\Delta}, v_i^t)}$$

is either one or zero, depending on whether the deterministic evolution of contamination in time following equation (2) reproduces or not the observed pattern. This is a consequence of the deterministic assumption made for the dynamics of the network.

Therefore, the maximum of $Prob(P|O) \propto Prob(O|P)Prob(P)$ is found among the origins that are consistent with the observation, that we will symbolize as $P \to O$. In other words, we will estimate the origin of the contamination as the set of nodes and patterns satisfying

$$\hat{P} = \underset{P:P\to O}{\operatorname{argmax}}\ Prob(P).$$

Now we need to fix a prior for the patterns. Assuming that the contamination is a rare event, we will consider as more probable the explanations requiring the smallest amount of nodes and time laps involved in the original contamination, therefore, the smallest set P.

Using equation (2) recursively, we can write the consistency equations between the observation at any node, and the original contamination variables v_i^t at nodes above it in the graph,

$$s_i^t = M(v_{\partial^{++}i}^{t-\Delta_c}, v_i^t) \equiv v_i^t \bigvee_{j\in\partial^{++}i} \bigvee_{c\in C_{ij}} v_j^{t-\Delta_c^+}. \qquad (3)$$

The reason why this recursive substitution of eq. (2) in itself has an end point is that we are considering non cyclic networks, so the observation at node s_o^t can only come from a finite number of nodes and times in the graph. In expression (3), Δ_c^+ stands for the time delay of path c from node n_j to node n_i, while $\partial^{++}i$ stands for all nodes n_j that can access node n_i by some path c in graph $G(V, E, \Delta)$. The expression $v_{\partial^{++}i}^{t-\Delta_c}$ inside the definition of function $M(\cdot, \cdot)$ represents all element $v_j^{t-\Delta_c^+}$ on the right-hand side of the equation.

The condition $P \to O$, read as "*origins $P = \{p_1, p_2, \dots\}$ causes observation O*" is equivalent to the condition

$$\forall_{o\in O}\ \forall_t\ s_o^t = M(v_{\partial^{++}o}^{t-\Delta_c}, v_o^t).$$

III.1. Graph reduction

As we want the smallest set P, any node that is not present on the right-hand side of this equation, will be assumed not to be an origin $v_i^t = 0$, since it is irrelevant to explain the observation. Furthermore, the nodes that do appear on the right hand side of one of these equations, but whose left-hand side is a clean observation $s_o^t = 0$, are forced by this equation to also be clean, $v_j^{t-\Delta_c^+} = 0$. Therefore, in order to explain the observation we remain with the observed nodes variables v_o^t that are seen as contaminated $s_o^t = 1$ and those $v_j^{t-\Delta_c^+}$ connected to them by equation (3) that are not connected to non contaminated observations. All this reduction of the valid nodes is equivalent to what is done in the contamination source pruning of [18].
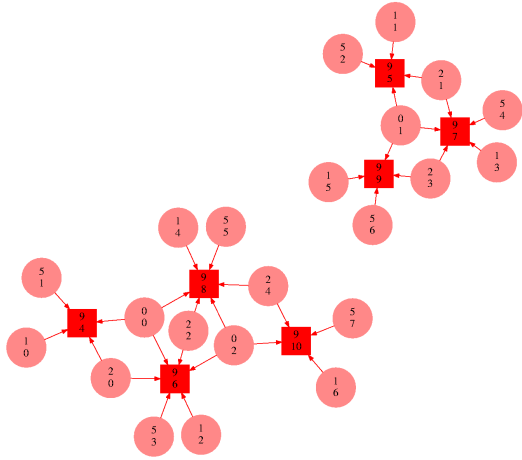

Figure 3. Each node is labeled by two numbers, the upper one is the original node number in the graph (Fig. 1) while the lower one stands for the time bin of the variable v_i^t it represents. Square nodes correspond to observed state variables v_o^t, while circular nodes correspond to the remaining variables (see text).

We can represent the relation between the remaining variables as a graph, in which each contaminated observed node variable v_o^t is connected to the non observed nodes that remained as possible explanations of the observed state, as in Figure 3. Furthermore, re-labeling the variables in this representation as y_k for the observed (square) nodes v_o^t and x_i for the non-observed (circular) nodes $v_j^{t-\Delta_c^+}$, the constraints imposed by eq. (3) can be stated as

$$\forall_{y_i \in Y} \quad y_i + \sum_{x_k \in \partial y_i} x_k \geq 1 \tag{4}$$

forcing each observed node itself, or one of its neighbors to be the origin of the contamination. The condition $\hat{P} = \underset{P:P\to O}{\mathrm{argmax}} \; Prob(P)$ now can be rewrite as

$$(\hat{x}, \hat{y}) = \underset{x,y}{\mathrm{argmin}} \left(\sum_i x_i + \sum_j y_j \right) \tag{5}$$

under the constraints given by eq. (4).

Both, the objective function and the constraints are linear, and therefore susceptible of being solved by integer/linear programming methods (LP). Furthermore, since coefficients are integers, more than one optimal solution might exist. The set of all solutions can be found by recursively slightly altering the coefficients of the variables in the objective function (see [23]). Once a solution is known, we get a set of variables y and x with value 1, while the rest is zero. Mapping back y_k and x_i to the original variables v_o^t and v_i^t, we claim this solution to be an optimal pattern of contamination.

At this point it is important to clarify that the optimization made by LP is only in time, meaning that any two solution with the same number of active times for the contamination are equivalent. In other words, the optimization does not distinguish whether the source nodes of the time-extended graph correspond to the same node in the water network, or to different nodes. This is relevant, since the basic assumption made, namely that contamination is a rare event, would also imply that solutions with one real node of the water network might be preferable to those involving more nodes, even in the cases where the latter involves less time-expanded nodes. The current procedure can not solve this issue neatly. We partially circumvent it by choosing among the solutions those involving less real nodes of the water distribution network.

We have shown how to transform the inference of contamination into a treatable optimization problem. Let us outline the whole procedure here:

1. **Input:** Start from the given WDN and the set of pipe times obtained from the solution of the Todini-Pilati equations for your system, and some observed contamination pattern. (For this step one probably will use EPANET or other similar software, as we did for the network of Modena city. In the case of random networks, this step was not required, since the time intervals were directly generated from a probability distributions.)

2. **Graph-time expansion:** Using the speeds along the pipes, and the distances of the pipes, compute the time delays Δ_{ji}. Use a suitable discretization of the times to create a time-extended graph as explained in the previous section. (We programmed this step in *Python*.)

3. **Graph reduction:** Starting from the time-extended graph, create a new graph by connecting every node s_i^t to all other nodes $s_k^{t_k}$ such that a path

$$(s_k^{t_k} \to s_{k1}^{t_{k1}}), (s_{k1}^{t_{k1}} \to s_{k2}^{t_{k2}}), \ldots, (s_{kn}^{t_{kn}} \to s_i^t)$$

exists. Reduce this graph (as in Figure 3) by retaining only the observed contaminated nodes and those connected to them that could possibly explain the contamination. (We programmed this step in *Python*)

4. **Optimization:** Use linear programming to find the most efficient explanation for the contamination observed. (We used *lpsolve* solver in Linux)

Next we discuss two typical situations where our method fails and succeeds. In the section after, we test the efficiency in three different situations: many unrealistic random topologies water distribution networks, a real city (Modena in Italy) and finally many realistic random cities statistically similar to Modena.

IV. SIMPLE SMALL EXAMPLES

Let us consider two separate examples, usign the toy network of Figure 1. The first example is the observation of a contamination at sensed node 7 (bottom left, painted in blue color), with pattern $s_7 = (0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$. Given that there is only one path in the network arriving at node 7, it is obvious that contamination should have happen along that path, this is, either at 7 itself, at node 2, at node 0, or at a combination of them. Our algorithm (and logic) will produce the following solutions (only representing the non zero nodes):

$$
\begin{aligned}
S \quad = \quad & \{(v_7^5, v_7^8), (v_2^2, v_2^5), (v_0^0, v_0^3) \\
& (v_7^5, v_2^5), (v_7^5, v_1^3) \\
& (v_2^2, v_7^8), (v_2^2, v_1^3) \\
& (v_1^0, v_2^5), (v_1^0, v_7^8)\}
\end{aligned}
$$

While all those solutions involve two nodes of the time-extended graph and are equivalent to our optimization code, only the first three involve one real node of the WDN. We call this solutions 1-node-solutions. Assuming that contamination is a rare event, we will select the subset of 1-node-solutions, whenever it is not empty, as the most probable contamination source. In this case, however, neither the algorithm nor logic can reduce our uncertainty among which of the three nodes $(7, 2, 0)$ was the most probable origin. This is a difficulty that arise also in many other contamination inference algorithms [10, 13, 16, 19].
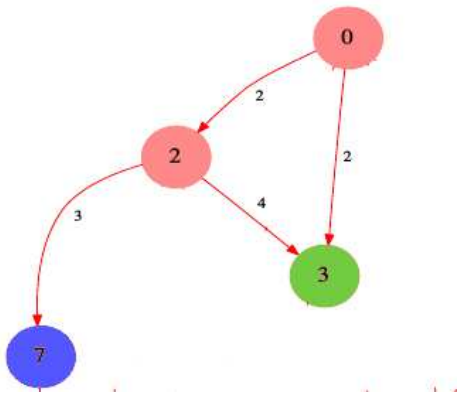

Figure 4. Part of Graph from figure 1. Blue (node 7, bottom left) and green (node 3, bottom right) nodes are those where two different contamination events are sensed.

The second example is the observation of a contamination at sensed node 3 (bottom right, painted in green color), with pattern $s_3 = (0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0)$. Now there are two paths in the network arriving at node 3. As before, we can explain the contamination with single node solutions as

$$
S \quad = \quad \{(v_0^3), (v_2^1, v_2^5), (v_3^5, v_3^9)\}
$$

The outcome of our algorithm, however, will be only (v_0^3), since it minimizes the number of nodes in the time extended graph required to explain the contamination effect. This illustrates both, the power and the risk of our algorithm. We expect our optimization procedure to return non-trivial inference whenever there is more than one path connecting the real contamination source with the observation nodes. At the same time, complex in time patterns (with contamination lasting for more than one time interval) are susceptible of being over-simplified. That would be the case of the previous example, if the real contamination origin had been node 2, a situation that our algorithm would disregard as sub-optimal.

We could study mathematically the probability of such situations appearing in water distribution networks, but we leave this for a future study on the optimal sensor distribution. For the optimality of sensor placement, exploiting multi-path will be crucial. We next characterize "experimentally" our method by trying it on many artificial networks.

V. RANDOM DISTRIBUTION NETWORKS

In the following we will only treat contaminations events that take place in a single node of the network. Therefor if the true contamination patterns is a solution of our optimization problem, then it is in the 1-node-solutions. Despite already being a reduction, the 1-node-solutions can contain many different explanations for our contamination. For instance if we have a WDN that is simply a single chain of nodes with water moving along, and a sensor at the end of it, the method will find all nodes in this line are equally good 1-node explanations of the observed pattern.

We define the efficiency of our method in two different ways. First, by the probability of finding the real contamination pattern in the 1-node-solutions. Second, by the probability that the real contamination pattern is one randomly selected 1-node solution. This second measure is more demanding, since in order to be efficient, the method should find few optimal solutions, while the first measure is insensitive to the multiplicity of solutions.

To explore the efficiency of the method we first test it on random graphs with WDN characteristics. These random graphs (as the one in figure. 1) are Erdos-Renyi graphs [24] in which a directionality and a time delay are enforced in each pipe. Figure 5 shows the efficiency of the method versus the number of nodes in the system. The origin of the contamination is taken to be a single randomly selected node i, and we spread from it a compact contamination pattern (contamination in consecutive times steps, lasting a randomly selected time between 2 and 10) at $t = 0$ $v_i = (1, 1, 1, 0, \ldots)$. The selection of sensors is made in the following way: the sensed node is picked randomly from the set of nodes touched by the contamination but guaranteeing that is not connected with previously chosen sensors (because

in this case the contamination in one of them can explain the contamination in the other). If this condition can not be satisfied, then the second sensor is picked randomly from the nodes through which contamination does not passes (the information of a node not being contaminated is also informative, since it allows the removal of all its neighbors from the set of possible explanations).
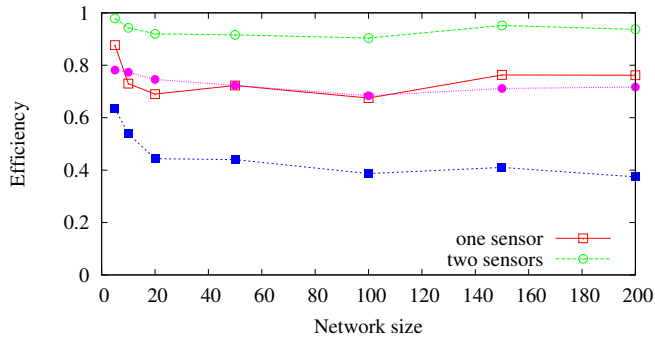

Figure 5. Efficiency of the inference method with one and two sensors in a system with a compact contamination pattern. Squares correspond to one sensor cases, while circles correspond to two sensors. Open points correspond to the first efficiency measure, while the full points correspond to the second (see first paragraph in this section). Each data point is an average of 1000 different contamination events in different systems.

We could have done simulations for bigger graphs, but the trend seems already to be stable in the range $20 - 200$. Let us underline that 200 nodes is already a size comparable to that of small cities, or at least of some industries. However, we did not push forward in this direction, since an Erdos-Renyi random graph is hardly a credible topology for real water distribution networks (see Modena and Modena-like random cities in the next sections).
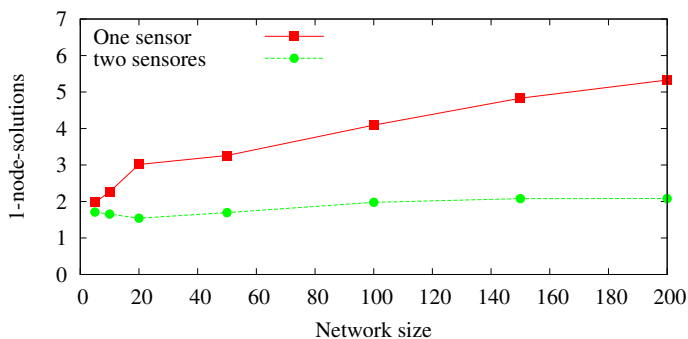

Figure 6. Comparison between the number of solutions which involved only one node of the Graph G with one or two sensors in the system. Each data point is an average of 1000 different contamination events on different systems.

The efficiency practically remains constant for 50 or more nodes. With two sensors in the system the method is nearly perfect since practically in every case of study the real contamination pattern was obtained among the optimal solutions found. However, in both cases, the inference method finds many 1-node solutions for the optimization problem, reducing the second efficiency measure.

In any case the number of solutions found by the method

is also a relevant parameter to characterize its efficiency. As we mentioned at the end of Section III the method can not determine the solutions which involve less number of nodes the real water distribution network (graph G) directly from optimization (5), but rather those involving less nodes of the time-extended graph. Given a set of optimal solutions we can further reduce this set by selecting among them those involving fewer real nodes in the network. In Figure 6 we show the number of optimal solutions which involve only one real node of Graph G.

Observe that the number of 1-node-solutions is never higher than ten, this means that with one sensor in a network of 200 nodes the number of nodes which are possible source of the contamination are around 10, which is significantly smaller than the network size, but still not a reliable estimation method. However, with two sensors the average number of optimal solutions is between 1 and 2, which is quite a reduction of the uncertainty.

VI. MODENA CITY

To get more realistic, we studied a real city network: the one of Modena, Italy. As can be seen in the top-left panel of Figure 3, real cities are far from being random E-R graphs, since they are mostly planar graphs. We took Modena topology as well as pressure and demand data from the Internet [25], and used EPANET to solve the stationary state of the network. With the resulting fluid velocities in pipes, we computed the delay times in each network pipe using a discretization of 30 seconds, which is smaller than the fastest pipe (41 seconds), and much smaller than the average pipe delay (~ 1100 seconds, more than 30 times the discretization). Then we constructed our abstract graph representing the city and studied many contamination events on random locations.
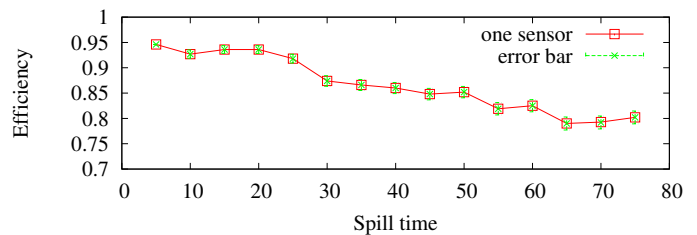

Figure 7. Efficiency of the method as a function of the duration of the contamination event. The time units is 30 seconds, so the largest spill time considered is 37.5 minutes.

We cannot make an efficiency *vs* size study, since the size of the network is fixed to 268 nodes, but we can study the efficiency with respect to the duration of the source contamination. Notice in Figure 7 that for the range from 1 to 75 discrete time intervals, the efficiency remains above 75 % . It seems to be a decreasing function of the contamination time. This is consistent with the fact that the proposed optimization method relies on the assumption that contamination is a rare event. This also suggest that a rule of thumb for assuming a good discretization of time is one that is smaller, but not many times smaller than the typical

duration of a contamination event. Of course, the meaning of "typical" here will be case specific.

VII. RANDOM MODENA-LIKE CITIES

We would also like to have statistic results over many different city networks. However, information about the water distribution network of real cities is normally restricted for security reasons. Specially since September 11th terrorists attacks, there has been a growing concern on concealing information that could be used to plan terrorist acts with the highest impact. Unfortunately, such comprehensible precaution affects the testability of studies like this, intended to protect citizens from intentional and non intentional contamination diffusion.
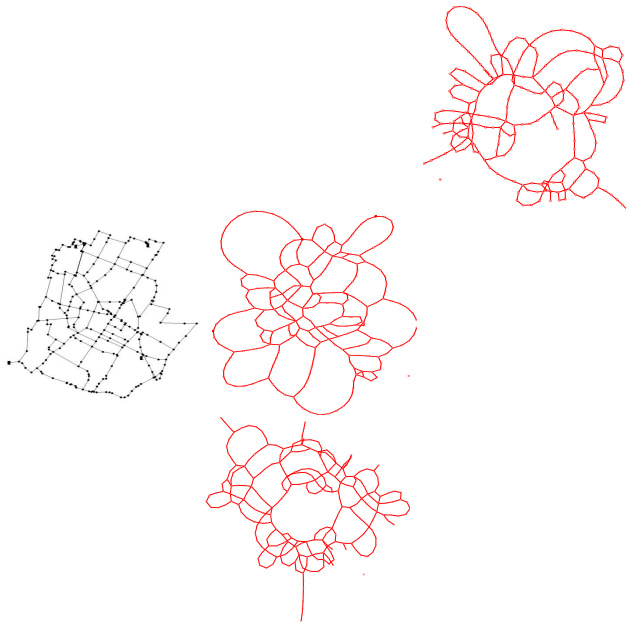

Figure 8. Modena and Modena-Like graphs in different views. 1. Graph of Modena city created with EPANET. 2. Graph of Modena city in a different view. 3,4 Graphs of Modena-Like cities.

To tackle this we devised a plausible algorithm to create random planar graphs similar to Modena. We start by a square 2D lattice graph of smaller size than Modena. Then we randomize the graph by the following two steps:

1. (add nodes) take randomly some edge (s_i, s_j) and replace it by a new node s_k and two new edges (s_i, s_k) and (s_k, s_j). Do this until you end up with a number of nodes similar to that in Modena.

2. (remove edges) take randomly some edges of the graph and remove them.

The process is fine-tuned to achieve a similar number of nodes as well as degree distribution as that observed in Modena. The fraction of new nodes introduced to achieve a degree distribution as in Modena is then kept fixed when producing bigger or smaller sizes of networks. Both steps proposed will preserve the planarity of the original 2D lattice.

In Figure 8 you can see the similarity between Modena and our random versions of it. The top-left panel represents the real topology of the Modena network, while the top-right one is also Modena, but plotted with a standard graph plotting software. The lower panels are random Modena-like cities produced as explained before.

After creating the artificial Modena-like topology, we randomly assign a rank of pressures to the nodes, and give orientation to the fluid in each pipe accordingly. In the resulting model, we apply a contamination of typical time duration 10 time units to a random node and evolve it forward in time. Then we try to infer its origin with our optimization method taking the observation at one other random point of the system among those that are touched by the contamination.

With only one sensor, in sizes from 50 to 500 nodes we found that the correct contamination origin was found among the optimal solutions nearly every time (¿99 %). However, the true solution is not the only one given by the algorithm. In the case with one sensor due to the low connectivity of real networks, and specifically in Modena and Modena-like networks, our method finds many optimal solutions, since almost each time the source and the sensed node are connected by a single path.
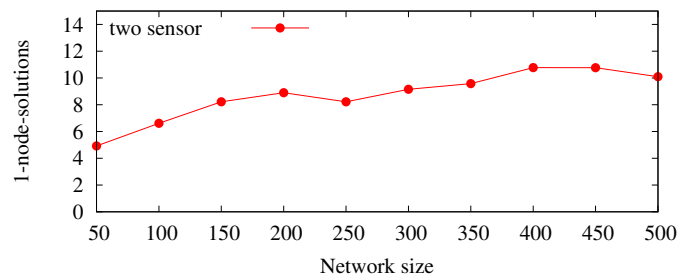

Figure 9. Number of solutions which involve only one node from Graph G in Modena-like networks with 2 sensors. Each data point is an average of 1000 different contamination events in different systems.

With two sensors, again, the method finds the true contamination pattern almost every time. In Figure 9 we show the number of optimal 1-node solutions found, but "normalizing" it. As we mentioned, when we have a single line of nodes the method can not distinguish which among them is the most probable source of contamination. For this reason, we decided to remove from the networks those nodes that have only two neighbors, and replace both pipes by an equivalent larger pipe. The number of 1-node solutions of this reduced version of the network are shown in Figure 9.

From the analysis of Modena and Modena like cities, we should emphasize that the method is good at finding the true origin among the solutions, but bad at reducing the number of equivalently good solutions. The multiplicity of optimal solutions is inevitable when the sensor node and the true origin of the contamination are connected by a single path. Let us underline, however, that we are testing our method in random settings, while in real life we expect the positioning of the sensors to have been carefully optimized to reduce the chances of multiple solutions.

VIII. CONCLUSIONS

In this paper we have presented a method to infer the contamination sources in fluid networks. The method relies in the plausible assumption that contamination events are rare. The inference was formulated in a Bayesian approach and it was transformed into a linear optimization problem with linear restrictions which can be solved using linear programming methods. We emphasize that the Bayesian step is used to formally derive an optimization problem, and not for the treatment of uncertainty in the network parameters (like flow, pipe diameters, etc) as has been done before. Through the simulation of many random events of contamination in random cities, Modena city and Modena-like cities, it was possible to determine the efficiency of the method (above 70 %), proving the method useful.

The method proposed could be readily extended to the case of non stationary states in the network. If the velocities in the pipes are changing with the demand during the day, for instance, it is not a problem to apply the procedure described to this case. It suffices to create the extended graphs connecting time-space nodes according to the time of the day each node represents.

However, the method is too simplistic in many aspects, some of which could be improved. Time discretization is one simplification, but is among those we would rather keep in future treatments. Binary contamination, however, seems a stronger simplification, since a lot of information could be present in the intensity pattern. To keep in the realm of discrete optimization problems, an improvement could be achieved by having more than two discrete contamination levels.

Furthermore, we consider that the main drawback of the method for its applicability, is its rigidity. Contamination could pass undetected by observation nodes, specially if highly diluted. In the current setting, any observed non-contaminated time-space node, implies the definite elimination of all the upstream nodes as possible origins. We have seen that when using our method with EPANET realistic diffusion, this implies losing the real sources many times. We think that the best improvement over the current method will come from applying stochastic methods of inference, like Bayesian networks, to the problem. We are currently moving in this direction.

REFERENCES

[1] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, R. Zecchina, Phys. Rev. Lett. 112, 118701 (2014).

[2] W. Luo, W. P. Tay and M. Leng, IEEE J. Sel. Top. Signa. 61, 2850 (2013).

[3] Ch. Milling, C. Caramanis, Sh. Mannor and S. Shakkottai, 2012, 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (2012).

[4] B. M. Spinelli, LCA3, I&C, EPFL (2009).

[5] K. Zhu, and L. Ying, Information Theory and Applications Workshop (2013).

[6] W. Luo, W. P. Tay and M. Leng, IEEE T. Signal. Proces. 61, 2850 (2014).

[7] A. Y. Lokhov, M. Mézard, H. Ohta and L. Zdeborová, Phys. Rev. E. 90, 012801 (2014).

[8] J. Jiang, IEEE Commun. Surv. Tut. 19, 465 (2017).

[9] I. Butera, F. Boano, R. Revelli and L. Ridolfi, J. Water. Res. Pl. 139, 418 (2012).

[10] A. Preis and A. Ostfeld, J. Water. Res. Pl. 132, 263 (2006).

[11] L. Perelman and A. Ostfeld, In Water Distribution Systems Analysis 426 (2010).

[12] C. D. Cristo and A. Leopardi, A. J. Water. Res. Pl. 134, 197 (2008).

[13] J. J. Huang and E. A. McBean, J. Water. Res. Pl. 135, 466 (2009).

[14] Ch. Hu, J. Zhao, X. Yan, D. Zeng and S. Guo, Ad Hoc Netw. 35, 116 (2015).

[15] J. Guan, M. M. Aral, M. L. Maslia, W. M. and Grayman, J. Water. Res. Pl. 132, 252 (2006).

[16] T. Tao and Y-j. Lu, X. Fu and K-l. Xin, J. Zhejiang. Univ-SC.A. 13, 559 (2012)

[17] H. Wang and K. W. Harrison, J. Water. Res. Pl. 139, 13 (2011).

[18] M. Propato, F. Sarrazy, M. and Tryby, J. Water. Res. Pl. 136, 483 (2009).

[19] E. Salomons and A. Ostfeld, Identification of possible contamination sources using reverse hydraulic simulation, 12th Annual International Symposium on Water Distribution Systems Analysis, Tucson, Arizona, USA, published on CD, (2010).

[20] E. Todini and S. Pilati, Computer applications in water supply: vol. 1—systems analysis and simulation (Research Studies Press Ltd., 1988).

[21] O. Giustolisi, D. Laucelli and A. F. Colombo, J. Water. Res. Pl. 135, 117 (2009).

[22] L. A Rossman, EPANET 2: users manual. (US Environmental Protection Agency. Office of Research and Development. National Risk Management Research Laboratory 2000).

[23] E. Ortega, "Diploma Thesis", Physics Faculty, University of Havana, 2015.

[24] P. Erdos and A. Rényi, Publ. Math. Debrecen. 6, 290 (1959).

[25] University of Exeter, Free Modena EPANET model. http://emps.exeter.ac.uk/media/universityofexeter/emps/research/cws/downloads/data/3-epanet/MOD.inp. (2014).