Enabling State Estimation for Fault Identification in Water Distribution Systems under Large Disasters

Qing Han*, Ronald T. Eguchi‡, Sharad Mehrotra*, Nalini Venkatasubramanian*
$*$ University of California, Irvine, $\ddagger$ ImageCat Inc., CA, US.

Abstract—**We present a graphical model based approach for on-line state estimation of water distribution system failures during large-scale disasters. Water distribution systems often exhibit extreme fragilities during large-scale disasters (e.g., earthquakes) resulting in massive pipe breaks, water contamination, and disruption of service. To monitor and identify potential problems, hidden state information must be extracted from limited and noisy data environments. This requires estimating the operating states of the water system quickly and accurately. We model the water system as a factor graph, characterizing the nonlinearity of fluid flow in a network that is dynamically altered by leaks, breaks and operations designed to minimize water loss. The approach considers a structured probabilistic framework which models complex interdependencies within a high-level network topology. The proposed two-phase approach, which begins with a network decomposition using articulation points followed by the distributed Gauss-Newton Belief Propagation (GN-BP) based inference, can deliver optimal estimates of the system state in near real-time. The approach is evaluated in canonical and real-world water systems under different levels of physical and cyber disruptions, using the Water Network Tool for Resilience (WNTR) recently developed by Sandia National Lab and Environmental Protection Agency (EPA). Our results demonstrate that the proposed GN-BP approach can yield an accurate estimation of system states (mean square error 0.02) in a relatively fast manner (within 1s). The two-phase mechanism enables the scalability of state estimation and provides a robust assessment of performance of large-scale water systems in terms of computational complexity and accuracy. A case study on the identification of "faulty zones" shows that 80% broken pipelines and 99% loss-of-service to end-users can be localized.**

I. Introduction

Water utilities are critical infrastructure and are considered an important lifeline to all local communities, whether regional or worldwide. Often, infrastructure systems that capture, deliver and store water are many decades old, and have become increasingly complex and vulnerable to a wide variety of natural, technological and man-made hazards [1]. Natural disasters and other types of hazards have resulted in different types of water service disruptions, and caused financial, social, environmental and human health consequences [2]. The ability to maintain delivery of water supplies during and after catastrophic events is critical to ensure public safety and welfare.

In this paper, earthquakes are particularly concerning since buried water pipelines are extremely vulnerable to damage from earthquake-caused ground failures [3]. For example, the 1994 magnitude 6.7 Northridge Earthquake (US) damaged over 1,000 distribution pipes and caused 7 days water outages; the 2010 magnitude 8.8 Chile Earthquake damaged 3,000 distribution pipes and caused over a month of water outages; the 2011 magnitude 9.0 Tohoku Earthquake (Japan) damaged thousands of distribution pipes and caused several months of water outages. Once pipeline networks are damaged, water service areas will immediately shutdown via closure of valves. In the mean time, potable water is distributed to customers using mobile water tankers, while service crews are dispatched to repair and restore the system to normal operating conditions. This scenario highlights the need for a holistic and efficient state estimation that produces estimates of the current operating states, and helps detect, locate and prevent possible secondary failures in the water system. An efficient hydraulic state estimation enables timely countermeasures that can mitigate and limit failure propagations, e.g., cascading failures such as release of waste, flooding, and possible contamination.

Current status of water facilities: Despite promising applications, the actual implementation of a real-time monitoring and measurement platform that adapts to perturbations caused by disruptive events is lacking. One reason is that water flow and pressures are generally not monitored in real-time at an individual customer level (i.e., households). Water is a relatively inexpensive resource. Consequently, most water networks are metered only for billing purposes, and there is no intelligent supervisory control and data acquisition (SCADA) system on distribution pipelines. However, civil engineers advocate that the next generation water networks will not be passive water delivery systems, but active highly-distributed event-based control systems [4]. Such a dynamical system will heavily reply on an efficient operating state estimation to facilitate effective water management under dynamic and nondeterministic environmental changes.

Challenges on water system state estimation: The analysis of hydraulic behaviors requires an accurate representation of network topology as well as real-time measurements of water flows and pressures. However, instrumenting the entire system of underground pipelines with sensing devices (pressure transducers and/or flow meters) is both unfeasible (inaccessibility of locations) and expensive. Also urban water systems are densely connected and complex networks crossing diverse geologic conditions, where system performance measurements are highly correlated. It is non-trivial to infer operational states even with a complete observation. When limited numbers of metering devices are available, probabilistic state estimation can serve as a useful technique to "fill-in" missing performance data as well as "smooth-out" noisy measurements. Upon convergence, the optimization should reflect the current state of the water network which, in turn, should allow the prioritization of immediate responses and after-event repairs, and eventually restoration of the system.

As far as we are aware we are the first to provide the systematic study of water system performance estimation under

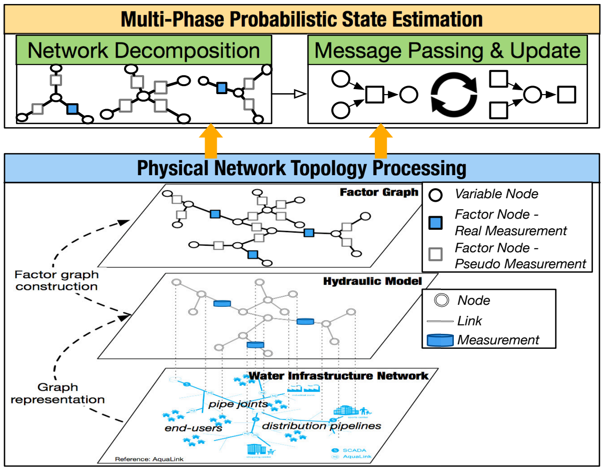


Fig. 1: Design of probabilistic model based state estimation.

limiting conditions, i.e., less than fully instrumented network and noisy data environment, and the first to combine this with distributed graphical model to identify failures caused by disasters. This is a key step to the aforementioned distributed event-based control system. Though inspired by water system resilience under earthquakes, the proposed methodology is designed for generic state estimation and analysis for other pipe based cyber-physical systems and beyond.

Contributions of this paper:

•**Network topology processing (Fig.1):** Design of a methodology that formulates the water system as a hydraulic model with measurement configurations, and transfers it into a factor graph representation to incorporate non-linear hydraulic principles within a structured probabilistic framework - (Sec.III).

•**Probabilistic state estimation (Fig.1):** A novel two-phase approach for improving the speed and accuracy of state estimation on the constructed factor graph: (I) split the water network into conditional independent components using articulation points; (II) estimate the hydraulic states using a distributed Gauss-Newton Belief Propagation based approach - (Sec.IV).

•**Real-world water systems evaluation:** Design of a series of experiments to explore the performance with respect to the time complexity and accuracy of the proposed approach on real-world water systems provided by EPA and Washington Suburban Sanitary Commission (WSSC) - (Sec.V).

•Extensive evaluations under different levels of physical and cyber disruptions, and a case study on the faulty zones identification - (Sec.V).

II. Related Work

Water system resilience: Due to increasingly failure-prone community water services, there has been a lot of research that concentrates on: the design and enhancement of the water systems to reduce the likelihood and impact of asset failures [5–7]; the detection, identification and control of failures during disasters to reduce the cascading impacts [8–10]; the recovery of the infrastructure to return the system to normal operating conditions [11]. With respect to failure identifications, our recent work [9] shows that the multi-leak localization is a non-trivial problem due to highly correlated performance measurements, and in the context of disasters, the unpredictable environmental changes will make it more difficult due to the lack of prior knowledge and the increased number and severity of damages.

Hydraulic simulator: The commonly used demand-driven (DD) hydraulic simulator, like EPANET [12], assumes that customer demands are always met even if the pressure is insufficient to provide the demand. They were not designed to handle sudden failures resulting in inadequate pressure or rapid changes in the system operation. In reality, however, disasters can lead to low pressure conditions that reduce the amount of water delivered to customers. This paper uses WNTR hydraulic simulator - a recently developed water network tool for assessing the resilience of drinking water systems to disasters [13]. WNTR has the pressure-driven demand (PDD) model where the demand supplied to the end-user is a function of the pressure at that node:

$$d = \begin{cases} 0 & p \le P_0 \\ D_f\sqrt{\frac{p-P_0}{P_f-P_0}} & P_0 \le p \le P_f \\ D_f & p \ge P_f \end{cases} \qquad (1)$$

where d is the actual volume delivered to the end-users $(\mathrm{m}^3/\mathrm{s})$, D_f is the expected demand $(\mathrm{m}^3/\mathrm{s})$, p is the gauge pressure inside the pipe (Pa), P_f is the pressure above which the customer receives the expected demand (Pa), and P_0 is the pressure below which they cannot receive any water (Pa).

Graphical model based inference: Graphical models are used to represent the conditional independence relationships among a set of random variables. It has been successfully deployed in many fields, such as computer visions [14], medical diagnostics [15], communication systems [16], and recently power grids [17]. Belief propagation (BP) is an efficient message-passing algorithm that gives exact inference results in linear time for tree-structured graphs [18]. Though widely used, tree-structured models possess limited modeling capabilities, and many stochastic processes arising in real-world applications cannot be well-modeled by cycle-free graphs [19]. Loopy belief propagation (LBP) is an application of BP on loopy graphs, however, the convergence and correctness of LBP are not guaranteed in general. LBP has fundamental limitations when applied to graphs with cycles: local message-passing cannot capture the global structure of cycles, and thus can lead to convergence problems and inference errors. [20] presented a feedback message passing algorithm for an efficient inference in loopy models, which makes use of a special set of vertices whose removal results in a cycle-free graph. Inspired by the exciting results made available by graphical models, we consider a graph representing the water distribution system as a probabilistic model.

III. Modeling Water System in Stochastic Manner

An efficient water system state estimator requires to provide optimal estimations under dynamic and nondeterministic operational and environmental changes. In order to properly model the system stochastic properties and to conduct computationally-tractable inferences, we propose a graphical model description of the water system, which can discover and analyze desired informative data by abstracting the physical nature into a cyber network of nodes and links such that nodes interact with each other along their incident links in a distributed message-passing manner. Specifically, we model hydraulic heads at water nodes as state random variables on the graph vertices, and the edges of the graph determine the interaction of state variables according to the hydraulic physical law (i.e. Hazen–Williams equation [12]). Viewed together, the graphical model is specified by the joint density of hydraulic head random variables in the network for state estimation, subject to the constraints imposed by the fundamental fluid mechanics.

A. Graphical Model of Water Systems

A water system is defined as a undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ (water can flow in both directions) with vertices $\mathcal{V} = \{1, ..., n\}$ that represent nodes (end-users – nodes with demand, and junctions – pipe joints), and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ that represent transmission/distribution pipelines. The set of measurements is defined as $\mathcal{M}$ that is connected to the graph $\mathcal{G}$. There are two kinds of measurements, real measurements and pseudo measurements denoted by $\mathcal{M}_{\mathrm{R}}$ and $\mathcal{M}_{\mathrm{P}}$ respectively, where $\mathcal{M}_{\mathrm{R}} \subseteq \mathcal{M}$, $\mathcal{M}_{\mathrm{P}} \subset \mathcal{M}$, $\mathcal{M}_{\mathrm{R}} \cup \mathcal{M}_{\mathrm{P}} = \mathcal{M}$ and $\mathcal{M}_{\mathrm{R}} \cap \mathcal{M}_{\mathrm{P}} = \emptyset$. Because the number of real measurements will be limited by the cost of installation and maintenance of sensing devices, and to initiate the state estimation algorithm, pseudo measurements will be added in order for the entire system to be "observable". The initial values of pseudo measurements are assigned based on the knowledge of real measurements, and usually with large noise variances. It is worth noting that in the context of seismic hazards, there is no enough priori knowledge (e.g., historical data) that can be used for an appropriate initialization.

The probabilistic measurement model of hydraulic system state estimation is expressed as

$$\mathbf{z} = \mathbf{g}(\mathbf{x}) + \mathbf{u} \tag{2}$$

where the vector $\mathbf{x} = (x_1, ..., x_n)$ represents the probabilistic water system states; the vector $\mathbf{u} = (u_1, ..., u_k)$ where u_i is the additive measurement noise assumed to be independent Gaussian random variable with zero mean, i.e. $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, Σ is a diagonal matrix with the i^{th} diagonal element σ_i^2; $\mathbf{z} = (z_1, ..., z_k)$ is the vector of measurement readings such as flow rate and hydraulic head; and $\mathbf{g} = (g_1(\mathbf{x}), ..., g_k(\mathbf{x}))$ is the vector of non-linear functions associated with each measurement following hydraulic physical laws. Each measurement $\mathcal{M}_i \in \mathcal{M}$ is associated with measured value z_i, measurement noise u_i, and measurement function $g_i(\mathbf{x})$.

The probabilistic state estimator aims to find an estimate $\hat{\mathbf{x}}$ of the true states $\mathbf{x}$ that achieves the maximum posteriori probability (MAP), given the measurement set $\mathbf{z}$ and the priori state information of $\mathbf{x}$ according to the measurement model in (2). It is mathematically expressed as

$$\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} \tag{3}$$

where $p(\cdot)$ represents the probability density function. Assuming that the prior probability distribution $p(\mathbf{x})$ is uniform, and given that the measurement probability distribution $p(\mathbf{z})$ does not depend on $\mathbf{x}$, MAP solution of (3) reduces to maximization of the likelihood function $\mathcal{L}(\mathbf{z}|\mathbf{x})$, which is defined via likelihoods of k independent measurements:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathcal{L}(\mathbf{z}|\mathbf{x}) = \arg\max_{\mathbf{x}} \prod_{i=1}^{k} \mathcal{N}(z_i|\mathbf{x}, \sigma_i^2) \tag{4}$$

One can find the solution of (4) by weighted least squares (WLS) estimator [21]:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_{i=1}^{k} \frac{(z_i - g_i(\mathbf{x}))^2}{\sigma_i^2} \tag{5}$$

To obtain the WLS estimate in (5), we need to first obtain a proper formulation for $\mathbf{g}(\mathbf{x})$ (Sec.III-B), and employ an efficient algorithm to conduct marginalization over $p(\mathbf{x}|\mathbf{z})$ in (3) with respect to $\mathbf{x}$ (Sec.IV).

B. Hydraulic Network Model

The hydraulic system model is defined using the non-linear measurement functions $\mathbf{g}(\mathbf{x})$ that follow the physical laws to connect measured variables with state variables. This model takes hydraulic head denoted by $\mathbf{h}$ as state variables $\mathbf{x}$ (i.e., $\mathbf{x} \equiv \mathbf{h}$), since hydraulic head measurements are essential pieces of information that are required for determining water service availability. Hydraulic head represents the mechanical energy per unit weight of fluid in the system, and is defined on water node i as $h_i = p_i + e_i$, for $i \in \mathcal{V}$, where p_i is the pressure head and e_i is the elevation head at node i. The typical set of measurements $\mathcal{M}$ in water systems includes: the status of valves V_{ij} (open or closed) and flow rates Q_{ij} (cubic meter per second, cms or m^3/s)) at pipes $(i, j) \in \mathcal{E}$, and the hydraulic head h_i (meter, m) at special nodes $i \in \mathcal{V}$ (e.g., reservoir, pump and tank). That is $\mathcal{M} = \{\mathcal{M}_{V_{ij}}, \mathcal{M}_{Q_{ij}}, \mathcal{M}_{h_i}\}$ for $(i, j) \in \mathcal{E}$ and $i \in \mathcal{V}$, where $\{\mathcal{M}_{h_i}\}$ is referred to as the direct measurement $\mathcal{M}_{\mathrm{dir}}$ since it measures state variables directly, and $\{\mathcal{M}_{V_{ij}}, \mathcal{M}_{Q_{ij}}\}$ is referred to as the indirect measurement $\mathcal{M}_{\mathrm{ind}}$. Noted that real/pseudo measurements ($\mathcal{M}_{\mathrm{R}}$ and $\mathcal{M}_{\mathrm{P}}$) and direct/indirect measurements ($\mathcal{M}_{\mathrm{dir}}$ and $\mathcal{M}_{\mathrm{ind}}$) are just two different classifications of measurements that are defined for the convenience to present the proposed approach. For reasons of completeness, a short elaboration of the hydraulic background is given following. Reader may safely skip this part. The measurement functions used in the PDD model are specified based on the measurement types and readings.

For flow-rate measurement $z_{Q_{ij}}$:

$$g_{Q_{ij}}(\cdot) = (1/R_{ij})^{\frac{1}{1.852}} \cdot |h_{L_{ij}}|^{\frac{1}{1.852}} \quad \text{if } z_{Q_{ij}} > 0.0004 \tag{6a}$$

$$g_{Q_{ij}}(\cdot) = (1/(R_{ij} \cdot m)) \cdot h_{L_{ij}} \quad \text{if } z_{Q_{ij}} < 0.0002 \tag{6b}$$

$$g_{Q_{ij}}(\cdot) = a'(h_{L_{ij}}/R_{ij})^3 + b'(h_{L_{ij}}/R_{ij})^2 + c'(h_{L_{ij}}/R_{ij}) + d' \quad \text{if } 0.0002 \leq z_{Q_{ij}} \leq 0.0004 \tag{6c}$$

For hydraulic head measurement z_{h_i}:

$$g_{h_i}(\cdot) = h_i \tag{7}$$

Here $h_{L_{ij}} = |h_i - h_j|$ is the headloss in the pipe (m), and $R_{ij} = 10.667 C^{-1.852} d^{-4.871} L$ is the pipe resistance coefficient (unitless) [12] where C is the Hazen-Williams roughness coefficient (unitless), d is the pipe diameter (m) and L is the pipe length (m). Constant m=0.001 in (6b), and constants a'=$1.524 \cdot 10^{15}$, b'=$-2.530 \cdot 10^{9}$, c'=$1.830 \cdot 10^{3}$, d'=$-7.695 \cdot 10^{-5}$ in (6c), which are calculated using polynomial curve fitting. In (6), different functions are used according to the values of flow-rate measurements. Because when $Q_{ij} \approx 0$, it can cause the Jacobian of the set of hydraulic equations to become singular, and [13] proposed to split the domain of Q into several segments to create a piecewise smooth function.

C. Factor Graph Construction

To solve the optimization problem in (3), we instead need to find an optimal solution of (4) in an efficient manner. We first construct a factor graph to describe a factorization of the likelihood function $\mathcal{L}(\mathbf{z}|\mathbf{x})$. Factor graphs comprised of the set of variable nodes and factor nodes have been widely used to represent factorization of a probability distribution function, enabling efficient computations [22].

As shown in Fig.1, a factor graph can be formed from the hydraulic model, where the variable node characterizes the probability distribution of the hydraulic head at nodes, and the factor node is determined by the set of measurements. The pseudo measurements will be filled in based on the real measurement readings, to make the entire system "observable". That is, the vector of state variables $\mathbf{h}$ defines the set of variable nodes $\mathcal{X} = \{h_1, ..., h_n\}$, while the set of measurements $\mathcal{M}$ defines the set of factor nodes $\mathcal{F} = \{f_1, ..., f_k\}$. A factor node f_i connects to a variable node $x_s \in \mathcal{X}$ if and only if the state variable h_s is an argument of the corresponding measurement function $g_i(\mathbf{x})$ according to (6) and (7). In this manner, hydraulic head and flow rate are modeled separately, and their correlations can be captured in the corresponding factor nodes.

D. Leak Event Model

Leaks can cause large changes in network hydraulics, and we use WNTR to model the pipe breaks [13]. A new junction is added onto each pipeline to model the event node, and if a pipe breaks, its event node will be used as the leaky point. In WNTR, the mass flow rate of fluid through the hole, d^{leak}, is expressed as:

$$d^{\text{leak}} = C_d A \sqrt{2\rho} p^{\alpha} \tag{8}$$

where C_d is the discharge coefficient (unitless) with default value 0.75, A is the area of the hole (m^2), ρ is the density of the fluid (kg/m^3), p is the pressure (Pa) computed using elevation and hydraulic head (m), fluid density and acceleration due to gravity (m/s^2), and α is set to 0.5 for large leaks out of steel pipes.

IV. A Multi-phase Probabilistic State Estimation

The industry paradigm is shifting from the traditionally deterministic model based centralized monitoring architecture to probabilistic model based highly distributed interactive data and resource management. Therefore, a multi-phase, distributed implementation of the state estimator is likely to be the preferred approach, which enables a fast and accurate hydraulic behavior assessment.

The belief propagation algorithm efficiently calculates the marginal distribution for each state variables by passing messages (a) from a variable node x_s to a factor node f_i and (b) from a factor node f_i to a variable node x_s. Variable and factor nodes locally process the incoming messages and calculate outgoing messages in a distributed manner. Under the assumption that measurement errors $\{u_i\}$ follow a Gaussian distribution, the probability density function of $\{x_s\}$ and $\{f_i\}$ are Gaussian. The passing-message can then be characterized by mean and variance. The marginal inference provides marginal probability distributions $p(\mathbf{x}|\mathbf{z})$ that is used to find an estimate $\hat{\mathbf{x}}$ of the true states $\mathbf{x}$. It is well-known that the key assumption of the BP algorithm converging to the optimal solution is that the applying graph has no cycle, i.e., tree-structured [22]. Such assumption often does not hold for most water distribution networks, which are locally dense and contain loops. In addition, due to the non-linearity of the measurement functions, the BP based approach will be sequentially applied over the factor graph until the stop criterion is satisfied, which increases the time complexity of convergence. Our experience shows that the inference on a large-scale water systems can take more than 30min to converge, which is too slow for an on-line state estimation, especially under seismic events. Ideally, state estimation should run at the scanning rate or at least less than the sampling rate of industrial metering devices (15min) [23], to handle the new measurement as soon as it is delivered from telemetry to the computational unit. To overcome these limitations, we proposed a two-phase approach: (I) decomposes the hydraulic network into conditional independent connected components, and (II) performs the GN-BP based inference on each of them. It is worth noting that Phase II can be done in parallel for all components that can further reduce the time complexity.

A. Phase I: Network Decomposition by Articulation Points

Phase I aims to generate several disjoint connected components where each one has a moderate size and they are conditional independent given specific nodes being observed. Articulation points (APs) are vertices in an undirected connected graph, whose removal along with the removal of their incident links disconnects the graph. The APs can divide a graph into several biconnected components, where a biconnected component is a connected and "nonseparable" subgraph, meaning that if any one of vertices is removed, the subgraph will remain connected. Noted that a biconnected graph has no APs. In the context of the water system, it is important to know the relatively structural prominence of nodes or links to identify key elements in the network. A water node that is an articulation point often has higher information centrality and will be considered as the critical location - a single point whose failure would cause network disconnection [24]. A water network subzone that is a biconnected graph is considered as more resilient and less susceptible to damage and perturbation [5]. The existing of the alternative supply

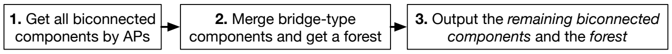
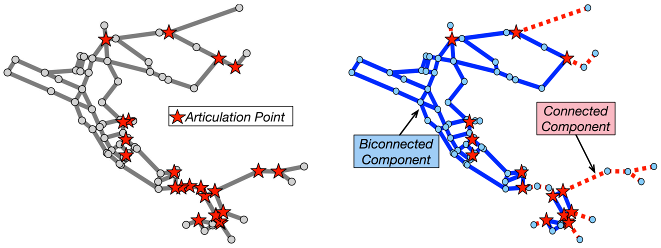

1. Get all biconnected components by APs
2. Merge bridge-type components and get a forest
3. Output the remaining biconnected components and the forest

Fig. 2: The work flow of Phase I - network decomposition.



(a) Novato water network (b) Network decomposition

Fig. 3: Decomposition of (a) Novato water network into (b) 3 biconnected components (solid lines) and 9 (bi)connected components (dashed lines).

paths provides a two-fold redundancy and improve system robustness and resilience by avoiding critical locations and network bottlenecks. This redundancy, in turn, may improve the performance of state estimations, since the incorrect measurement on one path may be compensated by measurements from alternative paths, and an estimate on the variable node can be derived by the cooperation of messages from multiple incident links. The work flow of Phase I is shown in Fig.2, and summarized in following:

1. Find the articulation points and biconnected components of the water network. For example, in Fig.3a, this is the hydraulic network of the water system operated by North Marin Water District, and nodes that are labeled by stars are APs, which are used to generate all biconnected components.

2/3. One observation on branched water networks is that there are many bridge-type biconnected components, that consist of a single edge. The purpose of the network decomposition is to reduce the time complexity while obtain the optimal state estimates. This will require to use as many sensing devices as the number of split points to make disjoint components conditionally independent. Without using many sensors, we do not want to split the network into many small parts (components with a single edge). Thus, we merge those bridge-type biconnected components and obtain a disjoint union of trees (it can be proved by contradiction, which is not shown to save space). For example, in Fig.3b, the network decomposition outputs 3 biconnected components that are not bridge-type, and 9 (bi)connected components after merging.

Without losing the structural information after decomposition, the node that is the intersection of 2 or more generated disjoint connected components needs to be observable. In Fig.3b, nodes labeled by stars are intersection points and will be observed. That is $\mathcal{M}_{h_i} \in \mathcal{M}_{\mathrm{R}}$ if node i is such a point. The paper focuses on efficient state estimation by utilizing limited and noisy measurements. The problem to find a minimum amount of measurements that are required to deliver optimal system states is out of the scope. The corresponding study is referred to as observability analysis.

B. Phase II: Hydraulic State Inference

In Phase II, we describe an efficient Gauss-Newton Belief Propagation (GN-BP) based hydraulic state estimation algorithm, which converges to the optimal inference results for all water nodes in a reasonable time. To speed up the convergence time on loopy networks, we then introduce a feedback vertex set (FVS) selection criterion to break all loops, and a modified version of the proposed algorithm to use FVS.

1) GN-BP based on-line inference: The BP based algorithm allows the state of end-users to be estimated in a distributed, message-passing manner with the neighboring end-users where flow meters are located. Then the aggregated information is communicated in a bottom-up way to the backbone system operator for driving system control. However, due to the non-linearity of measurement functions $\mathbf{g}(\cdot)$ in (6), the closed-form expressions for certain classes of BP messages cannot be obtained. Therefore we integrate Gauss-Newton (GN) method with BP based inference to solve the WLS problem in (5). The GN algorithm is used to solve non-linear least squares problem by minimize a sum of squared function values and it has advantage that second derivatives, which can be challenging to compute, are not required.

Based on k number of measurements $\mathcal{M}$, the solution of (5), which is a vector of n state variables $\hat{\mathbf{x}} \equiv \hat{\mathbf{h}}$, can be found using the GN method [21]:

$$[\mathbf{J}(\mathbf{x}^{(\nu)})^T\mathbf{W}\mathbf{J}(\mathbf{x}^{(\nu)})]\cdot\Delta\mathbf{x}^{(\nu)} = \mathbf{J}(\mathbf{x}^{(\nu)})^T\mathbf{W}\mathbf{r}(\mathbf{x}^{(\nu)}) \tag{9a}$$

$$\mathbf{x}^{(\nu+1)} = \mathbf{x}^{(\nu)} + \Delta\mathbf{x}^{(\nu)} \tag{9b}$$

where $\nu = \{1,2,3,...\}$ is the iteration index, and at each iteration step ν, $\Delta\mathbf{x}^{(\nu)} \in \mathbb{R}^n$ is the vector of increments of state variables $\mathbf{x}$, $\mathbf{J}(\mathbf{x}^{(\nu)}) \in \mathbb{R}^{k\times n}$ is the Jacobian matrix of measurement functions $\mathbf{g}(\mathbf{x}^{(\nu)})$, $\mathbf{W} \in \mathbb{R}^{k\times k}$ is a diagonal matrix containing inverses of measurement variances, i.e. $\mathbf{W} = \Sigma^{-1}$, and $\mathbf{r}(\mathbf{x}^{(\nu)}) = \mathbf{z} - \mathbf{g}(\mathbf{x}^{(\nu)})$ is the vector of residuals, i.e., the difference between measured and estimated values. The Jacobian expressions corresponding to $g_{Q_{ij}}(\cdot)$ and $g_{h_i}(\cdot)$ can be computed based on (6) and (7).

Consider the GN method in (9) where, at each iteration ν, the algorithm returns a new estimate $\hat{\mathbf{x}}$, and (9a) represents the minimization problem:

$$\min_{\Delta\mathbf{x}^{(\nu)}} ||\mathbf{W}^{1/2}[\mathbf{r}(\mathbf{x}^{(\nu)}) - \mathbf{J}(\mathbf{x}^{(\nu)})\Delta\mathbf{x}^{(\nu)}]||_2^2 \tag{10}$$

Hence, the probability measurement model (2) can be re-defined as a group of linear equations:

$$\mathbf{r}(\mathbf{x}^{(\nu)}) = \phi(\Delta\mathbf{x}^{(\nu)}) + \mathbf{u} \tag{11}$$

where $\phi(\Delta\mathbf{x}^{(\nu)}) = \mathbf{J}(\mathbf{x}^{(\nu)})\Delta\mathbf{x}^{(\nu)}$ comprises linear functions. The MAP solution of (3) can be reduced to maximum likelihood problem, and the equation (4) can be re-defined as an iterative optimization problem:

$$\begin{aligned}\Delta\hat{\mathbf{x}}^{(\nu)} &= \arg\max_{\Delta\mathbf{x}^{(\nu)}} \mathcal{L}(\mathbf{r}(\mathbf{x}^{(\nu)})|\Delta\mathbf{x}^{(\nu)})\\ &= \arg\max_{\Delta\mathbf{x}^{(\nu)}} \prod_{i=1}^{k} \mathcal{N}(r_i(\mathbf{x}^{(\nu)})|\Delta\mathbf{x}^{(\nu)}, \sigma_i^2) \qquad (12a)\end{aligned}$$

$$\mathbf{x}^{(\nu+1)} = \mathbf{x}^{(\nu)} + \Delta\hat{\mathbf{x}}^{(\nu)} \tag{12b}$$

Next, we show that the solution of (12) can be efficiently obtained using BP based algorithm applied over the underlying factor graph introduced in Sec.III-C. The factor graph constructed by the factorization of the likelihood function in (12a) is slightly different from the one in (4). The set of variables nodes is defined as the increments of state variables instead of the state variable itself, i.e. $\mathcal{X} = \{\Delta h_1, ..., \Delta h_n\}$, while the set of factor nodes is defined as before based on the measurements $\mathcal{M}$, i.e. $\mathcal{F} = \{f_1, ..., f_k\}$. The factor node f_i connects to a variable node Δx_s if and only if Δx_s is an argument of the corresponding function $\phi_i(\Delta\mathbf{x})$, that is if the state variable h_s is an argument of the measurement function $g_i(\mathbf{x})$. The BP algorithm on factor graphs proceeds by passing two types of messages along the edges: from variable nodes to factor nodes and from factor nodes to variable nodes. BP messages represent "beliefs" about variable nodes, thus a message that arrives or departs a variable node is a probability distribution of the random variable associated with this node. The "beliefs" will be iteratively updated by incoming messages and propagated by outgoing messages until the stopping criterion is satisfied.

Message from a variable node to a factor node: Since we consider the Gaussian graphical model, the message from a variable node Δx_s to a factor node f_i at iteration step τ can be characterized by mean $r^{(\tau)}_{\Delta x_s \to f_i}$ and variance $\sigma^{2\ (\tau)}_{\Delta x_s \to f_i}$:

$$r^{(\tau)}_{\Delta x_s \to f_i} = \left(\sum_{f_a \in \mathcal{F}_s \backslash f_i} \frac{r^{(\tau-1)}_{f_a \to \Delta x_s}}{\sigma^{2\ (\tau-1)}_{f_a \to \Delta x_s}} \right) \cdot \sigma^{2\ (\tau)}_{\Delta x_s \to f_i} \tag{13a}$$

$$\frac{1}{\sigma^{2\ (\tau)}_{\Delta x_s \to f_i}} = \sum_{f_a \in \mathcal{F}_s \backslash f_i} \frac{1}{\sigma^{2\ (\tau-1)}_{f_a \to \Delta x_s}} \tag{13b}$$

where $\mathcal{F}_s$ is a set of factor nodes incident to Δx_s, and $\mathcal{F}_s \backslash f_i$ is a subset by excluding the factor node f_i. The incoming messages used for calculation are obtained in previous iteration $(\tau - 1)$.

Message from a factor node to a variable node: Similarly, the message from a factor node f_i to a variable node Δx_s can be characterized by mean $r_{f_i \to \Delta x_s}$ and variance $\sigma^2_{f_i \to \Delta x_s}$:

$$r^{(\tau)}_{f_i \to \Delta x_s} = \frac{1}{C_{i,\Delta x_s}} \left(r_i - \sum_{\Delta x_b \in \mathcal{X}_i \backslash \Delta x_s} C_{i,\Delta x_b} \cdot r^{(\tau)}_{\Delta x_b \to f_i} \right) \tag{14a}$$

$$\sigma^{2\ (\tau)}_{f_i \to \Delta x_s} = \frac{1}{C^2_{i,\Delta x_s}} \left(\sigma^2_i + \sum_{\Delta x_b \in \mathcal{X}_i \backslash \Delta x_s} C^2_{i,\Delta x_b} \cdot \sigma^{2\ (\tau)}_{\Delta x_b \to f_i} \right) \tag{14b}$$

where $\mathcal{X}_i$ is a set of variable nodes incident to f_i, and $\mathcal{X}_i \backslash \Delta x_s$ is a subset by excluding the variable node Δx_s. $C_{i,\Delta x_p}$ for $\Delta x_p \in \mathcal{X}_i$ are Jacobian elements of the measurement function $g_i(\cdot)$ associated with f_i:

$$C_{i,\Delta x_p} = \frac{\partial g_i(\cdot)}{\partial x_p} \tag{15}$$

Marginal inference: The marginal of the state variable is the estimated value of the increment, which will be calculated when $r_{f_i \to \Delta x_x}$ and $\sigma^2_{f_i \to \Delta x_x}$ converge:

$$\Delta \hat{x}_s = \left(\sum_{f_i \in \mathcal{F}_s} \frac{r_{f_i \to \Delta x_x}}{\sigma^2_{f_i \to \Delta x_x}} \right) \cdot \left(1 \Big/ \sum_{f_i \in \mathcal{F}_s} \frac{1}{\sigma^2_{f_i \to \Delta x_x}} \right) \tag{16}$$

The GN-BP based inference subroutine is summarized in Algorithm 1. To present the algorithm precisely, we define different types of factor nodes based on the measurements $\mathcal{M}$. The factor nodes that correspond to real measurements $\mathcal{M}_{\mathrm{R}}$ are real factor nodes $\mathcal{F}_{\mathrm{R}} \subseteq \mathcal{F}$, and similarly pseudo factor nodes $\mathcal{F}_{\mathrm{P}} \subset \mathcal{F}$ are associated with pseudo measurements $\mathcal{M}_{\mathrm{P}}$. The direct/indirect measurements $\mathcal{M}_{\mathrm{dir}}$ and $\mathcal{M}_{\mathrm{ind}}$ are represented by direct/indirect factor nodes $\mathcal{F}_{\mathrm{dir}} \subseteq \mathcal{F}$ and $\mathcal{F}_{\mathrm{ind}} \subseteq \mathcal{F}$ respectively. In Algorithm 1, the outer loop stops when the difference on estimated values is less than a very small number ϵ_O, and the inner loop stops when the difference on BP messages is less than a very small number $\epsilon_I(\nu)$ that varies with the outer iterations.

Algorithm 1 The distributed GN-BP based inference

```
Input factor graph $G_f(\mathcal{X}, \mathcal{F})$, state variables $\mathbf{x}$ with initial
    values, state threshold $[x_L, x_H]$, measured values $\mathbf{z}$
Output estimated states $\hat{\mathbf{x}}$

    /* Outer state update loop $\nu = 1, ...;\ \tau = 0$ */
while $|\mathbf{x}^{(\nu)} - \mathbf{x}^{(\nu-1)}| < \epsilon_o$ do
    for $f_i \in \mathcal{F}$ do
        $r_i^{(\nu)} = z_i - g_i(\mathbf{x}^{(\nu)})$ using (6) and (7)
    end for
    for $\Delta x_s \in \mathcal{X}$ do
        if $f_{x_s} \in \mathcal{F}_{\mathrm{R}}$ then
            $r^{(\nu,\tau=0)}_{\Delta x_s \to f_i} = r_s^{(\nu)};\ \sigma^{(\nu,\tau=0)}_{\Delta x_s \to f_i} = \epsilon_\sigma$ for $f_i \in \mathcal{F}_s$
        else
            $r^{(\nu,\tau=0)}_{\Delta x_s \to f_i} = \epsilon_r;\ \sigma^{(\nu,\tau=0)}_{\Delta x_s \to f_i} = \infty$ for $f_i \in \mathcal{F}_s$
        end if
    end for
    /* Inner message update loop $\tau = 1, ...$ */
    while $|r^{(\tau)}_{f \to \Delta x} - r^{(\tau-1)}_{f \to \Delta x}| < \epsilon_I(\nu)$ do
        Compute $r^{(\tau)}_{f_i \to \Delta x_s}$, $\sigma^{2\ (\tau)}_{f_i \to \Delta x_s}$ using (14)
        Compute $r^{(\tau)}_{\Delta x_s \to f_i}$, $\sigma^{2\ (\tau)}_{\Delta x_s \to f_i}$ using (13)
    end while
    /* Marginal inference */
    Compute $\mathbf{x}^{(\nu+1)} = \mathbf{x}^{(\nu)} + \Delta\hat{\mathbf{x}}^{(\nu)}$ using (16)
    /* State validation */
    for $\Delta x_s \in \mathcal{X}$ do
        if $x_s^{(\nu+1)} \notin [x_L, x_H]$ then $x_s^{(\nu+1)} = \bar{\mathbf{x}}^{(\nu+1)}$
        end if
    end for
end while
```

2) Feedback vertex set selection: The non-bridge biconnected components generated by Phase I contain loops - it provides path redundancy for resilience but can also add difficulty for BP inference. We consider a particular set of nodes called a feedback vertex set (FVS) denoted by $\mathcal{F}$ whose removal breaks all the cycles and results in a cycle-free graph, which is inspired by [20]. The algorithm proposed in [20] runs in time $\mathcal{O}(m^2 n)$ where m is the number of feedback nodes and n is the total number of nodes. When m is bounded by a small number, this is a significant reduction from $\mathcal{O}(n^3)$ of LBP.

Many of water systems in US have a hybrid network topology (a combination of loops and branches), and thus it is possible to find a FVS with a reasonable size to remove all loops in a water network [25]. Without losing much structural information in terms of the removing nodes, the goal is to find a minimum FVS to break all cycles and enable a fast LBP convergence on the remaining graph. After Phase I, all non-cycle-free graphs are biconnected, meaning that all nodes in the graph are part of a cycle. To find an optimal FVS with a small size, we propose a greedy heuristic algorithm: one feedback node is chosen at each iteration, and at each stage we examine the graph excluding the nodes already included in the FVS $\mathcal{F}$ and select the node with the largest degree. We then remove the node along with its incident edges and put it into $\mathcal{F}$. The same procedure will be continued on the remaining graph $\mathcal{T}$ until it is empty. The motivation for this method is given that since the number of cycles is reduced with the removal of nodes, it makes sense to choose nodes with the highest degrees to remove more cycles at each iteration. The selection algorithm is summarized in Algorithm 2.

To utilize FVS, we use a special update scheme for feedback nodes. Consider the loopy graph in Fig.4a, FVS identified by Algorithm 2 contains two feedback nodes. (1) Algorithm 1 is first applied on the cycle-free graph G_T by removing feedback nodes and its incident edges (Fig.4b). We obtain inaccurate "partial states" for the nodes in the cycle-free graph. (2) We then compute the inference results for the feedback nodes by applying Algorithm 1 on the subgraph of feedback nodes, their incident edges and neighbors G_T (Fig.4c). (3) Last, we make corrections to the "partial states" of the non-feedback nodes by running Algorithm 1 on the cycle-free graph again (Fig.4d). Optimal inference results are obtained for all nodes. Noted that in (2) and (3), the initial states of nodes that are neighbors of feedback nodes (e.g., nodes appear both in Fig.4c and 4d) are determined by the previous stage.

Algorithm 2 The FVS selection criterion

```
Input biconnected component $G$ generated by Phase I
Output a FVS $\mathcal{F}$
Objective find a optimal $\mathcal{F}$ to break all cycles in $G$

Let $\mathcal{F} = \emptyset$ and $\mathcal{T} = G$
while $\mathcal{T}$ is not empty do
    (a) Get node degrees of $\mathcal{T}$
    (b) Put the node with the highest degree into $\mathcal{F}$ and
        remove it with its incident edges from $\mathcal{T}$
    (c) Clean up $\mathcal{T}$ by eliminating all tree branches.
end while
```

3) Hydraulic state estimation: Given a water network, Phase I first splits it into several disjoint connected components, whose hydraulic states are then estimated by Phase II. To execute the GN-BP based inference on a graph $G(V, E)$, the state variables on nodes V need to be initialized and the pseudo indirect measurements on edges E need to be set. The initial values of observed state variables equal to their real measured values, while for those that are not observed, their initial values are set to the average of the observed states. The value of a pseudo indirect measurement is set to the same value as its closest real indirect measured value. The state estimator is summarized in Algorithm 3.

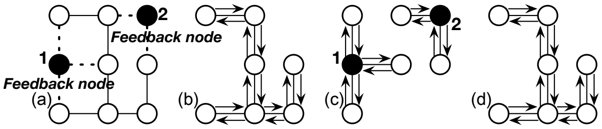


Fig. 4: The message update scheme with 2 feedback nodes. The nodes here represent variable as well as factor nodes.

Algorithm 3 The multi-phase hydraulic state estimation

```
Input water network $\mathcal{G}$, measurements $\mathcal{M}$
Output estimated hydraulic heads of water nodes

/* After Phase I */
for each connected component $G_c(V, E)$ of $\mathcal{G}$ do
/* Initialization */
    For $\forall i \in V$: if $\mathcal{M}_i \in \mathcal{M}_{\mathrm{R}}$, $x_i = z_i$; else $x_i =$ average
    value of $\mathcal{M}_i \in \mathcal{M}_{\mathrm{dir}} \cap \mathcal{M}_{\mathrm{R}}$.
    For $\forall j \in E$: if $\mathcal{M}_j \in \mathcal{M}_{\mathrm{P}}$, $z_j = z_c$, where $c =$
    $\arg\min_l$ distance$(j, l)$ and $\mathcal{M}_l \in \mathcal{M}_{\mathrm{ind}} \cap \mathcal{M}_{\mathrm{R}}$.
/* Inference */
    if Use FVS then
        Find FVS of $G_c$, and get $G_T$ and $G_F$.
        Alg.1 on factor graphs $G_f(\mathcal{X}, \mathcal{F})$ of $G_T$ and $G_F$.
    else
        Alg.1 on factor graph $G_f(\mathcal{X}, \mathcal{F})$ of $G_c$.
    end if
end for
```

V. Experimental Study

In this section, we examine the effectiveness of the proposed multi-phase state estimation algorithm on a small-scale canonical water network and two real-world water systems where they have different configurations on topology, network size and pressure range (Fig.5). The inference approaches we compared include: GN-BP - directly estimate the states of the entire network by Phase II; FVS+GN-BP - find FVS and estimate the states using FVS by Phase II; Decomp+GN-BP - split the network by Phase I and estimate the states of each component by Phase II; Decomp+FVS+GN-BP - find FVS for each components generated by Phase I and estimate the states using FVS by Phase II. We begin by describing the setup under which the experiments are conducted, and introduce the performance metrics and the results.

A. Experimental Setup

Figures 5b/5c show the real-world water systems that are used to evaluate the scalability of the proposed approach.

NET3 is the service area, containing multiple pressure zones, of North Marian Water District (NMWD) provided by EPA, and WSSC-SUBNET is a single pressure zone of WSSC service area provided by WSSC. Figure 6 illustrates that NET3 has relatively large variances on both hydraulic heads and flow rates, where the variance on hydraulic heads is $5.989\mathrm{m}$ compared with $1.386 \cdot 10^{-4}\mathrm{m}$ of WSSC-SUBNET, and the variance on flow rates is $1.5 \cdot 10^{-2}\mathrm{m}^3/\mathrm{s}$ compared with $2.755 \cdot 10^{-6}\mathrm{m}^3/\mathrm{s}$ of WSSC-SUBNET.

We use WNTR to simulate the earthquake impacts on the water distribution system by generating a earthquake event with magnitude 5.5 (unitless) and shallow depth $5000\mathrm{m}$ at a random location. The pipe failure probabilities are then calculated using the attenuation model of peak ground acceleration (PGA) where $\mathrm{PGA} = 403.8 \times 10^{0.265\mathrm{M}}(\mathrm{R}+30)^{-1.218}$ and the fragility curve that defines the probability of exceeding a damage state as a function of PGA (Fig.7). Here M is the earthquake magnitude and R is the distance to the epicenter (km). The leak diameter of the broken pipe is generated following the uniform distribution between 0.15 and 0.3 of the pipe diameter. Each pipeline may have different lengths and diameters. We assume that 50% pipelines are instrumented by flow meters with varied noise variances, and the critical points are instrumented by SCADA monitoring systems with a small noise variance, since, in reality, SCADA systems are less susceptible to physical and cyber failures. The critical points include reservoirs, tanks, pumps, and those articulation points that are used to split the network in Phase I. There are 19 ($\approx$ 8%) and 48 ($\approx$ 8%) critical nodes in NET3 and WSSC-SUBNET respectively.

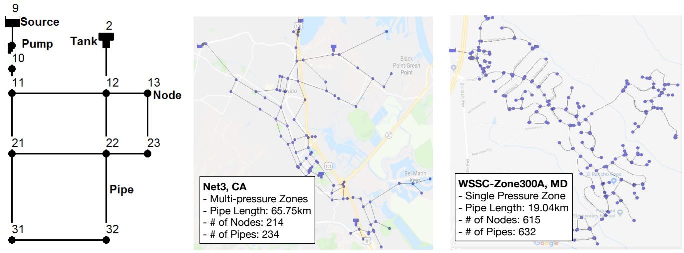
B. Performance Metrics

The effectiveness of the hydraulic state estimator is first evaluated in terms of the **time complexity** and the **accuracy**. The sampling rate of industrial flow meters is 15min, meaning that the estimator needs to infer the current system states in less than 15min before new measurements arrive. The faster the estimation converges to the optimal values, the quickly the countermeasures can be adopted. The time complexity is evaluated by the number of iterations to converge, and the accuracy is evaluated by the mean square error (MSE). We also consider two resilience metrics: **pipe damage state** and **water service availability**. There are two damage states: "no damage" and "break", and the goal is to identify faulty zones where pipes are in the "break" state. To demonstrate this identification performance, we define True Positive (TP) as the number of predicted broken pipes within the distance threshold to the leaky points divided by the number of true broken pipes, and False Positive (FP) as the number of predicted broken pipes not in the distance threshold divided by the number of predicted broken pipes. A higher TP with a lower FP means a better performance. Water service availability at each node is computed as $V_i/\hat{V}_i$ for $i \in \mathcal{V}$, where V_i the actual water volume (m^3) received at node i and $\hat{V}_i$ is the expected water volume (m^3) received at node i. The water service availability can be influenced by failures and operational changes after a disaster, and it is important to estimate the received volume at end-users to localize the areas where they loss the access to the supply and facility. Precision and recall are used to demonstrate this estimation performance.

C. Complexity and Accuracy of Hydraulic State Estimation

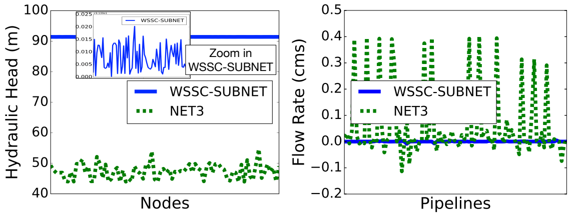
In this section, the proposed state estimation approach is validated through a detailed simulation study on a small-scale canonical water network and two real-world water systems. We first demonstrate the performance of the approach on the small-scale water network (Fig. 5a), where the flow rate measurements have a large noise variance ($5\mathrm{m}^3/\mathrm{s}$). Figures 8a/8b illustrate that our approach can generate an accurate estimation on hydraulic heads ($\mathrm{MSE} = 0.02$) with a flat start of $305\mathrm{m}$ in a fast manner (converge in 1s). The reason for simulating a flat start is to mimic the water system environment with strong state variation, making typical initial guess method (from the last static state estimation) less informative for the new estimation process. Figure 8c shows the fact that the proposed GN-BP based estimation approach does not rely heavily on the initial guess.

In the following set of simulations, we study the performance under different levels of sensing and infrastructure disruptions, with respect to the number of iterations to converge and the mean square error of the headloss at pipes. Headloss is the absolute difference on hydraulic heads at the end nodes of a pipeline (6), and it can be used to identify pipe failures, since the headloss of a broken pipeline will increase due to the leaking. An earthquake event is generated on NET3, and it causes 10% (≈ 20) pipe failures with varied leak volumes. Different percentages of the disrupted sensing devices are simulated based on the locations of broken pipelines, where those disrupted meters are considered with a large noise variance $2.5 \cdot 10^{-3}\mathrm{m}^3/\mathrm{s}$ compared with the noise variance $10^{-6}\mathrm{m}^3/\mathrm{s}$ of non-disruptive meters. Figures 9a/9b illustrate that compared with FVS+GN-BP, the proposed two-phase approach (Decomp+) can dramatically reduce the time complexity using less number of iterations and improve the accuracy with lower errors. Because the network decomposition separates the network into several subnets, and each of them has a relatively small size. In addition, each subnet can yield a better initialization by capturing its local information compared with the initialization over the entire network, since the initial states of unobserved state variables are set to the average of the directly observed values in its connected component. The performance of GN-BP (without FVS and Decomp) on NET3 is not shown, because it takes more than $30\mathrm{min}$ to converge and makes the estimation too long to be meaningful. The approach of Decomp+FVS+GN-BP can further decrease the computational complexity but with the cost of less accuracy. Because the removal of feedback nodes breaks the original graph structure, which can result in the loss of topological information. Figures 9c/9d show the estimation results on WSSC-SUBNET where there are 5% (≈ 30) pipe failures, and the noise variances of disrupted and non-disruptive meters are $10^{-4}\mathrm{m}^3/\mathrm{s}$ and $10^{-8}\mathrm{m}^3/\mathrm{s}$ respectively. Likewise, the two-



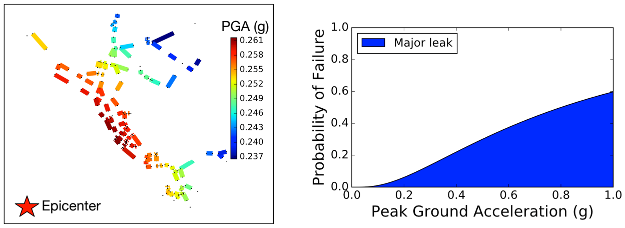
(a) Sample Net (b) NET3 (c) WSSC-SUBNET

Fig. 5: (a) A sample network provided by EPA, and two real-world water systems: (b) water distribution system operated by NMWD and (c) a single pressure zone operated by WSSC.



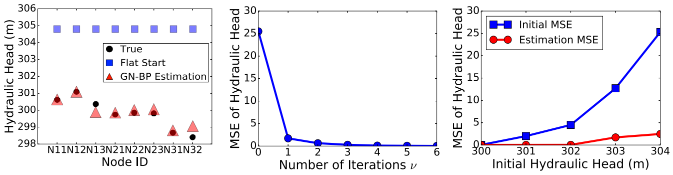
(a) Hydraulic Head (b) Flow Rate

Fig. 6: Comparisons on (a) hydraulic head (m) and (b) flow rate (m^3/s) of NET3 and WSSC-SUBNET water systems at 11am under normal condition.



(a) Peak Ground Acceleration (b) Fragility Curve

Fig. 7: (a) PGA of pipelines for a magnitude of 5.5 earthquake. (b) Fragility curve for pipe damage.



(a) Hydraulic Head (b) MSE (c) MSE vs Initial

Fig. 8: Sample Network - (a) estimated hydraulic heads with a flat start 305m and (b) mean square error versus number of outer iterations. (c) Errors versus initial head values.

phase approach yields a better performance with less time complexity and high accuracy.

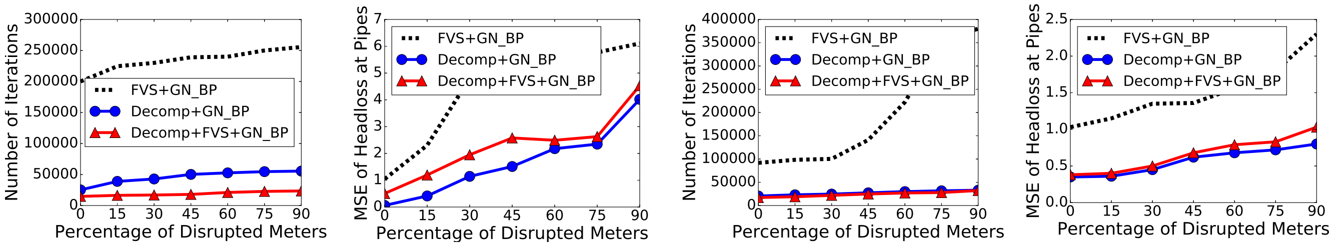
In Fig.10, the comparison is performed on NET3 under different levels of infrastructure (pipe) failures, where the noise variance of meters is $10^{-6}\mathrm{m}^3/\mathrm{s}$. It can be seen from Fig.10a, that as the number of pipe failures increasing, our approach is able to converge in approximately same amount number of iterations. Fig.10b shows that Decomp+GN-BP can achieve a 0.71 MSE of headloss when 9% (≈ 22) pipe breaks.

D. Faulty Zones Identification

This section explores the hydraulic state estimation for the faulty zones identification. To enable the ability of a system to minimize disruptions and return to the normal function after disruptive incidents, it is important to quickly detect and localize faulty regions such that this information can be used by water agencies and city planners for damage control, community notifications and evacuation plans. We use WSSC-SUBNET in this case study, since the coordinates of its water nodes are their true geo-locations. In this study, 5 major leak events with different large leak volumes are generated at random locations, which are indicated by circles in Fig.11a. According to the modeling of leak events (Sec.III-D), each pipeline is split into two segments, and the leaky point on a pipe can cause the difference on the headloss between these two segments. The key observation is that the true headloss can be used to identify the broken pipes with $\mathrm{TP} = 1$, $\mathrm{FP} = 0$. In Fig.11a, our approach is able to localize damaged pipes with $\mathrm{TP} = 0.8$ of a distance threshold $200\mathrm{m}$. It can be seen that though the predicted locations are not the exact leaky points, it can help narrow down and target the potential faulty regions such that detailed examinations can be executed effectively. Water service availability is another important resilience metric to identify the places where they loss the access to the facility. The delivered water volume can be calculated using (1) where the node pressure is computed using its elevation and estimated hydraulic head. Figure 11b shows that 99% of loss-of-service nodes can be localized. With this information, the portable water can be delivered for drinking and other sanitary purposes.

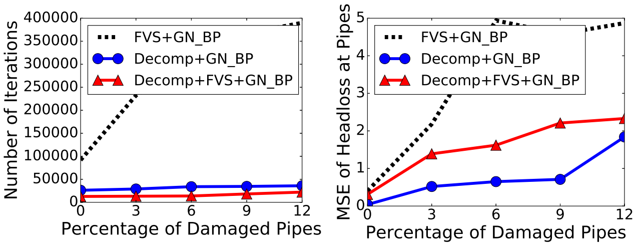
VI. Concluding Remarks

This paper presents a novel probabilistic state estimation approach for fault identification, which combines physical constraints with structured nondeterministic information into a single cyber-physical graphical model, for water distribution systems. We consider the real-world water systems under disasters, where different percentages of physical infrastructures and monitoring devices are disrupted, and the proposed two-phase mechanism is scalable for state estimation in large-scale water networks. This paves the way for distributed control for next generation water systems. One can imagine that the estimated damaged state information is fed into a control decision process, where the pipe network would be reconfigured using remotely controlled valves to save both water and customer demand issues [8]. In future work, we aim toward a systematic middleware design and implementation, which will leverage dynamic data from multiple information sources including failure models and fragility curves as a source for priori knowledge, hydraulic state estimates and human inputs as a source for posterior information, and simulation/modeling



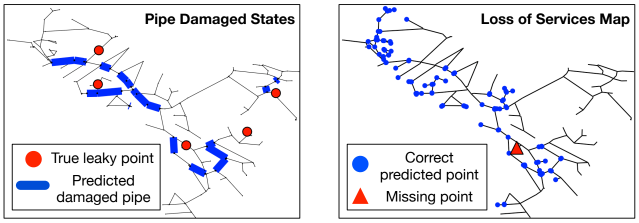
(a) NET3: Number of Iterations (b) NET3: Mean Square Error (c) WSSC-SUBNET: Iterations (d) WSSC-SUBNET: MSE

Fig. 9: Total number of iterations to converge and mean square error of headloss at pipes versus the percentage of sensor failures on NET3 and WSSC-SUBNET.



(a) Number of Iterations (b) Mean Square Error

Fig. 10: NET3 - (a) number of iterations to converge and (b) error of headloss at pipes versus percentage of pipe failures.



(a) Pipe Damage States (b) Loss of Services Nodes

Fig. 11: WSSC-SUBNET - (a) predicted damaged pipes with $\mathrm{TP} = 0.8, \mathrm{FP} = 0.4$ of distance threshold $200\mathrm{m}$; (b) predicted loss-of-service nodes with $\mathrm{precise} = 1$ and $\mathrm{recall} = 0.99$.

engines, that ultimately will enhance the resilience of future community water distribution systems.

Acknowledgements: This work was funded under NSF Awards No. CNS-1143705, CNS-0958520, and CNS-1450768.

References

[1] S. L. Cutter, J. A. Ahearn, et al. Disaster resilience: A national imperative. *Science & Policy for Sustainable Development*, 2013.

[2] K. A. Klise, R. Murray, and L. T. N. Walker. Systems measures of water distribution system resilience. Technical report, 2015.

[3] J. Eidinger and C. A. Davis. Recent earthquakes: implications for us water utilities. *Water Research Foundation*, 2012.

[4] Nec and imperial college london smart water project: https://wp.doc.ic.ac.uk/aese/project/smart-water/. 2016.

[5] A. Yazdani, R. A. Otoo, and P. Jeffrey. Resilience enhancing expansion strategies for water distribution systems: A network theory approach. *Environmental Modelling & Software*, 2011.

[6] Q. Shuang, M. Zhang, and Y. Yuan. Node vulnerability of water distribution networks under cascading failures. *Reliability Engineering & System Safety*, 2014.

[7] J. M. Torres, L. Duenas-Osorio, et al. Exploring topological effects on WDS performance using graph theory and statistical models. *Water Resources Planning & Management*, 2016.

[8] S. Kartakis, W. Yu, R. Akhavan, and J. A. McCann. Adaptive edge analytics for distributed networked control of water systems. In *IOT Design and Implementation*, 2016.

[9] Q. Han, P. Nguyen, R. T. Eguchi, K.-L. Hsu, and N. Venkatasubramanian. Toward an integrated approach to localizing failures in community water networks. In *Distributed Computing Systems*, 2017.

[10] P. Venkateswaran, Q. Han, R. T. Eguchi, and N. Venkatasubramanian. Impact driven sensor placement for leak detection in community water networks. In *Cyber-physical Systems*, 2018.

[11] A. D. González, L. Dueñas-Osorio, et al. The interdependent network design problem for optimal infrastructure system restoration. *Computer-Aided Civil & Infra Engineering*, 2016.

[12] L. A. Rossman. Epanet 2 users manual. 2000.

[13] K. A. Klise, M. Bynum, et al. A software framework for assessing the resilience of drinking water systems to disasters. *Environmental Modelling & Software*, 2017.

[14] S. Milan, H. Vaclav, and B. Roger. Image processing, analysis, and machine vision. 2002.

[15] A. V. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 2006.

[16] H. El Gamal and A. R. Hammons. Analyzing the turbo decoder using the gaussian approximation. *Information Theory*, 2001.

[17] Y. Weng, R. Negi, and M. D. Ilic. Graphical model for state estimation in electric power systems. In *Smart Grid Communications*, 2013.

[18] S. L. Lauritzen. *Graphical models*, 1996.

[19] M. I. Jordan et al. Graphical models. *Statistical Science*, 2004.

[20] Y. Liu, V. Chandrasekaran, A. Anandkumar, and A. S. Willsky. Feedback message passing for inference in gaussian graphical models. *Signal Processing*, 2012.

[21] M. Cosovic and D. Vukobratovic. Distributed gauss-newton method for ac state estimation: A belief propagation approach. In *Smart Grid Communications*, 2016.

[22] D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in gaussian graphical models. *Machine Learning Research*, 2006.

[23] M. Allen, A. Preis, et al. WDS monitoring and decision support using a wireless sensor network. In *Software Eng., Artificial Intelligence, Networking & Parallel/Distributed Computing*, 2013.

[24] U. Brandes and D. Fleischer. Centrality measures based on current flow. In *Theoretical aspects of computer science*, 2005.

[25] E. Hernadez, S. Hoagland, and L. Ormsbee. Water distribution database for research applications. In *World Environmental and Water Resources Congress*, 2016.