Leak Event Identification in Water Systems Using High Order CRF

Qing Han, Wentao Zhu and Yang Shi

***Abstract*— Today, detection of anomalous events in civil infrastructures (e.g. water pipe breaks and leaks) is time consuming and often takes hours or days. Pipe breakage as one of the most frequent types of failure of water networks often causes community disruptions ranging from temporary interruptions in services to extended loss of business and relocation of residents. In this project, we design and implement a two-phase approach for leak event identification, which leverages dynamic data from multiple information sources including IoT sensing data (pressure values and/or flow rates), geophysical data (water systems), and human inputs (tweets posted on Twitter). In the approach, a high order Conditional Random Field (CRF) is constructed that enforces predictions based on IoT observations consistent with human inputs to improve the performance of event identifications.**

Considering the physical water network as a graph, a CRF model is built and learned by the Structured Support Vector Machine (SSVM) using node features such as water pressure and flow rate. After that, we built the high order CRF system by enforcing twitter leakage detection information. An optimal inference algorithm is proposed for the adapted high order CRF model. Experimental results show the effectiveness of our system.

I. INTRODUCTION

Water is a critical resource and a lifeline service to communities worldwide; they are essential for sustaining the economic and social viability of a community. Often the infrastructures that capture, deliver and store water in cities and communities are many decades old; with the rise in urban populations, these infrastructures have become more increasingly complex and vulnerable to failures due to natural, technological and manmade events. For example, pipe breakage is one of the most frequent types of failure of water networks and often causes community disruptions. Based on a report from Los Angeles Department of Water and Power (LADWP), Los Angeles (LA) has been experiencing an unusual increase in pipe beaks and leaks, mainly in old pipes that are susceptible to corrosion problems and pipe joint displacements caused by surface deformations. Pipe bursts may also cause transportation network collapse and water loss often lead to additional energy expenditures for transporting water from natural resources to the end users. Extreme weather and rainfall events (e.g. Hurricane Sandy, El Niño 2016) stresses already weakened pipes to the point of causing major pipe breaks and significant increases in leak rates and thus major pipe breaks of failures. It represents a very high cost vulnerability and is associated with public health implications and wastage of a limited resources.

Pipe breaks or bursts often reduce pressure heads and increase flow rates at failure point. IoT sensing data from water infrastructures can track the changes of the network in a timely manner, and reflect a certain level of failures in the network. However, these measurements are limited by (a) sensor locations (static sensors), (b) the number of sensors installed (due to high cost), and (c) they are highly correlated with each other. Therefore, it is hard to isolate the damaged pipes by these data itself, but aggregating with external sources will be helpful. Human reports related to leak events may complement the limitations of IoT observations. Because human sensing are more dynamic, reliable, and accessible. The key of combination/fusion is that information from different data sources indicate the presence of a problem at different

Conditional Random Field (CRF) [1] has been successfully used in structured prediction problems in the undirected graphic models. The main advantage of CRF is that it tries to model $p(\mathbf{y}|\mathbf{x})$ instead of $p(\mathbf{x}, \mathbf{y})$ given the observation $\mathbf{x}$. However, the CRF model only allows the adjacent relationship of $\mathbf{y}$ due to the markov property of CRF. The high order CRF model [2] extends it to exploit high-order dependencies, which provides substantial performance improvement. In this paper, we first use the CRF to model the leak event in the water system. Then a high-order CRF model is used by enforcing human inputs.

The rest of this paper is organized as follows. Related work on leak detection, CRF and Structured Support Vector Machine (SSVM) are introduced in Section II. We describe the proposed approach in Section III that is extensively validated using data from the hydraulic simulator in Section IV. Section V concludes the paper.

II. RELATED WORK

There has been substantial work on single leak detection. The primary method to determine a leak event is based on the measurement of leak-related vibro-acoustic phenomena with the help of expensive and sophisticated devices whose efficiency largely depends on the operator skills [3], [4]. Machine learning techniques have been suggested for leak identification problems, such as maximum likelihood methods [5], [6], SCEM-UA algorithm [7], and neural network [8], [9]. The results, however, are not solid because these approaches were evaluated by specific use cases that are not sufficient to address the general performance.

Weng et. al. [10] developed a fast on-line state estimation in electrical system. Belief propagation and variational belief propagation are applied to the proposed graphical model. The gain of using these methods is not only in computation accuracy but also in computation time. Beyond these, the estimation algorithm can scale up well. However, the model is only for electrical pressure estimation from the noisy

observation, and can not be integrated with the outside information directly.

Ahmed et. al. [11] considered a gasoline leakage problem. By defining different stages of the system, i.e. healthy, minor fault and faulty stages. A hidden markov model is used in the system. However, they didn't attempt to build the estimation for the leakage location, and the model cannot be applied to undirected or direction changeable problems, such as water systems.

Recently, high order conditional random field models are emerging [12], [13]. These models enforce label consistency in the CRF model and obtain better performance than previous models. We adapted the high order CRF model to the leakage detection in the water system. Based on the special features we have in the waterpipe network, we proposed an optimal solution for the problem.

III. Two-phase Approach to Leak Event Identification

Pipe leaks or bursts often lead to changes in pressure heads and flow rates, which can be used to obtain critical information on which parts of the system are suffering the effects of water pipe failures. However, the IoT sensing data itself (pressure and flow rate) may not enough to locate all leak events due to (*a*) limited observations (inaccessible locations and high cost) and (*b*) highly correlated features (densely connected network). In real world, the damage to underground infrastructures is often hidden, and most pipe failures are silent until they are noticed by people. Thus, human inputs are integrated in the inference process to complement limitations of the IoT observations.

A. Problem Formulation

A water system is modeled as an undirected graph $G(\mathcal{V}, E)$ (water can flow in both directions) with vertices $\mathcal{V}$ that represent the nodes (joint of pipes), and edges E that represent pipelines. $|\mathcal{V}|$ equals to the number of nodes in the network. A set of pressure and/or flow rate sensors $\mathcal{A}$ are simulated using the hydraulic simulator. $\mathcal{C}$ is a set of subsets which contains relevant vertices from human sensing. We consider X as a set of observations, i.e. the measurements of pressure values and/or flow rates collected from sensors, and Y as a set of event variables, i.e. the leakage states (leaking or not) of each node that we wish to identify. An arbitrary assignment to X is denoted by a vector $\mathbf{x} = \{x_a : a \in \mathcal{A}\}$. Similarly for Y, an assignment $\mathbf{y} = \{y_v : v \in \mathcal{V}\}$ is a vector of labels taking from the label set $\mathcal{L} = \{0, 1\}$ where $y_v = 1$ indicates a leak at location v. Note that the leak event is assumed to occur at the joint of pipes for simplicity.

B. Two-Phase Approach

In the first part, We will discuss how to build and learn the CRF model. After that, we will explore the high order CRF model by adding human report in the built CRF model. At last, a greedy inference method is proposed for the high order CRF model.

*1) **Phase I: Learning IoT Observations**:* Conditional Random Field (CRF) is an undirected graphical model that has been widely used for structured prediction [1]. Based on the formulation, (X, Y) is a conditional random field because the random variable Y is indexed by the vertices of G and y_v conditioned on $\mathbf{x}$ obey markov property with respect to the graph G: $p(y_v|\mathbf{x}, y_{v'}, v' \neq v) = p(y_v|\mathbf{x}, y_{v'}, (v, v') \in E)$ [14]. The conditional distribution $p(\mathbf{y}|\mathbf{x})$ can then be modeled and trained by using machine learning based techniques.

Structured Support Vector Machine (SSVM) can be used for the learning and inference of CRFs by generalizing the Support Vector Machine (SVM) classifier to do structured learning.

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \max_{\mathbf{y}_n \in Y} (\Delta(\hat{\mathbf{y}}_n, \mathbf{y}_n) + \mathbf{w}^T \theta(\hat{\mathbf{y}}_n, \mathbf{x}_n) - \mathbf{w}^T \theta(\mathbf{y}_n, \mathbf{x}_n)) + \frac{C}{2} \|\mathbf{w}\|^2 \tag{1}$$

where $\Delta(\hat{\mathbf{y}}_v, \mathbf{y}_v) = \sum_{v \in \mathcal{V}} \mathbb{1}[\hat{\mathbf{y}}_v \neq \mathbf{y}_v]$ is the hamming loss between the prediction $\mathbf{y}_v$ and leakage observation $\hat{\mathbf{y}}_v$, C is the cost factor related to the trade off between the empirical loss and regularization, N is the number of training samples.

In the inference period, we use a structured linear predictor based on the learned SSVM model:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in Y} \mathbf{w}^T \theta(\mathbf{x}, \mathbf{y}) \tag{2}$$

where $\hat{\mathbf{y}}$ is a vector of predicted structured labels, $\mathbf{w}$ are parameters that are learned from data, and θ is defined by the user-specified structure of the model [15]. To compute the $argmax$, several inference solvers, e.g. Quadratic Pseudo-Boolean Optimization (QPBO) and Alternating Directions Dual Decomposition (AD3), can be applied.

The inference outcome of the model learned on IoT observations is $\mathcal{S} = \{v : \hat{y}_v = 1 \wedge v \in \mathcal{V}\}$, representing a subset of $\mathcal{V}$ which are predicted as leaking positions. This set will be updated after Phase II. Notice that, we can also obtain the label assignment probability in this model [16].

*2) **Phase II: High order CRF with Human Inputs**:* It is natural to think that there are higher possibilities for one subarea to have pipeline break if some human living around reported in social networks. To leverage human inputs, we bring in social media, Online Social Network (OSN), to incorporate human sensing. OSN has become a major platform for information sharing in which we can mine interested patterns [17]. We apply the Tweet Acquisition System (TAS) developed at UCI to selectively collect tweets relevant to leak events from Twitter, and use the associated geographic information to track and locate the risky area. The human input, however, is unable to specify the exact position of the damage due to various social behaviors. Therefore, it is considered as high order potentials [12] in the inference process to enforce event consistency. We assume that twitter information can reflect true events with high confidence.

That is, based on the content, location, and the number of tweets we can locate the faulty region at different levels of granularity.

Let $\mathcal{C} = \{c : c = \{v : |l_c - l_v| < \gamma \wedge v \in \mathcal{V}\}\}$ represent a set of subsets of $\mathcal{V}$ (i.e. a set of cliques) infered from human inputs. Here, $|l_c - l_v| < \gamma$ indicates that nodes v whose distance to the location of clique c (l_c) identified by the GeoTag of tweets is less than the threshold γ. That is, nodes within a certain distance from the location of the tweet are likely to leak. The high order potential $\Phi_c : \mathcal{L}^{|c|} \rightarrow \mathbb{R}$ is defined over this clique assigns a cost to each possible configurations (or labelings) of $\mathbf{y}$. Because we assume that the effects of human inputs on leak event identification is non-negative, we have

$$\Phi_c = \begin{cases} 0 & if \ \ \exists v \in \mathcal{S} \ \ for \ \ v \in c \\ Inf & o.w. \end{cases} \tag{3}$$

where S is the leakage set obtained by the above CRF model.

The high-order CRF model enforcing the human input can be obtained as

$$\min \sum_{v \in \mathcal{V}} \Phi_v(y_v) + \sum_{(v,v') \in E} \Phi_{v,v'}(y_v, y_{v'}) + \Phi_c \tag{4}$$

where the unary potential $\Phi_v(y_v)$ is defined as the minus entropy (corresponding to maximum entropy) of y_v if the vertex v is in the human reported leaking cliques and all the vertices in that clique are not in the set S. If the vertex v is not in the human reported leaking cliques or a vertex in the clique is in the S, $\Phi_v(y_v)$ can be defined as $-\mathbf{w}^T\theta(\mathbf{x}, \mathbf{y})$, which means the prediction for the vertex v is the CRF inference result as (2) since there is no extra information adding into the system. Because the leakage events among the vertices of G are conditionally independent given the observations X, the pairwise terms $\Phi_{v,v'}(y_v, y_{v'})$, in our case, are constant.

*3) **Inference**:* According to (3), an event inconsistency can push the energy to the infinity. Therefore, Algorithm 1, a greedy inference algorithm, is to update $\mathcal{S}$ by adding a node v^* from the clique with $\Phi_c = Inf$ into S if $v^* = \arg\max_{v \in c} H(y_v)$ (5) and $H(y_{v^*}) > \Gamma$.

$$H(y_v) = -\sum_{i=0}^{1} p_i(\hat{y}_v) \log p_i(\hat{y}_v) \tag{5}$$

Note that $p_i(\cdot)$ can be obatined by machine learning based techniques applied in Section III-B.1. In this manner, Algorithm 1 can remove the inconsistency between IoT observations and human inputs to minimize the energy function given in (4).

IV. Experiment and Results

In this section, we begin by presenting the datasets and parameters under which the simulations are conducted, and describe the implementation of the high order CRF model.

Algorithm 1 Greedy algorithm for the integration of human inputs (the high order potential)

Input $\mathcal{S}$, $\mathcal{C}$
Output *updated*($\mathcal{S}$)
Objective $\min_{\mathcal{C}} E[\mathbf{y}]$

$\mathcal{S} = \{v : \hat{y}_v = 1 \wedge v \in \mathcal{V}\}$
$\mathcal{C} = \{c : c = \{v : |l_c - l_v| < \gamma \wedge v \in \mathcal{V}\}\}$
for c in $\mathcal{C}$ **do**
  if $\Phi_c = 0$ **then**
    continue
  else
    $v^* = \arg\max_{v \in c} \mathrm{H}(y_v)$
    if $H(y_{v^*}) > \Gamma$ **then**
      $\mathcal{S} = \mathcal{S} \cup \{v^*\}$
    end if
  end if
end for

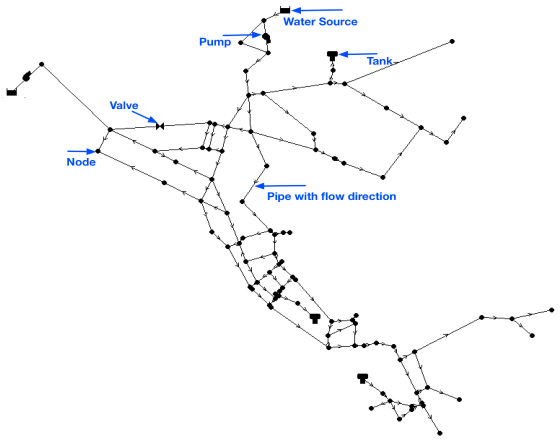

Fig. 1. A water network provided by EPANET with 118 pipes, 96 nodes, 2 pumps, a valve, 3 tanks, and 2 water sources.

A. Datasets and Experiment Setup

The IoT sensing data is generated using a commercial grade hydraulic simulator EPANET [18]. Figure 1 shows a real-world based water network provided by EPANET. The elevations of pipes varies with the topography, and each pipe has four attributes - length, diameter, roughness coefficient, and status (open or close controlled by valve). Each node has a pattern of time variation of the demand (i.e. consumption). A leak event is assumed to occur at nodes, since the interconnect points are more risky than others. We assume a fully IoT observations composed of pressure values and flow rates with the number of sensors $|\mathcal{A}| = 218$. That is, each data sample has 218 features.

Extensive simulations are run on EPANET to generate sufficient observations for training. The number of training sets and testing sets are $10,000$ and $2,000$ separately. For each simulation run, different numbers of leak event(s) are generated randomly with different locations, sizes, and starting time. The sampling frequency of the sensors is 15

TABLE I

HAMMING SCORE OF LEAK EVENT IDENTIFICATION WITH LEARNING USING SSVM AND INFERENCE WITH HIGH ORDER POTENTIAL.

p	γ **(unit: meter)**	Γ	**Hamming Score**
0.3	2	0	**0.9654**
0.3	3	0.04	**0.9622**
0.7	2	0	**0.9789**
0.7	3	0.04	**0.9701**

minutes.

B. Implementation

In Phase **I**, the IoT observations are trained using SSVM with penalty parameter $C = 0.25$. We use MATLAB with libsvm packet [19] to first predict possible leak event(s) with the leak probability that will be used in the inference process in Phase **II**.

As mentioned in Section III-B.2, tweets posted on Twitter are used for social media dataset that can complement the limitations of IoT observations. We simulate the human inputs by assigning a probability p that people will report a leak around the true event position. The confidence of the human reports can be simulated by changing the threshold γ. To avoid the inconsistency between the IoT observations and human inputs, the predicted label of the node with the highest entropy that is larger than Γ will be set to 1 as running Algorithm 1.

C. Result

We use Hamming score as the evaluation for the results. Hamming score is defined as $\frac{P \cap T}{P \cup T}$ in our case, where P is a set of locations predicted as the leak spots, and T is a set of true event locations. The score is bounded by 1, and the higher the score the better the performance.

In previous work, we applied random forest on this dataset. The hamming score without human inference is **0.65** and the one with high order potential is **0.74**. In this paper, the best hamming score we have by learning the model and inference using SSVM without the integration of high order potential is **0.9566**. Table I shows that the hamming score varies with the parameters after we consider human sensing. Clearly, the integration of high order potential improves the performance. With more accurate human reports, the performance is even better.

V. CONCLUSIONS

In the project, we constructed a CRF model to model the structure leak detection in the water system. The structured SVM is used for the built CRF model. To fully take advantage of human inputs, a high order CRF model is adapted to our problem. An optimal greedy inference algorithm is proposed for the high order CRF model in our case. Experimental results show the effectiveness of the proposed system and the desirable detection performance.

REFERENCES

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[2] N. Ye, W. S. Lee, H. L. Chieu, and D. Wu, "Conditional random fields with high-order features for sequence labeling," in *Advances in Neural Information Processing Systems*, 2009, pp. 2196–2204.

[3] G. Hessel, W. Schmitt, K. Van der Vorst, and F.-P. Weiss, "A neutral network approach for acoustic leak monitoring in the vver-440 pressure vessel head," *Progress in Nuclear Energy*, vol. 34, no. 3, pp. 173–183, 1999.

[4] J. Rajtar and R. Muthiah, "Pipeline leak detection system for oil and gas flowlines," *Journal of manufacturing science and engineering*, vol. 119, no. 1, pp. 105–109, 1997.

[5] Z. Poulakis, D. Valougeorgis, and C. Papadimitriou, "Leakage detection in water pipe networks using a bayesian probabilistic framework," *Probabilistic Engineering Mechanics*, vol. 18, no. 4, pp. 315–327, 2003.

[6] J. Rougier, "Probabilistic leak detection in pipelines using the mass imbalance approach," *Journal of Hydraulic Research*, vol. 43, no. 5, pp. 556–566, 2005.

[7] R. Puust, Z. Kapelan, D. Savic, and T. Koppel, "Probabilistic leak detection in pipe networks using the scem-ua algorithm," in *8th Annual Water Distribution Systems Analysis Symposium*, 2006, pp. 27–30.

[8] C. Ai, H. Zhao, R. Ma, and X. Dong, "Pipeline damage and leak detection based on sound spectrum lpcc and hmm," in *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*, vol. 1. IEEE, 2006, pp. 829–833.

[9] J. Mashford, D. De Silva, D. Marney, and S. Burn, "An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine," in *Network and System Security, 2009. NSS'09. Third International Conference on*. IEEE, 2009, pp. 534–539.

[10] Y. Weng, R. Negi, and M. D. Ilic, "Graphical model for state estimation in electric power systems," in *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*. IEEE, 2013, pp. 103–108.

[11] Q. Ahmed, A. Iqbal, I. Taj, and K. Ahmed, "Gasoline engine intake manifold leakage. diagonosis/prognosis using hidden markov model," in *International Journal of Innovative Computing, Information and Control*, 2012, pp. 4661–4674.

[12] P. Kohli, P. H. Torr *et al.*, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[13] J. Wegner, J. Montoya-Zegarra, and K. Schindler, "A higher-order crf model for road network extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1698–1705.

[14] Wikipedia, "Conditional random field - wikipedia, the free encyclopedia," 2016, [Online; accessed 8-June-2016]. [Online]. Available: \url{https://en.wikipedia.org/w/index.php?title=Conditional_random_field&oldid=722584204}

[15] A. C. Müller and S. Behnke, "Pystruct-learning structured prediction in python," *Journal of Machine Learning Research*, vol. 1, pp. 1–1, 2013.

[16] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[17] S. Kumar, F. Morstatter, and H. Liu, *Twitter data analytics*. Springer, 2014.

[18] L. A. Rossman *et al.*, "Epanet 2: users manual," 2000.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.