On Graph-based Feature Selection for Multi-hop Performance Characterization in Industrial Smart Water Networks

Athanasia Panousopoulou* and Panagiotis Tsakailides*†
Email:{apanouso, tsakalid}@ics.forth.gr
†Department of Computer Science, University of Crete, Heraklion GR-70013, Greece
*Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, GR-70013, Greece

***Abstract*—Recent deployments of Smart Water Networks in urban environments are causing a paradigm shift towards sustainable water resources management. Nevertheless, there exists a substantial gap on respective solutions for industrial water treatment. In such deployments the wireless network backbone would have to overcome limiting factors that span across different layers of a protocol stack. Incorporating data analytics for capturing multi-dimensional correlations could be extremely beneficial to the design of reconfigurable network protocols for industrial Smart Water Networks. In this work, we exploit recent findings in the arena of network measurements and we propose a graph-based unsupervised feature selection approach for extracting the dominant network conditions that affect the performance of user-defined links. We employ a real-life industrial Smart Water Network deployed in a desalination plant to evaluate the efficacy of our approach. Finally, we provide useful insights on how different locations in a desalination plant affect the performance of the network backbone.**

Keywords*—*Graph-based feature selection; Network Measurement and Analysis; Industrial smart water networks

I. Introduction

Over the past years Smart Water Networks (SWN) [1] have emerged as a multi-disciplinary research field, which blends, among others, Cyber-Physical Systems and Wireless Sensor Networks (WSN) with water infrastructures for optimizing the utilization of water resources. Successful SWN deployments address both the response in water-extreme situations [2], as well as the engagement of citizens in water consumption and sustainability. As such, in their majority existing SWN platforms are deployed in urban areas, yielding feedback on the condition of the water distribution network infrastructure and the residential monitoring, both in terms of consumption, as well as in terms of quality [3], [4]. Despite their efficacy, such solutions cannot address the engineering challenges that arise from the design and development of *industrial* SWN, which are deployed in water treatment plants, responsible for water purification.

The success of an industrial SWN is strictly aligned to the efficacy of the underlying network backbone to provide a reliable, continuous, and secure data flow between front-end sensing/actuating and decision making systems. As a consequence, typical WSN network imperfections met in realistic deployments [5] can significantly affect the performance of an industrial SWN. Except for the RF-harshness of the industrial environment, characterizing the network performance can be affected by numerous multi-dimensional factors, such as: application-driven positioning, the impact of the power constraints on the dynamic nature of the wireless connectivity, the ambient conditions, and the hardware characteristics. These challenges can be addressed by evaluating the performance of user-designated, end-to-end links, which are built upon multi-hop topologies. Opposed to point-to-point link layer characterization [6]–[8], one can exploit the inherited redundancy of network information available on application-driven, multi-hop higher layer links, and holistically map the dominant conditions that affect the network performance. This knowledge can further on be employed for designing truly reconfigurable network solutions that can compensate any degradation of network performance in highly dynamic and data-critical environments, like those met in industrial water treatment plants.

In this work, we address the problem of multi-hop network performance characterization by the means of feature selection, which allows the automatic calculation of the most prominent network conditions. We employ heuristic network metrics that are readily available in typical WSN deployments and are independent from customized protocol stacks. We exploit the framework presented in [9] for the collection of the essential network measurements, and the extraction of a set of statistical network features. We propose a new graph-based unsupervised feature selection algorithm that goes beyond pairwise comparisons for calculation of the dominant features. Our approach is evaluated against the current state-of-art in graph-based feature selection, using a dense data set (appr. 300,000 instances of end-to-end network traffic) generated during the deployment of an industrial SWN to a desalination plant. The results highlight both the efficacy of the proposed scheme, as well as how different parts of the desalination procedure affect the performance of the multi-hop network backbone.

II. Multi-hop WSN performance characterization

We consider a multi-hop WSN comprised of N nodes, which are deployed in an application-driven manner and operate in an unattended fashion over long periods of time.

At the Application Layer, pairs of sensor nodes i and j establish end-to-end links $i \rightarrow j$ over a network path $P_{ij} = \{i, \ldots, k, \ldots, j\}$, where i is the transmitter, j is the receiver, and each k-th node is a relay node between i and j. During the normal operation of the network, each node $k \in P_{ij}$ can monitor simple network metrics, which are related both to its own functionality, as well as the quality of the link $i \rightarrow j$. In addition, the Packet Reception Ratio $PRR_{ij} \in [0,1]$ is calculated on the side of the j-th node, and defined as the ratio between the packets received by the j-th node and the packets transmitted by the i-th node over fixed windows of observation of length equal to w. The value of PRR_{ij} can be classified into discrete qualitative labels l_{ij} (e.g., $\{Excellent, Good, Problematic, Poor\}$) based on empirically derived operational thresholds α_E, α_G, α_P, $\in [0,1]$, where $\alpha_E \geq \alpha_G \geq \alpha_P$ [6], [10].

The problem at hand is the automated calculation of the network factors that are responsible for classifying the performance of the link $i \rightarrow j$ to different values of l_{ij}. This can become a challenging task in realistic deployments, due to the combination of different factors, e.g., the nexus between energy and communications, irregular transmission patterns [11], the on-board temperature [12], interference and collisions at the MAC layer [13].

Driven by these challenges, we recently [9] formulated the problem of characterizing the performance of each $i \rightarrow j$ link as a feature selection problem. Our formulation considers as input an initial feature vector $\mathbf{f}_{ij}$ comprised of M features, and yields in return a subset of R dominant features ($R \leq M$) $\mathbf{f}^*_{ij} \subseteq \mathbf{f}_{ij}$, that are most relevant to inferring the network label l_{ij}, $\forall$ $i \rightarrow j$. Specifically, by employing a passive network monitoring procedure, that piggybacks heuristic network metrics as each data packet travels over P_{ij}, a batch of network measurements are collected at the side of the j-th node. This batch of measurements is split into smaller windows of observations with fixed length w, and the feature vector $\mathbf{f}_{ij}$ is derived from the first-order statistics (i.e., mean value μ and standard deviation σ) of the metrics presented in Table I.

TABLE I. THE NETWORK METRICS AND THE ASSOCIATED ATTRIBUTES EMPLOYED FOR FORMING THE FEATURE VECTOR $\mathbf{f}_{ij}$ [9].

Network Metric	Description	Attributes
PRX^*_{ij}	Receiver Power over path P_{ij} (dBm)	Mean value ($\mu(PRX^*_{ij})$) and standard deviation ($\sigma(PRX^*_{ij})$)
LQI^*_{ij}	Link Quality Indicator over path P_{ij}	Mean value ($\mu(LQI^*_{ij})$) and standard deviation ($\sigma(LQI^*_{ij})$)
NF^*_{ij}	Noise Floor over path P_{ij} (dBm)	Mean value ($\mu(NF^*_{ij})$) and standard deviation ($sigma(NF^*_{ij})$)
f^{tx}_i	Transmission rate at the MAC layer (bpm)	The value of f^{tx}_i over fixed windows of observation w
f^{rx}_i	Reception rate at the MAC layer (bpm)	The value of f^{rx}_i over fixed windows of observation w
$\lvert P_{ij} \rvert$	Length of path P_{ij}	Mean value ($\mu(\lvert P_{ij} \rvert)$) and standard deviation ($\sigma(\lvert P_{ij} \rvert)$
$\widehat{PRR}_{ij}$	The moving average (WMEWMA) of PRR_{ij} [14]	The value of $\widehat{PRR}_{ij}$ over fixed windows of observation w
T_i	On-board temperature of the i-th node (^oC)	Mean value ($\mu(T_i)$) and standard deviation ($\sigma(T_i)$)
H_i	Percentage of on-board humidity for the i-th node	Mean value ($\mu(H_i)$) and standard deviation ($\sigma(H_i)$)
V_i	Input power level for the i-th node (Volt)	Mean value ($\mu(V_i)$) and standard deviation ($\sigma(V_i)$)

Applying the same procedure on all observation windows yields D subsequent instances of the feature vector $\mathbf{f}^d_{ij}$ ($1 \times M$), where d =1,2,...M. These vectors formulate the features data set $\mathbf{A} \triangleq \left[\mathbf{f}^1_{ij}; \mathbf{f}^2_{ij}; \ldots; \mathbf{f}^D_{ij}\right]$ ($D \times M$), which is employed for calculating the vector of the most relevant features $\mathbf{f}^*_{ij}$ for each $i \rightarrow j$ link. While these dominant variables can be later on used for constructing a predictive model for calculating the value of l_{ij}, the remaining attributes can be eliminated since they contribute limited or redundant information.

III. Graph-based Feature Selection

Feature selection is in principle a search problem with the objective of reducing the dimensionality of a search space in typical data mining applications. Shifting to the WSN paradigm, it is considered essential to adopt on-line characteristics. As a consequence, numerous technical challenges raise. First of all, a-priori knowledge of the labels l_{ij} is not available and the existence of training periods is considered an unrealistic assumption. In addition, in order to encompass the multi-dimensional correlations between the factors that can affect the performance of end-to-end links, novel solutions that go well beyond the 2D (pairwise) characterization of redundancy and compression [9], [15] should be adopted. Finally, while accurate classification is the primary goal, the quality of the reduced feature vector in terms of compression is another important aspect for fertilizing the ground for in-network and distributed implementations.

The extraction of multi-dimensional correlations within the feature set, without any prior knowledge on the class labels, can be efficiently addressed by unsupervised graph-based feature selection techniques. The potential of graph-based learning feature level fusion has recently been unraveled in a plethora of classification problems [16], [17]. In a nutshell, in graph-based feature selection, the feature set is modeled as a weighted, fully connected graph; the vertices of the graph represent the features and the weighted edges denote their inter-similarity. The objective of graph-based feature selection is to exploit clustering techniques for grouping similar features together, and extracting a set of few, representative features from each cluster.

A. *Graph-based Feature Selection using Node Centrality and Representation Entropy*

Recent state of the art in graph-based feature selection incorporates the concept on node centrality, as the means of characterizing the potential of a feature to be included in the dominant features vector. As a representative example, the Graph Clustering with Node Centrality (NC) feature selection method [17] combines the following characteristics: (a) it employs the Pearson product-moment correlation coefficient as the similarity measure between features; (b) it uses a community detection algorithm for grouping the features into clusters; (c) it exploits the Laplacian centrality (LC) [18] for characterizing the connectivity and density around a feature, while taking into account both local as well as global characterization of its importance (centrality) within the cluster; (d) it proposes an iterative search strategy that uses NC for eliminating redundant features from each cluster.

The limitation of the graph with NC feature selection algorithm is that it does consider the centrality, yet, not the contribution of each feature to the overall entropy of the feature data set. As a consequence, the quality of compression, which is considered of primary importance for the WSN application domain, might be poor when the graph with NC method is adopted. Driven by recent findings in the Pattern Recognition arena [15], [19], [20], we herein propose a variation of graph with NC feature selection, that considers several improvements. *First*, we replace the Pearson correlation coefficient with the maximal information compression index λ_2 between two features [15] for calculating the edge-weighted graph of the feature data set $\mathbf{A}$. The value of λ is defined as the smallest eigenvalue of the covariance matrix between two variables and reflects the information loss in pattern reconstruction, when the pairwise data are projected to their principal direction. As such, highly relevant attributes would have a small value of λ_2, and as the redundancy between two features decreases, the value of λ_2 increases. *Second*, we employ the powerful, yet lightweight, concept of dominant sets [19] for grouping similar features into K clusters $\mathbf{c}_\kappa$, $\kappa = 1,2,\ldots,K$. In principle, dominant sets generalize the notion of maximal complete subgraph to weighted graphs, and simultaneously provide the property of intra-cluster homogeneity and inter-cluster heterogeneity; the overall similarity inside the cluster is higher than the one between internal and external nodes. *Third*, during the iterative search within the clusters of features, we additionally consider representation entropy (RE) as a criterion of characterizing features as dominant o redundant.

The representation entropy $H_{\mathbf{A}}$ quantifies the redundancy within the feature data set $\mathbf{A}$. It is defined as $H_{\mathbf{A}} = -\sum_{m=1}^{M} \tilde{\lambda}_m \log \tilde{\lambda}_m$, where $\tilde{\lambda}_m$ is the normalized value of the m-th eigenvalue λ_m of the $M \times M$ covariance matrix of the feature data set $\mathbf{A}$, with respect to the sum of all eigenvalues of the covariance matrix. It expresses the amount of information compression available in $\mathbf{A}$: when $H_{\mathbf{A}} \to 0$, the information of the data set $\mathbf{A}$ can be compressed into the information that corresponds to the dominating feature only. By contrast, when $H_{\mathbf{A}} \to \log M$ (i.e., its maximum value) all features are equally important and, thus, the redundancy of $\mathbf{A}$ is low.

Recent approaches in unsupervised feature ranking and pairwise clustering [9], [20] have employed the concept of RE to quantify the amount of redundancy that each feature contributes to the data set by defining the difference dH_m in the data uncertainty when the m-th feature is excluded from the feature vector. If the m-th feature corresponds to a principal component, then the value of dH_m will become high. On the other hand, if the m-th feature describes a pattern with predictable behavior, then $dH_m \to 0$.

The herein proposed approach combines the calculation of dH_m with the value of the laplacian centrality LC_m of the m-th feature in the form of their product $LC_m \times dH_m$. The outcome is employed to evaluate whether the respective data pattern both conveys the sufficient amount of redundancy, as well as exhibits the necessary cluster-wise characteristics, in terms of connectivity and density, i.e., whether $LC_m \times dH_m$ is greater/less than a threshold δ. Specifically, features with $LC_m \times dH_m < \delta$ are considered redundant and are subsequently eliminated from the cluster and the original feature vector $\mathbf{f}_{ij}$.

The feature selection method that combines the above characteristics is summarized in Algorithm 1. In a manner similar to [17], the termination of the iterative search for representative features depends on the threshold δ that dictates the degree of importance of each feature. Nonetheless, instead of employing a predetermined, user-defined value for δ, we consider that its value varies with respect to median value of M_κ of $LC_m \times dH_m$ across the entire cluster $\mathbf{c}_\kappa$.

Algorithm 1 Graph with NC-RE feature selection algorithm

```
Require: (a) The initial feature vector $\mathbf{f}_{ij}$ containing $M$ features, (b)
    the corresponding data set $\mathbf{A}^{D \times M}$, (c) a user-defined parameter
    $\alpha \in [0,1]$ .
Ensure: The reduced feature vector $\mathbf{f}^*_{ij} \subseteq \mathbf{f}_{ij}$, containing $R$ features
    ($R \leq M$).
 1: Initialize the reduced feature vector $\mathbf{f}^*_{ij}$ to $\emptyset$.
 2: Model $\mathbf{f}_{ij}$ as a fully connected undirected weighted graph, using
    the maximal information compression index $\lambda_2$ [15] for calculating the inter-similarity.
 3: Employ the recursive replicator dynamics for calculating the
    dominant sets of the graph [19], and group all features into $K$
    feature clusters $\mathbf{c}_\kappa$, $\kappa = 1,2,\ldots,K$.
 4: for $\kappa = 1$ to $\kappa = K$ do
 5:     For each feature $m \in \mathbf{c}_\kappa$ calculate $LC_m \times dH_m$, using [18] and
        [9], [20] respectively.
 6:     Calculate the median value $M_\kappa$ of $LC_m \times dH_m$ over $\mathbf{c}_\kappa$ and set
        $\delta \leftarrow \alpha \times M_\kappa$.
 7:     Find the subset $\rho_\kappa$ of the cluster $\mathbf{c}_\kappa$ s.t. $\rho_\kappa = \{m \in
        \mathbf{c}_\kappa | LC_m \times dH_m < \delta\}$.
 8:     if $\rho_\kappa = \emptyset$ or $|\mathbf{c}_\kappa| - |\rho_\kappa| \leq 1$ then
 9:         Go to Step 13.
10:     else
11:         Update $\mathbf{c}_\kappa \leftarrow \mathbf{c}_\kappa \setminus \rho_\kappa$ and return to Step 5.
12:     end if
13:     Update $\mathbf{f}^*_{ij} \leftarrow \mathbf{f}^*_{ij} \cup \mathbf{c}_\kappa$ and $\kappa \leftarrow \kappa + 1$.
14: end for
15: Return the reduced feature vector $\mathbf{f}^*_{ij}$ and stop.
```

IV. Evaluation Studies on Industrial Smart Water Networks

Algorithm 1, henceforth referred to as Graph-NC-RE, has been applied on multi-hop network traffic collected during the deployment of an industrial smart water network in a desalination plant to monitor and control the phenomenon of biofouling process, which is related to the accumulation of unwanted bacterial matter on the surface of the reverse osmosis membranes [21]. The respective multi-hop WSN topology was comprised of 10 nodes, with the objective to collect data from different parts of the plant to a sink node. Thus, all $i \to j$ links considered the same destination, i.e. $i = 1,2...10$ and $j =$ct. The smart water sensor nodes were deployed at user-designated locations, namely, the sea water intake, the pre-treatment, the security filters, and the reverse osmosis. As expected, due to the nature of the application, except for the metallic environment, additional RF challenges related to bulky water tanks and heavy machinery in operation, were introduced at the multi-hop network performance. In order to compensate the impact of the industrial environment on the network performance, a customized protocol stack was employed, featuring: (a) at the Physical Layer, the IEEE 802.15.4 standard [22]; (b) at the MAC layer a customized CSMA-based protocol, coping with interference [23]; and (c) at the Network Layer, a customized version of the Routing over Low Power and Lossy Networks IEFT standard (RPL) that balances the load traffic in popular relay nodes [24]. The data set considered for our evaluation studies represents 48 hours of non-stop operation of the industrial smart water network, which generated a data packet every 0.1 minutes, producing approximately 300,000 instances of end-to-end link traffic.

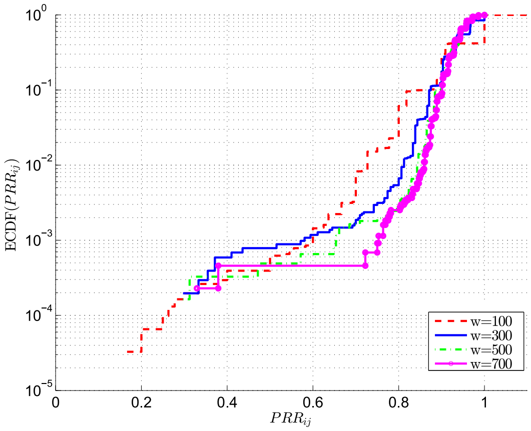

Fig. 1. The ECDF of PRR_{ij} for $w = \{100, 300, 500, 700\}$.

Figure 1 presents the estimated Cumulative Density Function (ECDF) of PRR_{ij} considering different cases for the length w of the window of observation, namely $w = \{100, 300, 500, 700\}$ corresponding to monitoring the network status every 10min, 30min, 50min, and 70 min respectively. We observe that the protocol stack employed yields in overall a satisfactory performance, since the probability of having $PRR_{ij} \leq 0.8$ equals to 0.07 for $w = 100$, and decreases as the value of w increases. This highlights the fact that as the real-time requirements of the network performance monitoring become more strict ($w \downarrow$) the sensitivity of PRR_{ij} to sporadic packet losses increases ($PRR_{ij} \downarrow$). As explained below, this aspect affects the performance of the feature selection process, responsible for calculating $\mathbf{f}^*_{ij}$.

Experimental Results

Based on the performance of the protocol stack, we set $\alpha_E = 0.98$, $\alpha_G = 0.8$, and $\alpha_P = 0$, and thereby consider 3 labels for the network performance, namely $l_{ij} = \{$Excellent, Good, Problematic$\}$. The Graph-NC-RE feature selection algorithm was implemented on Matlab, using the libraries provided at [25] for the recursive clustering of $\mathbf{f}_{ij}$ into dominant sets. Its efficacy is evaluated against the original algorithm that considers node centrality (Graph-NC) [17] and a variation of Graph-NC-RE that considers a fixed, predefined value for δ ($\delta = 0.01$). The metrics used for the evaluation process are: (a) the k-NN cross validation accuracy (CV, defined in [0,1]), where $k = \lceil \sqrt{D} \rceil$ [15], (b) the normalized value of the representation entropy $\overline{H}$ for the data patterns that correspond to $\mathbf{f}^*_{ij}$, with respect to its maximum value ($\log R$), and (c) the compression rate, defined as $\frac{M-R}{M}\%$. The CV metric quantifies the ability of $\mathbf{f}^*_{ij}$ for classification, while $\overline{H}$ describes redundancy within $\mathbf{f}^*_{ij}$, or equivalently the quality of the compression. Finally, CR expresses the percentage of compression achieved with respect to the initial feature vector.

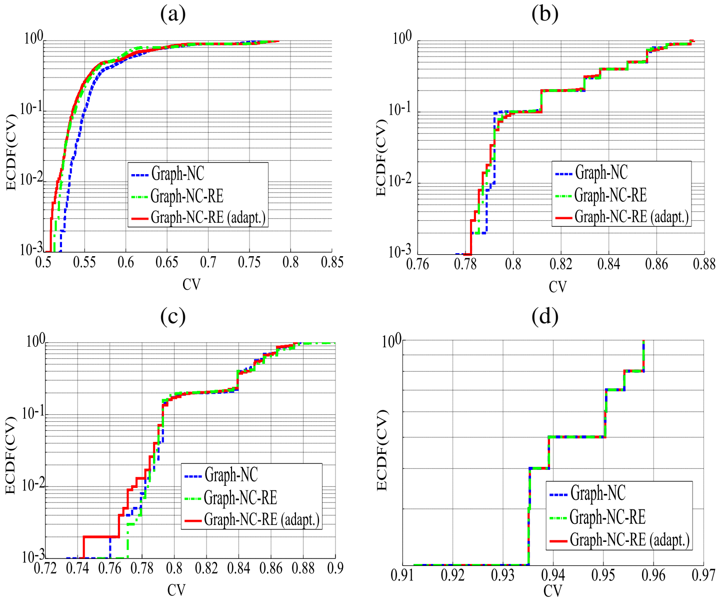

Fig. 2. The ECDF of CV for: (a) $w = 100$, (b) $w = 300$, (c) $w = 500$, (d) w= 700.

Figure 2 presents the ECDF of the CV metric for the graph feature selection algorithms and $w = \{100, 300, 500, 700\}$. We observe that when $w = 100$ the performance of all graph feature selection algorithms is poor and improves as the window size increases; the value of CV remains lower than 0.8 for $w = 100$, and goes beyond 0.9 for w=700. This is due to the fact that when w is small, the volume of raw input network streams is not sufficient for extracting a representative feature vector $\mathbf{f}_{ij}$. Nevertheless, the herein proposed Graph-NC-RE (with either fixed or adaptive value of δ) yields better results for the smaller values of w considered (100, 300, 500), since the respective ECDF curves reach faster 1.

The value of $\overline{H}$ for $\mathbf{f}^*_{ij}$ for all smart water sensor nodes is presented in Fig. 3 for the three different graph-based feature selection algorithms and the four different cases of w. The impact of including dH_m as a criterion for calculating the representative features within each dominant set is highlighted in all different cases of w considered; the value of $\overline{H}$ when Graph-NC-RE (with either fixed or adaptive value of δ) is applied is higher than the one provided when Graph-NC is utilized for the calculation of $\mathbf{f}^*_{ij}$. In addition, when the value of δ is adaptive, the quality of compression improves, thereby highlighting the benefit of dynamically adapting the criterion of removing redundant features from each dominant set. With regard to the impact of the value of w on the quality of compression, we observe that higher values of w yield better results for $\overline{H}$ when the Graph-NC-RE family is applied, in compliance to our observations above on the feature selection performance in terms of CV.

Similar observations can be derived when the compression rate CR is considered for all smart water sensor nodes, presented in Fig. 4 for the three different graph-based feature selection algorithms. The median value of CR when Graph-NC is considered does not exceed the value of 55%, and $w = \{100, 300, 500, 700\}$. By contrast, the Graph-NC-RE family yields considerably higher compression rate, which remains higher

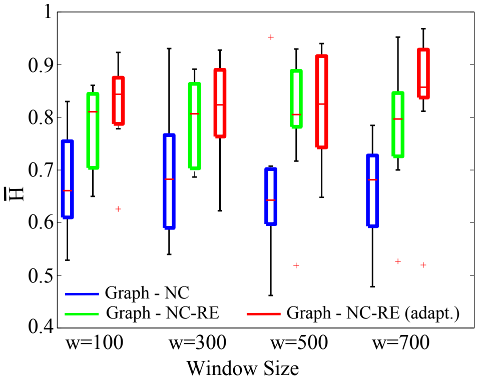

Fig. 3. The $\overline{H}$ of $\mathbf{f}^*_{ij}$ per sensor node and $w = \{100,300,500,700\}$.

than 65%, while slight improvements are additionally observed when the value of δ is adapted within the recursive search of dominant features.

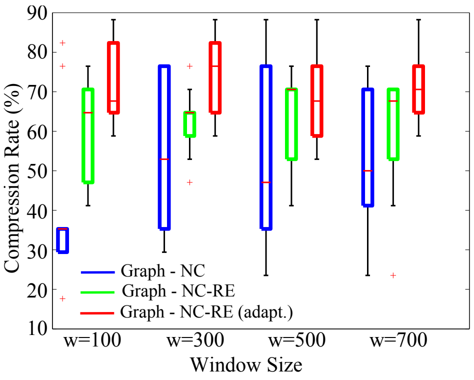

Fig. 4. The CR achieved per sensor node and $w = \{100,300,500,700\}$.

Dominant Network Features for Multi-hop Industrial Smart Water Networks: Table II summarizes the percentage of occurrence of the dominant features extracted when Graph-NC-RE with adaptive δ is applied, with respect to: (a) the value of w (i.e., $w = \{300,700\}$), and (b) the location of deployment at the desalination plant (i.e., sea water inlet, pre-treatment, security filters, reverse osmosis). With regard to the length of the window size considered for extracting the feature vector, we observe that at each sub-network deployed at the four different areas of the desalination plant the dominant features become more diverse as the window size increases from $w = 300$ to $w = 700$. This is accompanied by an increase on the number of dominant features; for instance at the sea water intake the reduced feature vector is comprised by 6 features when w =300 (i.e., $\sigma(NF^*_{ij})$: 12.5%, f^{rx}_i: 25%, $\mu(T_i)$: 25%, $\sigma(T_i)$:12.5%, $\mu(V_i)$:12.5%,$\sigma(V_i)$:12.5%), while its length increases to 8 (i.e., $\mu(PRX^*_{ij})$: 11.11%, $\sigma(NF^*_{ij})$: 11.11%, f^{rx}_i: 22.22%, $\mu(T_i)$: 11.11%, $\sigma(T_i)$:11.11%, $\mu(H_i)$: 11.11%, $\mu(V_i)$:11.11%,$\sigma(V_i)$:11.11%) when w =700. This highlights the fact that the higher the volume of information available for extracting each feature vector, the higher the granularity achieved in differentiating dominant features per sensor node located in the same part of the desalination plant.

With regard to the contents of $\mathbf{f}^*_{ij}$ when w =700, both significant variations and similarities can be observed for the multi-hop network performance at different locations of the desalination plant. First of all, both $\widehat{PRR}_{ij}$ and $\sigma(|P_{ij}|)$ are always compressed by another feature. Especially for $\sigma(|P_{ij}|)$ this implies that due to the limited connectivity options, the length of routing paths formulated in the industrial environment remains constant. In addition, ambient conditions are prominent in characterizing the network performance; the mean value of on-board temperature $\mu(T_i)$ appears in $\mathbf{f}^*_{ij}$ for the sub-networks at the four locations of the desalination plant, while especially for the locations where bulky water tanks are present (sea water inlet, security filters, and reverse osmosis) humidity becomes an additional key characteristic for the network performance. In addition, as expected the ambient RF noise over the routing paths is consistently present in the reduced feature vector, while its mean value $\mu(NF^*_{ij})$ and variance $\sigma(NF^*_{ij})$ are among the most popular features for the sub-network deployed at the machinery responsible for reserve osmosis, which is characterized as a dense metallic environment with a plethora of pumps and valves. Finally, the increased percentage of occurrence of f^{rx}_i at the sub-networks deployed at the sea water inlet and pre-treatment phases (22.22% and 20% respectively) highlight the increased relaying activity of the smart water sensor nodes deployed in those areas, as opposed to the sub-network located at the security filter which seems to have the longest routing paths to reach the sink node ($\mu(|P_{ij}|)$: 20%).

V. Conclusions and Future Work

In this work, we have addressed the problem of characterizing the multi-hop network performance by proposing a novel algorithm for graph-based, unsupervised feature selection. The evaluation studies on data collected from a real-life deployment in a water desalination plant emphasize the efficacy of our approach in terms of both quality and percentage of compression achieved. The results additionally highlight the importance of the volume of raw input streams considered for the extraction of the feature vector in improving the granularity of the differences between sensor nodes, at the expense of relaxing real-time constraints. Moreover, the discussion provided on the most popular dominant features highlights the fact that the network conditions that affect the network performance vary with respect to phases of the desalination process. Our current work concentrates on addressing the on-line and distributed limitations of the proposed algorithm,emphasizing both on the adaptive extraction of the original feature vector, as well as the incorporation of the network imperfections in the graph-based feature selection.

Acknowledgement

This work was funded by the DEDALE project contract no. 665044 within the H2020 Framework Program of the EC. We are also grateful to Acciona Agua (http://www.acciona-agua.com/) for providing the premises of La Tordera's desalination plant, as well as to Prof. Baltasar Beferull-Lozano and the Intelligent Signal Processing and Wireless Networks Lab (University of Agder, Norway), and Dr. Ioannis Glaropoulos for assisting with the WSN protocol stacks development and deployment at the industrial plant.

TABLE II. DOMINANT FEATURES FOR THE INDUSTRIAL SMART WATER NETWORK WHEN GRAPH-NC-RE WITH ADAPTIVE δ IS APPLIED

Features List		Sea Water Intake		Pre-treatment		Security Filters		Reverse Osmosis	
		w =300	w =700	w =300	w =700	w =300	w =700	w =300	w =700
1	$\sigma(PRX^*_{ij})$								
2	$\mu(PRX^*_{ij})$		11.11%	14.28%		11.11%	10%	15%	4.76%
3	$\sigma(LQI^*_{ij})$								4.76%
4	$\mu(LQI^*_{ij})$							5%	4.76%
5	$\sigma(NF^*_{ij})$	12.5%	11.11%		20%	11.11%	10%	5%	14.29%
6	$\mu(NF^*_{ij})$					11.11%		10%	14.29%
7	f^{tx}_i			28.57%	20%		10%	10%	14.29%
8	f^{rx}_i	25%	22.22%	28.57%	20%		10%	10%	9.52%
9	$\mu(\lvert P_{ij}\rvert)$				10%	22.22%	20%		
10	$\sigma(\lvert P_{ij}\rvert)$								
11	$\widehat{PRR}_{ij}$								
12	$\mu(T_i)$	25%	11.11%	14.28%	20%	22.22%	20%	10%	9.52%
13	$\sigma(T_i)$	12.5%	11.11%			11.11%	10%	10%	9.52%
14	$\mu(H_i)$		11.11%			11.11%	10%	10%	
15	$\sigma(H_i)$				10%				
16	$\mu(V_i)$	12.5%	11.11%	14.28%				15%	4.76%
17	$\sigma(V_i)$	12.5%	11.11%					10%	9.52%

REFERENCES

[1] "Smart water networks," http://www.swan-forum.com/. [Online]. Available: http://www.swan-forum.com/

[2] G. Xu, G. Q. Huang, and J. Fang, "Cloud asset for urban flood control," *Advanced Engineering Informatics*, 2015.

[3] T. T.-T. Lai, W.-J. Chen, K.-H. Li, P. Huang, and H.-H. Chu, "Triopusnet: Automating wireless sensor network deployment and replacement in pipeline monitoring," in *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, ser. IPSN '12. New York, NY, USA: ACM, 2012, pp. 61–72.

[4] N. Rapousis, M. Katsarakis, and M. Papadopouli, "Qowater: A crowdsourcing approach for assessing the water quality," in *Proceedings of the 1st ACM International Workshop on Cyber-Physical Systems for Smart Water Networks*, ser. CySWater'15. New York, NY, USA: ACM, 2015, pp. 11:1–11:6.

[5] M. Bertocco, G. Gamba, A. Sona, and S. Vitturi, "Experimental characterization of wireless sensor networks for industrial applications," *Instrumentation and Measurement, IEEE Transactions on*, vol. 57, no. 8, pp. 1537–1546, Aug 2008.

[6] N. Baccour, A. Kouba, C. A. Boano, L. Mottola, H. Fotouhi, M. Alves, H. Youssef, M. A. Ziga, D. Puccinelli, T. Voigt, K. Rmer, and C. Noda, *Radio Link Quality Estimation in Low-Power Wireless Networks*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer, 2013.

[7] T. Liu and A. E. Cerpa, "Data-driven link quality prediction using link features," *ACM Trans. Sen. Netw.*, vol. 10, no. 2, pp. 37:1–37:35, 2014.

[8] ——, "Temporal adaptive link quality prediction with online learning," *ACM Trans. Sen. Netw.*, vol. 10, no. 3, pp. 46:1–46:41, 2014.

[9] A. Panousopoulou, M. Azkune, and P. Tsakalides, "Feature selection for performance characterization in multi-hop wireless sensor networks," *Ad Hoc Networks*, 2016 (to appear, pending minor revision).

[10] C. Boano, M. Ziga, T. Voigt, A. Willig, and K. Romer, "The triangle metric: Fast link quality estimation for mobile wireless sensor networks," in *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International Conference on*, Aug 2010, pp. 1–7.

[11] V. Gungor, B. Lu, and G. Hancke, "Opportunities and challenges of wireless sensor networks in smart grid," *Industrial Electronics, IEEE Transactions on*, vol. 57, no. 10, pp. 3557–3564, Oct 2010.

[12] F. Schmidt, M. Ceriotti, N. Hauser, and K. Wehrle, "If you can't take the heat: Temperature effects on low-power wireless networks and how to mitigate them," in *Wireless Sensor Networks*, ser. Lecture Notes in Computer Science, T. Abdelzaher, N. Pereira, and E. Tovar, Eds. Springer International Publishing, 2015, vol. 8965, pp. 266–273.

[13] G. Anastasi, M. Conti, and M. Di Francesco, "A comprehensive analysis of the mac unreliability problem in ieee 802.15.4 wireless sensor networks," *Industrial Informatics, IEEE Transactions on*, vol. 7, no. 1, pp. 52–65, Feb 2011.

[14] A. Woo, T. Tong, and D. Culler, "Taming the underlying challenges of reliable multihop routing in sensor networks," in *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, ser. SenSys '03. New York, NY, USA: ACM, 2003, pp. 14–27.

[15] P. Mitra, C. A. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 301–312, Mar 2002.

[16] Z. Zhang and E. Hancock, "A graph-based approach to feature selection," in *Graph-Based Representations in Pattern Recognition*, ser. Lecture Notes in Computer Science, X. Jiang, M. Ferrer, and A. Torsello, Eds. Springer Berlin Heidelberg, 2011, vol. 6658, pp. 205–214.

[17] P. Moradi and M. Rostami, "A graph theoretic approach for unsupervised feature selection," *Engineering Applications of Artificial Intelligence*, vol. 44, pp. 33 – 45, 2015.

[18] X. Qi, E. Fuller, Q. Wu, Y. Wu, and C.-Q. Zhang, "Laplacian centrality: A new centrality measure for weighted networks," *Information Sciences*, vol. 194, pp. 240 – 253, 2012, intelligent Knowledge-Based Models and Methodologies for Complex Information Systems.

[19] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 167–172, Jan 2007.

[20] V. Rao and V. N. Sastry, "Unsupervised feature ranking based on representation entropy," in *Recent Advances in Information Technology (RAIT), 2012 1st International Conference on*, March 2012, pp. 421–425.

[21] G. Tzagkarakis, G. Tsagkatakis, D. Alonso, E. Celada, C. Asensio, A. Panousopoulou, P. Tsakalides, and B. Beferull-Lozano, "Signal and data processing techniques for industrial cyber-physical systems," in *Cyber Physical Systems: From Theory to Practice*, D. B. Rawat, J. Rodrigues, and I. Stojmenovic, Eds. CRC Press, USA, 2015.

[22] IEEE Std 802.15.4, *Part 15.4: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (LR-WPANs)*, 2011.

[23] D. Alonso-Roman, E. Celada-Funes, C. Asensio-Marco, and B. Beferull-Lozano, "Improving reliability and efficiency of communications in wsns under high traffic demand," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, April 2013, pp. 268–273.

[24] P. Di Marco, C. Fischione, G. Athanasiou, and P.-V. Mekikis, "Harmonizing mac and routing in low power and lossy networks," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 231–236.

[25] S. I. Dimitriadis, N. A. Laskaris, V. Tsirka, M. Vourkas, S. Micheloyannis, and S. Fotopoulos, "Tracking brain dynamics via time-dependent network analysis," *Journal of Neuroscience Methods*, vol. 193, pp. 145 – 155, 2010.