STEP: Semantics-Aware Sensor Placement for Monitoring Community-Scale Infrastructure

Andrew Chio
achio@uci.edu
University of California, Irvine
Irvine, California, USA

Jian Peng
jian.peng@ocpw.ocgov.com
Orange County Public Works
Orange, California, USA

Nalini Venkatasubramanian
nalini@uci.edu
University of California, Irvine
Irvine, California, USA

ABSTRACT

Built utility infrastructures provide essential services such as water, gas, and power to communities, and their resilient operation under anomalies and spurious events is critical. In this paper, we study the deployment of heterogeneous IoT sensors in geo-distributed infrastructure networks, using stormwater as a driving usecase. These systems are responsible for drainage and flood control, but in doing so, serve as conduits that carry pollutants to receiving waters. The timely detection of such events is challenging, due to the transient/random nature of pollutants, scarce historical data, and complexity of the system. We present STEP, an integrated framework for sensor placement that leverages the network structure and topology, behavioral properties (e.g., flow rate), and community semantics such as locations of facilities (e.g., commercial spaces, residential areas, and industrial plants, etc.). We identify key metrics to capture anomaly coverage and traceability, use past pollution incidents to inform sensor deployment, and model network operations through physics-based simulations and community-scale semantics. STEP is evaluated on six real-world stormwater networks, which show the efficacy of our approach over existing methods.

CCS CONCEPTS

• **Computing methodologies** → **Modeling and simulation**; • **Computer systems organization** → **Sensor networks**; • **Networks** → *Network structure*; *Network dynamics*.

KEYWORDS

sensor deployment, stormwater monitoring, heterogeneous anomalies, semantics-aware modeling

ACM Reference Format:
Andrew Chio, Jian Peng, and Nalini Venkatasubramanian. 2023. STEP: Semantics-Aware Sensor Placement for Monitoring Community-Scale Infrastructure. In *The 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23), November 15–16, 2023, Istanbul, Turkey*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3600100.3623752

BuildSys '23, November 15–16, 2023, Istanbul, Turkey

ACM ISBN 979-8-4007-0230-3/23/11.
https://doi.org/10.1145/3600100.3623752

1 INTRODUCTION

Communities rely on built utility infrastructures such as water, gas, and power as critical lifelines. These engineered systems are designed and maintained by municipal agencies and service providers. Recently, urban growth, climate change, and aging have given rise to multiple modes of failure, which are challenging to handle due to their continuous, transient, or sporadic natures [3, 32, 40]. Fortunately, Internet of Things (IoT) ecosystems and new data-driven methods have enabled smart monitoring solutions for improved operational efficiency and resilience. In this paper, we aim to design IoT/sensor placements to detect and trace anomalies for stormwater community lifelines. We model relevant system dynamics, and develop a framework for IoT deployment.

Stormwater networks, also called municipal separate storm sewer systems (MS4s), consist of catch basins, outfalls, and channels which drain rainwater and nuisance flows from urban areas to rivers, bays, and oceans. In doing so, they can transport pollutants that lead to water quality impairments that affect ecological processes, e.g., harmful bacteria, algal blooms, and ecotoxicity [4, 44]. In the US, regulations like the 1987 Clean Water Act amendment [16] prohibit such discharge into MS4s without permits. Here, enforcement and compliance is difficult, as discharges can enter via thousands of catch basins or unauthorized connections, impairing downstream waters [37]. Thus, rapid and effective detection of anomalies or large water quality deviations, is crucial to prevent illegal discharges and take corrective actions.

State-of-the-art approaches for detecting anomalies rely on inspections, citizen reports, and manual site visits [4, 6, 45]. This "grab sampling" covers a negligible area of the network, making it hard to capture illicit discharges, temporally and spatially. For example, Orange County Public Works conducts 5 annual site visits for 30 outfalls, where staff spent an hour per site for observation and testing. This costly and time-consuming procedure measures water quality using test kits and laboratory analysis, and take weeks to process. As a result, current approaches fail to provide comprehensive coverage and timely detection of potential issues. Fortunately, the rise of IoT smart city initiatives has opened opportunities for low-cost monitoring of utility lifelines [1]. Here, finding optimal sensor deployments for geo-distributed systems is hard, due to lack of fine grain knowledge. Understanding the evolution of anomalies and their root cause using grab samples is imprecise, and their transient nature requires sensor deployments to consider potential sources. Anomalies also consist of diverse chemicals and manifest through varying phenomena, such as turbidity, flow, and pH.

Several aspects of the infrastructure can help provide insight for suitable sensor deployments to monitor these anomalies. First, the network *structure* details physical aspects of MS4s (e.g., locations of catch basins, pipes, etc.); this can help establish a basis for candidate

deployment locations using properties like network centrality and distance. This is augmented with the *behavior* of a network, which captures the response to various stimuli (e.g. contaminants injected in a catch basin), and their impact. A third factor stems from the observation that the type of anomalies are influenced by the specific land use in a community (residential areas, commercial spaces, restaurants, etc.) - each with specific discharges and pollutants that vary in frequency, type, and chemical makeup. We refer to this as *community-level semantics*.

In this paper, we present STEP, a framework integrating *structural*, *behavioral*, and *semantic* aspects of an infrastructure network to gain insights into its operation. We use historical data and community-level semantics to construct realistic anomalies that inform a sensor placement optimization. We derive key topological and empirical properties from the network and physics-based simulations. To ensure feasibility in real settings, STEP provides tools for exploration and refinement of solutions. Key contributions include:

- A novel approach for sensor placement in built infrastructures using structure, behavior, and semantics (§2).
- A mathematical model to capture the structure and behavior of infrastructure, i.e stormwater networks, and compute notions of coverage and traceability (§3).
- A semantic approach for generating anomalies using historical data and community land use characteristics (§4).
- Formalizing and solving the heterogeneous sensor placement problem to address coverage and traceability, and a human-in-the-loop to refine placement (§5 and §7).
- A detailed evaluation of STEP on six real-world stormwater networks provided by domain experts (§6).

2 TRACING ANOMALIES

We address the role of sensor placement in isolating anomalies in community-scale infrastructures. In the stormwater setting, anomalies, i.e. contaminants, may be introduced sporadically at different locations/times; their propagation depends on physical attributes and may have a transient presence in the network. Such sporadic and transient anomalies must be captured using appropriate sensors at well-chosen locations. Our goal is to develop a sensor placement technique to select the types/locations of sensors for deployment, to detect and trace diverse anomalies. We review related work in fault/anomaly identification in water networks, address limitations, and then describe the STEP approach.

2.1 Related work and Limitations

The sensor placement problem has been studied in the drinking water domain, with the goal of finding suitable deployments to monitor a community network. Key objectives include: contaminant detection time [24, 30]; population impact [7, 39, 46]; and coverage [29, 41], among others. Early work [7] proposed mixed integer programming (MILP) to optimize a placement using limited sensors. This was extended in [8, 48] to address robustness when using failure-prone sensors. However, MILP often scales poorly, which has promoted greedy heuristics which exploit submodularity [17, 27, 29, 46]. Efforts to define multi-criteria objectives [27, 29], and the role of topology [17, 46] have also been explored. Others propose genetic algorithms, which can perform well with greedy heuristics, but trade off guarantees of optimality [24, 30, 39, 41]. Recently, methods have considered parallel architectures [14] and imperfect sensors [18].

After sensors are deployed, the next logical step examines the problem of anomaly detection. Prior work has utilized statistical tests, time-series analysis and machine-learning (ML) based methods for anomaly detection. Early techniques relied on scoring data and running statistical tests to determine outliers, including ANOVA [10], Grubb's test [15], PCA [11], and more. Time-series analysis methods, such as ARIMA [2, 5] were also developed, which leveraged moving averages and forecasting to capture and predict patterns/trends to detect anomalies. These techniques need significant human interpretation and tuning, making them difficult to apply at scale. Recently, this has given rise to ML models, which provide a framework for automation. Among these, were decision trees [25], support vector machines [12], Bayesian networks [19], and most recently, deep learning [31, 35].

A key limitation with several existing methods is the assumption of homogeneity in both sensors and contaminants, i.e., all sensors may detect the presence of all contaminants. In real-world cases, contaminants are varied and produce phenomena that propagate differently in the sensorized infrastructures. Sensor capabilities can also be varied in type and cost, which impact the quality of the resulting deployment. We argue that incorporating the appropriate type of sensor at meaningful locations is essential for instrumenting more of the network within a constrained budget. The feasibility of deployments proposed by algorithms also requires the involvement of domain experts and practitioners.

2.2 The STEP Approach and Architecture

We propose an integrated approach to the heterogeneous sensor placement problem, entitled STEP, which coalesces multiple aspects - *structural*, *behavioral*, and *semantic* - of the infrastructure to gain insight into a suitable deployment. As described earlier, network structure examines how nodes are positioned and interconnected, which can help derive metrics such as node centrality and distance. The operational aspects of the network capture the behavior of the system in response to stimulus (e.g., contaminant propagation through nodes), and are guided by physics-based principles. Then, semantics detail the correlation between properties of anomalies (e.g., frequency, type, environmental impacts), and their likelihood of occurrence when specific community infrastructures lie upstream (e.g., industrial plant). These *semantic land uses* capture how people interact with the community and land. Anomalies are observed by sensors, but budget limits of agencies dictate the extent of sensing and instrumentation. Thus, our goal is to find the best placement (sensor types/locations) to maximize the ability to cover and trace anomalies introduced in the network.

Fig. 1 shows the high-level framework of STEP. We assume that the underlying network structure and associated historical data are provided by domain experts a priori. Using domain knowledge and/or public data, e.g., OpenStreetMap [33], we obtain information on landuse. Together, this allows us to learn and model realistic anomalies which then informs a semantics-aware anomaly generation process. The constructed anomalies are used to observe network behavior via physics-informed simulations. STEP derives topological and empirical network properties, and proposes a sensor

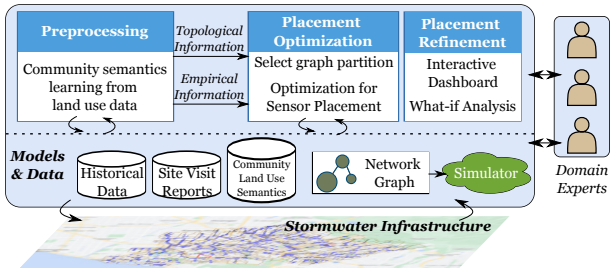


Figure 1: STEP Components and Workflow

placement to maximize the coverage and traceability of anomalies. Finally, we develop an interactive toolkit to facilitate effective deployments and incorporate domain-expert feedback.

3 MODELING NETWORK STRUCTURE AND BEHAVIOR

We model key elements of the infrastructure network and define several properties based on its structure and behavior.

Infrastructure Network Model. The geo-distributed infrastructure network is modeled as a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where node $v_j \in \mathcal{V}$ is a junction (candidate location for instrumentation) and edge $(v_i, v_j) \in \mathcal{E}$ is a conduit with flow from v_i to v_j. $\mathcal{G}$ is characterized by several properties to model the physics of flow/contaminant propagation. Let $pa(v_j)$ be the direct *parents* of v_j in $\mathcal{G}$, and $path(v_i, v_j)$ be the *path* of nodes observing flow from v_i to v_j.

Community Model. To model the community, let semantic land use $u_m \in \mathcal{U}$ express how citizens utilize and interact with the land (e.g., industrial, residential, etc.). We observe that each u_m can produce different types of anomalies in the network. For instance, industrial areas are more likely to release harmful chemicals than residential areas. For node v_j, let $\mathcal{U}(v_j)$ denote the semantic land uses near v_j, and $Area(v_j, u_m)$ denote the area near v_j with land use u_m. Each u_m is assigned a priority λ_m (defined by a domain expert) that represents the importance of monitoring anomalies from u_m.

Sensor Model. We define a sensor $s_l \in \mathcal{S}$ with the 3-tuple $(p_l, \epsilon_l^{acc}, c_l)$, which describes a sensor measuring phenomenon p_l with Gaussian error ϵ_l^{acc}. Its cost c_l includes purchasing hardware, deploying it in the field, and maintaining it over time. We let $\mathcal{S}(p)$ be the set of sensors measuring p.

Anomaly Model. We define a transient anomaly $\alpha_k \in \mathcal{A}$ using the 5-tuple $(v_k^*, t_k^s, t_k^e, \mathcal{P}_k, u_k)$, which describes an anomaly originating at node v_k^* with duration (t_k^s, t_k^e). The phenomena $\mathcal{P}_k$ are produced by α_k, and detected by a sensor s_l iff its measured phenomenon p_l is in $\mathcal{P}_k$. The anomaly is more likely to be produced by the land use u_k. We use $\mathcal{A}(v_i)$ to denote the set of anomalies whose origin node is v_i. We let $time(\alpha_k, v_k^*, v_j)$ be the time taken for α_k to propagate from v_k^* to node v_j (which is ∞ if flow from v_k^* cannot reach v_j).

Definition: Placement. We represent a candidate placement, $\mathcal{X}$, as a matrix whose entries $x_{lj} = 1$ iff a sensor s_l is *deployed* at node v_j, and 0 otherwise.

Definition: Node Coverage. A typical definition of network coverage is purely structural, i.e. based on how many nodes fall within sensor range. In contrast, we define coverage as the ability to capture a set of anomalies in the network. We look to optimize the node coverage COV, i.e., the proportion of nodes monitored by the placement $\mathcal{X}$, wrt. the set of anomalies $\mathcal{A}$, as in Eqn. 1a. A node v_i is *covered* by $\mathcal{X}$ if at least ρ% of the anomalies originating at v_i can be detected by downstream sensors in $\mathcal{X}$, as shown in Eqn. 1b. This implies that sensor s_l must observe anomaly α_k, i.e., $p_l \in \mathcal{P}_k$, and the propagation time to an instrumented node v_j is bounded by τ, i.e., $time(\alpha_k, v_k^*, v_j) \leq \tau$. We express these with $OB(l,k)$ and $PT(k,j)$. Note that v_j must lie downstream of v_k^* in $\mathcal{G}$ for $time(\alpha_k, v_k^*, v_j)$ to be bounded by τ. The indicator function "$\mathbb{1}[stmt]$" is 1 if statement $stmt$ is true, and 0 otherwise.

$$COV(\mathcal{X}, \mathcal{A}, \mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} covered(v_i, \mathcal{X}, \mathcal{A}(v_i)) \tag{1a}$$

$$covered(v_i, \mathcal{X}, \mathcal{A}(v_i)) =$$

$$\mathbb{1}\left[\sum_{\alpha_k \in \mathcal{A}(v_i)} \sum_{v_j \in \mathcal{V}} \sum_{s_l \in \mathcal{S}} x_{lj} OB(l,k) PT(k,j) \geq \rho \, |\mathcal{A}(v_i)| \right] \tag{1b}$$

Definition: Traceability. We define *traceability* to describe the degree to which sensor observations can help track the origin of an anomaly. Let $\mathcal{G}^{up}_{v_j}$ be the *upstream subgraph* of node v_j induced from the nodes $\mathcal{V}^{up}_{v_j}$ upstream of v_j. Then, let $\mathcal{G}^{up}_{v_j,\alpha_k,\mathcal{X}}$ be an induced subgraph of $\mathcal{G}^{up}_{v_j}$, consisting of nodes v_i where the anomaly α_k originating at v_k^* would *first* be observed by a sensor at v_j. Then, the traceability TR for placement $\mathcal{X}$ is the average proportion of nodes that lie in each $\mathcal{G}^{up}_{v_j,\alpha_k,\mathcal{X}}$, for the given anomalies $\mathcal{A}$, as in Eqn. 2.

$$TR(\mathcal{X}, \mathcal{A}, \mathcal{G}) = \frac{1}{|\mathcal{A}|} \sum_{\alpha_k \in \mathcal{A}} \sum_{s_l \in \mathcal{S}(\mathcal{P}_k)} \sum_{v_j \in \mathcal{V}} \left| \mathcal{V}^{up}_{v_j,\alpha_k,\mathcal{X}} \right| \Big/ |\mathcal{V}| \tag{2}$$

Definition: Betweenness Centrality. Identifying nodes through which more flow propagation occurs can indicate natural deployment candidates. The *betweenness centrality* of a node v_j empirically measures the number of anomalies observed at v_j before time threshold τ, as shown in Eqn. 3.

$$\mathcal{BTN}(v_j) = \sum_{\alpha_k \in \mathcal{A}} \mathbb{1}\left[time(\alpha_k, v_k^*, v_j) \leq \tau \right] \tag{3}$$

Definition: Branching Complexity. The complexity of *branching* in a network, i.e., merges/splits at nodes, is important in determining its traceability. Networks with high branching have more junctions through which flows combine, which then requires more sensors to monitor. We define branching complexity $\mathcal{BC}$ in Eqn. 4, which formalizes the notion that an upstream graph of a "chain" structure is easier to trace than a complex "tree" structure.

$$\mathcal{BC}(v_j) = \begin{cases} 1 & \text{if IsRoot}(v_j) \\ \mathcal{BC}^{max}_{pa(v_j)} + \sum_{v_i \in pa(v_j)} \dfrac{\mathcal{BC}(v_i)}{\mathcal{BC}^{max}_{pa(v_j)}} - 1 & \text{else} \\ \text{where: } \mathcal{BC}^{max}_{pa(v_j)} = \max_{v_i \in pa(v_j)} \mathcal{BC}(v_i) & \end{cases} \tag{4}$$

Physics-informed model for infrastructure behavior. Network behavior, as a response to stimuli (e.g., anomalies), are fundamentally governed by physical laws of the conservation of mass (Eqn. 5a) and momentum (Eqn. 5b). The Environmental Protection Agency Storm Water Management Model (EPA SWMM) [22] is a simulator developed by domain experts that models and solves these equations to portray the operation of the network. Multiple physical attributes are utilized, including the distance x, time t, flow area A, flow rate Q, hydraulic head H, friction slope S_f and gravity g. Details on solving these equations are provided in [22].

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0 \tag{5a}$$

$$\frac{\partial Q}{\partial t} + \frac{\partial (Q^2/A)}{\partial x} + gA\frac{\partial H}{\partial x} + gAS_f = 0 \tag{5b}$$

4 SEMANTICS-AWARE MODELING AND GENERATION OF ANOMALIES

We show how potential anomalies are extracted from past data to inform the sensor deployment process. Our workflow profiles anomalies from historical water quality grab sample data, and correlates potential origins with nearby community-level semantics. STEP then generates new potential anomalies based on the land uses in the network.

Let the historical water quality grab sample dataset $\mathcal{D}$ be a set of 5-tuples (t, v, f, o, p), representing a flow f that propagates water quality observations o of a phenomenon p at node v at time t. In MS4s, observed phenomena include turbidity, temperature, flow rate, etc. [9]. We assume such data is captured by water agencies for regulatory compliance.

Extracting anomalies from water quality data. We describe how physical features of anomalies are extracted from historical data, and used to learn a semantic map. This starts by constructing a uniformly distributed set of anomalies which may occur in a given network, and simulating their propagation using a physics-based simulator like [22]. Anomalies are then grouped based on their initial features and simulated behavior in the network using agglomerative clustering [20]. These groups represent profiles of anomalies with similar impact in the network. Each grab sample measurement in $\mathcal{D}$ is matched to the closest anomaly profile above. This is used to identify a set of potential origin nodes, and their corresponding start and end times. Note that the phenomena produced by the historical anomaly is given, and labeled by domain experts using simple thresholds.

The semantic map constructed using historical data depends on the origin node of the anomaly, and the semantic land uses that lie upstream. To identify a correlation between semantic land uses and different sets of phenomena, we construct a probabilistic semantic map based on the anomalies extracted from $\mathcal{D}$. For each node v at which historical data was captured, let $\mathcal{M}_v^{up}$ denote the total area of each semantic land use u that lies upstream of v. We then sum $\mathcal{M}_v^{up}$ across these nodes for each semantic land use, and normalize as needed. After this semantic map is constructed, we iterate through the set of historical anomalies, and assign a semantic land use "cause", based on the set of potential origin nodes.

Generating new anomalies using semantics We use the semantic map $\mathcal{M}$ constructed above to generate N new potential anomalies for the infrastructure network. We construct a new anomaly α_k by selecting a candidate semantic land use u_k to serve as the "cause" of the anomaly. Here, each u_k has weight equal to the number of historical anomalies that correlated with u_k in $\mathcal{M}$. This semantic land use is leveraged in selecting all other physical aspects of the new anomaly. In particular, the candidate origin node v_k^* is chosen with weight equal to the semantic land use area $Area(v_k^*, u_k)$, i.e., nodes with more area for u_k are chosen with higher likelihood. The time period (t_k^s, t_k^e) is chosen by sampling a normal distribution on the start/end time of historical anomalies caused by u_k. Lastly, the set of phenomena $\mathcal{P}_k$ produced is chosen based on the correlation to u_k in $\mathcal{M}$.

Algorithm 1: Generate Semantic Anomalies

Input: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, Historical Anomalies $\mathcal{A}_{hist}$, SemanticMap $\mathcal{M}$, int N
Output: Anomalies $\mathcal{A}_{sem}$

$\mathcal{A}_{sem} \leftarrow \emptyset$
for $i \leftarrow 1..N$ **do**
  $u_k \leftarrow PickLanduse(\mathcal{A}_{hist}, \mathcal{M})$
  $v_k^*, t_k^s, t_k^e \leftarrow PickOriginAndDuration(\mathcal{V}, u_k, \mathcal{M})$
  $\mathcal{P}_k \leftarrow PickPhenomena(u_k)$
  $\mathcal{A}_{sem} \leftarrow \mathcal{A}_{sem} \cup \{Anomaly(v_k^*, t_k^s, t_k^e, \mathcal{P}_k, u_k)\}$
return $\mathcal{A}_{sem}$

5 THE INTEGRATED STEP SOLUTION AND PLACEMENT ALGORITHM

The STEP approach solves the heterogeneous sensor placement problem by integrating structural, behavioral and semantic properties of an infrastructure network. We introduce a novel information theoretic abstraction called *semantic entropy*, and utilize it with the network properties defined in §3 to develop a scalable solution for coverage and traceability.

Definition: Semantic Entropy. The skewness in the distribution of upstream semantic land uses can provide insight into the potential causes of an anomaly. For instance, if an anomaly occurs in a region primarily consisting of a specific semantic land use, it becomes easier to attribute the anomaly to that particular land use. This can be utilized to develop a knowledge base for identifying the potential causes of new anomalies. Eqn. 6 defines the semantic entropy $\mathcal{SE}$ for semantic land uses $\mathcal{U}$; in the graph $\mathcal{G}_{v_j}^{up}$.

$$\mathcal{SE}(\mathcal{U}, \mathcal{G}_{v_j}^{up}) = \sum_{u_m \in \mathcal{U}(v_j)} \lambda_m \cdot \left(-P(u_m)\log P(u_m)\right)$$
$$\text{where: } P(u_m) = \sum_{v_i \in \mathcal{V}_{v_j}^{up}} \left(\frac{Area(v_i, u_m)}{\sum_{u_m \in \mathcal{U}} Area(v_i, u_m)} \right) \tag{6}$$

We adapt the information-theoretic definition of *entropy* in [42] to measure the skewness of semantic land uses. Let $P(u_m^{up})$ indicate the proportion of the total area allocated for land use $u_m \in \mathcal{U}$ in the nodes of $\mathcal{G}_{v_j,p,\chi}^{up}$, to the total area of regions in $\mathcal{G}_{v_j}^{up}$. Then, the *semantic entropy* is defined as the sum of $-P(u_m^{up})\log P(u_m^{up})$, over all semantic land uses $u_m \in \mathcal{U}$, weighted by priority λ_m of u_m. This represents the weighted average amount of "information" (i.e., bits) needed to describe the upstream distribution of semantic land uses.

5.1 Placement Optimization

The core optimization relies on mixed integer linear programming (MILP), to produce optimal sensor deployments for a given network. However, due to the large, geo-distributed scale of many utility infrastructures, pure MILP solutions can be prohibitively difficult to solve without reducing its complexity. Thus, we utilize a graph partitioning strategy to split the network into smaller pieces to solve, and merge partial solutions using heuristics and domain expert feedback.

Graph Partitioning. To ensure that the MILP is tractable, we first partition the infrastructure graph $\mathcal{G}$ into multiple smaller subgraphs using the network properties defined above. Here, it is important to find graph partitions that group key components of the network together, so that an optimum can be found within a partition.

To this end, we note the role that the betweenness centrality and branching complexity play in influencing the coverage and traceability of a proposed placement. Since the betweenness centrality measures the frequency of anomalies that pass through nodes and the branching complexity measures the upstream flow structure, partitioning on these metrics enables coverage and traceability requirements to be met. We define $\Delta\mathcal{BTN}(v_j)$, $\Delta\mathcal{BC}(v_j)$, and $\Delta\mathcal{SE}(v_j)$ to represent the total change between the respective metrics at node v_j and its parents $pa(v_j)$. This reflects the measured quantity at v_j without influence from upstream nodes of v_j, which we use to greedily partition the graph $\mathcal{G}$. That is, our graph partitioning selecting nodes to instrument which most decrease the mean betweenness centrality, branching complexity, and semantic entropy. The tradeoff between coverage and traceability is decided with a weight on $\Delta\mathcal{BTN}(\cdot)$, $\Delta\mathcal{BC}(\cdot)$ and $\Delta\mathcal{SE}(\cdot)$. We repeat this for the number of partitions desired, N_{part}. Note that placing all sensors in $\mathcal{S}$ at this point will help to minimize the worst case coverage and traceability. We detail this process in Alg. 2.

Formulating a MILP and Merging the Solution. For each subgraph, we formulate and solve the MILP in Eqn. 7. The objective function defined in Eqn. 7a considers both coverage and traceability, which implicitly capture the sensor placement's capacity to detect anomalies, and trace them to a set of potential sources, respectively. There are two primary constraints considered in the formulation: 7b limits the budget allowed, while 7c limits the number of sensors measuring a specific phenomenon at a node.

$$\max \sum_{\alpha_k \in \mathcal{A}} \sum_{x_{jl} \in \mathcal{X}} x_{lj} w_{cov} COV(x_{lj}, \alpha_k) + x_{lj} w_{tr} TR(x_{lj}, \alpha_k) \quad (7a)$$

$$\text{subject to :} \quad \sum_{s_l \in \mathcal{S}} \sum_{v_j \in \mathcal{V}} x_{lj} c_l \leq B^c \quad (7b)$$

$$\sum_{s_l \in \mathcal{S}(p)} x_{lj} \leq 1 \qquad \forall v_j \in \mathcal{V}, \forall p \in \mathcal{P} \quad (7c)$$

Lastly, we merge the placement solutions obtained by the MILPs. Then, each node in the placement is adjusted based on whether its migration to an adjacent node can improve the global coverage and traceability objectives.

Placement Refinement. While our algorithm generates an "ideal" solution, deploying sensors at these locations may be infeasible or ill-advised. For instance, external factors such as potential vandalism, location-specific communication issues, and physical barriers preventing human access can require changes to a proposed placement. To aid domain experts in constructing and refining a sensor placement solution, we develop a STEP interactive toolkit that provides user-level visualization for each step of our approach. This allows a human-in-the-loop (i.e., domain expert) to insert regional infrastructure networks, generate ideal placements, and alter the suggested placement as desired. Such what-if analysis can leverage domain expert feedback from the field and is critical in effective community scale deployments. More details on the toolkit are presented later in §7.

Algorithm 2: Sensor Placement

Input: Graph $\mathcal{G}$, Sensors $\mathcal{S}$, Anomalies $\mathcal{A}$, int N_{part}, Budget B
Output: Placement $\mathcal{X}$

$placements \leftarrow \emptyset$; $G_{subs} \leftarrow \emptyset$; $best \leftarrow 0$; $v_{part} \leftarrow null$
for $i \leftarrow 1..N_{part}$ **do**
  for $v_j \in \mathcal{V}$, *for each subgraph* **do**
    $score \leftarrow w_{cov}\Delta\mathcal{BTN}(v_j) + \frac{w_{tr}}{2}\Delta\mathcal{BC}(v_j) + \frac{w_{tr}}{2}\Delta\mathcal{SE}(v_j)$
    if $score < best$ **then** $best \leftarrow (score, v_j)$;
    $G_{subs} \leftarrow G.AddSplit(v_{part})$
$nodes \leftarrow GetPartitionNodes(G_{subs})$
$placement \leftarrow placement \cup Sensorize(nodes)$
for $\mathcal{G}' \leftarrow \mathcal{G}_{subs}$ **do**
  $\mathcal{A}' \leftarrow \bigcup_{v \in \mathcal{V}'} \mathcal{A}(v)$
  $B' \leftarrow B \cdot \frac{|\mathcal{A}'|}{|\mathcal{A}|}$; // Budget for subgraph
  $\mathcal{X}' \leftarrow$ Use MILP to solve Eqn. 7 with budget B'
  Add $\mathcal{X}'$ to $placements$
$\mathcal{X} \leftarrow AdjustPlacement(placement, hops = 5)$
return $\mathcal{X}$

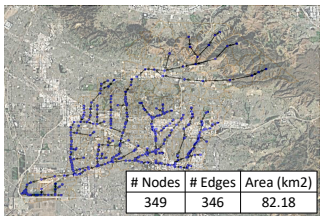
6 EXPERIMENTS

We evaluate the STEP framework for six real-world stormwater networks. We compare STEP against multiple baseline techniques for sensor placement, and analyze the number of anomalies detected, their traceability, and nodes coverage.

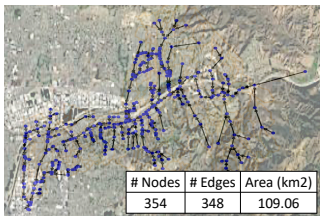
6.1 Experimental Setup

Real-world Networks. STEP is evaluated on six real-world stormwater networks covering cities in Southern California in the US. The networks were provided by Orange County Public Works (OCPW) and defined using EPA SWMM [22]. Fig. 2 visualizes the structure of the networks and summarizes basic properties. We leverage the definition of *subcatchments* within the EPA SWMM models to specify the region surrounding nodes in the network. Three categories of semantic land uses are defined: (i) high priority land uses with priority $\lambda_{=}3$: agriculture, commercial-service, industrial; (ii) medium priority land uses with priority $\lambda_{=}2$: mixed commercial and mixed urban; (iii) low priority land uses with priority $\lambda_{=}1$: hi-density residential, lo-density residential.

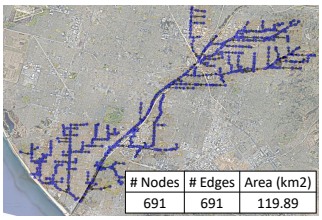
Historical Data. We use historical grab sample data provided by OCPW, which details instances where anomalous behavior was



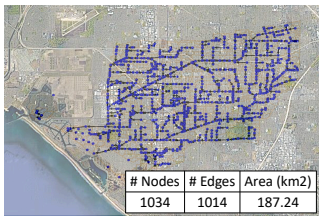
(a) Coyote Creek Upstream (Network - Small 1)



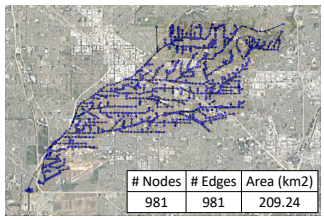
(b) Santa Ana Upstream (Network - Small 2)



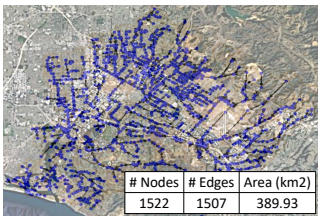
(c) Santa Ana Downstream (Network - Small 3)



(d) Anaheim (Network - Medium 1)



(e) Coyote Creek Downstream (Network - Medium 2)



(f) Newport Beach (Network - Large 1)

Figure 2: EPA SWMM Networks used for Evaluation

reported in a network and several water quality metrics were captured. The dataset contains 1292 historical grab samples from 30 different locations between 2006 and 2022 throughout each evaluated stormwater network.

Sensors. We specify the water quality sensors to deploy in Table 1. Each sensor measures a different phenomenon, and was developed by [13, 43, 47]. The sensor costs vary from $100 to $150 for hardware and deployment. Recurrent costs for continued operation and maintenance (e.g. cellular dataplan, battery replacements) range from $300 to $350 per year, based on rates at which these sensors stop logging data. The accuracy of the sensor is a constant or a percentage of the quantity of the measured phenomenon, and is empirically derived. In our simulations, we assume that sensors can only observe an anomaly if the percent difference between its observed value and its simulated value is under 30%.

Table 1: Sensors considered in placement

Phenomenon	Accuracy	Hardware & Depl. Cost	Op. Cost
Turbidity	11.6%	$100	$300
Depth	1 mm	$150	$350
Temperature	$0.5°C$	$200	$300
Electric Cond.	10%	$150	$300
Velocity	5 mm/s	$150	$350

Anomalies. To obtain the empirical measurements necessary to compute the metrics defined in §5, we construct two sets of anomalies. First, we define 5 anomaly instances uniformly across all nodes in each network. These anomalies have a random duration of 30±5 minutes and flow rate of 0.2±0.2 cfs. The set of phenomena produced is randomly sampled between the phenomena in Table 1. Then, a more realistic set of anomalies, based on historical data and semantic land uses was generated using the methodology in §4 for evaluating the proposed placements.

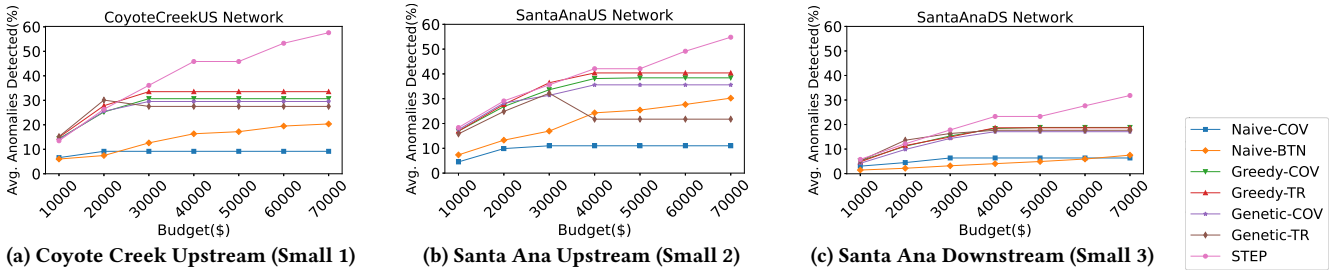
Comparison Algorithms. We compare STEP against two common baseline algorithms for sensor placement optimization. The Greedy Heuristic (Greedy) [26, 36, 38] is an algorithm that selects sensors to deploy based on their ability to maximize a given criteria. We consider two classes of baseline algorithms that operate greedily. The first class aims to only leverage structural, i.e., topological, network properties to perform the sensor placement. These two algorithms, Naive-COV and Naive-BTN select sensor deployment locations based on the radius-based definition of coverage, and the global betweenness centrality, respectively. The second class of greedy algorithms leverages both topological and empirical network properties, to maximize coverage (COV) and traceability (TR) directly as objectives, using the definitions provided in §3. We also explore another commonly used sensor deployment strategy: the Genetic Algorithm (Genetic) [21, 28, 34]. This technique searches for a sensor deployment by simulating the process of natural selection and evolution. In our experiments, we consider a population size of 1000, crossover rate of 0.8, and mutation rate of 0.01. We similarly use coverage and traceability as objectives. As mentioned previously, these baselines use the uniform distribution of anomalies for optimization, but are evaluated on the more realistic, semantic-aware distribution of anomalies.

Performance Metrics. We first evaluate the efficacy of our approach on the number of anomalies detected by each proposed placement. Then, we compare the coverage and traceability provided by the placements, as defined in Eqn. **??** and 2. For each comparison, we report the range of percent differences between our approach and each baseline. This value is computed using $|v(STEP) - v(CMP)|/v(STEP)$, where $v(STEP)$ and $v(CMP)$ represent the metric value from STEP and a comparison algorithm, respectively.

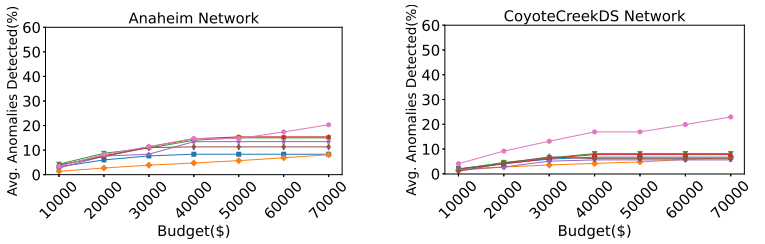
6.2 Experimental Results

We compare STEP on the number of anomalies detected, their traceability and the number of nodes effectively monitored.

Detected Anomalies. In Fig. 3, we first report the average number of anomalies that were captured by the proposed sensor deployment, for each of the evaluated networks as a function of budget. The results show that the STEP approach was generally able to outperform each of the other baselines wrt. the number of

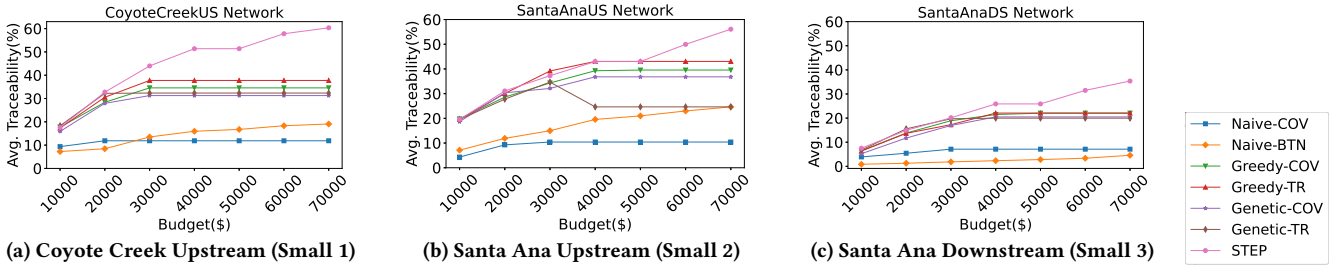


(a) Coyote Creek Upstream (Small 1) (b) Santa Ana Upstream (Small 2) (c) Santa Ana Downstream (Small 3)

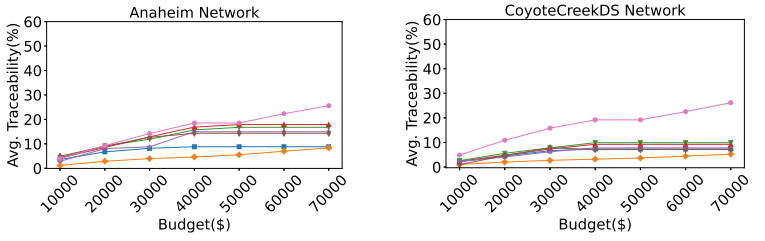


(d) Anaheim (Med 1) (e) Coyote Creek Downstream (Med 2) (f) Newport Beach (Large 1)

Figure 3: Number of anomalies detected in evaluation networks



(a) Coyote Creek Upstream (Small 1) (b) Santa Ana Upstream (Small 2) (c) Santa Ana Downstream (Small 3)

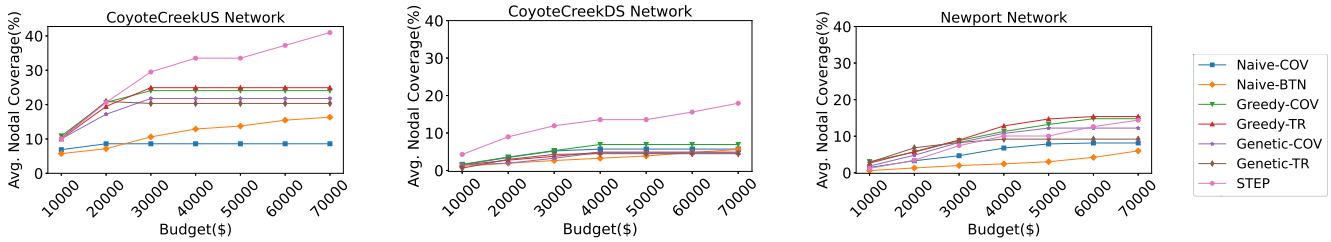


(d) Anaheim Network (Med 1) (e) Coyote Creek Downstream (Med 2) (f) Newport Beach (Large 1)

Figure 4: Evaluation of Network Node Traceability vs Budget

anomalies detected. The average percent difference between STEP and the baseline approaches for the highest budget limit evaluated, ranged between 35.66–528.13% for small networks, 32.08–309.10% for medium networks, and 0.74–206.90% for the large network. This implies that using the STEP approach, which leverages semantics in addition to structural and behavioral aspects of the network was effective for monitoring the semantically generated anomalies. We observe that the largest factor in determining the actual proportion of anomalies captured lies in the number of nodes in the network. That is, the smallest networks (Coyote Creek Upstream and Santa Ana Upstream) have the highest proportion of anomalies captured, while the largest network (Newport Beach) has the smallest. We note that since the budget range used in the evaluation is similar for each size of network, the placement proposed for the largest network was unable to be significantly distinguished from other solutions. In general, this result shows that our approach was able to propose a sensor deployment that can effectively monitor the network for anomalies.

Traceability. We next examine the degree to which anomalies can be traced to a potential source. Fig. 4 reports the traceability of anomalies observed in the network. We plot the average proportion of nodes that were eliminated as potential sources for each anomaly. Note that if an anomaly is undetected, no nodes can be eliminated. The average percent difference between the traceability enabled by STEP and the baselines for the largest budget limit, ranged from 30.30–671.65%, 43.12–400.36%, and 2.95–272.75%, for the small, medium, and large sized networks. We similarly find that the size of the network most impacts the traceability - since smaller networks were generally able to detect more anomalies, they were also able to apply reasoning to eliminate potential source nodes. This show that STEP deployed sensors in locations that balanced the tradeoff



(a) Coyote Creek Upstream (Small 1) (b) Coyote Creek Downstream (Med 2) (c) Newport Beach (Large 1)

Figure 5: Evaluation of Nodal Coverage vs Budget

between detecting anomalies, and tracing them back to potential origins.

Node Coverage. Lastly, we examine the number of nodes monitored by the proposed sensor deployments. Fig. 5 shows node coverage provided by the proposed placements for one of small, medium, and large sized networks. The results show that small and medium sized networks receive better node coverage than when compared to any of the baselines. The average percent difference for these networks range between 27.45–376.67% and 43.23–300.00%, respectively. However, we note that for the large network, this percent difference drops to 2.67–140.65%. Due to the limited budget explored for deployment in this large network, we can see that all placement algorithms perform sub-optimally wrt. the number of nodes covered, which also leads to a loss in traceability.

7 TOWARDS A STEP PROTOTYPE

To aid domain experts and other stakeholders in visualizing and refining a proposed sensor placement, we developed a prototype system to provide an interface to interact with STEP. The detailed architecture and prototype system is shown in Fig. 6. The workflow for our system fundamentally leverages the infrastructure graph, historical data, and semantics-level data described earlier, from which topological and empirical network properties are derived.

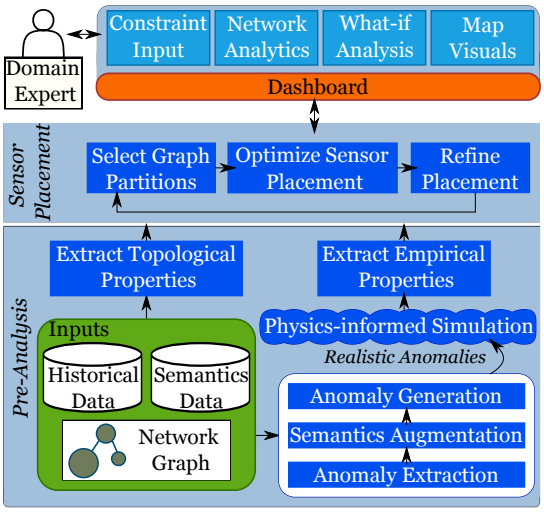


Figure 6: The STEP Prototype Architecture

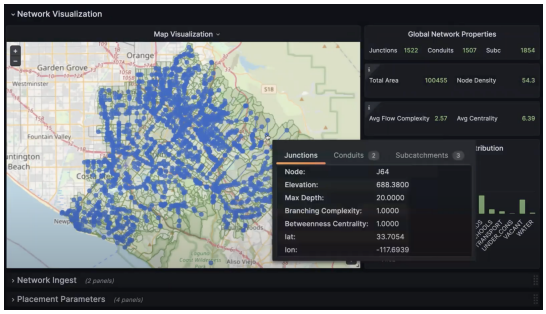


Figure 7: The STEP Interactive Dashboard

This informs a potential sensor deployment, which is provided to domain experts through our dashboard in Fig. 7.

We envision that this is used to explore local refinements to a proposed sensor placement when certain nodes are deemed unsuitable for instrumentation because of external factors (e.g., vandalism, communication issues, physical barriers to human access, etc.). Then, network analytics and what-if analysis for new coverage and traceability values when such changes occur are presented, alongside suggestions for possible refinements. We published the STEP prototype toolkit and dashboard on GitHub [23].

8 CONCLUSIONS AND FUTURE WORK

We presented STEP, a framework for the heterogeneous sensor placement problem which leverages structural, behavioral, and semantic aspects of community infrastructure. STEP relies on historical grab sample data and community-level semantics to learn and construct realistic anomalies for a specific infrastructure network. Evaluations performed on six real-world stormwater infrastructure networks show that STEP is able to balance the tradeoffs between coverage and traceability. To aid in the usability of STEP, we developed an interactive dashboard for visualizing and refining the proposed sensor placement. As part of our future work, we will integrate the proposed deployments into real-world stormwater networks for anomaly source identification. Our experiences here will be also used to identify elements for automation to reduce the effort required by domain experts.

ACKNOWLEDGMENTS

This work is supported by the UC National Laboratory Fees Research Program Grant No. L22GF4561, and National Science Foundation NSF Grants No. 1952247 and 2008993. Any opinions, findings, and conclusions or recommendations expressed in this material are

those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] M. Al-Hader and A. Rodzi. 2009. The smart city infrastructure development & monitoring. *TERUM* 4, 2 (11 (2009).

[2] P. Arumugam and R Saranya. 2018. Outlier detection and missing value in seasonal ARIMA model using rainfall data. *Materials Today: Proceedings* 5, 1 (2018).

[3] ASCE. 2021. Stormwater. https://infrastructurereportcard.org/cat-item/stormwater-infrastructure/

[4] A. Barbosa et al. 2012. Key issues for sustainable urban stormwater management. *Water research* 46, 20 (2012).

[5] D. Barrientos-Torres et al. 2023. Water Flow Modeling and Forecast in a Water Branch of Mexico City through ARIMA and Transfer Function Models for Anomaly Detection. *Water* 15, 15 (2023).

[6] B. Bernstein et al. 2009. Assessing urban runoff program progress through a dry weather hybrid reconnaissance monitoring design. *Environ. Monit. Assess.* 157 (2009).

[7] J. Berry et al. 2006. Sensor placement in municipal water networks with temporal integer programming models. *J. Water Resour. Plan. Manag.* 132, 4 (2006).

[8] J. Berry et al. 2009. Designing contamination warning systems for municipal water networks using imperfect sensors. *J. Water Resour. Plan. Manag.* 135, 4 (2009).

[9] J. Bertrand-Krajewski et al. 1998. Distribution of pollutant mass vs volume in stormwater discharges and the first flush phenomenon. *Water Res.* 32, 8 (1998).

[10] J. Blanch et al. 2009. Arima models for data consistency of flowmeters in water distribution networks. *IFAC Proc. Volumes* 42, 8 (2009).

[11] N. Branisavljević et al. 2011. Improved real-time data anomaly detection using context classification. *J. Hydroinformatics* 13, 3 (2011).

[12] A. Candelieri. 2017. Clustering and support vector regression for water demand forecasting and anomaly detection. *Water* 9, 3 (2017).

[13] S Catsamas et al. .. Characterisation and development of a novel low-cost radar velocity and depth sensor.

[14] C. Ciaponi et al. 2019. Reducing impacts of contamination in water distribution networks: a combined strategy based on network partitioning and installation of water quality sensors. *Water* 11, 6 (2019).

[15] T. A Cohn et al. 2013. A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. *Water Resour. Res.* 49, 8 (2013).

[16] C. Copeland. 1999. Clean Water Act: a summary of the law. Congressional research service, Library of Congress Washington, DC.

[17] S. Das and S. K Udgata. 2021. Sensor Placement for Contamination Source Detection in Water Channel Networks. In *IEEE ICC*. IEEE.

[18] C. de Winter et al. 2019. Optimal placement of imperfect water quality sensors in water distribution networks. *Comput. Chem. Eng.* 121 (2019).

[19] N. Ding et al. 2018. Multivariate-time-series-driven real-time anomaly detection based on bayesian network. *Sensors* 18, 10 (2018).

[20] D Eads. 2008. hcluster: Hierarchical clustering for SciPy. *URL http://scipy-cluster.googlecode. com* (2008).

[21] N. Ehsani and A. Afshar. 2010. Optimization of contaminant sensor placement in water distribution networks: multi-objective approach. In *Water Distrib. Syst. Anal. 2010*.

[22] EPA. 2023. EPA Stormwater Management Model (SWMM). https://www.epa.gov/water-research/storm-water-management-model-swmm

[23] https://github.com/andrewgchio/STEP. 2023. GitHub Repository.

[24] C. Hu et al. 2015. A MapReduce based Parallel Niche Genetic Algorithm for contaminant source identification in water distribution network. *Ad Hoc Netw.* 35 (2015).

[25] D. Jalal and T. Ezzedine. 2020. Decision tree and support vector machine for anomaly detection in water distribution networks. In *IWCMC*. IEEE.

[26] M. L Kansal et al. 2012. Identification of optimal monitoring locations to detect accidental contaminations. In *World Environ. Water Resour. Congr. 2012: Crossing Boundaries*.

[27] A. Krause et al. 2008. Efficient sensor placement optimization for securing large water distribution networks. *J. Water Resour. Plann. Manage.* 134, 6 (2008).

[28] A. Krause et al. 2009. Simultaneous placement and scheduling of sensors. In *IPSN*. IEEE.

[29] A. Krause and C. Guestrin. 2009. Optimizing sensing: From water to the web. *Computer* 42, 8 (2009).

[30] A. Kumar et al. 1999. Detecting accidental contaminations in municipal water networks. *J. Water Resour. Plan. Manag.* 125, 5 (1999).

[31] D. Li et al. 2019. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *ICANN*. Springer.

[32] Jason R Masoner et al. 2019. Urban stormwater: An overlooked pathway of extensive mixed contaminants to surface and groundwaters in the United States. *Environmental science & technology* 53, 17 (2019).

[33] OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

[34] A. Preis and A. Ostfeld. 2008. Multiobjective contaminant sensor network design for water distribution systems. *J. Water Resour. Plann. Manage.* 134, 4 (2008).

[35] K. Qian et al. 2020. Deep learning based anomaly detection in water distribution systems. In *ICNSC*. IEEE.

[36] S Rathi and R Gupta. 2014. Sensor placement methods for contamination detection in water distribution networks: A review. *Proc. Eng.* 89 (2014).

[37] J. Sage et al. 2017. Assessing the Effect of Uncertainties in Pollutant Wash-Off Dynamics in Stormwater Source-Control Systems Modeling: Consequences of Using an Inappropriate Error Model. *J. Environ. Eng.* 143, 2 (2017).

[38] R. Schwartz et al. 2014. Integrated hydraulic and organophosphate pesticide injection simulations for enhancing event detection in water distribution systems. *Water Res.* 63 (2014).

[39] R. Schwartz et al. 2014. Optimal sensor placement in water distribution systems for injection of chlorpyrifos. In *EWRI Congress*.

[40] A. Semadeni-Davies et al. 2008. The impacts of climate change and urbanisation on drainage in Helsingborg, Sweden: Suburban stormwater. *Journal of hydrology* 350, 1-2 (2008).

[41] E.Q. Shahra and W. Wu. 2020. Water contaminants detection using sensor placement approach in smart water networks. *JAIHC* (2020).

[42] C. Shannon. 1948. A mathematical theory of communication. *BSTJ* 27, 3 (1948).

[43] B. Shi et al. 2021. A low-cost water depth and electrical conductivity sensor for detecting inputs into urban stormwater networks. *Sensors* 21, 9 (2021).

[44] L. Skinner et al. 1999. Developmental effects of urban storm water in medaka (Oryzias latipes) and inland silverside (Menidia beryllina). *AECT* 37 (1999).

[45] Robert W Smith. 2002. The use of random-model tolerance intervals in environmental monitoring and regulation. *JABES* 7 (2002).

[46] P. Venkateswaran et al. 2018. Impact driven sensor placement for leak detection in community water networks. In *ICCPS*. IEEE.

[47] M Wang et al. .. An Innovative Low-cost Turbidity Sensor for Long-term Turbidity Monitoring in the Urban Water System. (.).

[48] J.P. Watson et al. 2009. Formulation and optimization of robust sensor placement problems for drinking water contamination warning systems. *J. Infrastruct. Syst.* 15, 4 (2009).