Physics-based Pollutant Source Identification in Stormwater Systems

Andrew Chio[1], Russell Bent[2], Andrey Y. Lokhov[2], Jian Peng[3], Nalini Venkatasubramanian[1]

European Control Conference 2024 - Stockholm, Sweden

June 26, 2024

[1] *Department of Computer Science, University of California, Irvine, California, US*

[2] *Theoretical Division, Los Alamos National Laboratory, US*

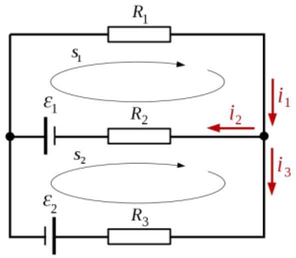
[3] *Orange County Public Works, California, US*

Infrastructure Network Design and Operation

Physics plays a big role in the design and operation of infrastructure systems

Electric Grid

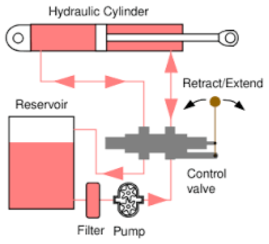
Driven by physics



Power Balance

Water Network

Driven by physics



Hydraulics

Management and Control

- Load balancing
- Fault Detection and Isolation
- Vulnerability Analysis
- Load shedding
- State estimation
- …

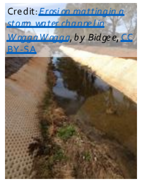
Driving Example: Stormwater Infrastructure Networks

Cities and Communities

Stormwater Network

Catch Basin

Network / Channels



Outfalls

Rivers, Bays, Ocean

Nuisance Flows

Rainfall *Excess Irrigation*

Unwanted Pollutants and Contaminants

Oils and Greases Illegal dumping

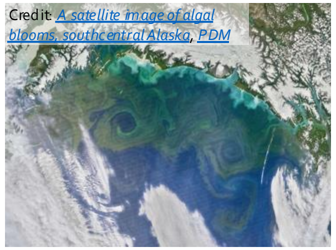
Pesticides *Industrial Chemicals*

Unpermitted wastewater

Water Quality Impairments

Fish Kills

Algal Blooms



Addressing Unwanted Pollutants: The Role of Physics

Detention and Retention

- *Temporarily store stormwater in storage basins*
- *Release slowly for flood control, peak flow reduction, erosion control*
- *<u>Control based on fluid dynamics and hydrology</u>*

Credit: *Trounce Pond*, by *Drm310*, *CC BY-SA*

Overflow Controls

- *Detect potential flooding conditions and impact on other areas of network*
- *<u>Dynamics based on gravity and network structure</u>*

Credit: *Close-up side view of Tainter gate*, by *Yangj28*, *CC BY-SA*

Green Infrastructure

- *Permeable pavements, rain gardens, etc. for managing stormwater to optimize water retention and pollutant removal*
- *<u>Design based on hydrological and soil physics</u>*

Credit: *Rain Garden Installation*, *PDM*

Dry Weather Flows (DWFs)

Illicit discharges into stormwater networks during dry weather / low flow conditions with no rain

Source: City of High Point Stormwater Division

Used wash water,
High Point, NC

Source: MRWC

Petroleum, solvents, oil,
Millers River Watershed Council

Source: Grand Prairie Texas Public Health and Environmental Quality

Diesel spill,
Grand Prairie, TX

Source: PWCVA

Illicit discharge,
Prince William County, VA

OC Public Works

Nature of Dry Weather Flows

- *Transient and Spontaneous*
  - *E.g., illegal dumping into a catch basin*
- *Varying pollutant loads*
- *Driven by physics of flow propagation*

On Monitoring Dry Weather Flows (DWFs)

Current techniques are inadequate

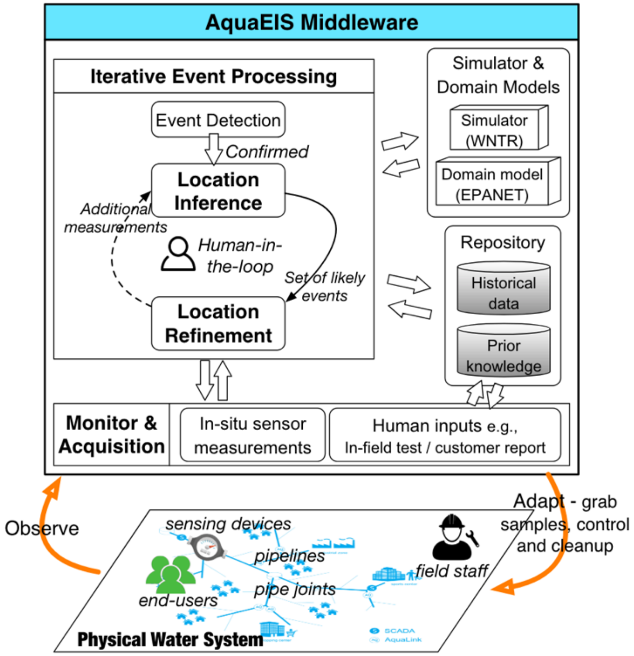
- Manual inspections, citizen reports, site visits in large, regional catchment areas
- Testing kits and laboratory analysis require 3-5 weeks for processing

OC Public Works

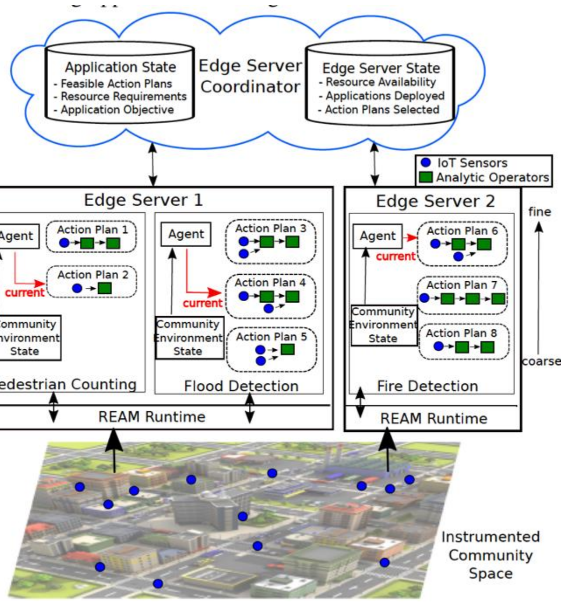
The Promise of the Internet-of-Things

- *New real-time monitoring capabilities*
- *Rapid detection and management of events*

AquaEIS[1]



REAM[2]



[1] Han et al., *AquaEIS: Middleware Support for Event Identification in Community Water Infrastructures. ACM Middleware 2019.*

[2] Venkateswaran et al., *REAM, A Framework for Resource Efficient Adaptive Monitoring of Community Spaces. PMC 2021.*

The DWF Source Identification Problem

Input:
Stormwater Network Sensor Observations from predeployed sensors[1]

→ DWF Source Identification →

Output:
Infer pollution sources and amounts, and their evolution over time

[1] Chio et al., *STEP: Semantics-aware Sensor Placement for Monitoring Community-Scale Infrastructure. ACM BuildSys '23*

On Effective Source Identification

Input:
Stormwater Network Sensor Observations from predeployed sensors

DWF Source Identification

Output:
Infer pollution sources and amounts, and their evolution over time

Traditional Method[1]
- Physical observations, sampling post hoc lab analysis
- Very low probability of success

Optimization-based Approaches[3]
- Greedy heuristics and evolutionary algorithms
- Faster, but sub-optimal

Bayesian Approach[2]
- Pollutant origin treated as random variable and likelihood updated using sensor data
- Expensive to run

ML/DL-based Approaches[4]
- Training models to identify source nodes through observed data
- Require large amounts of data
- May need heavy tuning + computation

[1] Bernstein et al, Environment Monitor Assess '09, Li et al., Environmental Pollution '23
[2] Snodgrass et al, Water Resource Research '97, Zeng et al, Advances in Water Resources '12
[3] Banik et al, Water '17, Han et al, J. Hydrology '20,
[4] Grbcic et al, J. Hydroinform '20, Mo et al, Water Resources Research '19

The Role of Physics in DWF Source Identification

Input:
Stormwater Network Sensor Observations from predeployed sensors

→ DWF Source Identification →

Output:
Infer pollution sources and amounts, and their evolution over time

Black Box Approach

- Leverages pre-defined set of inputs to cache
- Source identification searches through run results of prior physics-based simulations
- Quality varies depending on cached values
- Can require massive compute and memory resources

White Box Approach

- Modeling and solving for embedded physical equations in simulations are studied directly
- Exploit underlying computational model to examine effects of specific inputs
- Requires deep knowledge of the domain and underlying physics + computation model

Our Approach: Design of a Backwards Inference Model

*Construct an **efficient** and **effective** physics-based DWF **backwards inference model** by deriving a **close approximation of the physics** that govern **stormwater flow dynamics***

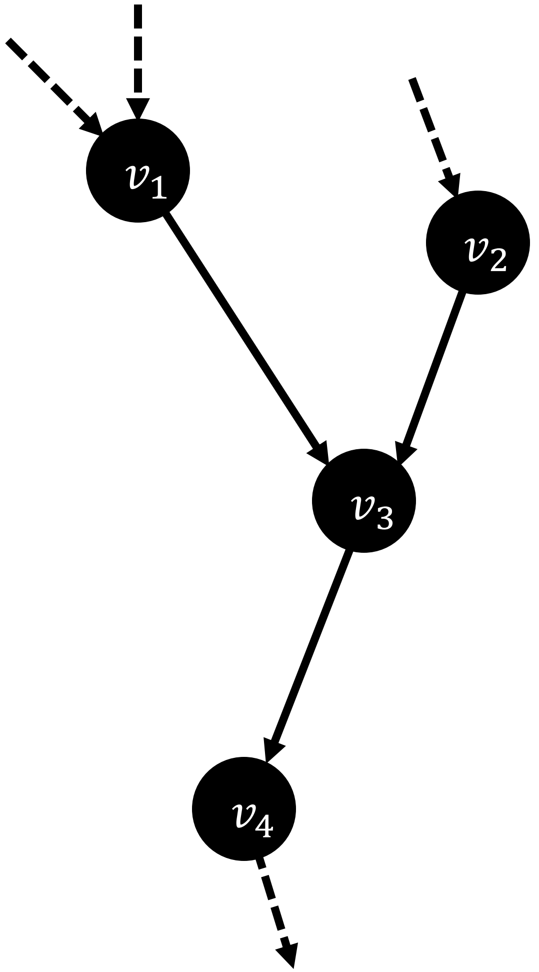
1. **Where?**

   Given sensor observations, where could a DWF anomaly originate from?

2. **How much?**

   How much flow should be expected at a suspected origin location?

Formulation: Stormwater Infrastructure Graph



- Directed **Graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- **Nodes** $v_j \in \mathcal{V}$ (Junctions)
  - Location $\left(x_j, y_j\right)$
  - Invert elevation z_j
- **Edges** $e_{ij} \in \mathcal{E}$ (Conduits)
  - Length L_{ij}
  - Frictional roughness f_{ij}
  - Shape S_{ij}
  - *Slope* m_{ij} *(derived)*

Flow Propagation within Stormwater Networks

Principles of Flow Propagation

- *Conservation of Mass*:

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

- *Conservation of Momentum*:

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

EPA SWMM[1] using Dynamic Wave Analysis

$$Q^{t+\Delta t} = \frac{Q^t + \Delta Q_{iner} + \Delta Q_{pres}}{1 + \Delta Q_{fric}}$$

$$\Delta Q_{iner} = 2\bar{U}\left(\bar{A}^{t+\Delta t} - \bar{A}^t\right) + \bar{U}^2 \frac{(A_{dn} - A_{up})}{L}\Delta t$$

$$\Delta Q_{pres} = -g\bar{A}\frac{(H_{dn} - H_{up})}{L}\Delta t$$

$$\Delta Q_{fric} = g\eta^2 \frac{|\bar{A}|\Delta t}{\bar{R}^{4/3}}$$

$$H^{t+\Delta t} = H^t + \frac{\Delta t}{2}\frac{\left(\sum Q^t + \sum Q^{t+\Delta t}\right)}{\left(A_{SN} + \sum A_{SL}\right)^{t+\Delta t}}$$

Difficult for inference: *Implementation of Dynamic Wave Analysis introduces non-differentiable equations*

$\boldsymbol{Q}$ = Flow Rate
$\boldsymbol{H}$ = Hydraulic Head

[1] EPA SWMM. https://www.epa.gov/water-research/storm-water-management-model-swmm

Approximations to Flow Propagation for Differentiability (1)

1. Dry Boundary Conditions
2. Critical and surcharged flows near maximum capacities
3. Backwards Flow Support
4. Measuring closeness to criticality in conduits

Intuition for Approximation:

Remove dry boundary conditions; always update $Q^{t+\Delta t}$ regardless of current cross-sectional flow area

$$Q^{t+\Delta t} = \begin{cases} \dfrac{Q^t + \Delta \mathrm{Q}_{\mathrm{iner}} + \Delta \mathrm{Q}_{\mathrm{pres}}}{1 + \Delta \mathrm{Q}_{\mathrm{fric}}} & if\ \bar{A} \geq \epsilon_A \\ Q^t & else \end{cases}$$

Approximations to Flow Propagation for Differentiability (2)

1. Dry Boundary Conditions
2. Critical and surcharged flows near maximum capacities
3. Backwards Flow Support
4. Measuring closeness to criticality in conduits

Intuition for Approximation:

Since dry weather flows are (by definition) occur in dry weather / low flow conditions, conduits will not reach critical or surcharged states

$Q^{t+\Delta t} = \min\{Q^{t+\Delta t}, Q_{norm}\}$ where

$$Q_{norm} = \frac{1.49}{n} A_{up} R_{up}^{2/3} \sqrt{L^2 - \left(H_{up} - H_{down}\right)^2}$$

Approximations to Flow Propagation for Differentiability (3)

1. Dry Boundary Conditions
2. Critical and surcharged flows near maximum capacities
3. Backwards Flow Support
4. Measuring closeness to criticality in conduits

Intuition for Approximation:

Backwards flow occurs when rate of flow introduction is critical; this does not occur in dry weather conditions

$$Q^{t+\Delta t} = \begin{cases} Q^{t+\Delta t} & if\ Q^{t} \cdot Q^{t+\Delta t} > 0 \\ 0.0001 \times sign\left(Q^{t+\Delta t}\right) & else \end{cases}$$

Approximations to Flow Propagation for Differentiability (4)

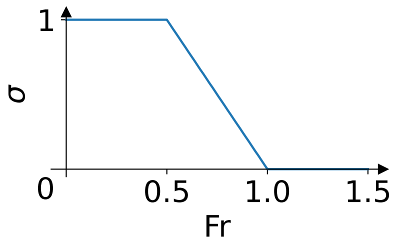
1. Dry Boundary Conditions
2. Critical and surcharged flows near maximum capacities
3. Backwards Flow Support
4. Measuring closeness to criticality in conduits

Intuition for Approximation:

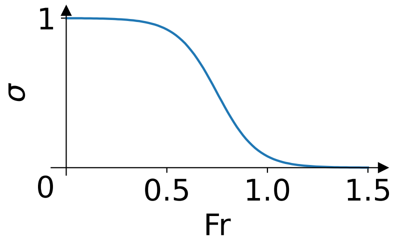
Differentiable approximation made for weight factors that are used for numerical stability of Dynamic Wave Analysis

$$\sigma = \begin{cases} 1 & if\, Fr \leq 0.5 \\ 2(1 - Fr) & if\ 0.5 < Fr \leq 1 \\ 0 & if\, Fr \geq 1 \end{cases}$$

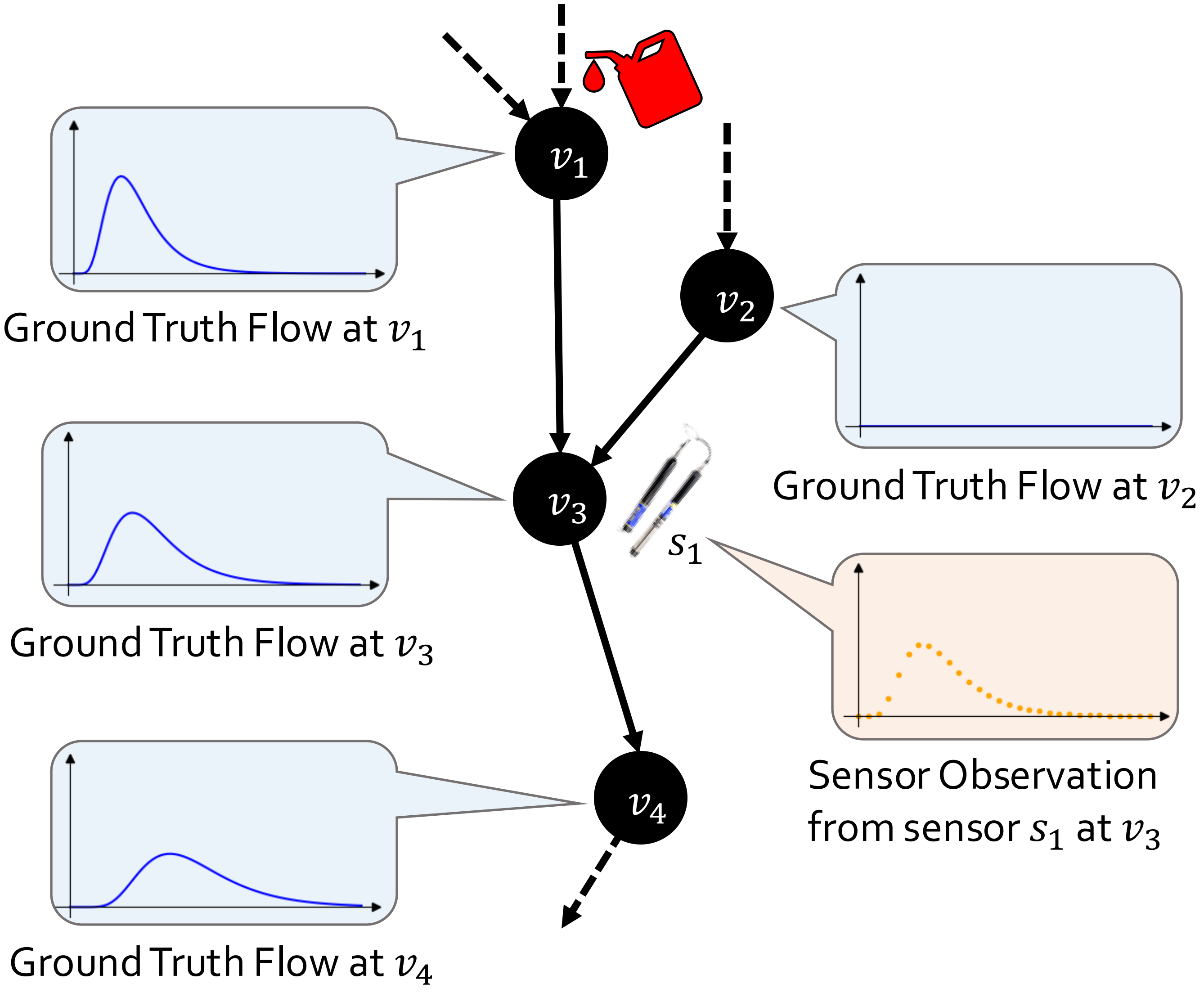
$$where\ Fr = |\bar{U}| / \sqrt{\left(\frac{g\bar{A}}{\bar{W}}\right)}$$



$$\sigma = \left(1 + \exp\left(10 * (Fr - 0.75)\right)\right)^{-1}$$



DWF Anomaly Flows and Sensor Observations



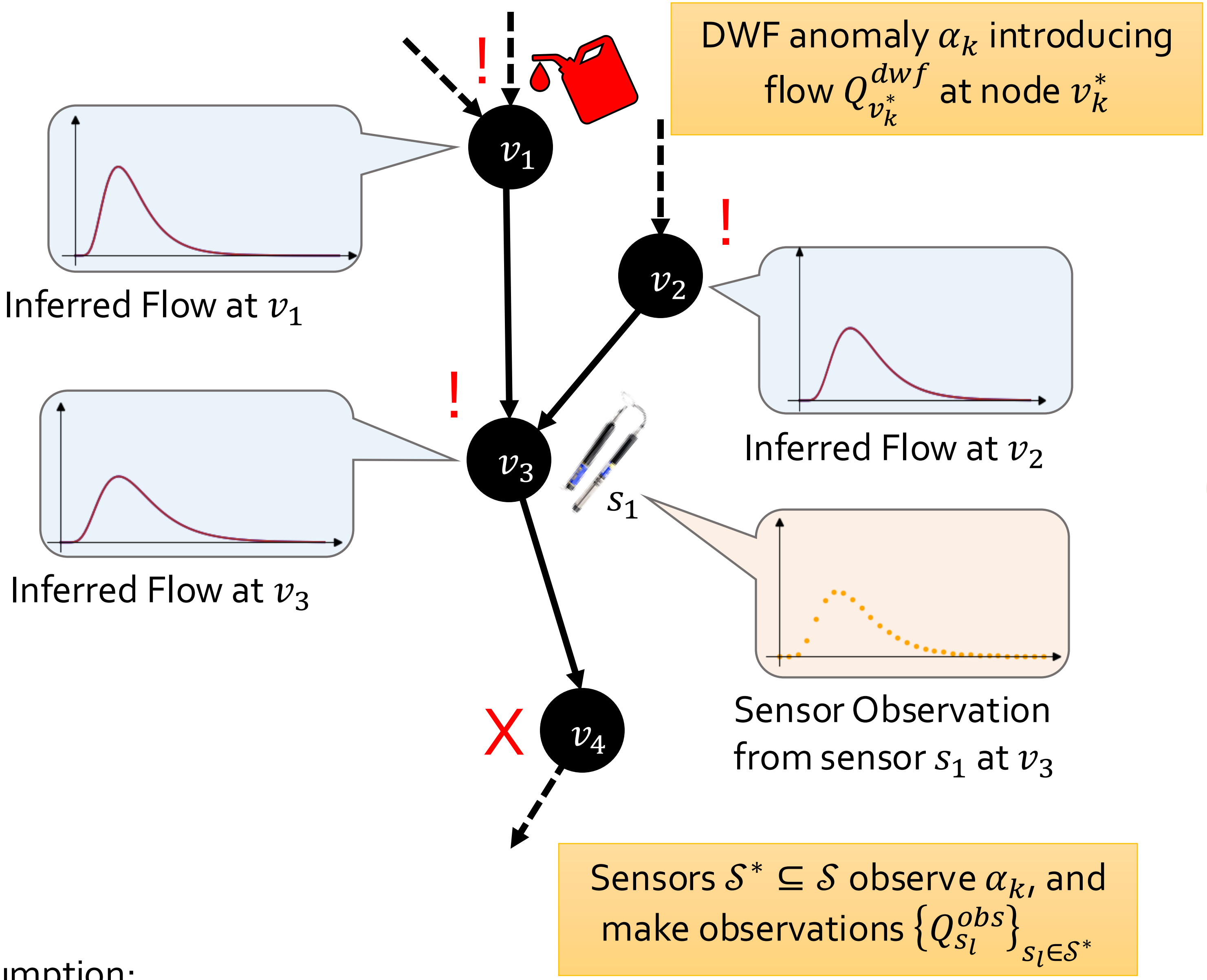
DWF Anomaly $\boldsymbol{\alpha_k} \in \boldsymbol{\mathcal{A}}$

- Origin node v_k^*
- DWF inflow curve $Q_{v_k^*}^{dwf}(t)$

Sensor $\boldsymbol{s_l} \in \mathcal{S}$

- Periodicity of measurement λ_l sec
- Flow observations $\left\{Q_{s_l}^{obs}(t)\right\}_{t \in T^*}$

Problem Statement: DWF Source Identification



Goal:

Infer the flow $\boldsymbol{Q^{inf}_{v^*}}$ to introduce at v^* that would most likely produce $\{\boldsymbol{Q^{obs}_{s_l}}\}_{\boldsymbol{s_l \in \mathcal{S}^*}}$ for the set of candidate nodes $\mathcal{V}^*$

Assumption:

Only one DWF occurs in the network at a time.

Our Approach: Design of a Backwards Inference Model

*Construct an **efficient** and **effective** physics-based DWF **backwards inference model** by deriving a **close approximation of the physics** that govern **stormwater flow dynamics***

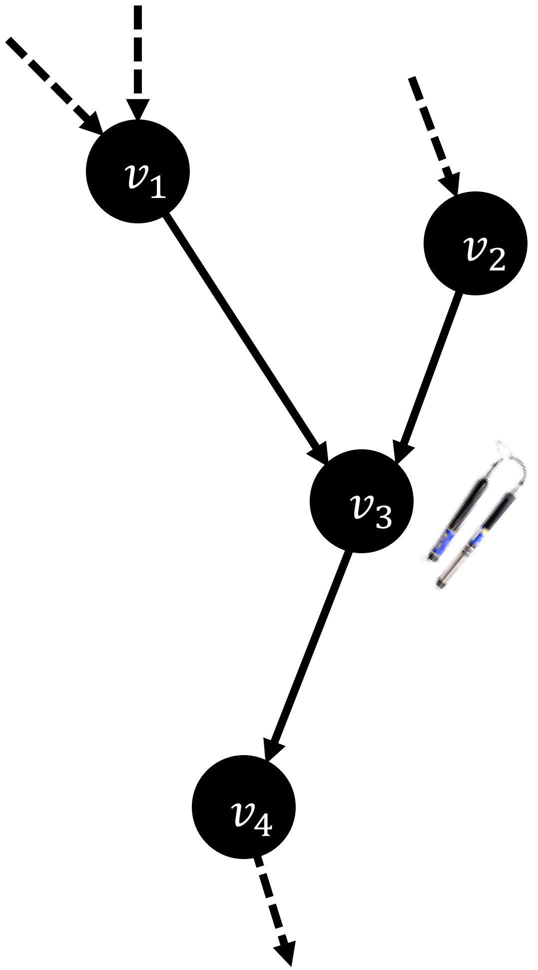
1 **Where?**

Given sensor observations, where could a DWF anomaly originate from?

2 **How much?**

How much flow should be expected at a suspected origin location?

Finding Initial Potential Sources for Anomalies

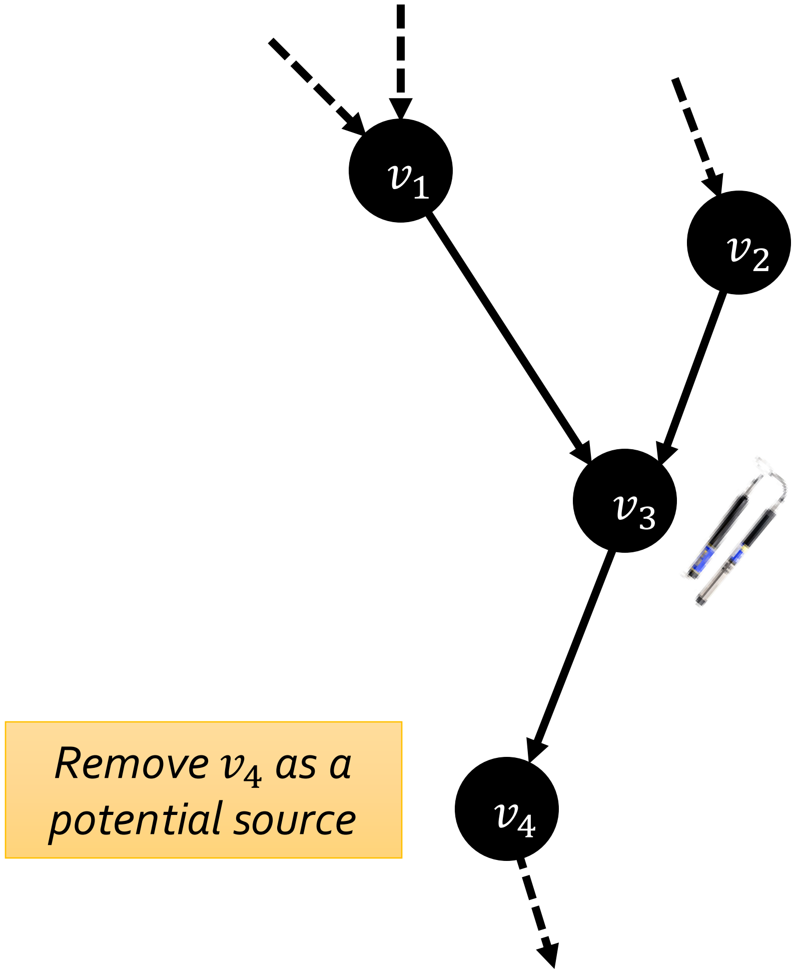


Solving for where: pruning set V^*

Potential Sources

$V^* = \{v_1, v_2, v_3, v_4\}$

Leveraging non-zero sensor observations



Solving for <u>where</u>: pruning set V^*

Intuition 1:

Non-zero sensor observations imply that DWF anomaly *must* lie *upstream*

$$\Downarrow$$

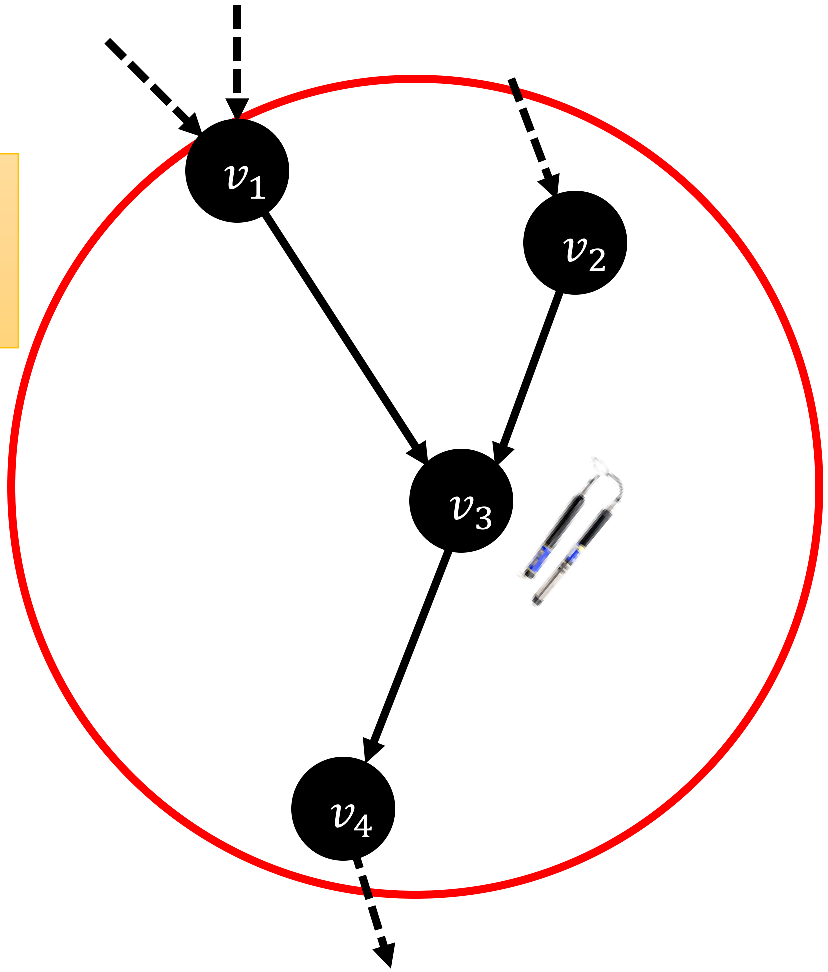
Constraint:

Eliminate nodes that do not lie upstream of the sensor's deployed node

Update: $V^* \leftarrow V^* \cap V^{up}$

Leveraging lack of sensor observations

Heuristically eliminate parents of v_1 and v_2



Solving for <u>where</u>: pruning set V^*

Intuition 2:

No sensor observations imply that *either:*

(1) the DWF anomaly does not lie upstream; or

(2) the DWF anomaly lies upstream, but is no longer detectable

$$\Downarrow$$

Constraint:

Eliminate nodes that lie within a threshold τ_d of the sensor's deployed node

Determine τ_d empirically using $[Q^{MIN}, Q^{MAX}]$

Update: $V^* \leftarrow V^* - V^{up}$

Our Approach: Design of a Backwards Inference Model

*Construct an **efficient** and **effective** physics-based DWF **backwards inference model** by deriving a **close approximation of the physics** that govern **stormwater flow dynamics***

1. **Where?**

   Given sensor observations, where could a DWF anomaly originate from?

2. **How much?**

   How much flow should be expected at a suspected origin location?

Formulating a Least Squares Regression

For each $\boldsymbol{v}^* \in \boldsymbol{V}^*$:

$$\underset{Q_{v^*}^{dwf}}{\arg\min} \sum_{s_l \in \mathcal{S}^*} \sum_{t \in T} \left(Q_{s_l}^{obs}(t) - Q_{s_l}^{simu}\left(t; Q_{v^*}^{dwf}\right) \right)^2$$

$$\text{s. t.}$$

$$Q^{MIN} \leq Q_{v^*}^{dwf}(t) \leq Q^{MAX} \qquad \forall t \in T$$

$$Q^{t+\Delta t} = \frac{Q^t + \Delta Q_{iner} + \Delta Q_{pres}}{1 + \Delta Q_{fric}} \qquad \forall t \in T$$

$$H^{t+\Delta t} = H^t + \frac{\Delta t}{2} \frac{\left(\sum Q^t + \sum Q^{t+\Delta t}\right)}{\left(A_{SN} + \sum A_{SL}\right)^{t+\Delta t}} \qquad \forall t \in T$$

Potential **DWF Flow Profile**

Sensor Observations at time t

Simulated value at sensor, given the DWF inflow $Q_{v^*}^{dwf}$

Constraints on the **min/max value** of $Q_{v^*}^{dwf}(t)$

Physics of flow propagation, *subject to the approximations made for differentiability*

Efficient optimization of regression using fast, non-linear solvers;
$Q_{s_l}^{simu}(\cdot)$ *provides expected flows resulting from* $Q_{v^*}^{dwf}(\cdot)$

The Overall Approach

Algorithm 1: Source Inference

Input: Graph $\mathcal{G}$, Sensors $\mathcal{S}$, Observations $\{Q^{obs}_{s_l}\}_{s_l \in \mathcal{S}}$, float τ_q, float τ_d, float τ_o

Output: Potential anomalies $\mathcal{A}^*$

`// Compute set of potential origin nodes`

$\mathcal{V}^* \leftarrow \emptyset$

for $s_l \leftarrow \mathcal{S}$ **do**

  $v \leftarrow GetDeployedNode(s_l)$

  $\mathcal{V}^{up} \leftarrow GetUpstreamNodes(v, \mathcal{G}, \tau_d)$

  if $\exists\, t : Q^{obs}_{s_l}(t) \geq \tau_q$ **then** $\mathcal{V}^* \leftarrow \mathcal{V}^* \cap \mathcal{V}^{up}$ **else** $\mathcal{V}^* \leftarrow \mathcal{V}^* - \mathcal{V}^{up}$;

`// Find flow curves for potential sources`

$M \leftarrow map()$

for $v^* \leftarrow \mathcal{V}^*$ **do**

  objval, Qinf $\leftarrow$ solve (3), assuming the source is v^*

  $M[src] \leftarrow (objval, Qinf)$

$minobjval \leftarrow \min_{objval,\_ \in M.values()} \{objval\}$

$\mathcal{A}^* \leftarrow \{(src, Qinf) : \forall (src, (objval, Qinf)) \in M : |minobjval - objval| \leq \tau_o\}$

return $\mathcal{A}^*$

Intuition

Phase 1: ***Where?***

Iterate through each deployed sensor

Eliminate improbable sources of the DWF

Phase 2: ***How much?***

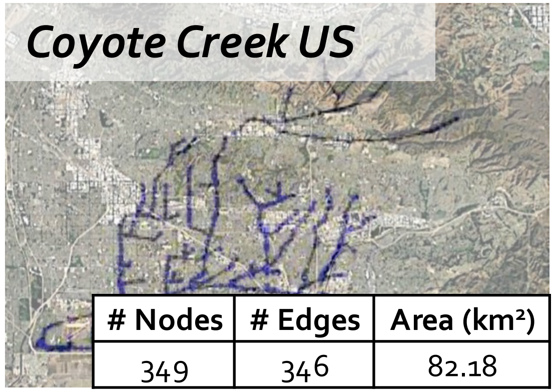
Model and solve the least squares regression as defined previously; cache results

Return set of likely potential sources within threshold τ_o

Experiments: 6 Real-World Stormwater Networks

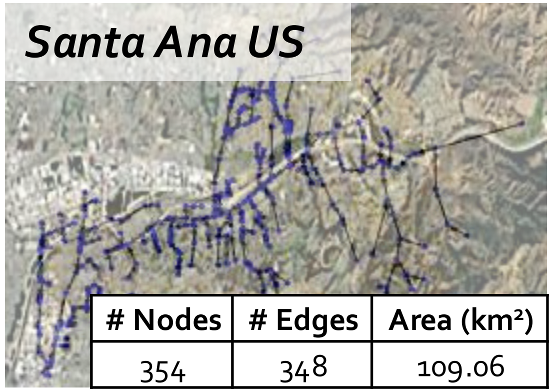
Real-World Networks

- 6 EPA SWMM[1] networks of stormwater systems in Southern California, USA of varying sizes
- *Provided by Orange County Public Works (OCPW)*



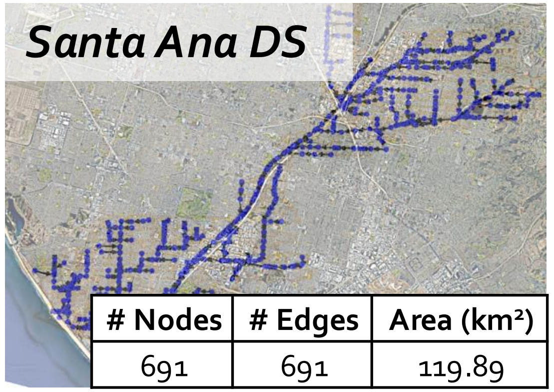
Coyote Creek US

# Nodes	# Edges	Area (km²)
349	346	82.18



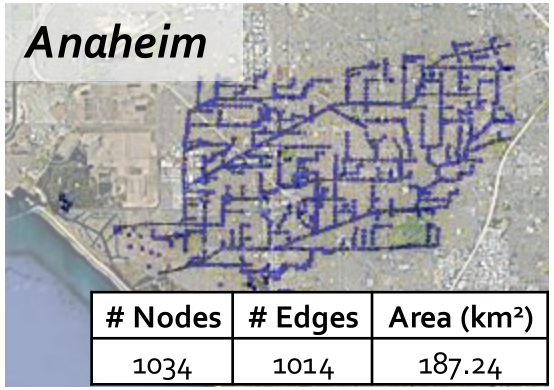
Santa Ana US

# Nodes	# Edges	Area (km²)
354	348	109.06



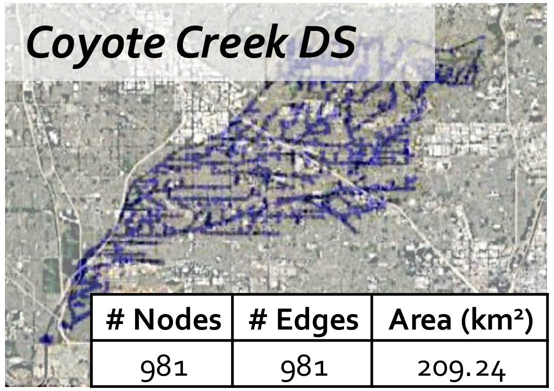
Santa Ana DS

# Nodes	# Edges	Area (km²)
691	691	119.89



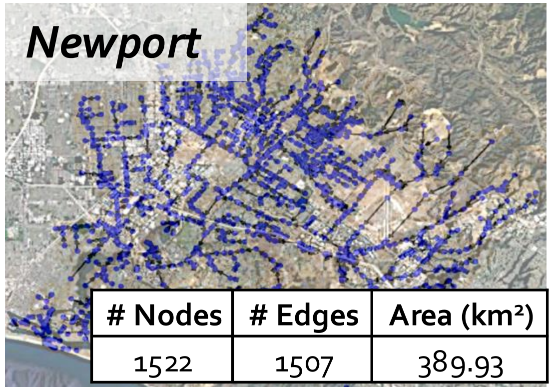
Anaheim

# Nodes	# Edges	Area (km²)
1034	1014	187.24



Coyote Creek DS

# Nodes	# Edges	Area (km²)
981	981	209.24



Newport

# Nodes	# Edges	Area (km²)
1522	1507	389.93

OC Public Works

[1] EPA SWMM. https://www.epa.gov/water-research/storm-water-management-model-swmm

Experimental Setup Details

Sensors

- Considered homogenous flow sensors that generate observations with periodicity $\lambda = 30$ sec
- *Assumption: Sensors are pre-deployed in network at varying levels of instrumentation (10%, 25%, 50%, 75%, 100%)*

DWF Anomalies

- 100 anomalies constructed randomly for each network:
    - Origin chosen randomly
    - Inflow curve with max. magnitude $|0.25 \pm 0.2|$ cfs
    - Start/end times range between 0 and 2 hours

Comparison Baseline

- Simulated and cached 10 anomalies uniformly across all junctions of each network
- An anomaly is "inferred" by searching for the best match to a given set of sensor observations

Implementation Details

- *M1 MacBook Pro (16 GB memory, 10 CPU cores)*
- Implemented in Julia using *JuMP interface* to *Ipopt*, and *MA57 solver* for optimization

Experiment 1: Impact of Approximations

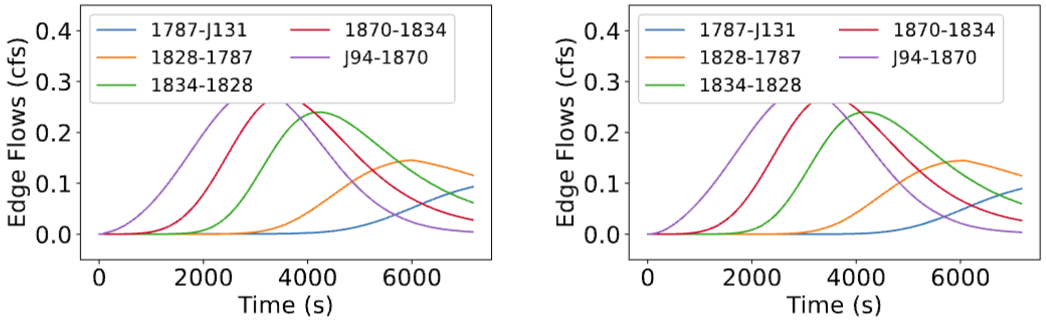
- Examine impact of the approximations made on the accuracy of modeling and solving for flow dynamics
- The average MSE between EPA SWMM and our differentiable version across all anomalies is ***negligible***.
- **Bottom figure** illustrates an example of a typical anomaly simulated using EPA SWMM, compared against our differentiable version

The approximations made to the physics of flow propagation are negligible – all MSE values are very small.

Impact of Approximations

Network	Avg MSE
Anaheim	2e-3 ± 2e-2
Coyote Creek Downstream	9e-6 ± 1e-5
Coyote Creek Upstream	3e-4 ± 2e-3
Newport	3e-6 ± 4e-6
Santa Ana Downstream	1e-2 ± 8e-4
Santa Ana Upstream	1e-5 ± 1e-5

Comparison of edge flows from a typical anomaly



(a) EPA SWMM Edge Flows (b) Differentiable Version

Experiment 2: Accuracy of the Backwards Inference Model

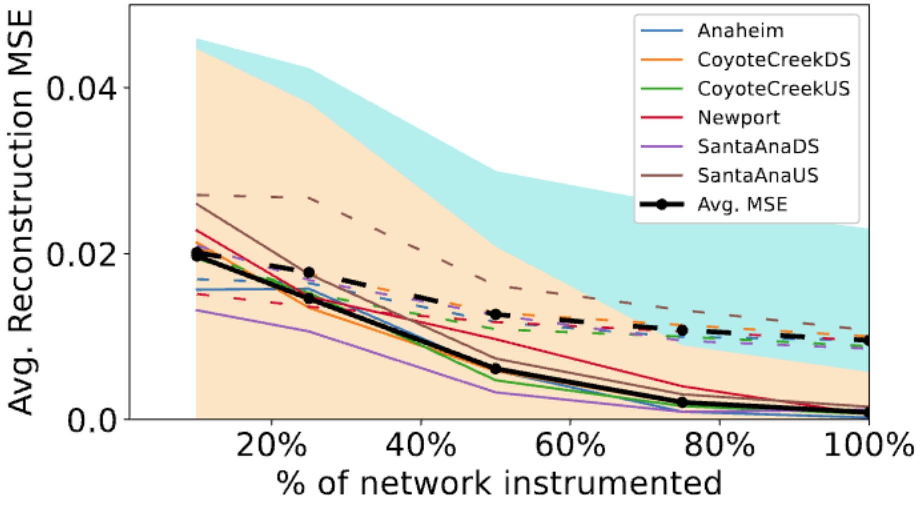
- The **average MSE** for anomalies reconstructed at the correct origin drops for all networks:
  - **10%** instrumentation yields ~0.02 MSE
  - **100%** instrumentation yields ~0.0 MSE

- Inferred flow for a typical anomaly shows very small differences for reconstructed anomalies.

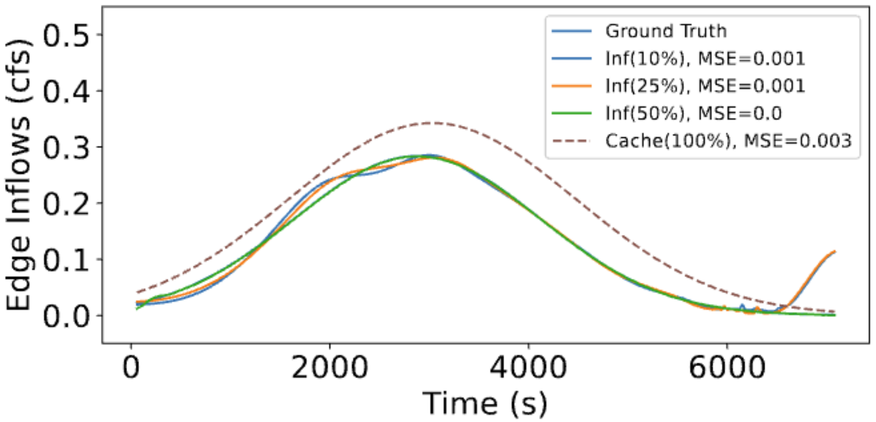
- Results from cache-based comparison baseline depend on quality of cache

Our model accurately infers flows for the anomalies evaluated, across different levels of instrumentation.

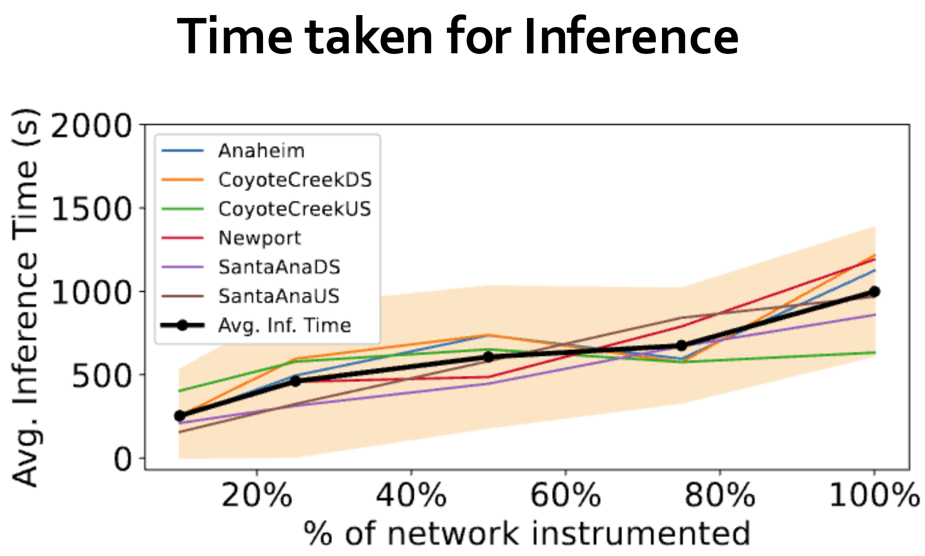
MSE of Backwards Inference Model



Example Inferred Flow for a Typical Anomaly



Experiment 3: Time taken by the Backwards Inference Model



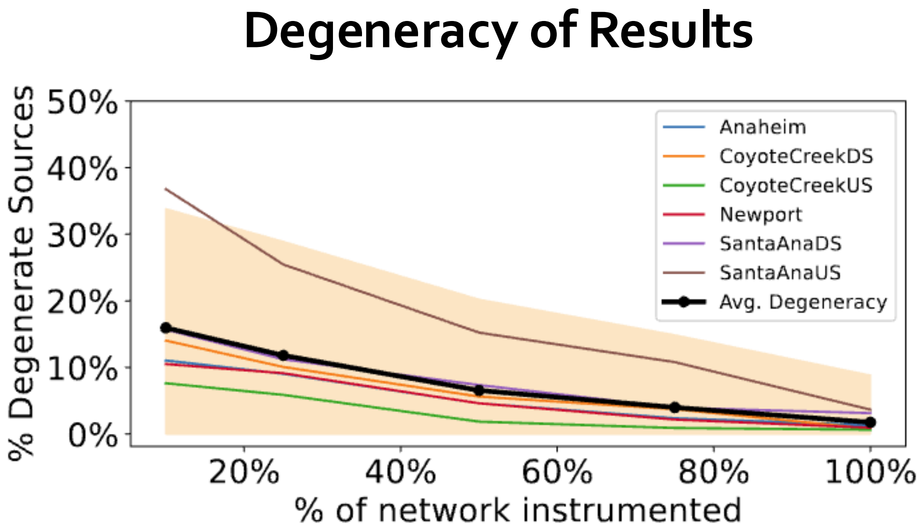
- The **average time taken** for a reconstruction result **increases** with higher level of instrumentation for all networks:
  - **10%** instrumentation: **~250 seconds**
  - **100%** instrumentation: **800-1300 seconds**

- Standard baseline caching approach too significantly less time, but **consumed ~14.3GB** of memory, generated **over ~3 days**

Our model infers flows in a timely manner, which is essential for supporting real-time control and management.

Experiment 4: Degeneracy of Results

- The **number of other equally-likely potential sources** generally **decreases** with higher levels of instrumentation for all networks:
  - **10%** instrumentation: **18%** degenerate sources
  - **100%** instrumentation: **~0%** degenerate sources



Our model helps eliminate improbable nodes as sources as captured data increases, and degree of uncertainty decreases.

Key Takeaways and Future Directions

- We presented a **physics-based backwards inference model** for stormwater networks
- **Several approximations of the physics** driving flow propagation were applied to allow compatibility with fast, non-linear solvers
- **Six real-world stormwater networks** were used for evaluation, showing the accuracy and timeliness of our backwards inference model over a standard black box caching approach
- Our code is publicly available on GitHub:
  - https://github.com/andrewgchio/SWMMBackwardsInference
- ***Ongoing and Future Work:***
  - Explainability of DWF anomalies
  - Multiple & concurrent potential contaminants at varying points of the network
  - Study the generalization of this technique to other real-world infrastructures

GitHub Link