Statistical Methods for Water Distribution Networks

Estéban Nocet-Binois

October 2024

Github page: CEE599B-Project-2024

- See docs for proposal and current progress.
- See src for source code (containing graph in R, methods, and eventually analysis).
- See data for original `NJ1.inp` file, as well as my data pipeline and the resulting `.csv` files.

1 Research Purpose

The goal is to explore statistical methods on graphs such as model selection and graph parameter estimation, uncertainty propagation, correlation and clustering in graphs, going beyond simple measures of graph topology (number of edges, degree of nodes, average path length, centrality, etc).

Most advances in this type of work focus on biological and brain networks, for tasks such as identification and correlation of functional sub-networks, or network classification (eg, sound or diseased) – see references in Section 3.1. The goal of this project is to explore the applicability of these tools to larger-scale, spatial networks such as water distribution systems (WDS). I have three points of interest:

- Hydraulic modeling in WDS that integrates real-world data and methods from graph theory (application to network reconstruction, model evaluation, and the estimation of error and bias). This will probably be the actual focus of this project, mainly revolving around points 1. and 4. of the Methodology section (below). However, depending on the course this project takes, I may also be interested in the following questions.
- Network design and expansion: tools from graph limits can model how the network might evolve with extension. Precedent point is also important for this task.
- One other interesting research question could also be that of subgraphs and boundary effects: while zonal / disctrict-based governance is a common approach for managing large scale distribution systems, it is crucial to estimate the information loss when sampling subgraphs (or graphets) from larger network (see Rheinwalt et al., 2012).

See references in Section 4.2.

Note: There is one major limitation to this work, namely that the analysis is mainly structural. Although there is a notion of temporality in the demand data (and induced pressure data) for the current dataset (see next section), the range (80 hours) is just enough to have information on the flow over the network. The reason for it is that the analysis focuses on stability and perturbation theory, with respect to the flow over the network, aiming at sensitivity quantification (as ordinal ranking rather than prediction) and decomposition tasks. However, if we want to explore, and predict, the actual behavior of the system under disruption or failure (leakages, pipe breaks, floods or extreme events), the data we have is just insufficient. The second problem is that there are few such datasets: we have the LeakDB generated dataset (of abrupt vs. incipient leaks), but no real-world examples and no data under extreme events.

These articles from researchers at the Università di Firenze Arrighi et al. (2017), Tarani et al. (2019) and Arrighi et al. (2021) can be interesting with that respect, however they also seem to perform quasi-static sensitivity analysis. One option could be to find a dataset of water precipitation in the same region as the distribution network, and estimate the network's response to potential flooding (by setting a threshold water level, past which disruption would be assumed). Yet the results can not be validated without real data on the state of the network (or at least on supply shortfalls at demand nodes).

2 Dataset

We need a dataset that contains:

- Geometry – *static*
  - Three types of nodes: demand nodes (hydrants), source nodes (reservoirs and tanks) and junctions (pipe connections)
  - Two types of edges: distribution pipes (mains and branches) and pumps (where pressure and flow can be controlled)
- Flow-volume through pipes – *temporal*
- Pressure at junctions – *temporal*
- Demand data – *temporal*
- Topography: elevation and spatial coordinates are optional, distances are needed
- No data on water quality is needed (hydrolic analysis only)

For this purpose the following water distribution systems (see here and here) can be used: Anytown, Balerma (447 nodes), FossPoly 1, Hanoi from LeakDB (32 nodes), Modena, ZJ (115 nodes), C-Town (396 nodes), D-Town (406 nodes) and L-Town (782 nodes), and the Kentucky dataset (in particular Ky1, Ky6, Ky8, and Ky10).

For our study, we will be considering the Dover Township (NJ) distribution system 05 NJ 1 that has been used in three other articles, by Maslia et al. (2000), Hoagland et al. (2015) and Hwang & Lansey (2017). It contains eight tanks, 12 pumps, and 483 miles of pipe. For more details on the dataset, checkout the research log.

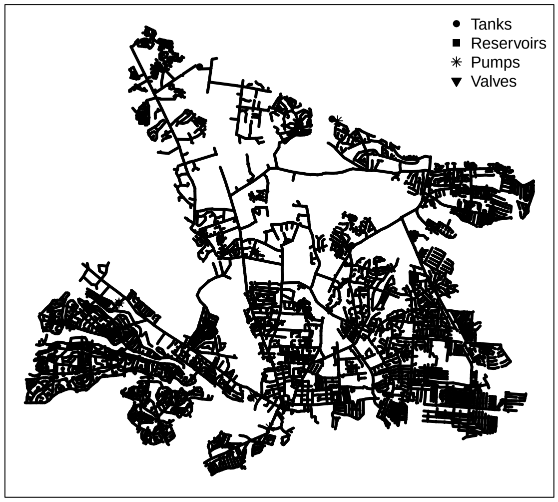

Figure 1: NJ1 network based on the Dover Township, NJ distribution system

References:

- Tello et al. (2024): "Large-Scale Multipurpose Benchmark Datasets for Assessing Data-Driven Deep Learning Approaches for Water Distribution Networks"
- Jolly et al. (2014): "Research Database of Water Distribution System Models"
- Adraoui et al. (2024): "Towards an Understanding of Hydraulic Sensitivity: Graph Theory Contributions to Water Distribution Analysis" – see Section 5.1

3 Methodology

3.1 Model Estimation

Motivation: Just like distribution fitting, graph fitting involves identifying the best model for a set of observations, where those observations are network connections rather than time series, hence modeling large-scale interdependencies as realizations of random processes on graphs. The estimated model can be used for scenario generation, deviation estimation, and spectral theory in general can help gain insights into the network's structural properties (e.g. finding bottlenecks).

See work from André Fujita and Daniel Yasumasa Takahashi, in particular article 4. (2020) that provides the algorithms.

Analyze the graph structure variability: given an adjacency matrix A, we derive

- the spectrum of G, that is, the set of eigenvalues $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n)$
- the spectral distribution of G
- the graph spectral entropy

Use Kullback-Leibler divergence to compare the performance between various graph models, such as the Erdos-Renyi random graph, Geometric random graph, K-regular random graph, Watts-Strogatz model, Barabási-Albert model, etc. For example, the Erdős-Rényi model can be estimated with the link density of the network only, while the Watts-Strogatz model can be estimated with the network's average degree, clustering coefficient and average path length.

Note: In brain networks, the spatial positioning of neurons or brain regions is not usually considered, whereas in water distribution, spatial layout is central to how the network functions. Hence, water networks will benefit from geometric / spatial models that incorporate distance. Also, the larger the network the most applicable this approach is.

References:

1. Fujita et al. (2017): "A Statistical Method to Distinguish Functional Brain Networks" – focus on autism and Asperger
2. Fujita et al. (2017): "Correlation between graphs with an application to brain network analysis" – focus on autism spectrum disorder
3. Takahashi et al. (2012): "Discriminating Different Classes of Biological Networks by Analyzing the Graphs Spectra Distribution" – foci on protein-protein interactions and ADHD
4. Fujita et al. (2020): "A semi-parametric statistical test to compare complex networks" – focus on PPI networks of enteric pathogens

3.2 Spectral Clustering

Motivation: Spectral clustering can help forming zones in the WDS for efficient management and control.

References:

1. Nascimento & de Carvalho (2011): "Spectral methods for graph clustering – A survey"
2. von Luxburg (2007): "A tutorial on spectral clustering"
3. Spielman (2007): "Spectral Graph Theory and its Applications"
4. Sarkar & Jalan (2006): "Spectral properties of complex networks"

See also these papers with code.

3.3 Graph Fourier Transform

Motivation: GFT can help reduce noise in sensor measurements and monitor the WDS with minimal data.

References:

1. Schultz et al. (2020): "Graph Signal Processing for Infrastructure Resilience: Suitability and Future Directions"

2. Wei et al. (2020): “Optimal Sampling of Water Distribution Network Dynamics Using Graph Fourier Transform”
3. Wei et al. (2019): “Monitoring Networked Infrastructure with Minimum Data via Sequential Graph Fourier Transforms”
4. Pagani et al. (2021): “Neural Network Approximation of Graph Fourier Transform for Sparse Sampling of Networked Dynamics”

3.4 Uncertainty Propagation

References:

1. Rebrova & Salanevich (2024): “On Graph Uncertainty Principle and Eigenvector Delocalization”, with code available
2. Arola-Fernandez et al. (2020): “Uncertainty propagation in complex networks: from noisy links to critical properties”
3. Ji et al (2023): “Signal propagation in complex networks”

3.5 GNN for State Estimation

References:

1. Pagani et al. (2021): “Neural Network Approximation of Graph Fourier Transform for Sparse Sampling of Networked Dynamics”
2. Li et al. (2024): “Real-time water quality prediction in water distribution networks using graph neural networks with sparse monitoring data”
3. Truong et al. (2024): “Graph Neural Networks for State Estimation in Water Distribution Systems”

4 Bibliography

4.1 General books

- van Mieghem (2011): *Graph Spectra for Complex Networks*
- Kolaczyk (2017): *Topics at the Frontier of Statistics and Network Analysis*
- Kolaczyk (2009): *Statistical Analysis of Network Data: Methods and Models*
- Crane (2018): *Probabilistic Foundations of Statistical Network Analysis*
- Bagrow & Ahn (2024): *Working with Network Data: A Data Science Perspective*
- Avrachenkov & Dreveton (2022): *Statistical Analysis of Networks*
- Kalyagin et al. (2020): *Statistical Analysis of Graph Structures in Random Variable Networks*
- See also Chapter 3 of “Network Science: An Aerial View”

4.2 Other articles on graphs and WDS

1. Adraoui et al. (2023): “Towards an Understanding of Hydraulic Sensitivity: Graph Theory Contributions to Water Distribution Analysis”
2. Yu et al. (2024): “A review of graph and complex network theory in water distribution networks: Mathematical foundation, application and prospects”

3. Herrera et al. (2016): "A Graph-Theoretic Framework for Assessing the Resilience of Sectorised Water Distribution Networks"
4. Zingali et al. (2024): "Application of Primary Network Analysis in Real Water Distribution Systems"
5. Price & Ostfeld (2014): "Optimal Water System Operation Using Graph Theory Algorithms"
6. Fragkos et al. (2021): "Decomposition methods for large-scale network expansion problems"
7. Anchieta et al. (2023): "Water distribution network expansion: an evaluation from the perspective of complex networks and hydraulic criteria"

5 Network Analysis in R

Checkout:

- https://ladal.edu.au/net.html
- https://bookdown.org/jdholster1/idsr/network-analysis.html
- https://cran.r-project.org/web/packages/statGraph/