Applied Network Science

RESEARCH

Open Access

Probabilistic measures of edge criticality in graphs: a study in water distribution networks

Andrea Ponti[1*], Antonio Candelieri[2], Ilaria Giordani[3] and Francesco Archetti[1]

*Correspondence:
a.ponti5@campus.unimib.it
[1] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy
Full list of author information is available at the end of the article

Abstract

The issue of vulnerability and robustness in networks have been addressed by several methods. The goal is to identify which are the critical components (i.e., nodes/edges) whose failure impairs the functioning of the network and how much this impacts the ensuing increase in vulnerability. In this paper we consider the drop in the network robustness as measured by the increase in vulnerability of the perturbed network and compare it with the original one. Traditional robustness metrics are based on centrality measures, the loss of efficiency and spectral analysis. The approach proposed in this paper sees the graph as a set of probability distributions and computes, specifically the probability distribution of its node to node distances and computes an index of vulnerability through the distance between the node-to-node distributions associated to original network and the one obtained by the removal of nodes and edges. Two such distances are proposed for this analysis: Jensen–Shannon and Wasserstein, based respectively on information theory and optimal transport theory, which are shown to offer a different characterization of vulnerability. Extensive computational results, including two real-world water distribution networks, are reported comparing the new approach to the traditional metrics. This modelling and algorithmic framework can also support the analysis of other networked infrastructures among which power grids, gas distribution and transit networks.

Keywords: Network analysis, Robustness, Vulnerability, Water distribution network, Spectral analysis, Jensen–Shannon divergence, Wasserstein distance

Introduction

Overview and motivation

Robustness and resilience describe the capability of the network to withstand failures and perturbations in its components and keep delivering services regardless of disruptive events, either random or malicious, as, in a water distribution network (WDN), failures in pumping stations or valves and severe bursts in the main pipes.

Resilience, robustness, reliability and vulnerability are terms strictly linked and often confusingly used. Scholz et al. (2012) gives a comprehensive analysis of the different contexts in which the above terms are used. Evaluating robustness of complex and interconnected networks requires the identification and mapping of critical components (i.e., nodes/edges) whose failure impairs the functioning of the network and the assessment of how much this impacts the ensuing increase in vulnerability. The resilient operation of complex networks depends on their structural connectivity i.e., the existence of redundant paths between pair of nodes.

The most used analysis of vulnerability is carried out by studying different measures of the connectivity of the graph as expressed by centrality indices. According to a widely used metric (Latora and Marchiori 2007) an increase in vulnerability is the loss of efficiency as a consequence of the failure of a set of nodes/edges and their removal from the network. The structure and functions of the network strongly rely on the existence of paths between pair of nodes: when nodes and/or links are removed the length of such paths will increase and eventually some couples of nodes will become disconnected. A closely related analysis can be carried out using spectral graph theory: algebraic connectivity, given by the smallest positive eigenvalue of the Laplacian matrix of the network, was introduced in Fiedler (1973). The larger is the algebraic connectivity and the more difficult is to cut the network into disconnected components.

The work presented in this paper focuses on analyzing the criticality of links towards the overall vulnerability of the network by combining notions from graph theory with probabilistic analysis of dissimilarity between networks.

While previously quoted approaches are based on average values of shortest paths the approach proposed in this paper is based on the node-to-node distance probability distributions. These node related distributions are then aggregated into a network wise distribution. At this point the similarity between graphs can be associated to a distance between distributions. The 2 measures considered are the Jensen–Shannon (JS) divergence, based on Kullback–Leibler divergence, and the Wasserstein-1 (WST-1, aka the Earth-Mover) distance based on optimal transport theory. Extensive computational results, including two real-world water distribution networks, are reported comparing the distance-based approach with the traditional robustness metrics based on centrality measures, the loss of efficiency and spectral analysis.

Related works

The analysis of WDNs using measures and analytical tools from complex network theory has been, at the authors' knowledge, first suggested in the seminal paper (Yazdani and Jeffrey 2011) and further developed in Shuang et al. (2014), Archetti et al. (2015), Soldi et al. (2015). A closely related approach, spectral analysis, has been also used in Candelieri et al. (2017), Diao et al. (2016), Herrera et al. (2016), Maiolo et al. (2018) for assessing the resilience in WDNs.

Di Nardo et al. (2018) focuses on spectral techniques and shows how spectral metrics and algorithms support critical tasks of WDN management by just using topological and geometric information. More recently Ulusoy et al. (2018) proposed a hydraulically informed measure of criticality called water flow edge betweenness centrality (WFEBC), built upon on shortest-path and random walks betweenness measures (Herrera et al. 2016, 2015). Shuang et al. (2019) is a wide survey of quantitative resilience methods of WDN including network base approaches. Diao (2020) extends this analysis to multiscale resilience in water distribution and drainage systems.

In more general terms the very issue of differential robustness hinges on a notion of distance or similarity between networks or graphs. The distributional approach to network dissimilarity has been suggested in Schieber et al. (2017).

Entropy based methods like Kullback–Leibler, and Jensen–Shannon are most widely used but can run into problems when the distributions have different supports. This limitation might be relevant when a network has been derived deleting nodes/edges. For this reason, we have introduced in this paper the use of the Wasserstein distance. Wasserstein distance in particular has become important in image analysis and text analysis (Bonneel et al. 2016). More references will be given in "Conclusions and perspectives" section. Here we only quote Deza and Deza (2009) for a general survey of distances, Cover and Thomas (2006) as a general introduction to information theoretic arguments and Villani (2008) for a mathematical framework of optimal transport problems which the Wasserstein distance belongs to.

The contributions of this paper

The key element in this paper is that nodes and graphs are represented as probability distributions: this allows a richer representation of the structural properties of the network than that enabled by the average values of shortest paths and other connectivity metrics. The probability distribution associated to each node, of the node to node distance, is a signal of the relative importance of that node in the graph. The node related distributions can then be aggregated into a distributional representation of the whole network.

The main contribution of this paper is the introduction of a new set of vulnerability metrics given by the distance between the probability distributions of node–node distances between the original network and that resulting from the removal of nodes/edges. These metrics allow the formulation of measures of criticality of single edges in the network.

Two distances have been analysed: Jensen–Shannon divergence based on information theoretic arguments and Wasserstein (WST) distance (aka the Earth-Mover) distance, based on optimal transport theory. The computational results confirm that the values of the distances JS and WST are strongly related to the criticality of the removed nodes/edges and can capture numerically the impact of the deletion of a specific node/edge. A key advantage of the Wasserstein distance is that, under quite general conditions specified in Peyré and Cuturi (2019), it is a differentiable function of the parameters of the distributions which makes possible its use to assess the sensitivity of the network robustness to distributional perturbations.

Organization of the paper

The structure of the paper is as follows: "Entropy based measures of distance between networks" section describes entropy-based measures of distance, in particular JS, between networks. "Wasserstein distance" section describes distances, in particular WST, based on transport theory. "Experimental setting and computational results" section describes the modeling and algorithmic structure of the analysis framework proposed and the computational results on the networks used in this study. Finally, in "Conclusions and perspectives" section some conclusions and perspectives are provided.

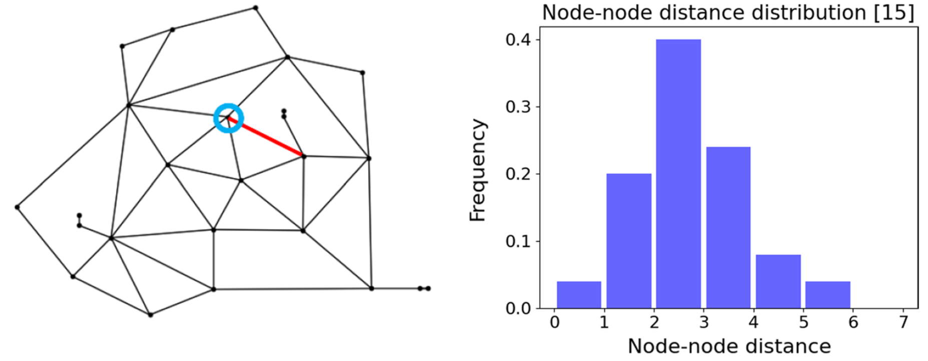

Fig. 1 A schematic representation of Anytown (left). The node–node distance distribution of node 15 (right), highlighted in light blue in the network

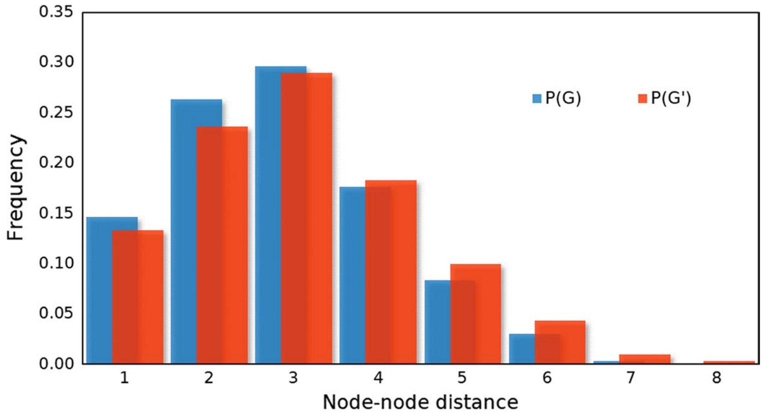

Fig. 2 The node–node distance distributions at network level of G and G'

To get a better focus on the main arguments, the background notions are contained in the appendices (A. graph basic definitions, network measures and spectral analysis, B. efficiency and vulnerability measures, and C. for the data and software resources).

Entropy based measures of distance between networks

Given a graph $G(V, E)$ we associate to each node $i = 1, \dots, n$ a discrete probability distribution

$$P_i(k) = \frac{n_{i,k}}{n-1} \tag{1}$$

as the fraction of nodes which are connected to i at a distance k. The distribution is displayed in Fig. 1. Figure 1a displays Anytown, used in the literature as a benchmark (Farmani et al. 2005): the associated graph G consists of 25 nodes and 44 edges.

The support of this distribution is $1, \dots, D(G)$ where D(G) is the diameter of G. The distance distribution over the whole network is represented by the global histogram in Fig. 2.

$$P_G(k) = \mu_k = \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{n_{i,k}}{n-1} = \tfrac{1}{n}\sum_{i=1}^{n} P_i(k) \tag{2}$$

Let $G\prime$ be the graph without the red edge and p and p′ the distributions $P_G(k)$ and $P_{G'}(k)$. For Anytown $D(G) = 8$ and the two histograms are displayed in Fig. 2.

$$\begin{aligned} P_G = [&0.147, 0.263, 0.297, 0.177, & P_{G\prime} = [&0.133, 0.237, 0.290, 0.183,\\ &0.083, 0.030, 0.003, 0] & &0.100, 0.043, 0.010, 0.003] \end{aligned}$$

The most widely used distance measure is the Kullback–Leibler (KL) divergence

$$KL\left(p|\frac{p+p'}{2}\right) = \int p \log \frac{2p}{p+p'} dx \tag{3}$$

which has the drawback of being asymmetric and possibly infinite when there are points such that $p = 0$ and $p\prime > 0$.

The Jensen–Shannon divergence is built on KL and is symmetric and always definite.

$$JS(p|p') = \frac{1}{2}KL\left(p|\frac{p+p'}{2}\right) + \frac{1}{2}KL\left(p\prime|\frac{(p+p')}{2}\right) \tag{4}$$

The use of *Jensen–Shannon* divergence in computing the dissimilarity between networks has been considered in (Schieber et al. 2017) along with the distance D_{JS}:

$$D_{JS}(P_G, P_{G'}) = \sqrt{JS(P_G(k), P_{G'}(K))} \in [0, 1]. \tag{5}$$

Wasserstein distance

Wasserstein distance is a measure of the distance between two probability distributions. It is also called Earth Mover's (EM) distance from its informal interpretation as the minimum cost of moving and transforming a pile of sand in the shape of one probability distribution to the shape of the other distribution. The cost is quantified by the amount of sand moved times the moving distance. If the distribution domain is continuous the formula for the Earth-Mover (EM) distance is:

$$W(p, p') = \inf_{\gamma \in \Pi(p,p')} \mathbb{E}_{(x,y)\sim\gamma}\left[||x-y||\right] \tag{6}$$

where $\Pi(\mathrm{p}, p\prime)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively p and $p\prime$. One joint distribution $\gamma(x, y) \in \Pi(\mathrm{p}, p\prime)$ describes one transport plan: intuitively $\gamma(x, y)$ indicates how much mass must be transported from x to y in order to transform the distribution p into the distribution $p\prime$. Therefore, the marginal distribution of over x adds up to p $\sum_x \gamma(x, y) = p'(y)$ and analogously $\sum_y \gamma(x, y) = p(x)$. If x is the starting point and y the destination the total amount of sand moved is $\gamma(x, y)$ and the traveling distance is $||x - y||$ and thus the cost is $\gamma(x, y)\,||x - y||$.

The expected cost averaged over all the (x, y) pairs can be computed as:

$$\sum_{x,y} \gamma(x,y)\,||x-y|| = \mathbb{E}_{(x,y)\sim\gamma}\left[||x-y||\right] \tag{7}$$

Finally, we take the minimum among the costs of all sand moving solutions as the EM distance: the EM distance is the cost of the optimal transport plan.

This paper is concerned with the Wasserstein metric for discrete distributions—specifically one-dimensional histograms-in the Euclidean space. We focus on concepts related to the practical calculation of this metric. The Wasserstein metric is motivated by the classical optimal transportation problem first proposed in Monge (1781).

This problem received its modern linear programming formulation by Kantorovich (1942). There are two sets of points x_i with $i = 1, \ldots, m$ and y_j with $j = 1, \ldots, n$. The cost of transporting one unit from x_i to y_j is given by c_{ij}. The transport plan can be expressed in the form of an $m \times n$ matrix where the element γ_{ij} represent the amount (of sand) transported from x_i to y_j. The central problem is to find the transport plan that minimizes the cost of total transportation cost which is the sum of the cost on the available roots $\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\gamma_{ij}$.

Recently the Wasserstein distance has become a key tool in image processing (Bonneel et al. 2016) and machine learning (Frogner et al. 2015) and it has been used also for the generation of adversarial networks (Arjovsky et al. 2017; Weng 2019).

The formulation, of the WST distance for general probability measures requires sophisticated mathematical models (Villani 2008). Its computation raises challenging mathematical and computational problems (Peyré and Cuturi 2019).

In particular when the cost matrix is 0 on the diagonal and 1 elsewhere, Earth Mover distance is given by the l_1 norm (Manhattan distance) of the difference between histograms.

In the case of water distribution networks, the distributions of node–node distances are discrete and 1-dimensional, as shown in Fig. 2, and the computation of WST distance reduces to the comparison of two 1-D histograms: the corresponding Wasserstein distance can be computed very efficiently by using a simple sorting.

There are two key advantages of WST over JS: also in the cases when the distributions are supported in different spaces, even without overlaps, WST can still provide a meaningful representation of the distance between distributions. Furthermore, a key advantage of WST is its differentiability. Both points are illustrated in the following example (Fig. 3).

Let's consider $Z = U(0,1)$ the uniform distribution on the unit interval. Let P be the distribution of $(0, Z)$ (0 on the x-axis and the random varable Z on the y axis and $P_\theta = (\theta, Z)$.

- $KL(P, P_\theta) = +\infty$ if $\theta \neq 0$ and 0 if $\theta = 0$.
- $JS(P, P_\theta) = \log 2$ if $\theta \neq 0$ and 0 if $\theta = 0$.
- $W(P, P_\theta) = \theta$ if $\theta \neq 0$ and 0 if $\theta = 0$.

Therefore, Wasserstein provides a smooth measure which is useful for any optimization and learning process using gradient descent (Arjovsky et al. 2017).

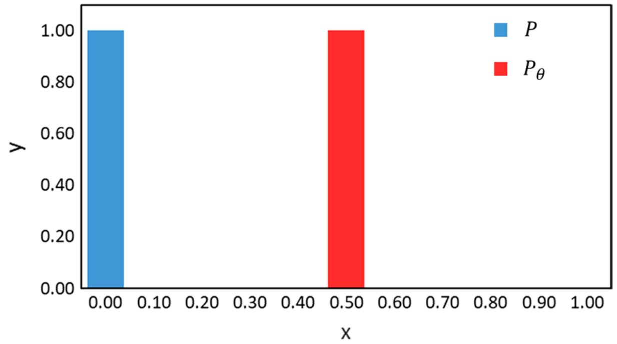

Fig. 3 Two probability distributions P and P_θ. If $\theta \neq 0$ there is no overlap between P and P_θ

In our WDN's the support of $P_G(k)$ is given by the integers $k = 1, \ldots, D(G)$ where $D(G)$ is the diameter of G ("Appendix A.1") (analogously for $G\prime$). When $G\prime$ is derived from G removing some edges, we have $D(G\prime) \geq D(G)$. Since we are in the particular case where the distributions are represented by histograms one can extend to G the support of $G\prime$ setting $\mu_G(k) = 0$ for $k = D(G) + 1, \ldots, D(G\prime)$.

$$\begin{aligned} P_G = [&0.147, 0.263, 0.297, 0.177, \quad P_{G'} = [0.133, 0.237, 0.290, 0.183, \\ &0.083, 0.030, 0.003, 0] \qquad\qquad\quad 0.100, 0.043, 0.010, 0.003] \end{aligned}$$

The informal interpretation of WST, from which its name Earth Mover is derived, is captured by the formula.

$$\delta_k = \delta_{k-1} + P_G(k) - P_{G'}(k) k = 1, \ldots, \max\left(D(G), D\left(G'\right)\right)$$

$$W\left(G, G'\right) = \sum |\delta_i| = 0.1767, \quad D_{JS}\left(G, G'\right) = 0.0718$$

Experimental setting and computational results

In our problem the network space is given by the basic network G and the subgraphs obtained by the removal of one or more edges. The elements of this space are represented as probability distributions of node-2-node distances (Eq. 1) and aggregated into a distributional representations of the basic network and of the subgraphs (Eq. 2). In this space it takes place the computation of the Wasserstein distance. The result of this computation is mapped back into the network space as measures of network dissimilarity and labels of criticality of individual components. In 4.1 the experimental setting is described.

Experimental setting

The networks considered are:

- Neptun is the WDN of the Romanian city of Timisoara, with an associated graph of 333 nodes and 339 edges.
- Abbiategrasso refers to a pressure management zone in Milan (namely, Abbiategrasso) with an associated graph consisting of 1213 nodes and 1391 edges.

The description of these networks is also given in “Appendix C”. As far as efficiency and algebraic connectivity measures are concerned, we have used the standard measures reported in the “Appendix B”.

The first step in the analysis is clustering in order to identify the specific edges whose removal induces a disconnection of the network. The number of clusters K is set according to context information about the districtualization adopted by the water utility. Failures affecting also only one pipe may imply a reduction in the efficiency of the network and an increase in vulnerability. The software used is described in “Appendix C.2”.

The two real-world WDNs analyzed are very sparse (with density q lower or equal to 0.006). The two real-world WDNs (given in “Appendix C.1”) are effectively planar and “almost” regular. This fact can be due to the fact that their structure is strongly constrained by spatial characteristics making a classification based on nodal degree distribution less meaningful.

In particular their degree distribution does not follow a power law and measures, given in Table 1, for Neptun and Abbiategrasso, really set them apart from other kinds of networks like transportation, communications and social. The usefulness of measures based on centrality or spectral measures in assessing the increase of vulnerability or the loss of robustness is quite limited, according to our results, in WDN of realistic size.

Computational results

The computational results, which are quite unique in the literature given the size of the networks analysed, demonstrate that probabilistic distance measures show better capacity to discriminate between different networks not only globally but also edge-wise. The results are organized around figures and tables. Figures 4 and 5 display (respectively for Neptun and Abbiategrasso) the output of clustering, that is critical edges whose removal generates a disconnection; these edges are highlighted in red.

Table 1 displays centrality measures. Table 2 efficiency, vulnerability, and algebraic connectivity. Table 3 the values of distributional distances compared with the loss of efficiency. The most relevant result is that using discrete probability distributions to represent the graph associated to the WSN, as well as all its elements, provides a novel and more effective analytical framework to assess similarity between two networks (Table 1), as well as the same network before and after some modifications, such as expansion, disruptions, etc. Especially in the case of disruptive events, the novel WST based analytical framework enable the definition and the computation of new and more discriminant criticality indices (comparison between Tables 2 and 3),

Table 1 Topological measures

Measure	Neptun	Abbiategrasso
Density (q)	0.0061	0.0019
Link-per-node ratio (e)	1.0180	1.1467
Central point dominance ($c_{b'}$)	0.2432	0.3100
Clustering coefficient (CC)	0.0000	0.0055
Diameter	57	83
Characteristic path length	23.7613	30.6126

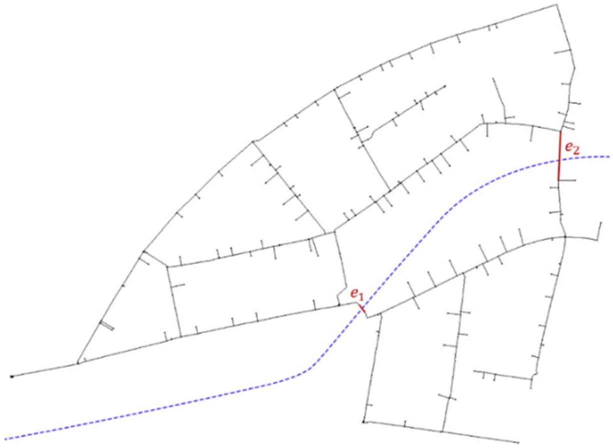

Fig. 4 Neptun WDN (clustering with K = 2): Critical edges (red) whose removal generates a disconnection

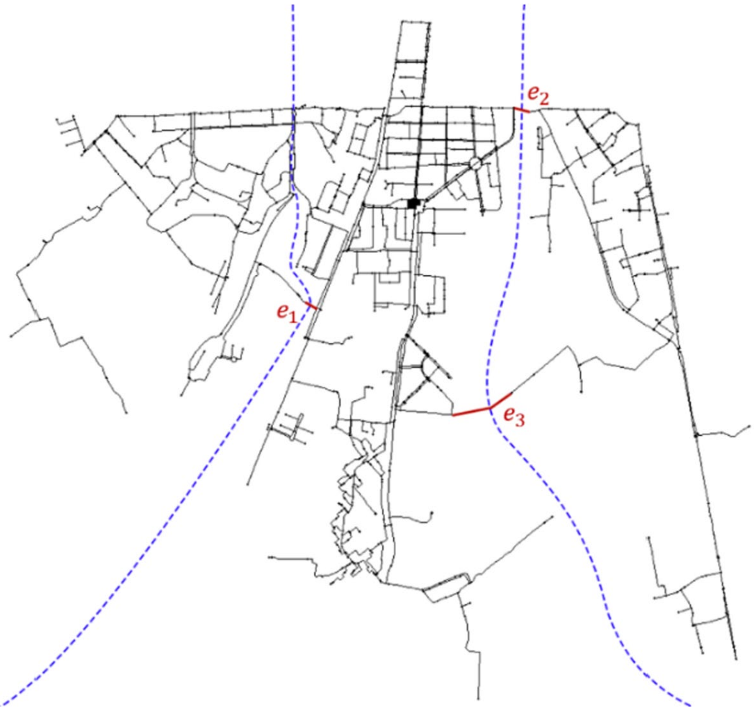

Fig. 5 Abbiategrasso WDN (clustering with K = 3): Critical edges (red) whose removal generates a disconnection

ideally applicable to any critical networked infrastructure (transportation networks, energy grids, communications networks, etc.). Another relevant benefit provided by the novel WST based analysis is that it can deal with distributions with different supports as well as sparsity, and it is not affected by different binning schemes. This is a relevant advance with respect to previous approaches based on probability distributions, for instance those based on the Kullback–Leibler or the Jenses–Shannon divergences, which require to adopt some drawbacks for dealing with the two mentioned

Table 2 Vulnerability measures

Neptun	E	V_{MEAN}	V_{MAX}	**Algebraic connectivity**
G	0.068608	0.018927	0.072646	0.0018
G'(*removing* e_2)	0.065390	0.024181	0.211362	0.0007
G''(*removing* e_1)	0.064486	0.024796	0.194813	0.0006
G'''(*disconnected*)	0.051924	0.016642	0.068246	0.0000
Abbiategrasso	E	V_{MEAN}	V_{MAX}	**Algebraic connectivity**
G	0.047557	0.003436	0.150390	0.0004
G'(*removing* e_2)	0.045019	0.003935	0.181174	0.0003
G''(*removing* e_3)	0.046385	0.003642	0.205294	0.0004
G'''(*removing* e_1)	0.040405	0.002628	0.060728	0.0000
G''''(*disconnected*)	0.031077	0.002251	0.057007	0.0000

Table 3 Probabilistic distances versus loss of efficiency

Neptun	**Jensen–Shannon**	**Wasserstein**	**Loss of efficiency**
G, G'	0.1677	3.3183	0.0469
G, G''	0.2456	5.4704	0.0601
G, G'''	0.3286	6.5542	0.2432
Abbiategrasso	**Jensen–Shannon**	**Wasserstein**	**Loss of efficiency**
G, G'	0.0935	3.1040	0.0534
G, G''	0.0528	1.5170	0.0246
G, G'''	0.1843	5.4871	0.1504
G, G''''	0.3633	8.8845	0.3465

issues. Even more important, the proposed WST based analysis allows for an intuitive visualization of the criticality of each edge, matching the typical requirement for a "priority ranking" of critical elements, driving the preparation of a budget-constrained rehabilitation plan.

For the WST distance it has been computed also the heat map of edge wise criticality for Neptun and Abbiategrasso. In Fig. 6, the colour associated to each edge (i,j) is an index of criticality given by the Wasserstein distance between G and $G\backslash(i,j)$.

Conclusions and perspectives

The main result of this paper is that probabilistic measures based on the probability distribution of node–node distances, yield a distance, between the original network and that resulting from the removal of some nodes/edges, which can provide a set of indicators of the increase in vulnerability. The computational results confirm that the value of the distances Jensen–Shannon and Wasserstein confirm the results of clustering.

Albeit the measures of vulnerability analyzed in this paper are based on topological arguments they are anyway important to narrow the set of critical components before moving to the hydraulic verification.

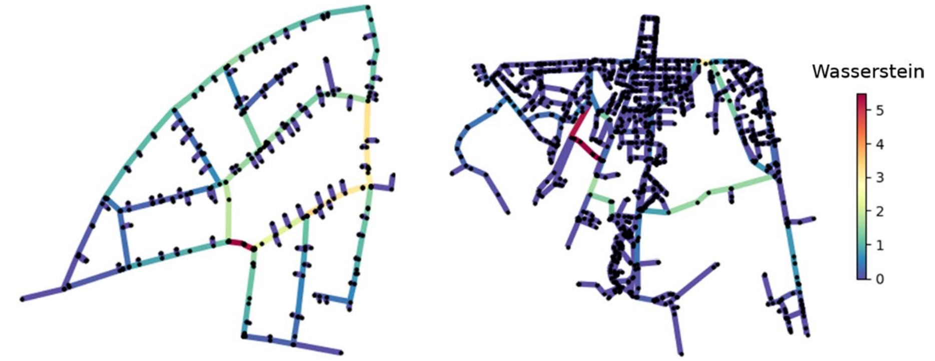

Fig. 6 Heatmap of edge wise criticality for Neptun (left) and Abbiategrasso (right). The colour associated to each edge (i, j) is an index of criticality given by the WST distance between G and $G \backslash (i, j)$

Significantly the differentiability of WST allows to include in the differential robustness analysis edges weighted by parameters derived from the flow dynamics in the network correlating connectivity analysis to hydraulic performance indicators.

This analysis framework supports decision making at design stage, to simulate alternative network layouts of different robustness, and also at operational stage where the decision to be taken can be, which nodes/edges are to temporarily be removed for maintenance and rehabilitation. Indeed, critical tasks of WDN management can be supported by just using topological and geometric information. The analysis framework also helps for the efficient and automatic definition of district metered areas and to facilitate the localization of water losses through the definition of an optimal network partitioning.

The modelling and algorithmic framework platform developed can be straightforwardly translated to many networked infrastructures among which power grids, transit networks but also global supply chains network whose vulnerability has been exposed in the recent COVID crisis.

Appendix A

Graph theory is the mathematical basis to provide a unifying language for the study of networks: with this in mind it is useful to give some basic definitions which will be used in the sequel. For a wide-ranging analysis of the role of graph theory in the analysis of networks the reader is advised to look at (Newman 2010).

A.1 Basic definitions

Let denote a graph with $G = (V, E)$, where V is the set of nodes and E is the set of edges. Each edge of G is represented by a pair of nodes (i, j) with $i \neq j$, and $i, j \in V$ and with $n = |V|$ and $m = |E|$. If $(i, j) \in E$, i and j are called adjacent nodes. A graph G is undirected if (i, j) and (j, i) represent the same edge. A graph G is simple if no self-loops are admitted (edges starting from a node and ending on the same node) and only one edge can exist between each pair of nodes (i, j), with $i \neq j$. The adjacency relationship between the nodes of G can be represented through a non-negative $n \times n$ matrix A (i.e., the adjacency matrix of G). The entry $a_{i,j}$ of the adjacency matrix A is 1 if i and j are adjacent nodes (i.e., $\left(i,j\right) \in E$), and 0 otherwise. Furthermore, $a_{ij} = a_{ji}$ if G is undirected and a_{ii} (entries on the diagonal) are 0 if G is simple.

- The degree of the node i, k_i is the number of edges having i as one of the two nodes on the edge: $k_i = \sum_{j=1}^{n} a_{i,j}$. Anyone of the edges having i as one of its nodes is called incident on i.
- When G is *directed*, meaning that the order of the two nodes of an edge is relevant for its definition, the k_i can be split into *out-degree* (number of edges having i as first node) and *in-degree* (number of edges having i as second node).
- A *path* in a graph is a sequence of nodes connected by edges the length of the path is the number of edges. A connected component is a maximal subgraph when all nodes can be reached from every other.
- *The shortest path* between i and j is the path with the smallest length. This length is called the distance between i and j $d_{i,j}$. The largest distance among each possible pair of nodes in G is named diameter $D(G)$.
- The characteristic path length is the average distance for every possible pair of nodes $\left(i,j\right)$.

$$L_g = \frac{1}{n*(n-1)} \sum_{j=1}^{n} \sum_{k \neq j} d\left(j,k\right)$$

A useful representation is to arrange the distances in the distance matrix $D = \left[d_{i,j}\right] i,j = 1,\ldots,n$. The maximum entry of row i $\max_{j=1,\ldots,n} d_{i,j}$ is also known as the *eccentricity* of node i. The maximum eccentricity among the nodes is equal to $D(G)$.

A subgraph $G' = (V', E')$ of G is a graph such that $V' \subseteq V$ and $E' \subseteq E$; a connected component of G is maximal if is the largest possible subgraph for which you could not find another node in the graph that could be added to the graph with all the nodes be still connected.

The core concept is centrality which addresses the question "which are the most important nodes in a network?". There are many centrality measures from the simplest like node degree, which can anyway be illuminating, to eigenvector-based measures like Page Rank.

A.2 Basic measures

The density of the network is the fraction of edges which are present in the network:

$$q = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}$$

The number of edges $m = \frac{1}{2} \sum_{i=1}^{n} k_i$.

If c is the mean node degree, $c = \frac{1}{n} \sum_{i=1}^{n} k_i$ we get $c = \frac{2m}{n}$ and $q = \frac{c}{n-1}$.

The density is in the range $(0, 1)$.

A *cut-set*, specifically a node cut-set, is a set of nodes whose removal disconnects i and j. A *minimum cut-set* is the smallest cut-set. Analogously for edge cut-set.

The *centrality measures* address the issue of the relative importance of nodes/edges. The most widely used measures are:

- *Closeness centrality* $C = \frac{n}{\sum d_{ij}}$ is based on the mean distance from i to j averaged on all nodes.
- *Betweenness centrality*: let be $\eta^i_{st} = 1$ if node i lies on the shortest path from s to t and 0 otherwise. Then, betweenness centrality is given by $b_i = \frac{1}{n^2}\sum_{s,t=1}^{n} \eta^i_{st}$. It measures the extent to which a node lies on the paths between other nodes. We can similarly define an *edge betweenness* that counts the number of shorter paths that run along the edge.
- *Link-per-node ratio (e)*, as the number of edges of a graph with respect to the number of its nodes. $e = \frac{m}{n}$.
- *Central point dominance* $c'_{b'}$ based on betweenness centrality is a measure for characterizing the organization of a network according to its path-related connectivity; $c'_b = \frac{1}{n-1}\sum_{i=1}^{n} (b_{max} - b_i)$ where b_i is the betweenness centrality of the node i and b_{max} is the maximum value of betweenness centrality over all the n nodes of the network.
- The *clustering coefficient (CC)* is the number of triangles with respect to the overall number of possible connected triples, where a triple consists of three nodes connected at least by two edges while a triangle consists of three nodes connected exactly by three edges:

$$CC = \frac{3N_{triangles}}{N_{triples}}$$

There are other definitions of CC for which the reader is addressed to Newman (2010).

To compute the centrality indices in this paper, the open-source software Cytoscape (http://www.cytoscape.org/) has been adopted (Morris et al. 2011). This point will be further discussed in "Wasserstein distance" section.

A.3 Spectral measures

Spectral graph theory studies the eigenvalues of matrices that embody the graph structure. One of the main objectives in spectral graph theory is to deduce structural characteristics of a graph from such eigenvalue spectra.

In case of undirected graphs, the adjacency matrix $A(G)$ is symmetric and all its eigenvalues are real. The eigenvalues $\mu_1(G) \leq \mu_2(G) \leq ... \mu_n(G)$ of $A(G)$ are called the spectrum of G. The largest eigenvalue of the adjacency matrix $\mu_n(G)$ is called spectral radius of G and is denoted by ρ(G). An important property is given by the following inequality.

$$\sqrt{\Delta(\mathrm{G})} \leq \rho(G) \leq \Delta(\mathrm{G}),$$

where $\Delta(\mathrm{G}) = \max k_i : i = 1, \ldots, n$ that relates the spectral radius with Δ(G), the maximum degree of the nodes.

The spectrum of A(G) allows to define the *Eigenvector centrality* of the node i, is $x_i = \rho(\mathrm{G})^{-1} \sum_{j=1, j \neq i}^{n} a_{ij} x_j$. *Katz centrality* and *Page Rank* algorithm are just parametrized version of eigenvector centrality (Newman 2010).

The difference $s(G) = \rho(G) - \mu_{n-1}(G)$ between the spectral radius of G and the second largest eigenvalue of the adjacency matrix $A(G)$ is called the spectral gap of G. A small value of $s(G)$ is usually observed through low connectivity, and the presence of bottlenecks and bridges whose removal cuts the graph into disconnected parts.

The Laplacian matrix of G is an n × n matrix $L(G) = D(G) - A(G)$, where $D(G) = diag(k_i)$. The matrix $L(G)$ is positive semi-definite in case of simple graph. The eigenvalues of $L(G)$ are called the Laplacian eigenvalues of G. The Laplacian eigenvalues $\lambda_1(G) = 0 \leq \lambda_2(G), \ldots \leq \lambda_n(G)$ are all real and nonnegative. The smallest eigenvalue is always equal to 0 with multiplicity equals to the number of connected components of G. The second smallest eigenvalue is called the algebraic connectivity of G which is one of the most widely used measures of connectivity. Larger values $\lambda_2(G)$ represent higher robustness against efforts to disconnect the graph, so the larger it is, the more difficult it is to cut a graph into independent components. An important inequality for the algebraic connectivity is given by

$$\lambda_2(G) \leq \frac{n}{n-1}\delta(G),$$

that relates it with the minimum degree of the nodes $\delta(\mathrm{G}) = \min_{i=1,\ldots,n} k_i$. In case of a connected graph, also the following inequality can be proved

$$\lambda_2(G) \geq \frac{4}{n \cdot D(G)}$$

that relates the algebraic connectivity with the diameter of the graph. Another spectral distance is based on the analysis of the eigenvectors of the Laplacian.

Appendix B

The performance of the network after the removal of nodes/edges is often evaluated as the change of the efficiency, as defined in as

$$E = \frac{1}{n(n-1)} \sum_{i,j \in V, i \neq j} \frac{1}{d_{ij}},$$

where the d_{ij} represent the distance between i and j. Normalization by $n(n-1)$ ensures that $E \leq 1$, in case of unweighted graph. The maximum value $E = 1$, is assumed if and only if the graph is complete.

B.1 Efficiency based vulnerability measures

A way to measure the vulnerability of the network is using the loss of efficiency observed when we remove some nodes/edges. The relative drop in the network efficiency caused by the removal of a node i from the graph is defined as

$$C^E_\Delta(i) = \frac{E(G) - E(G\backslash\{i\})}{E(G)},$$

where $G\backslash\{v\}$ denotes the network G without the node i. The loss of efficiency of G is defined as

$$V_{MAX}(G) = \max_{i \in V} C^E_\Delta(i), \tag{1}$$

$$V_{MEAN}(G) = \frac{1}{\mathrm{n}} \sum_{i \in V} C^E_\Delta(i).$$

The larger the value $V_E(G)$, the larger is the vulnerability. Analogous formulas can be written removing the edges.

B.2 Spectral vulnerability measures

There is no specific formula, contrary to those reported in the previous subsections, linking spectral analysis to a measure of vulnerability related to the removal of a node. However, both algebraic connectivity λ_2, the second smallest eigenvalue of the Laplacian $L(G)$ and spectral gap $s(G)$, the difference between the spectral radius of G and the second eigenvalue of the adjacency matrix $A(G)$, are indicators of difficulty to split the graph. The larger the algebraic connectivity, the more difficult it is to disconnect the graph. It is also related to the min-cut problem in spectral clustering. A large value of the spectral gap, together with a uniform degree distribution, results in higher structural sturdiness and robustness against node and link failures. On the contrary, low values of spectral gap indicate a lack of good expansion properties usually represented by bridges, and network bottlenecks. The larger the spectral gap the more robust is the network.

Appendix C

C.1 Data resources

- Neptun is the WDN of the Romanian city of Timisoara, with an associated graph of 333 nodes and 339 edges. This network has been a pilot in the European project Icewater.
- Abbiategrasso refers to a pressure management zone in Milan (namely, Abbiategrasso) with an associated graph consisting of 1213 nodes and 1391 edges. This network has also been a pilot in the European project Icewater.

In analyzing WDNs one must consider that most of the end-users are supplied by single connections. A preliminary preprocessing has been performed in order to identify the simple core of the network by removing all nodes with degrees smaller than two (also called leaves of the network) along with the parent edge connecting them to their associated root-node.)

C.2 Software resources

The tools for the analysis of connectivity and efficiency have been coded in Pythton.

For the network visualization it has been used the Python package WINTR, based on EPANET 2 in WDN's. which enables network editing and hydraulic simulation.

The software used for the computation of JS is the function JensenShannonDivergence from the Java library Mallet. The software used for the computation of the Wasserstein distance is the function EarthMoversDistance from the Java library Apache Commons Math. For the clustering it has been used an effective computational scheme which uses a data representation in the lower dimensional space spanned by the most relevant eigenvectors of the normalized Laplacian matrix. The basic steps are:

1. Construct the affinity matrix S(G) = I + A(G) whose eigenvalues are the same as $A(G) + 1$.
2. Construct the matrix $L_N(G) = D^{-\frac{1}{2}} S(G) D^{-\frac{1}{2}}$ where $D = [d_{ii}] = k_i$, where k_i is the degree of node i.
3. Compute the eigenvalues of $L_N(G)$ and the eigenvectors corresponding to the K largest eigenvalues of $L_N(G)$ and denote them by $u_1, u_2, \ldots, u_K u_1, u_2, \ldots, u_K$
4. We build the matrix U such that the kth column of U is u_k and normalize the rows such that each row has unit length.
5. Treating the rows as points in the K-dimensional space $\mathbb{R}^K$ and perform K-means clustering of these points in K clusters.

We used the implementation of this method given in scikit-learn.

Abbreviations

WDN: Water distribution network; KL: Kullback–Leibler; JS: Jensen–Shannon; WST: Wasserstein; EM: Earth mover's.

Acknowledgements

FA acknowledges the "contribution" of his son Davide whose early fascination with the "life of a droplet" might have kindled the father's enduring interest in the things water. We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources (DEMS—Department of Economics, Management and Statistics).

Authors' contributions

AP contributed to conceptualization, investigation, software, and writing. AC contributed to conceptualization, investigation, software, and writing. IG contributed to conceptualization, investigation, software, and writing. FA contributed to conceptualization, investigation, and writing. All authors read and approved the final manuscript.

Funding

This study has been partially supported by the Italian project "PERFORM-WATER 2030"—programma POR (Programma Operativo Regionale) FESR (Fondo Europeo di Sviluppo Regionale) 2014–2020, innovation call "Accordi per la Ricerca e l'Innovazione" ("Agreements for Research and Innovation") of Regione Lombardia, (DGR N. 5245/2016—AZIONE I.1.B.1.3—ASSE I POR FESR 2014–2020)—CUP E46D17000120009.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available because they contain information which could be related to the security of the water supply. They are available from the corresponding author on reasonable request on the basis of scientific research program.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy. [2]Department of Economics, Management and Statistics, University of Milano-Bicocca, 20126 Milan, Italy. [3]Oaks srl, Milan, Italy.

Received: 1 May 2021 Revised: 10 October 2021 Accepted: 13 October 2021
Published online: 30 October 2021

References

Archetti F, Candelieri A, Soldi D (2015) Network analysis for resilience evaluation in water distribution networks. Environ Eng Manag J 14:1261–1270

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein GAN. arXiv:1701.07875 [cs, stat]

Bonneel N, Peyré G, Cuturi M (2016) Wasserstein barycentric coordinates: histogram regression using optimal transport. ACM Trans Graph 35:71-1

Candelieri A, Giordani I, Archetti F (2017) Supporting resilience management of water distribution networks through network analysis and hydraulic simulation. In: 2017 21st international conference on control systems and computer science (CSCS). IEEE, pp 599–605

Cover TM, Thomas JA (2006) Elements of information theory (Wiley series in telecommunications and signal processing). Wiley, New York

Deza MM, Deza E (2009) Encyclopedia of distances. In: Encyclopedia of distances. Springer, Berlin, pp 1–583

Di Nardo A, Giudicianni C, Greco R, Herrera M, Santonastaso GF (2018) Applications of graph spectral techniques to water distribution network management. Water 10:45

Diao K (2020) Multiscale resilience in water distribution and drainage systems. Water 12:1521

Diao K, Sweetapple C, Farmani R, Fu G, Ward S, Butler D (2016) Global resilience analysis of water distribution systems. Water research 106:383–393

Farmani R, Walters GA, Savic DA (2005) Trade-off between total cost and reliability for Anytown water distribution network. J Water Resour Plan Manag 131:161–171

Fiedler M (1973) Algebraic connectivity of graphs. Czechoslovak Math J 23:298–305

Frogner C, Zhang C, Mobahi H, Araya-Polo M, Poggio T (2015) Learning with a Wasserstein loss. arXiv:1506.05439

Herrera M, Abraham E, Stoianov I (2016) A graph-theoretic framework for assessing the resilience of sectorised water distribution networks. Water Resour Manag 30:1685–1699

Herrera M, Abraham E, Stoianov I (2015) Graph-theoretic surrogate measures for analysing the resilience of water distribution networks. Procedia Eng 119:1241–1248

Kantorovich L (1942) On the transfer of masses. In: Doklady Akademii Nauk, pp 227–229. **(in Russian)**

Latora V, Marchiori M (2007) A measure of centrality based on network efficiency. New J Phys. https://doi.org/10.1088/1367-2630/9/6/188

Maiolo M, Pantusa D, Carini M, Capano G, Chiaravalloti F, Procopio A (2018) A new vulnerability measure for water distribution network. Water 10:1005

Monge G (1781) Mémoire sur la théorie des déblais et des remblais. In: Histoire de l'Académie Royale des Sciences de Paris

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinform 12:1–14

Newman M (2010) Networks: an introduction, networks. Oxford University Press, Oxford

Peyré G, Cuturi M (2019) Computational optimal transport: with applications to data science. Found Trends Mach Learn 11:355–607

Schieber TA, Carpi L, Díaz-Guilera A, Pardalos PM, Masoller C, Ravetti MG (2017) Quantification of network structural dissimilarities. Nat Commun 8:1–10

Scholz RW, Blumer YB, Brand FS (2012) Risk, vulnerability, robustness, and resilience from a decision-theoretic perspective. J Risk Res 15:313–330

Shuang Q, Zhang M, Yuan Y (2014) Performance and reliability analysis of water distribution systems under cascading failures and the identification of crucial pipes. PLoS ONE 9:e88445

Shuang Q, Liu HJ, Porse E (2019) Review of the quantitative resilience methods in water distribution networks. Water 11:1189

Soldi D, Candelieri A, Archetti F (2015) Resilience and vulnerability in urban water distribution networks through network theory and hydraulic simulation. Procedia Eng 119:1259–1268

Ulusoy A-J, Stoianov I, Chazerain A (2018) Hydraulically informed graph theoretic measure of link criticality for the resilience analysis of water distribution networks. Appl Netw Sci 3:1–22

Villani C (2008) Optimal transport: old and new. Springer, Berlin

Weng L (2019) From gan to wgan. arXiv:1904.08994

Yazdani A, Jeffrey P (2011) Complex network analysis of water distribution systems. Chaos Interdiscip J Nonlinear Sci 21:016111

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.