UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MACHINE LEARNING APPROACH TO MODEL INTERDEPENDENT NETWORK PERFORMANCE

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
MASTER OF SCIENCE

By
RAMINENI CHITTIBABU NAIDU GHANESHVAR
Norman, Oklahoma
2019

MACHINE LEARNING APPROACH TO MODEL INTERDEPENDENT NETWORK PERFORMANCE

A THESIS APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Kash Barker

Dr. Sridhar Radhakrishnan

Dr. Andrés González

ACKNOWLEDGEMENT

I would like to thank my advisor and guide, Dr. Kash Barker, for his guidance, support, encouragement, and giving me an opportunity to work under him. I would like to sincerely thank Mrs. Nafiseh Ghorbani Renani, PhD student, for her consistent help and cooperation. I would also like to thank the Data Science & Analytics Department for the opportunity provided. I want to use this opportunity to also thank my family like friends for helping me and providing moral support.

ABSTRACT

Resilience can be understood as the property of an object or system to recover from its state of disruption back to its complete functional stage as it was prior to the disruption event. There are different ways of measuring the resilience of a system and tracking the system performance is one of the methods. Thus, measuring the time taken by the system to recover to its original state is one of the parameters that can be considered. In this research work, we have focused on building a model(s) that predicts the time taken for an interdependent network to recover and function at 100%. In order to implement this idea, we present a case-study of the system of interdependent water, gas, and power utilities in Shelby County, TN. The model is trained using the train data set from the data set generated by running the optimization code multiple times and observing the time taken for the inter-dependent network to recover completely. The prediction of time to recover is made on the test data set using different models and the results are then compared.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Resilience

While our dependence on infrastructure systems such as electric power, water supply, and telecommunication networks continues to grow along with their dependence on each other, recent natural disasters and malevolent attacks have demonstrated how a single event can cripple such networks and the community that relies on them for an extended period (Ouyang, 2014). Due to the substantial economic and social impacts of disruptions to such interdependent infrastructure systems, ensuring their resilience is a major concern (Kettl, 2013) as even small disruptions can lead to substantial failures in such an interconnected system (Danziger et al., 2016; Eusgeld et al., 2011; Wu et al., 2016).

The concept of resilience has been quantified with various measures in different domains (Hosseini et al., 2016). In this work, we adopt a resilience paradigm based on the system performance across system states illustrated in Figure 1, adapted from Henry and Ramirez-Marquez (2012). In this model, the resilience measure is defined by two primary dimensions: vulnerability, or the drop in the performance after a disruptive event, e, and recoverability, or the timely restoration of the system performance to the desired level. From this model, resilience is quantified as the ratio of recovery at time t to loss noted by $Я_{\varphi}(t|e)$. Network performance at time t is represented with $\varphi(t)$ (e.g., the amount of demand met at a certain node).

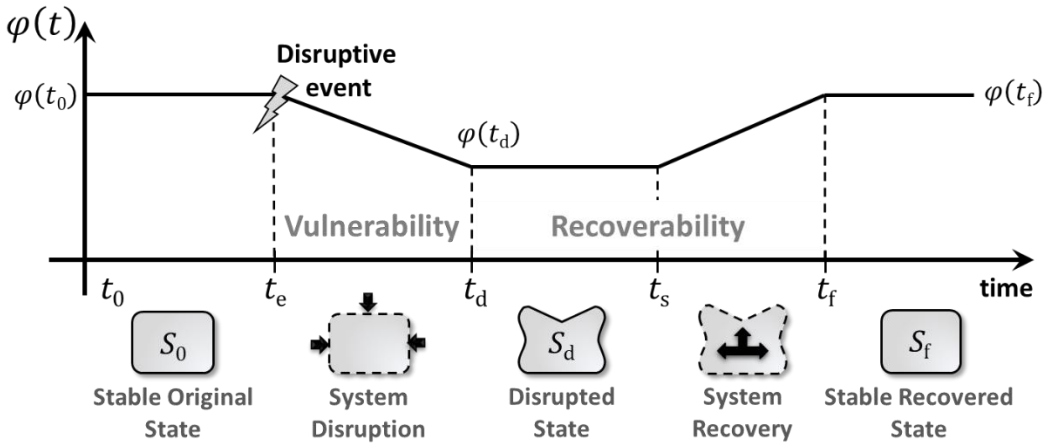

Figure 1. System performance across system states (adapted from Henry and Ramirez-Marquez (2012)).

1.2 Objective

The objective of this research is to predict (using different predictive techniques and compare) the time taken (in seconds) for an interdependent network to recover completely from an instance of interdiction of a set of randomly selected nodes, given the optimization objective of minimizing the cumulative weighted fraction of unsupplied demand and constraints of demand, supply, interdependency, restoration rates of the nodes, work crews available for restoring the network(s) among others.

1.3 Workflow

In this section, the basic work flow followed in achieving the objective is discussed. Firstly, a case-study of the system of interdependent water, gas, and power utilities in Shelby County, TN, is chosen for the research. An optimization code is modelled with the objective of minimizing the cumulative weighted fraction of unsupplied demand (discussed in Section 3.1.4) and constraints involving demand, supply, interdependency, restoration rates of the nodes, work crews available

for restoring the network(s), interdiction of elements of the network such as nodes and links and many others (discussed in Section 3.1.6). Then, a modification is made to this optimization code such that the model is executed multiple times and node attack information, node restoration rate information and the time taken for the network to restore is stored in the form of data frame. The data obtained is then used for predictive analysis and the results are then compared (discussed in Section 4).

2. BACKGROUND

The concept of interdiction on a network can be compared to a situation where a network (for e.g., a city) is interdicted (i.e., attacked) due to a natural disaster or a human-made attack that makes the network unstable or inefficient. Thus, being able to analyze the network's performance after an interdiction (which might cause damage to some of its elements) occurs is a way to be able to take precautionary measures. Predicting the time taken for the network to recover completely is one of the analysis that should be considered.

This section explains the background of machine learning and how it has been used in the field of networks and optimization. Optimization and Machine Learning are the two main concepts this work is built on. While optimization is concerned with exact solutions, machine learning is concerned with generalization abilities of learners (Munoz, 2014). Recently, machine learning has been used in every possible field to leverage its amazing power (Wang et al., 2018). Security is one of the main concerns facing the development of new projects in networking and communications. Another challenge is to verify that a system is working exactly as specified. On the other hand, advances in Artificial Intelligence (AI) technology have opened new markets and opportunities for progress in critical areas such as network resiliency, health, education, energy, economic inclusion, social welfare, and the environment (Hussein et al., 2018). In this study, the focus is on the network resiliency, mainly on the time taken or the speed of the network's recovery after an interdiction occurs. By combining optimization techniques and machine learning methods, the network performance is analyzed and discussed.

3. METHODOLOGY

In this section the methodology followed to implement the idea of building a model that will predict the time taken by our inter-dependent network after a disruptive event is discussed.

It is achieved in 3 steps, the first being developing an optimization code that will select a set of nodes from the network that are being disrupted and reduce the recovery time. Second step is generating data by running the optimization code multiple times for random nodes being disrupted each time and observing the time taken (in seconds) for the network to recover completely. The final step is building different predictive models by training the data (train data set) generated and testing it with the test data set. The predictive models' results are then compared.

3.1. Optimization code

The optimization model used in this work is an extension of the work from (Ghorbani Renani, et al., 2019). Hence the model's notations, parameters and assumptions are taken directly form the source. The only

3.1.1. Network description and representation

For this research, we present a case-study of the system of interdependent water, gas, and power utilities in Shelby County, TN, USA. Figure 2 shows the general illustration of (a) the water, (b) gas, and (c) power networks individually, along with (d) their superposition (adapted from González et al., 2016).

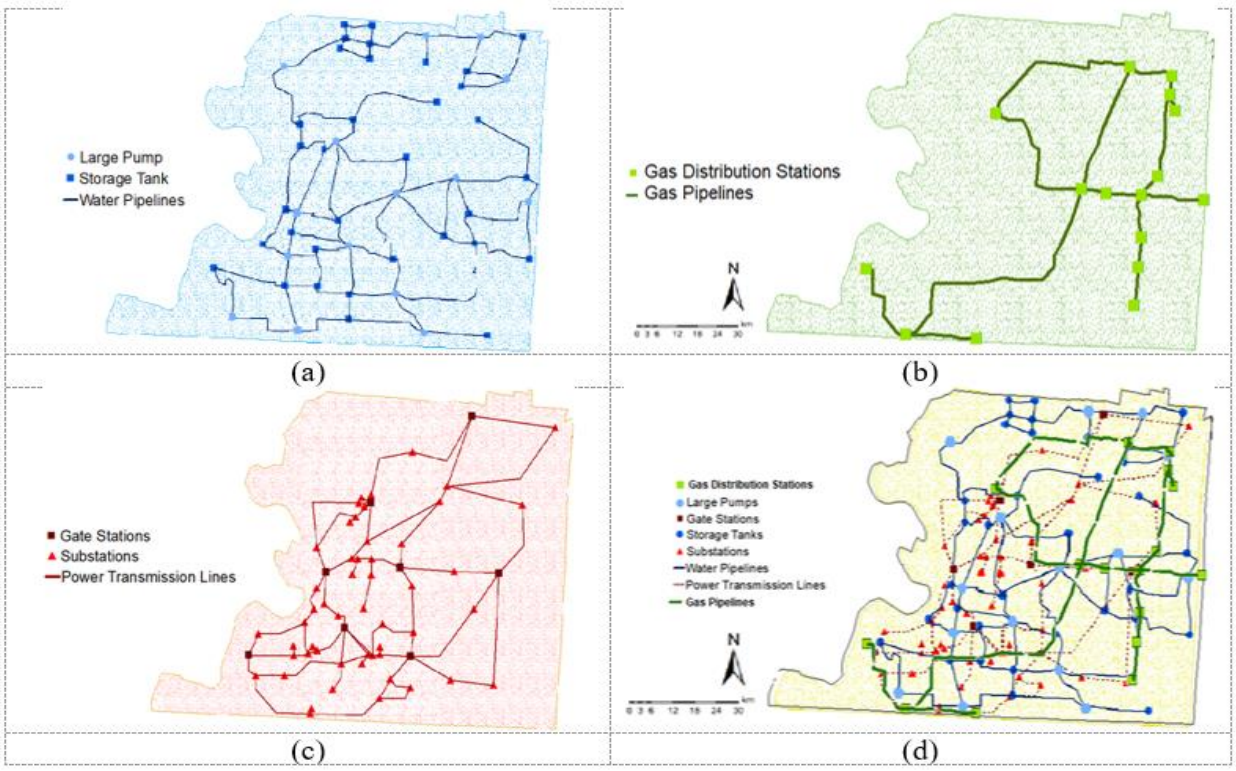

Figure 2. General depiction of the (a) water, (b) gas, and (c) power networks, and (d) the interdependency among networks in Shelby County, TN (adapted from González et al. (2016)).

This model considers the physical interdependency for networks (mentioned in Section 3.1.2) in which the functionality of a set of nodes in one or more networks enable the functionality of a node in another network. In this case study, the power is dependent on the water network, which enables cooling in the power network (Zhang et al., 2016). The interdependent system of networks consists of 125 nodes and 164 links.

Table 1. presents the general structure of the network components particularly in this study. In the water network, storage tanks and water pumps represent demand and supply nodes, respectively, and water pipelines are considered as the links. For the power network, substations are demand nodes, and gate stations are supply nodes, where the power transmission lines are the links. For the gas network, pipelines and gas distribution stations represent links and nodes, respectively.

Table 1. General formation of the water, gas, and power networks components in Shelby County, TN.

(adapted from (Ghorbani Renani et al., 2019)**)**

Network	Nodes	Supply nodes	Demand nodes	Links
Water	49	34	15	71
Gas	16	3	13	17
Power	60	37	9	76

From figures 3, 4 & 5 show the water, gas and power networks represented with links. The size of the nodes is directly proportional to the degree and the color codes are as follows, blue represents water, green represents gas, red represents power and the grey is used for nodes with degree >= 4 in all the networks.

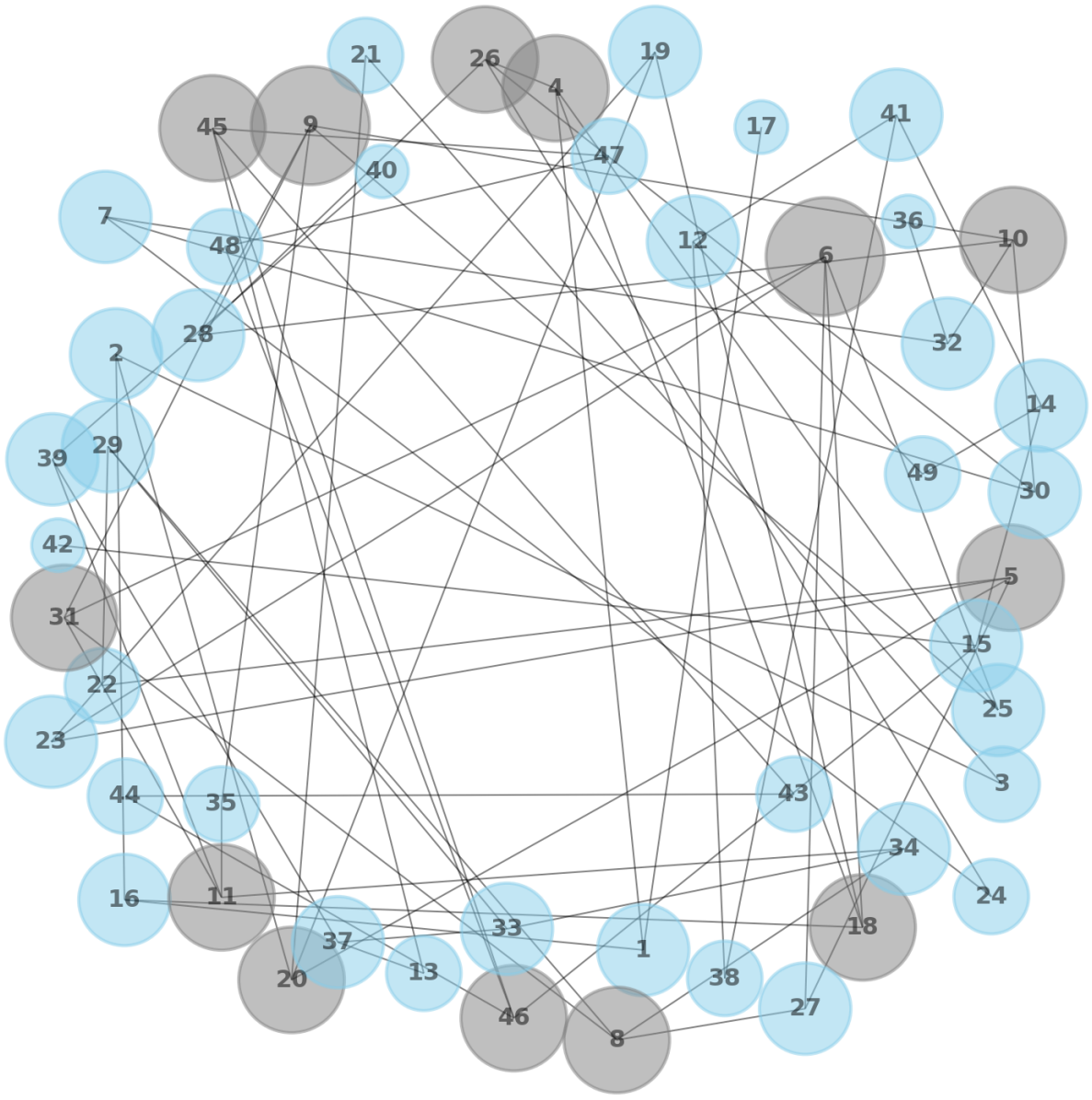

Figure 3. Water network with 49 nodes and 71 links.

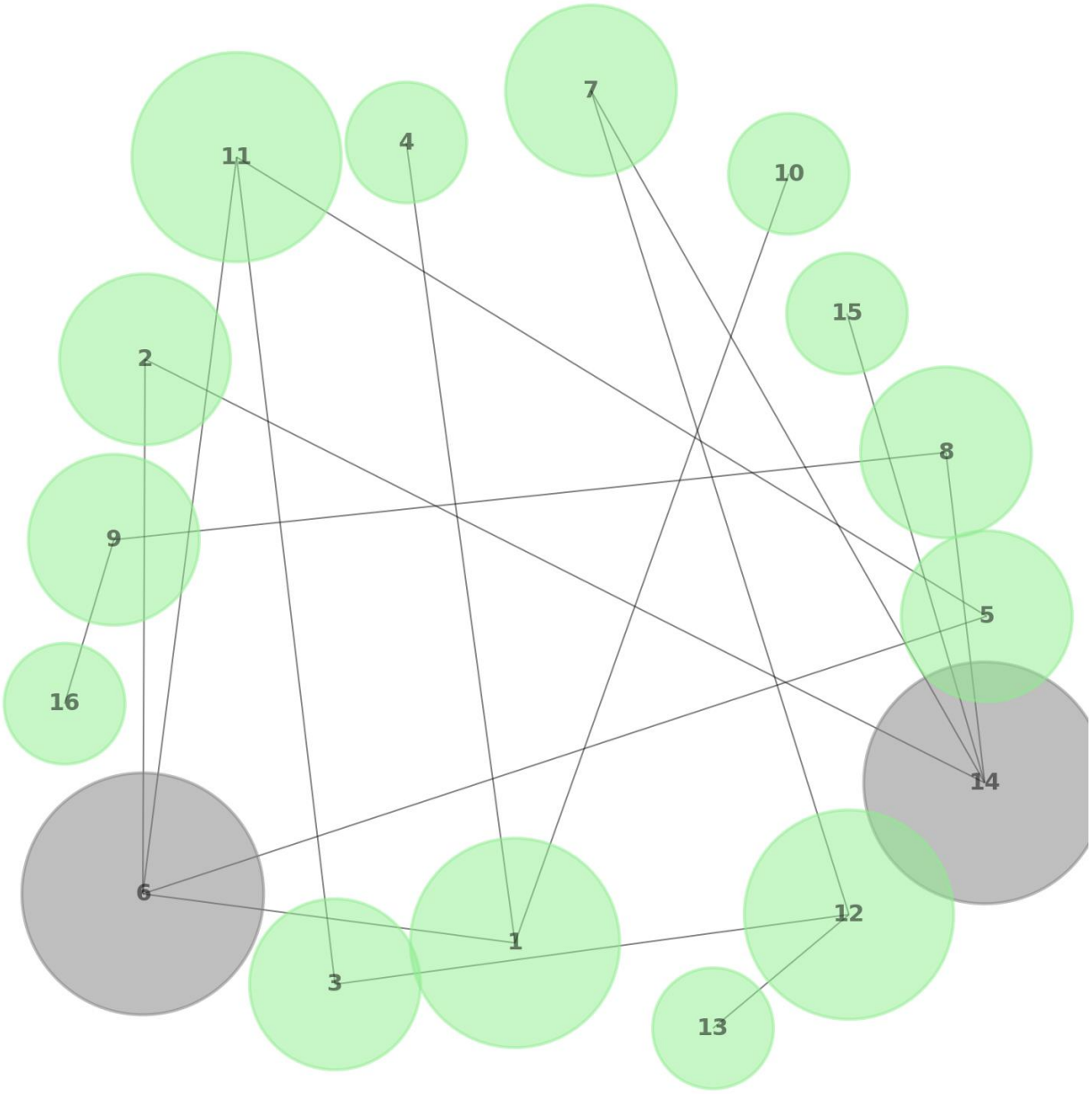

Figure 4. Gas network with 16 nodes and 17 links.

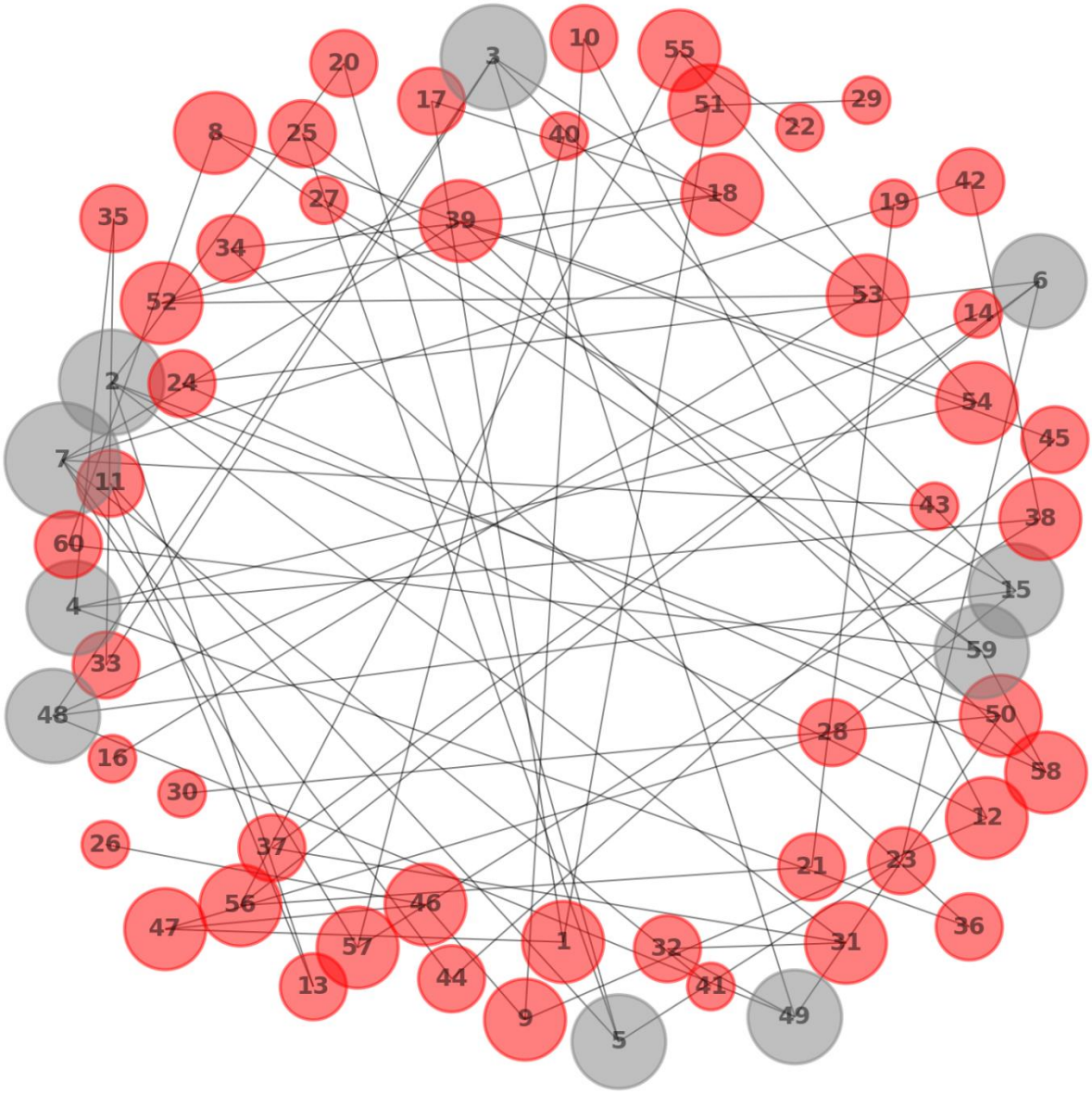

Figure 5. Power network with 60 nodes and 76 links.

3.1.2. Model assumptions

The proposed optimization model has the following underlying assumptions associated with the structure and operation of the system of networks and their interdependencies, and recovery, among others.

- Each infrastructure network consists of a set of nodes (including supply, demand and transshipment nodes) connected by a set of links, such that each supply node, demand node, and link have known supply capacity, demand, and flow capacity, respectively.
- There are work crews (work groups) responsible for repairing disrupted components in each infrastructure network, and this number can vary by network.
- A work crew can only work on a single disrupted component at a time. A disrupted component can be restored by a single work crew (once the disrupted component is assigned to them) until they become operational.
- There is a known restoration rate for each component, λ, representing the proportion of the component restoration per unit time by each work crew.
- Restoration time for each disrupted component is a function of both its failure and its restoration rate, both of which can vary by component.
- Infrastructure networks are physically interdependent such that every "parent" node must be operational for the dependent "child" nodes to be operational.

3.1.3. Notation

An undirected network is denoted by $G = (N, A)$, where N is the set of nodes, and A is the set of links. Assume a set K of networks, each with a set of nodes N^k such that $\bigcup_{k \in K} N^k = N$ and set of links A^k such that $\bigcup_{k \in K} A^k = A$. Nodes can consist of supply nodes ($N_s^k \subseteq N^k$), demand nodes ($N_d^k \subseteq N^k$) and transshipment nodes ($N^k \setminus \{N_d^k, N_s^k\}$) such that $N_s^k \cap N_d^k = \emptyset$. Sets $N'^k \subseteq N^k$ and

$A'^k \subseteq A^k$ are candidate nodes and links, respectively, in network $k \in K$, that can be interdicted or protected in the system of interdependent networks. Note that this model considers a single commodity flowing through each network, but it could be easily extended to a multicommodity model. Ψ represents interdependency among networks such that $\left((i,k),(\bar{\imath},\bar{k})\right) \in \Psi$ denotes node $i \in N^k$ in network $k \in K$ physically depends on node $\bar{\imath} \in N^{\bar{k}}$ in network $\bar{k} \in K$ where $N^k \cap N^{\bar{k}} = \emptyset$, $A^k \cap A^{\bar{k}} = \emptyset$ and $\forall\, k, \bar{k} \in K: k \neq \bar{k}$. Set R^k represents the available work crews in network $k \in K$. Index $t \in T$ provides the set of available time periods. Table 2 and

Table 3 outline the model parameters and decision variables, respectively.

Table 2. Model parameters.

$\eta^k_{it_e}$	Flow reaching node $i \in N^k_d$ in network $k \in K$ before the attack
w^k_i	Importance weight assigned to node $i \in N^k_d$ in network $k \in K$
S^k_i	Amount of supply in node $i \in N^k_s$ in network $k \in K$
d^k_i	Amount of demand in node $i \in N^k_d$ in network $k \in K$
u^k_{ij}	Capacity of link $(i,j) \in A^k$ in network $k \in K$
λ^k_{ij}	Restoration rate of the link $(i,j) \in A'^k$ in network $k \in K$
λ^k_i	Restoration rate of the node $i \in N'^k$ in network $k \in K$
E	An arbitrarily small positive number, $0 < \varepsilon < 1$
M	An arbitrarily large positive number

Table 3. Model decision variables.

η^k_{it}	Amount of demand met at node $i \in N^k_d$ in network $k \in K$ at time $t \in T$, continuous
x^k_{ijt}	Flow on link $(i,j) \in A^k$ in network $k \in K$ in time $t \in T$, continous
y^k_{ij}	Equal to 1 if link $(i,j) \in A'^k$ in network $k \in K$ is protected, binary
y^k_i	Equal to 1 if node $i \in N'^k$ in network $k \in K$ is protected, binary

z_{ij}^k	Equal to 1 if link $(i,j) \in A'^k$ in network $k \in K$ is interdicted, binary
z_i^k	Equal to 1 if node $i \in N'^k$ in network $k \in K$ is interdicted, binary
F_{ij}^k	Equal to 1 if link $(i,j) \in A'^k$ in network $k \in K$ is disrupted, binary
F_i^k	Equal to 1 if node $i \in N'^k$ in network $k \in K$ is disrupted, binary
α_{ij}^k	Equal to 1 if link $(i,j) \in A'^k$ in network $k \in K$ is operational, binary
α_i^k	Equal to 1 if node $i \in N'^k$ in network $k \in K$ is operational, binary
α'^{kr}_{ijt}	Equal to 1 if link $(i,j) \in A'^k$ in network $k \in K$ is restored by work crew $r \in R^k$ in time $t \in T$, binary
α'^{kr}_{it}	Equal to 1 if node $i \in N'^k$ in network $k \in K$ is restored by work crew $r \in R^k$ in time $t \in T$, binary
β_{ijt}^k	Equal to 1 if link $(i,j) \in A'^k$ in network $k \in K$ is reactivated at time $t \in T$, binary
β_{it}^k	Equal to 1 if node $i \in N'^k$ in network $k \in K$ is reactivated at time $t \in T$, binary

3.1.4 Objective function

In particular, the model aims to minimize the (weighted) fraction of unsupplied demand over the planning horizon. For this purpose, let us define $\zeta(t)$ as the weighted proportion of unmet demand (relative to the met demand before the disruption) at time t, as shown in Eq. (1), where η_{it}^k represents demand being met at node i in network k at time t, $\eta_{it_e}^k$ represents the amount of demand met prior to the disruption at time t_e (from Figure 1), and w_i^k is the relative importance of node i in network k. As mentioned in Section 1, network performance at time t is represented with $\varphi(t)$ which equals to $1 - \zeta(t)$.

$$\zeta(t) = \frac{\sum_{i \in N_d^k} \sum_{k \in K} w_i^k \left(\eta_{it_e}^k - \eta_{it}^k\right)}{\sum_{i \in N_d^k} \sum_{k \in K} w_i^k \eta_{it_e}^k} \quad (1)$$

Then, the proposed objective function, which seeks to minimize the cumulative weighted fraction of unsupplied demand over the planning horizon for the worst-case disruption scenario, would be defined in Eq. (2).

$$\xi = \min_{\eta,x,F,\alpha,\alpha',\beta} \sum_{t\in T} \zeta(t) \tag{2}$$

Constraints (3)-(6) present the nature of the first and second level decision variables.

$$y_{ij}^{k} \in \{0,1\} \qquad \forall\, (i,j) \in A'^{k}\,, \forall\, k \in K \tag{3}$$

$$y_{i}^{k} \in \{0,1\} \qquad \forall\, i \in N'^{k}\,, \forall\, k \in K \tag{4}$$

$$z_{ij}^{k} \in \{0,1\} \qquad \forall\, (i,j) \in A'^{k}\,, \forall\, k \in K \tag{5}$$

$$z_{i}^{k} \in \{0,1\} \qquad \forall\, i \in N'^{k}\,, \forall\, k \in K \tag{6}$$

3.1.6. Restoration Level

This level of the model is related to the second defender actions to plan the restoration of interdicted components. In this level, the failure of each disrupted component and its operationality status is determined for the recovery process. Flow balance constraints enable decision variables in this restoration level to connect with the objective function. Therefore, restoration scheduling is automatically set to return the system of networks to a stable operation as rapidly as possible.

The restrictions associated with the restoration level comprise constraints (3)-(44). Constraints (3) and (4) generally deliver failure status for candidate links $A'^k \subseteq A^k$ and candidate nodes $N'^k \subseteq N^k$ respectively, depending on the protection and interdiction strategies. Constraints (5)-(8) determine the operationality status of each link and node based on their failure, where the binary variable α is 1 when its associated link or node is operational (either protected or not interdicted) after the disruption. Constraints (9) and (10) ensure that operational links and nodes, respectively, are not restored. Constraints (11) and (12) state that nonoperational links and nodes cannot be functional at period 1, since they require at least one time unit to be reactivated. Constraints (13)-(16) represent the flow balance constraints at node $i \in N^k$ in infrastructure network $k \in K$ at time $t \in T$. Constraints (17) represent the capacity restriction for each link $(i,j) \in A^k$. Constraints (18)-(20) ensure that a positive flow through any given link can be attained at a period $t \in T$ only if such a link, along with its starting and ending nodes, are operational or were already recovered. Constraints (21) ensure that the restoration task of a disrupted link is continued without interruption once it has commenced. Likewise, constraints (24) accomplish this for a disrupted node. Constraints (22) and (23) calculate the total time that a specific work crew should be assigned to restore a nonoperational link. Note that constraints (22) and (23) together help to deliver integer value for the total required recovery time of an element. Similarly, constraints (25) and (26) calculate this time for a nonoperational disrupted node. Constraints (27) and (28) ensure that once the nonoperational disrupted link and node, respectively, are fully restored at time $t \in T$, they are labeled reactivated from the next period ($t+1 \in T$) to the end of the time horizon of the model. Constraints (29) ensure that once the restoration of a disrupted link commences by a work crew at time $t \in T$, that specific work crew completes restoration of that link. Similarly, constraints (30) accomplish this same restriction for the disrupted nodes. Constraints (31) and (32) state that, at the

given time $t \in T$, only one work crew can work on the restoration task of a specific nonoperational disrupted link or node, respectively. Constraints (33) ensure that, at the given time $t \in T$, only one nonoperational disrupted component can be restored by a given work crew. Constraints (34) establish the interdependency among networks, ensuring that the positive flow through a link can be only available if their corresponding related parent nodes (in other networks) are operational. Finally, constraints (35)-(44) represent the nature of the decision variables for the restoration level.

$$y_{ij}^k + \left(1 - z_{ij}^k\right)\left(1 - y_{ij}^k\right) = 1 - F_{ij}^k \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{3}$$

$$y_i^k + \left(1 - z_i^k\right)\left(1 - y_i^k\right) = 1 - F_i^k \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{4}$$

$$\alpha_{ij}^k \le 1 - F_{ij}^k \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{5}$$

$$\alpha_{ij}^k + F_{ij}^k \ge \varepsilon \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{6}$$

$$\alpha_i^k \le 1 - F_i^k \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{7}$$

$$\alpha_i^k + F_i^k \ge \varepsilon \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{8}$$

$$\alpha_{ij}^k \le 1 - \beta_{ijt}^k \qquad \forall\, (i,j) \in A'^k\,, \forall t \in T, \forall\, k \in K \tag{9}$$

$$\alpha_i^k \le 1 - \beta_{it}^k \qquad \forall\, i \in N'^k, \forall t \in T\,, \forall\, k \in K \tag{10}$$

$$\beta_{ij1}^k = 0 \qquad \forall\, (i,j) \in A'^k, \forall\, k \in K \tag{11}$$

$$\beta_{i1}^k = 0 \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{12}$$

$$\sum_{(i,j)\in A^k} x_{ijt}^k - \sum_{(j,i)\in A^k} x_{jit}^k \le S_i^k \qquad \forall\, i \in N_s^k\,, \forall\, t \in T, \forall\, k \in K \tag{13}$$

$$\sum_{(i,j)\in A^k} x_{ijt}^k - \sum_{(j,i)\in A^k} x_{jit}^k = 0 \qquad \forall\, i\ \in N^k \backslash \left\{N_d^k, N_s^k\right\}, \forall\, t \in T, \forall\, k \in K \tag{14}$$

$$\sum_{(i,j)\in A^k} x_{ijt}^k - \sum_{(j,i)\in A^k} x_{jit}^k = -\eta_{it}^k \qquad \forall\, i \in N_d^k\,, \forall\, t \in T, \forall\, k \in K \tag{15}$$

$$\eta_{it}^k \le d_i^k \qquad \forall\, i \in N_d^k\,, \forall\, t \in T, \forall\, k \in K \tag{16}$$

$$x_{ijt}^k \le u_{ij}^k \qquad \forall\, (i,j) \in A^k\,, \forall\, t \in T, \forall\, k \in K \tag{17}$$

$$x_{ijt}^k \le u_{ij}^k\left(\, \alpha_{ij}^k + \beta_{ijt}^k \,\right) \qquad \forall\, (i,j) \in A'^k\,, \forall\, t \in T, \forall\, k \in K \tag{18}$$

$$x_{ijt}^k \le u_{ij}^k\left(\, \alpha_{i}^k + \beta_{it}^k \right) \qquad \forall\, (i,j) \in A^k\,, \forall\, i \in N'^k, \forall\, t \in T, \forall\, k \in K \tag{19}$$

$$x_{ijt}^k \le u_{ij}^k\left(\, \alpha_{j}^k + \beta_{jt}^k \right) \qquad \forall\, (i,j) \in A^k\,, \forall\, j \in N'^k, \forall\, t \in T, \forall\, k \in K \tag{20}$$

$$\sum_{s=1}^{t} \alpha'^{kr}_{ijs} \le M\,\left(1 - \left(\alpha'^{kr}_{ij,t+1} - \alpha'^{kr}_{ijt}\right)\right) \qquad \forall\, (i,j) \in A'^k\,, \forall\, t \in T\,, \forall\, k \in K, \forall\, r\ \in\ R^k \tag{21}$$

$$\sum_{r\,\in\,R^k}\ \sum_{t\in T} \alpha'^{kr}_{ijt} \ge \frac{F_{ij}^k}{\lambda_{ij}^k} - M\alpha_{ij}^k \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{22}$$

$$\sum_{r\,\in\,R^k}\ \sum_{t\in T} \alpha'^{kr}_{ijt} < \left(\frac{F_{ij}^k}{\lambda_{ij}^k} + 1\right) + M\alpha_{ij}^k \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{23}$$

$$\sum_{s=1}^{t} \alpha'^{kr}_{is} \leq M\left(1-\left(\alpha'^{kr}_{i,t+1}-\alpha'^{kr}_{it}\right)\right) \quad \forall\, i \in N'^{k}, \forall\, t \in T, \forall\, k \in K, \forall\, r \in R^{k} \tag{24}$$

$$\sum_{r \in R^{k}} \sum_{t \in T} \alpha'^{kr}_{it} \geq \frac{F_i^k}{\lambda_i^k} - M\alpha_i^k \quad \forall\, i \in N'^{k}, \forall\, k \in K \tag{25}$$

$$\sum_{r \in R^{k}} \sum_{t \in T} \alpha'^{kr}_{it} < \left(\frac{F_i^k}{\lambda_i^k}+1\right) + M\alpha_i^k \quad \forall\, i \in N'^{k}, \forall\, k \in K \tag{26}$$

$$\frac{\sum_{r \in R^{k}} \sum_{s=1}^{t-1} \alpha'^{kr}_{ijs}}{F_{ij}^k/\lambda_{ij}^k} \geq \beta_{ijt}^k \quad \forall\, (i,j) \in A'^{k}, \forall\, t \in T \mid t \neq 1, \forall\, k \in K \tag{27}$$

$$\frac{\sum_{r \in R^{k}} \sum_{s=1}^{t-1} \alpha'^{kr}_{is}}{F_i^k/\lambda_i^k} \geq \beta_{it}^k \quad \forall\, i \in N'^{k}, \forall\, t \in T \mid t \neq 1, \forall\, k \in K \tag{28}$$

$$\sum_{\substack{s \in R^{k} \\ s \neq r}} \sum_{t \in T} \alpha'^{ks}_{ijt} \leq M\left(1-\alpha'^{kr}_{ijt}\right) \quad \forall\, (i,j) \in A'^{k}, \forall\, t \in T, \forall\, k \in K, \forall\, r \in R^{k} \tag{29}$$

$$\sum_{\substack{s \in R^{k} \\ s \neq r}} \sum_{t \in T} \alpha'^{ks}_{it} \leq M\left(1-\alpha'^{kr}_{it}\right) \quad \forall\, i \in N'^{k}, \forall\, t \in T, \forall\, k \in K, \forall\, r \in R^{k} \tag{30}$$

$$\sum_{r \in R^{k}} \alpha'^{kr}_{ijt} \leq 1 \quad \forall\, (i,j) \in A'^{k}, \forall\, t \in T, \forall\, k \in K \tag{31}$$

$$\sum_{r \in R^{k}} \alpha'^{kr}_{it} \leq 1 \quad \forall\, i \in N'^{k}, \forall\, t \in T, \forall\, k \in K \tag{32}$$

$$\sum_{(i,j) \in A'^{k}} \alpha'^{kr}_{ijt} + \sum_{i \in N'^{k}} \alpha'^{kr}_{it} \leq 1 \quad \forall\, t \in T, \forall\, k \in K, \forall\, r \in R^{k} \tag{33}$$

$$x_{ijt}^{k} \leq u_{ij}^{k}\left(\alpha_{\bar{\imath}}^{\bar{k}} + \beta_{\bar{\imath}t}^{\bar{k}}\right) \quad \forall (i,j) \in A^{k}, \forall k, \bar{k} \in K, \forall \bar{\imath} \in N'^{\bar{k}} \mid \left((i,k),(\bar{\imath},\bar{k})\right) \in \Psi \text{ or } \left((j,k),(\bar{\imath},\bar{k})\right) \in \Psi, \forall t \in T \tag{34}$$

$$\eta_{it}^{k} \geq 0 \quad \forall i \in N_{d}^{k}, \forall t \in T, \forall k \in K \tag{35}$$

$$x_{ijt}^{k} \geq 0 \quad \forall (i,j) \in A^{k}, \forall t \in T, \forall k \in K \tag{36}$$

$$F_{ij}^{k} \in \{0,1\} \quad \forall (i,j) \in A'^{k}, \forall k \in K \tag{37}$$

$$F_{i}^{k} \in \{0,1\} \quad \forall i \in N'^{k}, \forall k \in K \tag{38}$$

$$\alpha_{ij}^{k} \in \{0,1\} \quad \forall (i,j) \in A'^{k}, \forall k \in K \tag{39}$$

$$\alpha_{i}^{k} \in \{0,1\} \quad \forall i \in N'^{k}, \forall k \in K \tag{40}$$

$$\alpha'^{kr}_{ijt} \in \{0,1\} \quad \forall (i,j) \in A'^{k}, \forall t \in T, \forall k \in K, \forall r \in R^{k} \tag{41}$$

$$\alpha'^{kr}_{it} \in \{0,1\} \quad \forall i \in N'^{k}, \forall t \in T, \forall k \in K, \forall r \in R^{k} \tag{42}$$

$$\beta_{ijt}^{k} \in \{0,1\} \quad \forall (i,j) \in A'^{k}, \forall t \in T, \forall k \in K \tag{43}$$

$$\beta_{it}^{k} \in \{0,1\} \quad \forall i \in N'^{k}, \forall t \in T, \forall k \in K \tag{44}$$

Before implementing the proposed solution algorithm, nonlinear constraints (3), (4), (27), and (28) are linearized. To linearize constraints (3) and (4), we define three different sets of linear constraints (Taha, 1998). Following this idea, constraint (3) can be replaced by constraints (45)-(47).

$$y_{ij}^k + (1 - z_{ij}^k) + F_{ij}^k \geq 1 \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{45}$$

$$F_{ij}^k \leq z_{ij}^k \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{46}$$

$$y_{ij}^k \leq 1 - F_{ij}^k \qquad \forall\, (i,j) \in A'^k\,, \forall\, k \in K \tag{47}$$

Similarly, constraint (4) can be replaced by constraints (48)-(50).

$$y_i^k + (1 - z_i^k) + F_i^k \geq 1 \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{48}$$

$$F_i^k \leq z_i^k \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{49}$$

$$y_i^k \leq 1 - F_i^k \qquad \forall\, i \in N'^k\,, \forall\, k \in K \tag{50}$$

Finally, by adopting the big-M method (Taha, 1998), constraints (27) and (28) can be replaced by constraints (51) and (52).

$$1 - \left(\frac{\frac{F_{ij}^k}{\lambda_{ij}^k} - \sum_{r \in R^k} \sum_{s=1}^{t-1} \alpha'^{kr}_{ijs}}{M} \right) \geq \beta_{ijt}^k \qquad \forall\, (i,j) \in A'^k, \forall\, t \in T \mid t \neq 1\,, \forall\, k \in K \tag{51}$$

$$1 - \left(\frac{\frac{F_i^k}{\lambda_i^k} - \sum_{r \in R^k} \sum_{s=1}^{t-1} \alpha'^{kr}_{is}}{M} \right) \geq \beta_{it}^k \qquad \forall\, i \in N'^k, \forall\, t \in T \mid t \neq 1, \forall\, k \in K \tag{52}$$

Parameter M in constraints (21)-(26) and (51)-(52) only needs to be greater than the maximum required time for restoring the disrupted components.

3.2 Data generation

3.2.1 Understanding the data of Approach I

The optimization model is modified to run multiple times and each time the nodes attacked are randomized and the restoration rates for each node are also randomized mostly between [0.7-0.99]. The time taken for the restoration of the entire network is also noted. As mentioned earlier the total number of nodes in the network are 125, (water network – 49, gas network – 16 and power network – 60), each row in the data consists of 251 data cells, that is 125 cells with nodes' data of whether a node is attacked (represented by 1) and or not (represented by 0), 125 cells of restoration rates of these nodes in the same order and the last cell represents the time for total restoration of the network. Below is a table of how the data set of size n*251 looks.

Table 4. Template of the data generated from the optimization code.

Node1	Node2	…	Node125	Rest_rate1	Rest_rate2	…	Rest_rate125	Time
Row 1								
Row 2								
…								
Row n								

The column names used are in the format: nwx_ny that represents network x node y. So, the column names are as follows, nw1_n1 to nw1_n49 for network 1 (that is the water network), nw2_n1 to nw2_n16 for network 2 (that is the gas network), nw3_n1 to nw3_n60 for network 3

(that is the power network). Similarly, for the columns with restoration rates, rr_nwx_ny represents the restoration rate for node y of network x. The snaps of the actual data are shown below.

The entire data set can be found at https://sooners-my.sharepoint.com/:f:/g/personal/ghaneshvar_ou_edu/EocFD7mhoEpEtMVVyBNsicYBIlHk24Q54mHEOF5wl01RTA?e=YUeYXQ in the folder Data > Original Data.

Table 5. An extract of the data generated with the status of the nodes.

nw1_n1	nw1_n2	nw1_n3	nw1_n4	nw1_n5	nw1_n6	nw1_n7	nw1_n8	nw1_n9
0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1
0	0	1	0	0	0	0	0	0
0	0	1	0	0	1	1	1	0
0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	1	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	1	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0

Table 6. An extract of the data generated with restoration rates along with the target value, time, in seconds.

rr_nw3_n53	rr_nw3_n54	rr_nw3_n55	rr_nw3_n56	rr_nw3_n57	rr_nw3_n58	rr_nw3_n59	rr_nw3_n60	time
0.81	0.77	0.73	0.82	0.75	0.84	0.7	0.7	9
0.91	0.76	0.86	0.76	0.83	0.76	0.91	0.82	29
0.87	0.82	0.92	0.91	0.8	0.97	0.93	0.94	17
0.82	0.96	0.91	0.84	0.97	0.92	0.92	0.97	33
0.8	0.83	0.73	0.88	0.75	0.85	0.71	0.77	9
0.63	0.53	0.7	0.74	0.51	0.56	0.72	0.59	11
0.74	0.94	0.66	0.83	0.77	0.93	0.93	0.91	9
0.74	0.84	0.73	0.82	0.77	0.85	0.7	0.8	9
0.79	0.89	0.89	0.77	0.86	0.75	0.81	0.76	15
0.94	0.88	0.85	0.95	0.91	0.82	0.88	0.87	23
0.75	0.87	0.89	0.87	0.86	0.91	0.75	0.83	29
0.96	0.93	0.98	0.97	0.89	0.88	0.92	0.9	23
0.76	0.8	0.86	0.88	0.9	0.92	0.85	0.94	25
0.85	0.83	0.87	0.78	0.86	0.72	0.7	0.82	11
0.37	0.21	0.46	0.33	0.44	0.33	0.49	0.39	7
0.76	0.82	0.7	0.75	0.75	0.83	0.77	0.89	13
0.85	0.97	0.97	0.93	0.96	0.97	0.84	0.88	35
0.28	0.47	0.47	0.27	0.38	0.37	0.24	0.48	10
0.65	0.83	0.62	0.72	0.61	0.88	0.73	0.64	13
0.76	0.78	0.75	0.89	0.81	0.76	0.87	0.84	9
0.81	0.82	0.77	0.85	0.85	0.7	0.82	0.79	27
0.7	0.75	0.84	0.86	0.82	0.81	0.87	0.81	13
0.88	0.85	0.74	0.74	0.8	0.75	0.7	0.76	11
0.88	0.83	0.87	0.71	0.8	0.88	0.84	0.71	7
0.87	0.76	0.8	0.77	0.81	0.78	0.85	0.84	21
0.84	0.75	0.72	0.8	0.85	0.79	0.78	0.81	9
0.86	0.75	0.85	0.87	0.83	0.94	0.87	0.84	21

Now investigating the distribution of number of nodes attacked (for the 10,000 iterations considered for analysis) in water, gas and power networks individually and the whole interdependent network.

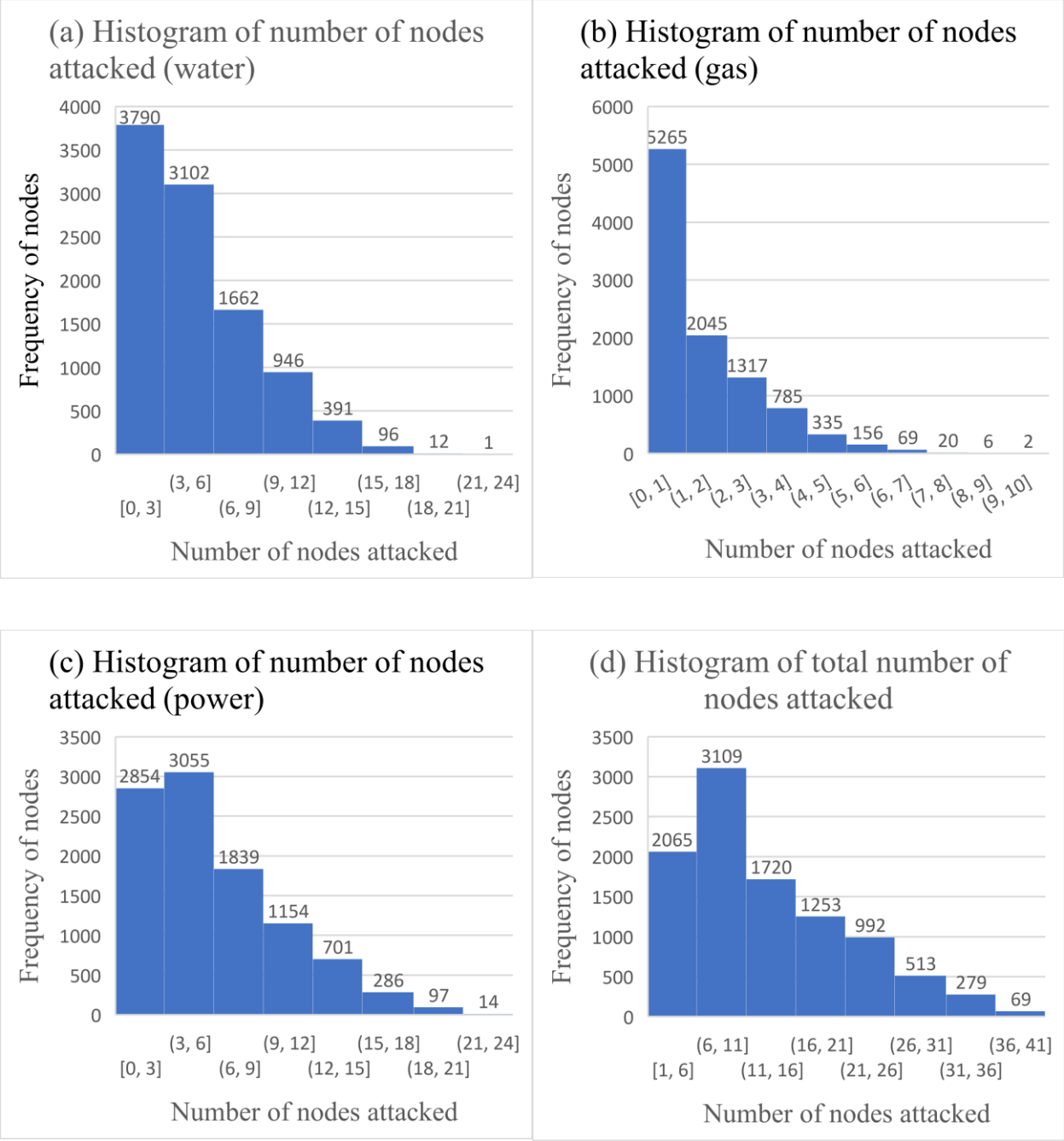

Figure 6. Histograms of number of nodes attacked for a) water, b) gas, c) power networks and d) the whole network.

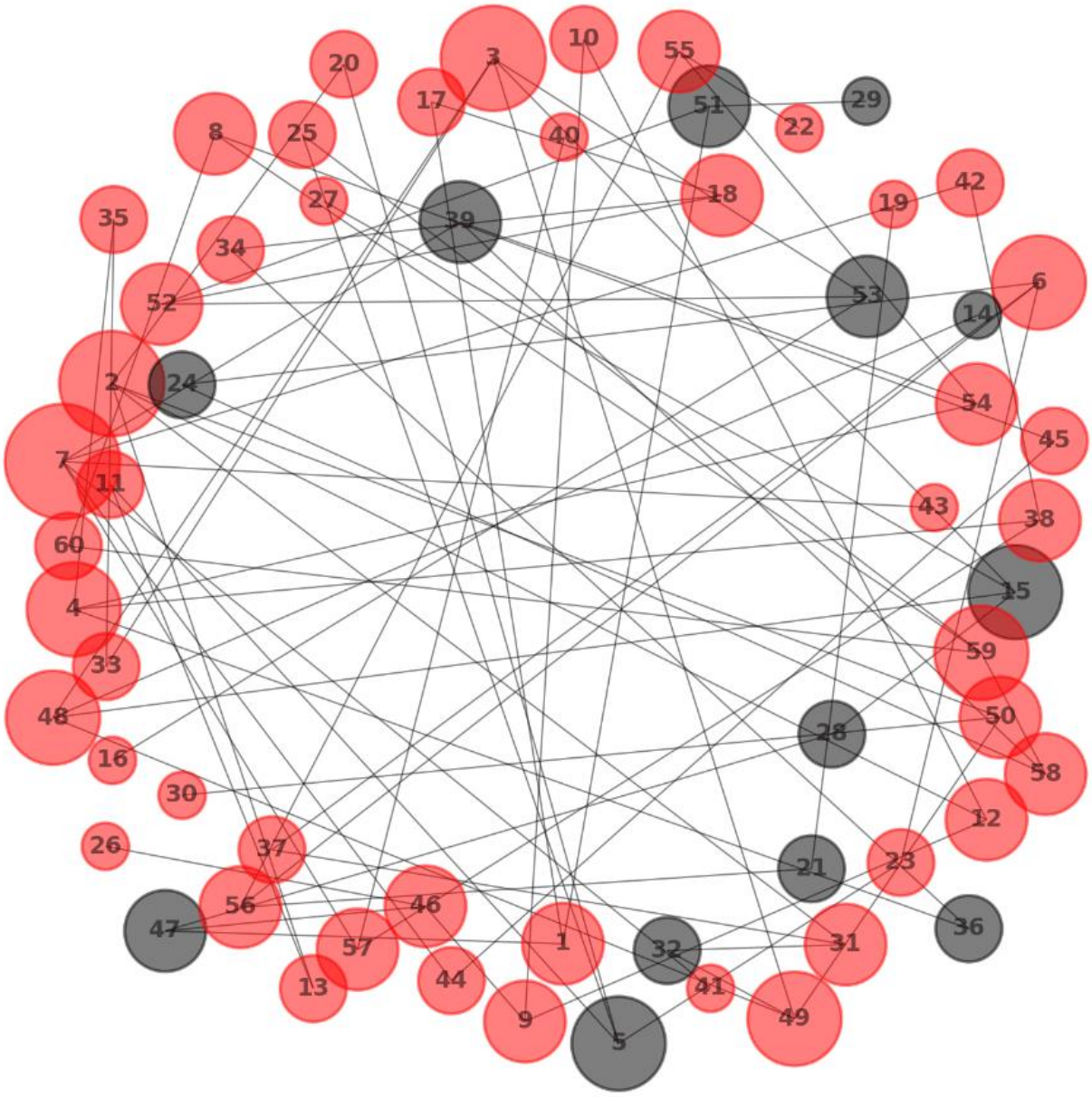

Figure 7. Screenshot of one of the frames of the GIF created for the power network.

To represent the randomness of the nodes selected for being attacked, GIFs were created from 100 iterations selected at random from the data set, as shown in figure 9. The gray color indicate that those nodes have been attacked.

3.2.2 Addition of features, extension of Approach I

A few features were added to the existing data as a part of feature engineering. These features include 3 columns that represent the total number of nodes attacked in water, gas and power networks individually and these columns are named as ‘total_nw1’, ‘total_nw2’ and ‘total_nw3’ respectively. A column that represents the total number of nodes attacked in the whole network is also added and named as ‘total’. Addition of these four features improved the prediction accuracy and the comparison for one of the predictive methods is discussed in Section 4.

3.2.3 Approach II

Hence, an alternate method was thought of and implemented. This method includes formatting of the data obtained from the optimization code. It includes the multiplication of the first 125 columns with the next 125 columns, that is the AND product of the node attack data and the restoration rate data of the respective columns. The time column remains unchanged. This results in a data set with 126 columns having restoration rate of an attacked nodes and zero otherwise, with last column still representing time. The entire data set can be found at https://sooners-my.sharepoint.com/:f:/g/personal/ghaneshvar_ou_edu/EocFD7mhoEpEtMVVyBNsicYBIlHk24Q54mHEOF5wl01RTA?e=YUeYXQ in the folder Data > Formatted Data.

This approach helps n keeping only the relevant data required for the prediction, that is the restoration rates of only the nodes attacked and the rest are zeros. An advantage of using this approach is that the computational time taken to train the dataset obtained is considerably reduced, as the dataset size is reduced by half (from 251 columns to 126 columns). As the size of the train dataset is only 7,000 rows, the time difference between training the datasets from approach I and

approach II is in seconds. When larger datasets are trained, the time difference could be in hours or days.

Table 7. An extract of the dataset after formatting.

nw3_n53	nw3_n54	nw3_n55	nw3_n56	nw3_n57	nw3_n58	nw3_n59	nw3_n60	time
0	0	0	0	0	0	0	0	9
0	0.76	0	0	0	0	0	0	29
0	0	0	0	0	0.97	0	0	17
0.82	0	0	0.84	0	0	0.92	0	33
0	0	0	0	0	0.85	0	0	9
0	0	0	0	0	0	0	0	11
0	0	0	0	0.77	0	0	0	9
0	0	0	0	0	0	0	0	9
0	0	0	0	0	0	0	0.76	15
0	0	0	0	0	0	0	0	23
0.75	0	0	0.87	0	0	0	0	29
0	0	0	0	0	0	0	0	23
0	0.8	0	0	0.9	0	0.85	0	25
0	0	0	0	0	0	0	0	11
0	0	0	0	0	0	0	0	7
0	0.82	0	0	0	0	0	0	13
0	0	0	0.93	0	0.97	0	0	35
0	0	0	0	0	0	0	0	10
0	0	0	0	0	0	0	0.64	13
0	0	0.75	0	0	0	0	0	9
0	0	0	0	0	0	0	0	27
0	0	0	0	0	0	0	0.81	13
0	0	0	0	0	0	0	0	11
0	0	0	0	0	0	0	0	7
0.87	0	0	0	0	0.78	0.85	0	21
0	0	0	0	0	0	0	0	9
0	0	0	0	0	0	0	0	21
0.84	0	0.89	0	0	0	0	0	33
0	0	0	0	0.84	0	0	0	11
0	0	0	0	0	0	0.85	0	29
0	0.34	0	0	0.22	0	0	0	23

3.2.4 Limitations

As observed from the above histogram d) of figure 8, the maximum number of nodes attacked in the network is 40 (Figure, d)). This is because of the computational limits of the software package used to solve the optimization model, Gurobi. In order to increase the number of nodes considered to be attacked, the possible time required for the network (which is a parameter used in the code) to restore should also be increased. This increases the computational time exponentially. For example, keeping the possible time taken to restore to 75 seconds and the number of nodes attacked to 50 could possibly take more than 24 hours to execute just one iteration. Hence keeping in mind, the time constraint, maximum number of nodes attacked were restricted to 40.

Proportion of the number of rows of the data generated to the total possible combinations of the number of nodes attacked is almost 0. This is because there are 125 nodes and even with the consideration of fixing the number attacked nodes to 25 (out of 125), the total combinations are $1.3*10^{26}$ ($^{n}C_{r} = {}^{125}C_{25}$), which is a huge number. The total number of unique combinations (rows) possible would be $2^{125} = 4.2*10^{37}$. Whereas the size (rows) of the data generated is 10,000. Thus, the ratio is very minimal and hence a very small portion of the combinations are considered (at random) for the analysis. The time taken to create 10,000 rows of data is around 12 days.

3.3 Predictive modelling

For the predictive analysis, a data set of 10,000 rows and 251 columns (as described above) is generated. This data set is divided into train data set (70%, 7000 rows) and test data set (30%, 3000 rows).

Four different models were used to predict the time taken for the network to restore, namely, Random Forest Model, Gradient Boosting Model, Linear Model and Decision Trees Model. The predicted time values are the compared with actual values and the RMSE (root mean square error) and correlation are compared. RStudio software was used to make the predictive analysis and compare the results.

4. RESULTS

4.1 Discussion of approach I

On comparing the prediction results and validating (RMSE and correlation values) them with test data made using the Random Forest technique applied on the original data and the data after the addition of the 4 features (discussed in Section 3.2.2) produced from the optimization code, the following observation is made. The train dataset consists of 7,000 rows and the test data set consists of 3,000 rows.

Table 8. Comparison of RMSE and correlation for Random Forest model.

	Before feature engineering	After feature engineering
RMSE value	4.92	1.05
Correlation	0.81	0.99

Comparing and analyzing the four different predictive models, namely Random Forest, Gradient Boosting, Linear Model and Decision Trees that were used for predicting and comparing.

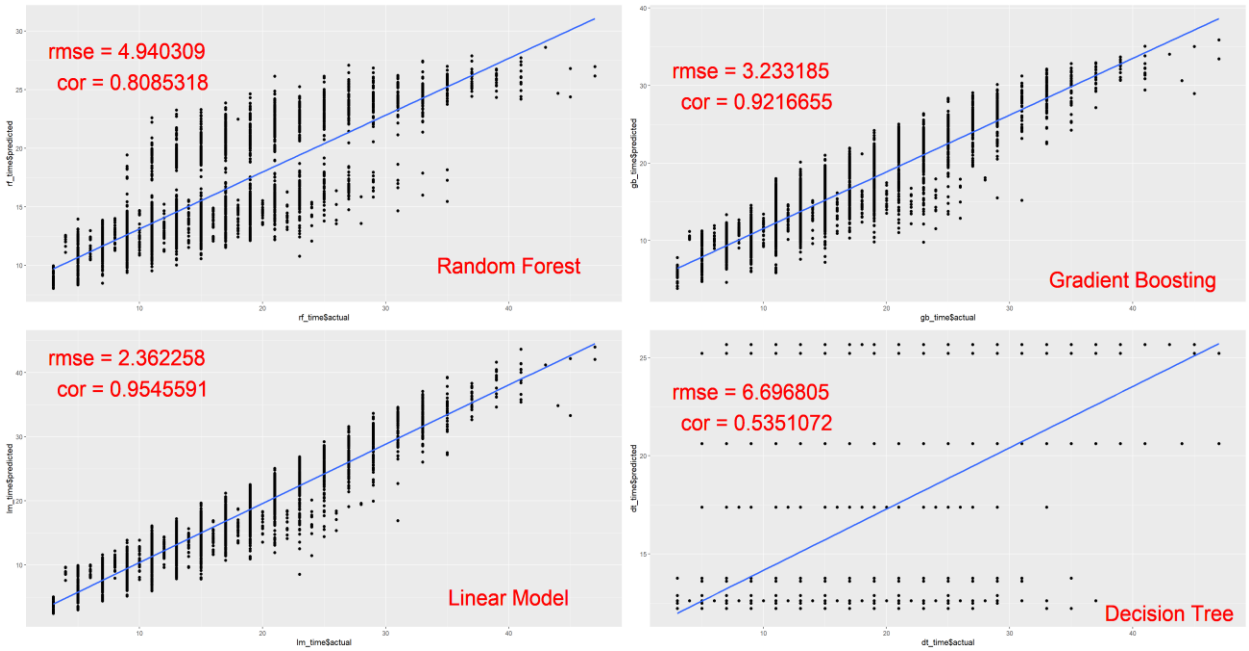

Figure 8. Comparison of the models: Random Forest, Gradient Boosting, Linear Model and Decision Tree for the original data.

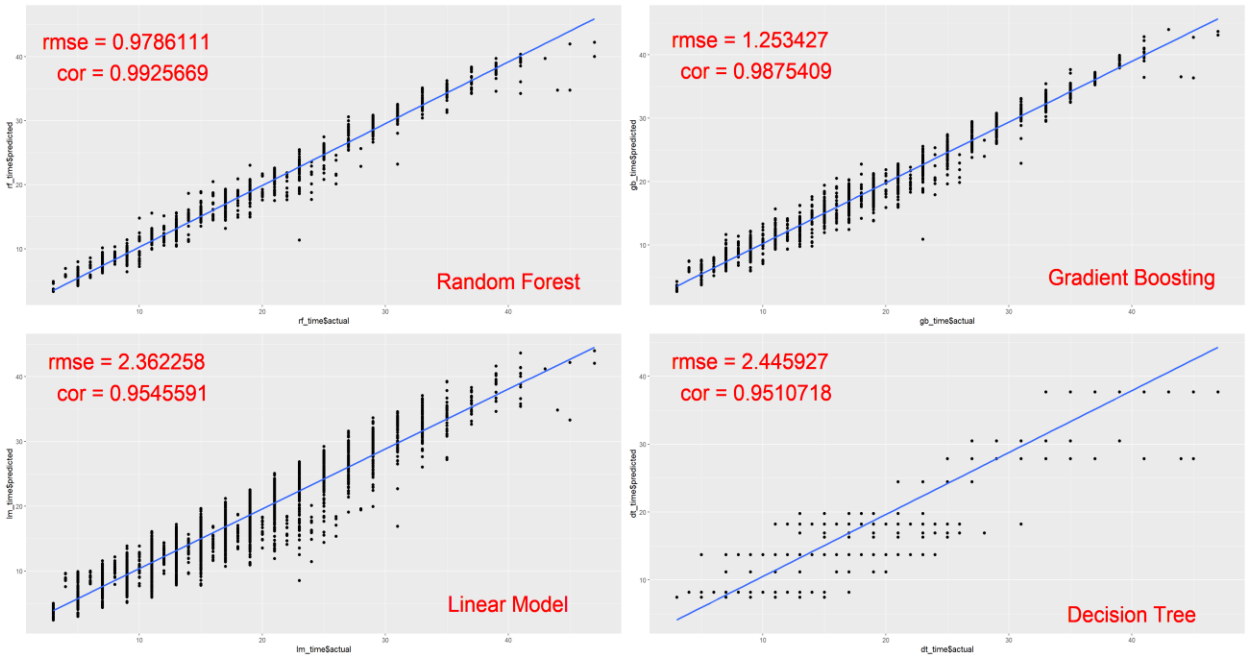

Figure 9. Comparison of the models: Random Forest, Gradient Boosting, Linear Model and Decision Tree after the addition of the four features.

It can be observed that there is clear improvement in the RMSE and correlation values of the prediction of time before and after adding the features.

An analysis was made on how the addition of the four features (representing the number of nodes attacked) impacted the predictions and it was observed that these features alone were able to predict with a correlation of 0.95 (for the Random Forest model). This was because of high restoration rates of the nodes that were produced randomly were limited mostly between 0.7 and 0.99. This means that all the nodes that fall under these restoration rates (0.5-0.99) take 2 seconds to recover (as 1/0.5 = 2 seconds and 1/0.99 = 1.01 ≈ 2 seconds). Thus, most of the nodes were assigned a restoration rate that results in 2 seconds to be repaired or restored. In that way the addition of features affects the prediction and do not help us in providing a value that is dependable despite having high accuracy.

4.2 Discussion of approach II

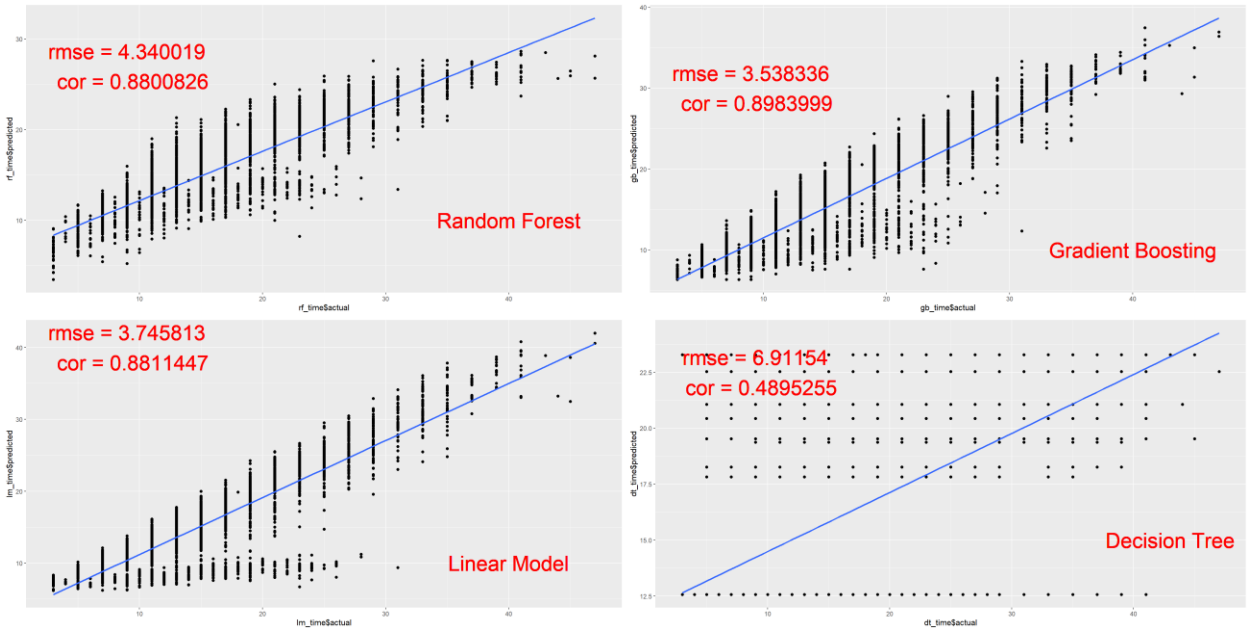

Figure 10. Comparison of the models: Random Forest, Gradient Boosting, Linear Model and Decision Tree for the formatted data.

This data generated using approach II was used for similar analysis and the results are compared as shown in figure 13.

4.3 Comparison of approach I & II

Table 9 shows the comparison of the three different data sets used for the predictive analysis.

Table 9. Comparison of the results

Predictive Method	Original data (Approach I)	After feature engineering	After data formatting (Approach II)
Random Forest	RMSE = 4.94 Correlation = 0.81	RMSE = 0.98 Correlation = 0.99	RMSE = 4.34 Correlation = 0.88
Gradient Boosting	RMSE = 3.23 Correlation = 0.92	RMSE = 1.25 Correlation = 0.99	RMSE = 3.54 Correlation = 0.90
Linear Model	RMSE = 2.36 Correlation = 0.95	RMSE = 2.36 Correlation = 0.95	RMSE = 3.75 Correlation = 0.88
Decision Tree	RMSE = 6.70 Correlation = 0.54	RMSE = 2.45 Correlation = 0.95	RMSE = 6.91 Correlation = 0.49

Comparing the original data (blue) and formatted data (green) predictions from Table 9, it is observed that the prediction from Random Forest model is better for the formatted data. Whereas for rest of the models, Gradient Boosting, Linear model and Decision Tree the original data yields better prediction.

5. CONCLUSION

It is important to know how resilient a system is as it defines the strength of the system to recover to its original functioning state and can be used as a method of assessing how strong it is towards a disruptive event. From resilience point of view, time taken by a system to recover from a disruptive event is one of the measures of evaluating the system's performance. This assessment can be used in taking precautionary measures towards the safety of the system.

The objective of this research is to be able to predict the time taken for the interdependent network to recover completely from an interdiction of a set randomly selected nodes. This is done by training the data obtained from an optimization model, whose objective is to minimize the recovery time given supply, demand, restoration rates and other constraints. The validation is made by comparing the prediction results of the test data.

In order to implement this concept, we present a case study based on the interdependent network consisting of water, gas, and power networks in Shelby County, TN, USA. The results show the predicted time taken for the network to recover completely to its original state and the comparison of the results obtained from different models used for the prediction.

From the comparison Table 9, it can be observed that the predictions for the data set with the addition of the four features (highlighted in yellow) are highly correlative compared to the predictions for original data (blue) and the formatted data (green). Meaning these four columns are dominating the prediction process due the restoration rate limitations as discussed in Section 4.1. Thus, these prediction results (in yellow) representing the data after feature engineering cannot be considered for comparison for this work.

Comparison of the predictions of approach I (blue) and approach II (green) from Table 9, shows that the prediction from Random Forest model is better for the formatted data, that is approach II. Whereas for rest of the models, Gradient Boosting, Linear model and Decision Tree the approach I yields better prediction.

Compelling conclusions cannot be made on which approach among approach I or approach II is better, or which predictive modelling gives better results compared to others as the dataset generated and used for training and predicting is only a very tiny fraction of what can be produced without limitations and hence the conclusions made on the results obtained are not substantial. But the objective of predicting the time taken for the network to recover is achieved and the validation results show the accuracy of the prediction.

6. FUTURE SCOPE

Due to the time and software constraints, the depth of the research was limited. The directions in which this work can be explored in a broader way is discussed in this section. The future work can be explored on the following:

- Focusing on the size of the data used for the analysis, bigger the data size, more the reliability of the analysis.
- Increasing the number of nodes considered to be attacked or interdicted (discussed in Section 3.2.3) will provide a data set that is less sparse when compared to lesser nodes considered.
- Decreasing the of restoration rates of the nodes, that is considering the nodes take long time to restore, for example consider a node with a restoration rate of 0.12, the node will need $1/0.12 = 8.3 \approx 9$ seconds to recover. This ensures the range of the time taken for the independent network to restore is high, thus the data is diversely spread.

7. REFERENCES

Danziger, M. M., Shekhtman, L. M., Bashan, A., Berezin, Y., & Havlin, S. (2016). Vulnerability of Interdependent Networks and Networks of Networks. In A. Garas (Ed.), *Interconnected Networks* (pp. 79–99). https://doi.org/10.1007/978-3-319-23947-7_5

Devanandham, H., & Ramirez-Marquez, J. E. (2012). Generic metrics and quantitative approaches for system resilience as a function of time. *Reliability Engineering and System Safety*, *99*, 114–122. https://doi.org/10.1016/j.ress.2011.09.002

Eusgeld, I., Nan, C., & Dietz, S. (2011). System-of-systems approach for interdependent critical infrastructures. *Reliability Engineering and System Safety*, *96*(6), 679–686. https://doi.org/10.1016/j.ress.2010.12.010

Ghorbani Renani, N., González, A., & Barker, K. (2019). Tri-Level Optimization for Enhancing Resilience in Interdependent Networks. *IISE Annual Conference & Expo 2019*. Retrieved from https://eventscribe.com/2019/IISE/fsPopup.asp?Mode=presInfo&PresentationID=549349

González, A. D., Dueñas-Osorio, L., Sánchez-Silva, M., & Medaglia, A. L. (2016). The Interdependent Network Design Problem for Optimal Infrastructure System Restoration. *Computer-Aided Civil and Infrastructure Engineering*, *31*(5), 334–350. https://doi.org/10.1111/mice.12171

Hosseini, S., Barker, K., & Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. *Reliability Engineering and System Safety*, *145*, 47–61. https://doi.org/10.1016/j.ress.2015.08.006

Hussein, A., Chehab, A., Kayssi, A., & Elhajj, I. H. (2018). Machine learning for network resilience: The start of a journey. *2018 5th International Conference on Software Defined Systems, SDS 2018*. https://doi.org/10.1109/SDS.2018.8370423

Kettl, D. F. (2013). *System Under Stress: Homeland Security and American Politics*. CQ Press.

Munoz, A. (2014). Machine Learning and Optimization. *Courant Institute of Mathematical Sciences*.

Taha, H. A. (1998). Operations Research: An Introduction. *Journal of Manufacturing Systems*, *17*(1), 78.

Wang, M., Cui, Y., Wang, X., Xiao, S., & Jiang, J. (2018). Machine Learning for Networking: Workflow, Advances and Opportunities. *IEEE Network*. https://doi.org/10.1109/MNET.2017.1700200

Wu, B., Tang, A., & Wu, J. (2016). Modeling cascading failures in interdependent infrastructures under terrorist attacks. *Reliability Engineering and System Safety*, *147*, 1–8. https://doi.org/10.1016/j.ress.2015.10.019

Zhang, Y., Yang, N., & Lall, U. (2016). Modeling and simulation of the vulnerability of interdependent power-water infrastructure networks to cascading failures. *Journal of Systems Science and Systems Engineering*, *25*(1), 102–118. https://doi.org/10.1007/s11518-016-5295-3