Modelling and Analysis of Critical Infrastructure for Situational Awareness Applications

Samir Puuska

Master's Thesis
UNIVERSITY OF HELSINKI
Department of Computer Science

Helsinki, 26th May 2016

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Laitos — Institution — Department Department of Computer Science

Tekijä — Författare — Author
Samir Puuska

Työn nimi — Arbetets titel — Title

Modelling and Analysis of Critical Infrastructure for Situational Awareness Applications

Oppiaine — Läroämne — Subject

Computer Science

Työn laji — Arbetets art — Level Master's Thesis	Aika — Datum — Month and year 26th May 2016	Sivumäärä — Sidoantal — Number of pages 46

Tiivistelmä — Referat — Abstract

Critical infrastructure forms an interdependent network, where individual infrastructure sectors depend on the availability of others in order to function. In such environment, faults easily propagate through the interlinked systems causing cascading failures. In order to effectively respond to incidents at national scale, it is necessary to maintain situational awareness by creating a common operational picture over all infrastructure sectors. A suitable way of modelling critical infrastructure and the interdependencies is required for building a system capable of delivering the needed information for obtaining robust situational awareness.

This thesis presents a model of critical infrastructure for national scale situational awareness applications, as well as analysis methods for estimating current and future infrastructure status. The model uses directed graphs in conjunction with finite state transducers to present dependencies and operational status of critical infrastructure systems. Analysis method utilising graph centrality measures was developed for quantifying both system specific and infrastructure wide impact of disruptions. Additionally, an entropy based analysis method was created for estimating operational status of infrastructure systems in situations, where current data is not available.

The electric grid and mobile networks of a coastal area of Finland were modelled using the presented methods. Dataset of system failures observed during a storm, in conjunction with simulation tools were used to evaluate the suitability of the framework for situational awareness tasks. Results indicate, that the proposed modelling and analysis methods are suitable for real time situational awareness applications.

Avainsanat — Nyckelord — Keywords
critical infrastructure, situational awareness, graph theory, automata theory, entropy

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpula Science Library

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction

In modern societies, people are surrounded by advanced networks of infrastructure systems providing the essential services for everyday life. Water, electricity and food distribution are just few of the things we rely on a daily basis, often without thinking. The essential systems necessary for the vital societal functions are called *critical infrastructure* (CI). Recent events, such as Fukushima Daiichi nuclear disaster (2011), India power blackouts (2012), and Duqu 2.0 APT campaign (2015), have shown that this infrastructure is relatively vulnerable to various natural and man-made effects[23]. The malfunction or destruction of this infrastructure or any of its subsystems may cause major economic losses, pose various hazards to the environment and people, and – in extreme cases – result in human casualties. These systems are interdependent in nature; a failure in one system may quickly affect the operation of other, connected systems.

The field of critical infrastructure protection in its modern form is relatively new. In its infancy the focus was almost exclusively on natural disasters and recovery procedures. The groundwork for a more holistic approach was laid during the 1990s, when various counter-terrorism related laws, acts, and directives were issued in the United States of America. The most notable directive concerning critical infrastructure was Presidential Decision Directive of 1998 (PDD-63), where the definition of critical infrastructure was revised and division to different sectors was defined[6]. The directive called for extensive CI vulnerability assessments, as well as creation of preventative measures and recovery plans. Additionally, the then relatively new, topics of cyber security and cyber threats were noted and recognised as a valid concern.

Although awareness about the risks existed before the 2000s, there was relatively little systematic academic interest towards man-made threats facing critical infrastructure. After the attacks of September 11, 2001 the situation changed. Resources and interest were directed to researching ways to protect critical infrastructure from various threats, often with considerable amount of governmental support. Since then, critical infrastructure and the protection of essential assets have become a major field of study[18]. This interest has partly been sparked by the apparent interdependent nature of critical infrastructure systems, which leaves critical infrastructure systems open to cascading failures where malfunctions and errors propagate through multiple infrastructure systems and layers[32].

1.1 Research Problems and Scope

Recent cyber threats and natural catastrophes have shown that our current infrastructure is highly vulnerable to various failures and malfunctions. The importance of critical infrastructure, and the observed fragility of it, have increased the demand for a more active monitoring both at industrial and governmental level.

The Finnish government has addressed this need by publishing a Security strategy for society 2010, a document detailing the preparedness levels and requirements for every layer of society[3]. In order to realise the end goals defined in the strategy, the Finnish government has launched and funded scientific research on critical infrastructure protection.

This thesis was written as a part of two such research projects: Digital Security of Critical Infrastructures (DiSCI), and Situational Awareness on Critical Infrastructure -project (VN TEAS)[2]. VN TEAS is a part of the Government's analysis, assessment and research activities, coordinated by the Prime Minister's Office of Finland[1]. The Finnish Funding Agency for Technology and Innovation funded the DiSCI reseach project.

DiSCI project aimed to find solutions for threats facing the CI at national level. One of the research goals was to create a demonstration environment of centralised common operational picture framework which supports situational awareness in complex networked environment. During the project, a set of requirements for real-time monitoring system for the critical infrastructure was developed. The Situational Awareness of Critical Infrastructure and Networks (SACIN) software framework was developed for evaluation of these CI monitoring concepts.

VN TEAS project investigates the underlying interdependencies in Finnish critical infrastructure, and their impact on its performance under both normal operation as well as in serious crises. A simulation model is created for analysing dependencies of electricity distribution and telecommunication networks in situations where multiple incidents affect their performance simultaneously.

In this thesis we propose a solution for nationwide CI modelling and analysis for common operating picture (COP) situational awareness purposes. The work detailing the software architecture and front-end components were published as separate articles[38,33].

The two key research questions examined in this thesis are presented below:

I. How to model critical infrastructure efficiently at national scale to support real-time SA applications?

II. What methods can be developed for analysing the CI to estimate current situation and predict future states?

We firstly collect and analyse the requirements for building a model and analysis framework for critical infrastructure, tailored for situational awareness applications. After constructing the model based on the requirements, a suitable set of analysis methods are developed to complement the model. Since situational awareness is highly time dependent and tied to the human element, the resulting methods and complexity must be adjusted for this particular application. Earlier versions of the proposed model presented in this thesis were published as separate original articles[29,30,16].

A set of requirements for Critical Infrastructure Monitoring Operator was developed during the DiSCI project[34]. The goal this thesis was to satisfy the essential functions and information requirements defined therein. Critical infrastructure model and a set of analysis methods were constructed based on these requirements. One important design aspect was to achieve real time analysis capability, needed for keeping pace with rapidly evolving conditions that may occur in infrastructure networks. The methods were evaluated by building a test system as a part of the SACIN framework. Finally, the core concepts were implemented as a part of VN TEAS simulation and analysis environment, where further evaluations could be made.

1.2 Related Work

Modelling and analysis of critical infrastructure is a notable field of contemporary research. Consequently, various different modelling and simulation approaches have been published in modern literature. Ouyang conducted an extensive review of state-of-the-art models for critical infrastructure in 2014[25]. The review covered many different modelling formalisms that have been employed for studying different aspects of CI functionality. The approaches range from purely empirical analysis to economic theory, network theory and system dynamics, among others more esoteric ones.

There has not been, however, much research on the real-time capable modelling and analysis frameworks for critical infrastructure. Especially from a situational awareness viewpoint, there exist a gap in research on this subject. The current models often require extensive information about the internal structure of CI systems and material or resource flows. As well as requiring substantial amount of information about the infrastructure systems and their dependencies, real-time performance has not been considered. The work presented in this thesis attempts to fill this gap by specifically considering the requirement to model a large number of systems while accounting for real-time aspects of SA. The limited availability of information on the internal operation of various CI systems has also been addressed.

1.3 Limitations

In this thesis we build a framework based on pre-collected requirements. The available datasets are used to draw conclusions whether or not the modelling formalisms and analysis methods are feasible. Since situational awareness necessarily includes humans, the actual impact to SA should be assessed by conducting user tests, where the methods are incorporated as part of the operators work flow. In such tests, other factors like user interface design also play a major role. The methods presented in this thesis should be seen as necessary, but not sufficient building blocks for a complete situational awareness system.

2 Background

This section describes the key concepts and definitions used in this thesis. Firstly, the definition and description of critical infrastructure is given. Secondly, the concept of sensor fusion is explored. Thirdly, the concept of situational awareness is presented. Finally, a brief overview of the main mathematical concepts is given.

2.1 Critical Infrastructure

Critical infrastructure (CI), as defined by the European Council, is

> an asset, system or part thereof located in Member States which is essential for the maintenance of vital societal functions, health, safety, security, economic or social well-being of people, and the disruption or destruction of which would have a significant impact in a Member State as a result of the failure to maintain those functions;[11].

Critical infrastructure can be divided into different sectors and layers according to their relative importance. One such system, proposed by Lewis, divides critical infrastructure into eleven sectors (Table 1)[18]. The sectors are arranged to form three levels, where higher tiers generally depend on lower ones to function, although interdependencies also occur between the layers (Figure 1).

Sector	Examples
Agriculture and food	Grocery stores, plantations
Public health	Hospitals and related services
Emergency services	Police, ambulance services, fire brigade
Defence industry	Ammunition, repair services, logistics
Telecommunications	Internet, phone lines, fibre lines
Energy & Power	Power plants and delivery systems, fuel resources
Transportation	Trains, buses, aeroplanes
Banking and finance	Commercial and (inter)governmental banks
Chemical industry	Fertiliser, disinfectants
Postal and shipping	Import and export, regular mail
Water	Fresh water delivery, waste water services

Table 1: List of critical infrastructure sectors, as defined by Lewis[18].

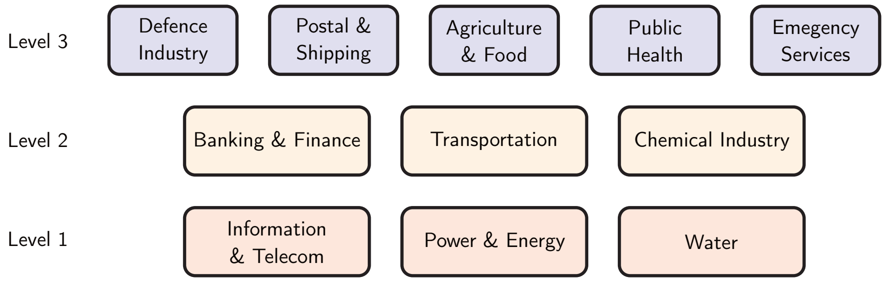


Figure 1: Three layers of interdependencies between critical infrastructure sectors, adapted from Lewis.[18, p. 57]

Critical infrastructure is usually described as highly interdependent[32]. Different infrastructure sectors depend on the availability of others in order to function. Due to this interconnectedness, cascading failures that span multiple infrastructure sectors are possible. Multi-system failures can cause large capacity shortages, leading to financial losses, equipment damage, and human casualties.

A study by Luiijf et al. on the interdependencies in European CI ecosystems found that 60% of all inter-sectional CI failures originated from the energy sector, 28% from telecommunication–internet sector, and the remaining 12% from other sectors[19]. It was also noted, that while CI is highly interdependent, the majority of failure routes were focussed and directional. This was due to the fact, that most sectors are dependent on the electric grid.

The failures may be caused by both man-made and natural disturbances. For example, storms or floods can cause major disruptions on large geographical area. One such case is the Finnish Tapio -storm, where, at its worst, over 300 000 customers of electricity companies were without power. The storm also affected water delivery and waste water processing systems, because they were not equipped with emergency power. Cellular networks also experienced coverage losses over wide areas.

Regional State Administrative Agency for Southwestern Finland compiled a detailed report on the Tapio storm[14]. The report states that the storm caused significant problems to the VIRVE network, a Terrestrial Trunked Radio -based telecommunication system used by the Finnish authorities. At the worst, 50% of the base stations were non-operational. This was a major problem especially for leadership at the operative level. Emergency power generators were available, but the effective locations for the devices were not known.

The Tapio storm caused one of the largest multi-sector service outage in recent history. The observations in the report were therefore extremely useful for guiding the requirement formulation.

2.2 Sensor Fusion

Sensor fusion refers to the process of combining the outputs of two or more sensors in order to gain information that would be impossible to obtain by just looking the outputs of the sensors separately. In a military setting, for example, combining the outputs of radar and optical sensors, enables identifying targets and their trajectories even when the individual sensor readings are of poor quality.

The Joint Directors of Laboratories (JDL) data fusion model is a framework for combining a set of procedures and algorithms for refining sensor data in order to improve current situational awareness. The JDL sensor fusion model was originally presented in 1988, and later refined in 1999[37]. The model consists of five different levels, which each refine and combine the data from previous levels in order to create more informed predictions and analyses (illustrated in Figure 2 and Table 2). Level 0 fusion process is responsible for aligning the raw input data into a common format. Level 1 combines the pre-processed data and identifies different objects, such as systems, attacks, and malfunctions. Level 2 forms a system-level perspective for the current situation, after which level 3 attempts to predict the future state of the system in question. Level 4 manages the sensors and allows the refinement of the fusion process, by, for example, shutting down damaged or captured sensors. Level 5 is the interface between the fusion system and the human operator, where the situation is finally assessed by combining the automated analysis and operator expertise.

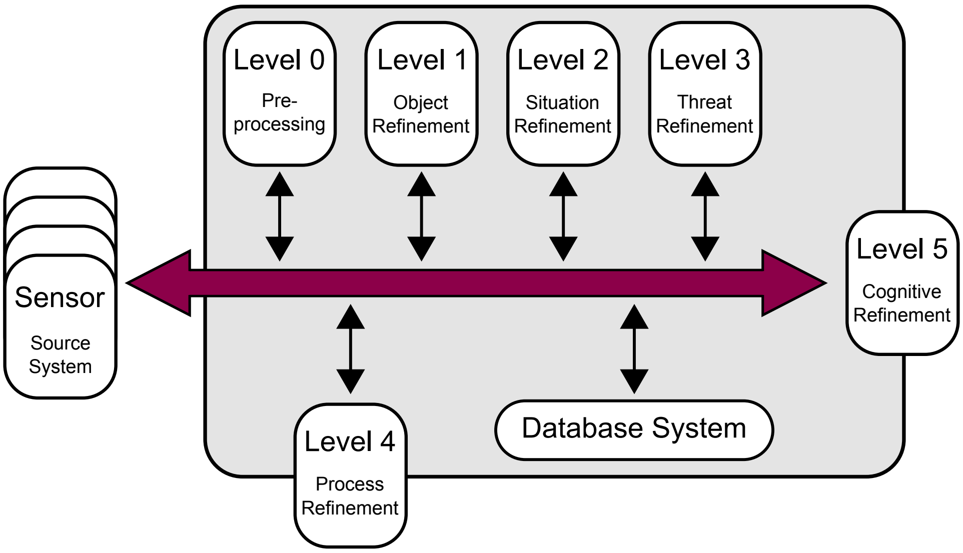


Figure 2: JDL model, adapted from[13,37]

Even though sensor fusion is often understood as a way to improve the prediction quality of ordinary physical sensors, some of the frameworks are also suitable for

Level	Name	Description
0	Pre-processing	Raw sensor data processing
1	Object Refinement	Objects and their position is identified
2	Situation Refinement	Relation among entities is assessed
3	Threat Refinement	Future state predicition
4	Process Refinement	Resources allocation
5	Cognitive Refinement	Human operator interprets the situation

Table 2: List of JDL levels

cyberphysical sensor systems[13]. The sensors in a cyber setting could contain e.g. intrusion detection systems, host health monitoring sofware and network flow analysers. The output of these sensors can be easily combined and refined using sources such as vulnerability databases for further analysis.

Since critical infrastructure consists of various systems of different type, they may be seen as separate types of sensors. JDL model offers a framework for handling the sensors in order to utilise the data they produce. In SACIN framework the JDL model is used to refine data provided by cyber-sensors, called Agents[38]. The various CI systems utilise SACIN Agent middleware for providing information to the system, and central software components further process the various JDL steps before presenting the information to the monitoring operator[17].

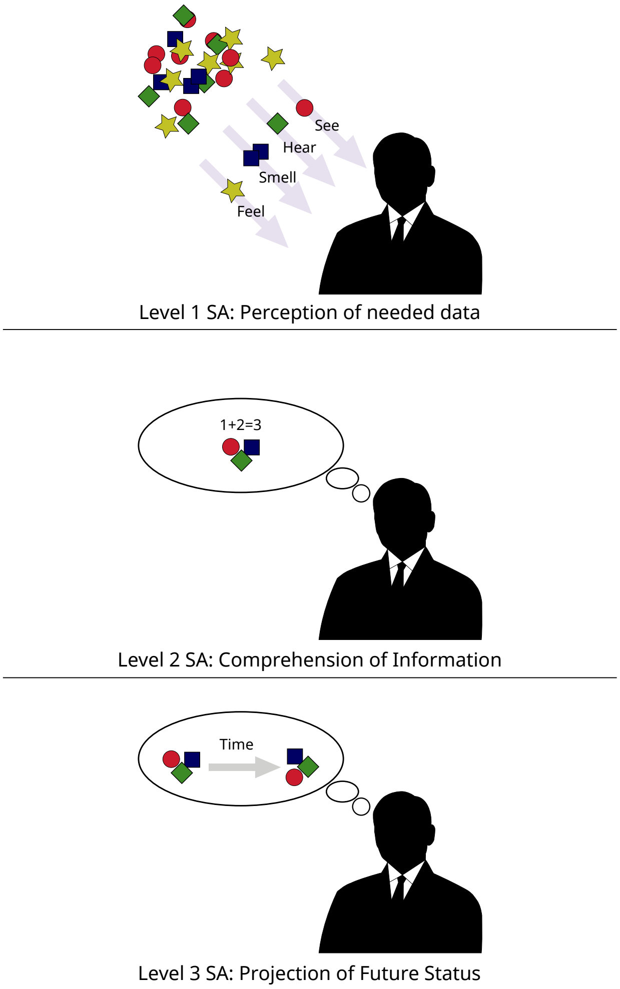
2.3 Situational Awareness

Situational awareness (SA) refers to the information, processing methods, – and ultimately – to the mental picture that a person is required to have in order to accomplish a specific task or procedure. Endsley defines situational awareness as "being aware of what is happening around you and understanding what that information means to you now and in the future"[10].

The concept of situational awareness was first introduced in the field of military aviation, where the information requirements of the fighter pilots were studied[9]. Since the conception, situational awareness oriented thinking has made its way outside aviation circles.

Situational awareness, as defined by Endsley, consists of three levels of understanding: *perception* of the elements, *comprehension* of the current situation, and *projection* of future status (Figure 3)[10]. Critical infrastructure forms a complicated networked system where humans struggle to maintain clear picture about the current and future state. Maintaining a robust situational awareness is, however, necessary in order to detect disruptions and faults as early as possible. The modelling and analysis techniques must, ultimately, support the monitoring operator in his or hers endeavour for obtaining situational awareness.

Obtaining situational awareness requires models and analysis methods which support the human operator in this endeavour. For a critical infrastructure monitoring system, the SA level 1, perception, means collecting essential information about the state of the infrastructure and presenting them to the user. SA level 2, comprehension, requires the operator to understand how the observed information affects the goals or mission. SA level 3, projection of future status, requires that the operator has both the information and mental model for accurately deducing future state from available information.

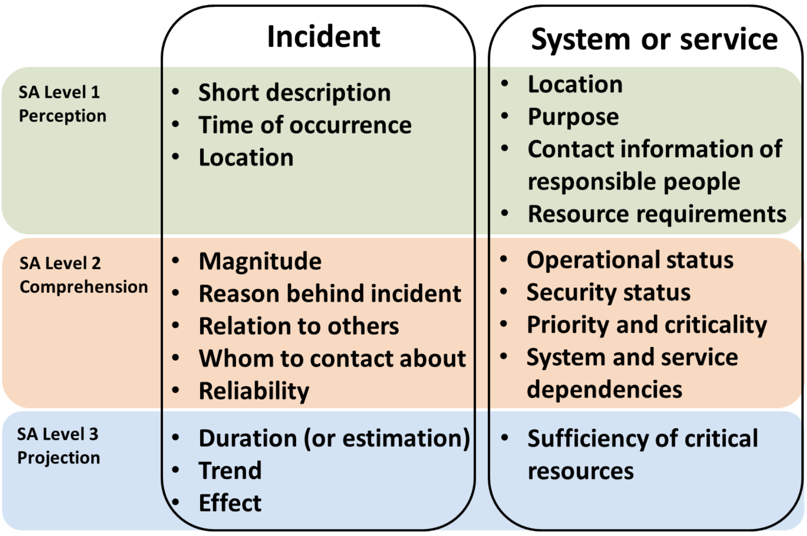


Level 1 SA: Perception of needed data

Level 2 SA: Comprehension of Information

Level 3 SA: Projection of Future Status

Figure 3: Levels of situational awareness, adapted from Endsley et al[10].



	Incident	System or service
SA Level 1 Perception	• Short description • Time of occurrence • Location	• Location • Purpose • Contact information of responsible people • Resource requirements
SA Level 2 Comprehension	• Magnitude • Reason behind incident • Relation to others • Whom to contact about • Reliability	• Operational status • Security status • Priority and criticality • System and service dependencies
SA Level 3 Projection	• Duration (or estimation) • Trend • Effect	• Sufficiency of critical resources

Figure 4: Summary of SA requirements for Critical Infrastructure[30].

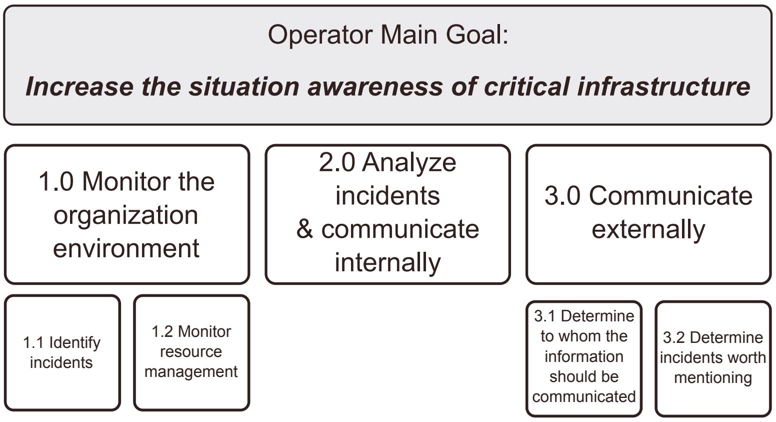

Operator Main Goal:
Increase the situation awareness of critical infrastructure
1.0 Monitor the organization environment
2.0 Analyze incidents & communicate internally
3.0 Communicate externally
1.1 Identify incidents
1.2 Monitor resource management
3.1 Determine to whom the information should be communicated
3.2 Determine incidents worth mentioning

Figure 5: GDTA tree of operator requirements, as defined by Rummukainen et al.[33].

During the DiSCI project, information requirements in all SA layers were formulated. These requirements were defined for both incidents and systems / services[30]. The requirements are illustrated in Figure 4. This information was collected by interviewing subject matter experts, and observing the monitoring operator's work.

Rummukainen et al. compiled a structured representation, known as the *Goal Directed Task Analysis* tree (Figure 5)[34]. The aim of GDTA is to collect requirements for a specific task in a structured fashion, in order to extract parts that are needed for performing said task. The modelling and analysis tools should be built to reflect the requirements.

2.4 Graph Theoretic Concepts

A graph G is an ordered pair $G=(V,E)$ where V is a set of vertices (or nodes), and $E \subseteq [V]^2$ is a set of edges. A graph may be directed or undirected, depending on how the edge set is defined; undirected graphs consider edges $\{a,b\}$ and $\{b,a\}$ $a,b \in V$ equivalent, whereas directed graphs treat edges as ordered pairs, thus (a,b) and (b,a) are different edges with direction from first tuple member to the second[8].

2.4.1 Centrality

Graphs can be used to represent networks, both man-made and naturally occurring. In these networks some nodes can represent things that are more important in rank than others. This ranking is known as centrality. Various different *centrality measures* exist to form a ranking, each having a different concept of what constitutes as important. Centrality measures generally rank nodes based on reachability, amount of leaving and incoming edges, or some combination of these metrics, but additional attributes may be used if they are added to graph.

One of the simplest centrality measure is the degree centrality. Node's centrality is simply defined by how many edges it has. The degree centrality (C_D) of node v is defined as

Definition 1 (**Degree centrality**).

$$C_D(v) = deg(v)$$

where *deg*() is the number of edges.

As a node with relatively low degree can act as a bridge between two large segments, it is apparent that the nodes should be examined in relation to more nodes than their immediate neighbours. Betweenness centrality attempts to rectify this shortcoming by using shortest paths[12]. Betweenness centrality is calculated by examining shortest paths between every other vertex than v. A ratio between shortest paths including and excluding v is calculated for each pair, and the sum of all ratios is the betweenness centrality.

Definition 2 (**Betweenness centrality**).

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the sum of shortest paths from s to t and $\sigma_{st}(v)$ is the sum of shortest paths that pass through v.

Centrality measures often take long to calculate for larger networks. For example, the time complexity of betweenness centrality can be $O(|V|^3)$ in worst-case. For this reason, the graph centralities can be calculated in advance, and weighed with additional parameters to reflect changed situation.

2.4.2 Graph Topology in Man-Made Networks

The topology of man-made networks has been under intensive research for decades. This research has yielded an observation of phenomena, where man-made infrastructure networks seem to follow a particular structure. The results indicate that the degree distributions tend to follow a *power-law distribution*[24,39,4]. The approximate knowledge of the distribution is useful, when modelling the infrastructure systems, since it allows the estimation of graph densities and other topological properties. These properties can be used to choose a centrality measure that provides meaningful results in CI specific graphs.

Graphs, whose degree counts follow the power-law distribution, are called *scale-free* networks. More formally, in a scale-free graph, the probability an item of size x (or node with a degree count of x), is

$$p(x) = Cx^{-\alpha}$$

where $\alpha > 0$ is the exponent of the power law[24]. The constant C is determined once α is known, such that the distribution sums to 1.

The graph shown in Figure 6 is a randomly generated scale-free graph, with its degree distribution in Figure 7. A network like this could be a part of a small power grid in a remote location: few transformer substations deliver electricity to smaller transformers, which are then connected to the individual houses. The graph in Figure 6 was created using the Barabási–Albert method presented in[4], a process for generating graphs obeying exponential distribution. The degree centrality is visualised by indicating a higher degree with a red hue.

Real-world networks also exhibit scale-free structure. A dependency graph containing all transformers and substations of the Åland island was constructed using National Land Surveys topographic database[20]. The resulting graph has 812 nodes and 832 edges, with average degree distribution of 2.049. The graph is presented in Figure 8, and the corresponding degree distribution in Figure 9. Exponentially distributed data should fall on a straight line in a log-log plot. Figure 10 contains this fitted line to further illustrates the exponential nature of the degree distribution.

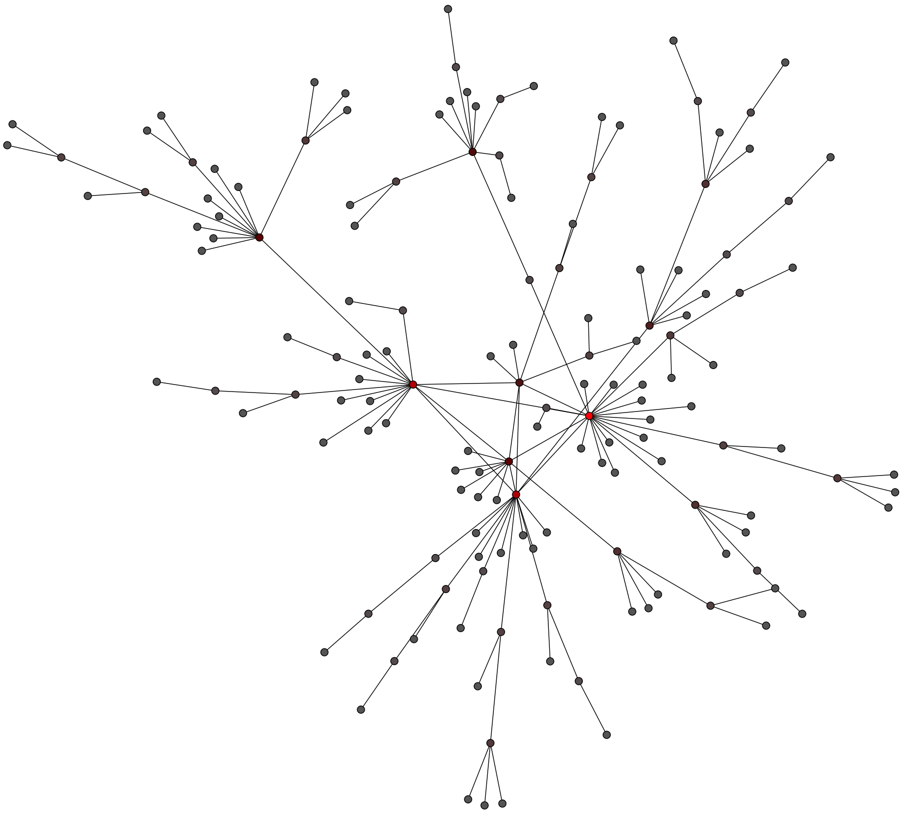


Figure 6: A randomly generated Barabási–Albert graph with $n = 150$ nodes and $m_0 = 5$ initial nodes. The red colour indicates a higher degree centrality.

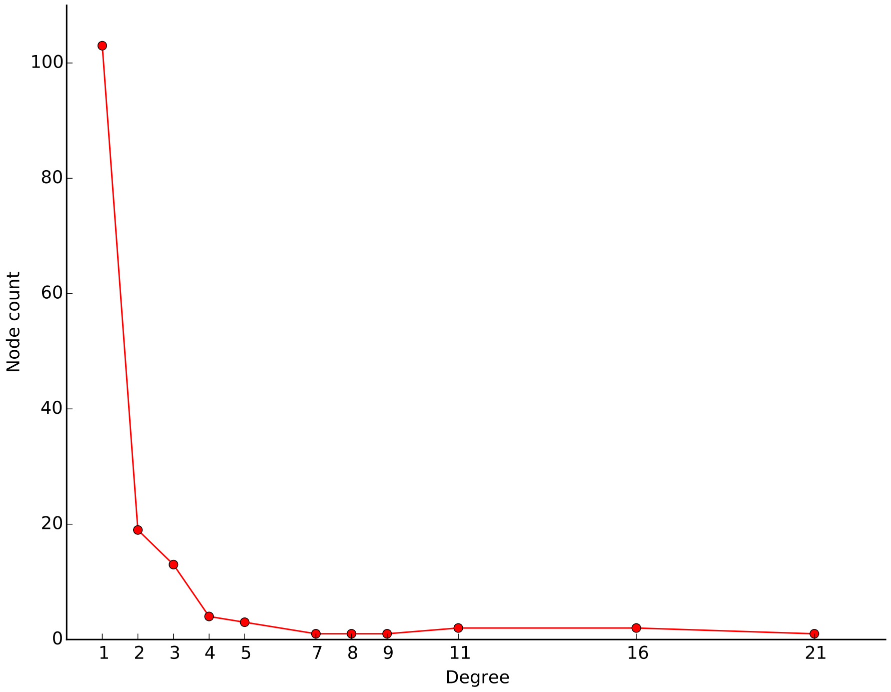


Figure 7: The degree distribution for graph in Figure 6. There are over 100 nodes with degree of 1, and only one node with a degree of 21.

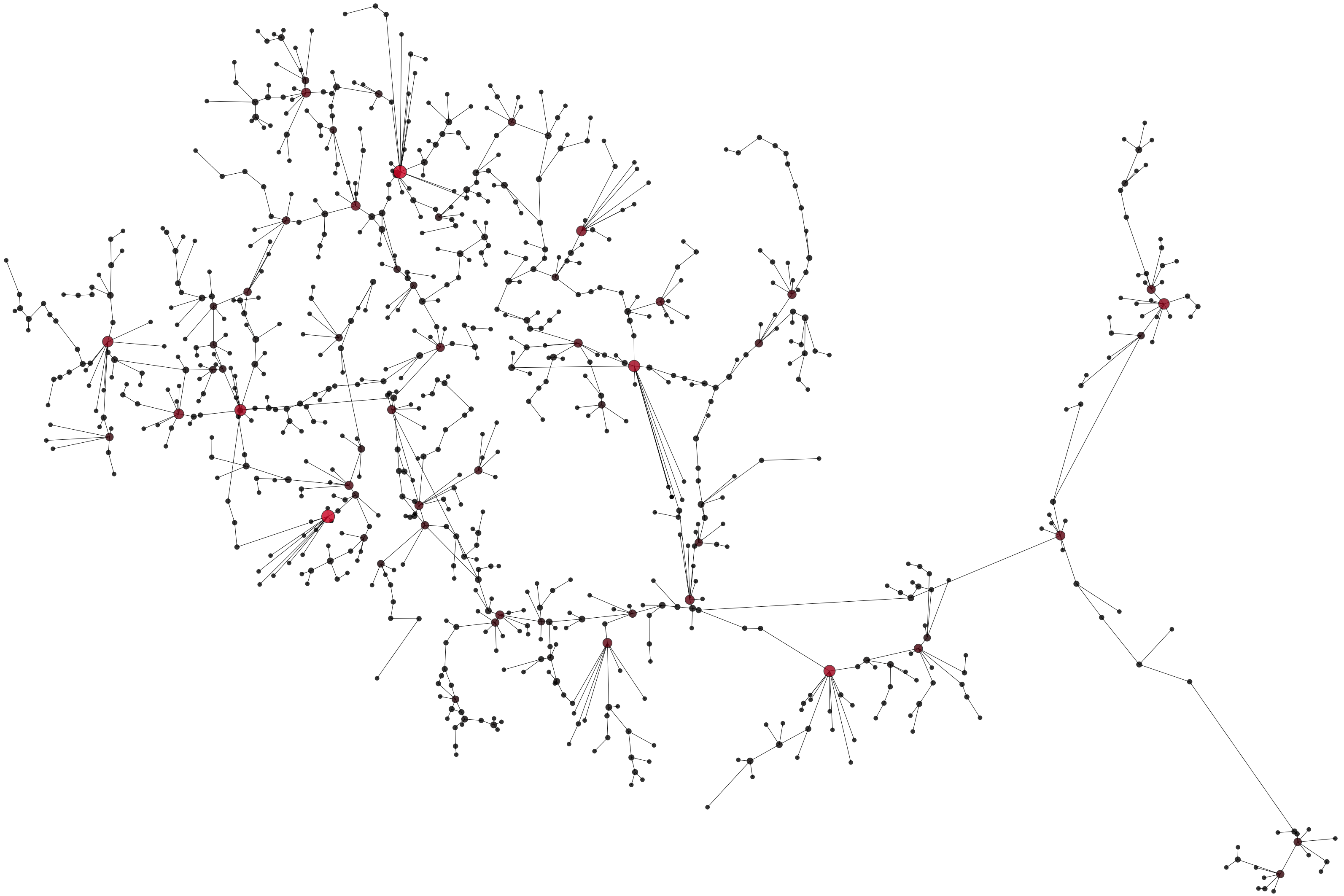


Figure 8: A graph view of the Åland island power grid.

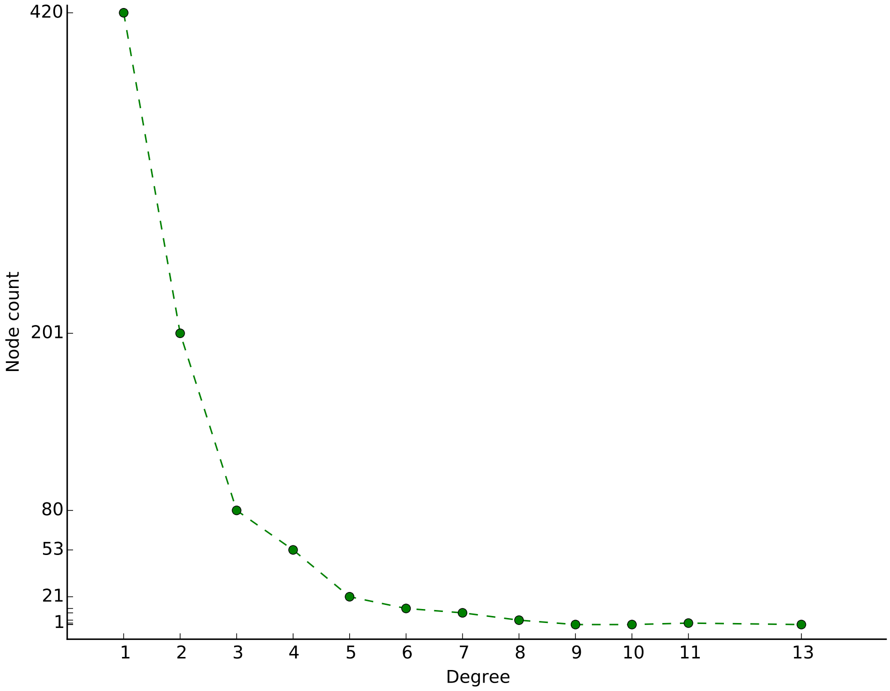


Figure 9: A plot of the degree distribution of the Åland island graph. Red color indicates higher degree centrality.

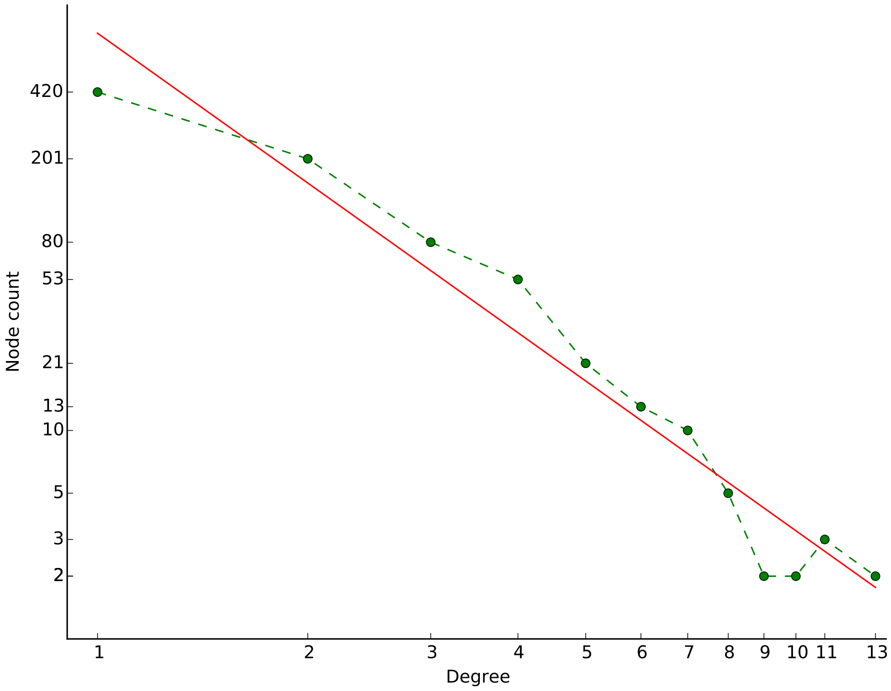


Figure 10: A log-log plot of the degree distribution of the Åland island graph, with a fitted line illustrating the logarithmic nature of the distribution.

2.5 Information Theory and Entropy

Information theory considers the concept of information from a mathematical standpoint. The roots of information theory lie at the advent of electronic communication systems. The telegraph called for a more formal system to quantify various properties of communication systems, such as optimal channel capacity and information loss[7].

In order to understand information, a model for *information source* needs to be examined. An intuitive approach is to imagine a telegraph. The purpose of a telegraph is to produce telegrams that are of interest to the receiver. A telegraph is a sequence of characters of a particular alphabet. Each letter of the alphabet has a know frequency, i.e. probability of appearing. The receiver knows how common each letter is, but the actual message is unknown. Information source, such as the telegraph, can be modelled as a stochastic process, since the output stream can be thought as a sequence of random variables[28].

Definition 3 (**Discrete Memoryless Source**)**.** The discrete memoryless source (DMS) is a simple model of an information source. DMS consists of a random variable X and an associated probability mass function, and symbol emission speed. The DMS is then a random process where output is a sequence of independent and identically distributed random variables.

The receiver wanting to quantify the amount of information any single observed symbol contains. Intuitively, since the probability of observing each symbol is known, observing an unlikely symbol contains more information than a common one. For example, observing a source capable of transmitting only a single symbol provides no information. The logarithm based definition of information has been chosen due to its useful and intuitive properties[36].

Definition 4 (**Information**)**.**

$$I(x) = -\log P(x_i)$$

The concept of entropy is tightly related to information. It is defined as the expected value of information, or more formally

Definition 5 (**Entropy**)**.**

$$H(X) = -\sum_i P(x_i) \log P(x_i)$$

Entropy is usually measured by using logarithms of base 2. This means, that the resulting value will have its units in bits. Imagine, if you will, an uniformly distributed random variable X over 8 outcomes. Since $\log_2 8 = 3$, it means that three bits is enough to represent all outcomes.

$$H(X) = -\sum_{i=1}^{8} P(i) \log_2 P(i) = -\sum_{i=1}^{8} \frac{1}{8} \log_2 \frac{1}{8} = \log_2 8 = 3$$

Since entropy tells how many bits, on average, it takes to represent an outcome, the calculated entropy agrees with the assessment above. A non-uniform distribution with eight outcomes $\left(\frac{4}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}\right)$ would then have a smaller entropy:

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{8}\log_2\frac{1}{8} - 6(\frac{1}{16}\log_2\frac{1}{16}) = 1.375$$

As expected, the entropy is smaller because this random variable does not have as much uncertainty as the uniformly distributed counterpart. In other words, the entropy of a probability distribution can therefore be seen as a characterisation of unpredictability.

3 Modelling Critical Infrastructure

This section discusses about modelling of critical infrastructure from a situational awareness viewpoint. We start by defining the requirements and continue by presenting a graph-based model.

3.1 Requirements

Critical infrastructure has been described as highly interconnected and interdependent system of systems, that consists of thousands of devices, services and processes with complex and sometimes unknown relationships[32]. Information retrieval is often possible from only a subset of the systems, and the availability of historical data may be uncertain.

The possibility that a fault in one system can cause a cascading failure that propagates through the whole CI must be taken into account. This propagation might not be instant, and it may require a depletion of backup resources. Time component must therefore be incorporated into the model to account for this behaviour. Flexibility and the ability to use different levels of abstraction in the model are considered beneficial features. It is likely that all of the dependencies are not known in advance.

For situational awareness purposes, the system's current and future ability to produce or provide a certain service is more crucial than the exact modelling of internal functionality. By modelling the infrastructure as a network of services dependent on each other, the information requirements can be kept on a manageable level. The model must be capable of operating, with limited accuracy if necessary, in the situations where information is scarce. Due to the limitations in available data, the information requirements of the model must be relatively modest. This requires that all domain specific details, like electricity flows should be omitted if possible, in order to simplify the model and make it suitable large-scale deployment, where running heavy models at that scale computationally prohibitive. The relations between each system may not be contingent on knowing material flows or other technical details of the coupling, since such data is usually not available.

Rummukainen et al. have identified a set of requirements for pieces of information that must be present in order for the model to be a suitable tool for CI SA systems[34]. As a minimum, the model must include following:

a. *Operational status* must exist for every system. Each object must be accompanied with a tag explaining current or best known status.

b. Each system must have *criticality* or *priority*. There must exist a measure which allows the ranking of the systems based on their relative importance in the CI.

c. *Dependencies* between systems must be, at least partially, known. If a correct operation of a system is dependent on some other system, this relationship needs to be modelled.

d. Sufficiency of critical *resources* must be modelled. If there is a resource (e.g. backup power) stored onto some system, the depletion process must be present.

Rummukainen et al. also define other essential information requirements; location, purpose, contact information, and security status. These are not necessary core parts of the mathematical model, as no operations are done with them. They are added as tags to each object on data modelling stage during software implementation, and made available on the user interface.

3.2 Dependency Graph

Directed graphs are extremely suitable for presenting dependencies between objects. Graphs have been used as an aid for analysing critical infrastructure[39,26], but the focus has been on the static analysis of topology and dependency chains, without active functionality.

Due to the interconnectedness of CI, we have chosen dependency-heavy modelling formalism. Critical infrastructure can be modelled as a dependency graph, where nodes represent different CI systems and the edges dependency relationships between them. To model the interconnectedness and dependencies inside critical infrastructure, we construct a directed graph:

Definition 6 (**Critical infrastructure dependency graph**). A directed graph $G = (V, E)$ where each vertex $v \in V$ represents CI system and each edge $e \in E$ a dependency relation between two systems.

A relation may exist between any two systems, and the systems may be depended on each other as is often the case in real-world systems. For example, many GSM base stations fail after 3 hours of continuous power outage, so the stations are dependent on electric grid[14]. However, a next generation power grid may also depend on cellular networks for remote adjustment, and malfunction without it. These systems form a bidirectional dependency relation with each other. A failure in either system will cause the dependent system to fail after certain time, which is not necessarily the same to both directions. Each node on the critical infrastructure dependency graph presents one system. There is no definite guidelines as to what constitutes as a system in critical infrastructure setting. For example, a cellular radio tower can be modelled as a single system providing GSM and LTE services, or the base stations could be modelled as two separate systems. The appropriate abstraction level is left for the persons implementing the monitoring framework for specific systems. In most cases, the distinction should be relatively straightforward.

Critical infrastructure devices and systems are usually monitored by the companies or other organisations that own the devices. When a significant change is operational state occurs, the operators are usually notified of this *event* via a status message sent by the affected device. Status messages are often also sent periodically, even when no fault is detected to ensure that the devices can not silently fail. Monitoring is usually done by some automated system that is supervised by a human operator using a computerised SA system. For simplicity, it can be said that a system can send events, even though it would be more accurate to say that a system sends status messages containing details of the occurred event.

An event that causes a malfunction or interrupts a service or process is called *incident*. Incidents may trigger further events and incidents on other systems. When coupled with the knowledge about the dependencies, it is possible to model the critical infrastructure as a set of communicating systems. Even if some systems are not actively monitored, we can indirectly observe their possible state by observing systems connected to them.

3.3 The System Model

The states and state changes are naturally captured by using a *finite state machine (FSM)*. Since CI systems are expected to influence others on state change, we model this behaviour by expanding the Mealy machine, a variant of FSM that adds output values to state transitions based on current state and input[22]. In the model, the output alphabet represents outgoing effects. A change in system's state is expected to affect the operational status of the dependent systems. This is modelled by using the output alphabet as an input feed for dependent systems automatons, and in turn for its dependencies for cascading failure effect. In this thesis, we use the terms state and status interchangeably.

Definition 7 (System state machine). The system state machine is an 8-tuple $SSM = (Q, \Sigma_i, \Sigma_o, T, O, D, S, q0)$, where

Q is a finite set of (capability) states;

Σ_i is a finite set of input events (alphabet);

Σ_o is a finite set of output events (alphabet);

$T : Q \times \Sigma_i \to Q$ is the transition function;

$O : Q \times \Sigma_i \to \Sigma_o$ is the output function;

$D : Q \times \Sigma_i \to Q \times \mathbb{R}^+$ is a delayed state transition function;

$S : Q \to [0,1]$ is a status function;

$q0$ is the initial state.

The state machine represents systems *operational status*. The status function S maps each state to a real-valued variable, which represents the current percentage of operational capability.

The pure state machine does not intend to fully capture all aspects of a CI system. The key idea is to include all relevant failure stages from 100% to 0%. For example, a primary reactor coolant circuit of a nuclear power plant could be modelled as a 4-state machine (Figure 11). In a reactor, incidents like severe coolant overheating or overpressure may require extreme emergency measures, which leave the reactor unusable[35].

Timed transitions are used, when another state change is known to follow fist change after a period of time. This functionality is implemented using countdown timers. After moving to a state where delayed transition is defined, a timer is initialised to correspond function D. If no state change has occurred before timer expires, the state is automatically changed. The timer is deleted, if another state change occurs before timer expiration.

The System state machine can be represented by several matrices and indexes; By presenting the transition and output functions as a lookup table, the current state can be then presented with a single index $Q_{current}$. If there are multiple instances of the same automata, they can share the same lookup table, and only use this index for determining their state. Timed transitions also require automata-specific countdown clock for timed state change. Since we expect to see many instances of the same automata, we can efficiently implement the automata transition function as a table. The amount of states in one FSM is assumed to be quite small, as the idea is to capture the failure pattern rather than try to model the whole process accurately. Instances of the same automata can use a shared lookup table, and only keep track of their current state. This potentially results in improvements on both memory and CPU usage in cases where the modelled infrastructure is relatively homogeneous.

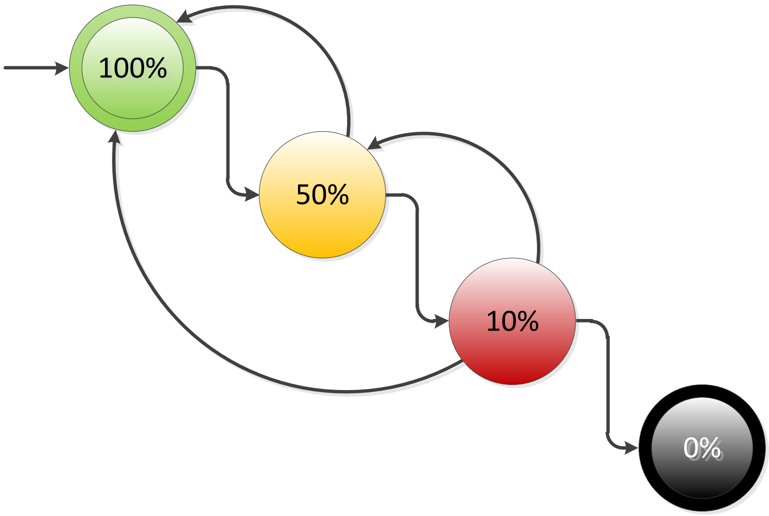


Figure 11: An exaple of a system state machine, where 0% status means unrecoverable failure. In reality, the state machine representing a nuclear reactor would be more complex.

3.4 Critical Infrastructure Model

The CI system is modelled as a set of system state machines that communicate with each other via messages. These actors correspond to the nodes in critical infrastructure dependency graph, which may communicate via unidirectional first-in first-out channels, that correspond to the edges of the dependency graph, honouring their direction.

Updating system state machines by traversing the graph can be done in $O(|V|+|E|)$ time complexity by expanding breadth- first traversal, as shown in Algorithm 1. $|E|$ is typically much smaller than $|V|^2$, since critical infrastructure dependency graphs are relatively, as shown in Section 2.4.2. The update algorithm is run every time a new event arrives, or any SSM changes state due to a delayed transition.

The algorithm utilises two queues (Q and P), that keep track of nodes in two levels. Since a node can depend on multiple parent nodes, the status of a node cannot be determined before all of its parents are at their correct state. Furthermore a node can be dependent on multiple other systems. In case there are several possibilities, the worst one (given by the S) is chosen as the new state by UpdateLevel function.

The graph is processed breadth-first in a fashion which forms directed acyclic graph originating from the affected node. Back edges to already processed nodes are ignored, since the accuracy of the model, and the CI graph topology are not suited for cycle analysis.

Algorithm 1: SSM update algorithm

```
Input : A graph $G$ and source vertex $s$
Result: Updated graph $G$
begin
    depth← 0;
    for each vertex $v \in G$ do
        color[$v$] ← WHITE;
        d[$v$] ← $\infty$ ;
        sym[] ← NIL;
    color[$s$] ← GREY;
    d[$s$] ← 0 ;
    Q← $\emptyset$;
    P← $\emptyset$;
    ENQUEUE($Q$,$s$);
    ENQUEUE($P$,$s$);
    while $Q \neq \emptyset$ do
        $u$ ← DEQUEUE($Q$);
        for each $v \in child[u]$ do
            if ($color[v] = $ WHITE) then
                color[$v$] ← GREY;
                d[$v$] ← d[$u$]+1;
                ENQUEUE($Q$, $v$);
                if $d[v] > depth$ then
                    UpdateLevel();
                    depth+1;
                v.sym[] ← u.symbol;
                ENQUEUE($P$, $v$);
            else if ($color[v] = $ GREY) then
                if $d[u] < d[v]$ then
                    v.sym[] ← u.getSymbol;
        color[$u$] ← BLACK;
    UpdateLevel();
```

Function UpdateLevel

```
begin
    while $P \neq \emptyset$ do
        $u$ ← DEQUEUE($P$);
        test each $s \in$ u.sym[] and set $ASM_u$ to state with lowest value given by
        function $S$;
```

4 Real-Time Analysis

This section discusses about the the analysis methods for critical infrastructure models. Firstly, the requirements for achieving desired analysis capability are examined. Secondly, two analysis tools are presented, one for quantifying impact of incidents, and a second one for estimating operational status under uncertain conditions.

4.1 Requirements

Situational awareness is formed through the use of sensors, data fusion, and automated systems, but it is fundamentally connected to the human element. Therefore the analysis should focus on delivering metrics that are useful to the human operator. Situational awareness tools do no usually perform any sort of automated response, as any action is orchestrated using human-in-the-loop -paradigm

Analysis methods were selected to conform to the requirements defined by Rummukainen et al[34]. The most important analysis capability is a method for quantifying the impact of single event, on both the system it affected and the CI network as a whole. For each incident or event, following information is needed:

a. *Magnitude* must be quantifiable at both system and CI level.

b. *Relation* between incidents and systems.

c. *Duration* must be quantifiable at both system and CI level.

The analysis results should be produced in real-time from the perspective of the monitoring operator. Since the operator might want to conduct multiple variations of the presented analysis techniques under restricted time, the importance of quick operation is further heightened.

The design philosophy mirrors the same principles used in the model. The analysis should produce estimates of the operational status using information provided by the model and the received event. Estimates of operational status should also be provided even when current data is not available.

4.2 Topology Based Measures on Critical Infrastructure Model

To assess the impact of an event, it is necessary to account for the current situation, and report how much it was changed. Since the critical infrastructure is modelled as a graph, it is possible to leverage pre-calculated centrality measures to assess how important a certain node is in a given situation. Depending on the centrality measure used, it is possible to analyse impact of different cascading failure patterns.

We continue by defining a Downstream Weighted Impact Sum (DWIS), as first described in[29]. It attempts to estimate the impact of a single event by summing a change in status function value, weighed by the centrality. More formally, DWIS is defined as:

Definition 8 (Downstream Weighted Impact Sum).

$$\mathrm{DWIS}(v) = \sum_{A_i \in \mathrm{T(v)}} \Delta S(A_i) \cdot C_i$$

where v is the starting node, $T(v)$ is the set of all nodes reachable from v, $\Delta S(A_i)$ is the difference of SSM status value function before and after the state transition caused by the event, and C_i is the (normalised) centrality of the node i. DWIS can be calculated in $O(|V| + |E|)$, since it requires only one pass through each affected node.

The formula allows us to estimate the effects of a single event on the whole infrastructure, as it gives higher results if important nodes are affected. This allows the estimation of the *magnitude* of the disruption. If one of the systems is known beforehand to be extremely crucial, there is the possibility of weighting it more during during modelling a particular CI cofiguration. The DWIS is highly dependent on the chosen centrality measure, as well as the scaling function, if any, used to normalise the measures. A suitable scaling function can be chosen to achieve balance between nodes with small and large centrality.

The DWIS explicitly uses the difference between old and new state, $\Delta S(A)$. The method provides results based on the actual change in operational condition. It can also be used to estimate what would theoretically happen, if the status of particular component is altered, for example by repairing it. It allows the framework to suggest which component would be the best candidate for repairs. Since the maximum sum of centrality is known, current situation can also be assessed against the fully operational state.

4.3 Entropy Measures on Critical Infrastructure Model

The previously presented model and analysis methods are event-driven and discrete. Although the delayed state transitions add a time component to the model, it does not otherwise account for passage of time[16]. The goal of the probabilistic analysis is to leverage existing knowledge about the target system for estimating its current state. For this purpose, we associate each system with a probability distribution, that associates each possible FSM state with a probability.

For example, let M be a finite state machine with 3 states (operational [O], marginally operational [M], and not-operational [N]) such that

$$P(X = x) = \begin{cases} a, & \textit{when } x = O \\ b, & \textit{when } x = M \\ c, & \textit{when } x = N \end{cases}$$

The probabilities a, b, and c may have been collected by observing the operation of the system for a certain time period, or they may have been defined by the operator. The associated probability distribution should be chosen to model how the system reacts when it is unable to communicate with the monitoring framework. Some systems might experience intermittent communication delays which do not cause any disruptions. Other systems, such as networked remote control solutions, might only stop communicating if they are malfunctioning. The chosen distribution can now be used to estimate system status without status messages.

We may now view the system as an information source that produces events based on the defined probability distribution. After each event, the probability distribution must be altered. This models the fact that major faults often prevent the system from reaching normal operational capacity without repairs or other manual intervention.

In case we get a sensor reading not-operational [N], we define the new probability distribution as follows:

$$P(X=x)=\begin{cases} a\cdot(1-\mathrm{e}^{-kt})\cdot S(N), & \textit{when } x=O \\ b\cdot(1-\mathrm{e}^{-kt})\cdot S(N), & \textit{when } x=M \\ 1-(a+b)\cdot(1-\mathrm{e}^{-kt})\cdot S(N), & \textit{when } x=N \end{cases}$$

where t denotes the elapsed time since the event, constant k is an operator-definable parameter, and $S(N)$ is the output of the status function as defined for the automata M. The new probability distribution is calculated using the severity of last-known status, elapsed time since receiving said status, and the original distribution associated with the system. By using the status function for weighing the probability, the distribution accounts for the fact that recovering from "worse" states to "better" is more improbable than other way around. The distribution allows us to calculate the expected value of the function S, as well as the entropy.

The general case for n-state system is defined as

$$P(X=x)=\begin{cases} a_1\cdot(1-\mathrm{e}^{-kt})\cdot S(A_j), & \textit{when } x=A_1 \\ a_2\cdot(1-\mathrm{e}^{-kt})\cdot S(A_j), & \textit{when } x=A_2 \\ \vdots & \\ 1-(\sum_{i\neq j} a_i)\cdot(1-\mathrm{e}^{-kt})\cdot S(A_j), & \textit{when } x=A_j \\ \vdots & \\ a_n\cdot(1-\mathrm{e}^{-kt})\cdot S(A_j), & \textit{when } x=A_n \end{cases}$$

In Figure 12 a plot of three-state FSM representing a base station is shown. At time $t=0$ the base station is known to be marginally operational, due to lack of power. During normal operation, the station sends heartbeat event once a minute. System operator has indicated, that even a brief delay in status messages is a symptom of a fault. This causes the not-operational state likelihood to increment. There is a fair amount of uncertainty associated to probability distributions before $t=150$, since no state is extremely unlikely. In Figure 13 the fast growth of entropy can be seen, while the expected value of S remains relatively stable.

At $t = 150$ the station is able to send 'battery depleted' status message before losing power. This causes the probability for not-operational state to become one. Since the value of S for non-operational state is almost zero, a recovery is highly improbable. The growth of entropy after $t = 150$ also illustrates the differences between distributions.

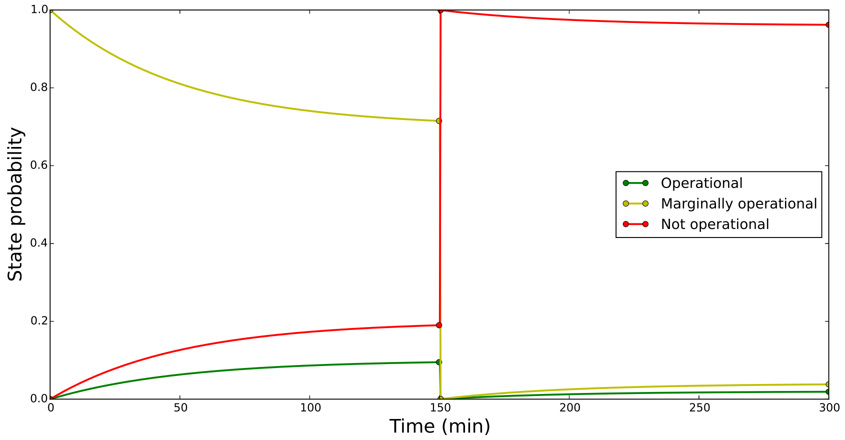


Figure 12: Probabilities for each three operational state in base station FSM. At $t = 150$ a message indicating battery depletion is received, causing a state change.

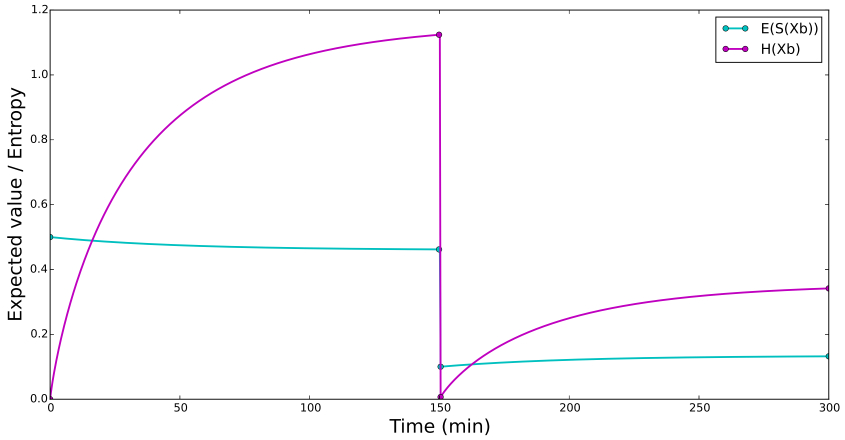


Figure 13: Expected value of $S(X_b)$, and entropy $H(X_b)$ (measured in bits) of a random variable X_b representing a status of a cellular base station.

By defining the probability distribution using both pre-collected knowledge of failure patterns, and last-known state, an estimate of current status can be made without new information. The entropy associated with the calculated probability distribution informs how "unreliable" the expected value currently is. For example, the expected value (or nearest possible state) of a system can be displayed as a coloured marker in the user interface. Entropy can be used to turn this indicator slowly to gray, when the estimate is considered too unreliable for use in decision making.

5 Evaluation

The proposed model and analysis techniques were evaluated against the requirements presented by Rummukainen et al.[33], as well as running several benchmarks over several generated and real-world graphs. Various scenarios were designed based on reports compiled by Regional State Administrative Agencies, using datasets supplied by the VTT Technical Research Centre of Finland Ltd (VTT) in conjunction with Caruna Ltd.

5.1 Requirement Based Evaluation

The main goal of the proposed modelling formalism was to to satisfy the requirements set by Rummukainen et al., as presented in Section 3.1. Requirement *a* – *operational status* for a CI system can be determined by the assigned automaton and the status function S. The automaton keeps track of the current status and the S function associates a numerical value for operational capability. Requirement *b* – *criticality* is assigned to each node by using a graph centrality measure. In case, where a particular node is known to be extremely important, the calculated measure can be boosted to include this fact. Also the centrality measure and centrality normalisation function can be adjusted for increased accuracy. Requirement *c* – *dependencies* are expressed by using a directed graph, where the edge directions mark dependency relation between systems. Finally, requirement *d* – *resource sufficiency* is modelled by allowing time-delayed transitions in automata. The requirements and corresponding model capabilities are shown in Table 3.

Requirement	Model capability
Operational status	Automata reflects the best-known status
Criticality	Graph centrality is calculated for each node
Dependencies	Directed graph expresses the dependency relations
Resource sufficiency	Timed transitions to model e.g. resource depletion

Table 3: List of SA requirements and model capabilities

The analysis methods were also created based on requirements by Rummukainen et al., as presented in Section 4.1. For each event, the proposed analysis methods can provide estimates of magnitude at both system and infrastructure level. The change S function output directly shows the magnitude for any single system. Utilising the centrality and S function, magnitude at infrastructure level can be estimated by calculating a centrality weighed sum of all affected systems. In case there is no new data available, a probabilistic entropy based measure can be used to make less reliable estimates. Duration was not directly addressed by the proposed analysis methods. The probability distribution used in the entropy based analysis method addresses the time component, but it does not provide estimates on repair time or fault duration. The requirements and corresponding analysis capabilities are shown in Table 4.

Requirement	Analysis capability
Magnitude	Impact sum utilising centrality and change in operational status
Relation	Graph structure shows all affected systems
Duration	Not directly addressed by the analysis methods

Table 4: List of SA requirements and analysis capabilities

Some elements of the requirements are best addressed at software implementation stage. The model and analysis methods should be complemented with well-thought visualisation methods, map data, and additional information, as the raw numbers and values might not be intuitively meaningful for human operators.

5.2 Modelling and Analysis Using Simulation Tools

A simulation environment was created for use in VN TEAS project using real failure data and base station coverage measurements. Caruna Ltd. provided a dataset containing the location, type, and partial dependency structure of various electric grid components for use in this research project. They also provided a fault log detailing a storm that affected the provided grid, spanning several weeks. In addition, a cellular coverage mapping of the area was conducted by VTT. The environment and the datasets were used to test the model and analysis methods by simulating realistic failure data and replaying actual failure logs.

5.2.1 Dataset

A dataset covering both electric grid and mobile phone communication networks found in a coastal area of Finland was used. A total of 2391 components from the electric grid were used in building of the model. Components found in the dependency graph include primary and secondary substations, as well as disconnectors, as shown in Table 5. This ’Service area’ is linked to the nation-wide high-voltage core grid, and forms a mid-voltage distribution network belonging to a local utility company. Component positions and logical dependency information was further augmented by using open access data provided by the National Land Survey of Finland (NLS)[20].

A total of 151 GSM/UMTS/LTE radio towers found in the area were connected to nearest secondary substation. There were total of 1747 base stations attached to the towers. The base stations were added as attributes for the hosting radio tower, and were not modelled as separate nodes.

Finally, buildings (N 119206) in the area, as indicated by NLS data, were attached to the nearest secondary substation using QGIS (2.12 Lyon)[31]. Due to the sensitive nature of the dependency data, accurate visualisations of the geographical locations or dependencies are not shown in this thesis. The data does not have dependencies from base stations to electric grid components, since final cellular coverage maps were not available.

Feature	Count
Primary substation	23
Disconnector	54
Secondary substation	1390
Radio tower	151
Building	119206
Total no. of features	120824

Table 5: The features included in the building of final graph model.

The final dependency graph has 68080 nodes and 68079 edges. The degree distribution exhibits power-law properties, as illustrated in Figure 14, and agrees with our assumption of CI graph topology. The building nodes were omitted from the picture, since they all have degree of one. Each node representing a power grid component was given a two-state FSM (Figure 15a). Base stations, which contain a backup power source are assigned with a three-state FMS (Figure 15b). The buildings were also assigned type (a) automatons, to model whether or not they are receiving power. Operational states (green) were assigned $S(O) = 1$, Marginally operational states $S(M) = 0.6$, and Not operational states $S(N) = 0.1$. Degree distribution and centrality measures for the graph were calculated using Gephi (0.8.2)[5].

Custom format parsers and scripts were developed to integrate all datasets into formats beneficial for the framework. Geographical data on the above-ground segments of the electric grid lines (NLS) was unified to form a connected line segments. Since NLS database is intended for constructing maps, it does not have actual models of the power lines. Instead, the lines are presented as polyline segments, which do not necessarily form a continuous line even though the actual line is continuous. Furthermore, a crossing of two or more lines, or connections to buildings or relay fields created irregularities on the grid network as some of the essential structures are underground[21]. These issues were corrected by hand, consulting aerial photos, also provided by the NLS database. In the case of large transformer structures, the aerial photos were used to estimate how the outputs of a particular components were configured, and in the case of islands or islets, undersea cable entry and exit points. This process resulted in one unified undirected graph representing the power lines of the target area.

The Caruna dataset contained a tree-like representation on the dependencies

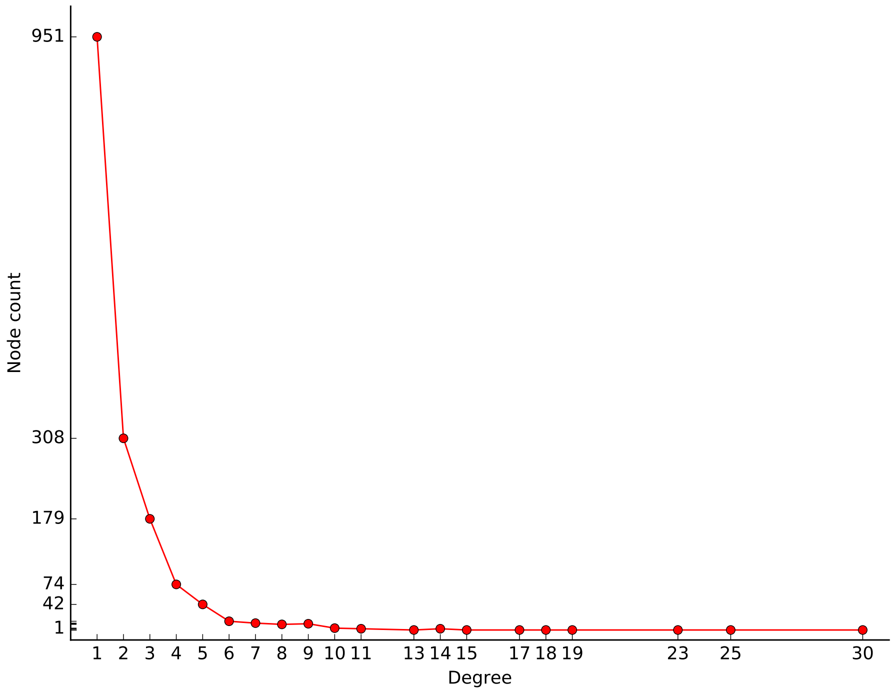


Figure 14: The degree distribution for graph of the modelled area, excluding building nodes.

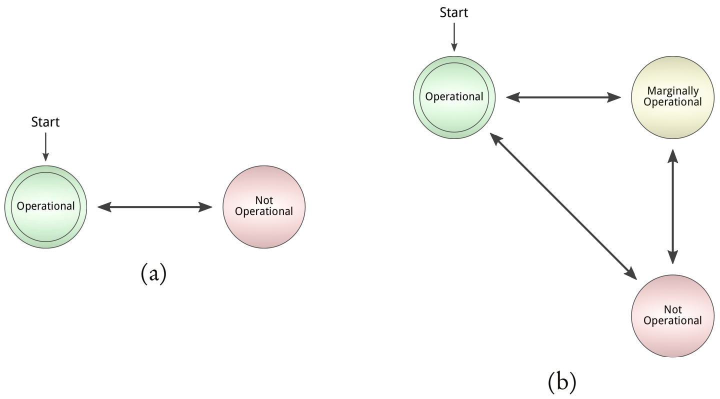


Figure 15: Two state diagrams representing the operation of (a) electric grid component and (b) base station. Green state represents [O]perational, yellow [M]arginally operationa, and red [N]ot operational state.

of grid components (as well as their geographical location and type), but lacked critical inter-component dependency information; the dataset could indicate that a primary substation was powering three secondary stations, but would lack the information that the dependent stations were connected as series on the same power line, such that a failure in first component would render the two remaining non-operational. To rectify this situation, the grid components on Caruna -dataset were connected to the nearest line segment of the NLS data via geoprocessing. Using both Caruna and NLS data, the final dependency graph was created by utilising shortest-path routes and spanning tree algorithms, until the resulting graph respected the constraints imposed by both datasets. The nodes representing buildings were connected to the nearest secondary substation, if the distance did not exceed 3 Kilometres. The dataset was hand-pruned after the operation to account various geographical factors such as bodies of water, major roads and other confounding factors. Convex hulls for each building group associated with substation were calculated, in order to visualise areas affected by substation faults (Figure 17). In addition, some buildings, such as hospitals, were marked 'critical' (N 117).

A mapping between the geographical data and the nodes found in the graph model was retained for visualisation and future analysis purposes. Final data formats utilised the Graph Markup Language for graph model and a set of ESRI shapefiles for geographical data.

5.2.2 Software Components

A software implementation was constructed to assess the feasibility of the modelling and analysis techniques presented in this thesis. Furthermore, a suitable user interface for visualisation was implemented for user-level situational awareness tests, which are to be conducted at the end of the research project. The resulting software framework is powered by the GraphStream library[27], as well as QGIS geoprocessing and visualisation components and uses both Java and Python programming languages.

VTT conducted a survey of base station radiation patterns in the target area. Using VTTs Network Planning Tool (NPT)[15], the obtained dataset can provide estimates of cellular coverage on the area, when a base station is non-operational. Using the model presented in this Thesis, the combined framework can be used to estimate how the cellular grid is affected by the loss of power in electric grid segment.

5.2.3 Simulation Results

The target area was successfully modelled with the graph based approach using datasets obtained from Caruna Ltd. and National Land Survey of Finland. Using the software tool and CI model, the extent of disruptions and repair processes can be visualised. NLS data was also used in creating additional map layers (Figure 16).

A series of artificial faults were analysed using the framework. Fault one was total failure of the core electric grid. Fault two was a mid-sized fault situation spanning several days, created using the real storm data. Fault three was a smaller

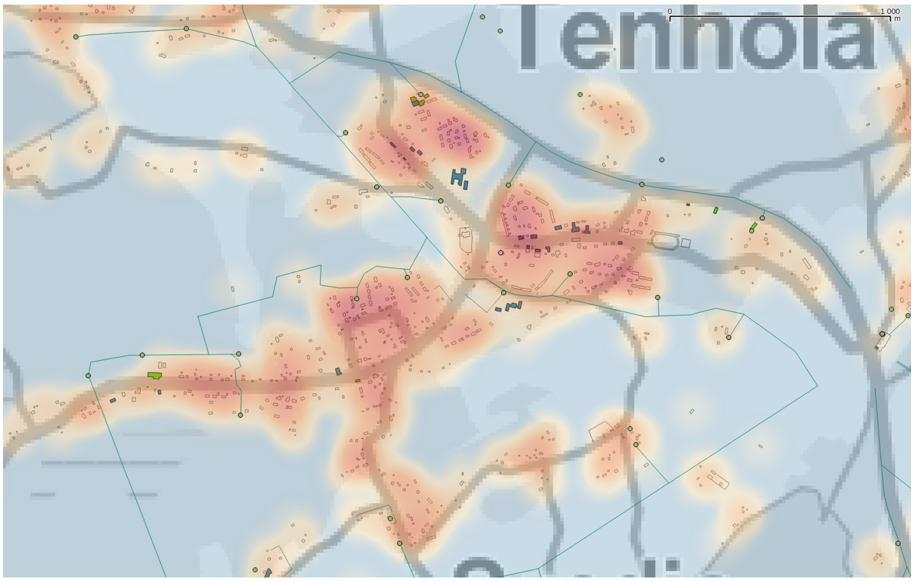


Figure 16: A subsection of the modelled area. Green lines represent power lines and green dots secondary substations. Heat map corresponds to population density. The image is a screenshot of the developed visualisation tool.

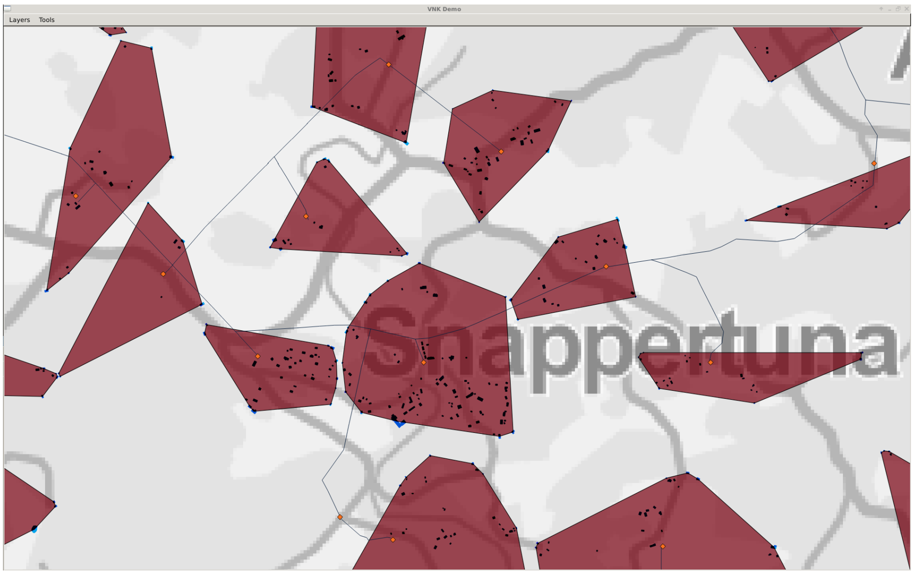


Figure 17: A screenshot of the developed software package. Red convex hulls represent areas without electricity, yellow points are secondary substations.

subsection, affecting 30 electric grid components, one base station, and one critical feature (Figures 19 and 20). All of the faults were caused by a total failure in an electric grid component, where the state was set to Not Operational.

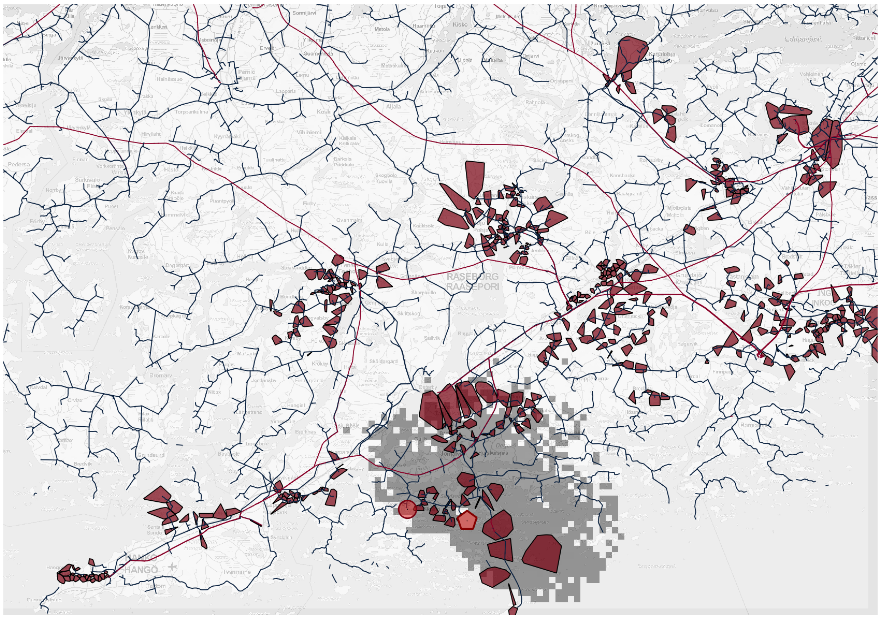


Figure 18: A screenshot of the full area. The grey mosaic layer is an example of cellular coverage disruption. The visualised electric grid disruption is based on the real storm data combined with the CI model.

As indicated by Table 6, the DWIS values increase rapidly, when central components are affected.

Actual storm data was used to calculate the entropy -based measures for one base station. Storm data was collected for a period of 10 days 3 hours and 17 minutes. During this time, eight faults occurred to the secondary substation powering the base station. The target probability distribution was chosen to be $P(O) = 0.2$, $P(M) = 0.4$ and $P(N) = 0.4$, due to the fact that communication loss to base station during storms indicate high fault probability. The selected base station did not have backup batteries, and enters to Not Operational state when the secondary substation powering it fails and powers itself on when the power is restored.

Severity, or the values of S function were $S(O) = 0.1$, $S(M) = 0.5$ and $S(N) = 0.9$; $k = 0.007$. The probabilities for each state are illustrated in Figure 21, and the corresponding entropy measures and expected value in Figure 22. The entropy plot indicates, that the operational state is associated with high uncertainty, indicating unstable operating condition.
The implemented software framework was benchmarked using the described dataset. The update operation for all components takes less than a second, indicating suitable performance with modest hardware. Benchmarks were conducted using commodity laptop with 8 Gigabytes of RAM and Intel i5 mobile processor.

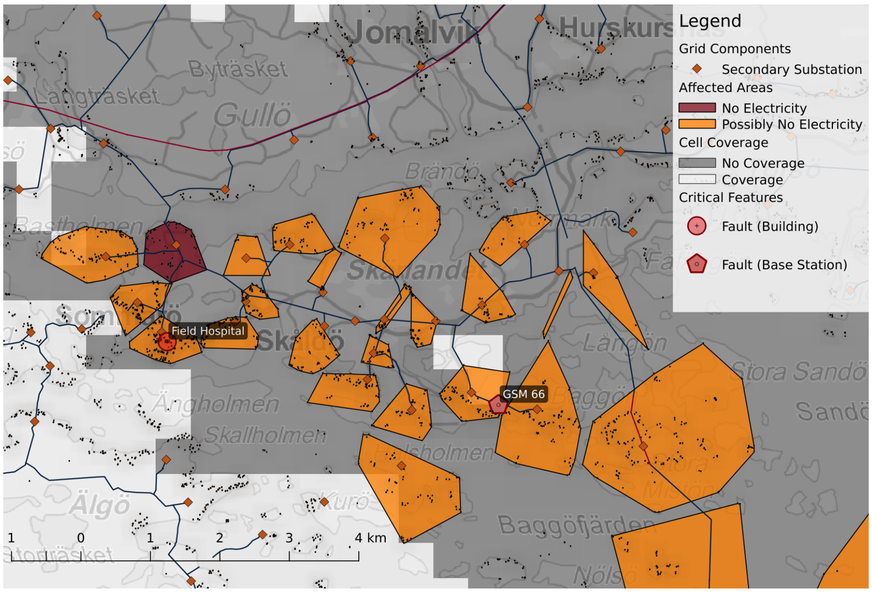


Figure 19: A screenshot of a fault affecting one base station and one critical feature. Note the power loss and communication fault.

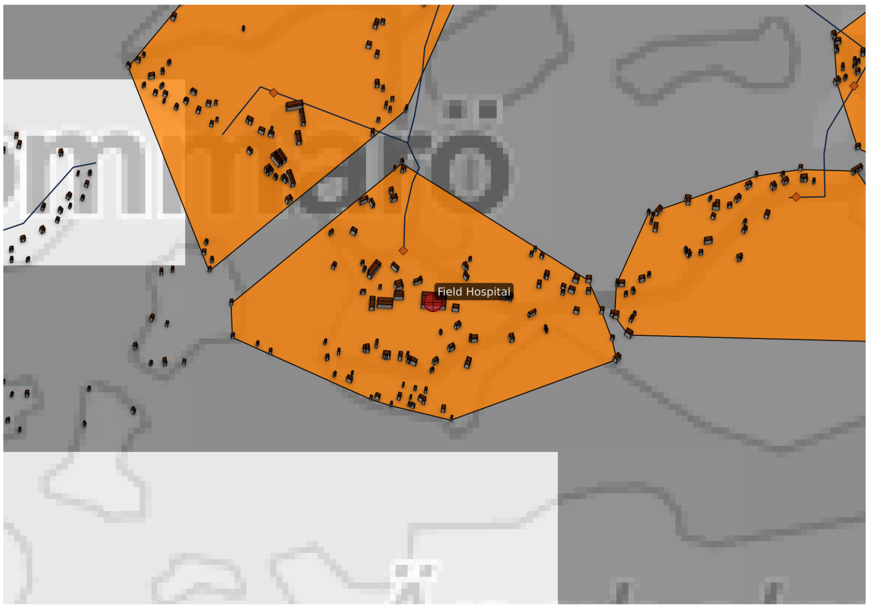


Figure 20: A more detailed view of one critical feature. The software can pinpoint both the exact location and status for each critical feature.

Fault Size	Affected Nodes	DWIS
Full Area	1620	88964.09
Medium	83	5636.69
Small	30	907.12

Table 6: List of DWIS values for faults of differing size. Buildings are not counted towards affected nodes.

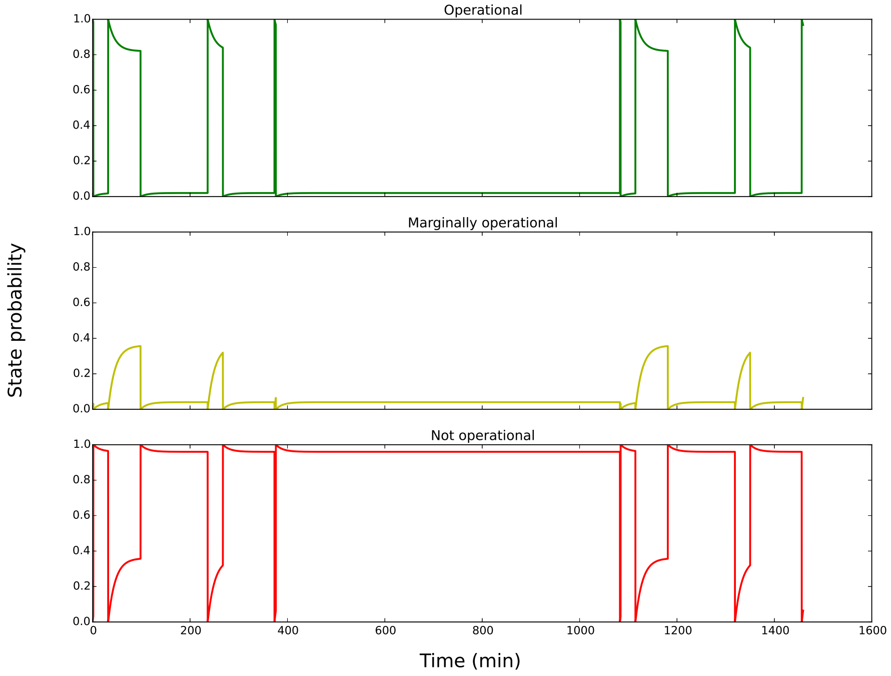


Figure 21: Probabilities for each three operational state for base station during a storm. The base station enters to state [N] at time $t = 0$.

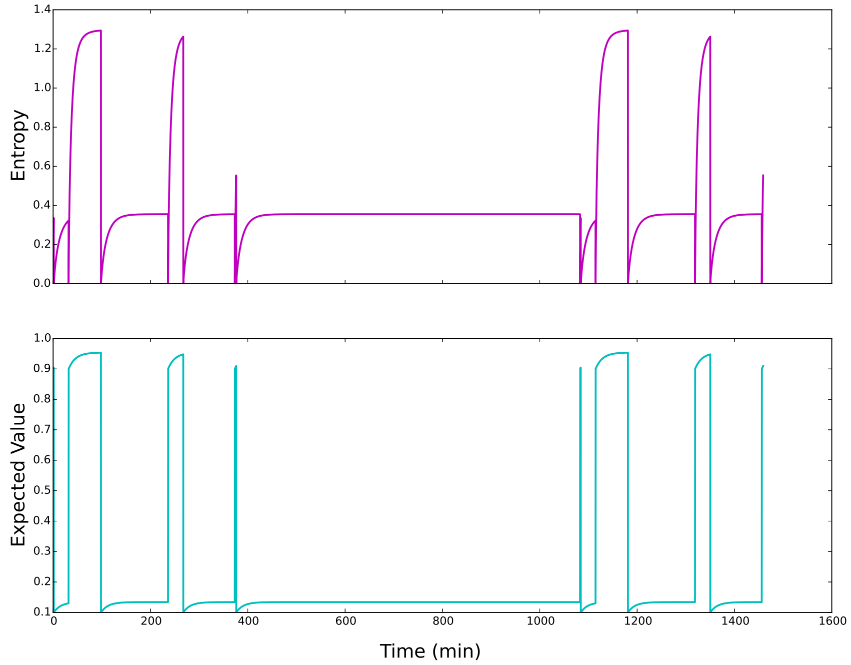


Figure 22: Expected value of S, and entropy H (measured in bits) for base station during a storm.

6 Results and Discussion

In this thesis we have presented a novel approach for CI modelling and analysis by combining graphs and finite state machines. Additionally, we explored the possibility of using entropy -based measures to estimate operational status of systems that are off-line or otherwise unable to send status information.

The design goals were set based on a set of information requirements for critical infrastructure monitoring operator, as defined by Rummukainen et al[34]. Additional requirements imposed by the need for real-time monitoring capability were also honoured. Model data requirements and suitable abstraction levels were also taken into account.

Several man-made networks were modelled as graphs, and analysed to find suitable modelling formalisms. Based on the observed topological structure, a graph based approach combined with finite state machines was used to model infrastructure dependencies and operational status. Other necessary elements were incorporated as part of the modelling formalism for achieving the required modelling power.

Analysis methods suitable for real-time operation were developed to complement the created modelling tools. Using the graph topology and relative changes in operational status, a method was developed for quantifying both system-specific and infrastructure-wide impact of disruptions. Entropy -based analysis method was created for situations, where current data is not available, and an estimate based on previously observed events must be made.

The modelling and analysis methods were tested using data collected during a real storm, and by simulation tools. A coastal area of Finland was modelled using both public and non-public datasets. Modelling of the area proved cumbersome, but doable using semi-manual script assisted processing. Using both public and non-public dataset, a suitable graph model could be constructed. Based on the evaluation results, the presented methods capture the extent and magnitude of different types of disruptions, and can operate on limited input data.

Overall, meeting the set requirements and simulation based evaluations indicate that the presented methods can be used in real-time situational awareness applications. As the information requirements were kept relatively modest, and it is likely that infrastructure operators may be willing to share data at this abstraction level.

6.1 Datasets

There has been much discussion about the interdependencies between CI systems and sectors. Although this is known, there has been relatively little success in obtaining the datasets that describe those dependencies. Indeed, some authors claim that it is impossible to map the interdependencies at the national scale. There exists only a few datasets concerning the dependencies, and essentially none at all contain several CI sectors. Although modelling and analysis of CI has been a hot topic for quite a while, public datasets that are suitable for evaluating the work are still almost non-existent. In some cases, the companies that run CI systems are unwilling to share their data, for academic purposes. This is somewhat understandable, because this data is often thought as company confidential, and sharing is not seen as directly beneficial. Even though the governmental authorities are able to access this information, it is usually classified, and can't be readily used in academic research.

For this thesis, a major part of the dataset was constructed form what the author believes is the best and onliest public data source available for this purpose in Finland; the National Land Surveys database. The dataset contains many features useful for building maps, but the relationships between objects are not included. Moreover, the dataset is not accurate or consistent enough for building e.g. graphs form the electric grid automatically, so the process is essentially manual. Furthermore, the data only gives limited insights to the interdependencies between objects, and it does not contain information about possible failures and their propagation. Open-source efforts have been launched to crowd-source e.g. the locations of base stations, but the data is somewhat inaccurate especially in less-travelled places.

The lack of data makes many data mining and other advanced approaches that demand large datasets unsuitable. It would be advantageous for society if governments actively pushed for open data, which could then further be distributed to the scientist. Data sharing should be seen as beneficial between industry sectors, so that every participant can benefit from advanced data fusion and common operating picture.

6.2 Centrality Measures and Other Parameters

In this thesis the betweenness centrality was used to rank infrastructure components. However, a purely topological centrality value does not provide the best estimate, since additional data on the importance of components is often available. A suitable way to weight the centrality against a list of known important components would provide more accurate impact estimates. Furthermore, the scaling of the chosen centrality measure may prove to be as important as the centrality measure itself.

6.3 User Tests

Situational awareness is ultimately tied to the human element. The ability to provide data that can be transformed into actionable information is the ultimate benchmark for a decision support system. User test should be conducted for assessing the actual impact of the proposed methods towards situational awareness.

7 Acknowledgments

The work presented in this thesis was funded by TEKES, The Prime Minister's Office of Finland, and National Defence University. Their support is gratefully acknowledged. The author also wishes to thank VTT, Caruna Ltd., and The National Land Survey of Finland for generously providing valuable datasets for scientific use.

Bibliography

[1] *Government's analysis, assessment and research activities.* http://tietokayttoon.fi/en/putting-knowledge-to-use. Accessed: 2016-04-11.

[2] *Kriittisen infrastruktuurin tilannetietoisuus [situational awareness on critical infrastructure].* http://tietokayttoon.fi/hankkeet/hanke-esittely/-/asset_publisher/kriittisen-infrastruktuurin-tilannetietoisuus. Accessed: 2016-04-11.

[3] *Security strategy for society 2010.* www.yhteiskunnanturvallisuus.fi/en/materials/doc_download/26-security-strategy-for-society.

[4] Albert, Réka and Barabási, Albert László: *Statistical mechanics of complex networks.* Rev. Mod. Phys., 74:47–97, Jan 2002. http://link.aps.org/doi/10.1103/RevModPhys.74.47.

[5] Bastian, Mathieu, Heymann, Sebastien, and Jacomy, Mathieu: *Gephi: An open source software for exploring and manipulating networks.* 2009. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[6] Clinton, W: *Presidential decision directive 63.* The White House, Washington, DC, 1998. http://fas.org/irp/offdocs/pdd/pdd-63.htm.

[7] Cover, Thomas M and Thomas, Joy A: *Elements of information theory.* John Wiley & Sons, 2012.

[8] Diestel, Reinhard: *Graph Theory.* Springer, Berlin, 4th edition, 2010.

[9] Endsley, Mica R: *Toward a theory of situation awareness in dynamic systems.* Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1):32–64, 1995.

[10] Endsley, Mica R and Jones, Debra G: *Designing for situation awareness: An approach to user-centered design.* CRC Press, 2012.

[11] European Parliament and Council of the European Union: *COUNCIL DIRECTIVE 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection.* Official Journal of the European Communities, 2008.

[12] Freeman, Linton C: *A set of measures of centrality based on betweenness.* Sociometry, pages 35–41, 1977.

[13] Giacobe, Nicklaus A: *Application of the jdl data fusion process model for cyber security*. In *SPIE Defense, Security, and Sensing*, pages 77100R–77100R. International Society for Optics and Photonics, 2010.

[14] Horelli, Ilkka: *Tapaninpäivän 26.12.2011 myrskytuhot Lounais-Suomessa*. Technical report, Lounais-Suomen aluehallintovirasto, February 2012, ISBN 978-952-5882-00-1. `https://www.avi.fi/documents/10191/56990/Myrskyraportti+8.6.2012+LSAVI.pdf/5feb9ee3-426c-4806-99f7-220c2dd59955`.

[15] Horsmanheimo, Seppo, Maskey, Niwas, Kokkoniemi-Tarkkanen, Heli, Savolainen, Pekka, and Tuomimaki, Lotta: *A tool for assessing interdependency of mobile communication and electricity distribution networks*. In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pages 582–587. IEEE, 2013.

[16] Klemetti, M, Puuska, S., and Vankka, J.: *Entropy measures in critical infrastructure graphs*. Proceedings of the 7th conference of the International Society of Military Sciences, 2015.

[17] Lääperi, L. and Vankka, J.: *Architecture for a system providing a common operating picture of critical infrastructure*. In *Technologies for Homeland Security (HST), 2015 IEEE International Symposium on*, pages 1–6, April 2015.

[18] Lewis, Ted G.: *Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation*. Wiley-Interscience, 2006, ISBN 0471786284.

[19] Luiijf, Eric, Nieuwenhuijs, Albert, Klaver, Marieke, Eeten, Michel van, and Cruz, Edite: *Empirical findings on critical infrastructure dependencies in europe*. In Setola, Roberto and Geretshuber, Stefan (editors): *Critical Information Infrastructure Security*, volume 5508 of *Lecture Notes in Computer Science*, pages 302–310. Springer Berlin Heidelberg, 2009, ISBN 978-3-642-03551-7. `http://dx.doi.org/10.1007/978-3-642-03552-4_28`.

[20] Maanmittauslaitos: *Maastotietokanta*. `http://www.maanmittauslaitos.fi/en/digituotteet/topographic-database`, 2014. [Online; accessed 10-2014].

[21] Mason, C. Russell: *The art and science of protective relaying*. `http://www.gegridsolutions.com/multilin/notes/artsci/artsci.pdf`.

[22] Mealy, George H.: *A method for synthesizing sequential circuits*. Bell System Technical Journal, The, 34(5):1045–1079, Sept 1955, ISSN 0005-8580.

[23] Murray, Alan T and Grubesic, Tony H: *Overview of reliability and vulnerability in critical infrastructure*. In *Critical Infrastructure*, pages 1–8. Springer, 2007.

[24] Newman, Mark EJ: *Power laws, pareto distributions and zipf's law*. Contemporary physics, 46(5):323–351, 2005.

[25] Ouyang, Min: *Review on modeling and simulation of interdependent critical infrastructure systems*. Reliability Engineering & System Safety, 121(0):43 – 60, 2014, ISSN 0951-8320. `http://www.sciencedirect.com/science/article/pii/S0951832013002056`.

[26] Pederson, Peter, Dudenhoeffer, D, Hartley, Steven, and Permann, May: *Critical infrastructure interdependency modeling: a survey of us and international research*. Idaho National Laboratory, pages 1–20, 2006.

[27] Pigné, Yoann, Dutot, Antoine, Guinand, Frédéric, and Olivier, Damien: *Graphstream: A tool for bridging the gap between complex systems and dynamic graphs*. CoRR, abs/0803.2093, 2008. `http://arxiv.org/abs/0803.2093`.

[28] Proakis, John G and Salehi, Masoud: *Communication systems engineering*. Pearson, 2001.

[29] Puuska, Samir, Kansanen, Kasper, Rummukainen, Lauri, and Vankka, Jouko: *Modelling and real-time analysis of critical infrastructure using discrete event systems on graphs*. In *Technologies for Homeland Security (HST), 2015 IEEE International Symposium on*, pages 1–5, April 2015.

[30] Puuska, Samir, Rummukainen, Lauri, Timonen, Jussi, Lääperi, Lauri, Klemetti, Markus, Oksama, Lauri, and Vankka, Jouko: *Nationwide critical infrastructure monitoring framework and its implementation*. Preprint submitted to The International Journal of Critical Infrastructure Protection, September 30, 2015.

[31] QGIS Development Team: *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2009. `http://qgis.osgeo.org`.

[32] Rinaldi, S.M., Peerenboom, J.P., and Kelly, T.K.: *Identifying, understanding, and analyzing critical infrastructure interdependencies*. Control Systems, IEEE, 21(6):11–25, Dec 2001, ISSN 1066-033X.

[33] Rummukainen, Lauri, Oksama, Lauri, Timonen, Jussi, and Vankka, Jouko: *Visualizing common operating picture of critical infrastructure*. In *SPIE Sensing Technology+ Applications*, pages 912208–912208. International Society for Optics and Photonics, 2014.

[34] Rummukainen, Lauri, Oksama, Lauri, Timonen, Jussi, and Vankka, Jouko: *Situation awareness requirements for a critical infrastructure monitoring operator*. In *Technologies for Homeland Security (HST), 2015 IEEE International Symposium on*, pages 1–6, April 2015.

[35] Säteilyturvakeskus: *Reactor Coolant Circuit of a Nuclear Power Plant*. Regulatory guide, Säteilyturvakeskus, November 2013, ISBN 978-952-309-083-5. `http://www.finlex.fi/data/normit/41775-YVL_B.5e.pdf`.

[36] Shannon, C. E.: *A mathematical theory of communication*. The Bell System Technical Journal, 27(3):379–423, July 1948, ISSN 0005-8580.

[37] Steinberg, Alan N, Bowman, Christopher L, and White, Franklin E: *Revisions to the jdl data fusion model*. In *AeroSense'99*, pages 430–441. International Society for Optics and Photonics, 1999.

[38] Timonen, Jussi, Lääperi, Lauri, Rummukainen, Lauri, Puuska, Samir, and Vankka, Jouko: *Situational awareness and information collection from critical infrastructure*. In *Cyber Conflict (CyCon 2014), 2014 6th International Conference On*, pages 157–173. IEEE, 2014.

[39] Watts, Duncan J and Strogatz, Steven H: *Collective dynamics of 'small-world' networks*. nature, 393(6684):440–442, 1998.