A Complex Network Analysis Approach for Risk Increase Factor Prediction in Nuclear Power Plants

Mouna Rifi[1,2], Mohamed Hibti[1] and Rushed Kanawati[2]
[1]*EDF R&D, 7, Boulevard Gaspard Monge, Palaiseau, France*
[2]*LIPN, Université Paris 13, 99, avenue Jean-Baptiste Clément, Villetaneuse, France*

Keywords: Nuclear Safety, Probabilistic Safety Assessments (PSA), Complex Networks, Attribute Prediction, Centralities.

Abstract: We explore applying network based metrics to predict safety metrics of components in Nuclear Power Plants (NPP). We first show how to model accident sequences as complex networks, then we conduct a statistical study over the main network metrics to show that these are highly correlated with the RIF (Risk Increase Factor) which is a very popular metric in nuclear safety studies.

1 INTRODUCTION

A Nuclear Power Plant is a complex system for which the safety studies drive all the life cycle, starting from the design phase and ending with the dismantling step passing through both licencing and power generation phases. These studies are mainly based on deterministic approach. However, probabilistic safety studies are now a recognized tool to achieve a more global and systematic verification using quantitative insight. This probabilistic approach is called Probabilistic Safety Assessment (PSA) (WAS, 1975).
Notwithstanding the success that has PSA as a tool for analyzing, priorizing and quantifying the risk in NPP, it presents mainly two limitations:

- PSA computations are based on the reduction to its normal disjunctive form of a boolean formula corresponding to the so called Master Fault Tree representing all the paths leading to an undesired events. This problem is known to be an NP hard (NP referring to non-deterministic polynomial time) problem (Bollig and Wegener, 1996) (Friedman and Supowit, 1990). Therefore, many approximations are made to get the cutset list: eleminating negligible cutsets and using approximative formulae (i.e. Min Cut Upper Bound or Sylvestre-Poincarré of first order to sum of disjunct products).

  Some critical components may then be neglected in the study if they dont appear in these frequent cutsets.

- Even with approximations, PSA computation times stay instable for big models (a small modification can make computation time change from polynomial to exponential), which is not appropriate in an industrial context.

In (Hibti et al., 2016), some similiarities where showed between complex network metrics (e.g. betweenness centrality) and the multiple occurrences of component failures in the cutset list wich is also an indicator of safety importance.
In this paper, we propose to use complex networks analysis to identify components that might be important based on their position on the system using the different network centralities. We first provide a method of modeling a set of sequences into networks, then we study the association between the Risk Increase Factor (RIF) and different vertices centralites in the obtained complex networks. Our aim here is to predict high values of RIF which is a PSA metric representing the increase of the risk. The remainder of the paper is organized as follows: in the second section we present some PSA notions; in the third section we describe how to model a Functional Requirement Diagram (FRD) as a directed network, model a real case of PSA, we give some specificities of the obtained network. In the next section we perform some statistical approaches (Tree classification and Logistic regression) to study the RIF using the directed network centralities, in other words, we aim to predict an attribute using topological attributes of the obtained complex network vertices.

Rifi, M., Hibti, M. and Kanawati, R.
A Complex Network Analysis Approach for Risk Increase Factor Prediction in Nuclear Power Plants.
DOI: 10.5220/0006700000230030
In *Proceedings of the 3rd International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2018)*, pages 23-30
ISBN: 978-989-758-297-4

2 PROBABILISTIC SAFETY ASSESSMENT AND ITS IMPORTANCE METRICS

In this section, we explain the most basic notions of the Probabilistic Safety Assessments, then we present some risk importance metrics mainly used in the industry.

Probabilistic Safety Assessment (PSA) also called "Probabilistic Risk Assessment", with its systematic investigation approach, is used as a complement to the deterministic approaches. The genesis of modern PSA methodology was the Reactor Safety Study (RSS) known as WASH-1400 (WAS, 1975).

It allows to prioritize the undesired event into initiating events, and establish for each initiating event accident sequences resulting from the success or failure of the mitigation actions.

By studying the NPP as an integrated system, including both technical and socio-organisational aspects, PSA supports risk management, identify the accident sequences or scenarios, determinates how likely these scenarios occur, establish for each one the potential consequences. Thus, it provides an input to for risk informed decision making (maintenance activities, plant modifications, graded quality assurance...).

PSA studies are organized in three levels, we focus on the level one which estimates the likelihood that a reactor core could be damaged (melt).

A PSA study starts by identifying the spectrum of initiating events that could possibly lead to the undesired event (for instance Reactor core melt or massive and early release).
An initiating event could be an important component failure, an internal hazard like a fire or an external hazard like a flood in the plant site. For each initiating event an accident sequence analysis is performed regarding the plants response to the initiating event effects.
All the senarios starting from the initiating event and going through the success or failures of the safety systems or human actions dedicated to mitigate the effects of the initiating event are identified.

These sequences are summarized in the form of event sequence diagrams (Functional Requirement Diagrams), which are converted into Event Trees (for each initiating event).
The sequences of the Event Trees are successions of succes or failure of mitigation missions expressed in Faul Trees. Those leading to unacceptable consequences are evaluated (their frequencies are calculated and their cutsets are listed) using Fault Tree Analysis.

Fault Trees Analysis (Ericson, 1999) models the plant systems in detail. It is used by the analysts to identify the causes of a system failure (which is a cascade of OR and/or AND gates of potential component failure). It also compute the failure probability of each mission. To quantify the global risk, the *Boolean Fusion* (Knuth, 2007) is used to get the normal disjunctive form. Thus, the risk is the sum through different minimal cut sets' probability.

A PSA study involves different input data such as:

- Failure probabilities of different data plant's components. These probabilities are either fixed by feedback records or by statistical methods.
- Initiating events frequencies, which are the occurrence frequency of a fault or hazard that may lead to the undesired event.
- Common Cause Failure, which is the failure of multiple components, belonging to the same system, due to a single specific cause.
- Human Error Probabilities, represents failure of an operator to conduct a manual actions. These probabilities are estimated via Human Reliability Analysis.

2.1 PSA Component's Importance Metrics

PSA has some importance measures used to identify the role of components in the risk of a NPP. If Fussell-Vesely (FV) measure is usually used to quantify the risk importance of a component, Risk Increase Factor (RIF) is mostly used in measuring the safety-importance of this component (Van der Borst M., 2001). In table 1 we present the definitions of these importance measures for a given component x_i.

The following definitions are used in table 1:

$R(x_i = 0)$ is the risk estimation if we assume that the component x_i is perfectly reliable; $R(x_i = 1)$ is the risk estimation supposing that the component x_i is failing or absent. R_{base} represents the risk estimation without any assumption on the component.

3 CONSTRUCTION OF THE NETWORK OF AN EVENT SEQUENCE DIAGRAM

As mentionned before, one of the PSA's steps is the accident sequence analysis where accident contexts are defined and the sequences of successive mitigation functions are described. The progression of the accident could go depending on the success or failure

Table 1: Risk Importance Measures.

Measure	Abbreviation	Pinciple
Risk Decrease	$RD(x_i)$	$R_{base} - R(x_i = 0)$
Fussell-Vesely	$FV(x_i)$	$\frac{R_{base} - R(x_i=0)}{R_{base}}$
Risk Decrease Factor	$RDF(x_i)$	$\frac{R_{base}}{R(x_i=0)}$
Criticality Importance	$CR(x_i)$	$\frac{R(x_i=1) - R(x_i=0)}{R_{base}}$
Risk Increase	$RI(x_i)$	$R(x_i = 1) - R_{base}$
Risk Increase Factor	$RIF(x_i)$	$\frac{R(x_i=1)}{R_{base}}$

of the corresponding mitigation function either to a state where the plant is considered safe or at least controlled (acceptable consequence) or to an undesired outcome such as the plant core melt or any degraded state (unacceptable consequence).

These functions represent the system missions, instrumentation and control missions that are initiated either automatically or manually by the operators, in addition to human factors representing failure or success of human actions. It considers the consequences of the potential failures on the NPP's safety, and produce a set of Functional Requirement Diagrams (FRD) (Swaminathan and Smidts, 1999). The figure 1 represents a simplified FRD corresponding to an initiating event (IE), involving 3 mitigation missions, with 2 consequences (acceptable: AC and unacceptable: UA). A red transition is a failure of a mission and a green one represents a success.

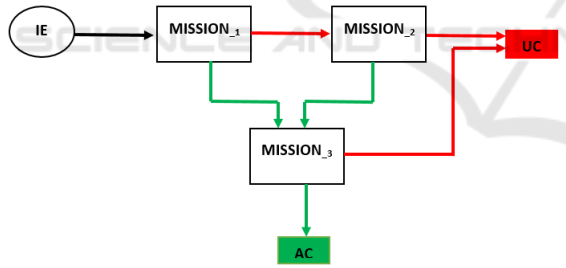

Figure 1: A simplified example of a Functional Requierement Diagram.

3.1 Methodology

In this section we describe the methodology to build a network of an accident sequence. An accident sequences analysis in PSA is summarized in a FRD. So our method aims to produce a network for each FRD. We first produce a skeleton network using as edges the different transitions leading to an unacceptable consequence (UA) and as "macro-vertices" the differents missions. Then we develop each "macro-vertex" into a network. We remind that each mission corresponds to a particular NPP system in a particular configuration. Thus, the network of a "macro-vertex" represents the system's components (pumps, diesel, valve, acquisition,...) as vertices, and the different flows (fluid, electrical, and signal) between these components are the directed edges, in addition to the directed edges previously introduced which represents the transitions of the FRD. These different components are the vertices and these flows are the edges.

So, we obtain a network which models all the safety systems used in the accident sequence mitigation for an initiating event in a state of the NPP; in other words, a network for each FRD.

The obtained network involves different types of edges (hydraulic, electrical, signal,..) and different types of vertices (pumps, valves, diesels,...). Each type of the modeled components has a specific caracterizing attributs that we choose to consider as attributes of the vertices. So, the vertices are the safety systems components, and the edges are the differents flows between those).

For this study, we choose to consider the network as simple (no type of edges) and directed since the edges are oriented and we take in consideration as attributes for the vertices, only the RIF and the network centralities of the vertices.

3.2 A Real Case of Study

As an application case, we model the mitigations actions performed in order to conduct the reactor to a safe state in the wake of the occurence of the "Uncontrolled Level Drop" (ULD) initiating event into a network for the European Pressurized Reactor in a shutdown state.

To mitigate the effects of the initiator we have to ensure three safety functions: The first one is the inventory control of the primary fluid, the second one is the evacuation of the residual heat and the third one is maintaining the integrity of the containment.

Using the method described before, the directed network of the ULD has the following specificities: The obtained network is relatively small. It characteristics are summarized in the table 2. We recall that the clustering coefficient of a graph measures the probability that the adjacent vertices of a vertex are connected. Network analysis metrics reveal some char-

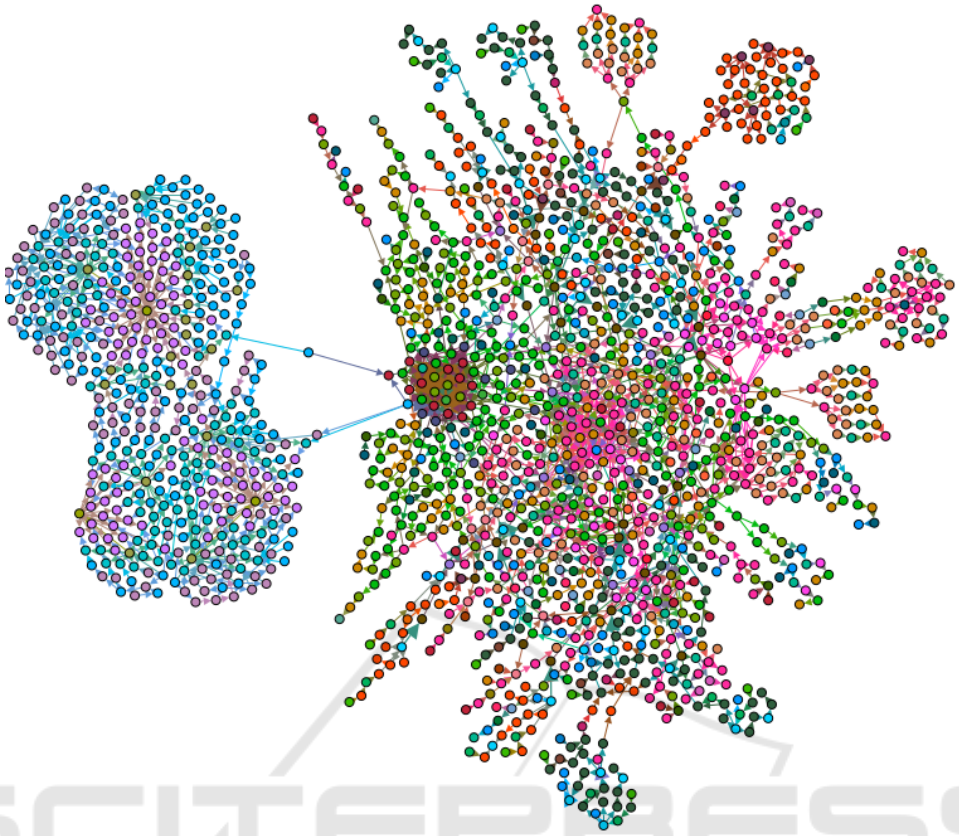

Figure 2: Network of the study “Uncontroled Level Drop”; colors represent the different type of components.

Table 2: Summary of the ”Uncontroled Level Drop” network caracteristics.

Network	Directed & Attributed
# Vertices	$N = 1700$
# Edges	$M = 2700$
Density	0.009
Diameter	33

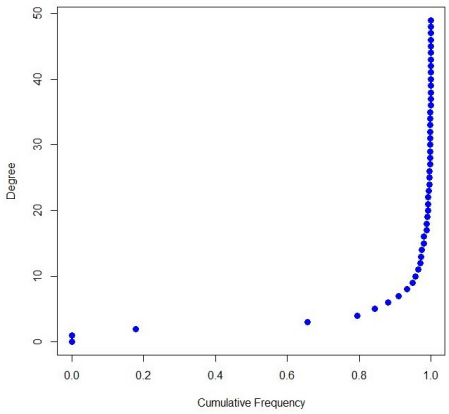

Figure 3: Degree distribution in the network of “Uncontroled Level Drop”.

acteristics of the structure of our systems, typically types of components which are important for each metrics, and component’s systems beloging for each metrics. The figure 2 represents this network where colors represents the type of component modeled.

The degree distribution of this network is illustrated in the figure 3.

4 STATISTICAL STUDIES OF THE ATTRIBUTE RIF USING NETWORK CENTRALITIES

We aim here to predict the RIF variable, which as mentionned before is a measure of the safety-importance of a component in PSA, using network centralities.

The considered centralities are In-Degree, Out-Degree, In-Closeness, Out-Closeness, Betweeness (Freeman, 1979), Page-Rank (Brin and Page, 1998), Hub-Score and Authority (Kleinberg, 1999).
So, our objective is to find:

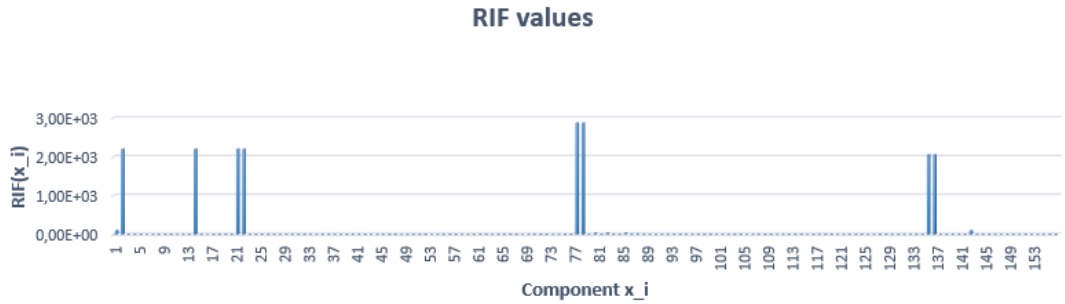

Figure 4: RIF observation values.

RIF(x)=f(In-Degree(x), Out-Degree(x), In-Closeness(x), Out-Closeness(x), Betweeness(x), Page-Rank(x), Hub-Score(x), Authority(x))

The RIF formula for each component x_i is presented before in table 1. This computation is intractable since it involves computing the core damage risk under the assumption that the studied component is failing (or absent) which is already instable. So, computing RIF values of many components is laborious, therefore, we have decided to build a small sample composed by vertices found to be important by each network centrality (highly ranked). We selected 20 best ranked vertices by each centrality. The sample contains 156 observations obtained as explained above from around 1700 individuals (the network vertices). To perform the prediction, we use two famous methods of classification in datamining which are "Classification Tree" and "Logistic Regression"..

4.1 The RIF Variable

The RIF values taken over the sample are represented in figure 4.

We observe in the figure 4 that the RIF variable takes the same value for most of the observations except some outliers. These outliers correspond to NPP components whose failure have a high impact on the Core Melting Frequency of the NPP (safety-important components).

Moreover, the standard deviation of the RIF variable is 4 times of the mean, this indicates that this variable is very scattered and thus, it is difficult to adjust.

We can also based on the figure 4 simplify our model by discretizing the RIF variable into two modalities:

- high RIF ($RIF > 2$) becomes the category**RIF=1**;
- low RIF ($RIF < 2$) becomes the category **RIF=0**.

4.2 Reduction and Discretization of Predictors

In this part, we want firstly to identify the possible correlations between the predictors using a graphical analysis of the values distributions which are represented in the figure 5.

According to the histograms shown in the figure 5, the In-Degree, the Out-Degree and the Betweenness appear to havea similar values distribution, Hub Score and Authority seem to be similar too.

Therefore, we take as an input for the logistic regression performed later only the variables In-Degree, In-Closeness, Page Rank and the Authority.

For the sake of precision, we choose to discretize the selected predictors according to the discretization revealed by their histograms, tables of RIF by each predictors allow us to merge some categories for some predictors. Thus, we obtain 2 modalities for each predictor. We recode the selected variable as presented in table 3.

Table 3: Discretization of In-Degree, In-Closeness, Page Rank and the Authority.

In-Degree	In-Degree.b
(0.0000,0.0005]	1
else	0
In-Closeness	**In-Closeness.b**
[0.0000,0.000783]	1
else	0
Page-Rank	**Page-Rank.b**
[0.0000,0.00026]	1
else	0
Authority	**Authority.b**
[0.0001, 0.0251]	1
else	0

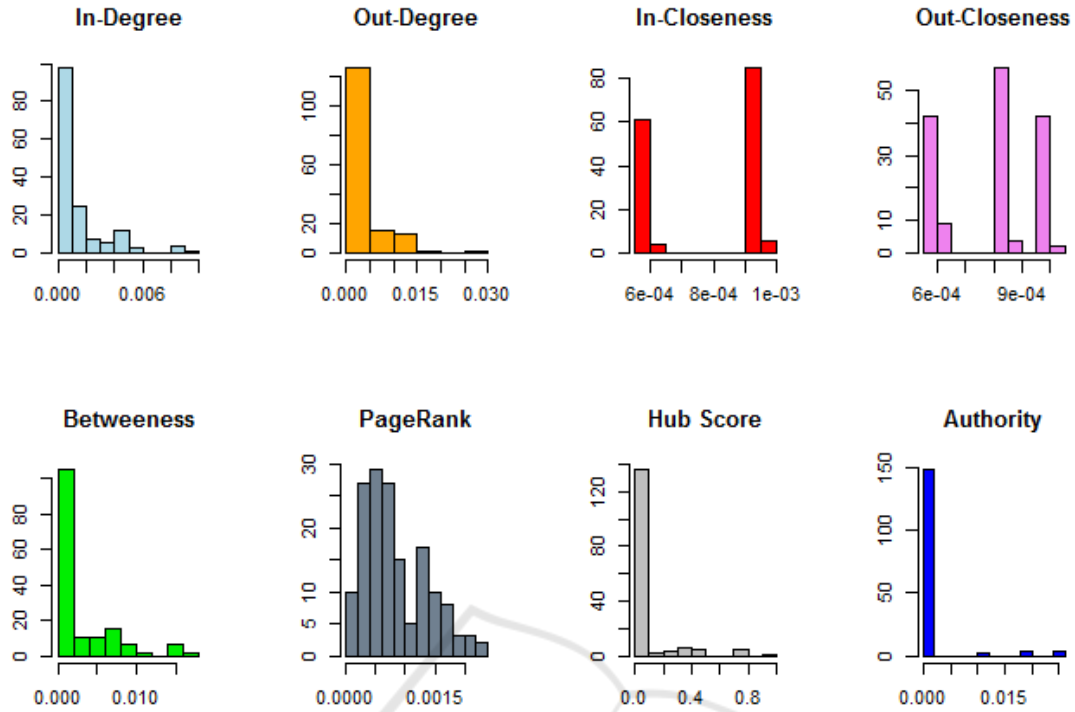

Figure 5: Values distributions of each network metric.

4.3 Classification Tree

The Classification Tree or "Decision Tree" divides the predictor variables into different regions so that the dependent variable ("target variable") can be predicted more accurately. This method is one of the Statistical Learning methods (Hastie et al., 2001) used when the target variable is categorial, by contrast with the Regression Tree which is used when the target variable is continuous.

We apply this method using all the predictors (continuous values) and without any variable elimination in order to predict the RIF categories (0 and 1).

We denote that this method considers that the RIF's predictors are the In-Degree, Out-Degree, the In-Closeness, the Out-Closeness, the Hub-Score and the Authority and however, eleminates the Betweeness and the Page-Rank.

Table 4: Confusion matrix of the Classification Tree.

From / To	0	1	Total	% Correct answer
0	148	0	148	100%
1	5	3	8	**37.50%**
Total	153	3	156	96.79%

The confusion matrix for this method shows a success of 96.79% to predict the value of RIF which is good for a global prediction of RIF classes. However, we focus on the $RIF = 1$ category (high values of RIF) where this method detects only 3 out of 8 values 1 which represents a recall of 37.50%.

4.4 Logistic Regression Study

Our aim, in this part, is to detect the high values of RIF using vertices topological metrics (In-Degree, Out-Degree, In-Closeness, Out-Closeness, Betweenness, Page Rank, Hub Score and Authority) which describe the components of the sample. We use in this part the logistic regression in order to construct the model of classification.

To perform the logistic regression, we consider the selected predictors discretized as described in the section 4.2.

It is known that the Logistic Regression is one of the most common model analysis for multivariate problems. It is also a method of Statistical Learning (Hastie et al., 2001). It gives the possibility to measure the relation between an event's occurrence, which is a qualitative dependent variable, and the potential predicting variables. The choice of explicative variables to be consideried in the logistic regression model is based on prior knowledge of the studied phenomena, which is in our case high values of the RIF, and the statistical association between the variable and the RIF.

Table 5: Training Result of the logistic regression.

RIF / RIF_{Pred}	0	1
0	91	57
1	0	8

Table 6: Confusion Matrix of the logistic regression.

From / To	0	1	Total
0	91	57	148
1	2	6	8
Total	93	63	156

Considering the small size of our sample and the fact that the class to predict is rare, we decide to use the Leave One Out Cross Validation method to validate this model in order to have more data to build the model, and many testing samples. We recall that Leaveone_out cross validation method consists of leaving one observation for the validation and building the model over the other $(n-1)$ observations; and performing this for each observation of the initial sample.

Training Study. Let's call $Mpred$ the matrix of prediction of P($RIF = 1/X$) where X is an observation. The column c_k is obtained from the logit model where the training sample is obtained by excluding the k considered for the test. Thus, $Mpred$ corresponds to $n = 156$ logit models. We observe that values of $Mpred$ are too small, in other words if we compare to a $sensitivity = 1/2$ we will obtain just values 0. We suggest to set the sensitivity of our model to 0.4; we also compare the value predicted, for each observation, to the mean value of the predictions obtained over the different training samples. for example: for an observationX_i=(InDegree(x_i), InCloseness(x_i), PageRank(x_i), Authority(x_i)) we compute the mean $m(X)$ of all RIF prediction obtained by each training sample (156 sample), if the prediction is $>= m(X)$ we consider that it is a high value that we replace by one, otherwise we replace it by 0. In this way, we obtain the binary prediction matrix$Mpred_b$. Finally, for an observation X_i, we consider that the prediction obtained for RIF is 1 if over the (n-1) training prediction the majority are 1 otherwise the training prediction takes 0.

The average training error obtained for the model is 36.5%. But this model doesn't make any misclassification for the class $RIF = 1$ which is good for our study.

Test Study. Concerning the prediction obtained by the test samples, they are represented in $Mpred$ in the diagonal $Mpred[k,k]$. Values are also small, we use the same method used in the training (compare the prediction obtained to the mean of prediction obtained) if the prediction is higher then this mean it is considered as 1 otherwise it is considered 0. the logistic model has the confusion matrix represented in the table 6.

In the safety studies, we focus more on the recall then on the precision. The recall($RIF = 1$) obtained by this method is 75% which is a good result, the test error obtained for this model is 38%(misclassification($RIF = 1$) is 25%. We performed the same study on the preductors with continuous values the recall(RIF=1) is almost the same, which mean that our discretization is good.

For our study, the logistic regression model gives the best result since the recall of $RIF = 1$ is the highest (75%). This classification study can certainely be improuved using a bigger sample.

5 CONCLUSION

In this paper, we describe how to model accident sequences into directed networks, we apply this network-based method to a real case the obtained network is a complex network whith a low diameter which make it a "smallworld network" where the centralities are interested to study. We give some specificities of the network.
We also perform some datamining approaches of classification in order to predict high values of RIF (nonstable variable) using network centralities applied over a set of vertices, we observed that the classification tree gives a global good prediction but lower on the high RIF (which is our target class), the logistic regression is then performed using an adequate discretization of the predictors we could obtain an excellent model with a recall for the class $RIF = 1$ 75% we conclude that the best prediction is obtained by the logistic regression using as predictiors the variables (discretized) In-Degree, In-Closeness, Page Rank and the Authority. In the future works, we will extend the study to a bigger sample to impove the classification, and then predict unknown RIFs. We also plan to test other classification methods like SVM.

REFERENCES

(1975). Reactor safety study. an assessment of accident risks in u. s. commercial nuclear power plants. appendix xi. analysis of comments on the draft wash-1400 report. [pwr and bwr]. Technical report, United States.

Bollig, B. and Wegener, I. (1996). Improving the variable ordering of obdds is np-complete. *IEEE Trans. Comput.*, 45(9):993–1002.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

Ericson, C. A. (1999). Fault tree analysis: A history. *International System Safety Conference*.

Freeman, L. (1979). Centrality in social networks i: Conceptual clarification. *Social Networks,*, Vol. 1:215239.

Friedman, S. J. and Supowit, K. J. (1990). Finding the optimal variable ordering for binary decision diagrams. *IEEE Trans. Comput.*, 39(5):710–713.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hibti, M., Oudjit, A., and Marechal, A. (2016). Exploring "loca" with complex network based analysis: A mean-centrality based criteria. *PSAM*.

Kleinberg, J. (1999). Authorative sources in a hyperlinked environment. *Journal of the ACM*.

Knuth, D. (2007). The art of computer programming. *Addison-Wesley*.

Swaminathan, S. and Smidts, C. (1999). The event sequence diagram framework for dynamic probabilistic risk assessment. *Reliability Engineering & System Safety*, 1:73–90.

Van der Borst M., S. H. (2001). An overview of psa importance measures. *Reliability Engineering & System Safety*, 72:241–245.