Catch and Keep: Malicious URL site detection and mitigation through Machine Learning and Formal Concept Analysis

Zvikomborero Chitiyo[1]zvikochitiyo@gmail.com, Courtney Terence Yon[2]courtneyterrence9@gmail.com, and Indurain Jostein van Staden[3]indurain193@gmail.com

Department of Information Science, Stellenbosch University, South Africa

Abstract. Malicious URLs (Unified Resource Locator) can cause chaos in all types of communities we live in today. Google Inc. reveals that it detects large numbers of new and risky web pages most of which are compromised and authorized web pages. The detection of malicious URLs is popular amongst researchers due to the dangerous nature of malicious URLs. The digital age demands organizations to keep up with this technological pace and therefore, this study focuses on Machine Learning in the form of logistic regression for the accurate prediction and detection of malicious and benign URLs. Logistic regression has become prevalent over the last decade, making numerous amounts of research in higher education journals. This study uses Formal Concept Analysis (FCA) to test whether it is a viable tool when researching cybersecurity datasets which offers approach. Lastly, this study makes a recommendation on the state of the South African IT infrastructure and Heuristic classification.

Keywords: Machine Learning, FCA, Cyber Security, URL Detection, Logistic Regression, Python

1 Introduction

Malicious URLs detection is still one of the most important concerns in cybersecurity. The emergence of present-day telecommunication has led to more sophisticated ways of sending information and communication, in turn, it allowed for the expansion of firms, corporations, and organizations that transcended towards social networks, e-commerce, and online banking.
Since, there are numerous cyber invasions, coupled with the possibilities of novel attack types, society can find it difficult to detect and design for these types of invasions. The urgency to close the gap on cyber criminals becoming increasingly important. This is due to the nature of the technological growth rate as the hasty changes of the latest Information Technology, gadgets, applications and reliance increases. Furthermore, the majority of these cybercrime activities are seen through the proliferation of malicious URLs. The following research paper uses Machine Learning, to detect, classify and mitigate malicious URLs.

Through the use of the logistic regression model, the paper shows the accuracy of the models, discusses the trends and patterns and classification through feature selection. Since FCA studies how objects can be hierarchically grouped with their common attributes this paper utilises the tool to compare malicious sites which adds a new perspective to this domain offering a fresh outlook for Machine Learning.

1.1 Research Aims and Objectives

The aims and objectives help provide more insights into a logistic regression, machine learning detection with malicious URL sites and more importantly FCA. The study tests whether FCA is a viable method to analyse cybersecurity, malicious URL data, the testing for the categorization of features, and the types of malicious URL attacks that took place and the testing of the model's accuracy.

2 Related Works

According to [7] in terms of detection (automation), there has been substantial research and headway into the field, however, the detection of malicious URLs using machine learning is still a troublesome issue. Studies have shown that there seem to be signs of possible future success in terms of the classification of URLs. [13] suggests a method of detection for malicious websites through the use of host-based and lexical features of the URLs due to the rich data and the nature of websites. Their methods allow the viewing of thousands of features as well as the identification of significant URL metadata and elements with no need to use "heavy domain expertise". This method was assessed from 30 000 instances and produced significant results for example - the extremely high rate of classification rate with a 95 - 99 percent coupled with a low false-positive rate. [15] tackled large amounts of 'evolved' URL features through the use of online algorithms. They used a system that collected real-time URL features. This was compared to categorised URLs taken from a big web-mail source. The data saw a classification rate of 99 percent through the use of confidence weighted learning. [18] used Cost-Sensitive Online Active Learning (CSOAL) framework to firstly, negate the necessity of using a large number of training data. Secondly, it was used to negate the class disproportion in the malicious URL detection issue. FCA is an efficient tool for the formalization of model classification as well as machine learning. The lattices allow for exploratory knowledge discovery. According to [12], FCART is a beneficial tool compared to other FCA tools because the scientific community has confidence in it with regards to the preprocessing and analysis with regards to big data with no need for external preprocessing applications. [6] measured the hypothesis based classification in an array of areas with no extra alterations for the respective data sets - utilizing a simple binarisation of numerical data. They concluded with the following: The bi-clustering application which included parameter optimization had a better quality of classification in contrast with the new FCA algorithm. A large sum of classification mistakes took place and an

excessive rate of classification failures. It is possible for the FCA algorithm to generate accurate classification however struggles to class the greater number of their samples.

A study conducted by [2] investigated the popularity of phishing attacks whereby the researchers divided the attacks into four sections namely: target devices, attack techniques, communication media attacks, and countermeasures. They then noted that it was difficult to identify these attacks respectively. [20] discusses how logistic regression has become prevalent over the last decade, making numerous amounts of research in higher education journals. This trend is because academics have available top software for statistical tools for high-end analyses. [5] believes that it is expected that logistic regression will continue to improve in popularity.

3 Datasets and Motivation for Machine Learning

URLs are pivotal components in the domain of online criminal activities. To counter the issues of online cyber warfare, the security community has aided its efforts in developing a vast number of techniques for blacklisting malicious URL sites. The study detected and categorized malicious URLs according to their attack type. Thereafter, it features machine learning mechanisms of predicting which attack class a URL site falls into. The data set used in the study is obtained from the Canadian Institute of Cybersecurity, which is the (ISCX-URL-2016) of 2016[17].
The second dataset is used for the testing of FCA and checks its suitability and efficiency of detection of malicious and benign URLs[19]. Lastly, the study tests for the formalization of model classification and to check for positive results as a potential machine learning classifier.

4 Machine Learning Terms and Notation

4.1 Data Vectorisation

Data vectorisation is the process of converting algorithms from operating on a single value at a time to operating on multiple or a set of values (vectors) in one instance. The focal purpose of vectorising data as researched by[14], is to speed up Python code with the use of a loop function which assists in minimizing the running code of time efficiently. This feature is common across machine learning techniques with a close relation towards tokenization which as studied by [5], and divides texts into words or smaller sub-texts. Furthermore, this enables a good generalized relationship between texts and labels. This generalization within their relationship provides is a vocabulary of the dataset, set to be used which acts as a set of unique tokens present within the data.

4.2 Logistic Regression Model and Confusion Matrix Classifier

Commonly used as an appropriate regression analysis on evaluating a dichotomous dependent variable. The logistic regression model underlies itself as a central mathematical concept with logit as the natural logarithm as an odds ratio [14]. Given the nature of the study, it considers the instance whereby the distribution of a divided outcome variable such as a list containing five different URL attack types undergo machine learning performance measures; whilst paired with a divided predictor variable such as the labelled attack score. [15] states that the logistic regression model is best suited for describing and testing hypotheses concerning relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. The simplest case of a regression model for this study is potentially composed around one continuous predictor X, the list of URL sites that undergo evaluation in being labelled as either a benign, defacement, malware, phishing or spam attacks; and one dichotomous outcome variable of Y, a URL label attack score on whether it: 1. Benign (0): attacks that do not pose any inherent security threat;
2. Spam (1): attacks coming in the form of unsolicited commercial emails that are immediately redirected or spotted as spam;
3. Defacement (2): attacks that alter the visual appearance of websites and web pages;
4. Phishing (3): attacks that come across as fraudulent attempt in obtaining sensitive information of users; or
5. Malware (4): attacks delivered via email, text-messages or pop-up messages.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

- True Positive: corresponds to examples that are correctly labelled as positives
- False Positive: refer to negative examples incorrectly labelled as positive
- False Negative: refer to positive examples incorrectly labelled as negative
- True Negative: correspond to negatives correctly labelled as negative

Metric	Formula
Recall	TP / (TP + FN)
Accuracy	TP / (TP + FP)
Precision	(TP + TN) / (TP + TN + FP + FN)
F-measure	(2 * Recall * Precision) / (Recall + Precision)

4.3 Training; Testing; and Validation of Dataset

The following Machine Learning notations are linked to creating a supervised model that can generalize input samples that it has never seen before [10]. Furthermore, they help provide an interpretation regarding the sample test of the machine learning model regarding malicious URL detection. However, for a conducive use of the separate datasets in the Machine Learning model, requires an understanding of the process of distinguishing between the three datasets. The steps undertaken in differentiating between the datasets are as follows:

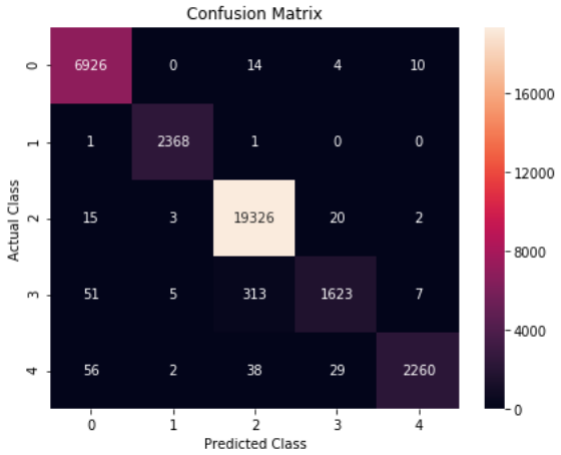
- Step 1: Make the model examine the dataset.
- Step 2: Make the model learn from its mistakes.
- Step 3: Conclude how well the model performs.

1. Training Dataset: a sample of the data utilized to fit the model. Additionally, the training set corresponds to step 1 as it includes a set of input examples that the model is expected to fit into - or trained on - by adjusted parameters [9].
2. Validation Dataset: a sample of the data utilized in providing an evaluation of a model fit on the training dataset while tuning hyperparameters of the model. This dataset additionally corresponds to step 2, where it is periodically evaluated. This is conducted to tune the parameters based on a frequent evaluation of results that are based on the validation set. 3. Testing Dataset: a sample of the data utilized in providing an unbiased evaluation of the final model fit on the training dataset. Lastly, this dataset corresponds with the final evaluation in step 3 of the model of the training phase which consists of using the training and validation sets is completed. The importance of the final step should however not be understated; as it is significant in testing how generalized the model is for us to obtain a working accuracy of the model[9].

Split ratio of the dataset The corresponding datasets in terms of their evaluation process and how they are split are dependent on two factors as researched by[11]. The first factor concerns the total number of samples in the data and the second factor concerns the actual model that is being trained upon. With certain models needing a substantial amount of data to be trained upon, the recommended route for such a model would be to optimize large training sets. For models with few hyperparameters will come to ease invalidating and tuning the dataset; as opposed to a scenario whereby the hyperparameters are many in terms of their nature which would call for a larger validation set.

5 Methodology and Detection

5.1 Data Mapping and Transformation



With a close reference to Appendix A, the labelled attack scores correctly predicted instances of: 1. Benign: 0 cases of spam URLs, 14 cases of defacement URLs, 4 cases of phishing URLs, and 10 cases of malware URLs 2. Spam: 1 case detected as benign and/or defacement URLs and, 0 cases of phishing and malware URLs 3. Defacement: 15 cases of benign URLs, 3 cases of spam URLs, 2O cases of phishing URLs, and 2 cases of malware URLs 4. Phishing: 51 cases of benign URLs, 4 cases 313 cases of defacement URLs and, 7 cases of malware URLs, and 56 cases 5. Malware: 56 cases of benign URLs, 2 cases of spam URLs, 38 cases of defacement URLs and, 29 cases of phishing URLs

5.2 Heuristic Classification

The approach is an improvement of the black-list method whereby the idea is to create a "blacklist of signatures" [20]. The essence of the signatures is to determine a match and test whether there is a correlation between a new URL with that of an existing malicious URL. Intrusion Detection Systems (IDS) scans sites for signatures and flags suspicious sites. To determine whether its a malicious site it is required to visit the site in an environment that is disposable such as a virtual machine. The main advantage of the IDS is that it is flexible and expandable. This is an improvement compared to Blacklisting as this classification detects threats from new URLs. Consequently, this approach is limited to common threats as it cannot cope with evolving attack techniques.

5.3 Tools Used

The main tool utilised in our analysis was Python, that aided the creation of a data vectorisation function that splits slashes in the studied URLs and removes

the .com component of the URL before tokenizing each URL a result of its constant occurrence. For this study, the extracted features of malicious URLs with Machine Learning algorithms are a list of five different URL attack types. After the URL list undergoes the vectorisation process, the extracted features concerning the test data utilised to go through a model-building procedure. The model building occurs in the form of the logistic regression model of predicting the above mentioned tokenized URL list and distinguishing the attack label each URL categorizes under.
Formal Concept Analysis Research Toolbox is a software application that was created for knowledge discovery through the FCA framework. There are many tools available for lattice creation and FCA which include: Conexp-clj, Tockit, Lattice Miner, Toscana, and ConExp.

6 Formal Concept Analysis

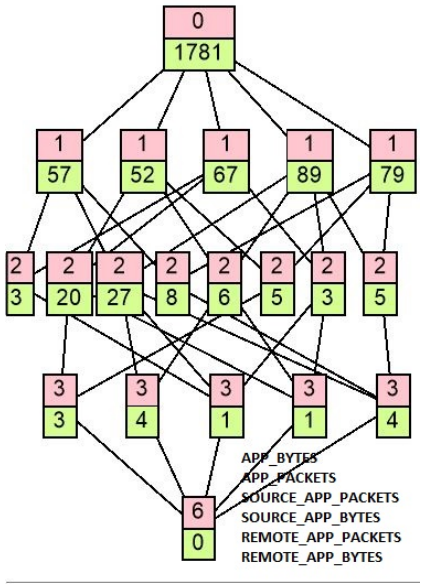
6.1 Association Rules in FCA

The analysis of concept lattices in the study of FCA on malicious and benign URLs requires an understanding of association rules. A concept lattice is comprised of many different associations that may be observed in the mentioned lattices that consist of implied associations. For every formal concept of (A, B) consists of a set of objects of A and a set of attributes of B. The relation that these formal concepts relate with one another is dependent on the different confidence levels and also, the different levels of support they incur. The confidence level is calculated by determining the probability of nodes with specific attributes to have other attributes as well. The support level is calculated by evaluating how much of a certain statement is backed up by its frequency of pattern in the lattice. In a scenario whereby one is given two sets of attributes Y and Z, the confidence of association rule $Y \rightarrow Z$ is defined as:

$$conf\,(Y \rightarrow Z) = \frac{|(Y \cup Z)'|}{|Y'|}$$

Since the FCA method is a culmination between lattices grouped in ordered sets this makes the visualization a proper method to link to the dataset the researhers used. This makes it conducive to perform analysis on lattice theory made possible by the understanding of object attribute matrices. With the case of our study with FCA, noteworthy associations being addressed below, the dataset used was split into two official concept lattices for a deeper understanding of relationships between the corresponding attributes and objects. The split of attributes and objects within the concept lattices have a direct relation to the variables that have been used in the Spearman correlation matrix analysis section which follows.

6.2 Noteworthy Associations



- Source application packets and application packets share a notable positive correlation with source application bytes.
- Source application packets and remote application bytes share a notable positive correlation source application bytes and remote application bytes.
- Remote application bytes and application packets share a notable positive correlation with remote application packets, remote application bytes and application packets.

6.3 Remarks regarding Noteworthy Associations

The proposed study of FCA on malicious and benign URLs by the researchers is still at a primary stage of analysis. However, what has been uncovered by the researchers with reference to the association rules and the relations between attributes and objects is the evenly spread constructs of the above-mentioned concept lattices. This potentially suggests there not being a substantial relationship between certain attributes as seen in the above figure.

6.4 Spearman Correlation Matrix

The Spearman Correlation Matrix (Appendix C) presents itself as a statistical coefficient measure of the strengths of relationships between paired data. With regards to the case of our second dataset consisting of malicious and benign websites, is denoted by rs, with the design constraints suggested to be:

$$-1 \leq r_s \leq 1$$

Further interpretation of the Spearman correlation is its similarity to that of the Pearson correlation in terms of how the closeness, the strong the relationship is between the matrix attributes. Due to correlation matrices presenting the effect sizes of similar attributes, the strength of the correlation in relation to the absolute values. Values between 0.00 - 0.19 are rated as "very weak"; Values between 0.20 - 0.39 are rated as "weak"; Values between 0.40 - 0.59 are rated as "moderate"; Values between 0.60 - 0.79 are rated as "strong"; and Values between 0.80 - 1.00 are rated as "very strong". Appendix D presents intercorrelations among all features of the second dataset. The matrix shows features of TCP_CONVERSATION_PACKAGE, SOURCE_APP_PACKETS, REMOTE_APP_PACKETS, SOURCE_APP_BYTES, REMOTE_APP_BYTES and APP_PACKETS, all have near perfect and monotonic relationships with each other. Research suggests that strong correlations between independent features is deemed undesirable. As a result of this, it is recommended to limit to only features that demonstrate high performance in the detection of malicious and benign URLs.

7 Findings and Discussion

7.1 Classification Report of Accuracy of the Model

	Recall	Precision	F1-score	Support
0 - Benign	0.98	1.00	0.99	6954
1 - Spam	1.00	1.00	1.00	2370
2 - Defacement	0.99	1.00	0.99	19366
3 - Phishing	0.97	0.81	0.88	1999
4 - Malware	0.99	0.95	0.97	2385
Accuracy			0.98	33074
Macro Average	0.98	0.95	0.97	33074
Weighted Average	0.98	0.98	0.98	33074

The sole purpose of the classification report is to represent the main classification metrics on a per-class basis. This aids the opportunity of gathering a deeper understanding of the behaviour of the classifier as stated by [17] and provides a global accuracy measure that masks any potential functional weaknesses of a class within the studied model. The metrics of the classification report are defined in terms of true and false positives, and true and false negatives.

7.2 Mitigation of access to malicious URL sites

Lexical-based features Features that are obtained from the URL string name. It is important to know that URLs are made up of two main components that

include the protocol identifier and the resource name which specifies the domain name where the link is located [14]. These two components are separated by a colon and two forward slashes. Figure: Depicts the breakdown of the URL Syntax:

It is evident that the malicious URL has a distinguishable pattern in the way the URL text is written, especially that of phishing attacks. The lexical features include the average domain token length and branch name whilst phishing URLs have a different pattern. An example of a phishing URL attack is targeted at well-known brand names, therefore the reseacrhers employ binary feature to examine whether the brand name is contained within URL tokens, not in its second-level domain (SDL).

Host-based features Features that require information that is requested from the Domain Name System (DNS) server. The feature describes the properties extracted from the site host as identified by the host portion of the URL[8][14]. Subsequently, this feature determines where the malicious sites are hosted, who owns the sites and how they are managed. The list below examines the properties to the host-based feature:

- WHOIS : The information retrieved includes details of the registration of domain names and the expiration date of the registrars. With this information available the classifier can determine whether domain name belongs to individuals listed as malicious URLs.
- IP address location: Each device that has access to the computer network is assigned an IP address in which one can derive the host's graphic location.
- Domain name: It refers to Internet Protocol (IP) resource, as mentioned above, allows the host computer to gain access to the Internet.

However, it should be mentioned that this feature has its drawbacks including the availability and connection speed. Whilst data is collected and prediction is ongoing in DNS the serves would not be available, consequently, the data would contain missing values that affect the quality of the classifier.

UBlock (adblocker) Top tier Internet companies have been collaborating to initiate, such as Google and Facebook along with start-ups to build a single platform for one motive to prevent users from accessing malicious URLs. Many web-based companies store large amounts of URLs on databases that are redefined regularly and that contains up-to-date URLs. This is not a feasible solution to the problem [3]. The need for human intervention to maintain URL is one of the limiting factors in this method albeit the platform performance with greater accuracy.

7.3 Limitations

1. A substantial volume of data: The dataset is significantly large as the study comprises of a combination of two URL datasets in which the researchers can establish a more accurate result. Subsequently, the datasets consisted of multi-class labels which included white spaces. The resulting implications of the above mentioned are time consumed for loading and processing of the data.
2. Redirected URL sites: Legitimate URL sites are susceptible to becoming compromised as an additional request by a user, unknowingly redirects them to a malicious evasion code. It is considered as a limitation in the sense that the URL is passed as a legitimate site.

8 Recommendation and Proposed Solution

In terms of the heuristic classification issues that arises, this method includes the time it takes to filter a new URL through the blacklists. Another issue that arises is having a machine capable of storing and processing large amounts of data as malicious URLs are stored within a list as a reference for any new URL sites. The problem arises whereby the machine is delayed users would have already become a victim. Adblock is a proposed solution to mitigate malicious URLs for major Internet companies like Facebook, Google and Twitter but the reason the researchers propose this within our revised model is that most organisations don't have the resources or income.

9 Conclusions and Future Work

This study incorporated both Machine Learning techniques in the form of Logistic Regression as well as Formal Concept Analysis. The combination of the two covered both statistical and Qualitative knowledge exploration discovery. The analysis using FCART on cybersecurity data affirmed that the FCA tool is able to obtain knowledge which was previously unknown ahead of the research. This study makes a contribution to malicious URL sites and cybersecurity data and proves that Formal Concept Analysis is a fairly effective method in understanding Malicious URL site data which opens up further research regarding cybersecurity research.

Machine Learning through the use of Logistic regression allowed for the categorization features, the detection of the types of Malicious URL attacks through Heuristic classification, the researchers propose a deeper investment into South African internet-based companies in order to mitigate and lessen the number of attacks due to higher levels of security. The research has allowed for the prospect for future work since there is a limited amount of research regarding FCA within Cyber Security there can be ways where similar research can be done and extended.

References

1. Al-Janabi, M., de Quincey, E., Andras, P.: Using supervised machine learning algorithms to detect suspicious URLs in online social networks. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1-8, (2017). https://dl.acm.org/citation.cfm?id=3116201
2. Aleroud, A., Zhou, L.: Phishing Environments, Techniques, and Countermeasures: A Survey, pp. 1 – 45. (2017)
3. Choi, H., Zhu, B B., Lee, H.: Detecting Malicious Web Links and Identifying Their Attack Types, pp. 1 – 12 (2011)
4. Confusion Matrix – Online Calculator,https://confusionmatrixonline.com/. Last accessed 11 Oct. 2019
5. Google Developers,https://developers.google.com/machine-learning/guides/text-classification/step-3label_vectorization. Last accessed 25 Sep 2019
6. Kuznetsov, S.: Machine Learning on the Basis of Formal Concept Analysis. In: Proceeding Automation and Remote Control 2001, vol. 62 (No 10), pp. 1543 – 1564. Russia (2001)
7. Ma, J., Saul, L K., Savage, S., Voelker G M.: Beyond blacklists: learning to detect malicious websites from suspicious URLs. pp. 1245–1254, New York, USA (2009)
8. Ma, J., Saul, L K., Savage, S., Voelker, G M.: Identifying Suspicious URLs: An Application of Large-Scale Online Learning. In: Proceedings of the 26th International Conference on Machine Learning, pp. 1 – 8, USA (2009)
9. Machine Learning Mastery, https://machinelearningmastery.com/difference-test-validation-datasets/. Last accessed 25 Sep 2019
10. Medium Towards Data Science, https://medium.com/datadriveninvestor/what-are-training-validation-and-test-data-sets-in-machine-learning-d1dd1ab09bae. Last accessed 25 Sep 2019
11. Medium Towards Data Science, https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7. Last accessed 25 Sep 2019
12. Neznanov, A A., Kuznetvov, S O.: Information Retrieval and Knowledge Discovery with FCART. pp. 74 - 82 (2013)
13. Patil, D R., Patil, J B.: Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification. In: The ISC Int'l Journal of Information Security, vol. 10(2), pp. 141 – 162 (2018)
14. Sahoo, D., Liu, C., Hoi,.S: Malicious URL Detection using Machine Learning: A Survey, vol.1 (No 1), pp. 1 -37 (2019)
15. Sharma, S., Parveen, S.: Efficient Malicious URL based on Feature Classification. In: International Journal of Engineering Research and General Science, vol. 3(3), pp. 113 – 120 (2015)
16. University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html. Last access 4 July 2019
17. Verma, R., Das, A.: What's in a URL: Fast Feature Extraction and Malicious URL Detection. In: IWSPA '17 Proceedings, pp. 55-63, USA (2017)
18. Yellowbrick, https://www.scikit-yb.org/en/latest/api/classifier/classification-report.html. Last accessed 12 Oct 2019
19. DropBox, https://www.dropbox.com/s/f7c8ijhhp4joaig/dataset.zip?dl=0. Last accessed 12 Oct 2019

A Appendix

A.1 Appendix A

Attack Label	Comment
0 - Benign	0 cases of Benign URLs being taken as Spam URLs 14 cases where Benign URLs were taken to be Defacement URLs 4 cases where Benign URLs were taken to be Phishing URLs 10 cases where Benign URLs were taken to be Malware URL
1 - Spam	1 case of Spam URLs being taken as Benign URLs 1 case of Spam URLs being taken as Defacement URLs 0 cases of Spam URLs being taken as Phishing URLs 0 cases of Spam URLs being taken as Malware URLs
2 - Defacement	15 cases where Defacement URLs were taken to be Benign URLs 3 cases where Defacement URLs were taken to be Spam URLs 20 cases where Defacement URLs were taken to be Phishing URLs 2 cases where Defacement URLs were taken to be Malware URLs
3 - Phishing	51 cases where Phishing URLs were taken to be Benign URLs 5 cases where Phishing URLs were taken to be Spam URLs 313 cases where Phishing URLs were taken to be Defacement URLs 7 cases where Phishing URLs were taken to be Malware URLs
4 - Malware	56 cases where Malware URLs were taken to be Benign URLs 2 cases where Malware URLs were taken to be Spam URLs 38 cases where Malware URLs were taken to be Defacement URLs 29 cases where Malware URLs were taken to be Phishing URLs

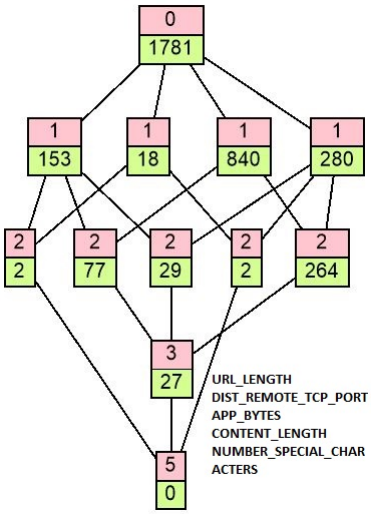
A.2 Appendix B

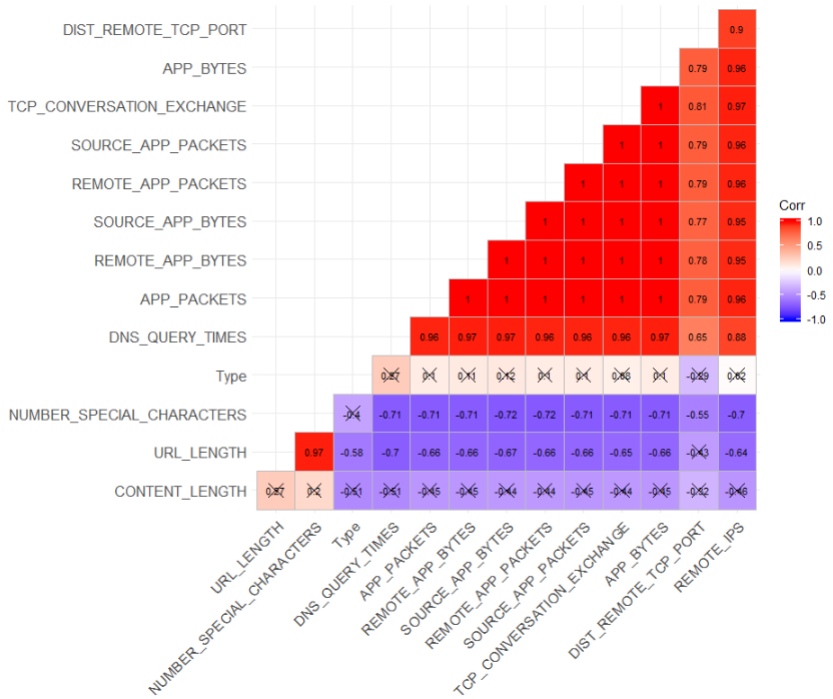
Relationship Definitions in FCA

Relationship Types	**Expression**
Strong Positive Association	$conf(Y \rightarrow Z) \geq 0.85$
Notable Positive Association	$conf(Y \rightarrow Z) \geq 0.75$
Notable Negative Association	$conf(Y \rightarrow Z) \leq 0.25$
Strong Positive Correlation	$conf(Y \rightarrow Z) \geq 0.85$ *and* $conf(Z \rightarrow Z) \geq 0.85$
Notable Positive Correlation	$conf(Y \rightarrow Z) \geq 0.75$ *and* $conf(Z \rightarrow Y) \geq 0.75$
Notable Negative Correlation	$conf(Y \rightarrow Z) \leq 0.25$ *and* $conf(Y \rightarrow Z) \leq 0.25$

A.3 Appendix C

FCART Concept Lattice model



A.4 Appendix D

Spearman Correlation Matrix