Correlation and Dependence Analysis on Cyberthreat Alerts

John M.A. Bothos, Konstantinos-Georgios Thanos, Dimitris M. Kyriazanos, George Vardoulias, Andreas Zalonis, Eirini Papadopoulou, Yannis Corovesis, Stelios C.A. Thomopoulos

National Centre for Scientific Research "Demokritos" (NCSRD), Greece

ABSTRACT

In this paper a methodology for the enhancement of computer networks' cyber-defense is presented. Using a time-series dataset, drawn for a 60-day period and for 12 hours per day and depicting the occurrences of cyberthreat alerts at hourly intervals, the correlation and dependency coefficients that occur in an organization's network between different types of cyberthreat alerts are determined. Certain mathematical methods like the Spearman correlation coefficient and the Poisson regression stochastic model are used. For certain types of cyberthreat alerts, results show a significant positive correlation and dependence between them. The analysis methodology presented could help the administrative and IT managers of an organization to implement organizational policies for cybersecurity.

Keywords – Correlations, cyberattacks, dependencies, network, time series

1. INTRODUCTION

Today most organizations in the world heavily depend on IT infrastructure such as computer networks, servers, databases and information systems, to carry out their daily activities. This infrastructure has been the target of cyberattacks which aim to disrupt the ability of an organization to perform its activities, steal data or even put it out of business. According to [1] and [2], after a security breach, organizations are affected in fields such as, operations, finance systems, brand reputation and customer retention. Cyberattacks cause various direct or hidden costs to an organization's tangible and intangible assets, jeopardizing even its sustainability in some cases. This emphasizes the need for organizations to prioritize cybersecurity so as to minimize the risk of a cyberattack being successful. An organization that can reduce its administrative costs by optimizing its cybersecurity defense mechanisms, can divert more monetary resources to other investments for business growth. Due to the high costs involved in adopting and implementing a proactive cybersecurity policy, organizations usually develop ineffective cybersecurity solutions as reactions to cyberattack incidents [3]. Implementing an effectively proactive information security policy makes the IT infrastructure more productive, increases its availability and guarantees an organization's activities to continue uninterrupted. To implement such an effective line of cyber-defense, an organization not only has to determine the value of its assets, but also the cyberthreat environment, by determining correlations and dependencies between various types of cyberattacks and malware.

Research on cyberattack pattern recognition in network traffic has been going on for quite some time. Relevant research approaches have been made in the scope of finding satisfactory predicting mathematical models for such incidents. Empirical modelling of cyber-alerts relates mainly to the study of time-series models for efficient forecasting of cyberattacks. In [4], Markov models on time-series data of communications were used to highlight the importance of detecting types of anomalies in a computer network traffic flow in identifying types of intrusions, in the network. In [5] and [6], ARFIMA and FIGARCH models were used on time-series data of network traffic, to predict whether detected anomalies are indications of real cyberattacks or just false alarms and to detect cyberattacks on a DDoS network. In [7] predictive time-series models were used to forecast vulnerabilities of web browsers, while in [8], a dynamic risk assessment stochastic model is used to identify inventory-enhancement opportunities for critically disrupted systems.

Our study contributes to the relevant research by applying mathematical methods for the detection of significant correlation and dependence between different types of cyberthreat alerts. In order to determine the degrees of these correlations and

dependencies, Spearman's correlation coefficient and Poisson regression stochastic modelling are used. Significant correlations and dependencies among certain types of cyberthreat alerts are distinguished that can be used for event count predictions of such incidents.
Through our contribution we aspire to enrich the variety of scientific methods that have been employed so far for the analysis of cyberattack pattern recognition in computer networks' traffic. As well as the use of *Markov,ARFIMA* and *FIGARCH* time-series models, we propose the use of probabilistic relevant time-series models, like the *Poisson* stochastic model, in order to determine valid dependence relations between different types of cyber-alerts.

2. METHODOLOGY

2.1 Experiment set up: network description and data mining approach

The alert logs of an intrusion detection system (IDS), already deployed in the network of an organization, were used for the formation of the analysis dataset. This IDS is placed in the entry point next to the border router of the network of the organization. The network serves about 1000 users categorized in certain organizational units. Each unit serves a different scientific discipline or support division. Each such unit is protected by a dedicated firewall and VLAN segmentations, operates the local infrastructure under its own management and has absolute control of the firewall and UTP cabling.
The IDS operation is at the front of all individual firewalls and the only communicated addresses concerned the external IP of each firewall. All user traffic was NATed on incoming or outgoing directions. The collaborating network operation center personnel have the absolute clearance to manage and operate the border routers, the centrally enforced access lists and the central IDS. The IDS alert data were collected in a database of events and that data was communicated to the research analyst team following a privacy impact assessment and anonymization procedure from the authorized network administration personnel. As a result, the database of events and all research data processed in the context of this paper, contain no information about individual IP addresses that correspond to user workstations or any other content that could be used to directly or indirectly identify a network user, i.e. by exclusion, narrowing down to a very small number of possible subjects or correlation and cross-matching with other public information. A thorough analysis on the privacy and legal challenges of network research can be found in [9].
The IDS system used is the *Suricata* intruder detection system [13] based on the *Oinkmaster* ruleset [14]. The log file records of cyberthreat alerts were related to the most frequent and costliest types of cyberattacks. Our dataset consisted of the following selection of cyberthreat alert protocols, detected by the IDS.

Types of cyberthreat alerts

Type of alert	Description of cyberthreat
WORM/TROJAN ()	This alert is emerged from traffic that is related to the propagation of viruses and worms and contamination of systems
TOR	Alert that indicates that illegal communications are taking place, such as espionage, criminal communications, illegal financial transactions, etc.
GPLSNMP	Alert emerged from traffic related to leakage information, mainly exploiting vulnerabilities of protocols, such as SNMP.
VOIP	This alert is related to attempts to exploit vulnerabilities for illegal usage of Internet telephony servers.
SQL	Multiple attacks related to the vulnerabilities of databases.
GPLRPC	Alerts related to the exploitation of the RPC PROTOCOL, e.g. malicious software injection.
IPMI	Alert that signals possible exploitation of system consoles at a very low level.
MOBILE	Mobile malware alert about installation of software to mobile search that exploit end-user devices, i.e. phishing, etc.
CNC	Command and control which indicates traffic related to the management of botnets that carry out cyberattacks such as DDOS.
DNS	Alert about suspicious queries to the domain servers related to information leakage.
SPAMHAUS	Unsolicited email traffic.
SCAN	Network activity related to external attempts to reconnaissance topologies,

	network services, operating systems, in order to exploit vulnerabilities.
MALWARE	Alert about transformed normal software with malicious parts in order to exploit unaware users.
DDOS	Cyberattacks related to denial of services, unable to operate.
COMPROMISED	Alert about systems that have been penetrated.

Records included every cyberthreat alert in the network traffic flow, during working days and hours, from Monday to Friday and from 8 a.m. until 8 p.m., for a 60-day period. At the end of the test period, this volume of log files was processed with data mining methods in order to filter the incidents by type of cyberthreat alert. For each cyberthreat alert category, we aggregated the number of relevant incidents by a time step of an hour, in order to form an adequately large sample size, so that valid and unbiased statistical results be produced. This resulted in fifteen (15) time series of the number of incidents by type, each one corresponding to one of the above-mentioned cyberthreat alerts, respectively. Consequently, each time series consists of an hourly number of incidents recorded for each specific cyberthreat alert.

2.2 Correlation analysis

Usually IDS outputs of detected cyberthreat alerts about suspicious signatures are huge, requiring an awful lot of network operator's attention and systematic analysis as part of cybersecurity actions. Correlation analysis offers an optimizing solution to the problem of limited machine computation power, reducing the quantity of data that needs to be processed, in order to extract useful information, without losing the overall situational picture. Detecting potential correlations among the various kinds of cyberthreat alerts is a prerequisite in order to proceed to modelling the dependencies between them. Due to the nature of our dataset, composed by discrete variables consisted of count data and not continuous random variables we could not resort to the Pearson correlation coefficient, so we preferred instead to use the Spearman's correlation coefficient:

$r_s = 1 - \{(6 * \Sigma d_i^2) / [n * (n^2 - 1)]\}$, where

Σ : Sum

d_i : differences between the ranks of pairwise cyberthreat alerts

n : number of cyberthreat alerts(sample size)

for the calculation of the correlation coefficients between these fifteen (15) different types of cyberthreat alerts, each by every other. As a result, a 15x15 correlation matrix emerged, in which each cell contained the Spearman's correlation coefficient value, for the respective cyberthreat alerts. Setting a threshold of significant positive correlation at 40% and over, results yielded that with a 95% probability or $\alpha = 0.05$ level of statistical significance

SQL and *VOIP* cyberthreat alerts were correlated by 61.9%, *SQL* and *SCAN* cyberthreat alerts were correlated by 47.1%, COMPROMISED and SQL cyberthreat alerts were correlated by 51.7% and *COMPROMISED* and *SCAN* cyberthreat alerts were correlated by 92%.

2.3 Dependence analysis

Exploring further the significantly correlated cyberthreat alerts, we made use of regression stochastic modelling to estimate the potential dependence between them.

We formed the following functions about the dependence between the significantly correlated cyberthreat alerts, with over 40% positive correlation between them.

SQL=f(VOIP), VOIP=f(SQL),
SCAN=f(SQL), SQL=f(SCAN),
COMPROMISED=f(SQL),
SQL=f(COMPROMISED),
COMPROMISED=f(SCAN),
SCAN=f(COMPROMISED).

So the respective regressions that had to be run were the following:

SQL cyberthreat alert on *VOIP* cyberthreat alert.
VOIP cyberthreat alert on *SQL* cyberthreat alert.
SCAN cyberthreat alert on *SQL* cyberthreat alert.
SQL cyberthreat alert on *SCAN* cyberthreat alert.
COMPROMISED cyberthreat alert on *SQL* cyberthreat alert.
SQL cyberthreat alert on *COMPROMISED* cyberthreat alert.
COMPROMISED cyberthreat alert on *SCAN* cyberthreat alert.
SCAN cyberthreat alert on *COMPROMISED* cyberthreat alert.

For each type of cyberthreat alert, we estimated the dependence of the mean number of its emergence in the network at an hour *t*, on the number of incidents of other cyberthreat alerts that are over 40% ($r_s \geq +0.5$) positively correlated, at the same hour *t* and the previous 2 hours *t-1*, *t-2*. We selected this time lapse window, based on the intuition of imminent action against a cyberthreat alert from the network's administrator in order to prevent and/or mitigate as much as possible potential damages and losses from a cyberattack. A maximum 2-hour interval provides an adequate time window of action by the network

operators to apply response counter measures. Smaller windows include much more noise and bigger windows would probably miss important events and offer less opportunity to respond.

For this dependence analysis, we made use of a Poisson stochastic model. Taking the Poisson Probability Distribution Function

$Prob(Y=y) = (e^{-\lambda}*\lambda^{y})/y!\ ,\ y=0,1,2,....$ [10], where

λ : distribution parameter of *Poisson Probability Distribution Function* concerning the emergence of cyberthreat alerts in the network traffic flow and

$y!$: observed counts of emergence of each cyberthreat alert in the network traffic flow.

Considering the mean number of cyberthreat alerts of type *i* happening at hour *t*, depending linearly on the number of cyberthreat alerts of another type *j* at the same hour t_0 and/or at the previous 2 hours *t-1*, *t-2* as $E(y_{it}|\boldsymbol{x}_{jt})=\lambda_t=e^{x'_{jt}*\boldsymbol{bj}}$ [11],

we finally take our model concerning the dependence of the probability the mean number of cyberthreat alert type *i* emerging at hour *t*, on the number of cyberthreat alert type *j* at the same hour *t* and/or the previous two hours *t-1*, *t-2* in the form of $Prob(y_{it}=\lambda_{it})=\{[(e)^{-\,e\,(x'_{jt}*\boldsymbol{bj})}*(e^{x'_{jt}*\boldsymbol{bj}})^{yit}]/y_{it}!\}\ +\ u_{it}$ [12].

According to [11], 'such models are estimated with maximum likelihood methodology, with the log-likelihood function being

$ln(y,b)=\Sigma_{t=1}^{n}(-e^{x'_{jt}*\boldsymbol{bj}}+y_{it}*\boldsymbol{x'_{jt}}*\boldsymbol{b_j}-lny_{it}!)$ [11]

and the relevant likelihood equations being

$\theta ln(y,b)/\theta\boldsymbol{b_j}=\Sigma_{t=1}^{n}(y_{it}-e^{x'_{jt}\boldsymbol{bj}})*\boldsymbol{x}_{jt}=\boldsymbol{0}$ [11]

formatting the following Hessian matrix

$\theta^2 ln(y,b)/\theta\boldsymbol{b_j}*\theta\boldsymbol{b_j}'=-\Sigma_{t=1}^{n}(e^{x'_{jt}*\boldsymbol{bj}}*\boldsymbol{x_{jt}}*\boldsymbol{x_{jt}}')$ [11]

The asymptotic estimator covariance matrix is in the form of

$[\Sigma_{t=1}^{n}(e^{x'_{jt}*\boldsymbol{bj}})_{est}*\boldsymbol{x_{jt}}*\boldsymbol{x_{jt}}')^{-1}$. [11]

Testing for the statistical significance of the model's hypotheses about the estimator, involves use of the LR statistic in the form of

$LR=2*\Sigma_{t=1}^{n}[ln(P_{iest}/P_{iestrestricted})]$ [11].

2.4 Results

The results of the regressions are shown in the following tables.

Table 1. Dependence of *SQL* cyberthreat alert on *VOIP* cyberthreat alert $SQL = f(VOIP)$

Dependent variable: SQL Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	12.22246	0.000136	89964.57	0.0000
VOIP	0.000374	2.17E-06	172.2328	0.0000
VOIP(-1)	-8.15E-05	2.93E-06	-27.83379	0.0000
VOIP(-2)	0.000352	2.13E-06	164.9999	0.0000
LR statistic 453225.2 Prob(LR statistic) 0.000000				

Table 2. Dependence of *VOIP* cyberthreat alert on *SQL* cyberthreat alert $VOIP = f(SQL)$

Dependent variable: VOIP Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.755649	0.050791	-14.87762	0.0000
SQL	1.13E-05	5.64E-07	20.02671	0.0000
SQL(-1)	9.10E-06	6.33E-07	14.36263	0.0000
SQL(-2)	4.05E-06	3.77E-07	10.73018	0.0000
LR statistic 14786.20 Prob(LR statistic) 0.000000				

Table 3. Dependence of *SCAN* cyberthreat alert on *SQL* cyberthreat alert $SCAN = f(SQL)$

Dependent variable: SCAN Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	7.550100	0.003718	2030.727	0.0000
SQL	4.26E-07	3.37E-08	12.64695	0.0000
SQL(-1)	5.63E-06	4.35E-08	129.4269	0.0000
SQL(-2)	2.84E-06	2.99E-08	94.77684	0.0000
LR statistic 320368.3 Prob(LR statistic) 0.000000				

Table 4. Dependence of *SQL* cyberthreat alert on *SCAN* cyberthreat alert $SQL = f(SCAN)$

Dependent variable: SQL Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	12.29104	9.05E-05	135792.3	0.0000
SCAN	1.88E-07	2.45E-09	76.95380	0.0000
SCAN(-1)	-1.02E-08	2.74E-09	-3.715406	0.0002
SCAN(-2)	1.38E-07	2.45E-09	56.15487	0.0000

LR statistic 10436.97				
Prob(LR statistic) 0.000000				

Table 5. Dependence of *COMPROMISED* cyberthreat alert on *SQL* cyberthreat alert *COMPROMISED = f (SQL)*

Dependent variable: COMPROMISED				
Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	6.509554	0.008114	802.2647	0.0000
SQL	4.79E-06	7.47E-08	64.05926	0.0000
SQL(-1)	1.12E-06	8.13E-08	13.75862	0.0000
SQL(-2)	1.99E-07	5.79E-08	3.438938	0.0006
LR statistic 32068.19				
Prob(LR statistic) 0.000000				

Table 6. Dependence of *SQL* cyberthreat alert on *COMPROMISED* cyberthreat alert *SQL = f (COMPROMISED)*

Dependent variable: SQL				
Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	12.13649	0.000292	41520.08	0.0000
COMPROMISED	0.000322	7.60E-07	424.0071	0.0000
COMPROMISED(-1)	-0.000179	8.30E-07	-216.0809	0.0000
COMPROMISED(-2)	-8.15E-05	5.04E-07	-161.6978	0.0000
LR statistic 492892.9				
Prob(LR statistic) 0.000000				

Table 7. Dependence of *COMPROMISED* cyberthreat alert on *SCAN* cyberthreat alert *COMPROMISED = f (SCAN)*

Dependent variable: COMPROMISED				
Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	7.847009	0.000829	9465.598	0.0000
SCAN	2.71E-07	2.18E-08	12.42193	0.0000
SCAN(-1)	2.47E-08	2.44E-08	1.013099	0.3110
SCAN(-2)	2.40E-07	2.17E-08	11.02043	0.0000
LR statistic 352.6056				
Prob(LR statistic) 0.000000				

Table 8. Dependence of *SCAN* cyberthreat alert on *COMPROMISED* cyberthreat alert *SCAN= f (COMPROMISED)*

Dependent variable: SCAN				
Method: ML/QML - Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	9.165064	0.001187	7718.030	0.0000
COMPROMISED	-0.000475	3.22E-06	-147.6575	0.0000
COMPROMISED(-1)	0.000737	3.76E-06	196.1082	0.0000
COMPROMISED(-2)	-0.000129	1.95E-06	-66.38216	0.0000
LR statistic 126924.1				
Prob(LR statistic) 0.000000				

The estimated coefficients by the Poisson regressions $Prob(y_{it}=\lambda_{it})=\{[(e)^{-e^{(x'_{jt}*b_j)}}*(e^{x'_{jt}*b_j})^{y_{it}}]/y_{it}!\} + u_{it}$ [12] in the above tables give the exact form of the dependence of the emergence of the mean number of one type of cyberthreat alert *i* at an hour *t*, on the emergence of another highly correlated (*>40%*) cyberthreat alert *j* at the same hour t and the previous 2 hours *t-1, t-2*. They also provide for the calculation of the relevant probabilities, through $E(y_{it}|\boldsymbol{x}_{jt})=\lambda_t=e^{x'_{jt}*b_j}$ [11].

The LR statistic of the relevant error probabilities $LR=2*\Sigma_{t=}{}^{n}[ln(P_{iest}/P_{iestrestricted})]$[11] denotes the statistical significance of the model's hypotheses about the estimators.

3. CONCLUSION

In this paper we applied mathematical methods to highlight possible relationships between different types of cyberthreat alerts in a network system. Our goal was to contribute to the enhancement of network cyber-defense policies by improving the effectiveness of IT systems' intelligence.

With the use of Spearman correlation analysis and Poisson regression stochastic modelling we tried to distinguish significant correlations and dependencies among certain types of cyberthreat alerts that can be used for forecasting such incidents.

Correlation analysis denoted a significant positive correlation of over 40% for the following pairs of cyberthreat alerts:
SQL and *VOIP* (61.9%), *SQL* and *SCAN* (47.1%), COMPROMISED and SQL (51.7%) and *COMPROMISED* and *SCAN* (92%).
The above significant correlations implied a strong degree of similar pattern of emergence and possible existence of a significant relationship between them. Based on these results, we proceeded further with the estimation of potential dependence between these cyberthreat alerts, regarding the dependence of the mean number of each type's emergence in the network at an hour *t*, on the number of the emergence of its over 40% positively correlated other type of cyberthreat alerts, at the same hour *t* and the previous 2 hours *t-1*, *t-2*.
Results of the dependence analysis denoted that *SQL* cyberthreat alert emergence significantly depended on *VOIP* cyberthreat alert emergence at the same, one and two hours before,
VOIP cyberthreat alert emergence significantly depended on *SQL* cyberthreat alert emergence at the same, one and two hours before,
SCAN cyberthreat alert emergence significantly depended on *SQL* cyberthreat alert emergence at the same, one and two hours before,
COMPROMISED cyberthreat alert emergence significantly depended on *SQL* cyberthreat alert emergence at the same, one and two hours before,
SQL cyberthreat alert emergence significantly depended on *COMPROMISED* cyberthreat alert at the same, one and two hours before.
COMPROMISED cyberthreat alert emergence significantly depended on *SCAN* cyberthreat alert at the same and two hours before, but not at one hour before
SCAN cyberthreat alert emergence significantly depended on *COMPROMISED* cyberthreat alert at the same, one hour and two hours before.
For all our dependence models the respective LR statistics denoted statistical significance of the models' hypotheses about the estimators.
The results of our research can be used as suggestions to IT managers, in order to apply and implement more efficiently cybersecurity strategies and cyber-defense tactics, without the need of monitoring all incidents emerging and so reduce data storing consumption and network capacity overload.

ACKNOWLEDGEMENT
The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 740829 - SAINT project.

REFERENCES

[1] Cisco, *"Annual Cybersecurity Report"* (2017).

[2] Deloitte, *"Beneath the Surface of a Cyberattack: A Deeper Look at the Business Impacts"*.

[3] Shim, *"Agency Problems in Information Security: Theory and Application to Korean Business"*, The Journal of Internet Electronic Commerce Research, Vol. 15 (2015).

[4] Neil, Hash, Brugh, Fisk, Storlie *"Scan Statistics for the Online Detection of Locally Anomalous Subgraphs"*, Technometrics, (2013).

[5] Andrysiak, Saganowski, Choras, Kozik *"Network Traffic Prediction and Anomaly Detection Based on ARFIMA Model"*, International Joint Conference SOCO'14-CISIS'14-ICEUTE'14.

[6] Andrysiak, Saganowski, Maszewski, Grad "*Long-memory dependence statistical models for DDoS attacks detection"*, Image Processing & Communications, vol. 20, (2015).

[7] Roumani, Nwankpa and Roumani *"Time series modelling of vulnerabilities"*, Computers and Security 51, (2015).

[8] Resurreccion and Santos *"Uncertainty modelling of hurricane-based disruptions to interdependent economic and infrastructure systems"*, Nat Hazards 69 (2013).

[9] Ohm, Sicker, Grunwald, *"Legal Issues Surrounding Monitoring During Network Research"*, IMC '07 (2007).

[10] Davidson, McKinnon, *"Econometric Theory and methods"*, (1999).

[11] Green, *"Econometric Analysis"*, (2002).

[12] Gujarati, Porter *"Basic Econometrics"*, (2008).

[13] Suricata Open Source IDS
https://suricata-ids.org/

[14] Oinkmster
http://oinkmaster.sourceforge.net/about.shtml