PROGETTAZIONE E SVILUPPO DI UN SISTEMA DI RILEVAMENTO MALWARE NEI DISPOSITIVI MOBILE TRAMITE ANALISI IBRIDA

Tesi di Laurea Magistrale in Ingegneria Informatica

G. De Cristofaro

Relatore: Prof. Giuseppe Lo Re

Correlatore: Prof. Alessandra De Paola

PROGETTAZIONE E SVILUPPO DI UN SISTEMA DI RILEVAMENTO MALWARE NEI DISPOSITIVI MOBILE TRAMITE ANALISI IBRIDA

Tesi di Laurea di

Giuseppe De Cristofaro

Relatore:

Prof. Giuseppe Lo Re

Correlatore:

Prof.ssa Alessandra De Paola
Ing. Antonio Bordonaro

Sommario

I dispositivi mobile, quali tablet e smartphone, fanno ormai parte della quotidianità di tutti gli utenti che li posseggono. Questi dispositivi forniscono strumenti utili agli utenti tramite dei software, chiamati comunemente applicazioni. L'estrema semplicità con cui gli possono installare nuove app sui propri dispositivi riduce l'attenzione gli utenti stessi pongono in questa azione, e ciò consente facilmente ad un malintenzionato di far eseguire un software potenzialmente malevolo ad un utente ignaro.

Questo progetto di tesi rivolge il suo sguardo in particolar modo alle applicazioni eseguibili nel sistema operativo Android. L'obiettivo di questo lavoro è lo sviluppo e la progettazione di un sistema che possa rilevare i malware in questi dispositivi tramite un sistema di analisi ibrida, che combina tecniche di analisi statica e analisi dinamica. La classificazione è eseguita addestrando algoritmi di machine learning sull'insieme delle caratteristiche che ha ottenuto i risultati migliori tra quelli presi in considerazione. L'analisi sperimentale descritta in questo lavoro di tesi ha consentito di individuare il miglior approccio per la costruzione del sistema di analisi ibrida, combinando in maniera più efficace possibile l'analisi statica e quella dinamica.

Sommario

1. Introduzione

Al giorno d'oggi, i dispositivi *mobile*, quali smartphone e tablet, sono largamente utilizzati dalle persone nella loro vita quotidiana: infatti, il numero di utenti che, nel 2021, sfrutta le potenzialità di questi dispositivi è raddoppiato nell'arco di 5 anni, passando da circa tre miliardi nel 2016 a poco più di sei miliardi, ipotizzando un ulteriore incremento fino a sette miliardi e mezzo nel 2026 [1], come mostrato nella Figura 1.

Le ragioni che hanno portato all' uso intensivo di dispositivi mobile risiedono proprio nella loro capacità di poter fornire dei servizi in maniera facile, veloce ed efficiente, generando soprattutto un netto risparmio di tempo e di risorse per gli utenti che usufruiscono di tali servizi.

Tramite i dispositivi mobile è infatti possibile ottenere qualsiasi servizio e in qualsiasi luogo e momento.

Con riferimento alla prima caratteristica, bisogna considerare che dispositivi quali smartphone e tablet permettono di installare e di utilizzare migliaia di software di ogni tipo. È possibile ottenere sia software che racchiudono servizi più importanti e che manipolano dati sensibili, quali informazioni bancarie, dati biometrici e servizi di geolocalizzazione, sia software utilizzati come svago o per operazioni più semplici, come ad esempio giochi o applicazioni di intrattenimento.

La seconda caratteristica è dovuta alla natura intrinseca di questi dispositivi, che sono portatili e consentono quindi di rimuovere qualsiasi limitazione che potrebbe imporre l'utente di dover richiedere un certo servizio in un certo luogo. Si pensi ad esempio alla possibilità di dover effettuare un bonifico ad una persona, o di dover mandare una email di lavoro mentre si è fuori casa: i dispositivi mobile permettono, in entrambi i casi, di rimuovere questa limitazione.

Questo enorme utilizzo porta, di conseguenza, ad una gargantuesca produzione di informazioni proveniente dai singoli dispositivi e, in particolar modo, dalla possibilità di poter condividere molto facilmente i software che possono essere scaricati e utilizzati dagli utenti.

Questi software, che comunemente prendono il nome di *applicazioni* o *apps*, vengono condivisi, gratuitamente o a pagamento, in dei negozi virtuali, chiamati *stores*, i quali non sempre possono fornire certificazioni sulla sicurezza delle applicazioni che si trovano al loro interno.

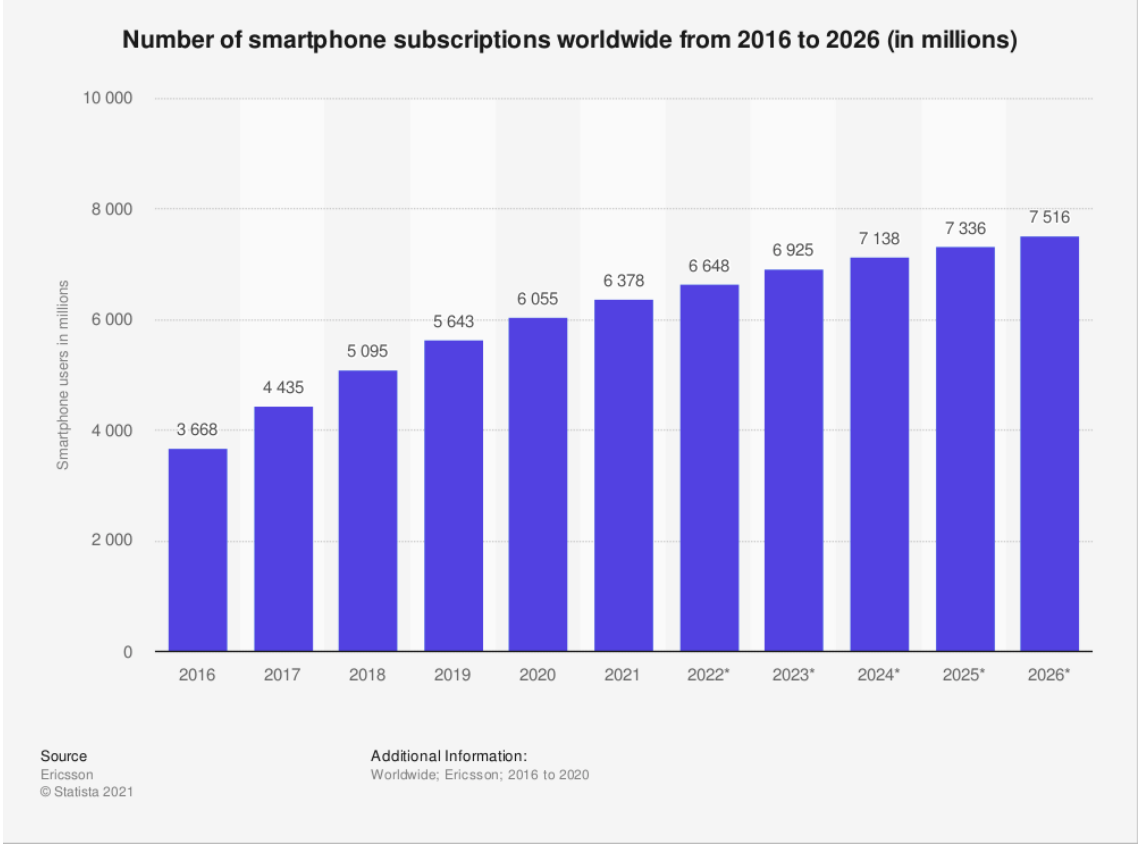

Figura 1. Evoluzione del numero degli utenti smartphone in [1]

Proprio a causa della mole di informazioni e di applicazioni, e dalla facilità con cui queste possono essere condivise, non risulta difficile mettere in circolazione delle applicazioni che potrebbero risultare dannose per l'utente che decide di installarle sul proprio dispositivo.

L'utente potrebbe infatti scaricare un'applicazione malevola inconsapevolmente, fidandosi dello *store* in cui essa si trova, oppure potrebbe voler cercare delle "varianti" di un'applicazione, le quali gli permetterebbero di ottenere dei benefici che non potrebbe ottenere da quella originale. Un esempio di quest'ultimo caso sono varianti di app che, all'insaputa dell'utente, mandano degli sms a numeri ad alto costo tariffario, portando così al continuo esaurimento del credito del malcapitato [2].

I motivi che portano gli utenti malintenzionati a mettere in circolazione questi software malevoli possono essere diversi e unificabili alle stesse motivazioni degli *hacker* che hanno come obiettivo i dispositivi fissi, quali pc, server, workstation, etc. Tuttavia, rispetto ai dispositivi fissi, uno smartphone può non solo risultare più debole nella compromissione a causa delle sue risorse limitate, ma può contenere dati di carattere

personale, quali dati biometrici, coordinate bancarie, dati di geolocalizzazione, nomi utenti, password e tante altre tipologie. L'unione di queste due caratteristiche fa sì che i dispositivi mobili risultino delle prede appetibili ai malintenzionati che vogliono rubare queste informazioni personali.
Le tipologie di applicazioni malevole possono essere suddivise in categorie sulla base del comportamento malevolo che attuano nei dispositivi.
Le categorie, individuate da *Google Play Protect*, sono:

- **Backdoor:** porzioni di codice che permettono l'esecuzione di operazioni non volute e potenzialmente dannose che vengono eseguite da remoto su un dispositivo.
- **Billing fraud:** codice che addebita l'utente in modo ingannevole. Questa categoria si può suddividere a sua volta in *SMS Fraud*, che invia SMS ad un costo elevato o che sottoscrive abbonamenti tramite SMS senza che l'utente se ne accorga, *Call Fraud*, che effettua chiamate a numeri a pagamento senza che l'utente se ne accorga, e *Toll Fraud*, che racchiude tutti i tipi di abbonamenti effettuati tramite diversi canali (esclusi SMS e chiamate) come la *WAP Fraud*, la più diffusa, che può far cliccare all'utente un bottone non visibile in una WebView.
- **Stalkerware o Commercial Spyware:** è un codice che trasmette informazioni personali senza un adeguato consenso e senza notificare che questo trasferimento sta avvenendo. Ad esempio, tramite dei form di uso legittimo, è possibile per i genitori tenere sotto controllo i propri figli. Tuttavia, può essere usato in modo improprio per seguire una persona senza il suo consenso.
- **Denial of Service (DoS):** codice che esegue un attacco di *denial of service* o che prende parte ad un attacco di *distributed denial of service* contro un sistema, mandando ad esempio un'elevata quantità di richieste HTTP per saturare la memoria dei server.
- **Hostile downloaders:** codice che di per sé non è ostile, ma che può scaricare applicazioni potenzialmente dannose. Un'applicazione può essere considerata una *hostile downloader* qualora il 5% delle applicazioni scaricate, con un minimo di 500, siano potenzialmente malevole.

- **Non-Android Threat:** in applicazioni di questo tipo, il codice contenuto non ha minacce per i dispositivi Android, ma può essere potenzialmente malevolo per altri dispositivi.
- **Phishing:** codice che finge di provenire da fonti sicure, richiedendo credenziali dell'utente o dati bancari per mandarli invece a terze parti. Il *phishing* è anche l'azione di intercettare la trasmissione delle credenziali degli utenti.
- **Elevated privilege abuse:** questo tipo di codice compromette l'integrità del sistema, ottenendo privilegi più alti così da poter cambiare o disabilitare le funzioni di sicurezza del dispositivo. Alcuni esempi sono applicazioni che violano il principio dei permessi di Android o che rubano credenziali da altre applicazioni, applicazioni che fanno in modo di non essere disinstallate o bloccate e applicazioni che effettuano il cosiddetto *root* del dispositivo senza il permesso dell'utente.
- **Ransomware:** questi malware prendono controllo di una parte o della totalità dei dati nel dispositivo con l'intento di chiedere un riscatto agli utenti per sbloccare i loro dati. Alcuni *ransomware* cifrano i dati e chiedono il riscatto per decifrarli mentre altri levano il controllo dell'utente dal dispositivo e chiedono il riscatto per ripristinare la funzionalità.
- **Rooting:** mentre le applicazioni che non hanno intenti malevoli informano l'utente dell'operazione di rooting e, successivamente, non eseguono altre operazioni potenzialmente pericolose, i malware di questa categoria invece agiscono eseguendo il rooting senza informare l'utente, oppure informano l'utente ma dopo eseguono operazioni potenzialmente malevole.
- **Spam:** vengono mandati messaggi insoliti ai contatti dell'utente o il dispositivo viene usato per inoltrare email di spam.
- **Spyware:** trasmettono dati personali all'infuori del dispositivo senza un'adeguata notifica o consenso da parte dell'utente.
- **Trojan:** sono applicazioni che appaiono come benigne ma che eseguono azioni potenzialmente malevole. Solitamente, questa categoria è messa in combinazione con altre categorie precedentemente descritte.
- Un'ulteriore categoria, non presente nella lista fornita da *Google Play Protect*, è quella degli **Adware**. Queste applicazioni presentano un uso intensivo di

pubblicità a fini commerciali, che però possono generare rallentamenti e delle occupazioni di memoria nel dispositivo non indifferenti.

La problematica delle applicazioni malevole è quindi molto diffusa: infatti, alla fine del 2020, sono stati registrati poco più di due milioni di pacchetti software malevoli in circolazione [3].

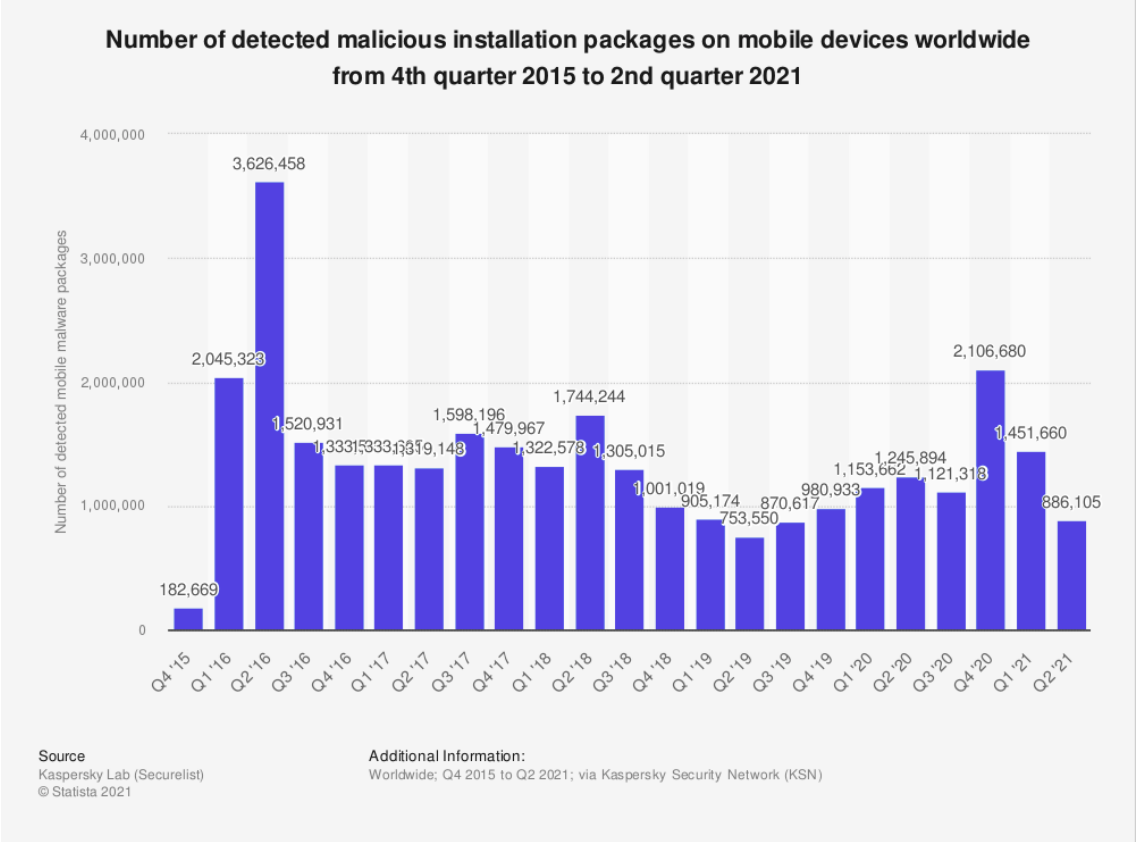

Figura 2. Numero di pacchetti malevoli nei dispositivi mobile in [3]

Il sistema operativo più colpito dagli attacchi delle applicazioni malevole è il sistema operativo *Android*: il motivo principale è che questo sistema operativo è *open source* e quindi permette a chiunque di poter utilizzare e modificare il codice sorgente del sistema operativo. In questo modo, per gli *hacker* (coloro che progettano e mettono in circolo le applicazioni malevole) è più facile individuare le vulnerabilità del sistema operativo da poter così sfruttare a proprio vantaggio.

Anche la diffusione nel mercato è un fattore che va ad incidere sulla predominanza dei malware all'interno di Android: infatti, alla fine del 2016, Android deteneva circa l'81%

di unità vendute in tutto il mercato globale [4]. Questa grande percentuale ha fatto sì che la superficie di attacco in cui gli hacker possono sferrare i loro attacchi corrisponda alla maggioranza dei dispositivi in circolazione.

La sicurezza nei dispositivi mobile è una caratteristica realizzabile in maniera meno immediata rispetto a quella dei dispositivi fissi, quali i computer. Questa differenza è data dal fatto che, mentre i dispositivi fissi possiedono risorse che permettono una elaborazione computazionalmente più onerosa, nei dispositivi mobile si lavora con delle risorse limitate, come una minore potenza di calcolo ed un consumo di energia che, rispetto alla loro controparte fissa, questi dispositivi non si possono permettere.
In questo ambito è quindi categorico cercare di sviluppare dei software di rilevamento dei malware capaci di individuare delle potenziali minacce nel minor tempo possibile e sfruttando la minore quantità di risorse.
Questo lavoro di tesi si colloca in un ambito in costante crescita ed evoluzione, in cui le strade da percorrere sono molteplici e non sono ancora state esplorate tutte. In particolare, si è sviluppato un sistema di riconoscimento di malware nei dispositivi *mobile* basato su tecniche di Machine Learning che permette di individuare le applicazioni potenzialmente pericolose per gli utenti prima che queste vengano installate nel dispositivo.
Il sistema in questione estrae ed elabora alcune delle caratteristiche statiche, ottenute osservando unicamente il file di installazione dell'applicazione, ed alcune delle caratteristiche dinamiche, ottenute eseguendo l'applicazione e osservando il suo comportamento.
Con questo lavoro ci si pone quindi l'obiettivo di fornire un ulteriore studio e strumento per difendere gli utenti che, ignari del pericolo, rischiano di compromettere i loro dati e i loro dispositivi.

Il Capitolo 2 fornisce una visione dell'attuale stato dell'arte, elencando alcuni lavori che hanno fornito dei contributi rilevanti e che hanno dato spunti per lo sviluppo di questo progetto.
Il Capitolo 3 fornisce una panoramica sugli algoritmi di machine learning che verranno usati per effettuare la classificazione in questo lavoro di tesi.

Il Capitolo 4 fornisce una descrizione del sistema proposto, spiegando la struttura del lavoro e le scelte progettuali effettuate.
Il Capitolo 5 fornisce una panoramica dei dataset che sono stati individuati e che sono stati tenuti in considerazione per lo sviluppo del progetto.
Il Capitolo 6 contiene gli strumenti utilizzati, quali dispositivi e tecnologie utilizzate, gli esperimenti effettuati e i risultati ottenuti dal sistema in ogni sua parte e nella sua interezza.
Il Capitolo 7 contiene le conclusioni e i possibili miglioramenti che potrebbero essere attuati in questo lavoro di tesi.

Bibliografia

[1] https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[2] The Evolution of Android Malware and Android Analysis Techniques.

[3] https://www.statista.com/statistics/653680/volume-of-detected-mobile-malware-packages/

[4] https://en.wikipedia.org/wiki/Usage_share_of_operating_systems#Mobile_devices

[5] (Li, 2018) Significant Permission Identification for Machine-Learning-Based Android Malware Detection

[6] (Qi Fang, Xiaohui Yang, Ce Ji, 2019) A Hybrid Detection Method for Android Malware

[7] (Daniel Arp, Michael Spreitzenbarth, Malte Hübner, Hugo Gascon, Konrad Rieck, 2014) - DREBIN : Effective and Explainable Detection of Android Malware in Your Pocket

[8] Android Security: A Survey of Issues, Malware Penetration, and Defenses

[9] (Liu, 2016), A Hybrid Malware Detecting Scheme for Mobile Android Applications

[10] (Arshad, 2018), SAMADroid A Novel 3-Level Hybrid Malware Detection Model for Android Operating System

[11] (A. De Paola, 2018), A Hybrid System for Malware Detection on Big Data

[12] (Xue, Li, 2019), Malware Classification Using ProbabilityScoring and Machine Learning

[13] (Kang, 2016), N-opcode Analysis for Android Malware Classification and Categorization

[14] (Mahima Choudhary, 2018), HAAMD:Hybrid Analysis for Android Malware Detection

[15] (Zhang, 2018), A Novel Android Malware Detection Approach Using Operand Sequences

[16] (Shen , 2019), Android Malware Detection Using Complex-Flows

[17] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon. AndroZoo: Collecting Millions of Android Apps for the Research Community. Mining Software Repositories (MSR) 2016

[18] Deep Learning - Ian Goodfellow, Yoshua Bengio, Aaron Courville

[19] https://virusshare.com/

[20] https://androguard.readthedocs.io/en/latest/#

[21] https://scikit-learn.org/stable/modules/svm.html

[22] https://scikit-learn.org/stable/modules/ensemble.html

[23] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon. AndroZoo: Collecting Millions of Android Apps for the Research Community. Mining Software Repositories (MSR) 2016.

[24] David Sean Keyes, Beiqi Li, Gurdip Kaur, Arash Habibi Lashkari, Francois Gagnon, Frederic Massicotte, "EntropLyzer: Android Malware Classification and Characterization Using Entropy Analysis of Dynamic Characteristics", Reconciling Data Analytics, Automation, Privacy, and

[25] Security: A Big Data Challenge (RDAAPS), IEEE, Canada, ON, McMaster University, 2021

[26] Abir Rahali, Arash Habibi Lashkari, Gurdip Kaur, Laya Taheri, Francois Gagnon, and Frédéric Massicotte, "DIDroid: Android Malware Classification and Characterization Using Deep Image Learning", 10th International Conference on Communication and Network Security (ICCNS2020), Pages 70–82, Tokyo, Japan, November 2020

[27] Samaneh Mahdavifar, Andi Fitriah Abdul Kadir, Rasool Fatemi, Dima Alhadidi, Ali A. Ghorbani; Dynamic Android Malware Category Classification using Semi-Supervised Deep Learning, The 18th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC), Aug. 17-24, 2020.

[28] Samaneh Mahdavifar, Dima Alhadidi, and Ali A. Ghorbani (2022). Effective and Efficient Hybrid Android Malware Classification Using Pseudo-Label Stacked Auto-Encoder, Journal of Network and Systems Management 30 (1), 1-34

[29] Sen Chen, Minhui Xue, Lingling Fan, Shuang Hao, Lihua Xu, and Haojin Zhu, "When Big Data Meets Cybersecurity: Adversarial Detection of Mobile Applications", Elsevier Computers & Security, 2017.

[30] *Martín, A., Lara-Cabrera, R., & Camacho, D. (2018). Android malware detection through hybrid features fusion and ensemble classifiers: the AndroPyTool framework and the OmniDroid dataset.*

[31] *Martín, A., Lara-Cabrera, R., & Camacho, D. (2018). A new tool for static and dynamic Android malware analysis.*

[32] *https://www.virustotal.com/*

[33] *Python Machine Learning - Sebastian Raschka*

[34] *https://koodous.com/*

[35] *https://scikit-learn.org/stable/modules/cross_validation.html*

[36] *https://developers.google.com/android/play-protect/phacategories*

[37] *Agate V., De Paola A., Ferraro P., Lo Re G., Morana M., SecureBallot: A secure open source e-Voting system, (2021) Journal of Network and Computer Applications, 191, art. no. 103165, DOI: 10.1016/j.jnca.2021.103165*

[38] *Agate V., De Paola A., Lo Re G., Morana M.A Simulation Software for the Evaluation of Vulnerabilities in Reputation Management Systems (2021) ACM Transactions on Computer Systems, 37 (1-4), art. no. 3458510 DOI: 10.1145/3458510*

[39] *Concone F., Lo Re G., Morana M., SMCP: a Secure Mobile Crowdsensing Protocol for fog-based applications, (2020) Human-centric Computing and Information Sciences, 10 (1), art. no. 28, DOI: 10.1186/s13673-020-00232-y*

[40] *Bordonaro A., De Paola A., Lo Re G., Morana M., Smart Auctions for Autonomic Ambient Intelligence Systems, (2020) Proceedings - 2020 IEEE International Conference on Smart Computing, SMARTCOMP 2020, art. no. 9239687, pp. 180 – 187, DOI: 10.1109/SMARTCOMP50058.2020.00043*

[41] *Agate V., De Paola A., Lo Re G., Morana M. DRESS: A distributed RMS evaluation simulation software, (2020) International Journal of Intelligent Information Technologies, 16 (3), DOI: 10.4018/IJIIT.2020070101*

[42] *Agate V., Curaba M., Ferraro P., Lo Re G., Morana M., Secure e-voting in smart communities, (2020) CEUR Workshop Proceedings, 2597, pp. 1 – 11*

[43] *Agate V., De Paola A., Lo Re G., Morana M. A Platform for the Evaluation of Distributed Reputation Algorithms, (2019) Proceedings of the 2018 IEEE/ACM 22nd International Symposium on Distributed Simulation and Real Time Applications, DS-RT 2018, art. no. 8601020, pp. 182 – 189, DOI: 10.1109/DISTRA.2018.8601020*

[44] De Paola A., Gaglio S., Lo Re G., Morana M. A hybrid system for malware detection on big data, (2018) INFOCOM 2018 - IEEE Conference on Computer Communications Workshops, pp. 45 – 50, DOI: 10.1109/INFCOMW.2018.8406963

[45] De Paola A., Favaloro S., Gaglio S., Lo Re G., Morana M., Malware detection through low-level features and stacked denoising autoencoders, (2018) CEUR Workshop Proceedings, 2058

[46] Concone F., De Paola A., Lo Re G., Morana M., Twitter analysis for real-Time malware discovery, (2017) 2017 AEIT International Annual Conference: Infrastructures for Energy and ICT: Opportunities for Fostering Innovation, AEIT 2017, 2017-January, pp. 1 – 6, DOI: 10.23919/AEIT.2017.8240551

[47] Agate V., De Paola A., Lo Re G., Morana M., Vulnerability evaluation of distributed reputation management systems, (2017) ValueTools 2016 - 10th EAI International Conference on Performance Evaluation Methodologies and Tools, pp. 235 – 242, DOI: 10.4108/eai.25-10-2016.2266868

[48] Agate V., De Paola A., Gaglio S., Lo Re G., Morana M., A framework for parallel assessment of reputation management systems, (2016) ACM International Conference Proceeding Series, 1164, pp. 121 – 128, DOI: 10.1145/2983468.2983474

[49] Agate V., de Paola A., Lo Re G., Morana M., A simulation framework for evaluating distributed reputation management systems, (2016) Advances in Intelligent Systems and Computing, 474, pp. 247 – 254, DOI: 10.1007/978-3-319-40162-1_27