# UNIVERSIDAD DE MURCIA

## ESCUELA INTERNACIONAL DE DOCTORADO

# Enhancing DGA-Based Botnet Detection Beyond 5G with On-Edge Machine Learning

# Aprendizaje Automático en el Edge para la Mejora de la Detección de Botnets DGA en Redes 5G y Futuras

**D. Mattia Zago**

**2021**

UNIVERSITY OF MURCIA

FACULTY OF COMPUTER SCIENCE

# Enhancing DGA-Based Botnet Detection Beyond 5G with On-Edge Machine Learning

Aprendizaje Automático en el Edge para la Mejora de la Detección de Botnets DGA en Redes 5G y Futuras

Author

**Mattia Zago**

Thesis supervisors

Assoc. Prof. **Manuel Gil Pérez**, *Ph.D.*
Prof. **Gregorio Martínez Pérez**, *Ph.D.*

Murcia, June 2021

The following PhD Thesis is a compilation of the next published articles, being the PhD student the main author in all of them:

- Mattia Zago, Manuel Gil Pérez, and Gregorio Martínez Pérez. "**Scalable detection of botnets based on DGA: efficient feature discovery process in machine learning techniques**." *Soft Computing* 24 (2020): 5517-5537.
  DOI: `10.1007/s00500-018-03703-8`
  J.I.F 2019: 3.050 (Q2)

- Mattia Zago, Manuel Gil Pérez, and Gregorio Martínez Pérez. "**UMUDGA: a dataset for profiling DGA-based botnet**." *Computers & Security* 92 (2020): 101719.
  DOI: `10.1016/J.COSE.2020.101719`
  J.I.F 2019: 3.579 (Q2)

- Mattia Zago, Manuel Gil Pérez, and Gregorio Martínez Pérez. "**Early DGA-based botnet identification: pushing detection to the edges**." *Cluster Computing*, in press.
  DOI: `10.1007/s10586-020-03213-z`
  J.I.F 2019: 3.458 (Q1)

# Contents

PhD Thesis Summary

Publications composing the PhD Thesis

# Acknowledgements

Writing this dissertation marks the conclusive act of this Spanish arc of my life, which would have never been possible without the invaluable support of countless people. This document is their achievement as much as it is mine: my sincerest gratitude goes to all of you. Nonetheless, here follows in no particular order a collection of people that I wish to separately thank.

- To my family, for providing all the means and support that I needed.
- To Sabrina, for enduring all these years despite the distance.
- To Pantaleone, for being always there when I need him most.
- To the people in the research group, for making the university feel like a second home.
- To Gregorio and Manuel, for believing in me when nobody else did it.
- To the Erasmus Student Network volunteers, for teaching me the importance of balancing personal and professional lives.
- To Alexandra, Giorgia, Javier, Jorge, Lee, María, Paola, Sander, and Sergio for demonstrating that the borders are just a line on a map.
- To Stefano, Fabio, and Giuseppe, for being the best friends I could have hoped for.

Finally, this work has been supported by a predoctoral grant from the Spanish National Cybersecurity Institute (INCIBE) within the program "Ayudas para la Excelencia de los Equipos de Investigación Avanzada en Ciberseguridad" ("Grants for the Excellence of Advanced Cybersecurity Research Teams") with code `INCIBEI-2015-27353`.

# Abstract

Notwithstanding the scientific community's efforts and results, malwares are still wreaking havoc of computer networks; among these threats, botnets are growing at an alarming rate and have been responsible for dangerous attacks. Indeed, in the past five years, notorious botnets such as Mirai, Roboto, or Kraken have been a primary target of the cybersecurity community. However, independently from the purposes of these malwares, the botnets are characterised by a common point of failure, *i.e.*, the communication channel. Infected devices need to reach out to the Command & Control (C&C) servers to download second-stage infections, perform malicious actions, or await further commands. As the infected devices are already connected to the internet, TCP/IP connections have been widely abused, notwithstanding the providers' efforts in blacklisting IPs and sinkholing fully qualified domain names (FQDNs). Domain generation algorithms (DGAs) have grown to a conventional approach to elude detection algorithms by generating pseudo-random rendezvous-points, *i.e.*, the C&C servers FQDNs. Although many machine learning (ML)-oriented frameworks have been theorised to identify and intercept DGAs, the problem is yet to be solved. As such, this PhD thesis's scope is to analyse the DGAs' outputs, known as algorithmically generated domains (AGDs), to provide a set of ML tools and privacy-aware methodologies that help identify these evasive patterns.

To be more precise, the objectives achieved throughout this research are twofold. On the one hand, this thesis aims to provide a characterisation of the DGAs aspects, including, among others, a comprehensive survey of previous literary contributions, data sources, and ML-based approaches for DGA-based botnet detection. On the other hand, it aims to integrate and improve the state-of-the-art by providing methods, strategies, and technologies to enable DGA-based botnet detection at scale. Specifically, signature patterns are identified in malicious AGDs using Natural Language Processing (NLP) techniques, and, the resulting learning models are designed as services to be dynamically deployed anywhere on the network.

As a result, this research encompasses literary survey, theory and framework crafting, experiments design and evaluations, and knowledge gaps identification and discussions. Under the compendium modality, the three chapters composing this PhD dissertation are outlined as follows.

- Firstly, a state-of-the-art survey on ML approaches to DGA-based botnet detection; the first chapter reports on supervised and unsupervised algorithms, their features sets, the definition of use cases and experiments, and, ultimately, the outline of

multiple research challenges to guide the thesis. Eventually, the experimental findings lay the foundations for AGDs formal and verifiable study.

- Secondly, a comparative analysis of the data sources to power ML frameworks; the second chapter reports on the published datasets by providing a formal comparison and discussion on multiple orthogonal properties. In the same article, the University of Murcia Domain Generation Algorithm Dataset (`UMUDGA`) is introduced as the most complete, balanced, and up-to-date collection of DGA-related data, featuring 50 malware classes for a total of 30+ million FQDNs. Eventually, the exploratory analysis reported in the article suggests that ML solutions to precisely pinpoint the malware variant based on AGDs pattern recognition are feasible.

- Thirdly, a proof-of-concept framework where the detection of DGA-based botnets is deployed as a security service on edge; the third chapter compares and examines architectural edge intelligence (EI) approaches to enable scalable detection in fifth generation (5G) networks and beyond. In the article, the experimental evaluation demonstrates that AGD detection is not only reasonable and achievable, but it is also plausible to expect to have deployed such detection capabilities on the networks' edges and eventually on the user equipments (UEs).

In summary, the chapters composing this PhD dissertation promote cohesive research exploring, analysing, and ultimately tackling the DGA-based botnets. Following this Ariadne's thread, each chapter is self-contained and provides critical insights on the research challenges from a different perspective; together, these contributions depict a clear description of the research niche summarised in the thesis. However, although conclusive on the explored subjects, some questions mooted by this research remain unsolved. Prime among them is whether it will be feasible to provide anonymous, exchangeable, and trustworthy profiles of AGDs to enable collaborative and federated detection models without harming users' privacy.

# Resumen

Ahora que "todo" se ofrece como servicio, maximizar el rendimiento y minimizar la latencia son requisitos fundamentales de cualquier proyecto, ya sea éste de investigación y desarrollo o comercial. Esto es especialmente relevante considerando los millones de dispositivos que van a estar conectados gracias a las redes de quinta generación (5G), y que solo representan una parte de los que se esperan que lleguen a estar conectados a las redes futuras (del inglés, Beyond 5G o B5G). Sin embargo, en este entorno la ciberseguridad es una necesidad que sigue estando infravalorada: no hay normas de obligado cumplimiento en los estándares, en particular cuando se hace referencia a los dispositivos más pequeños y limitados. Como tales, los volúmenes de datos que cada vez son mayores, y que en muchos casos están incorrectamente manejados, hacen que los ingenieros y los investigadores dedicados a la ciberseguridad tengan dificultades en encontrar soluciones capaces de ofrecer servicios de ciberseguridad dentro de entornos y escenarios con recursos limitados. De hecho, en la última década se han visto programas con software malicioso, o malware por su abreviatura en inglés, sembrando el caos en las redes de computadores, desde las pequeñas y medianas empresas (PYME) hasta las grandes corporaciones globales. En consecuencia, los expertos de todo el mundo están preocupados por el hecho de tener millones de dispositivos conectados a Internet y protegidos de forma inadecuada. Entre esas amenazas cibernéticas, las botnets están creciendo a un ritmo alarmante, siendo responsables de peligrosos ataques contra objetivos tales como las infraestructuras críticas de cualquier país. A modo de ejemplo, en los últimos cinco años, botnets como Mirai, Conficker o Kraken han sido un objetivo principal de la comunidad de ciberseguridad.

A pesar de los esfuerzos y resultados de la comunidad científica, el malware sigue causando pérdidas y daños en las redes informáticas. Aun siendo cambiantes en su comportamiento y omnipresentes, no todos los malwares son idénticos; de hecho, difieren en alcance, en técnicas aplicadas y en efectividad. Sin embargo, existen funcionalidades que son universales y que comparten las distintas familias de malware. A modo de ejemplo, los programas maliciosos necesitan contactar con el atacante que está dirigiendo el ataque para recibir comandos (*p.ej.*, botnets), para llevar a cabo la exfiltración de datos (*p.ej.*, software espía) o para proporcionar el acceso no autorizado (*p.ej.*, troyanos de acceso remoto – del inglés Remote Access Trojan, RAT). En particular, los dispositivos infectados por una botnet necesitan comunicarse con los servidores de Comando y Control (C&C) para descargar infecciones de segunda etapa, realizar acciones maliciosas o quedar a la espera de nuevos comandos.

Durante la década pasada, la tendencia era utilizar diversas formas de comunicación (o canales de contacto) con el grupo de criminales cibernéticos detrás del software malicioso; entre ellos, como los dispositivos infectados ya están conectados a Internet, las conexiones TCP/IP han sido objeto de un abuso generalizado. Desde la perspectiva del criminal cibernético, se ha demostrado que las conexiones directas entre los dispositivos infectados y las direcciones IP de los servidores de C&C son ineficaces. De hecho, poner esas direcciones IP en una lista negra para que sean bloqueadas es una técnica de bajo costo y fácil de aplicar, y muchas de esas listas están disponibles públicamente y actualizadas a diario (*p.ej.*, Spamhaus). Sin embargo, históricamente, los criminales cibernéticos han diseñado malwares que cuentan con varias direcciones IP, generalmente disponibles bajo un nombre de dominio totalmente calificado (del inglés Fully Qualified Domain Name, FQDN) específico y codificado dentro del propio programa malicioso, que cambian y rotan dinámicamente durante la vida útil del malware para evitar los mecanismos de protección basados en listas negras. Estos cambios rápidos de direcciones IP también son fácilmente abordables; los nombres de dominio configurados en el malware pueden ser bloqueados en unas pocas horas (una técnica conocida en inglés como sinkholing). Por lo tanto, los criminales cibernéticos han llegado a introducir módulos dedicados a contener algoritmos pseudoaleatorios como, por ejemplo, los algoritmos de generación de nombres de dominio (del inglés Domain Generation Algorithms, DGAs). Estos algoritmos incluyen fragmentos de código que, aunque diferentes en los detalles de implementación, sirven para generar nombres de dominio pseudoaleatorios conocidos como AGDs (por sus siglas en inglés de Algorithmically Generated Domain), los cuales podrían ser registrados por los criminales cibernéticos para actuar como puntos de encuentro entre los servidores C&C y los dispositivos infectados. Este paso intermedio permite que los cibercriminales puedan generar potencialmente millones de FQDNs de forma dinámica y sin tener la necesidad de registrarlos todos, ya que solo uno de ellos es suficiente para permitir la conexión entre el dispositivo y los servidores de C&C. El uso de estos DGAs se ha convertido en un enfoque recurrente y muy eficaz para eludir los equipos de ciberseguridad y los algoritmos de detección.

En la última década, la comunidad científica se ha dedicado a explorar soluciones de detección de botnets realizando herramientas basadas tanto en sistemas de reglas y heurísticas como en sistemas de aprendizaje automático. Actualmente, son tres las categorías principales de técnicas de detección de estas ciberamenazas que se suelen diseñar y desplegar en entornos tanto en producción como de investigación. Brevemente, las características de cada una de ellas se presentan a continuación:

*i)* En primer lugar, se encuentran los sistemas basados en reglas y heurísticas, que permiten analizar una gran cantidad de datos de manera rápida y representan la solución ideal para identificar vulnerabilidades y ataques bien conocidos. A pesar de estos beneficios, representan un sistema más bien reactivo, es decir, que necesitan que alguien defina el conjunto de reglas a añadir para detectar cada amenaza, bien conocidas de antemano.

*ii)* En segundo lugar, están los sistemas basados en el análisis e inspección profunda de paquetes (del inglés Deep Packet Inspection, DPI), que, a pesar de ser muy eficaces en algunos escenarios, resultan ser muy invasivos con respecto a la privacidad de los usuarios y casi completamente ineficaces en ecosistemas donde habitualmente se cifran las comunicaciones. Además, tampoco es una técnica apropiada para los grandes volúmenes de tráfico que caracterizan las redes actuales 5G, y aún menos para los previstos para las redes B5G.

*iii)* Finalmente, en la tercera categoría se localizan los sistemas basados en inteligencia artificial, que, a pesar de la necesidad de entrenar los modelos con grandes cantidades

> de datos, resultan ser muy eficaces en la detección con una cierta confianza tanto en las amenazas conocidas como en las desconocidas. Gracias a estas propiedades, las herramientas basadas en inteligencia artificial se pueden convertir en soluciones proactivas, es decir, soluciones que, dentro de un determinado grado de confianza, son capaces de identificar amenazas nuevas o desconocidas.

La comunidad científica se ha dedicado a explorar soluciones basadas en inteligencia artificial y aprendizaje automático (del inglés Machine Learning, ML), y aunque se hayan teorizado muchos sistemas orientados al ML para la detección de los DGAs, el problema aún no ha sido resuelto del todo. Como tal, el alcance de esta tesis doctoral se centra en analizar los nombres de dominio generados por los DGAs, de cara a proporcionar un conjunto de herramientas de aprendizaje automático y metodologías respetuosas con la privacidad que ayuden a identificar estos patrones evasivos.

Antes de desarrollar más los objetivos y los avances en el estado del arte proporcionados por esta tesis doctoral, es necesario comprender la amenaza expuesta por parte de los DGAs. Así, resulta imprescindible elaborar algo más sobre su magnitud, ya que, según los informes tecnológicos más recientes el número de servidores de C&C sigue aumentando. Hay que destacar también que, desde la perspectiva de los atacantes, los DGAs son prácticos y eficientes porque causan un esfuerzo totalmente asimétrico, entre los recursos necesarios para obtener una conexión positiva entre el bot y el servidor C&C en comparación con los recursos necesarios para bloquear todos los posibles nombres de dominio maliciosos. Por un lado, el programa malicioso puede generar millones de AGDs por día, y el criminal cibernético detrás de la red solo tendrá que registrar y activar un par de ellos de forma aleatoria. Por otro lado, en contraposición, los equipos de ciberseguridad necesitan una manera inteligente de comprobar cada nombre potencial de dominio antes de decidir si la conexión es legítima o maliciosa, y finalmente aplicar la respuesta o contramedida más adecuada. Es precisamente en este escenario asimétrico donde los algoritmos de ML, como el reconocimiento de patrones, aportan una importante contribución. A este respecto, existen diversas herramientas de ML que implementan numerosos algoritmos y técnicas de aprendizaje que pueden ser desplegados para identificar los AGDs y sus eventuales “firmas” (o la ausencia de ellas).

Así, no debería de sorprender que la cantidad de tráfico generado en redes a gran escala como las redes 5G y las B5G influya directamente sobre la elección de las soluciones de detección de ciberamenazas basadas en algoritmos de ML. Por ejemplo, no es razonable tener un sistema de detección de intrusiones único y centralizado, en el pasado se ha demostrado ampliamente que las soluciones descentralizadas basadas en servicios pueden tratar de manera eficiente y efectiva volúmenes mucho mayores de datos generados por los dispositivos de los usuarios. Más en detalle, la ciberseguridad en 5G y en B5G involucra soluciones de autoorganización y autoprotección a través de servicios de seguridad desplegados en redes y entornos virtuales mediante el uso de tecnologías como SDN (del inglés, Software Defined Networks) y NFV (del inglés, Network Function Virtualization). Estos servicios de seguridad, que se definen como SECaaS (del inglés Security as a Service), se despliegan como parte de un ciclo automatizado que incluye los procesos de detección, de análisis y de mitigación de las ciberamenazas de forma colaborativa, escalable y descentralizada. Una de las características fundamentales de estos enfoques de protección es de ser independentes del entorno de despliegue. En otras palabras, en cuanto se garanticen los recursos adecuados, tanto en tema de recursos de computación como de soporte en tema de software y librerías de código, no importa si el lugar efectivo de actividad es un entorno en la nube (Cloud), en los perímetros de las redes (Edge) o en los dispositivos de los usuarios.

En el caso concreto de los SECaaS que hagan uso de ML, hay que comentar que efectivamente hay algunas fases que no son viables de ejecutar en entornos donde los recursos son limitados. Por lo tanto, es oportuno diferenciar entre las fases de entrenamiento, evaluación e inferencia. De hecho, a lo largo de esta contribución científica se han desarrollado componentes independientes del entorno de despliegue que apuntan a maximizar la diferenciación de estas fases, de manera que se puedan garantizar las condiciones más adecuadas para cada una de ellas. Así, usando por ejemplo técnicas de aprendizaje federado, cada fase puede ser optimizada en función del lugar de despliegue, ya sea en la nube, en el perímetro de la red o en los mismos dispositivos de los usuarios. De esta manera también se estarían garantizando que los datos obtenidos de los dispositivos no se transmitan más allá de lo necesario, ya que, en ciberseguridad, uno de los temas más en auge es precisamente el de limitar, en la medida de lo posible, la difusión de los datos personales a la nube.

Por lo tanto, en resumen, esta tesis doctoral se centra en el estudio de las botnets basadas en DGA a través de herramientas de ML para desplegarse como SECaaS en los perímetros de las redes 5G y B5G, con el objetivo de contribuir a la mejora del estado del arte en la identificación de aquellos elementos que permitan distinguir actividades sospechosas en entornos altamente dinámicos. En concreto, se han utilizado herramientas propias de múltiples ramas de conocimiento dentro de la inteligencia artificial y del ML, como el reconocimiento de patrones comunes en los malwares que integran los DGAs. Por aportar un ejemplo, se han estudiado, implementado y evaluado técnicas de análisis del lenguaje natural (del inglés Natural Language Processing, NLP) a fin de identificar similitudes y diferencias en la sintaxis de los AGDs. Así, esta investigación multidisciplinar abarca estudios del estado del arte, elaboración de teorías y modelos basados en aprendizaje automático, y diseño de experimentos y evaluaciones, así como la identificación y discusión de brechas de conocimiento.

Fijando como objetivo principal de la tesis doctoral el avance del conocimiento en tema de ciberseguridad y botnet basadas en DGAs, en esta tesis doctoral se han realizado dos contribuciones principales. Por un lado, esta tesis doctoral proporciona una caracterización de los aspectos de los DGAs incluyendo, entre otros, un estudio completo de contribuciones anteriores presentes en el estado del arte, fuentes de datos y enfoques basados en aprendizaje automático para la detección de botnets basadas en DGA. Por otro lado, en esta tesis doctoral se ha conseguido el objetivo aún más ambicioso de integrar y mejorar el estado del arte en términos de técnicas y literatura, proporcionando métodos, estrategias y tecnologías para permitir la detección de botnets basada en DGA a gran escala, es decir, en redes 5G y en las B5G, a través de técnicas avanzadas de ML. Específicamente, los patrones de firma que se han podido identificar en los AGDs usando técnicas de NLP y resultando en modelos de aprendizaje diseñados para ser implementados en SECaaS para que luego se puedan desplegar dinámicamente en cualquier ubicación de la red.

Metodológicamente se ha utilizado la modalidad de compendio de publicaciones para la consecución de esta tesis doctoral, siendo tres los artículos que la componen y cuyo resumen se ofrece a continuación.

- El primer artículo presenta un estudio del arte exhaustivo de los enfoques de aprendizaje automático para la detección de redes de bots basadas en DGAs. Este artículo analiza y propone algoritmos supervisados y no supervisados, sus conjuntos de características, la definición de casos de uso y experimentos, y en última instancia, el esquema de múltiples desafíos de investigación para guiar la tesis doctoral. Los hallazgos experimentales que surgen de este primer artículo sientan las bases para un estudio formal y verificable de los nombres de dominio generados por los DGAs, llamados AGDs.

- El segundo artículo aporta un análisis comparativo de las fuentes de datos para impulsar los modelos de aprendizaje automático. Este artículo informa sobre los conjuntos de datos publicados, proporcionando una comparación formal y una discusión sobre múltiples propiedades ortogonales. En el artículo también se presenta el dataset UMUDGA como la colección más completa, equilibrada y actualizada de datos relacionados con los DGAs hasta el momento de su publicación, con 50 clases de programas maliciosos para un total de más de 30 millones de FQDNs. Además, el análisis exploratorio reportado en el artículo sugiere que son factibles las soluciones de aprendizaje automático basadas en el reconocimiento de patrones y NLP para identificar con precisión variantes de programas maliciosos.

- El tercer artículo presenta una prueba de concepto donde la detección de botnets basadas en DGA se implementa como un servicio de seguridad en los perímetros más lejanos de la red. Este artículo compara y examina enfoques arquitectónicos de Edge Computing para permitir la detección escalable en redes 5G y futuras. En el artículo, la evaluación experimental demuestra que la detección de los nombres de dominio maliciosos no solo es razonable y alcanzable, sino que también es plausible esperarse a que tales capacidades de detección sean desplegadas en los perímetros de las redes, e incluso en los dispositivos de los usuarios finales.

En resumen, los artículos de investigación que componen esta tesis doctoral promueven una investigación que explora, analiza y, en última instancia, aborda las redes de bots basadas en DGA. Siguiendo este hilo conductor, cada artículo es autónomo y proporciona información crítica sobre los desafíos de la investigación desde una perspectiva diferente. En conjunto, estas contribuciones representan una descripción clara del nicho de investigación resumido en la tesis doctoral. Sin embargo, aunque concluyentes sobre los temas explorados, algunas cuestiones planteadas por esta investigación necesitan de mayor esfuerzo para su resolución. El principal de ellos es si será factible proporcionar perfiles anónimos, intercambiables y confiables para los nombres de dominio maliciosos a fin de permitir modelos de detección colaborativos y federados sin perjudicar la privacidad de los usuarios.

# PhD Thesis Summary

# 1 Introduction and motivation

Fast performances and low latency are strict requirements for every commercial solution and research proposal, especially now that “everything” is being offered as on-demand service. Additionally, despite the millions of devices connected since the advent of commercial fifth generation (5G) platforms, a radical increase is to be expected with beyond 5G (B5G) networks [1, 2]. The problem with these devices, however, is that their cybersecurity is often overlooked. There are no strict and enforceable specifications regarding the minimum security requirements of devices, especially resource-constrained ones. As such, the ever-growing volumes of poorly handled data make researchers and engineers struggle in finding innovative-yet-reliable solutions capable of delivering cybersecurity services in resource-constrained scenarios. Indeed, the last decade has seen malwares wreaking havoc of computers networks, from small and medium enterprises (SMEs) to massive global corporations: all around the globe, the experts have been, and still are, concerned with the risks of having billions of inadequately protected devices connected to the internet [3].

Besides being diffuse and pervasive, malwares differ in scope, applied techniques, and effectiveness. However, there exists a single universal functionality that is shared among them, *i.e.*, any malware needs to reach out to the owner for commands (*e.g.*, botnets), exfiltrate data (*e.g.*, spyware), or provide unauthorised access (*e.g.*, remote access trojans (RATs)) among others. The past decade trend was to use various communication channels to contact the cybercriminal group behind the malware; among them, HTTP(s)-based connections are the most common technique. Although effective in some scenarios, deep packet inspections (DPIs) and other content-based techniques are impractical in an ecosystem where connections are encrypted more often than not. Essentially, they are not suitable for the large volumes that characterise today’s 5G networks [4], let alone the envisioned numbers that will characterise B5G networks [1].

From the cybercriminal perspective, direct IP connections between the infected devices and the Command & Control (C&C) servers have been proved ineffective. Indeed, blacklisting such addresses is a low-cost, practical, and well-known technique; furthermore, daily-updated offenders’ lists are publicly available (*e.g.*, Spamhaus [5]). Cybercriminals have historically designed malwares that feature multiple IP addresses (generally available under a specific and hardcoded fully qualified domain name (FQDN)) that dynamically change and rotate during the malware’s expected lifespan to bypass this protection mechanism. However, these fast-flux IP addresses are also somewhat easy to tackle, as the rendezvous domain name can be blocked in just a few hours (a technique known as sinkholing). Hence, cybercriminals came up with the dynamic generation of FQDNs via pseudo-random generation modules within the malware code. Such modules contain a domain generation algorithm (DGA), a fragment of code that, although different in the implementation details, serves to generate pseudo-random domain names that might be registered by the cybercriminals to act as *rendezvous-points* between C&C servers and infected devices. This intermediate step permits the cybercriminals to generate millions of FQDNs dynamically without the necessity to register all of them; in fact, one available domain name is just enough to permit the connection between the infected device and the C&C servers [6].

To grasp the threat, it is imperative to understand its magnitude before all else; in fact, according to the most recent tech reports [5, 7], the number of botnets C&C is still increasing. From the cybercriminal point of view, DGAs are practical and efficient because of the asymmetrical effort required in contrasting them. On the one hand, a single malware variant can generate millions of algorithmically generated domains (AGDs) per day, having only a few of them registered and active; however, on the other hand, the security teams need to check each FQDN and decide the most appropriate response. In this precise scenario, artificial intelligence (AI), and precisely machine learning (ML) algorithms such as patterns recognition, show their potential. The ML toolbox offers a set of tools and techniques that can be deployed to identify the DGAs variants' signature (or their lack of); section I will further discuss the subject. Indeed, in 2019 the number of AGDs registered plummed [5], confirming the excellent research results published in the previous five years [6, 8, 9, 10, 11, 12, 13]. Nonetheless, domain registrars' defensive policies fall short when vetting the registration of AGDs, hence permitting the abuse to continue [7].

This PhD thesis focuses on studying these generation algorithms with machine learning tools to identify elements that can distinguish suspicious activities in highly dynamic 5G/B5G environments. Indeed, ML has been studied and successfully deployed to recognise common patterns in generated domains, often leveraging syntactic analysis from Natural Language Processing (NLP). As such, the first chapter of this thesis (Scalable detection of botnets based on DGA (Article 1–`SoCo`)) focuses on surveying the literature aiming to establish a trend in algorithms and, in general, machine learning applications. The second chapter (UMUDGA: A dataset for profiling DGA-based botnet (Article 2–`CoSe`)), however, focuses on the data sources used to train and validate these models; after a thorough review of the available references, it unveils the University of Murcia Domain Generation Algorithm Dataset (`UMUDGA`), a collection of 30+ million domain names and 50 malware variants. Finally, the third chapter (Early DGA-based botnet identification (Article 3–`Clus`)) further advances these subjects by focusing on architectures that can maximise massive detection performance while minimising privacy leakage.

Indeed, we deem necessary to differentiate two research questions, namely identifying malicious and legitimate domain names using ML techniques, and the capability to do so at scale. The aspects related to the former research question (*i.e.*, the identification of AGDs) are discussed and analysed in section I, while, the latter's challenges (*e.g.*, the identification of AGDs at scale) require introducing the edge intelligence (EI) paradigm [14, 15] and how DGAs detection might benefit from its usage. The EI's concept has been widely discussed under different names (*e.g.*, mobile computing and fog computing, among others, as reported in the third chapter of this thesis, Article 3–`Clus`). The key is that its fundamental principle still applies to the subject at hand, *i.e.*, the detection process as a collection of decentralised micro-services that benefit from a shared knowledgebase [15, 16]. At the core of the EI paradigm, to put it differently, there is exploring the synergies between the cloud services, the decentralised and often automated edges, and the user equipments (UEs). In the application boundaries defined by the DGA-detection, this collaboration outlines the ML components' separation into virtual services, available on-demand, that shares data or trained models. As such, legitimate and suspicious FQDNs might be gathered and identified locally and then shared (anonymously) with the network of detection modules or a centralised authority. Shared data can be used to improve a pre-trained model's detection performances, either centralised in the cloud or spread over several independent on-edge detectors.

In the third chapter (Article 3–`Clus`), various architectural approaches to achieve such a synergy have been studied and discussed. In essence, a detection framework should consider to discuss and eventually balance the desired detection performances, users' privacy, and agility required to face new and unknown threats. As a point of fact, the learning models' quality relies on the quality of the collected data, which ultimately needs to be detached and anonymised regarding the user base to prevent privacy-related issues, without losing the capability to represent the deployment environment.

Ideally, a transparent view of the data will result in models with improved detection rates capable of identifying known anomalies and new threats. Despite the anonymisation introduced in such a collaborative environment, it remains unclear if it is possible to profile and uniquely identify the users, thus increasing the risk of personal data exposure. On paper, provided that the data security is guaranteed, collaborative learning models permits to achieve the scalability required by the volumes involved with 5G networks (and beyond), without losing the agility needed to face newly identified threats. These concepts are further discussed and analysed in section II.

## I Domain name identification as a machine learning task

Differentiate malicious AGDs from legitimate FQDNs can be seen as a pure ML task, independent from the actual use case or application environment. In such a scenario, key performances are the classification ones (*e.g.*, precision and recall, among others) rather than training and testing time or resources requirements. Under this prism, deep learning (DL) techniques provide good results without requiring too much work on the data preprocessing; however, their nature of "black box" algorithms make them of difficult interpretation, especially when it comes to making sense of the outcome results. As the literature suggests (*cf.*, Article 1–`SoCo`, Section 2) and experiments demonstrate (*cf.*, Article 1–`SoCo`, Section 3.3, Article 2–`CoSe`, Section 4, Article 3–`Clus`, Section 3), the Neural Networks (NNs)' sophistication is not required to tackle the DGA-related challenges. In other words, the collected AGDs call for a straightforward feature engineering process, rather than the automated, self-learning approach where DL stands out.

Furthermore, in a real-world application scenario where data sharing has to be examined (and its ramification discussed), collecting vast amounts of data required for training DL models could be difficult. On the contrary, classical ML approaches suffer less from the lack of sufficiently large, precisely labelled datasets. For another thing, as previously stated, one could argue that DL solutions could potentially work without features, hence simplifying the preprocessing steps required to process the data; however, the features identification process will still be carried out by the first layers of the network, that, in turn, require training. The DL training phase might also carry challenges related to required training resources and the amount of data required to feed the model. In the end, the data itself has to be encoded and eventually scaled or normalised before usage, thus impairing the benefits of avoiding a predefined feature extraction process. Thus, the application of classical ML algorithms, rather than DL ones, has been deemed more suitable to the task at hand.

Besides, one could argue that the DGA-related challenges have been sufficiently discussed in the literature, and thus solved. However, as demonstrated in this thesis, this is not the case. Essentially, each chapter of this thesis unveil critical shortcomings of the previously published frameworks and results, ultimately pointing out at yet-unsolved research challenges. Indeed, three main themes should be evaluated, namely *i)* the feature engineering process, *ii)* the data sources, and *iii)* the applied learning model. These aspects will be discussed in the following paragraphs.

**Feature analysis** — One of the essential aspects drawn from this thesis' feature analysis has been identifying two families of features, namely the context-aware and the context-free ones. In other words, the gathered features have been divided into two families, depending on whether they include (or not) users' personal and behavioural data. For example, packet-inspection -related features (such as time-based ones) are included in the context-aware family; on the contrary, NLP features (such as the ratio between characters) are included in the context-free family. The first chapter of this thesis (Article 1–`SoCo`) is dedicated to the literature review of these features' aspects; however, the feature formalisation and implementation has been studied in the second chapter of the thesis (*cf.*, Article 2–`CoSe`, Section 3).

Across this thesis, though, it has been proved that most of the gathered features are not adequate in describing the data. To be precise, in the first chapter (*cf.*, Article 1–`SoCo`, Section 3), several feature selection and feature extraction algorithms have been applied, leading to demonstrate that, in general, only a handful of features are needed to classify the AGDs correctly. Furthermore, in the second chapter (*cf.*, Article 2–`CoSe`, Section 4), it has been shown that, albeit using the full set of features, AGDs generated by some variants are indistinguishable. Likewise, the third chapter (*cf.*, Article 3–`Clus`, Table 2 and Article 3–`Clus`, Table 3), pointed out that limiting the feature set to the top 10 most informative ones significantly improves the resource consumption without unreasonably hindering the classification performances.

**Data sources** — For another aspect, there is a general lack of publicly available and extensive datasets regarding AGDs (and FQDNs in general). As firstly identified in the first chapter of this thesis (Article 1–`SoCo`), a quantitative comparison between the proposed frameworks is unfeasible, mainly due to the lack of shared data. The second chapter of this thesis (*cf.*, Article 2–`CoSe`, Table 1) focuses on surveying published and well-recognised data sources related to network traffic to address this shortcoming. In the article, nine orthogonal metrics have been designed, discussed and studied to present a formal comparison of the data sources available in the literature. The analysis led to the publication of a new data source that satisfies all the identified properties. Specifically, the second chapter presents both the data collection framework and the methodology followed to create the dataset (*cf.*, Article 2–`CoSe`, Section 3).

**Learning models** — Finally, several attempts have been made on the subject of the ML models for AGDs identification. As presented in the first chapter of this thesis, the literary review manifests the community's interest in exploring several potential algorithms (*cf.*, Article 1–`SoCo`, Section 2). In the article, several ML frameworks have been identified from the literature, leveraging supervised, unsupervised and mixed approaches. However, across the scale surfaced a general lack of reproducibility in terms of data, features and algorithms configurations.

For instance, as reported in the first chapter of this thesis (Article 1–`SoCo`), most literature claim to achieve good to excellent classification results, without providing information regarding either the data, the preprocessing, or the models' hyperparameters. As such, the first chapter pivots on two main contributions: providing a complete list of tested approaches (in the form of a comparative analysis of algorithms and claimed results) and an exhaustive list of studied features.

On the contrary, the second chapter analyses the dataset with five among the most common ML algorithms. The resulting exploratory analysis provides useful insights into the dataset composition, classes, and properties (*cf.*, Article 2–`CoSe`, Section 4).

Finally, the third chapter (Article 3–`Clus`) explores the capabilities of lightweight, tree-based classifiers at scale. To be precise, a collaborative framework is theorised, designed, and implemented leveraging the first two chapters' results (Article 1–`SoCo` and Article 2–`CoSe`). In such a scenario, the main focus is on the users' privacy; indeed, a shared-intelligence system can be designed based on the federated learning theory to provide anonymous and shareable knowledge without having to share users' data.

## II Cybersecurity as a service

Besides the challenging aspects of the ML tasks, the application use cases offer another perspective. Indeed, when deploying intrusion detection systems (IDSs), it is imperative to find the balance between detection performances and resources requirements. In other words, it is often convenient to accept a reduction in the detection rate in return for higher traffic volumes. In such a context, metrics like the classification yield or the resources consumption assume a higher relevance than in the pure ML tasks, and further optimisations are required to determine the equilibrium.

Indeed, the detection probes' locations influence, as expected, the amount of traffic that can be successfully inspected. For example, it is unreasonable to have a single and centralised detection module, especially when considering the volumes and the mobility requirements of 5G networks and beyond. On the contrary, multiple and decentralised IDSs are often deployed to scale the protection modules to cover a higher volume of connected devices. To be precise, the scalability-related challenges have been widely explored in the past, each time regarding the application scenario's specific requirements–mobile and not. However, with the explosive increase of connected devices related to the 5G and B5G networks, the paradigm of the security-as-a-service (SECaaS) becomes critical.

Traditionally, the cybersecurity approach to 5G involved the network components' self-organisation using software-defined network (SDN) and network functions virtualization (NFV) technologies. The virtualised services are in a cycle where the processes of detection, analysis, and mitigation of security threats work simultaneously and in a coordinated fashion. In such a scenario, this thesis aligns with the detection and analysis services by exploring, theorising, and discussing architectural designs to offer DGA-botnet detection as a dynamic module compatible with modern networks' strict requirements. Therefore, the concept of EI has been investigated and combined with the federated learning theory, ultimately proving that the detection process is not only feasible on the networks' farthest edges, but also on UEs. Indeed, the third and final chapter of this thesis (Article 3–`Clus`) focuses on these aspects of detecting and identifying DGA-based malwares leveraging the EI theory.

In the third article (*cf.*, Article 3–`Clus`, Section 2.2), several EI-compatible architectures are identified and studied, aiming to establish a knowledgebase useful to deploy DGA-detection modules by decoupling the training and testing phases from the model inference process. By virtualising these processes as services, mixed cloud-edge-device scenarios become available via technologies such as the SDN and NFV. Indeed, throughout this thesis, the main design principle has been to provide modular services that can be plugged in as independent components in a complex framework such as the ones proposed in 5G/B5G researches. By guaranteeing the separation between the learning process elements and phases, the compatibility with the SECaaS paradigm is achieved, and, as a consequence, whether the detection modules are designed and realised internally or outsourced becomes a secondary question. Indeed, as pointed out in the first chapter of this thesis (Article 1–`SoCo`), the Context-Free features used throughout the research can be extracted from the collected FQDNs independently from the actual network traffic, inspection techniques or technology implementation. Similarly, the second chapter's principal contribution (Article 2–`CoSe`) (*i.e.*, the `UMUDGA` dataset) might benefit any SECaaS provider by providing *i)* the high-quality labelled data to train ML models, and, *ii)* the testing data to be injected to evaluate the framework's performances.

As a consequence of the SECaaS paradigm compatibility, properties such as the managed execution and isolation can be guaranteed at any moment. The former ensures that the services can be deployed independently and where they are needed the most. On this subject, the third chapter of the thesis (*cf.*, Article 3–`Clus`, Figure 2) explores the different configurations from cloud to edge deployment, and eventually inferred on-device compatibility. Similarly, the latter property can be enforced by carefully deploying the services through containerisation, preventing unauthorised access to the model and the data itself. In this context, the potential capabilities hinted by developing federated learning solutions might permit to decouple the private data and models from the shared re-trained models. This subject has been defined and discussed, along with the privacy-related aspects and challenges, in the third chapter (Article 3–`Clus`).

## III Objectives

Identifying AGDs at scale in complex-environment, such as the ones offered by 5G/B5G scenarios, pointed out several challenges and multiple criticalities that had to be addressed.

As such, the first objective, defined as follows, aims to identify where the state-of-the-art is in terms of DGA-based botnet detection, with special attention to ML approaches, data sources, and published frameworks.

**Objective 1: State-of-the-art (`O1-SoTA`).** Outline and study the aspects of DGA-based botnet detection in 5G and B5G scenarios.

As the first objective is broad in both scope and potential approaches, four sub-objectives have been outlined; to be precise, the first three objectives aim to identify and explore the state-of-the-art under different prisms, while the last one aims to draw the needed conclusions necessary to identify the research path.

***Objective 1.1.*** Study and present a critical revision of researches on DGA-based botnet detection in 5G/B5G scenarios previously published in high-quality journals and conferences.

***Objective 1.2.*** Study and present a critical revision of publicly available data sources to power ML detection frameworks.

***Objective 1.3.*** Collection and analysis of published ML solutions to identify samples of AGDs.

***Objective 1.4.*** Identify and present research gaps, challenges, and potential future lines.

Following the identification and discussion of the characteristics of the problem at hand, the second stage of this PhD thesis focuses on achieving a more hands-on objective. Indeed, starting from state-of-the-art ML solutions and approaches, combining them with the obtained data sources, and having the 5G/B5G as the primary use case, a novel, scalable, and service-oriented framework has been designed to achieve the next objective:

**Objective 2: Identification framework (`O2-FRMW`).** Theorise, design and implement a proof of concept ML-based and SECaaS-compatible identification framework for DGA-based botnet detection in 5G and B5G scenarios.

Similarly, four steps and milestones have been identified to achieve this second objective:

***Objective 2.1.*** Obtain and maintain a curated data source to enable reproducible ML experiments.

***Objective 2.2.*** Study and present architectural approaches to the detection of DGA-based botnets in modern networks.

***Objective 2.3.*** Design, implement, and evaluate ML approaches to DGA-based botnet identification.

***Objective 2.4.*** Design, implement, and evaluate a SECaaS container to enable ML detection in a collaborative environment.

Besides the two primary objectives just identified, a series of methodological principles to maintain the research on a scientific and reproducible track throughout the PhD thesis has been identified and formalised as follows.

*Principle* 1*:* **Reproducibility of the research** (`P1-REPR`). Each research contribution shall be studied, formalised with specific attention to those details that enable to replicate and validate the findings and experiments.

*Principle* 2*:* **Open science** (`P2-OPEN`). Each research contribution shall be publicly released, including knowledge, data, and source code.

*Principle* 3*:* **Keep-it-simple** (`P3-KIS`). Each research contribution shall be unravelled and reduced to essential components, to facilitate reproducibility, evaluation, and usage of the obtained result.

*Principle* 4*:* **Evaluation and validation of the results** (`P4-EVAL`). Each research contribution shall be published in high-ranking peer-reviewed venues.

# 2 Methodology

This PhD thesis has been conducted following a scientific approach based on researching the state-of-the-art, from which key points, challenges, and theorised solutions have been proposed. As a result of the first initial literary review, it has been made clear that several issues made it impracticable to reproduce and validate numerous published results. Therefore, any subsequent effort has focused on enabling research outcomes' quantitative and qualitative rigorous analyses and comparisons. Throughout the research, the methodology aimed at adhering to the three principles defined in the previous section, namely P1-REPR, P2-OPEN[1], and P3-KIS. By publishing these results, this thesis also heeds to the fourth principle, *i.e.*, P4-EVAL.

It is well known that it is imperative to match originals conditions to replicate any experiment or result. When depicted in ML solutions, the P1-REPR principle implies to deploy models with the same configurations using the same data sources (both raw data and preprocessing). Under such a prism, and aligned with the P1-REPR and P2-OPEN principles that guide this thesis, the first two chapters (Article 1–SoCo and Article 2–CoSe) address precisely the data sources, their elaboration and their analysis using ML models. Hence, to achieve the O1-SoTA objective, the first chapter identifies and collects several research articles published in the previous five years on the subject of DGA-based botnet detection.

The analysis of the state-of-the-art, as requested by the O1-SoTA objective, led to the definition and formalisation of the features sets proposed in the literature, divided into two general families (*cf.*, Article 1–SoCo, Section 2), *i.e.*, those that rely on users' data (Context-Aware features) and those that rely only on the domain name itself (Context-Free features), being the latter the most common one deployed. These Context-Free features have been studied, reimplemented, and evaluated in the first chapter of this thesis (Article 1–SoCo, reaching O1.1 sub-objective), and eventually formalised and published within the UMUDGA dataset (Article 2–CoSe, reaching sub-objectives O1.3 and O2.1). The conclusions that have been drawn from the knowledge acquired on this subject hinted that authors have previously theorised new features without pondering whether they provide enough information to justify the computation resources needed to calculate them (sub-objective O1.4). For example, as reported in the first chapter (Article 1–SoCo), considering metrics related to $n$Grams distributions with $n \leq 2$ requires to add a non-trivial amount of resources to the feature extraction process as there are $|S|^n$ valid symbols per distribution, where $S$ is the set of valid ASCII symbols for domain names and $n$ is the size of the analysed $n$Gram.

[1] Although the research has not been published under the Open Access modality, each article has been legally released without restrictions as pre-print copy in the appropriate repositories.

Furthermore, as the sequence of characters taken into account increases, the probability of having such combination represented in a domain name decreases, leading to mostly zeroed distributions.

Once again, the keep-it-simple principle (P3-KIS) is reflected in the methodology and results. Indeed, all three chapters (Article 1–SoCo, Article 2–CoSe, and Article 3–Clus) demonstrated that privacy-aware, syntactic analysis of the domain names is enough to achieve outstanding classification performances, without the need to explore elaborated metrics that require profiling the users' behaviour. In the first chapter of this thesis, an exploratory analysis of the data using different feature selection and extraction techniques is carried out to prove the results (sub-objectives O1.3 and O2.3). In the experiments, six among the most famous (and used in the literature) classifiers have been deployed to identify the DGAs variants, suggesting that most of the identified features are indeed not providing a sufficient amount of information. In each chapter (Article 1–SoCo, Article 2–CoSe, and Article 3–Clus) the data, algorithms, and evaluation results have been thoroughly described and formalised (P1-REPR principle). However, it is yet to be discussed if solutions using Context-Aware features can outperform the already excellent results obtained in the experiments (O1.4 sub-objective).

The literary review also pinpointed that the already published researches were carried out mostly without releasing the data sources (sub-objective O1.3 and principle P2-OPEN). Hence, this thesis's second chapter focuses on the formal analysis between the published resources and the newly presented UMUDGA dataset (sub-objectives O1.3 and O2.1, P2-OPEN principle). Furthermore, a comparison framework has been designed to achieve sub-objective O1.2; in the framework, nine orthogonal metrics summarise the different properties, advantages and disadvantages of each data source, ultimately highlighting their lack of completeness. Among the metrics presented in the second chapter (Article 2–CoSe), it is possible to pinpoint the sources' verifiability, the data extensibility, and the ML readiness. Among others, these metrics suggested once again the critical importance of reproducible (P1-REPR principle) and usable open data (P2-OPEN principle). Indeed, the data collection process has been carried out by collecting the numerous malware variants from publicly available sources (such as previous researches, tech blogs, or vendor-specific bulletins, to cite a few) and executed it with predefined seeds (to control the output deterministically) to collect the resulting AGDs (sub-objective O2.1).

Furthermore, to complete the achievement of the O1-SoTA objective, a literary review of different architectural designs for DGA-based botnet detection at scale has been conducted. The last chapter (*cf.*, Article 3–Clus, Section 2.2) provides a technical overview of the different SECaaS architectures for ML-based detections on the network edges (sub-objectives O1.1 and O1.3). The article explores the different properties of the detection probes' location, providing the critical points to foster the discussion regarding the trade-off between classification capabilities, resource constraints, and limitations in user data usage.

Similarly, the O2-FRMW objective requested components and modules can be identified in each of the chapters composing this thesis. Indeed, following a bottom-up approach, the first and foremost element can be represented by the malwares' DGAs source code. As described in the second chapter (*cf.*, Article 2–CoSe, Section 3), and to achieve sub-objective O2.1, 50 malware variants have been collected; hence, their DGAs have been reimplemented and executed to generate at least 10.000 AGDs per variant (most DGAs, however, have been used to collect one million samples). As a result, the collected data, renamed as the "UMUDGA dataset", has been processed and publicly released after a peer-review process (Article 2–CoSe) and have been used to feed any experiment publicly released in this thesis (sub-objective `O2.1`, `P1-REPR` and `P2-OPEN` principles). Besides, with the formalisation of the studied features (Article 2–`CoSe`), each methodological principle can be identified: the designed methodologies and deployed procedures have been formalised in all research contributions and have been evaluated by the scientific community (`P1-REPR` and `P4-EVAL` principles); the source code has been released for each step of the research, including proof-of-concepts services and algorithms (`P1-REPR` and `P2-OPEN` principles); multiple approaches and solutions have been studied and ultimately formalised in the publications, proving that straightforward approaches can achieve excellent results without the need for obscure or uninterpretable constructs (`P1-REPR` and `P3-KIS` principles); each contribution has been published in competitive, high-quality, and peer-reviewed journals (`P4-EVAL` principle), namely Soft Computing (Article 1–`SoCo`), Computers & Security (Article 2–`CoSe`), and Cluster Computing (Article 3–`Clus`).

Indeed, the experimental detection modules have been studied and developed since the first (Article 1–`SoCo`) and second chapters (Article 2–`CoSe`), in which six among the most used classifiers have been described, implemented, trained and tested (`O2.3`). In these publications, the detection modules are examined as machine learning solutions, thus focusing on the classification performances in different scenarios and with different data sources. However, the third chapter (*cf.*, Article 3–`Clus`, Section 3) investigates the detection module as a service (`O2.4`), highlighting and discussing properties such as the requirements and performances, service deployment location, and eventually, how the detection probes relate to the users' privacy (`O2.2`). Finally, throughout these contributions, it is possible to identify the keep-it-simple principle (`P3-KIS`); for example, although deep and convoluted DL architectures have been proven effective, we have demonstrated that many intelligible and straightforward solutions can be equally valid.

# 3

# Conclusions and future work

Formally, this PhD dissertation aims at providing a study on domain generation algorithms (DGAs), and specifically on the techniques that can be used to identify them in the wild. However, this research's unwritten objective is to untangle the amount of machine learning (ML)-based contributions (that claims to solve the algorithmically generated domains (AGDs) identification problem) to describe a straightforward, working, and scalable approach that does not jeopardise users' privacy. As such, guided by keep-it-simple (`P3-KIS`)'s principle, this PhD thesis surveyed the literature regarding algorithms, data sources, and frameworks to DGA-based botnet detection in fifth generation (5G) networks and beyond. Indeed, security-as-a-service (SECaaS) and edge intelligence (EI) have been studied and applied to provide the required dynamicity characteristic of these modern environments.

Among the leading contributions achieved, three specifically outshine the others:

*i)* firstly, the formalisation of the designed features used to build ML solutions oriented to the detection of DGA-based botnet, with particular attention to those that do not require users' profiling (*cf.* Article 1: Scalable detection of botnets based on DGA);

*ii)* secondly, the collection and public release of a complete and up-to-date dataset for DGA-based malwares, including 50 DGAs and over 30 million AGDs (*cf.* Article 2: UMUDGA: A dataset for profiling DGA-based botnet); and,

*iii)* thirdly, the proof-of-concept AGDs detection module's design and implementation devised to be deployed as on-edge SECaaS (*cf.* Article 3: Early DGA-based botnet identification).

As per the evaluation and validation of the results (`P4-EVAL`) principle, these primary contributions have been published in top-tier journals, namely two Q2 (*Soft Computing* for Article 1 and *Computers & Security* for Article 2) and a Q1 (*Cluster Computing* for Article 3). Nowadays, research in computer science, particularly in cybersecurity and machine learning, still suffers from the lack of reproducibility. Indeed, many results and solutions claim to achieve exceptional results, and while that might be the case, the replicability and validation processes are often overlooked. As such, each contribution presented in this PhD thesis adhered to the principles of reproducibility of the research (`P1-REPR`) and open science (`P2-OPEN`). In doing so, the achieved results might be evaluated, reimplemented, and eventually improved by the scientific community and future researchers.

Together, they deliver the PhD dissertation's established objectives; however, some challenges are yet to be solved.

For one thing, this research focuses on analysing those features that do not require users profiling (context-free); however, their context-aware counterpart has been suggested as equally valuable in the literature as, intuitively, each malware variants behaves differently. As such, future works might reveal that a combination of these two families upholds the key to identify the most advanced malwares. Besides the potential advantages offered by context-aware features, it is imperative to consider that they require more invasive techniques than their context-free analogue. As such, future research should focus on how to provide AGDs detection services without harming users' privacy.

Furthermore, following the privacy subject, a specific issue strikes out: any given model needs to be trained with real-world data to learn its environment's characteristics. In such a context, sharing knowledge becomes a critical feature of collaborative detection frameworks; in particular, we hinted at the federated learning capabilities, without delving too much into it. Noteworthy future researches might explore this and other cooperation paradigms to unveil innovative solutions for identifying AGDs without having to share users data.

Last but not least, a remark is needed. Frameworks and proposals are designed, analysed, tested, and discussed in specific minimal use cases and scenarios. Despite this condition, research often overlooks the validation process and comparison with other previously published results. As such, the research area is in great need of a proper formal suite of validation benchmarks, to which new solutions should adapt.

# Bibliography

[1] Q. V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5G and beyond: fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020. DOI: 10.1109/access.2020.3001277

[2] C. Benzaid and T. Taleb, "AI for beyond 5G networks: a cyber-security defense or offense enabler?" *IEEE Network*, vol. 34, no. 6, pp. 140–147, Nov. 2020. DOI: 10.1109/mnet.011.2000088

[3] European Union Agency for Network and Information Security, "Malware threat landscape 2020," European Union Agency for Network and Information Security, Tech. Rep., 2020. [Online]. Available: https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/etl-review-folder/etl-2020-malware

[4] D. Papamartzivanos, F. Gómez Mármol, and G. Kambourakis, "Introducing deep learning self-adaptive misuse network intrusion detection systems," *IEEE Access*, vol. 7, pp. 13 546–13 560, 2019. DOI: 10.1109/access.2019.2893871

[5] Spamhaus Malware Labs, "Botnet threat report 2019," The Spamhaus Project SLU., Tech. Rep., 2019. [Online]. Available: https://www.spamhaustech.com/botnet-threat-report-2019/

[6] A. Khormali, J. Park, H. Alasmary, A. Anwar, M. Saad, and D. Mohaisen, "Domain name system security and privacy: a contemporary survey," *Computer Networks*, vol. 185, p. 107699, 2021. DOI: 10.1016/j.comnet.2020.107699

[7] Spamhaus Malware Labs, "Botnet threat update: Q1-2020," The Spamhaus Project SLU., Tech. Rep., 2020. [Online]. Available: https://www.spamhaus.org/news/article/798/spamhaus-botnet-threat-update-q1-2020

[8] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla, "A comprehensive measurement study of domain generating malware," in *25th USENIX Security Symposium*, Austin, TX, Aug. 2016, pp. 263–278. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/plohmann

[9] X. Luo, L. Wang, Z. Xu, J. Yang, M. Sun, and J. Wang, "DGASensor: fast detection for DGA-based malwares," in *5th International Conference on Communications and Broadband Networking*, Bali, Indonesia, Feb. 2017, pp. 47–53. DOI: 10.1145/3057109.3057112

[10] H. Mac, D. Tran, V. Tong, L. G. Nguyen, and H. A. Tran, "DGA botnet detection using supervised learning methods," in *8th International Symposium on Information and Communication Technology*, Nha Trang City, Viet Nam, Dec. 2017, pp. 211–218. DOI: 10.1145/3155133.3155166

[11] D. Tran, H. Mac, V. Tong, H. H. A. Tran, and L. G. L. Nguyen, "A LSTM based framework for handling multiclass imbalance in DGA botnet detection," *Neurocomputing*, vol. 275, pp. 2401–2413, 2018. DOI: 10.1016/j.neucom.2017.11.018

[12] R. R. Curtin, A. B. Gardner, S. Grzonkowski, A. Kleymenov, and A. Mosquera, "Detecting DGA domains with recurrent neural networks and side information," in *14th International Conference on Availability, Reliability and Security*, Canterbury CA, United Kingdom, Oct. 2019, pp. 1–10. DOI: 10.1145/3339252.3339258

[13] C. Catania, S. García, and P. Torres, "Deep convolutional neural networks for DGA detection," in *Computer Science – CACIC 2018*, Tandil, Argentina, Oct. 2019, pp. 327–340. DOI: 10.1007/978-3-030-20787-8_23

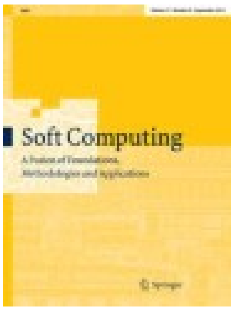
[14] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: the confluence of edge computing and artificial intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020. DOI: 10.1109/jiot.2020.2984887

[15] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020. DOI: 10.1109/comst.2020.2970550

[16] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: a comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020. DOI: 10.1109/comst.2020.2986024

# Publications composing the PhD Thesis

# 1 Scalable detection of botnets based on DGA



| | |
|---|---|
| **Title**: | Scalable detection of botnets based on DGA: *efficient feature discovery process in machine learning techniques* |
| **Authors**: | Mattia Zago, Manuel Gil Pérez, Gregorio Martínez Pérez |
| **Journal**: | Soft Computing |
| **J.I.F.**: | 3.050 Q2 (2019) |
| **Publisher**: | Springer |
| **Volume**: | 24 |
| **Pages**: | 5517–5537 |
| **Year**: | 2020 |
| **Month**: | January |
| **DOI**: | `10.1007/s00500-018-03703-8` |
| **Status**: | Published |

## Abstract

Botnets are evolving and their covert modus operandi, based on cloud technologies such as the virtualization and the dynamic fast-flux addressing, has been proved challenging for classic Intrusion Detection Systems and even the so-called Next-Generation Firewalls. Moreover, Dynamic Addressing has been spotted in the wild in combination with pseudo-random Domain names Generation Algorithm (DGA), ultimately leading to an extremely accurate and effective disguise technique. Although these concealing methods have been exposed and analysed to great extent in the past decade, the literature lacks some important conclusions and common ground knowledge, especially when it comes to Machine Learning solutions. This research horizontally navigates the state-of-the-art aiming to polish the feature discovery process, which is the single most time-consuming part of any Machine Learning approach. Results show that only a minor fraction of the defined features are indeed practical and informative, especially when considering zero day (0-day) botnet identification. The contributions described in this article will ease the detection process, ultimately enabling improved and more scalable solutions for DGA-based botnets detection.

## Keywords

Botnet · Domain Generation Algorithm · DGA · Machine Learning · Natural Language Processing

# 2

# UMUDGA: A dataset for profiling DGA-based botnet

| | |
|---|---|
| **Title**: | UMUDGA: A dataset for profiling DGA-based botnet |
| **Authors**: | Mattia Zago, Manuel Gil Pérez, Gregorio Martínez Pérez |
| **Journal**: | Computers & Security |
| **J.I.F.**: | 3.579 Q2 (2019) |
| **Publisher**: | Elsevier |
| **Volume**: | 92 |
| **Pages**: | 101719 |
| **Year**: | 2020 |
| **Month**: | May |
| **DOI**: | `10.1016/j.cose.2020.101719` |
| **Status**: | Published |

## Abstract

Advanced botnet threats are natively deploying concealing techniques to prevent detection and sinkholing. To tackle them, machine learning solutions have become a standard approach, especially when dealing with Algorithmically Generated Domain (AGD) names. Nevertheless, machine learning state-of-the-art is non-specialist at best, having multiple issues in terms of rigorousness, reproducibility and ultimately credibility. This research focuses on the first critical step of the training phase, that is, the collection of data suitable for being analysed by algorithms. We have detected a common lack of scientific rigorousness in the literature regarding the aforementioned AGD analysis and, therefore, we advocate two major contributions in this article: *i)* a thorough analysis of the cyber panorama in terms of botnets that make use of Domain Generation Algorithms (DGAs) as evasive techniques, that flows into *ii)* a full-fledged machine-learning-ready labelled dataset that features over 30 million AGDs sorted in 50 malware variant classes. This mature dataset aims to fill the gap in the comparability between the different researches published in the literature. Lastly, two minor contributions are also included in this article: *iii)* we designed an exploratory analysis of the proposed dataset to provide both data characteristics and potential future research lines, which eventually emerges as *iv)* a collection of suggested guidelines. When proposing a machine learning solution, researchers should adhere to it in order to achieve scientific rigorousness.

### Keywords

Domain Generation Algorithm (DGA) · Natural Language Processing (NLP) · Machine Learning · Botnet · Network Security

# 3

# Early DGA-based botnet identification

| | |
|---|---|
| **Title**: | Early DGA-based botnet identification:<br>*pushing detection to the edges* |
| **Authors**: | Mattia Zago, Manuel Gil Pérez, Gregorio Martínez Pérez |
| **Journal**: | Cluster Computing |
| **J.I.F.**: | 3.458 Q1 (2019) |
| **Publisher**: | Springer |
| **Volume**: | — |
| **Pages**: | — |
| **Year**: | 2021 |
| **Month**: | — |
| **DOI**: | `10.1007/s10586-020-03213-z` |
| **Status**: | Published |

## Abstract

With the first commercially available 5G infrastructures, worldwide's attention is shifting to the next generation of theorised technologies that might be finally deployable. In this context, the cybersecurity of edge equipment and end-devices must be a top priority as botnets see their spread remarkably increase. Most of them rely on algorithmically generated domain names (AGDs) to evade detection and remain shrouded from intrusion detection systems, via the so-called Domain Generation Algorithm (DGA). Despite the issue, by applying concepts such as distributed computing and federated learning, the cybersecurity community has prototyped and developed dynamic and scalable solutions that leverage the increased capabilities and connectivity of edge devices. This article proposes a lightweight and privacy-preserving framework that pushes the intelligence modules to the edges aiming to achieve early DGA-based botnet detection in mobile and edge-oriented scenarios. Experimental results prove the deployability of such architecture at all levels, including resource-constrained end-devices.

## Keywords

Domain Generation Algorithm (DGA) · Machine Learning · 5G · Cybersecurity · Edge Artificial Intelligence · Federated Learning