Scientific Research Publishing

Journal of Computer and Communications, 2020, 8, 12-28
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

The Application of Epidemiology for Categorising DNS Cyber Risk Factors

Jessemyn Modini, Timothy Lynar, Elena Sitnikova, Keith Joiner

School of Engineering and Information Technology, University of New South Wales at the Australian Defence Force Academy, Canberra, Australia
Email: j.modini@adfa.edu.au, t.lynar@adfa.edu.au, e.sitnikova@adfa.edu.au, k.joiner@adfa.edu.au

How to cite this paper: Modini, J., Lynar, T., Sitnikova, E. and Joiner, K. (2020) The Application of Epidemiology for Categorising DNS Cyber Risk Factors. *Journal of Computer and Communications*, **8**, 12-28.
https://doi.org/10.4236/jcc.2020.812002

Received: October 26, 2020
Accepted: November 30, 2020
Published: December 3, 2020

Abstract

This Epidemiology can be applied to cybersecurity as a novel approach for analysing and detecting cyber threats and their risks. It provides a systematic model for the analysis of likelihood, consequence, management, and prevention measures to examine malicious behaviours like disease. There are a few research studies in discrete cybersecurity risk factors; however, there is a significant research gap on the analysis of collective cyber risk factors and measuring their cyber risk impacts. Effective cybersecurity risk management requires the identification and estimation of the probability of infection, based on a comprehensive range of historical and environmental factors, including human behaviour and technology characteristics. This paper explores how an epidemiological principle can be applied to identify cybersecurity risk factors. These risk factors comprise both human and machine behaviours profiled as risk factors. This paper conducts a preliminary analysis of the relationships between these risk factors utilising Domain Name System (DNS) data sources. The experimental results indicated that the epidemiological principle can effectively examine and estimate cyber risk factors. The proposed principle has a great potential in enhancing new machine learning-enabled intrusion detection solutions by utilising this principle as a risk assessment module of the solutions.

Keywords

Epidemiology, Cybersecurity, Artificial Intelligence, Internet of Things (IoT), Epidemiological Security Analysis, Machine Learning

1. Introduction

The cyber terrain continues to expand at a rapid pace. From vehicles to fridges,

items are increasingly internet-connected and networked. This significantly expands the count of cyber-physical features and consequently the number of entry points for potential exploitation. Artificial intelligence (AI) and machine learning (ML) provide revolutionary means to analyse and respond to behavioural patterns across complex Internet of Things (IoT) ecosystems, in computer speed. It can provide exceptional facilitation of big-data correlation and pattern recognition across many complex factors. An IoT is a network of devices, for instance, vehicles, machines, and home appliances [1] which use sensors and network connectivity to transmit information [2]. Although IoT appliances are placed behind firewalls or routers with Network Address Translation (NAT), attackers would gain access to IoT systems using advanced and complex attacking techniques because of non-standard protocols of IoT devices-based Internet Protocol (IP) [3]. A botnet attack is one of the complex hacking techniques against IoT networks, which denotes a set of linked computers cooperating to implement suspicious and repetitive events to corrupt the resources of a victim such as DNS amplification attacks.

Cybersecurity systems, especially intrusion detection and prevention variants which exist in the industry are mostly discovering abnormal behaviours using methods that use anomaly-based, signature-based, heuristics-based, or hybrid-based [4]. The methods are effective at discovering well-known malicious activities attacks and known botnets. The detection methods often fail at recognizing new variants of attacks and new botnet families. These methods demand domain experts' knowledge to cope with the new types of botnets [5]. The existing methods need a manual update to their blacklists; therefore, they need more computational power, and cannot discover new attack families. One of the recent methods used is machine learning-based intrusion detection that attempts to understand the abnormal behaviours from data and classify them [6].

Machine learning methods are also vulnerable to adversarial attacks that would exploit the learning process [7]. We attempt to develop a new methodology that enhances the detection procedure of botnets and new attack families by assessing the progression of cyber threats. The new methodology depends on epidemiology which is a study that examines disease distribution and progression [8]. We propose utilising a novel epidemiology-based cyber-risk detection approach for understanding cybersecurity risk. This paper examines current literature on discrete cybersecurity risk factors and identifies the research gap in the analysis of collective cyber-security risk factors. This paper explores how epidemiological principles can be applied to determine a range of factors. These factors comprise both human and machine behaviours and characteristics profiled as risk factors. This paper conducts a preliminary analysis of the relationships between these risk factors utilising DNS data. DNS data contains a strong indication of human and machine behaviour indicators. Hence, it is a relevant data type to explore the relationship between people, devices, data and process; all fundamental elements of IoT.

The rest of this paper is structured as follows. Section 2 describes the background of epidemiology and how it would be used in cybersecurity applications. This is followed by a detailed description of epidemiology and related studies applied to cybersecurity, as explained in Section 3. Applications of epidemiology to DNS are presented in Section 4. Finally, in Section 5, the paper is summarised and future research direction is described.

2. Epidemiology and Cyber Security

Over recent decades humans have become increasingly connected; both in physical communities and in technology. The concept of epidemiology first emerged around 300 BCE out of the Hippocratic philosophy which began to shift public health from "mysticism to patient-oriented empiricism" [9]. Epidemiology is contemporarily defined as "the study (scientific, systematic, data-driven) of the distribution (frequency, pattern) and determinants (causes, risk factors) of health-related states and events (not just diseases) in specified populations (patient is community, individuals viewed collectively), and the application of (since epidemiology is a discipline within public health) this study to the control of health problem" [8]. It is a relationship and pattern-driven disciple, aimed at "comparisons to establish cause-effect relationships, evaluate information and make good decisions that will improve outcomes" [10]. The author illustrates that "human disease does not occur at random; there are factors or determinants which can increase or decrease the likelihood of developing disease". Therefore, an infection can be determined through a calculation of risk, where risk comprises likelihood multiplied by consequence.

Epidemiologists study root cause, community burden, history, impact, prevention, and management for diseases. To determine the risk of a particular disease or diseases on a person or community, epidemiologists study a range of "risk factors" [11]. These risk factors include genetic profiles, environmental factors [8], behaviours and health status including nutrition and inoculation history. These concepts can be directly applied to cybersecurity, where the human disease is equivalent to computer compromise, and human communities are equivalent to networks or the internet of everything (IoE). Cybersecurity is becoming more prevalent and critical by the day. The internet of everything is evolving, and the likelihood and consequences of cybersecurity attack are increasingly devastating. Cybersecurity attacks are a contemporary and human-led "disaster". Attacks can destroy individual livelihoods, businesses and whole economies, as well as weaken Nation states economically and militarily [12]. They can take down critical infrastructure, from communications to water and power [13]. The director of Homeland Security at the Center for Strategic and International Studies, David Heyman summarised cybersecurity risk in that, "we have a great sense of vulnerability, but no sense of what it takes to be prepared" [14].

Applications of epidemiology to cybersecurity have been prevalent for dec-

ades. In 1983, the technical computer virus was defined as "a program that can 'infect' other programs by modifying them to include a possibly evolved copy of itself. With the infection property, a virus can spread throughout a computer system or network using the authorizations of every user using it to infect their programs. Every program that gets infected may also act as a virus and thus the infection grows" [15]. This naming convention initiated the biological theme which has expanded to other forms of malware including "worms". The impact of cybersecurity incidents can be measured using epidemiological terminology, through "prevalence" and "cost of illness", where prevalence is the "number of existing cases of a disease in a population at a given time" [16] and cost of illness is likened to the cost for remedy including lack of productivity, costs of replacing hardware, software, potential reputational damage, etc.

These elements can be applied to "provide a systematic framework for the application and analysis of disease causes, spread and consequence, which can then be assessed to inform effective prevention and management methodologies" [17]. Cybersecurity experts are faced with constantly evolving threats. Actor tools, techniques, strategies and targets change by the day, resulting in a significant range of "risk factors" for consideration. These risk factors range from individual hardware and software attributes to configurations, networks, environments and behaviours. Hence, epidemiology provides a novel approach for the systematic analysis of these numerous risk factors. As epidemiologists monitor and prepare for public health disasters such as COVID-19, cybersecurity professionals can apply equivalent principles for planning, prevention and response to security disasters [18]. Epidemiological approaches are also highly effective for "allocating limited resources to obtain maximal benefits in disaster situations" [19]. Provided the resourcing issues that the cybersecurity is currently facing, this is highly pertinent for supporting efficient cybersecurity incident response. **Figure 1** demonstrates the linkages between epidemiology and cybersecurity.

The internet was first considered in similarity to biological systems in 1999 [20]. Since then, the immense likeness between the "propagation of pathogens (viruses and worms) on computer networks and the proliferation of pathogens in cellular organisms (organisms with genetic material contained within a membrane-encased nucleus)" [21] has inspired researchers to apply concepts of epidemiology to information and communications technology (ICT). The vast majority of research into epidemiological applications to cybersecurity is focused on the spread of malware across technology elements. Epidemiology has also been applied as a mathematical modelling technique for virus propagation analysis through fully connected networks. This is very limited as it will not determine "the effect of the topological structure of the Internet on the spread of computer viruses" [22]. Endpoint computers are the only risk factor elements considered in this method. Researchers have explored propagation inspired by biological paradigms over standard computer networks [24], peer-to-peer networks [23] [24] [25], IoT devices [26] and WiFi routers [27].

Figure 1. Schematic diagram comparing epidemiological principles to cyber security principles.

Extant literature focuses on intrusion detection and prevention methodologies for technical machine vulnerabilities (hardware and software). Literature is largely focused on a singular node and link network vulnerabilities. As such there is "limited systematic understanding of the factors that determine the likelihood that a node (computer) is compromised", in aggregate [28]. Scholars are noting a critical gap in "cyber epidemiology" research, which "treats individuals as highly distinct, independent, and important agents within a socio-technical system", and "advocate an approach to understanding how cybercrime thrives due to a failure to develop the understanding needed for effective behavioural control measures that are presented at the right place and the right time" [29]. Research into macro risk determination methods based on a range of factors is

scarce [28]. This research is essential for the detection, prevention, response and recovery of cyber attacks in an increasingly complex and interconnected world. Contemporary vulnerability and intrusion detection technologies are limited to known and published threat signatures and intelligence. These technologies, along with most of the research in this field, concentrate on singular indicators of compromise based on technical signatures. Intrusion detection and prevention technologies (known as IDS and IPS) are forms of finite perimeter defence and are not fully reliable for the mitigation of evolving cyber threats. They do not analyse human behaviour, nor contextualise a range of risk factors for the determination of risk. The risk of infection should be based on a range of system or network configurations and characteristics [29], plus elements of human behaviour which utilise artificial intelligence to evolve. Comprehensive and reactive models will "provide a scalable, resilient, and cost-effective mechanism that may keep pace with constantly evolving security needs" [25].

The closest form of truly comprehensive aggregate risk factor analysis can be seen in research on cognitive modelling of "dynamic simulations involving attacker, defender, and user models to enhance studies of cyber epidemiology and cyber hygiene" [30]. The researchers contend the importance of "wargaming" and simulation in both health care and cybersecurity, highlighting that "just as simulations in healthcare predict how an epidemic can spread and how it can be contained, such simulations may be used in the field of cyber-security as a means of progress in the study of cyber-epidemiology" [30]. The authors argue that epidemiology can be applied for the simulation of pandemic or disease outbreak, though prediction models based on existing behavioural data of threats. These simulations provide "realistic synthetic users for full-scale training/wargame scenarios", which will "enable much-needed research in cybersecurity and cyber-epidemiology" [30]. Such simulations need to include methods to inject non-deterministic behaviours [31].

3. Technical Applications of Epidemiology to Cybersecurity—DNS

Epidemiological concepts can be applied to extant research findings to form an aggregate risk profile. In recent research, a DNS Anomaly Detection tool (BotDAD) was proposed to detect a bot-infected machine in a network using DNS fingerprinting [32]. This technique analyses host DNS fingerprints on an hourly basis and identifies anomalous behaviour that diverges from standard machine behaviour [32]. Panza *et al.* [16] used clustering to group DNS domains basing on the similarity between their users' activity, then compared these groups by using Association Rules. This methodology identified patterns and trends in human behaviours. Other work focused on more niche human behavioural patterns in DNS data, including the profiling of behaviour ambiguity and behaviour polymorphism through analysis techniques including pattern upward mapping and multi-scale random forest classification [14] [33] [34] [35]. Concepts of

epidemiology demonstrate that both the likelihood and consequence of compromise can be determined for a comprehensive range of risk factors. These comprise environmental factors, human behaviour, and machine configuration and behaviour. Almost all security incidents contain multiple human elements across threat and defensive controls. As such, risk analysis should comprise both human and machine factors for realistic results. This paper looks to build on these approaches by analysing a comprehensive range of "risk factors" within a DNS dataset. These risk factors include normal machine behaviour, anomalous machine behaviour, normal human behaviour, and anomalous human behaviour.

4. Applications of Epidemiology to DNS

4.1. DNS Risk Factors

DNS data contains a strong indication of human and machine behaviour indicators. Hence, it is a relevant data type to explore the relationship between people, devices, data and process, which is form the fundamental elements of the Internet of Everything (IoE). It has a high volume of data, user types, host machine configurations and is encryption free. DNS is simply the machine-aided mechanism for resolving a word-based domain to an internet protocol address for any host on the internet [28]. It is a "yellow pages" for the internet, as humans understand and can remember English worded domain names (e.g., google.com), while computers understand numbers (IP addresses). DNS queries are generated when someone sends an email or visits a website. The DNS system leverages the DNS precursor, root name server, TLD name server and the authoritative name server to identify and route the end-user to the IP address that supports the domain that was searched for. The standard DNS function is characterised as normal machine behaviour.

Human behaviour often initiates the DNS query. This is most frequently through an action such as opening a browser, clicking a link, or typing a domain address or search query. These actions have an associated risk factor. For example, a user looking to access "http://google.com" will usually have a lower risk of compromise over a user looking to access a Dark Web host such as "http://hss3uro2hsxfogfq.onion/". This behaviour is classified as normal human behaviour, with a sub-classification of risk based on normal and anomalous activity type. These activity types can be intentional, or unintentional. DNS has inherent security weaknesses and vulnerabilities. These can be described in two categories: protocol attacks and server attacks. Protocol attacks compromise the DNS function. The first form of protocol attack is DNS cache poisoning, which allows malicious actors to "poison" the records and trick the DNS to re-direct and resolve malicious domains.

The second form of protocol attack is often referred to as "DNS spoofing", conducted in conjunction with "DNS ID hacking", where a malicious actor "spoofs" the packet's source address and ID fields, to answer a legitimate DNS request meant for a legitimate DNS server and impersonate the DNS reply. This

allows malicious actors to misdirect the requesting client, often to a malicious domain. The tactics, techniques, and procedures (TTPs) for DNS server attacks are rapidly evolving. There are two general types of DNS server attacks; those that exploit bugs in DNS software implementation and services on the DNS server, and denial of service attacks. Internal DNS servers will maintain a list of all server names and IP addresses for their managed domains. Any external query can gather this information. DNS also often relays query information from internal workstations to external servers which can provide hidden paths for exfiltration. DNS can also be used by attackers for reconnaissance activities, through DNS zone transfer attacks [32]. Malware also utilises Domain Generation Algorithms (DGA) to periodically generate several domain names that can be used for command and control servers [31]. These behavioural patterns can be characterised as malicious human behaviour, as a human is most often required to undertake these exploits, as shown in **Figure 2**.

4.2. Epidemiological Approaches to DNS Attach Analysis

The dataset used for this analysis was sourced from [32]. The authors utilised the dataset to create a DNS Anomaly Detection tool, called BotDAD, designed to use DNS fingerprinting to detect machines infected with botnets [32]. BOTDAD is an enterprise approach to anomaly detection and aims to build on extant approached which analysed based off failed queries. Similar DNS datasets have also

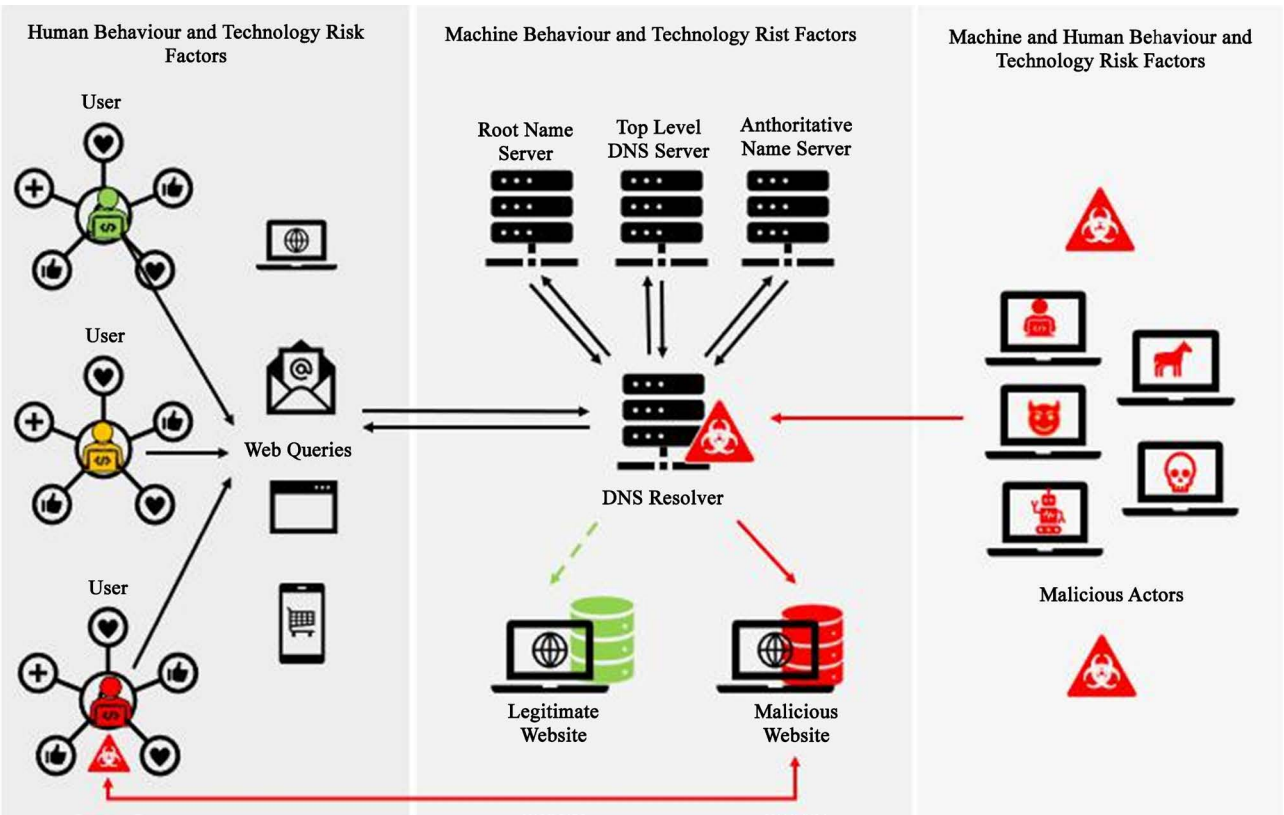

Figure 2. DNS human and machine based risk factors.

been used for the classification of malicious and non-malicious domains [18] [31] [32] [33] [36] and classification of IP flows [24]. This dataset has been used as an exemplar for rich DNS data. This data comprises a campus' DNS network traffic consisting of more than 4000 active users (in peak load hours) for random days in the month of April-May 2016. This set comprises DNS data (.pcap) from 23 April 2020 to 9 May 2016. 10 days of data was sufficient for a proof of concept, and contains enough data for a relational analysis and profiling of behaviour. A preliminary analysis was conducted on this dataset, to identify the categories of risk factors and the relationships between these risk factors.

4.3. Data Summary

Approximately 1.3% (7546 queries out of 601,092) of the DNS traffic was associated with malicious behaviour. **Figure 3** presents the number of clean versus malicious DNS queries over the 10 days of data. The malware was identified in every day of captured PCAP data. Nine variants were seen in total, broken down by percentage for each day of data, as depicted in **Figure 4**.

4.4. Data Features

Feature engineering underpins Artificial Intelligence (AI) and Machine Learning (ML) models. A feature engineering approach has been utilised to determine key data that contributes insight towards the heuristic categorisation of features and interactive features for the analysis of human and machine behaviours, or risk factors. Concepts of epidemiology will inform feature generation. The dataset comprised a range of malware variants, including IRC backdoor botnets, ransomware, worms, trojans and trojan downloaders. The data comprised the

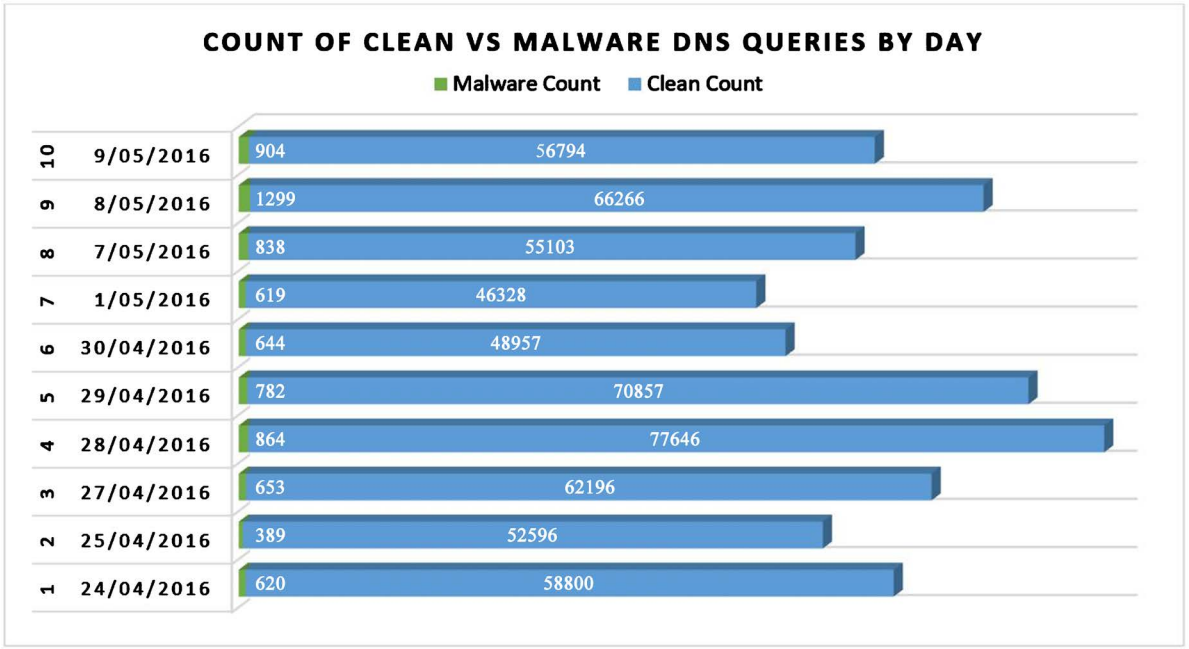

Figure 3. Count of clean and malicious DNS by day.

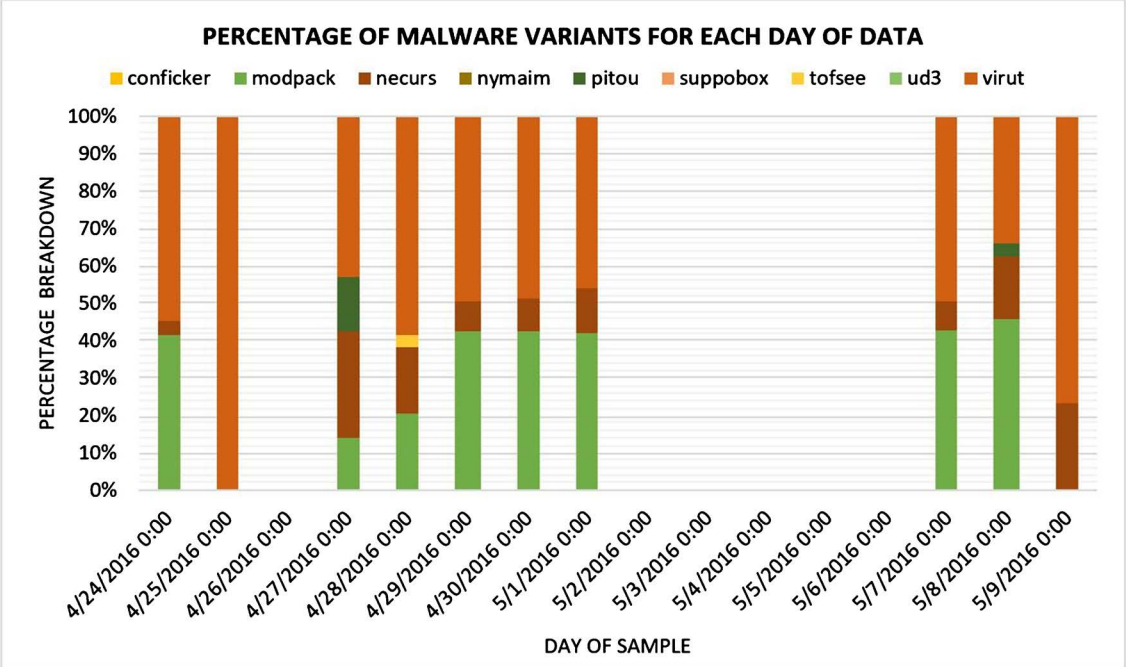

Figure 4. Percentage breakdown of malware variants present for each day of data.

following features:

1) Host IP Address
2) Time and Date of Query
3) Count of DNS Requests
4) Count of Distinct DNS Requests
5) Average Request per Minute
6) Count of IPV4 Address (A) DNS Requests
7) Count of Mail Exchanger (MX) DNS Requests
8) Count of Name Server Record (NS) DNS Requests
9) Count of Pointer Record (PTR) DNS Requests
10) Count of Distinct Requests to Top Level Domains (TLD)
11) Count of Distinct Requests to Second Level Domains (SLD)
12) Count of Distinct Requests to DNS Server
13) Count of Responses
14) City of IP Address
15) Country of IP Address
16) Count of Successful Responses
17) Count of Failed Responses
18) Average Time to Live (TTL) value
19) High Time to Live (TTL) Value
20) Classification: Clean or Infected
21) Botnet Variant
22) DNS Domain Name associated with Botnet DNS Queries

This DNS data can be categorised into four high-level categories of behaviour, as listed in Table 1. Using the DNS dataset, the following relationships between

Table 1. Risk factors in DNS data.

Risk Factor Categories	Example in DNS Dataset
Machine Behaviour—Normal	Normal machine behaviour comprises processes, queries and patterns that are expected elements based on system configuration. In the context of DNS, this is the true resolution of legitimate queries to the corresponding legitimate IP address. This also includes legitimate queries sent by machine services to support user applications.
Machine behaviour—normal	Anomalous machine behaviour is when a process, query or pattern error occurs. Machines do not make "mistakes". In DNS data, this occurs when a packet is lost, or a DNS server is unable to resolve the query. This can be caused by a loss of confidentiality (machine query compromised due to leaked password), availability (server is down) or integrity (protocol attacks). Malicious behaviour includes illegitimate queries sent by malware to command and control servers. This behaviour can indicate that the compromised machine is being utilised as an infrastructure to conduct further malicious activity.
Human behaviour—normal	Normal human behaviour comprises non-malicious queries to support legitimate internet browsing. In DNS data, this is when legitimate users are using the DNS service as it is designed. The majority of DNS query activity in the dataset was used for website browsing and email transmission.
Human behaviour—anomalous or malicious	Anomalous human behaviour is most often caused by human error. An example of this in the DNS dataset is where the text query has contained a typographical error that is still a recognised domain name, resulting in the DNS resolver pointing to a domain that is inconsistent with the user's intent. This domain could be high-risk. Malicious human behaviour is where an actor intentionally compromises the DNS service. This occurs initially through protocol and server attacks. Malicious human behaviour is also seen in the command and control of malware to spread infection and commandeer additional infrastructure within the network for malicious use.

risk factors were identified:

There is a large range of device types on the network. This diversity in IoT devices attracts a range of potential for compromise based on technical configuration and human behaviour (usage).

Device features (hardware, software and configuration) determine a range of risk factors. This dataset comprises a range of machine configurations, which have different risk profiles based on their susceptibility to compromise; both individually (specific hardware, software vulnerabilities) and in aggregate (configuration, or spread of compromise). The devices on the network comprise a range of operating systems. This is assumed by the range of Time to Live (TTL) values which differ by default e.g. MacOS is 64 for UDP, Windows is 128 for UDP, Linux can be 255 or 64 for UDP [37].

Operating System (software) is a key machine-type risk factor. The preliminary analysis of the DNS dataset demonstrates that there is a relationship between operating system type and risk of malware infection. All forms of malware identified were variants known to target and infect Windows Operating Systems only [31]. There is no evidence of macOS or Linux malware, or that macOS and Linux operating systems were compromised in this dataset.

There is a relationship between browsing high-risk websites and high-risk

downloading (from non-reputable torrent sites), and infection. A sub-sample of the DNS dataset saw users (172.31.149.56, 172.31.151.8) browsing BitTorrent and uTorrent sites minutes before initial evidence of malware presented. BitTorrent and uTorrent sites are known for supporting illegal file downloads. The evidence of malware comprised multiple and persistent hosts queries to known malicious command and control domains. It is highly likely that the high-risk user behaviour has resulted in a compromise.

There is a strong relationship between browsing frequency and infection. Within the DNS dataset, 5 malicious files (modpack downloads) were identified in top 30 IP addresses with the highest number of DNS requests. This equates to approximately 16.6% (0.166666). 4840 modpacks in all traffic (601,092 lines) equates to approximately 0.80% (0.00805). This demonstrates a high correlation between human behaviour in browsing frequency and the chance of downloading a malicious file. These files were downloaded on different days by a range of different IP addresses.

These relationships prove the concept of risk factor categorisation and contextual analysis to determine risk of compromise. This is explored in Table 1.

4.5. Malware Spread

Figure 5 illustrates the spread of malware through the university network, over time and by category. This demonstrates that a "mudpack" was the first instance of malware, sighted on day 1 (24 April 2016), as demonstrated in Table 2. From here, conficker was evidenced, followed by a surge in modpacks, which was followed by necurs, nymaim, pitou and suppobox malware variants.

From analysing Figure 5 and Table 2 and Figure 6, it appears that Conficker,

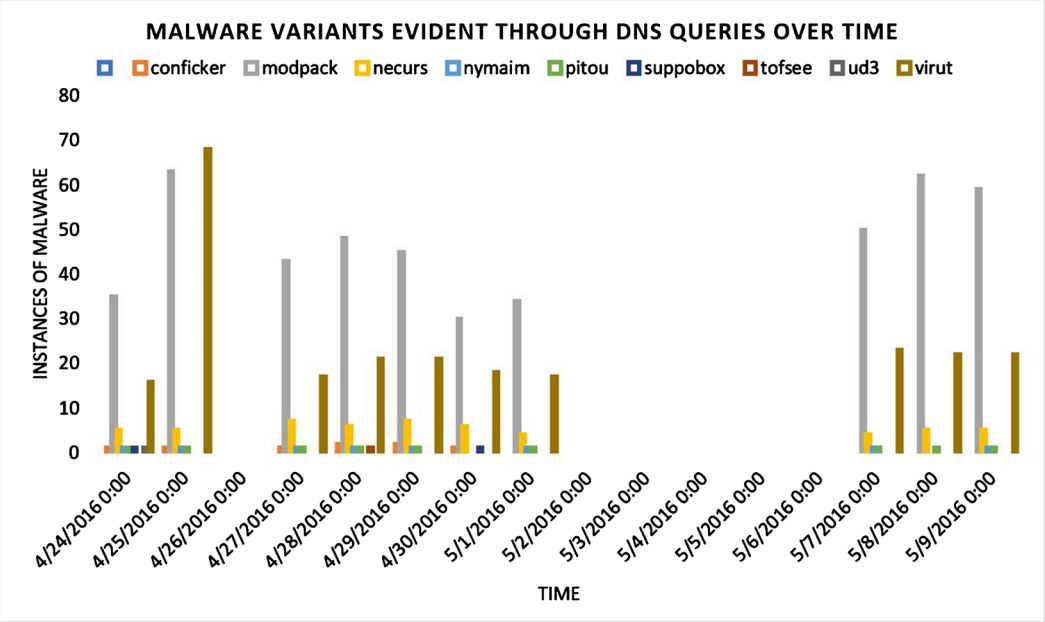

Figure 5. Malware variants evident through DNS queries over time.

Table 2. Counts of malware-related DNS queries by variant.

Day	Date	Conficker	Modpack	Necurs	Nymaim	Pitou	Suppobox	Tofsee	Ud3
1	24/04/2016	10	358	45	3	2	1	0	1
2	25/04/2016	9	204	21	1	9	0	0	0
3	27/04/2016	5	399	71	3	4	0	0	0
4	28/04/2016	7	494	72	3	10	0	1	0
5	29/04/2016	3	459	66	2	10	0	0	0
6	30/04/2016	3	399	43	0	0	1	0	0
7	1/5/2016	0	389	37	3	6	0	0	0
8	7/5/2016	0	564	32	1	4	0	0	0
9	8/5/2016	0	916	75	0	10	0	0	0
10	9/5/2016	0	612	61	1	6	0	0	0

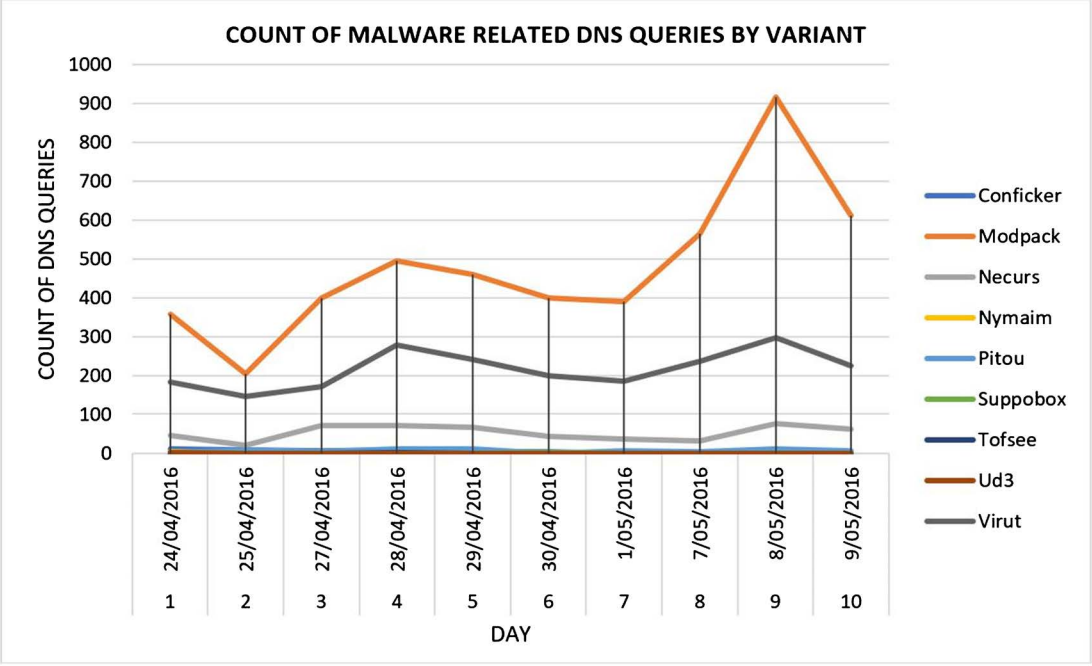

Figure 6. Graph count of malware-related DNS queries by variant.

Suppobox, Tofsee and UD3 could have potentially been defeated and have no more occurrences. This cannot be determined with certainty due to the limited dataset. Modpack is seen to increase over time to a peak on day 9, then start to reduce on day 10. Necurs, Virut and Pitou are seen to remain reasonably constant over time, while Nymaim starts strong and decreases slightly over time.

4.6. Epidemic Parameters

The DNS dataset contains data features that map directly to epidemiological dynamics, as illustrated in Table 3.

Table 3. Epidemiological dynamic and relevant DNS data features.

Epidemiological Dynamic	Relevant DNS Data Features
Population	Number of hosts within the network.
Number of initial infections	Count of the initial (first, by date and time) infected host (s) on the network.
Transmission Time: Length of incubation period	Measured in days, the incubation period is from when a host is initially compromised (e.g. malware downloaded), to when the malware commences the action.
Transmission Time: Duration Host is Infectious	Time (days) in which the malware is undertaking actions, and/or has the potential to undertake actions to infect other hosts.
Fatality Rate	Percentage of hosts that have been irreversibly compromised or unable to perform its function and/or damaged beyond repair.
Time from end or incubation to death	Time (days) from when the malware commences an action, to the host becoming irreversibly compromised ("bricked").
Recovery time	Time (days) from when the malware commences an action, to host recovery back to normal operations and a status of "not infected".
Length of hospital stay	Time (days) in which the malware is undertaking actions.
Hospitalisation rate	Rate (count as a percentage) of hosts that demonstrate indicators of compromise (IOCs) that are confirmed infected for some time where the malware is undertaking action.
Time to hospitalisation	Time (days) from hosts demonstrate indicators of compromise (IOCs) to when they are confirmed infected due to evidence of malware action.

5. Conclusions and Future Work

This paper has discussed the applications of epidemiology to cybersecurity. It has explored how epidemiological principles can be applied to estimate the probability and spread dynamics of infection based on a comprehensive range of factors. These factors comprise both human and machine behaviour and characteristics profiled as risk factors. This paper has demonstrated a preliminary analysis of the relationships between these risk factors utilising DNS data. There is a demonstrated research gap in the aggregation and analysis of both human and machine risk factors over time. This research provides a meaningful contribution to the profiling of these risk factors by presenting a taxonomy underpinned by epidemiological principles.

This research will manifest into a range of considerable contributions to cybersecurity. Further research is underway to utilise Artificial Intelligence and Machine Learning models to monitor and automate the analysis of these risk factors in aggregate. Further research is also underway to utilise DNS data features, related to epidemic dynamics, and apply these to epidemiological models to analyse the spread patterns of different malware variants including the reproduction number. This work will be also extended in developing the epidemiology principle as a risk assessment model for enhancing the performances of machine learning-based intrusion detection systems.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] SAP, the Internet of Things Definition, 2020. https://www.sap.com/australia/trends/internet-of-things.html

[2] Nord, J., Koohang, A. and Palisskiewicz, J. (2019) The Internet of Things: Review and Theoretical Framework. *Expert Systems with Applications*, **133**, 97-108. https://doi.org/10.1016/j.eswa.2019.05.014

[3] Koroniotis, N., Moustafa, N. and Sitnikova, E. (2020) A New Network Forensic Framework Based on Deep Learning for Internet of Things Networks: A Particle Deep Framework. *Future Generation Computer Systems*, **110**, 91-106. https://doi.org/10.1016/j.future.2020.03.042

[4] Marsden, T., Moustafa, N., Sitnikova, E. and Creech, G. (2017) Probability Risk Identification Based Intrusion Detection System for SCADA Systems. In: *International Conference on Mobile Networks and Management*, Springer, Berlin, 353-363. https://doi.org/10.1007/978-3-319-90775-8_28

[5] Gross, G. (2020) Intrusion Detection Techniques Methods and Best Practices. AT&T Business. https://cybersecurity.att.com/blogs/security-essentials/intrusion-detection-techniques-methods-best-practices

[6] Alazab, A., Abawajy, J., Hobbs, M., Layton, R. and Khraisat, A. (2013) Crime Toolkits: The Productisation of Cybercrime. 2013 12*th IEEE International Conference on Trust*, *Security and Privacy in Computing and Communications*, Melbourne, 16-18 July 2013, 1626-1632. https://doi.org/10.1109/TrustCom.2013.273

[7] Biggio, B. and Roli, F. (2018) Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning. *Pattern Recognition*, **84**, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

[8] Centres for Disease Control and Prevention (2020) Principles of Epidemiology in Public Health Practice, Third Edition, An Introduction to Applied Epidemiology and Biostatistics, Definition of Epidemiology.

[9] Rothman, K.J. (2012) Epidemiology: An Introduction. Oxford University Press, Oxford.

[10] Brantly, A. (2017) Public Health and Epidemiological Approaches to National Cybersecurity: A Baseline Comparison. In: *US National Cybersecurity*: *International Politics*, *Concepts and Organization*, Routledge, London, Chapter 7. https://doi.org/10.4324/9781315225623-7

[11] LaMorte, W. (2019) The Evolution of Epidemiologic Thinking. Boston University School of Public Health, Boston.

[12] Heinl, C.H. (2016) The Potential Military Impact of Emerging Technologies in the Asia-Pacific Region: A Focus on Cyber Capabilities. In: Bitzinger, R.A., Ed., *Emerging Critical Technologies and Security in the Asia-Pacific*, Palgrave Macmillan, Hampshire, 123-137. https://doi.org/10.1057/9781137461285_10

[13] Sun, C.A., Hahn, A. and Liu, C.C. (2018) Cyber Security of a Power Grid: State-of-the-Art. *International Journal of Electrical Power and Energy Systems*, **99**, 45-56. https://doi.org/10.1016/j.ijepes.2017.12.020

[14] Li, J., Ma, X., Luo, X., Zhang, J., Li, W., *et al.* (2018) Can We Learn What People Are Doing from Raw DNS Queries? *IEEE INFOCOM 2018 IEEE Conference on Computer Communications*, Honolulu, 15-19 April 2018, 2240-2248. https://doi.org/10.1109/INFOCOM.2018.8486210

[15] Cohen, F. (1987) Computer Viruses: Theory and Experiments. *Computers & Security*, **6**, 22-35. https://doi.org/10.1016/0167-4048(87)90122-2

[16] Panza, M., Madariaga, D. and Bustos-Jiménez, J. (2019) Revealing User Behavior by Analyzing DNS Traffic. In: *International Conference on Machine Learning for Networking*, Springer, Berlin, 212-226. https://doi.org/10.1007/978-3-030-45778-5_14

[17] Mahjoub, D. and Mathew, T.M. (2019) Domain Classification Based on Domain Name System (DNS) Traffic. US Patent 10,185,761.

[18] Goyal, S., Jabbari, S., Kearns, M., Khanna, S. and Morgenstern, J. (2016) Strategic Network Formation with Attack and Immunization. *International Conference on Web and Internet Economics*, Montréal, 11-14 December 2016, 429-443. https://doi.org/10.1007/978-3-662-54110-4_30

[19] Novick, L.F. (2005) Epidemiologic Approaches to Disasters: Reducing Our Vulnerability. *American Journal of Epidemiology*, **162**, 1-2. https://doi.org/10.1093/aje/kwi164

[20] White, T., Pagurek, B. and Bieszczad, A. (1999) Network Modeling for Management Applications Using Intelligent Mobile Agents. *Journal of Network and Systems Management*, **7**, 295-321. https://doi.org/10.1023/A:1018723428983

[21] Goel, S. and Bush, S.F. (2004) Biological Models of Security for Virus Propagation in Computer Networks. *Login*, **29**, 49-56.

[22] Yang, X. and Yang, L.X. (2012) Towards the Epidemiological Modeling of Computer Viruses. *Discrete Dynamics in Nature and Society*, **2012**, Article ID: 259671. https://doi.org/10.1155/2012/259671

[23] Ramachandran, K. and Sikdar, B. (2006) Modeling Malware Propagation in Gnutella Type Peer-to-Peer Networks. *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, Rhodes Island, 25-29 April 2006, 8 p. https://doi.org/10.1109/IPDPS.2006.1639704

[24] Zhou, L., Zhang, L., McSherry, F., Immorlica, N., Costa, M. and Chien, S. (2005) A First Look at Peer-to-Peer Worms: Threats and Defenses. In: *International Workshop on Peer-to-Peer Systems*, Springer, Berlin, 24-35. https://doi.org/10.1007/11558989_3

[25] Feng, C., Qin, Z., Cuthbet, L. and Tokarchuk, L. (2008) Propagation Model of Active Worms in P2P Networks. *The 9th International Conference for Young Computer Scientists*, Hunan, 18-21 November 2008, 1908-1912. https://doi.org/10.1109/ICYCS.2008.237

[26] Jerkins, J.A. and Stupiansky, J. (2018) Mitigating IoT Insecurity with Inoculation Epidemics. *Proceedings of the ACMSE* 2018 *Conference*, Association for Computing Machinery, New York, 29 March 2018, 1-6. https://doi.org/10.1145/3190645.3190678

[27] Hu, H., Myers, S., Colizza, V. and Vespignani, A. (2009) WiFi Networks and Malware Epidemiology. *Proceedings of the National Academy of Sciences*, **106**, 1318-1323. https://doi.org/10.1073/pnas.0811973106

[28] Gil, S., Kott, A. and Barabási, A.L. (2014) A Genetic Epidemiology Approach to Cyber-Security. *Scientific Reports*, **4**, Article No. 5659. https://doi.org/10.1038/srep05659

[29] Camp, L.J., Grobler, M., Jang-Jaccard, J., Probst, C., Renaud, K. and Watters, P. (2019) Measuring Human Resilience in the Face of the Global Epidemiology of Cyber Attacks. *Proceedings of the* 52*nd Hawaii International Conference on System Sciences*, MAUI United States, 8 January 2019, 4763-4772. https://doi.org/10.24251/HICSS.2019.574

[30] Veksler, V.D., Buchler, N., Hoffman, B.E., Cassenti, D.N., Sample, C. and Sugrim, S. (2018) Simulations in Cyber-Security: A Review of Cognitive Modeling of Network Attackers, Defenders, and Users. *Frontiers in Psychology*, **9**, 691. https://doi.org/10.3389/fpsyg.2018.00691

[31] IBM (2020) Submission to: Australia's Cyber Security Strategy 2020.

[32] Singh, M., Singh, M. and Kaur, S. (2019) Detecting Bot-Infected Machines Using DNS Fingerprinting. *Digital Investigation*, **28**, 14-33. https://doi.org/10.1016/j.diin.2018.12.005

[33] Shahriar, H. and Zulkernine, M. (2012) Mitigating Program Security Vulnerabilities: Approaches and Challenges. *ACM Computing Surveys*, **44**, 1-46. https://doi.org/10.1145/2187671.2187673

[34] Antonakakis, M., Perdisci, R., Lee, W., Vasiloglou, N. and Dagon, D. (2011) Detecting Malware Domains at the Upper DNS Hierarchy. *USENIX Security Symposium*, **11**, 1-16.

[35] Liu, Z., Zeng, Y., Zhang, P., Xue, J., Zhang, J. and Liu, J. (2018) An Imbalanced Malicious Domains Detection Method Based on Passive DNS Traffic Analysis. *Security and Communication Networks*, **2018**, Article ID: 6510381. https://doi.org/10.1155/2018/6510381

[36] Zhauniarovich, Y., Khalil, I., Yu, T. and Dacier, M. (2018) A Survey on Malicious Domains Detection through DNS Data Analysis. *ACM Computing Surveys*, **51**, 1-36. https://doi.org/10.1145/3191329

[37] Ivenson, S. (2020) IP Time to Live (TTL) and Hop Limit Basics-Packet Pushers August 2020.