Institution: Imperial College London

Unit of Assessment: 10 – Mathematical Sciences

Title of case study: B10-7 Improving global cyber-security through statistical methods

Period when the underpinning research was undertaken: 2010-2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Professor Nick Heard	NH: Chair in Statistics	NH: 2004-present
Professor Niall Adams	NA: Professor of Statistics	NA: 2002-present

Period when the claimed impact occurred: 1 August 2013- 31 December 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

Threats to cyber-security are a global problem, affecting business, government and society. Researchers at Imperial have developed statistical methods and data science techniques which have helped strengthen cyber-security across these domains to safeguard the privacy, wealth and wellbeing of all members of society.

These methods and techniques have been integrated into: 1) Microsoft Defender Antivirus, part of the Microsoft Windows operating system that detects 5 billion cyber threats on devices worldwide every month, and 2) PathScan and Credential Analytics, security and malware offerings from Ernst & Young Global Limited (EY), which are licenced to businesses worldwide. Microsoft enterprise security products are used by “90% of the Fortune 500” companies and one billion Windows 10 users. The EY security platform has “clients across the globe”. This impact benefits Microsoft and EY commercially and hundreds of millions of users worldwide benefit from improved security and safety.

2. Underpinning research

Research and development from the Statistical Cyber-Security Research group of the Department of Mathematics at Imperial, led by Nick Heard and Niall Adams, has driven the innovation of statistical models for network host behaviours and traffic patterns which can be observed within an enterprise computer network.

This ongoing work began in 2010, originally through studying anomaly detection for dynamic graphs, and has evolved into several strands of research for modelling computer networks at different levels of granularity. The microscopic level research has been concerned with building probabilistic models for the traffic passing along an edge between two nodes (hosts) in a computer network **[1,3]**, identifying automated and human traffic, learning temporal behaviour patterns and clustering hosts and users based on similarities in their connectivity patterns **[6]**. At the macroscopic level, research has investigated whole-network analyses using graph theory and spectral methods, concerned with predicting new connections in a network **[2]**; and in changepoint analysis, monitoring traffic for sudden deviations **[4]** like the WannaCry ransomware attack of 2017 **[B]**.

Besides explicit cyber modelling, related research has also branched into more theoretical work on changepoint analysis and meta-analysis. Interest in the former stems from the need to run anomaly detection techniques which can adapt to the heterogeneity and ever-changing nature of a typical enterprise network; adaptive forgetting-factor changepoint analysis techniques originating from Imperial **[4]** have been particularly influential.

Similarly, research on meta-analysis has been driven by empirical understanding that detection of a network intruder requires the combination of weak signals from several statistical analytics or techniques, as no single network event or occurrence can necessarily imply the presence of an attack (otherwise, network protocols will simply be configured to block such actions). For combining information sources to detect cyber-attacks, research from Imperial on selecting optimal methods for combining p-values **[5]** has been pivotal; this contribution is also noted in a letter of support from Microsoft.

The research in cyber-security has been driven by key collaborators in the application domain who have helped to shape the direction of innovation and have led to implementation of methods for contracted projects and commercial exploitation through patents and software licencing as follows.

Collaboration with Ernst and Young Global Ltd (EY) has been conducted via a three-way relationship with Los Alamos National Laboratory, who have a formal collaborative research and development agreement with EY in the US. Student internships each year since 2011 have led to co-creation of patentable technologies.

Direct collaboration with Microsoft has been established through research visits from NH and NA and annual summer internships for Imperial PhD students from the Statistical Cyber-Security Research group. The research directions in anomaly detection and information synthesis at Imperial lead by NH and NA have been strongly influenced by domain understanding established through this collaboration.

The research efforts of NA and NH were substantially supported by five and six-year secondments respectively to the Heilbronn Institute for Mathematical Research **[i, ii]**. Other funding institutions and collaborators include:

- the National Cyber Security Centre,
- the Alan Turing Institute and
- Los Alamos National Laboratory.

3. References to the research

[1] Turcotte, M. J. M., Heard, N. A. and Kent, A. D. (2015) Modelling user behaviour in a network using computer event logs. In Dynamic Networks in Cybersecurity. Imperial College Press, doi:10.1142/9781786340757_0003.
[2] Turcotte, M. J. M., Moore, J., Heard, N. A. and McPhall, A. (2016) Poisson Factorization for Peer-Based Anomaly Detection. In proceedings of IEEE Intelligence and Security Informatics Conference (ISI2016), Cybersecurity and Big Data, doi:10.1109/ISI.2016.7745472.
[3] Turcotte, M. J. M., Heard, N. A. and Neil, J. (2014) Detecting Localised Anomalous Behaviour in a Computer Network. In Advances in Intelligent Data Analysis XIII, 321–332, doi:10.1007/978-3-319-12571-8_28.
[4] Plasse J. and Adams N., (2019) Multiple changepoint detection in categorical data streams, Statistics and Computing, 29, 1109-1125, doi:10.1007/s11222-019-09858-0.
[5] Heard, N. A. and Rubin-Delanchy, P. T. G. (2018) Choosing Between Methods of Combining p-values. Biometrika, 105, 1, 239-246, doi:10.1093/biomet/asx076.
[6] Heard, N. A., Palla, K. and Skoularidou, M. (2016) Topic modelling of authentication events in an enterprise computer network. In proceedings of IEEE Intelligence and Security Informatics Conference (ISI2016), Cybersecurity and Big Data, doi:10.1109/ISI.2016.7745466.

Funding programmes:
[i] Secondment of NA to Heilbronn Institute for Mathematical Research (2011-2016, £380,690)
[ii] Secondment of NH to Heilbronn Institute for Mathematical Research (2013-2019, £385,432)

4. Details of the impact

Imperial College research has been incorporated into Microsoft Defender, which is part of all modern Microsoft Windows versions, and into cyber-analytic software from EY.

Global critical national infrastructures face a growing and evolving threat from cyber-attacks, and our modern networked economy relies heavily on maintaining secure environments for

exchanging, storing, and protecting business and consumer data and intellectual property. There is now a consensus amongst government and industry experts that statistical and machine learning techniques have an important role to play in current and future security defences. Providing statistical cyber-security solutions to industry effectively strengthens the UK and world economies. Cyber-security is therefore an emerging and strategically important field in statistics and the Statistical Cyber-Security Research group at Imperial is at the forefront, working closely with both government and industry, developing methods tailored to the most pressing security problems.

A single cyber breach, depending on the country and industry, has been estimated to cost on average $3.86 million, rising to an average of $8.64 million for the US **[A]**. The WannaCry attack of 2017 indiscriminately targeted Microsoft Windows operating systems that were not patched with respect to a known exploit. In the short period the attack was active, an estimated 230K computers were compromised globally **[B]**. Besides causing financial losses reaching approximately $4bn, WannaCry impacted the computer systems of NHS hospitals and surgeries, disrupting care for those in urgent need.

The Credential Analytics software of EY uses research from papers **[1,2]**, which provide statistical models for detecting the actions of an intruder on an enterprise work using authentication and computer event logs. The EY PathScan analytic uses research from paper **[3]**, concerned with detecting anomalous network traversal.

To enable this, three joint patent applications (two granted **[C, D]**, one pending) have been filed by Imperial in collaboration with Los Alamos National Laboratory, a US research lab funded by the US Department of Energy. Two pieces of statistical software utilising these patents have been licensed by EY, for commercial distribution in their cyber-security platform called PathScan, which has been deployed within large international corporations. Quoting the letter of support from EY **[E]**,

> "*EY have used the research from the Imperial-authored papers and patents listed below to develop two of our offerings in Cyber Analytics: PathScan and Credential Analytics […] In the 4+ years since, we have invested roughly 27 person-years to develop a scalable, sustainable product. We have clients across the globe, and have found particular success in APAC and the Middle East. EY is [also in] active discussions to license the IP associated with Credential Analytics. Again, with our LANL partners, we invested 4 person-years in 2018*".

Microsoft Defender Antivirus claims to deliver "comprehensive, ongoing, and real-time protection against software threats like viruses, malware, and spyware across email, apps, the cloud, and the web" **[F]**. This software is deployed worldwide on all devices running modern versions of the Microsoft Windows operating system, including Windows 10, Windows 7 and Windows Server. To quantify the scale of this provision, it should be noted that there are more than 1 billion devices running Windows 10 **[G]**.

Microsoft have implemented solutions in their security product Microsoft Defender Antivirus using anomaly detection techniques from **[2,4,6]** for respectively modelling behavioural patterns, performing change detection with adaptive forgetting-factors, and latent feature modelling of hosts and users on enterprise networks. Furthermore, the anomaly scores from the range of analytics being run by Microsoft Defender are combined into a single measure of surprise using results in meta-analysis published by the Statistical Cyber-Security Research Group at Imperial **[5]**.

Quoting the letter of support from Microsoft **[H]**,

> "***Various methods described in these papers have been implemented in our products**, including:*
> - *P-value combination, which we use as part of our lateral movement detection, **leading to over 2 million detections of malicious behavior per month**.*
> - *Adaptive forgetting factors, which influence many of our streaming algorithms, and which represent **approximately 20% of our overall detection portfolio**."*

In the above, *"P-value combination"* refers to **[5]** and "Adaptive forgetting factors" refer to **[4]**.

Microsoft's letter continues:

"The scale at which we operate (again, over 6 trillion signals analyzed daily) means that any new production algorithms must be implemented in the face of enormous compute, memory, and storage challenges. The fact that research and development from Imperial has successfully been transitioned into production is a testament to the importance we place on the cutting-edge work coming from the College."

5. Sources to corroborate the impact

[A] IBM Cost of a Data Breach Report (Archived here)
- https://www.ibm.com/security/digital-assets/cost-data-breach-report
- https://www.ibm.com/security/data-breach
- https://digitalguardian.com/blog/what-does-data-breach-cost-2020

[B] Information about the Wannacry Ransom Attack (Archived here)
- https://www.kaspersky.co.uk/resource-center/threats/ransomware-wannacry
- https://www.telegraph.co.uk/technology/2018/10/11/wannacry-cyber-attack-cost-nhs-92m-19000-appointments-cancelled/

[C] EY Credential Analytics patent: M. Turcotte, N.A. Heard and A. Kent. Modelling Behavior In A Network Using Event Logs, 2016. US Provisional Patent filed with Los Alamos National Laboratory (Archived here)

[D] EY PathScan patent: M. Turcotte, N.A. Heard and J.C. Neil., December 12, 2013. WO Patent App. PCT/US2013/031,463. Los Alamos National Security, LLC and Imperial Innovations Limited. (Archived here)

[E] Letter of Support from Partner/Principal, EY Cyber

[F] Information about Microsoft Cyber Security Strategy (Archived here)
- https://www.microsoft.com/security/blog/2019/05/14/executing-vision-microsoft-threat-protection
- https://www.microsoft.com/en-gb/windows/comprehensive-security

[G] Number of Windows 10 devices https://news.microsoft.com/bythenumbers/en/windowsdevices (Archived here)

[H] Letter of Support from Principal Data Scientist Lead, Microsoft Defender Advanced Threat Protection, Microsoft Corp