IIR

Internet
Infrastructure
Review

Dec.2016
Vol. 33

Infrastructure Security

Mirai Botnet Detection and Countermeasures

Content Delivery

Industry Efforts to Unify Streaming Formats

Technology Trends

Wikipedia as a Language Resource

IIJ
Internet Initiative Japan

Internet Infrastructure Review

December 2016 Vol.33

Executive Summary 3

1. Infrastructure Security 4

1.1 Introduction 4

1.2 Incident Summary 4

1.3 Incident Survey 11

1.3.1 DDoS Attacks 11

1.3.2 Malware Activities 13

1.3.3 SQL Injection Attacks 16

1.3.4 Website Alterations 16

1.4 Focused Research 18

1.4.1 Mirai Botnet Detection and Countermeasures 18

1.4.2 Miscellaneous SSL/TLS Topics 24

1.5 Conclusion 29

2. Content Delivery 30

3. Technology Trends 34

3.1 Introduction 34

3.2 Writing “Unix Archaeology" 34

3.3 Drilldown Searches 34

3.4 Named Entity Recognition (NER) 35

3.5 NER in the Natural Language Toolkit (NLTK) 36

3.6 Statistical Natural Language Processing and Corpora 37

3.7 Creating a Named Entity Corpus from Wikipedia 38

3.8 Conclusion 39

Executive Summary

Although the events took place in another country, the election of Donald Trump in the U.S. Presidential Election caused a commotion around the world, and it is difficult to judge whether recent tumult in finance, economics, and foreign policy can be attributed to this being a favorable or unfavorable result. With talks between Prime Minister Shinzo Abe and Russian President Vladimir Putin coming up after that, it became even harder to drag our attention away from international politics. Under these circumstances, Internet traffic is continuing to grow at a steady pace. Online broadcasts related to the U.S. Presidential Election were consumed on a global scale, and factors such as spikes in traffic in the middle of the night in Japan have once again given us a real sense that the Internet is important social infrastructure that supports activities worldwide.

This report discusses the results of the various ongoing surveys and analysis activities that IIJ, as a service provider, carries out to support the Internet and cloud infrastructure, and enable our customers to continue to use them safely and securely. We also regularly present summaries of technological development as well as important technical information.

In Chapter 1, we focus on discussing incidents that occurred between July 1 and September 30. Each year we are extra vigilant in August and September due to dates of particular historical significance falling on these months, but over the past few years they have passed without major incident. There were some attacks related to pseudo-religious activity, and we have reflected on this three month period while also paying attention to historical events and political and social situations. We also took a look at the Mirai botnet malware. So-called IoT devices lack significant processing power on their own, but they are prone to malware infections when their security is not properly managed, and large numbers of these devices can form massive networks that are used to launch attacks. We have covered this topic here in the belief that care must be taken with regard to Internet infrastructure going forward.

In Chapter 2, we examine efforts in the Internet content distribution industry to unify streaming formats, and discuss measures for resolving issues with the methods currently used. Live streaming results in a slight delay compared to real time. Many people in the industry seek to stream live events in a form as close to real time as possible, and we expect initiatives to bridge this gap will gather momentum. We also believe more work will be done on standardization, including technology that enables the offline viewing of videos by downloading them to a mobile device when it is connected to Wi-Fi, as the use of offline playback is not progressing due to issues associated with the data bandwidth restrictions on mobile plans. We have provided an in-depth look at the inner workings of the streaming business, which is expected to see further development in the future.

In Chapter 3, we performed a scientific study of Wikipedia. It may be a hard concept to grasp, but a researcher studying Wikipedia discusses how they have been tackling the computerization of historical fact drilldown on a daily basis. They detail the process by which they came to the conclusion that the basic technologies built for analyzing Wikipedia could be applied to this after looking into the field of natural language research. In a sense, from a perspective similar to the analysis of big data and deep learning, this may serve as a method for uncovering historical facts from within the information stored on Wikipedia.

Through activities such as these, IIJ continues to strive towards improving and developing our services on a daily basis, while maintaining the stability of the Internet. We will keep providing a variety of services and solutions that our customers can take full advantage of as infrastructure for their corporate activities.

Yoshikazu Yamai
Mr. Yamai is an Executive Managing Officer of IIJ and Director of the Service Infrastructure Division.
Upon joining IIJ in June 1999, he was temporarily transferred to Crosswave Communications, Inc., where he was engaged in WDM and SONET network construction, wide-area LAN service planning, and data center construction, before returning to his post in June 2004. After his return he was in charge of IIJ's Service Operation Division. From April 2016 he joined the Infrastructure Operation Division, and now oversees the overall operation of corporate IT services at IIJ. He also heads IIJ's data center operations, and he played a key role in the establishment of the modular "Matsue Data Center Park," which was the first in Japan to use outside-air cooling.

Mirai Botnet Detection and Countermeasures

1.1 Introduction

This report is a summary of incidents that IIJ responded to, based on information obtained by IIJ for the purpose of operating a stable Internet, information obtained from observed incidents, information obtained through our services, and information obtained from companies and organizations that IIJ has cooperative relationships with. This volume covers the period of time from July 1 through September 30, 2016. In this period, a number of hacktivism-based attacks were once again carried out by Anonymous and other groups, and there were frequent incidents that included many DDoS attacks, information leaks caused by unauthorized access, and website defacements. There were also DDoS attacks of unprecedented scale carried out using botnets comprised of IoT devices infected with malware. As shown here, many security-related incidents continue to occur across the Internet.

1.2 Incident Summary

Here, we discuss the IIJ handling and response to incidents that occurred between July 1 and September 30, 2016. Figure 1 shows the distribution of incidents handled during this period*1.

■ Activities of Anonymous and Other Hacktivist Groups

Attack activities by hacktivists such as Anonymous continued during this period. In correspondence with various events and assertions, DDoS attacks and information leaks targeted various companies and government-related sites.

As a protest against the drive hunting of dolphins and small whales in Japan, there have been intermittent DDoS attacks since 2013 which are believed to be performed by Anonymous. A statement that this attack campaign would continue was made as the fishing season opened on September 1 (OpKillingBay/OpWhales/OpSeaWorld). Although some of the attack targets have differed, by and large the targets have remained the same since last year. From September a large number of DoS attacks were made against these targets, but as before there were also many attacks against websites not on the attack target list. Consequently, during the month of September, over 30 DoS attacks against various sites in Japan were observed. The attacks had not subsided by October, and the number of participants in the attack campaign is thought to be increasing, so continued vigilance is required.

Between late August and early September, a series of simultaneous DoS attacks were made against multiple sites in Japan, causing connection problems for services on many sites, among other issues. There are many unknown factors regarding why these sites were targeted and the goals of the attacker, so there is currently no clear picture of things, but multiple message board sites associated with the Koshinkyo pseudo-religious group were among the targets, so we believe this may have something to do with the attacker's intentions.

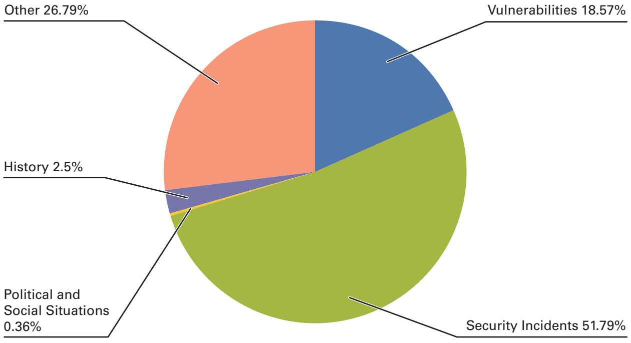


Figure 1: Incident Ratio by Category (July 1 to September 30, 2016)

Because September 18 is the day the Liutiaohu Incident that triggered the Manchurian Incident took place, the few days before and after this day is well known as a "historically significant date" upon which cyber attacks between Japan and China are more likely to occur. In particular, there were large-scale demonstrations such as the Chinese fishing boat collision incident that occurred near the Senkaku Islands in 2010, as well as the landing of activists on the Senkaku Islands and demonstrations in response to the nationalization of

[1] Incidents in this report are split into five categories: vulnerabilities, political and social situations, history, security incidents or other.
Vulnerabilities: Responses to vulnerabilities in network equipment, server equipment or software commonly used across the Internet or in user environments.
Political and Social Situations: Responses to attacks stemming from international conferences attended by VIPs and international conflicts, and other related domestic and foreign circumstances and international events.
History: Warnings/alarms, detection and response to incidents for attacks that occur on the day of a historically significant date that have a close connection to a past event.
Security Incidents: Unexpected incidents and related responses such as wide spreading of network worms and other malware; DDoS attacks against certain websites.
Other: Security-related information, and incidents not directly associated with security problems, including high traffic volume associated with a notable event.

the Senkaku Islands by the Japanese Government in 2012. During the same periods, a range of cyber attacks were made against many sites in Japan. Since 2013, however, attack activity during these periods has settled down, and again this year no particularly noteworthy attack activity was observed. Some cyber attacks are linked to real-life historical events and carry a historical context, so it is necessary to pay attention to political and social situations such as historically significant dates and current international affairs.

■ Vulnerabilities and Responses

During this period many fixes were released for Microsoft's Windows*[2]*[3]*[4]*[5]*[6]*[7], Internet Explorer*[8]*[9]*[10], Edge*[11]*[12]*[13], and Office*[14]*[15]*[16]. Updates were also released for Adobe Systems' Flash Player, Acrobat, and Reader. A quarterly update was provided for Oracle's Java SE, fixing many vulnerabilities. Several of these vulnerabilities were exploited in the wild before patches were released.

In server applications, a quarterly update was released by Oracle, fixing many vulnerabilities in the Oracle database server and many other Oracle products. A vulnerability that could allow remote privilege escalation or arbitrary code execution through SQL injection attacks was also discovered in Oracle's MySQL database server, and fixes were provided for various Linux distributions. Vulnerabilities were also discovered and fixed in the BIND9 DNS server, including those that could allow DoS attacks by an external party by exploiting issues in the processing of lightweight resolver queries, and the process that creates DNS responses. In SSL/TLS implementations, an attack method (SWEET32) was made public. It targets 64-bit block ciphers such as 3DES (or TripleDES), and recovers plaintext by finding collisions in ciphertext after intercepting a large amount of data encrypted with the same key. In response, the use of 3DES was restricted under default settings in implementations such as OpenSSL.

Researchers also demonstrated the possibility of a side channel attack that could allow off-path attackers to hijack TCP sessions even when unable to intercept the TCP communications, due to a vulnerability in the implementation of TCP in the Linux kernel. This vulnerability was patched in the latest Linux kernel*[17].

*2 "Microsoft Security Bulletin MS16-086 - Critical: Cumulative Security Update for JScript and VBScript (3169996)" (https://technet.microsoft.com/en-us/library/security/MS16-086).

*3 "Microsoft Security Bulletin MS16-087 - Critical: Security Update for Windows Print Spooler Components (3170005)" (https://technet.microsoft.com/en-us/library/security/MS16-087).

*4 "Microsoft Security Bulletin MS16-097 - Critical: Security Update for Microsoft Graphics Component (3177393)" (https://technet.microsoft.com/en-us/library/security/MS16-097).

*5 "Microsoft Security Bulletin MS16-102 - Critical: Security Update for Microsoft Windows PDF Library (3182248)" (https://technet.microsoft.com/en-us/library/security/MS16-102).

*6 "Microsoft Security Bulletin MS16-106 - Critical: Security Update for Microsoft Graphics Component (3185848)" (https://technet.microsoft.com/en-us/library/security/MS16-106).

*7 "Microsoft Security Bulletin MS16-116 - Critical: Security Update in OLE Automation for VBScript Scripting Engine (3188724)" (https://technet.microsoft.com/en-us/library/security/MS16-116).

*8 "Microsoft Security Bulletin MS16-084 - Critical: Cumulative Security Update for Internet Explorer (3169991)" (https://technet.microsoft.com/en-us/library/security/MS16-084).

*9 "Microsoft Security Bulletin MS16-095 - Critical: Cumulative Security Update for Internet Explorer (3177356)" (https://technet.microsoft.com/en-us/library/security/MS16-095).

*10 "Microsoft Security Bulletin MS16-104 - Critical: Cumulative Security Update for Internet Explorer (3183038)" (https://technet.microsoft.com/en-us/library/security/MS16-104).

*11 "Microsoft Security Bulletin MS16-085 - Critical: Cumulative Security Update for Microsoft Edge (3169999)" (https://technet.microsoft.com/en-us/library/security/MS16-085).

*12 "Microsoft Security Bulletin MS16-096 - Critical: Cumulative Security Update for Microsoft Edge (3177358)" (https://technet.microsoft.com/en-us/library/security/MS16-096).

*13 "Microsoft Security Bulletin MS16-096 - Critical: Cumulative Security Update for Microsoft Edge (3183043)" (https://technet.microsoft.com/en-us/library/security/MS16-096).

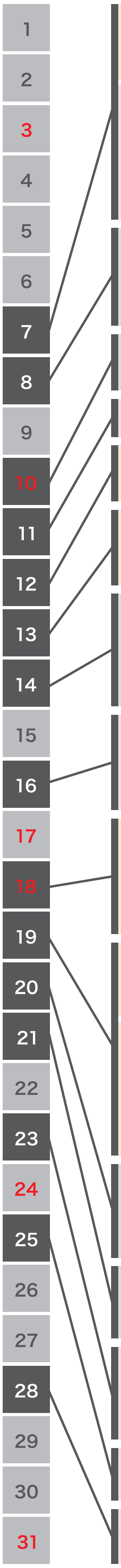
*14 "Microsoft Security Bulletin MS16-088 - Critical: Security Update for Microsoft Office (3170008)" (https://technet.microsoft.com/en-us/library/security/MS16-088).

*15 "Microsoft Security Bulletin MS16-099 - Critical: Security Update for Microsoft Office (3177451)" (https://technet.microsoft.com/en-us/library/security/MS16-099).

*16 "Microsoft Security Bulletin MS16-107 - Critical: Security Update for Microsoft Office (3185852)" (https://technet.microsoft.com/en-us/library/security/MS16-107).

*17 Yue Cao, et al., 25th USENIX Security Symposium, "Off-Path TCP Exploits: Global Rate Limit Considered Dangerous" (https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/cao).

July Incidents



V **7th: Multiple vulnerabilities in Adobe Acrobat and Reader that could allow illegal termination and arbitrary code execution were discovered and fixed.**
"Security Updates Available for Adobe Acrobat and Reader" (https://helpx.adobe.com/security/products/acrobat/apsb16-26.html).

S **7th: Unauthorized access to the website of TokyoZooNet led to defacement, and 21,688 email addresses registered to its mailing list leaked.**
"Unauthorized Access on the Metropolitan Zoo and Aquarium Website | Tokyo Metropolis" (http://www.metro.tokyo.jp/INET/OSHIRASE/2016/07/20q78400.htm) (in Japanese). "Unauthorized Access on the Metropolitan Zoo and Aquarium Website (follow-up) | Tokyo Metropolis" (http://www.metro.tokyo.jp/INET/OSHIRASE/2016/07/20q78600.htm) (in Japanese). "Notice of Metropolitan Zoo and Aquarium Official Website 'TokyoZooNet' Restoration and Apology | TokyoZooNet" (http://www.tokyo-zoo.net/topic/topics_detail?kind=&inst=&link_num=23843) (in Japanese).

O **8th: The Japan Tourism Agency held the 1st "Committee on Information Leakage in the Tourism Industry." It reviewed the report that summarized the string of information leaks that occurred in the tourism industry, and laid out the issues.**
"1st 'Committee on Information Leakage in Tourism Industry' to be held | 2016 | Press Release | News/Interviews | Japan Tourism Agency" (http://www.mlit.go.jp/kankocho/news06_000283.html) (in Japanese).

O **10th: Bitcoin reached 420,000 blocks, and the reward for miners was halved from 25 BTC to 12.5 BTC, but there was no significant change in its valuation.**

S **11th: The Twitter accounts of Twitter Co-founder Jack Dorsey and Yahoo! CEO Marissa Mayer were hijacked by the OurMine team.**

V **12th: Multiple vulnerabilities in Adobe Flash Player that could allow illegal termination or arbitrary code execution were discovered and fixed.**
"Security updates available for Adobe Flash Player" (https://helpx.adobe.com/security/products/flash-player/apsb16-25.html).

V **13th: Microsoft published their Security Bulletin Summary for July 2016, and released a total of eleven updates, including six critical updates such as MS16-084, as well as five important updates.**
"Microsoft Security Bulletin Summary for July 2016" (https://technet.microsoft.com/library/security/ms16-jul).

O **14th: The National Police Agency announced they had begun providing information to the international organization APWG (Anti-Phishing Working Group) to deter fraud and other damages caused by fake sites set up on overseas servers. Companies such as Web browser vendors participate in APWG.**
National Police Agency, "Regarding the supply of information on overseas fraudulent sites to the APWG" (http://www.npa.go.jp/cyber/pdf/APWG.pdf) (in Japanese).

S **16th: User information (user names, email addresses, and IP addresses) for around 2 million individuals leaked from the Ubuntu Forums website. There was an SQL injection vulnerability in the Forumrunner plug-in for vBulletin, which was exploited in this incident.**
"Notice of Ubuntu Forums breach; user passwords not compromised | Ubuntu Insights" (https://insights.ubuntu.com/2016/07/15/notice-of-security-breach-on-ubuntu-forums/).

V **18th: Apple released iOS 9.3.3, OS X El Capitan 10.11.6, and Security Update 2016-004, fixing multiple vulnerabilities, including those that could allow a remote attacker to execute arbitrary code. Also, tvOS 9.2.2 and watchOS 2.2.2 were released.**
"About the security content of iOS 9.3.3" (https://support.apple.com/en-us/HT206902). "About the security content of OS X El Capitan v10.11.6 and Security Update 2016-004" (https://support.apple.com/en-us/HT206903). "About the security content of tvOS 9.2.2" (https://support.apple.com/en-us/HT206905). "About the security content of watchOS 2.2.2" (https://support.apple.com/en-us/HT206904).

V **19th: Oracle released their quarterly scheduled update for multiple products including Java SE and Oracle Database Server, fixing a total of 276 vulnerabilities.**
"Oracle Critical Patch Update Advisory - July 2016" (http://www.oracle.com/technetwork/security-advisory/cpujul2016-2881720.html).

V **19th: The httpoxy vulnerability that had a wide-ranging impact on Web servers that use CGI (Common Gateway Interface) was disclosed. The National Police Agency issued an alert due to observations of access thought to target this vulnerability on its fixed-point observation systems.**
httpoxy, "A CGI application vulnerability for PHP, Go, Python and others" (https://httpoxy.org/). National Police Agency, "Regarding observations of access targeting a vulnerability (httpoxy) in Web servers that use CGI., etc." (http://www.npa.go.jp/cyberpolice/detect/pdf/20160720.pdf) (in Japanese).

O **20th: The "Pokémon GO" game for smartphones that uses location information was released in Japan, and taking into account the situation in countries such as the United States where it was released earlier, the National center of Incident readiness and Strategy for Cybersecurity issued an alert.**
"Alert regarding the 'Pokémon GO' location information game" (http://www.nisc.go.jp/active/kihon/pdf/reminder_20160721.pdf) (in Japanese).

O **21st: A hard fork was implemented by Ethereum to recover the ETH withdrawn fraudulently from The DAO virtual currency investment fund in June.**
"Hard Fork Completed - Ethereum Blog" (https://blog.ethereum.org/2016/07/20/hard-fork-completed/).

S **23rd: WikiLeaks published around 20,000 U.S. Democratic National Committee (DNC) internal emails on its website. Due to this, multiple executives including the Chairperson of the DNC resigned. Guccifer 2.0, the alias of a person thought to have compromised the DNC, also claimed to have provided the published email data himself.**
"WikiLeaks - Search the DNC email database" (https://wikileaks.org/dnc-emails/).

S **25th: Unauthorized access by an external party at the South Korean ticket reservation site INTERPARK led to the leak of personal information for approximately 10.3 million individuals.**

S **28th: Reuters reported on suspected unauthorized access by an external party at the U.S. Democratic Congressional Campaign Committee (DCCC) and stated that the FBI were investigating. The DCCC later admitted it had been the target of a cyber attack.**

*Dates are in Japan Standard Time

Legend					
	V Vulnerabilities	S Security Incidents	P Political and Social Situation	H History	O Other

The Largest-Scale DDoS Attacks in History

During this survey period a series of DDoS attacks of unprecedented scale were observed. The “Krebs on Security” blog run by Brian Krebs, a notable expert in the security industry, was hit by a 140 Gbps DoS attack after a series of articles about Israeli vDOS service were published there on September 9*[18] and September 10*[19]. vDOS is a DDoS-for-hire service (a service that conducts DDoS attacks on behalf of others), which is also known as a booter/stresser. It is thought that Krebs was targeted by DoS attacks as revenge for him investigating and exposing the internal affairs of this service, and due to the arrest of the two 18-year-old youths who had been running the vDOS service in Israel. The attacks subsequently became more severe, and it is said the blog was targeted in a large-scale attack of around 620 Gbps after another booter-related article*[20] was published on September 20*[21]. Prolexic (Akamai) had provided a DDoS protection service to Krebs on Security for free over the previous four years, but this service was withdrawn because of concern that attacks of such a large scale may have a significant negative impact on other paying customers. As a result, Krebs on Security could not be accessed temporarily. The blog would later begin receiving protection services from Google Project Shield*[22], and it was restored on September 25. An article Mr. Krebs published on September 20 revealed that the attack source was likely a botnet of IoT devices infected with malware designed to target devices such as routers, IP cameras, and digital video recorders. There have also been multiple reports from security vendors and other parties regarding DDoS attacks from botnets of IoT devices in the past*[23]. According to the observations of Arbor, a DDoS attack of around 540 Gbps that occurred during the Rio 2016 Olympics was also made using an IoT device botnet. It was reported that from a few months before the Olympic Games began communications targeting Telnet (23/TCP) rose dramatically, and that malware infections on IoT devices increased*[24].

The French hosting service OVH was also targeted by similar DDoS attacks at almost the same time as the attacks against Krebs on Security, with traffic in excess of 1 Tbps observed at peak times*[25]. The source code of Mirai, one of the malware presumed to have been at the root of these attacks, was published on an Internet message board site in October. See “1.4.1 Mirai Botnet Detection and Countermeasures” for more information about Mirai.

Many issues have been identified in IoT devices, including administrator ports such as Telnet being left open, default passwords being used without them being changed, and unchangeable administrator passwords being set. Because of this, an increasing number of vulnerable IoT devices such as these have been the target of malware infections in recent years, and these are currently being exploited and used as part of attack infrastructure.

Mass Password Information Leaks

During this period a string of mass password information leaks that occurred in the past at a variety of Internet-based services were once again disclosed. In each case the number of accounts that had leaked was extremely large, with around 25 million from Mail.ru, 68 million from Dropbox, 43 million from Last.fm, 98 million from Rambler.ru, 33 million from Qip.ru, and 500 million from Yahoo!. The leak from Yahoo! in particular was the largest ever from a single service provider. In some cases the data leaked included plaintext password information and unsalted MD5-hashed passwords that would be relatively easy to crack, so there was a high risk of immediate exploitation. Apart from a few exceptions, these leaks are thought to be based on data acquired by a Russian hacker group between 2011 and 2014, as with the leaks that had already been confirmed at services such as MySpace and LinkedIn. The information leaks were discovered when the data became available for general purchase on Russian message board sites and Dark Web marketplaces this year.

*[18] “Israeli Online Attack Service ‘vDOS’ Earned $600,000 in Two Years — Krebs on Security” (http://krebsonsecurity.com/2016/09/israeli-online-attack-service-vdos-earned-600000-in-two-years/).

*[19] “Alleged vDOS Proprietors Arrested in Israel — Krebs on Security” (http://krebsonsecurity.com/2016/09/alleged-vdos-proprietors-arrested-in-israel/).

*[20] “DDoS Mitigation Firm Has History of Hijacks — Krebs on Security” (http://krebsonsecurity.com/2016/09/ddos-mitigation-firm-has-history-of-hijacks/).

*[21] “KrebsOnSecurity Hit With Record DoS — Krebs on Security” (https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/).

*[22] “Google | Project Shield | Free DDoS protection” (https://projectshield.withgoogle.com/public/).

*[23] For example, Flashpoint and Level 3 reported on the BASHLITE botnet in blog posts at the end of August. “Attack of Things! - Beyond Bandwidth” (http://blog.level3.com/security/attack-of-things/).

*[24] “Rio Olympics Take the Gold for 540gb/sec Sustained DDoS Attacks!” (https://www.arbornetworks.com/blog/asert/rio-olympics-take-gold-540gbsec-sustained-ddos-attacks/).

*[25] “OVH News - The DDoS that didn’t break the camel’s VAC” (https://www.ovh.com/us/news/articles/a2367.the-ddos-that-didnt-break-the-camels-vac).

August Incidents

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

O **2nd: The National center of Incident readiness and Strategy for Cybersecurity (NISC) held the 3rd meeting of the Security-Minded Corporate Management Working Group, where the "Approach to Cyber Security for Corporate Management" the working group had written was published.**
"Security-Minded Corporate Management Working Group" (http://www.nisc.go.jp/conference/cs/jinzai/wg/index.html) (in Japanese).

S **3rd: A theft of bitcoin took place at the Bitfinex bitcoin exchange in Hong Kong, and transactions were temporarily suspended. The losses amounted to 120,000 BTC (equivalent to about 6 billion yen), and the price of bitcoin dropped about 10 percent as a result.**
"Security Breach - Bitfinex blog" (http://blog.bitfinex.com/announcements/security-breach/).

S **5th: Anonymous carried out protests in accordance with the Olympic Games held in Rio de Janeiro (OpOlympics/OpOlympicHacking/OpR10). During the Games, DoS attacks were made against websites of the local state of Rio de Janeiro.**

V **8th: Check Point Software Technologies announced there were four vulnerabilities named "QuadRooter" in the Qualcomm chipset used in many Android devices. These vulnerabilities were fixed along with others in the "Android Security Bulletin—September" (https://source.android.com/security/bulletin/2016-09-01.html) released on September 6.**
"QuadRooter: New Android Vulnerabilities in Over 900 Million Devices | Check Point Blog" (http://blog.checkpoint.com/2016/08/07/quadrooter/).

V **10th: Microsoft published their Security Bulletin Summary for August 2016, and released a total of nine updates, including five critical updates such as MS16-095, as well as four important updates.**
"Microsoft Security Bulletin Summary for August 2016" (https://technet.microsoft.com/library/security/ms16-aug).

S **13th: An entity calling themselves "The Shadow Brokers" released part of a group of files it claimed were from the Equation Group, and announced they would put the remaining files up for auction.**

S **14th: An incident of unauthorized login to the account of Russian athlete Yuliya Stepanova was confirmed in the Anti-Doping Administration and Management System (ADAMS) managed by the World Anti-Doping Agency (WADA). It was also revealed that multiple users had received phishing emails prior to this incident occurring.**
"WADA confirms illegal activity on Yuliya Stepanova's ADAMS account | World Anti-Doping Agency" (https://www.wada-ama.org/en/media/news/2016-08/wada-confirms-illegal-activity-on-yuliya-stepanovas-adams-account).

S **19th: The target list for the Anonymous attack campaign #OpKillingBay 2016 was released.**

O **19th: Twitter announced it had suspended 235,000 accounts over the past six months for alleged association with terrorist activities. It had already announced the suspension of 125,000 accounts in February, bringing the total number of accounts suspended to 360,000.**
"An update on our efforts to combat violent extremism | Twitter Blogs" (https://blog.twitter.com/2016/an-update-on-our-efforts-to-combat-violent-extremism).

V **25th: The INRIA group disclosed an attack method (SWEET32) that targets 64-bit block ciphers such as 3DES. It recovers plaintext by finding collisions in ciphertext after intercepting a large amount of data encrypted with the same key. This was dealt with by restricting the use of 3DES under default settings in implementations such as OpenSSL.**
"Sweet32: Birthday attacks on 64-bit block ciphers in TLS and OpenVPN" (https://sweet32.info/).

S **25th: It was revealed that the user information of approximately 25 million individuals had leaked from the Mail.ru message board site. A known SQL injection vulnerability in vBulletin was exploited.**
"LeakedSource Analysis of *.mail.ru Hack" (https://www.leakedsource.com/blog/mailru/).

S **28th: The DNS servers of SAKURA Internet were targeted in a DoS attack by an external party, affecting services. Rental servers and DNS servers were also later targeted by intermittent DoS attacks over a period of a few days.**
"Name server failure due to external DoS traffic" (http://support.sakura.ad.jp/mainte/mainteentry.php?id=20072) (in Japanese).

S **28th: The website of Gijutsu-Hyohron Co., Ltd. was targeted in a DoS attack by an external party, rendering their services unavailable. Intermittent DoS attacks were made later over a period of a few days.**
"Apology regarding intermittent connection failures due to DoS attacks" (https://gihyo.jp/news/info/2016/09/0901?ard=1472782622) (in Japanese).

S **30th: It was revealed that the FBI had issued alert information due to incidents of unauthorized access by an external party to voting systems in the states of Illinois and Arizona in the United States.**

S **31st: It was revealed that user email addresses and password information for approximately 68 million individuals who were registered users as of 2012 had leaked from Dropbox, and the passwords of affected users were reset.**
"Resetting passwords to keep your files safe | Dropbox Blog" (https://blogs.dropbox.com/dropbox/2016/08/resetting-passwords-to-keep-your-files-safe/).

S **31st: It was discovered that there were issues with the certificate issuance system at Chinese certificate authority WoSign that enabled fraudulent certificates to be issued.**
"The story of how WoSign gave me an SSL certificate for GitHub.com | Schrauger.com" (https://www.schrauger.com/the-story-of-how-wosign-gave-me-an-ssl-certificate-for-github-com). "WoSign Incidents Report (September 4th 2016)" (https://www.wosign.com/report/wosign_incidents_report_09042016.pdf).

*Dates are in Japan Standard Time

Legend										
	V	Vulnerabilities	S	Security Incidents	P	Political and Social Situation	H	History	O	Other

From past incidents we know that password information leaked in this way is at high risk of being exploited to target users that reuse the same password across multiple services. Attacks such as unauthorized logins to other sites for identity fraud, and the hijacking of SNS accounts of celebrities, other people, and organizations are common. Consequently, users need to take into account the possibility that password information may be leaked, and take preventative measures such as not using the same password across multiple services.

■ Government Agency Initiatives

The 9th assembly of the Cyber Security Strategic Headquarters was held by the National center of Incident readiness and Strategy for Cybersecurity (NISC) on August 31, and decided on "Cyber Security 2016" after considering the public comments to the draft put forward at the previous assembly*[26]. This is the second annual plan based on the Cyber Security Strategy approved by the cabinet in September 2015 that serves as the new national strategy indicating the basic future direction of cyber security policy. It also states concrete initiatives the government will implement in the 2016 fiscal year in line with the strategy. It incorporates a range of policies aimed at achieving the continuing development of Japan's economy, providing a safe and secure life for citizens, and maintaining peace in the international community. In addition to government agencies, critical infrastructure providers, companies, and individuals are also asked to cooperate and promote these initiatives.

In light of a series of information leaks in the tourism industry, in June of this year the Japan Tourism Agency established the "Committee on Information Leakage in Tourism Industry" to investigate the issues and put together measures to prevent recurrence, and its first assembly was held on July 8. At the second assembly on July 22, an interim report was created, and at the third assembly on September 16, the state of progress was confirmed. Also, the sharing of information within the tourism industry, as well as thorough information management and information sharing meetings to prevent information leaks from recurring, have been carried out jointly by the Japan Tourism Agency and the tourism industry intermittently from June*[27]. The interim report proposed measures that travel agencies and the Japan Tourism Agency should take going forward to prevent recurrence, and details of these have been made available to the general public through the information sharing meetings.

■ Other

On August 13, an entity calling themselves "The Shadow Brokers" released part of a group of files it claimed were from the Equation Group, and announced they would put the remaining files up for auction. The Equation Group is the name of an attack group that Kaspersky Labs reported on in 2015*[28]. Based on the characteristics of the attacks and similarities with other previous attacks, as well as the fact that some details match those listed in classified NSA documents taken by Edward Snowden, the Equation Group has been thought to be the Tailored Access Operations (TAO) department, which is chiefly responsible for attacks at the U.S. National Security Agency (NSA).

The published files included exploit code and malware that targeted the firewall products of Cisco and Juniper among others, and security researchers and vendors verified that the exploits actually worked. Some of the code exploited zero-day vulnerabilities that were not previously known. Vendors of the corresponding devices rushed to fix these vulnerabilities, but the exploit code known by the codename BENIGNCERTAIN was capable of remotely obtaining RSA private keys used in VPN encryption, and Cisco verified this not only affected the old PIX firewall products initially thought to be targeted, but also many network devices equipped with Cisco IOS software*[29].

The newly disclosed files contained a number of codenames listed in published documents that had been leaked by Mr. Snowden, and the MSGID included as an identifier for classified NSA documents newly-released by The Intercept was also used in binaries released by The Shadow Brokers*[30]. This means that it is extremely likely that the group of files published by The Shadow Brokers

*26 NISC, "9th meeting of the Cyber Security Strategic Headquarters" (http://www.nisc.go.jp/conference/cs/index.html#cs09) (in Japanese).

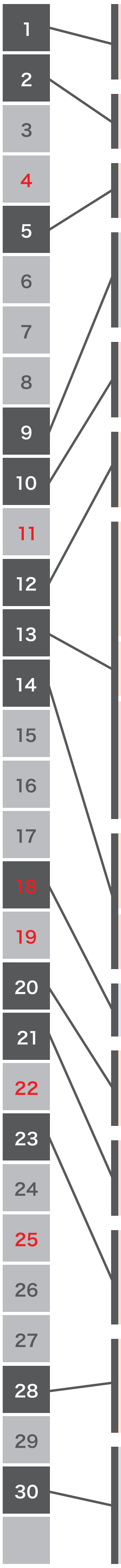
*27 "Overview of the '2nd Information Sharing Meeting of the Japan Tourism Agency and the Tourism Industry' | 2016 | Topics | News / Interviews | Japan Tourism Agency" (http://www.mlit.go.jp/kankocho/topics06_000080.html) (in Japanese).

*28 "Equation: The Death Star of Malware Galaxy - Securelist" (https://securelist.com/blog/research/68750/equation-the-death-star-of-malware-galaxy/).

*29 "IKEv1 Information Disclosure Vulnerability in Multiple Cisco Products" (https://tools.cisco.com/security/center/content/CiscoSecurityAdvisory/cisco-sa-20160916-ikev1).

*30 "The NSA Leak Is Real, Snowden Documents Confirm" (https://theintercept.com/2016/08/19/the-nsa-was-hacked-snowden-documents-confirm/).

September Incidents



S **1st: After a joint investigation by Denmark, Norway, Finland and the FBI, a 20-year-old Danish citizen was arrested on suspicion of involvement in attacks on many websites in Northern Europe that were part of the Anonymous #OpKillingBay attack campaign. He had used Twitter under the name RektFaggot, and it has been suggested he played an active role in conducting attacks against Japan in 2015.**

S **2nd: The website of mass electronics retailer Yodobashi Camera was targeted in a DoS attack by an external party, rendering their services unavailable. Intermittent DoS attacks were also conducted over a period of a few days.**

S **5th: An incident of unauthorized login by a third party occurred at the "Cecile Online Shop" shopping site of Dinos Cecile Co., Ltd.**
"Regarding unauthorized access to our 'Cecile Online Shop'" (http://www.cecile.co.jp/fst/information/20160905.pdf) (in Japanese).

O **9th: The U.S. government announced it had appointed its first Federal Chief Information Security Officer (CISO). The CISO was made responsible for advancing cyber security policy, planning, and implementation across U.S. government agencies.**
"Announcing the First Federal Chief Information Security Officer | whitehouse.gov" (https://www.whitehouse.gov/blog/2016/09/08/announcing-first-federal-chief-information-security-officer).

S **10th: The two founders of the vDOS DDoS-for-hire service were arrested in Israel.**
"Alleged vDOS Proprietors Arrested in Israel — Krebs on Security" (http://krebsonsecurity.com/2016/09/alleged-vdos-proprietors-arrested-in-israel/).

V **12th: A vulnerability in MySQL that could allow remote code execution and privilege escalation was discovered and disclosed by researchers.**
"MySQL <= 5.7.14 Remote Root Code Execution / Privilege Escalation (0day) (CVE-2016-6662)" (http://legalhackers.com/advisories/MySQL-Exploit-Remote-Root-Code-Execution-Privesc-CVE-2016-6662.html).

V **13th: Apple released iOS 10 and 10.0.1, fixing multiple vulnerabilities, including those that could allow a remote attacker to execute arbitrary code. Also, tvOS 10 and watchOS 3 were released.**
"About the security content of iOS 10" (https://support.apple.com/en-us/HT207143). "About the security content of iOS 10.0.1" (https://support.apple.com/en-us/HT207145). "About the security content of tvOS 10" (https://support.apple.com/en-us/HT207142). "About the security content of watchOS 3" (https://support.apple.com/en-us/HT207141).

V **13th: Multiple vulnerabilities in Adobe Flash Player that could allow illegal termination or arbitrary code execution were discovered and fixed.**
"Security updates available for Adobe Flash Player" (https://helpx.adobe.com/security/products/flash-player/apsb16-29.html).

S **13th: An entity calling itself the "Fancy Bears' Hack Team" obtained and released data from the Anti-Doping Administration and Management System (ADAMS) managed by the World Anti-Doping Agency (WADA). WADA also confirmed this fact.**
"American Athletes Caught Doping 2016-09-13" (http://fancybear.net/pages/1.html). "WADA Confirms Attack by Russian Cyber Espionage Group | World Anti-Doping Agency" (https://www.wada-ama.org/en/media/news/2016-09/wada-confirms-attack-by-russian-cyber-espionage-group).

V **14th: Microsoft published their Security Bulletin Summary for September 2016, and released a total of 14 updates, including seven critical updates such as MS16-104, as well as seven important updates.**
"Microsoft Security Bulletin Summary for September 2016" (https://technet.microsoft.com/library/security/ms16-sep).

S **14th: DC Leaks fraudulently obtained the Gmail data of Colin Powell, and published around two years' worth of emails between 2014 and 2016.**
"DC Leaks | Colin Luther Powell" (http://dcleaks.com/index.php/portfolio_page/colin-luther-powell/).

H **18th: Regarding attacks that have occurred around this day each year due to historical factors, this year saw no significantly notable organized attacks.**

V **20th: Apple released macOS Sierra 10.12, fixing multiple vulnerabilities, including those that could allow a remote attacker to execute arbitrary code.**
"About the security content of macOS Sierra 10.12" (https://support.apple.com/en-us/HT207170).

S **21st: Krebs on Security was targeted by a DDoS attack of around 620 Gbps, rendering it temporarily inaccessible.**
"KrebsOnSecurity Hit With Record DDoS — Krebs on Security" (https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/).

S **23rd: It was discovered that user email addresses and password information for at least 500 million individuals who were registered users as of late 2014 had leaked from U.S. Yahoo!.**
"An Important Message About Yahoo User Security | Yahoo" (https://yahoo.tumblr.com/post/150781911849/animportant-message-about-yahoo-user-security).

V **28th: A vulnerability in the process that creates DNS responses in BIND9 that could allow DoS attacks from an external source was discovered and fixed.**
"CVE-2016-2776: Assertion Failure in buffer.c While Building Responses to a Specifically Constructed Request | Internet Systems Consortium Knowledge Base" (https://kb.isc.org/article/AA-01419).

O **30th: The National center of Incident readiness and Strategy for Cybersecurity (NISC) announced the total budget for government cyber security. The requested budget amount for fiscal 2017 was 60.14 billion yen, an increase of around 10.3 billion yen over the initial budget of 49.83 billion yen for fiscal 2016. Note that 7.22 billion yen was added as part of the second supplementary budget for fiscal 2016.**
"Budget related to government cyber security (budgetary request for fiscal 2017 and secondary revision for the 2016 budget)" (http://www.nisc.go.jp/active/kihon/pdf/yosan2017.pdf) (in Japanese).

*Dates are in Japan Standard Time

Legend V Vulnerabilities S Security Incidents P Political and Social Situation H History O Other was not just the property of the Equation Group, but was also created by the NSA's TAO unit. There is a wide range of conjecture surrounding the true identity of The Shadow Brokers and their reasons for releasing these files, but the details are unclear and nothing definitive is known.

1.3 Incident Survey

1.3.1 DDoS Attacks

Today, DDoS attacks on corporate servers are almost a daily occurrence, and the methods involved vary widely. However, most of these attacks do not utilize advanced knowledge such as vulnerabilities, but aim to hinder or delay services by causing large volumes of unnecessary traffic to overwhelm network bandwidth or server processes.

■ Direct Observations

Figure 2 shows the state of DDoS attacks handled by the IIJ DDoS Protection Service between July 1 and September 30, 2016.

This shows the number of traffic anomalies judged to be attacks based on IIJ DDoS Protection Service criteria. IIJ also responds to other DDoS attacks, but these incidents have been excluded here due to the difficulty in accurately understanding and grasping the facts behind such attacks.

There are many methods that can be used to carry out a DDoS attack, and the capacity of the environment attacked (bandwidth and server performance) will largely determine the degree of impact. Figure 2 splits DDoS attacks into three categories: attacks against bandwidth capacity*[31], attacks against servers*[32], and compound attacks (several types of attacks against a single target conducted at the same time).

During these three months, IIJ dealt with 392 DDoS attacks. This averages out to 4.26 attacks per day, which is an increase in comparison to our prior report. Server attacks accounted for 62.76% of DDoS attacks, while compound attacks accounted for 31.89%, and bandwidth capacity attacks 5.36%.

The largest scale attack observed during this period was classified as a compound attack, and resulted in 2.97 Gbps of bandwidth using up to 890,000 pps packets. Of all attacks, 79.34% ended within 30 minutes of the start of the attack, 19.13% lasted between 30 minutes and 24 hours, and 1.53% lasted over 24 hours. The longest sustained attack for this period was a compound attack that lasted for five days, three hours, and 16 minutes (123 hours and 16 minutes).

We observed an extremely large number of IP addresses as the attack sources, whether domestic or foreign. We believe this is due to the use of IP spoofing*[33] and botnets*[34] to conduct the DDoS attacks.

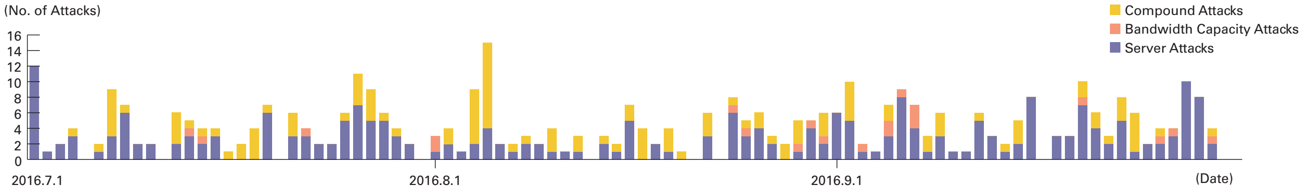


Figure 2: Trends in DDoS Attacks

*[31] Attack that overwhelms the network bandwidth capacity of a target by sending massive volumes of larger-than-necessary IP packets and fragments. When UDP packets are used, it is referred to as a UDP flood, while ICMP flood is used to refer to the use of ICMP packets.

*[32] TCP SYN flood, TCP connection flood, and HTTP GET flood attacks. In a TCP SYN flood attack, a large number of SYN packets that signal the start of TCP connections are sent, forcing the target to prepare for a large number of incoming connections, resulting in the waste of processing capacity and memory. TCP connection flood attacks establish a large number of actual TCP connections. In a HTTP GET flood a TCP connection with a Web server is established, and then a large number of GET requests in the HTTP protocol are sent, also resulting in a waste of processing capacity and memory.

*[33] Impersonation of a source IP address. Creates and sends an attack packet that has been given an IP address other than the actual IP address used by the attacker to make it appear as if the attack is coming from a different person, or from a large number of individuals.

*[34] A "bot" is a type of malware that after the infection, conducts an attack upon receiving a command from an external C&C server. A network made up from a large number of bots is called a botnet.

Backscatter Observations

Next we present DDoS attack backscatter observations*35 through the honeypots*36 of the IIJ malware activity observation project, MITF. Through backscatter observations, portions of DDoS attacks against external networks may be detectable as a third-party without intervening.

For the backscatter observed between July 1 and September 30, 2016, Figure 3 shows the source IP addresses classified by country, and Figure 4 shows trends in the number of packets by port.

The port most commonly targeted by DDoS attacks observed was port 80/TCP used for Web services, and accounted for 48.7% of the total. Attacks were also observed on 53/UDP used for DNS, 443/TCP used for HTTPS, 22/TCP used for SSH, and 27015/UDP that is sometimes used for gaming communications, as well as typically unused ports such as 19108/TCP, 8370/TCP, and 3306/UDP.

Communications at 53/UDP, which have been observed often since February 2014 and settled down on May 25th during the last reporting period, reoccurred around September 20th and have been observed at an average rate of 5,000 packets a day.

Looking at the source of backscatter packets by country thought to indicate IP addresses targeted by DDoS attacks in Figure 3, the United States accounted for the largest percentage at 30.0%, while China and France followed at 28.7% and 8.5%, respectively.

Now we will take a look at ports targeted in attacks where a large number of backscatter packets were observed. For attacks against Web servers (80/TCP and 443/TCP), there were attacks against a Bulgarian investigative journalism site from July 25 through July 28, and attacks against an online gaming site in China from July 28 through July 30. Attacks were also observed against a large number of servers for a CDN provider in the United States from August 11 through August 13, and against the official site of a shopping district for electronics in China on September 19 and September 30. Regarding other ports observed to have been affected, there were attacks against 6174/TCP targeting a specific IP address in China from July 21 through July 22, attacks against 19108/TCP targeting a specific IP address in China from July 31 through August 2, and attacks against 8370/TCP targeting a specific IP address in China from September 19 through September 20.

Notable DDoS attacks during the current survey period that were detected by IIJ's backscatter observations included attacks against WikiLeaks by a group calling themselves the "OurMine Team" on July 6, attacks against the Zimbabwe ruling party's site by Anonymous on July 7, and attacks against a mass electronics retailer site in Japan on September 3.

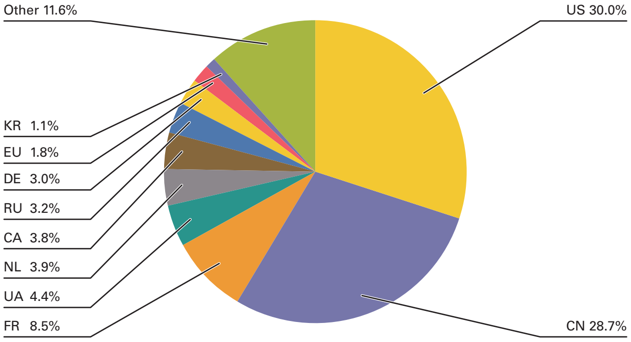


Figure 3: DDoS Attack Targets by Country According to Backscatter Observations

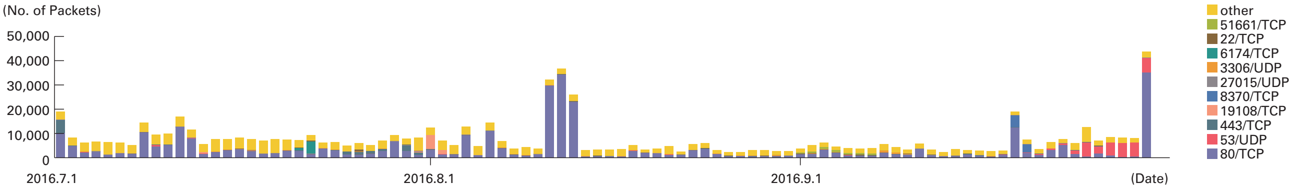


Figure 4: Observations of Backscatter Caused by DDoS Attacks (Observed Packets, Trends by Port)

*35 Honeypots placed by the MITF, a malware activity observation project operated by IIJ. See also "1.3.2 Malware Activities."

*36 The mechanism and limitations of this observation method, as well as some of the results of IIJ's observations, are presented in Vol.8 of this report (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol08_EN.pdf) under "1.4.2 Observations on Backscatter Caused by DDoS Attacks."

1.3.2 Malware Activities

Here, we will discuss the results of the observations of the MITF[*37], the malware activity observation project operated by IIJ. The MITF uses honeypots[*38] connected to the Internet in a manner similar to general users in order to observe communications that arrive over the Internet. Most appear to be communications by malware selecting a target at random, or scans attempting to search for a target to attack.

■ Status of Random Communications

Figure 5 shows the distribution of source IP addresses by country for incoming communications to the honeypots and Figure 6 shows the total volume (incoming packets) from July 1 through September 30, 2016. The MITF has set up numerous honeypots for its observations. Here, we have taken the average number per honeypot, and shown the trends for incoming packet types (top ten). Additionally, in these observations we made an adjustment so that multiple TCP connections to a specific port are counted as one attack, such as attacks against MSRPC.

Most of the communications that reached the honeypots during the survey period for this report were on 23/TCP used by Telnet, 1900/UDP used by SSDP, 22/TCP used by SSH, ICMP echo requests, 445/TCP used by Microsoft OSes, and 1433/TCP used by Microsoft's SQL Server.

Continuing the trend from the previous report, during the current survey period there was once again a high number of communications targeting 23/TCP used by Telnet, and increased even further from the latter half of September onwards. The National Police Agency and JPCERT/CC also reported that communications targeting 23/TCP had risen since the end of May this year[*39][*40][*41]. Meanwhile, incidents of dictionary attacks against Telnet targeting home routers and IoT devices (CCTV, DVR, NAS, etc.) that successfully compromised the devices and made them bots have been reported by multiple vendors[*42][*43][*44][*45][*46]. Bot infections that have been spreading include the Mirai bot, mentioned in 1.4.1, as well as Bashlite and Kaiten, which target Linux on IoT devices. This is leading IIJ to believe that most of these communications are scan attempts and infection activity targeting various IoT devices for which Telnet is enabled by default. During the period covered by this

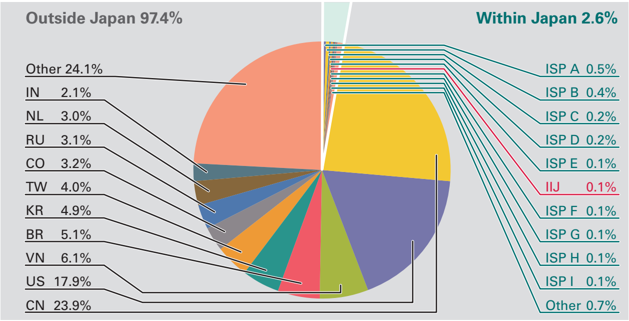


Figure 5: Sender Distribution
(by Country, Entire Period under Study)

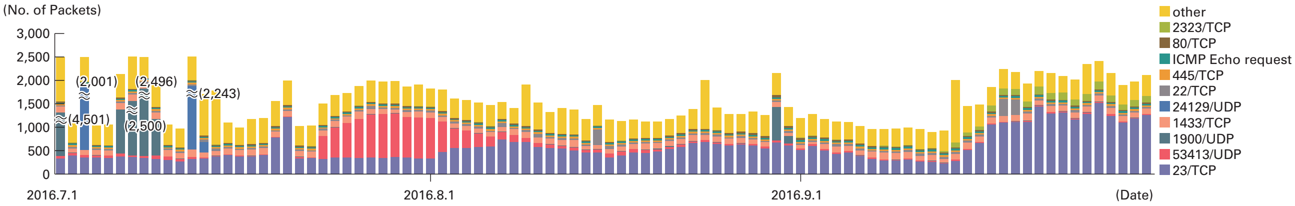


Figure 6: Incoming Communications at Honeypots (by Date, by Target Port, per Honeypot)

[*37] An abbreviation of Malware Investigation Task Force. The Malware Investigation Task Force (MITF) began its activities in May 2007, observing malware activity in networks through the use of honeypots in an attempt to understand the state of malware activities, to collect technical information for countermeasures, and to link these findings to actual countermeasures.

[*38] A system designed to record attacker and malware activities and their behavior by emulating vulnerabilities and simulating the damages caused by attacks.

[*39] "Internet observation results, etc. (June 2016)" (http://www.npa.go.jp/cyberpolice/detect/pdf/20160729.pdf) (in Japanese).

[*40] "Internet observation results, etc. (September 2016)" (http://www.npa.go.jp/cyberpolice/detect/pdf/20161020.pdf) (in Japanese).

[*41] "JPCERT/CC Internet Threat Monitoring Report [April 1, 2016 - June 30, 2016]" (http://www.jpcert.or.jp/english/doc/TSUBAMEReport2016Q1_en.pdf).

[*42] "CCTV DDoS Botnet In Our Own Back Yard" (https://www.incapsula.com/blog/cctv-ddos-botnet-back-yard.html).

[*43] "Attack of Things!" (http://blog.level3.com/security/attack-of-things/).

[*44] "IoT devices being increasingly used for DDoS attacks" (http://www.symantec.com/connect/blogs/iot-devices-being-increasingly-useddos-attacks).

[*45] "Large CCTV Botnet Leveraged in DDoS Attacks" (https://blog.sucuri.net/2016/06/large-cctv-botnet-leveraged-ddos-attacks.html).

[*46] "IoT Home Router Botnet Leveraged in Large DDoS Attack" (https://blog.sucuri.net/2016/09/iot-home-router-botnet-leveraged-in-large-ddos-attack.html).

report, the number of unique source IP addresses that targeted 23/TCP exceeded 1.4 million, demonstrating that a large number of devices may be infected with malware like these.

Access to 2323/TCP also increased from September. Figure 7 shows 2323/TCP access numbers by country. We can see that while next to no communications took place in July and August, they began to climb from September 6, and rose dramatically after September 14. The Mirai bot is known to have a characteristic where it communicates with 2323/TCP once every ten times, and when lining this up with the period in which communications began to increase, we believe the Mirai bot was responsible for most of it. Looking at the data by country, communications were received from IP addresses allocated to a wide range of countries, including Vietnam, China, Brazil, Colombia, and South Korea.

At the beginning of July, communications targeting the 1900/UDP used by the SSDP protocol increased. SSDP scanning requests were received from IP addresses allocated mainly to countries in the United States, China, France, South Korea, and Germany. These communications are thought to have been scanning for devices that could be used in DDoS attacks that utilize SSDP reflectors.

There was also a continued increase in communications to 1433/TCP. Upon investigation, we found that a large number of these communications were from IP addresses allocated to China, as well as many other IP addresses.

During the survey period for this report, there were once again a high number of communications to 53413/UDP. Our investigations found that these communications were attacks targeting a vulnerability in Netis and Netcore brand routers. The vulnerability was reported by Trend Micro in August 2014[47], and JPCERT/CC has reported there was a spike in attacks between April and June of 2015[48].

■ Malware Activity in Networks

Figure 8 shows the distribution of the source where malware artifacts were acquired from during the period under study, while Figure 9 shows trends in the total number of malware artifacts acquired. Figure 10 shows trends in the number of unique artifacts. In Figure 9 and Figure 10, the trends in the number of acquired artifacts show the actual number of artifacts acquired per day[49], while the number of unique artifacts is the number of artifact variants categorized in accordance with their hash digests[50]. Artifacts are also identified using anti-virus software, and a color-coded breakdown of the top 10 variants is shown along with the malware names. As with our previous report, for Figure 9 and Figure 10 we have detected Conficker using multiple anti-virus software packages, and removed any Conficker results when totaling data.

On average, 124 artifacts were acquired per day during the period under study, while there were 20 unique artifacts per day. After investigating the undetected artifacts more closely, included were multiple SDBOT families (a type of IRC bot) observed from IP addresses allocated to countries such as Taiwan, India, and Vietnam, as well as bitcoin mining tool downloaders.

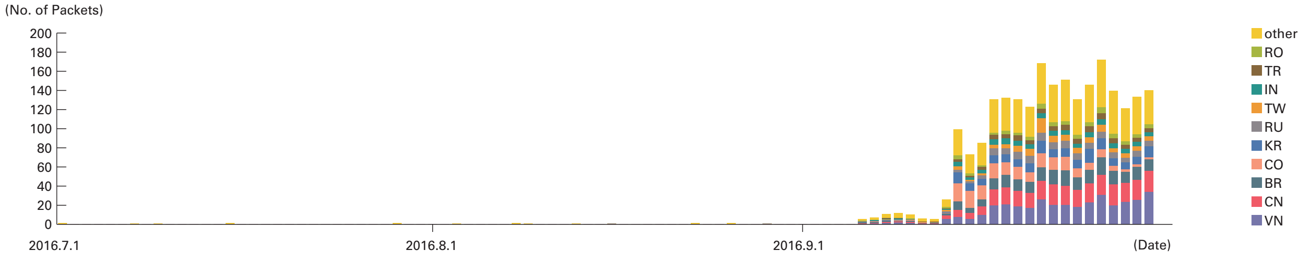


Figure 7: Incoming Communications at Honeypots (by Date, 2323/TCP, per Honeypot)

[47] "Netis Routers Leave Wide Open Backdoor" (http://blog.trendmicro.com/trendlabs-security-intelligence/netis-routers-leave-wide-open-backdoor/).

[48] "JPCERT/CC Internet Threat Monitoring Report [April 1, 2015 - June 30, 2015]" (http://www.jpcert.or.jp/english/doc/TSUBAMEReport2015Q1_en.pdf).

[49] This indicates malware acquired by honeypots.

[50] This value is calculated by utilizing a one-way function (hash function) that outputs a fixed-length value for each input. Hash functions are designed to produce a different output for practically every different input. We cannot guarantee the uniqueness of artifacts through hash values alone, given that obfuscation and padding may result in artifacts of the same malware having different hash values. The MITF understands this limitation when using this method as a measurement index.

About 61% of the undetected artifacts were in text format. Many of these text format artifacts were HTML, and 404 or 403 error responses from Web servers. We believe these were due to infection activities of old malware, such as worms continuing despite the closure of the download sites that newly-infected PCs attempted to access to download the malware. Another factor that impacted results for this reporting period was maintenance we performed on our honeypot environment to detect attacks over HTTP and FTP, which enabled us to acquire bots in PHP format and redirectors through .htaccess. A MITF independent analysis revealed that during the current period under observation 28.9% of malware artifacts acquired were worms, 61.3% were bots, and 9.8% were downloaders. The ratio of bots climbed significantly, but as mentioned previously this is due to many bots in PHP format being acquired since the honeypot environment was changed to make it possible to detect attacks over HTTP and FTP. In addition, the MITF confirmed the presence of 54 botnet C&C servers*[51] and 110 malware distribution sites. This is a large increase over the previous survey period, but this can be attributed to factors such as the increased number of malware acquired and updates to the analysis environment, the honeypot maintenance, as well as the detection of malware with DGA (domain generated algorithms).

■ Conficker Activity

Including Conficker, an average of 4,185 artifacts were acquired per day during the period under study for this report, representing 374 unique artifacts. Conficker accounted for 97.0% of the total artifacts acquired, and 94.8% of the unique artifacts. Since Conficker remains the most prevalent malware by far, we have omitted it from the figures in this report. Compared to the previous survey report, the total number of artifacts acquired this survey period has decreased by approximately 51%. This is believed to be due to the declining trend towards the latter half of the previous survey period, and because there were changes in the IP addresses of sensors as a result of the maintenance performed on the honeypots when transitioning to the current survey period. According to the observations by the Conficker Working Group*[52], as of October 2016 a total of 500,000 unique IP addresses are infected. This indicates a drop to about 16% of the 3.2 million PCs observed in November 2011, but shows that infections are still widespread.

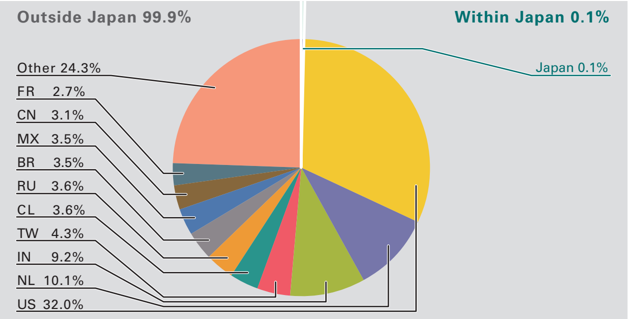


Figure 8: Distribution of Acquired Artifacts by Source
(by Country, Entire Period under Study, Excluding Conficker)

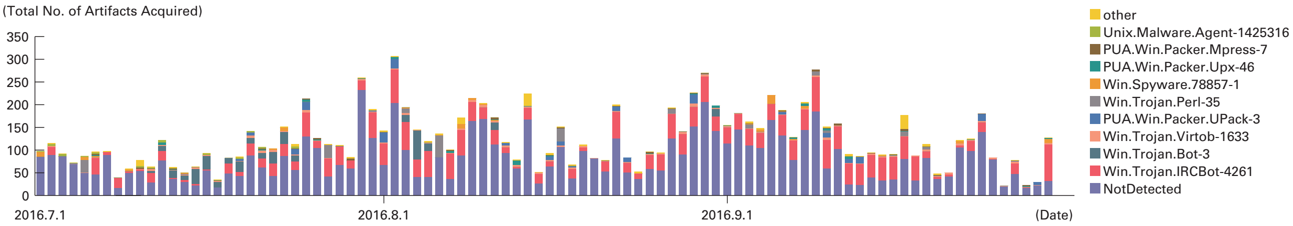


Figure 9: Trends in the Total Number of Malware Artifacts Acquired (Excluding Conficker)

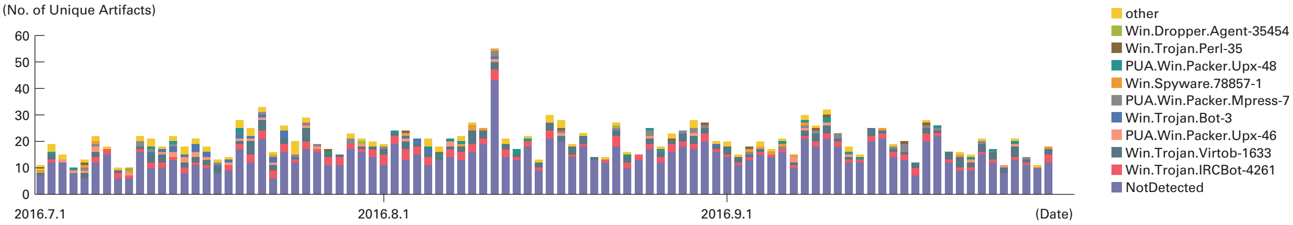


Figure 10: Trends in the Number of Unique Artifacts (Excluding Conficker)

*[51] An abbreviation of Command & Control server. A server that provides commands to a botnet consisting of a large number of bots.

*[52] Conficker Working Group Observations (http://www.confickerworkinggroup.org/wiki/pmwiki.php/ANY/InfectionTracking). Because no numerical data beyond January 7 is available within the current survey period, we have visually observed the highest value in the graph from early October, and used it.

1.3.3 SQL Injection Attacks

Of the different types of Web server attacks, IIJ is conducting ongoing investigations on SQL injection attacks*53. SQL injection attacks have been noted a number of times in the past, and continue to remain a major topic in Internet security. SQL injection attacks are known to attempt one of three things: the theft of data, the overloading of database servers, or the rewriting of Web content.

Figure 11 shows the source distribution of SQL injection attacks against Web servers detected between July 1 and September 30, 2016. Figure 12 shows the trend in the number of attacks. These are a summary of attacks detected through signatures in the IIJ Managed IPS Service. The United States was the source for 35.7% of attacks observed, while China and Japan accounted for 23.7% and 9.8%, respectively, with other countries following. The total number of SQL injection attacks against Web servers has remained nearly level with the previous report, but those from China have risen while those from Japan have decreased.

During this period, attacks from multiple sources in China directed at multiple targets took place on July 31. On August 6, there were attacks from a specific source in the United States directed at specific targets. On September 3, there were attacks from a specific source in the China directed at specific targets. On September 24, there were attacks from sources in a variety of regions directed at specific targets. Also, on September 25 there were attacks from Indonesia against specific targets. These attacks are thought to have been attempts to find Web server vulnerabilities.

As previously shown, attacks of various types were properly detected and handled within the scope of the service. However, attack attempts continue, requiring ongoing caution.

1.3.4 Website Alterations

Here we indicate the status of website alterations investigated through the MITF Web crawler (client honeypot)*54.

This Web crawler accesses hundreds of thousands of websites on a daily basis, focusing on well-known and popular sites in Japan. The number of sites that it accesses are added accordingly. In addition to this, we temporarily monitor websites that have seen short-term increases in access numbers. By surveying websites thought to be viewed frequently by typical users in Japan, it becomes easier to speculate trends for fluctuations in the number of altered sites, as well as the vulnerabilities exploited and malware being distributed.

For the period between July 1 and September 30, 2016, drive-by download attacks using the Neutrino Exploit Kit accounted

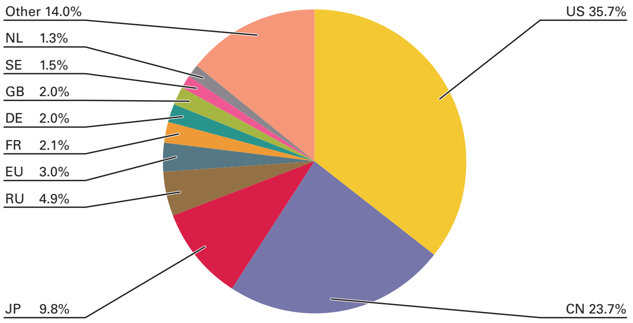


Figure 11: Distribution of SQL Injection Attacks by Source

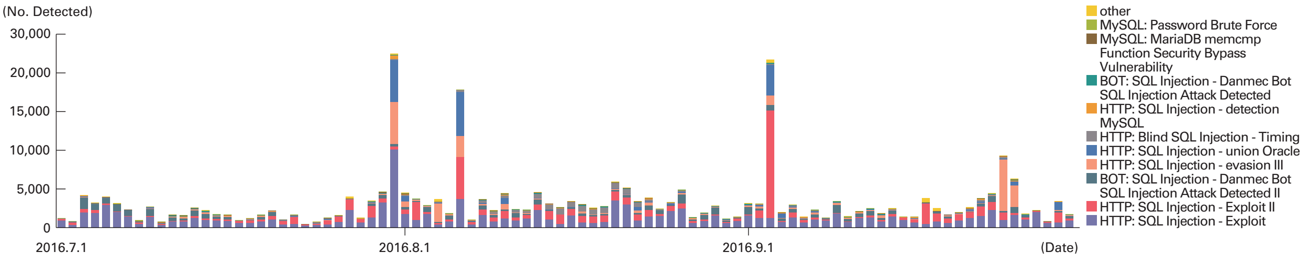


Figure 12: Trends in SQL Injection Attacks (by Day, by Attack Type)

*53 Attacks accessing a Web server to send SQL commands, and operating against an underlying database. Attackers access or alter the database content without proper authorization to steal sensitive information or rewrite Web content.

*54 Refer to "1.4.3 Website Defacement Surveys Using Web Crawlers" in Vol.22 of this report (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol22_EN.pdf) for a description of Web crawler observation methods.

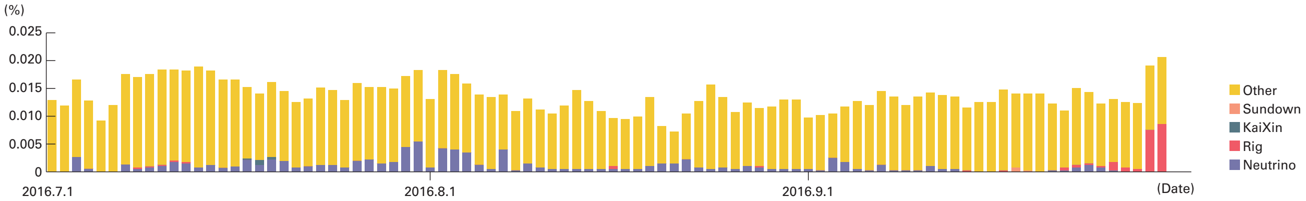
for the majority of passive attacks detected (Figure 13). This is a trend that had continued since the Angler Exploit Kit disappeared in June 2016. However, Neutrino was no longer observed at all around the end of September, with the Rig Exploit Kit instead being observed at a large scale[*55][*56]. Payloads such as Locky, Cerber, and Ursnif have been confirmed for these.

No points in common can be confirmed in the scope or content of the websites exploited as redirectors to exploit kits, but a number of websites using WordPress have been confirmed. We have observed many cases in which websites exploited as redirectors for Angler and Neutrino, as well as websites exploited for other fraudulent behavior, have functioned as redirectors for Rig. We have also confirmed that when accessing these redirector websites with a Mac OS X client, either you are not redirected to the landing page, or the landing page does not return a response. During this survey period no drive-by download attacks targeting Max OS X were observed[*57].

There continue to be a high number of observations where a dialog box implying the user has a malware infection is displayed on the browser screen, redirecting the user to a fraudulent site that prompts them to install a PUA[*58] or call a fake support center. These fraudulent sites also function in Max OS X environments.

Multiple cases have also been seen where attempts are made to force users to download executable files such as ransomware or PUA using elements like the Location header from the redirecting website. In these cases, the file is not executed automatically, but if a browser like Internet Explorer is being used, a dialog box asking whether or not to execute the file is displayed, which may lead users to execute it by mistake. In cases such as these where an attempt is made to force users to download executable files directly, the locations used to store the executable file included certain cloud storage services or the server where the website used to redirect users is hosted.

Because the number of drive-by download attacks has risen sharply from late September, we recommend implementing thorough vulnerability countermeasures, such as version management of the OS, applications, and plug-ins in browser environments, and the introduction of EMET[*59]. For website operators, it is essential to take measures against vulnerabilities by managing vulnerabilities in Web applications, frameworks, and plug-ins, and by also managing mashup content provided by external parties, such as advertisements and aggregation services.



*Covers several hundreds of thousands of sites in Japan. In recent years, drive-by downloads using exploit kits have been configured to change attack details and even whether or not to attack based on the client system environment or session information, source address attributes, and an attack quota such as the number of attacks. This means that results can vary wildly depending on the test environment and other circumstances.
*Threats based on passive attacks other than exploit kits, such as direct links to fraudulent sites and executable files, are classified as Other.

Figure 13: Rate of Passive Attack Incidence When Viewing Websites (%) (by Exploit Kit)

[*55] On September 29, we enhanced our Web crawler to improve the accuracy of detecting Rig EK. We believe these enhancements are a direct cause of the dramatic rise in Rig EK observations after September 29. However, the number of Rig EK observations had been increasing immediately prior to these enhancements. The large-scale Rig attack campaign has also been discussed in articles such as Malwarebytes Labs' "RIG exploit kit takes on large malvertising campaign" (https://blog.malwarebytes.com/cybercrime/exploits/2016/09/rig-exploit-kit-takes-on-large-malvertising-campaign/). For this reason, we estimate that Rig-based attacks began increasing from between mid-August and mid-September at the latest.

[*56] A quick report regarding the observation status of Rig EK between late September and mid-October 2016 is given under IIJ-SECT's "Alert regarding the growing number of Rig Exploit Kit observations" (https://sect.iij.ad.jp/d/2016/10/178746.html) (in Japanese).

[*57] The MITF Web crawler system conducts additional surveys using a Max OS X environment client honeypot when a website is observed behaving in a way that indicates the possibility of a passive attack via a Windows environment client honeypot.

[*58] An abbreviation of Potentially Unwanted Application. This is a generic term for applications deemed unnecessary for general work tasks, and thought to potentially lead to unwanted results for PC users and system administrators.

[*59] Examples include separating administrator privileges and applying application white lists. See Vol.31 of this report (http://www.iij.ad.jp/en/company/development/iir/031.html) under "1.4.2 Hardening Windows Clients Against Malware Infections" for more information.

1.4 Focused Research

Incidents occurring over the Internet change in type and scope from one minute to the next. Accordingly, IIJ works toward implementing countermeasures by continuing to conduct independent surveys and analyses of prevalent incidents. Here, we present information from the surveys we have conducted during this period, covering Mirai botnet detection and countermeasures, and discussing miscellaneous SSL/TLS topics.

1.4.1 Mirai Botnet Detection and Countermeasures

■ About the Mirai Botnet

The Mirai botnet (herein referred to as "Mirai") is malware that constructs a botnet by infecting IoT devices such as cameras and digital video recorders that can be connected to a network. Individually, IoT devices do not have significant processing capacity but when a large number of devices are used together, it becomes possible to launch an extremely powerful attack. The security of many IoT devices is not managed appropriately, and it can be said that it is relatively easy to form massive botnets using them.

In late September 2016, large-scale DDoS attacks were conducted against the U.S. security information site "Krebs on Security" and the French Internet service provider OVH*[60]*[61], and it is said that Mirai was used in these attacks. These DDoS attacks were of unprecedented scale, up to 665 Gbps against Krebs on Security, and 1 Tbps against OVH. Following this, in early October the creator of Mirai, who called himself Anna-senpai, abruptly released its source code. It is not currently possible to access the source code that Anna-senpai released, but it has been mirrored to various locations, and these mirrors can be accessed by anyone.

In this report we explain how Mirai operates based on the source code that was released, and discuss methods to detect Mirai, determine whether there is an infection and how to apply countermeasures*[62].

■ Mirai System Configuration and Operation

Figure 14 shows the minimum system configuration for Mirai. IoT devices often use non-x86 CPUs, so Mirai also includes shell script for cross-compiling ARM and MIPS binaries. It is possible to specify a variety of build options in this shell script, but because an "ssh" option is not implemented, only the "telnet" option has any meaning. Also, all communications associated with Mirai are carried out in plaintext.

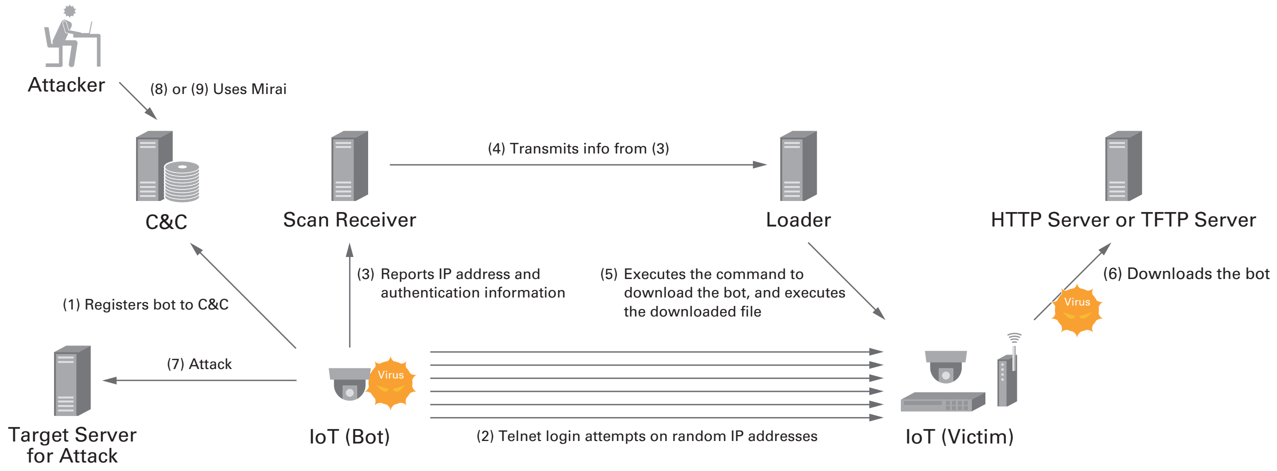

Attacker
(8) or (9) Uses Mirai
C&C
Scan Receiver
(4) Transmits info from (3)
Loader
HTTP Server or TFTP Server
(1) Registers bot to C&C
(3) Reports IP address and authentication information
(5) Executes the command to download the bot, and executes the downloaded file
(6) Downloads the bot
Virus
(7) Attack
Virus
Target Server for Attack
IoT (Bot)
(2) Telnet login attempts on random IP addresses
IoT (Victim)

Figure 14: Mirai Botnet System Configuration

*60 Krebs on Security, "KrebsOnSecurity Hit With Record DDoS" (https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/).

*61 A tweet reporting on the status of the DDoS on OVH (https://twitter.com/olesovhcom/status/779297257199964160).

*62 In this document, we refer to the source code at the following repository as of the end of October. GitHub - jgamblin/Mirai-Source-Code: Leaked Mirai Source Code for Research/IoC Development Purposes (https://github.com/jgamblin/Mirai-Source-Code).

IoT (Bot)

This is an IoT device that has been infected with Mirai and has become a bot. When Mirai infects an IoT device, the watchdog timer is disabled to prevent the device from automatically rebooting. It also checks whether or not it is possible to connect to 48101/TCP on the local host to determine if other Mirai bot instances are running. If running, the corresponding processes are terminated. Next, it changes its own process name to a random character string. If the processes up to this point succeed, "listening tun0" is written into the standard output as a sign that the launch was successful. The fact that the downloaded bot has been launched successfully is also confirmed by checking this string in the Communications (5) process mentioned later. Then, Mirai ends*[63] processes bound to 22/TCP, 23/TCP, and 80/TCP, and binds itself to these ports to prevent access to the management interface. Processes such as those containing ".anime" in the execution path, those for which the executable file has been deleted, and Qbot/Zollard/Remaiten processes are also terminated. Terminating the administrative processes and malware in this way prevents the Mirai-infected IoT device from being restored, or infected with other malware. Also, for obfuscation, the domain names and port numbers for the C&C server and Scan Receiver, as well as authentication information used in login attempts against other IoT devices, are XORed byte-by-byte with values derived from a key, named 0xdeadbeef.

Communications (1): A connection is made to the domain name and port number of the C&C server hard-coded into the bot (default: cnc.changeme.com, 23/TCP). After a TCP session is established, the bot sends data in the format shown in Table 1 to the C&C server and this bot becomes registered on the C&C server. Also, every 60 seconds a heartbeat is sent from the bot in the data format shown in Table 2. Meanwhile, attack commands are sent from the C&C server to the bot in the data format shown in Table 3 as necessary. This TCP session is persistent, and automatically reconnects even if it becomes disconnected.

Table 1: Data Format for C&C Connection

Packet	Data Length (Bytes)	Meaning
1	4	0x00 0x00 0x00 0x01 (a fixed byte string)
2	1	The data length sent in Packet 3 (*1)
3	The length specified in (*1)	The first parameter when the bot is launched

Table 2: Data Format for Heartbeat to C&C

Data Length (Bytes)	Meaning
2	0x00 0x00 (a fixed byte string) When the C&C server receives a heartbeat, it responds by sending a heartbeat back.

Table 3: Data Format for Attack Commands

Data Length (Bytes)	Meaning
2	Data length (up to 4096)
4	Attack execution time
1	Attack ID
1	Number of attack targets (combinations of IP addresses and netmasks for the attack targets are listed after this based on the number of targets specified here)
4	Attack target IP address
1	Netmask
1	Number of flags (combinations of flags are listed after this based on the number of flags specified here)
1	Flag ID
1	Length of data specified in flag (*2)
The length specified in (*2)	Data specified in flag (string)

*63 80/TCP does not end when using the bundled shell script compile option.

Communications (2): A SYN scan is performed on 23/TCP targeting a random IPv4 address. However, one in ten times a SYN scan is performed on 2323/TCP. This is thought to be because some devices targeted for infection provide Telnet over 2323/TCP instead of 23/TCP. Also, if the IP address matches one of those shown in Table 4, another random IP address is selected. After connecting to an IP address that responds to the SYN scan, the bot confirms whether it is possible to log in using the authentication information hard-coded into itself. There are 61 username and password combinations as shown in Table 5, and each is weighted, so the combinations are not attempted equally. These communications continue to occur while a device is infected with Mirai.

Communications (3): When a login attempt from (2) is successful, a connection is made with the Scan Receiver hard-coded into the bot (default: report.changeme.com, 48101/TCP). After establishing a TCP session, the authentication information used in the successful attempt is sent in the data format shown in Table 6.

Communications (7): When an attack command is received from the C&C server, an attack is made against the designated attack targets. See Table 7 for details regarding the attack methods used. Also, for HTTP floods, code used to recognize DOSarrest and CloudFlare is implemented. For DOSarrest, code thought to be intended to counteract the DDoS protection service has been implemented, but CloudFlare is only recognized, and no specially designed code for counteracting it has been implemented. The source code also contains the name "Proxy knockback connection" thought to be a planned attack implementation, and other attack methods that are implemented but not called.

Table 4: IP Addresses Not Subject to Attack

IP Address	Allocated Party
127.0.0.0/8	Loopback
0.0.0.0/8	Invalid address space
3.0.0.0/8	General Electric Company
15.0.0.0/7	Hewlett-Packard Company
56.0.0.0/8	US Postal Service
10.0.0.0/8	Internal network
192.168.0.0/16	Internal network
172.16.0.0/14	Internal network
100.64.0.0/10	IANA NAT reserved
169.254.0.0/16	IANA NAT reserved
198.18.0.0/15	IANA Special use
224.*.*.*+	Multicast
IP addresses with one of the following as the first octet: 6, 7, 11, 21, 22, 26, 28, 29, 30, 33, 55, 214, 215	Department of Defense

Table 6: Data Format for Authentication Information Sent to Scan Receiver

Data Length (Bytes)	Meaning
1	0x00 (fixed)
4	IP address
2	Port number
1	User name length (*3)
The length specified in (*3)	User name (string)
1	Password length (*4)
The length specified in (*4)	Password (string)

Table 5: Authentication Information Hard-Coded into Bot

User Name	Password	User Name	Password
root	xc3511	admin1	password
root	vizxv	administrator	1234
root	admin	666666	666666
admin	admin	888888	888888
root	888888	ubnt	ubnt
root	xmhdipc	root	klv1234
root	default	root	Zte521
root	juantech	root	hi3518
root	123456	root	jvbzd
root	54321	root	anko
support	support	root	zlxx.
root	(none)	root	7ujMko0vizxv
admin	password	root	7ujMko0admin
root	root	root	system
root	12345	root	ikwb
user	user	root	dreambox
admin	(none)	root	user
root	pass	root	realtek
admin	admin1234	root	00000000
root	1111	admin	1111111
admin	smcadmin	admin	1234
admin	1111	admin	12345
root	666666	admin	54321
root	password	admin	123456
root	1234	admin	7ujMko0admin
root	klv123	admin	1234
Administrator	admin	admin	pass
service	service	admin	meinsm
supervisor	supervisor	tech	tech
guest	guest	mother	fXXXXr (partially omitted)
guest	12345		

Scan Receiver

This is a server that receives scan results from the bot using 48101/TCP. It contains only basic functionality where it parses the information received and outputs it to a standard output. This output result is fed into the Loader that we discuss later.

Process (4): Information on IoT devices that can be logged into that is received from the bot is input to the standard input of the Loader. In the source code that was made available, no mechanism for automatically transferring information had been prepared.

Loader

This is the server that performs the actual infection activity based on the information on IoT devices that can be logged into that had been provided as input. After launching, it can create a large number of threads, and infect multiple IoT devices simultaneously.

Communications (5): Login to the IoT device is performed based on information received from the Scan Receiver. After login, the binary for the bot is downloaded and executed using a busybox wget command or tftp command. At this time, it is necessary to download a binary for the bot compatible with the CPU architecture of the IoT device, and the Loader determines this by analyzing the ELF header of the "/bin/echo" binary in the target IoT device. If wget or tftp cannot be used, the /bin/echo command in the IoT device is used to send and execute the bot downloader. When infection is successful, the same behavior listed under "IoT (Bot)" is observed.

Table 7: DDoS Attack List

Attack ID	Command	Attack Description	Attack Details
0	udp	UDP flood	Sends a large number of UDP packets.
1	vse	Valve source engine specific flood	Performs a UDP flood targeting the Source Engine.
2	dns	DNS resolver flood using the targets domain, input IP is ignored	Performs a DNS water torture attack against specified domain names. If cache DNS settings have not been configured on the IoT device, the following cache DNS servers are used. 8.8.8.8, 74.82.42.42, 64.6.64.6, 4.2.2.2
3	syn	SYN flood	Sends a large number of SYN packets.
4	ack	ACK flood	Sends a large number of ACK packets.
5	stomp	TCP stomp flood	An attack intended to bypass devices with DDoS protection. Sends a large number of ACK packets after establishing a TCP session.
6	greip	GRE IP flood	Sends a large number of GRE-encapsulated IP-UDP packets.
7	greeth	GRE Ethernet flood	Sends a large number of GRE-encapsulated ETH-IP-UDP packets.
8	None	Proxy knockback connection	Not implemented, so the details are unknown.
9	udpplain	UDP flood with less options. optimized for higher PPS	A UDP flood made faster by reducing the number of configured items.
10	http	HTTP flood	Sends a large number of HTTP requests such as GET. A random User-Agent is selected from Table 8.

Table 8: User-Agents Used in HTTP Floods

User-Agent
Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36
Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.0.2743.116 Safari/537.36
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.0.2743.116 Safari/537.36
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_6) AppleWebKit/601.7.7 (KHTML, like Gecko) Version/9.1.2 Safari/601.7.7

Communications (6) According to the information hard-coded in the Loader, the busybox wget command or tftp command is used to download the binary for the bot (wget default: 100.200.100.100, 80/TCP, file path: /bins/mirai.[arch] *64) (tftp default: 100.200.100.100, file name: mirai.[arch]). In the case of the bot downloader sent in (5), the bot is downloaded from the hard-coded HTTP server (default: 127.0.0.1, 80/TCP, file path: /bins/mirai.[arch]).

C&C

This is the server an administrator or user logs into to use the Mirai botnet. User accounts, attack execution history, and a whitelist of attack targets are managed through a database. The number of bots that can be used, maximum attack duration, and cooldown time is configured for each user, and after an attack is completed the next attack cannot be made until the set cooldown time has passed. Also, no commands have been prepared to stop the attacks or suspend bot activity.

Communications (8): The administrator or user connects to the C&C server over Telnet to use the botnet. A user name and password are used for authentication. Bots also connect to the C&C server over 23/TCP, and are identified as such using Telnet IAC directly after the 23/TCP connection is established. See Table 9 for the commands that can be used. Figure 15 shows an example attack command. Also, some of the messages when logging into the C&C server are in Russian, suggesting that either the creator or someone associated with the creator is from a Russian-speaking country (Figure 17).

Communications (9): The administrator or user can also use the botnet via API by connecting to 101/TCP instead of Telnet. Figure 16 shows an example attack command.

Identifying Infection Attempts and the Presence of Infections

Firewall Logs

If incoming access attempts from outside the firewall to 23/TCP and 2323/TCP are at a ratio of 9 to 1, this is a likely indication of exposure to Mirai infection activities. Because the target IP address is selected randomly, almost any organization could be targeted by these attacks. Conversely, if outgoing access attempts from inside the firewall to 23/TCP and 2323/TCP are at a ratio

Table 9: Commands Usable on the C&C Server

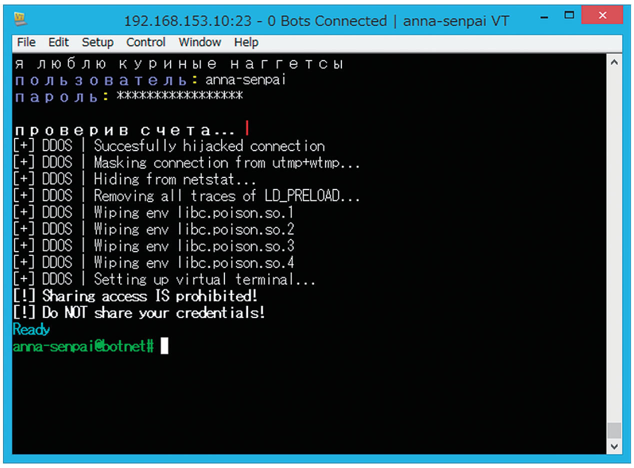
Command	Details
adduser	Adds a user. Can only be executed by an administrator.
botcount	Displays the number of bots connected to the C&C server. Can only be executed by an administrator.
-<BotCount>	Specifies the number of bots to use in an attack.
@<string>	Specifies the category of bots to use in an attack.
?	Displays help messages.
source	Can only be specified as an attack flag. Specifies the source IP address. Can only be executed by an administrator.

```
[-<BotCount> ]<atk cmd> <ip_addr>[/<mask>][,<ip_addr>[/<mask>]]...
<duration> [<flags>]
```

Figure 15: Attack Command Example (CLI)

```
<API Key>|[-<BotCount> ]<atk cmd> <ip_addr>[/<mask>][,<ip_addr>[/<mask>]]...
<duration> [<flags>]
```

Figure 16: Attack Command Example (API)



```
я люблю куриные наггетсы
пользователь: anna-senpai
пароль: ****************

проверив счета...
[+] DDOS | Succesfully hijacked connection
[+] DDOS | Masking connection from utmp+wtmp...
[+] DDOS | Hiding from netstat...
[+] DDOS | Removing all traces of LD_PRELOAD...
[+] DDOS | Wiping env libc.poison.so.1
[+] DDOS | Wiping env libc.poison.so.2
[+] DDOS | Wiping env libc.poison.so.3
[+] DDOS | Wiping env libc.poison.so.4
[+] DDOS | Setting up virtual terminal...
[!] Sharing access IS prohibited!
[!] Do NOT share your credentials!
Ready
anna-senpai@botnet#
```

Figure 17: C&C Login Screen

*64 [arch] indicates the CPU architecture.

of 9 to 1, this is almost certainly an indication that an internal device is infected with Mirai. The source IP address is not spoofed when performing the scans, so it is easy to identify any infected devices.

IDS/IPS

Mirai communications between systems are all carried out in plaintext, so it can be monitored using IDS/IPS. Figure 18 shows examples of a Snort signature that detects Mirai communications. This is based on the released source code, so the communications of a newer version of Mirai or a variant may not be detectable.

Countermeasures

IoT Device Manufacturers

When considering the security of IoT devices, the manufacturer plays an extremely important role. In the series of incidents related to Mirai, many news articles and reports have suggested changing the password of IoT devices as a countermeasure. While admittedly the fact that many users are not changing the default password is a direct cause contributing to the extremely large number of infected IoT devices, it has also been said that for some devices the password is hard-coded, meaning it is not possible to change the password or disable these accounts. In cases like these, users cannot take any measures to protect themselves, so it can be considered that it is the manufacturer’s responsibility to take some action. Manufacturers should revise their security designs taking into consideration the following items:

- Do not create backdoor accounts.
- Do not hard-code passwords.
- Do not limit the types of characters that can be used in passwords.
- Describe the entire management interface in manuals
  (and do not implement any backdoors).
- Avoid plaintext communications such as Telnet and HTTP.
- Make users change the password upon initial login, to prevent the use of the default password.

```
- Bot registration and heartbeat
alert tcp any any -> any 23 (msg:"Mirai Botnet: Register Bot with C&C"; flow:to_server,established; content:"|00 00 00 01|"; depth:4; sid:1000000; rev:1)
alert tcp any any -> any 23 (msg:"Mirai Botnet: Send Heartbeat from Bot to C&C"; flow:to_server,established; content:"|00 00|"; depth:2; pcre:"/^\x00\x00$/m";
sid:1000001; rev:1)
alert tcp any 23 -> any any (msg:"Mirai Botnet: Reply Heartbeat from C&C to Bot"; flow:from_server,established; content:"|00 00|"; depth:2; pcre:"/^\x00\x00$/m";
sid:1000002; rev:1)

- Bot downloader download
alert tcp any any -> any [23,2323] (msg:"Mirai Botnet: Download Bot Downloader via Telnet (echo)"; flow:to_server,established; content:"echo -ne '"; content:"' > upnp|3b|
/bin/busybox ECCHI"; sid:1000060; rev:1)

- Bot binary download command execution
alert tcp any any -> any [23,2323] (msg:"Mirai Botnet: Download Bot binary via Telnet (wget)"; flow:to_server,established; content:"/bin/busybox wget http://";
content:"/bins/mirai."; content:"-O - > dvrHelper|3b| /bin/busybox chmod 777 dvrHelper|3b| /bin/busybox ECCHI"; sid:1000070; rev:1)
alert tcp any any -> any [23,2323] (msg:"Mirai Botnet: Download Bot binary via Telnet (tftp)"; flow:to_server,established; content:"/bin/busybox tftp "; content:" -g -l
dvrHelper -r mirai."; content:"/bin/busybox chmod 777 dvrHelper|3b| /bin/busybox ECCHI"; sid:1000071; rev:1)

- Bot binary download communications
alert tcp any any -> any 80 (msg:"Mirai Botnet: Download Bot binary via HTTP"; flow:to_server,established; content:"GET /bins/mirai."; pcre:"/^GET
/bins/mirai\.(arm|arm7|m68k|mips|mpsl|ppc|sh4|spc|x86) HTTP/1\.[01]|0d 0a|$/mi"; sid:1000080; rev:1)
alert udp any any -> any 69 (msg:"Mirai Botnet: Download Bot binary via TFTP"; flow:to_server; content:"|00 01|mirai."; pcre:"/^\x00\x01mirai\.(arm|arm7|m68k|mips|mps-
l|ppc|sh4|spc|x86)\x00.+$/mi"; sid:1000081; rev:1)

- Bot execution
alert tcp any any -> any [23,2323] (msg:"Mirai Botnet: Run Bot binary (upnp & dvrHelper)"; flow:to_server,established; content:"./upnp|3b| ./dvrHelper telnet.";
content:"/bin/busybox IHCCE"; pcre:"/^\.\/upnp\; \.\/dvrHelper telnet\.(arm|arm7|m68k|mips|mpsl|ppc|sh4|spc|x86)\; \/bin\/busybox IHCCE/m"; sid:1000090; rev:1)
alert tcp any any -> any [23,2323] (msg:"Mirai Botnet: Run Bot binary (dvrHelper)"; flow:to_server,established; content:"./dvrHelper telnet."; content:"/bin/busybox IHCCE";
pcre:"/^\.\/dvrHelper telnet\.(arm|arm7|m68k|mips|mpsl|ppc|sh4|spc|x86)\; \/bin\/busybox IHCCE/m"; sid:1000091; rev:1)
```

Figure 18: Signature for Snort

Users

Because Mirai leaves no files behind, and only runs in memory, it is possible to restore an infected IoT device by rebooting it. However, it will be infected again right away if the default password is still used, so it is necessary to change the password.

This applies to products other than IoT devices as well, but it is essential to change the password to something other than the default. The default authentication information is listed in the manual, so it should be considered that attackers know this information. For this reason, it is not an exaggeration to say that using the default password is equivalent to not having any authentication at all. When changing the password, as often said, it is best to set a password that is as complex and as long as possible. Also, you should not connect IoT devices directly to the Internet unless absolutely necessary. If it is necessary to access the device via the Internet, apply appropriate access controls through either the IoT device itself or a firewall, etc.

1.4.2 Miscellaneous SSL/TLS Topics

Over the past few years, many new attacks against the SSL/TLS protocols and their implementations have been released, and we have seen major shifts in trends, such as migration requests due to the compromise of cryptographic algorithms, and the near completion of the new version, TLS 1.3. Therefore, in this section we will report on changing trends related to SSL/TLS, and discuss challenges we face as we move toward the IoT age.

The History of SSL/TLS Version Changes

SSL 2.0 was released by Netscape Communications in 1995, and after a number of extensions were added and a number of issues fixed, SSL 3.0*[65] was released the following year. SSL 2.0 had no function for preventing the alteration of the Handshake message portion (i.e. data integrity is not guaranteed), so MITM attacks were possible, and the protocol itself is recognized as vulnerable*[66]. Also, with the discovery of the POODLE attack*[67] in October 2014, padding oracle attacks against SSL 3.0 are now possible when the CBC cipher mode is used to encrypt messages, so it is currently recommended that SSL 3.0 not be used*[68].

TLS, the successor to SSL, now has three versions: TLS 1.0 (established in 1999), TLS 1.1 (established in 2006), and TLS 1.2 (established in 2008). Each of these protocols are still in widespread use. After TLS 1.0 was drawn up by the IETF based on SSL 3.0, TLS 1.1 was then designed to bolster its security by, for example, incorporating measures in its specifications beforehand to prevent the BEAST attack and its variants that the original protocol was vulnerable to when using CBC cipher mode. TLS 1.2 also enabled the use of authenticated encryption (AEAD: Authenticated Encryption with Associated Data)*[69]. But these protocols have been targeted in many attacks over the past few years. RFC7457*[70], which was issued in February 2015, summarizes the history of attacks against TLS that were known to the public by around 2014. It covers a wide variety of known attacks, pointing out vulnerabilities related to the RC4 stream cipher, which we will introduce next, and discussing downgrade attacks that force

*65 When SSL 3.0 was released in 1996 it was not a specification established under IETF guidance, but it was given RFC status as RFC6101 to serve as a historical resource (https://datatracker.ietf.org/doc/rfc6101/).

*66 RFC6176: Prohibiting Secure Sockets Layer (SSL) Version 2.0 (https://datatracker.ietf.org/doc/rfc6176/).

*67 An explanation of the POODLE attack is given in Vol.25 of this report published in November 2014 (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol25_EN.pdf), under "1.4.2 The POODLE Attack."

*68 RFC7568: Deprecating Secure Sockets Layer Version 3.0 (https://datatracker.ietf.org/doc/rfc7568/).

*69 RFC5116: An Interface and Algorithms for Authenticated Encryption (https://datatracker.ietf.org/doc/rfc5116/).

*70 RFC7457: Summarizing Known Attacks on Transport Layer Security (TLS) and Datagram TLS (DTLS) (https://datatracker.ietf.org/doc/rfc7457/). In March 2014, CRYPTREC published the CRYPTREC Cryptographic Technology Guideline - Countermeasures against recent attacks on TLS/SSL (http://www.cryptrec.go.jp/report/c13_kentou_giji02_r2.pdf) (in Japanese), which contains information on some of the attacks listed in RFC7457.

users to use a lower TLS version than expected, as well as timing attacks that occur when the compression function is enabled. For attacks after that period, portal sites such as CELLOS*[71] can be checked for information on major SSL/TLS vulnerabilities, but it has become extremely difficult to accumulate knowledge about the respective attacks and deal with them each time they appear.

As an example, let's look at cases involving the RC4 and TripleDES cryptographic algorithms. RC4 is a well-known stream cipher that has been used extensively to date, and is defined in the cipher suites within SSL/TLS. A wide range of attack models can be considered when attacking cryptographic algorithms, but for cryptographic protocols like SSL/TLS, there is a condition requirement called Broadcast setting, when considering a real use case. This is an assumption where a large amount of ciphertext can be obtained from the same plaintext (data before being encrypted) that is encrypted using multiple keys. When considering how SSL/TLS is used, this is a fairly realistic use case. A large amount of research and cryptanalysis based on this attack model has been performed since 2001*[72], and attacks where plaintext can be recovered through the bias of the stream keys generated by RC4 have been published. Specifically, a technique for generating large volumes of ciphertext by running malicious JavaScript in a browser has been written, and a paper presented at USENIX Security 2015 reported that it was possible to steal a cookie with a success rate of 94% by obtaining 9×2^{27} ciphertexts*[73]. In response to the various research and findings, the IETF considered this a real threat and issued a RFC in February 2015 to eliminate the use of RC4.*[74].

Meanwhile, the SWEET32 attack*[75] further reinforced the fact that TripleDES is vulnerable. This is not an attack method against a cryptographic algorithm itself, but a potential attack that could be successful when using the CBC cipher mode in SSL/TLS. This makes it impossible to prevent completely, and countermeasures by vendors all involve limiting or lowering the priority for use of TripleDES. Next, we will focus on the resources required to conduct this attack successfully. A paper presented at ACM CCS '16 stated that to restore a 2-block cookie would require capturing 785 gigabytes of ciphertext over a 38 hour period.*[76]. RC4 bias attacks are conducted against the RC4 stream cipher itself, but while SWEET32 attacks target a 64-bit block cipher, the use of CBC cipher mode is required, so we would hesitate to call this a direct attack against a cryptographic algorithm. When the SWEET32 attack appeared, an Internet draft*[77] suggesting that the use of TripleDES be terminated was reconsidered. Similar to RFC7465 that removed RC4 as a usable algorithm, discussions regarding TripleDES took place in the CFRG (Crypto Forum Research Group), but did not make it to RFC status.

*71 CELLOS consortium, Publication (https://www.cellos-consortium.org/index.php?Publication).

*72 See the following article for a summary of the RC4 bias attacks researched since 2001. Kenneth G. Paterson, "Big Bias Hunting in Amazonia: Large-Scale Computation and Exploitation of RC4 Biases", ASIACRYPT2014 Invited Talk (http://des.cse.nsysu.edu.tw/asiacrypt2014/doc/8_1_Big%20Bias%20Hunting%20in%20Amazonia%20Large-scale%20Computation%20and%20Exploitation%20of%20RC4%20Biases.pdf).

*73 Mathy Vanhoef, Frank Piessens, "All Your Biases Belong To Us: Breaking RC4 in WPA-TKIP and TLS" (https://www.rc4nomore.com/vanhoefusenix2015.pdf). A summary can also be read on the RC4 NOMORE (https://www.rc4nomore.com/) site.

*74 RFC7465: Prohibiting RC4 Cipher Suites (https://datatracker.ietf.org/doc/rfc7465/).

*75 Sweet32: Birthday attacks on 64-bit block ciphers in TLS and OpenVPN (https://sweet32.info/).

*76 Karthikeyan Bhargavan and Gaëtan Leurent, "On the Practical (In-)Security of 64-bit Block Ciphers: Collision Attacks on HTTP over TLS and OpenVPN", ACM CCS'16 (http://dl.acm.org/citation.cfm?id=2978423&CFID=697886415&CFTOKEN=82935453).

*77 B. Kaduk et al., Deprecate 3DES and RC4 in Kerberos (https://www.ietf.org/archive/id/draft-kaduk-kitten-des-des-des-die-die-die-00.txt).

More interestingly, the SWEET32 attack points out something worth looking at. TripleDES and the CBC cipher mode are both still included on the CRYPTREC E-Government Recommended Ciphers List[78]. Using both of these presumably “safe” primitives simultaneously has triggered a vulnerability. While this example is not necessarily a case of misusing a cryptographic algorithm, there are cryptographic algorithms that take into account potential misuses such as this. For example, in the AEAD CAESAR[79] competition, there is an algorithm proposed with a property called Nonce Misuse-Resistant, which assures security when the same nonce is used by accident even though it is assumed that a different nonce is to be used each time. Another misuse example is where the private key may be leaked when using the same parameters for signatures in DSA signatures[80].

Progress has also been made in techniques that can be categorized as timing attacks or side channel attacks, which differ from the attacks that attempt to recover text or keys at the cryptographic algorithm layer directly, such as those mentioned above. As a result, migration to TLS 1.2 is recommended for compatibility with AEAD, which cannot be used with TLS 1.0 and TLS 1.1. This recommendation to use AEAD was further strengthened when the Lucky13[81] attack demonstrated that timing attacks against TLS 1.2 were possible. Prior to that, it was thought that the BEAST attack and its variants could be prevented using TLS version 1.1 or higher, even when CBC cipher mode was used. Furthermore, there was increased urgency for migration to AEAD because RC4 can no longer be used, as mentioned above. However, we need to consider that each version has cipher suites that must be implemented. For example, TLS_DHE_DSS_WITH_3DES_EDE_CBC_SHA is mandatory in TLS 1.0, and the same goes for TLS_RSA_WITH_3DES_EDE_CBC_SHA in TLS 1.1. In both of these cases, the symmetric-key algorithm uses TripleDES in CBC cipher mode. Note that TLS_RSA_WITH_AES_128_CBC_SHA is a mandatory implementation in TLS 1.2. AES is used, so while there is a shift to use more secure algorithms, the use of CBC cipher mode is required in all versions of TLS. The point that can be taken from this fact is that padding oracle attacks against CBC cipher mode were not taken into consideration when the RFC was developed. Particularly on the server, cipher suites need to be appropriately prioritized, so that a cipher suite that uses CBC mode is not selected. Also, when selecting the cipher suite to use, a public key cryptographic algorithm that supports forward secrecy[82] is recommended. It is also necessary to exercise caution when configuring export-grade algorithms. It has been previously thought that it was safe to leave the weak cipher suites as part of the configuration, since the server would select the strongest available cryptographic algorithm. However, the FREAK attack made public in January 2015 and the Logjam attack[83] made public in May 2015 demonstrated attacks against configurations where Export-grade cryptographic algorithms are configured. Furthermore, in March 2016 the DROWN attack[84] was disclosed, and demonstrated that in situations where SSL 2.0 is enabled, ciphertext recovery attacks were possible even in environments that do not use Export-grade cryptographic algorithms.

[78] CRYPTREC, List of ciphers that should be referred to in the procurement for the e-Government system (CRYPTREC Ciphers List) (https://www.cryptrec.go.jp/english/list.html). However, TripleDES is rated as “permitted to be used for the time being.” Whether this can be interpreted as secure depends on the reader.

[79] CAESAR: Competition for Authenticated Encryption: Security, Applicability, and Robustness (https://competitions.cr.yp.to/caesar.html).

[80] RFC6979: Deterministic Usage of the Digital Signature Algorithm (DSA) and Elliptic Curve Digital Signature Algorithm (ECDSA) (https://datatracker.ietf.org/doc/rfc6979/).

[81] Nadhem AlFardan, Kenny Paterson, “Lucky Thirteen: Breaking the TLS and DTLS Record Protocols” (http://www.isg.rhul.ac.uk/tls/Lucky13.html).

[82] An explanation of forward secrecy is given in Vol.22 of this report published in February 2014 (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol22_EN.pdf), under “1.4.2 Forward Secrecy.”

[83] Yuji Suga, A Wonderful Encounter between Cryptography and Society: 2. Overview of SSL/TLS Protocol and the Security of Cryptographic Protocols - Our Permanent Battles against the Vulnerabilities of Secure Standard Protocols, “Information Processing” Bulletin Vol.56 No.11 (http://id.nii.ac.jp/1001/00145437/) (in Japanese).

[84] The DROWN Attack (https://drownattack.com/).

■ Examples of Configuration Criteria that Servers Should Meet

In the previous section, we mentioned that given all the attacks against SSL/TLS, it is very difficult to deal with each attack as they become public. In this section, we will introduce a number of documents that describe the kinds of evaluation criteria Web server administrators can use when going to apply their own Web settings. One of these is the configuration guidelines*[85] regarding cryptographic technology for SSL/TLS sites, published by CRYPTREC in May 2015. This documents configuration criteria required for SSL/TLS servers, including protocol versions, server certificates, and cipher suites. These guidelines were not written specifically for government systems alone, so they can be used as a reference document for improving the configuration of Web servers used in private enterprises. However, since around two years have passed since it was originally published, there are some parts that are not in line with the current state, so ideally, they should be updated or supplemented with some information. A detailed report has been published*[86] that discusses various products and appliances and whether they meet the requirements of these guidelines under default settings, or if they can be configured manually to do so, showing that information on practical countermeasures is becoming more readily available. Other information in Japanese on security assessments of cryptographic algorithms has been made available by CELLOS (Cryptographic protocol Evaluation toward Long-Lived Outstanding Security), a non-profit organization, and CRYPTREC had discussions in fiscal 2015 at the Priority Issues Assessment Task Force. These have resulted in the establishment of the Cryptographic Protocol Issue Assessment Working Group*[87] started in fiscal 2016 for further discussion.

Web browser vendors have also compiled information on Web server configuration*[88]*[89]. We will not provide details on each, but in addition to dealing with cipher suites, there are mentions of HSTS (HTTP Strict Transport Security), which is at the core in the shift to the full-time use of SSL/TLS. Qualys SSL Lab, which provides a test site for SSL/TLS servers, has also published a document*[90] regarding the best practices for configuration. It lists detailed evaluation criteria*[91], so that the reason behind the rating it provides can be understood and server configurations can be improved. Users can also view evaluation results by simply entering a FQDN, so configuration issues on SSL/TLS servers are essentially public information. However, one cannot determine whether the reason for a low rating is a misconfiguration or if the configuration is in place while tolerating risk. Despite this fact, this activity is a good thing since it is promoting the migration from weak cryptographic algorithms.

■ Changes in Browser Vendor Support Status

From the perspective of rating websites, it is now possible to know the status of a server easily through a web browser. One way this is done is through security indicators in the area where the URL is displayed. For example, a green bar is shown when a user accesses a server that implements EV SSL certificates. In Chrome, the method for displaying these security indicators changed in version 52 (for Macintosh desktops only; version 53 in other environments). The changes were made after a presentation at an international conference on usability security held in June 2016*[92], resulting in an improved interface being introduced. This paper took the approach of providing test subjects with several variations of the security indicator icons that indicate the status of a SSL/TLS connection, and had them choose what they felt was the best, and the test results were implemented into Chrome. The icons shown have meanings relating to the trustworthiness of connections and servers, and different icons are used accordingly. There

*[85] CRYPTREC, Guidelines for Cryptographic Configuration of SSL/TLS Implementations - For a Secure Website (cryptographic configuration measures) - (https://www.ipa.go.jp/security/vuln/ssl_crypt_config.html) (in Japanese).

*[86] Information-technology Promotion Agency, Japan, "Research report on cryptographic configuration methods for SSL/TLS appliance products, etc." published (http://www.ipa.go.jp/security/fy28/reports/crypto_survey/) (in Japanese).

*[87] CRYPTREC, CRYPTREC Report 2015 (https://www.cryptrec.go.jp/report/c15_prom_web.pdf).

*[88] Google Developers - Web Fundamentals, "Enabling HTTPS on Your Servers" (https://developers.google.com/web/fundamentals/security/encrypt-in-transit/enable-https).

*[89] Mozilla, "Security/Server Side TLS" (https://wiki.mozilla.org/Security/Server_Side_TLS).

*[90] Qualys SSL Lab, "SSL and TLS Deployment Best Practices Version 1.5 (8 June 2016)" (https://github.com/ssllabs/research/wiki/SSL-and-TLS-Deployment-Best-Practices).

*[91] Qualys SSL Lab, "SSL Server Rating Guide" (https://www.ssllabs.com/projects/rating-guide/). The latest version at the time of writing is 2009k (October 14, 2015) (https://www.ssllabs.com/downloads/SSL_Server_Rating_Guide.pdf), but in November 2016 it was announced that it will be updated in or after 2017: Announcing SSL Labs Grading Changes for 2017 (https://blog.qualys.com/ssllabs/2016/11/16/announcing-ssl-labs-grading-changes-for-2017).

*[92] Adrienne Porter Felt et al., Rethinking Connection Security Indicators", SOUPS2016 (http://research.google.com/pubs/pub45366.html).
The paper can also be found on the USENIX site (https://www.usenix.org/system/files/conference/soups2016/soups2016-paper-porter-felt.pdf).

are separate icons for a proper HTTPS communication with a valid EV SSL certificate, but also HTTPS communications with minor errors, and HTTPS communications with major errors. Of these, HTTPS communications with minor errors is shown as mixed content (HTTP content mixed with HTTPS content) where there is content that HTTP points to within the HTML content. Browser vendors have appealed to sites that are causing mixed content to rectify the matter*[93]*[94]. Prior to the aforementioned update, the icon provided a neutral impression, but the paper resulted in the selection of an icon that did not appear to be critically important but provided a negative impression to get the attention of users. Because Web administrators have not been able to catch up with these browser vendor changes, a number of SSL/TLS servers are unintentionally sending out HTTPS content that results in mixed content errors, so we recommend double-checking the status.

The list of root certificates and intermediate CA certificates that OSes and applications should retain, manage, and trust is referred to as the certificate store. As a general rule, SSL/TLS clients such as browsers refer to this certificate store to determine whether or not the corresponding Web server can be trusted, and it shows the status within the application. In most cases, it is possible to manually add and delete to the certificates stored in the certificate store, but users are generally not aware of the certificates when running an application for use. When obtaining a certificate, SSL/TLS server administrators must know the status of vendor certificate stores. This is important because the group of certificates contained in each certificate store differ slightly between the various OSes and applications, so there are cases where a particular certificate may be judged secure by a certain certificate store, but fail certificate verification in another certificate store when the chain of trust to the root cannot be traced. Some certificate authorities that issue certificates have been driven out of business after being unable to meet the various criteria from vendors. This demonstrates how strongly dependent this model is on the organizational entities that create trust anchors, thus creating a situation where the trends surrounding browser vendors are being brought to attention.

There is one other major problem in terms of certificate verification. In recent years, there have been an extremely large number of reports of a vulnerability in various Android applications that do not properly perform the process involving the certificate store, and cut short the certificate verification process. As this example shows, users are either unaware of or do not understand PKI, and only use the user interface provided by the application to view the status of a communication. In fact, as described above, users are now able to identify the status of a communication through security indicators due to the efforts of major browser vendors. This has led to more and more cases where users completely trust what is being displayed in their browser, without investigating other channels. Web browsers, which are a type of SSL/TLS client, need to have an interface that enables a user to quickly determine whether communications are dangerous or not, both when browsing through a standard PC and through devices with a smaller display area such as tablets and smartphones. Consideration must be given to security indicators on IoT products going forward, since they lack display capability.

■ TLS 1.3

Regarding TLS 1.3*[95], the successor version to TLS 1.2, the TLS working group last call announcement for its draft was made at IETF meeting 97*[96] in November 2016. It is at the last phase prior to issuing of the RFC in February 2017. In TLS 1.3, no cipher suites are specified as mandatory for implementation. CBC mode used in the past is not considered for use, and only cipher suites with AEAD algorithms that provide both encryption and MAC (data indicating that data has not been altered) are listed. AEAD is

*93 Google Developers - Web Fundamentals, Preventing Mixed Content (https://developers.google.com/web/fundamentals/security/prevent-mixed-content/fixing-mixed-content).

*94 Mozilla support, Mixed content blocking in Firefox (https://support.mozilla.org/en-US/kb/mixed-content-blocking-firefox).

*95 The Transport Layer Security (TLS) Protocol Version 1.3 (https://tlswg.github.io/tls13-spec/). At the time of writing, the Internet-Draft managed by the IETF is at Version 18 (https://datatracker.ietf.org/doc/draft-ietf-tls-tls13/18/).

*96 Eric Rescorla, TLS 1.3 (draft-ietf-tls-tls13-18) (https://www.ietf.org/proceedings/97/slides/slides-97-tls-tls-13-00.pdf).

included in 5 suites: AEAD_AES_128_GCM, AEAD_AES_256_GCM*[97], AEAD_AES_128_CCM*[98], AEAD_CHACHA20_POLY1305*[99], and AEAD_AES_128_CCM_8, and these can be used in place of the CBC cipher mode block ciphers and stream ciphers. TLS 1.3 has no interface to derive key data for MACs (message authentication codes) that were generated in previous versions, so in practice its use will be limited to AEAD alone.

As shown in the interconnectivity results reported at IETF meeting 97, the implementation of TLS 1.3 is progressing on both browsers and servers, and in some products users are able to confirm this*[100]*[101]. In part due to the lack of compatibility with TLS 1.2, a major version update (TLS 2.0, TLS 2, TLS 4) was also proposed, but at the meeting a majority of the people backed TLS 1.3, so the protocol will be promoted under this name. With the release of an RFC next year, we believe many implementations will support TLS 1.3. Meanwhile, because usage scenarios in which portable data formats are used to switch between multiple browsers are also anticipated, attacks targeting this functionality may become public. The application of post-quantum cryptography*[102] to TLS has also garnered a lot of attention*[103]. We believe that a chaotic situation involving activities on opposite sides of the spectrum will occur. Namely, issues with version migration on legacy devices and the addition of new functions will most likely continue for some time.

1.5 Conclusion

This report has provided a summary of security incidents that IIJ has responded to. This time we discussed Mirai botnet detection and countermeasures, and examined miscellaneous SSL/TLS topics. IIJ makes every effort to inform the public about the dangers of Internet usage by identifying and disclosing information on incidents and associated responses through reports such as this.

Authors:

Mamoru Saito

Director of the Advanced Security Division, and Manager of the Office of Emergency Response and Clearinghouse for Security Information, IIJ. After working in security services development for enterprise customers, in 2001 Mr. Saito became the representative of the IIJ Group emergency response team IIJ-SECT, which is a member team of FIRST, an international group of CSIRTs. Mr. Saito serves as a steering committee member for several industry groups, including ICT-ISAC Japan, Information Security Operation providers Group Japan, and others.

Masafumi Negishi (1.2 Incident Summary)
Tadashi Kobayashi, Tadaaki Nagao, Hiroshi Suzuki, Minoru Kobayashi, Hisao Nashiwa (1.3 Incident Survey)
Minoru Kobayashi (1.4.1 Mirai Botnet Detection and Countermeasures)
Yuji Suga (1.4.2 Miscellaneous SSL/TLS Topics)
Office of Emergency Response and Clearinghouse for Security Information, Advanced Security Division, IIJ

Contributors:
Yasunari Momoi, Hiroyuki Hiramatsu, Office of Emergency Response and Clearinghouse for Security Information, Advanced Security Division, IIJ

[97] RFC5288: AES Galois Counter Mode (GCM) Cipher Suites for TLS (https://datatracker.ietf.org/doc/rfc5288).
[98] RFC6655: AES-CCM Cipher Suites for Transport Layer Security (TLS) (https://datatracker.ietf.org/doc/rfc6655).
[99] RFC7539: ChaCha20 and Poly1305 for IETF Protocols (https://datatracker.ietf.org/doc/rfc7539/).
[100] Chrome Platform Status, TLS 1.3 (https://www.chromestatus.com/feature/5712755738804224).
[101] CloudFlare, Introducing TLS 1.3 (https://blog.cloudflare.com/introducing-tls-1-3/).
[102] An explanation of post-quantum cryptography is given in Vol.31 of this report published in June 2016 (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol31_EN.pdf), under “1.4.2 Trends in Post-Quantum Cryptography.”
[103] Chrome Platform Status, “CECPQ1 in TLS” (https://www.chromestatus.com/feature/5749214348836864).

Industry Efforts to Unify Streaming Formats

The technologies that comprise streaming delivery can be categorized into several classes (Table 1). These technologies were proposed by a variety of companies, and many of them are now international standards. Combining individual technologies to achieve optimal streaming makes it necessary to reconcile the standards involved. Currently a great shift is beginning to take place among the "streaming formats" class of technologies.

The HTTP Live Streaming (HLS) format advocated by Apple was announced in 2009. Ever since then, this key format has played a leading role in streaming. Its distinguishing characteristic is the adoption of segmented MPEG2-TS as the container format, with small files storing video and audio data delivered in a continuous stream. The use of HTTP/1.1 as the delivery protocol also had a major impact. HLS was initially a format designed for iOS, but it later became widely used on macOS, tvOS, and Android as well.

Apple has presented a standards proposal for HLS in Internet-Draft (I-D) form to the Internet Engineering Task Force (IETF) that promotes standards for the Internet. This document is titled "draft-pantos-http-live-streaming." I-Ds are intended to be work-in-progress documents, and although some may be published as an RFC, others may never progress beyond the proposal stage. Since Apple issued the first version on May 1, 2009, this I-D has continued to be submitted under the name of Roger Pantos, and as of September 2016 it is up to version 20.

The structure of HLS is extremely simple. An example of the file containing the instructions for playback (the manifest file) is shown below (Figure 1).

The essential part of this is the resources identified by URI. In this example, the three media files "first.ts," "second.ts," and "third.ts" must be located on the media.example.com server. It doesn't matter whether these files are mounted statically or dynamically. Then, as you can see from the scheme name, it is indicated that these media files will be distributed via HTTP.

This manifest file is also placed on a Web server and passed on to the media player. The media player reads the reference information and accesses the media files listed by URI in order from top to bottom. Comment rows contain reference information, and in the case of the example shown in Figure 1, EXT-X-TARGETDURATION indicates the maximum length of the media files, and EXTINF indicates the actual length of the next media file in seconds. In other words, we can see that the media specified by this playback instruction file has a total playback duration of 21.021 seconds.

No standardization activity is being carried out for this I-D. To publish it as an RFC it would need to be discussed by an IETF working group. But this I-D has not been brought up as an ongoing issue at any working group, and has yet to be discussed. It is treated as if it were a privately published document.

In what could be interpreted as a competing standard, MPEG-DASH (Dynamic Streaming over HTTP) was standardized by the MPEG (Moving Picture Experts Group) of the International Organization for Standardization (ISO), and published as ISO/IEC 23009-1 in 2012. MPEG is a group of experts who have been active since 1988, acting as a working group for standardization tasks.

Class	Technology
Presentations	HTML5, Flash, etc.
Codecs, metadata	H.264, H.265 (HEVC), AAC, WebVTT
Containers	MPEG2-TS, MP4
Streaming formats	HLS, MPEG-DASH, CMAF
Delivery protocols	HTTP/1.1
Transport protocols Network protocols	TCP/IP

Table 1: Streaming Delivery Structure

```
#EXTM3U
#EXT-X-TARGETDURATION:10
#EXTINF:9.009,
http://media.example.com/first.ts
#EXTINF:9.009,
http://media.example.com/second.ts
#EXTINF:3.003,
http://media.example.com/third.ts
#EXT-X-ENDLIST
```

Figure 1: HLS Manifest File Example

This group mainly focused on the area of standardization work for video and audio compression formats up until now, but as the streaming of video over the Internet became more popular they also branched out into drawing up streaming formats.

```
<MPD xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="urn:mpeg:dash:schema:mpd:2011"
xsi:schemaLocation="urn:mpeg:dash:schema:mpd:2011
DASH-MPD.xsd" mediaPresentationDuration="PT0H1M6.1S"
minBufferTime="PT1.5S"
profiles="urn:mpeg:dash:profile:isoff-on-demand:2011"
type="static">
  <Period duration="PT0H1M6.1S" start="PT0S">
    <AdaptationSet>
      <Representation bandwidth="4000000"
codecs="avc1.4d401e" height="1080" id="1"
mimeType="video/mp4" width=“1920">
        <BaseURL>video_4000.mp4</BaseURL>
        <SegmentBase indexRange="860-1023">
          <Initialization range="0-859" />
        </SegmentBase>
      </Representation>
    <Representation bandwidth="2400000"
codecs="avc1.4d401e" height="720" id="2"
mimeType="video/mp4" width="1280">
        <BaseURL>video_2400.mp4</BaseURL>
        <SegmentBase indexRange="859-1022">
          <Initialization range="0-858" />
        </SegmentBase>
      </Representation>
    </AdaptationSet>
  <AdaptationSet>
    <Representation bandwidth="128000" codecs="mp4a.40.2"
id="5" mimeType="audio/mp4">
        <BaseURL>audio_128.mp4</BaseURL>
        <SegmentBase indexRange="783-946">
          <Initialization range="0-782" />
        </SegmentBase>
      </Representation>
    </AdaptationSet>
  </Period>
</MPD>
```

Figure 2: MPD Example

The MPEG-DASH provisions call manifest files MPDs (Media Presentation Descriptions). Next we will show an example of an MPD (Figure 2).

You can see that it is structured using XML. The multiple Representation entries defined under AdaptationSet assume that clients will perform switching and playback dynamically. This example demonstrates that video streams of 4 Mbps and 2.4 Mbps have been prepared on the server side, enabling clients to select the optimal stream for their environment (bandwidth and CPU consumption, etc.).

The HLS, MPEG-DASH, Smooth Streaming, and HTTP Dynamic Streaming formats that are now widely used benefit greatly from adopting HTTP/1.1 as the delivery protocol. HTTP/1.1 was already quite prevalent, so it offered better scalability than using dedicated streaming protocols. The streaming formats carried over HTTP shared the idea of performing segmentation of the data and delivering small chunks from the server to the client.

However, during their design HLS, MPEG-DASH, Smooth Streaming, and HDS adopted manifest files and container formats individually, resulting in each featuring different combinations (Table 2). These circumstances caused confusion during content production and at CDN providers. When you want to support a wide range of clients, you need to create manifest files and container formats for all of them. If you are using HLS, MPEG-DASH, Smooth Streaming, and HDS, you have to create and manage four types of file. This increases the workload of production sites, and requires four times as

Name	Specification	Manifest File	Container Format	Standardization
HTTP Live Streaming	Apple	m3u8 An independent extension of the m3u standard	Segmented MPEG2-TS Latest version supports MP4	Not standardized However, it is used extensively even outside Apple
MPEG-DASH	ISO/IEC	MPD Written in XML	MP4, MPEG2-TS (MP4 is used most often)	International standard Detailed definition is carried out in various places (the MPEG Industry Forum, etc.)
Smooth Streaming	Microsoft	isml	MP4	None
HTTP Dynamic Streaming	Adobe	f4m Written in XML	f4f	None

Table 2: Common Streaming Formats

much storage to be prepared. It can also lead to decreased cache efficiency in the distribution systems of CDN providers. On top of that, the decision of which systems to implement on playback devices is a tricky issue. The adoption of Smooth Streaming and HDS has actually been dropping, with the HLS and MPEG-DASH systems chosen in most cases, but this doesn't change the fact that there are glaring inefficiencies.

Extensive structural changes have been made in the latest version 20 of HLS. It now supports MP4. Support for fragmented MP4 was added to the fragmented MPEG2-TS specified up until now. The new Packet Audio and WebVTT multimedia formats have also been added. Fragmented MP4 (fMP4) is a format also standardized by ISO/IEC that refers to a series of data sequenced into multiple files.

Support for this has created the possibility of being able to merge HLS and MPEG-DASH media libraries. If you had a single type of media library prepared using fragmented MP4, it could be distributed to multiple systems. Also from a content delivery business perspective it would provide for more efficient cache operation. This is because for both HLS and MPEG-DASH the most data-heavy part is the fragmented MP4 file group storing the video data.

Moves such as this by Apple correspond to the trend towards unifying streaming formats. In line with this, the Common Media Application Format (CMAF) has been proposed. The original draft was proposed by Apple and Microsoft, and it has been discussed by MPEG (The Moving Picture Experts Group). It is more realistic for this kind of streaming format standardization work to be carried out by MPEG rather than IETF. After all, MPEG is a community where experts in areas such as container formats and codecs gather.

The following text is given as the subtitle for the CMAF standard proposal.

Media Application Format optimized for large scale delivery of a single encrypted, adaptable multimedia presentation to a wide range of devices; compatible with a variety of adaptive streaming, broadcast, download, and storage delivery methods

It appears to comprehensively cover the technology involved, but CMAF carries over the results achieved using HLS and MPEG-DASH. To put it another way, you could say it covers all the streaming format issues that have cropped up to date.

Name	Manifest File	Containers
HTTP Live Streaming	m3u8	CMAF <CMAF internal structure> + CMAF Presentation + CMAF Selection Set (Can accommodate multiple different elements in the same content; camera footage, codecs, multi-lingual content, etc.) + CMAF Switching Set (Can accommodate multiple versions of the same content using different encoding formats) + CMAF Track + CMAF Fragments + CMAF Header
MPEG-DASH	MPD	

Table 3: Relationship Between HLS, MPEG-DASH, and CMAF

CMAF does not define manifest files, players, or delivery protocols. HLS and MPEG-DASH can be called up from the manifest files that each have.

CMAF has a hierarchical structure, and it is compatible with multiple languages and different bitrates, etc. (Table 3).

If media library creation is unified using CMAF, it will drastically reduce the work associated with creating content. Although it doesn't provide any direct benefits to users, it seems to be accepted that this is a necessary move to further popularize streaming media.

The significance of CMAF will come down to whether companies like Apple or Microsoft can set a major trend in the industry, instead of following their own paths. With this kind of framework, and the interactions between people cultivated through related discussions, a better response should be possible the next time a new challenge arises. You could say that innovation is more likely to originate from aiming to make further improvements based on what we have in place now, rather than from a unified protocol or format.

Finally, allow me to point out two challenges we face going forward.

The first is finding a way to reduce the time it takes for video viewing to start. Current live streaming usually lags behind real time by around 30 seconds or more. There are multiple reasons for this, such as the time it takes to encode and the network upload time, as well as the number of segments to buffer before initial playback. These aren't things that can be changed easily by configuring the encoder or delivery server, so they are hard to control on the stream generation side. But there is demand from people who would like to play back live event video in as close to real time as possible. This issue is widely recognized in the industry, and at some point in the future it is likely a proposal for improvement will be made.

The second challenge is offline playback. For the viewing of video on mobile devices, the download caps set by telecom carriers are a particularly large hurdle. Many people also shy away from video streaming on mobile devices. To cope with this, there is technology that downloads video to a mobile device while connected to Wi-Fi, enabling it to be viewed offline as well.

Google has added a system for offline playback to its YouTube app. You can save certain videos, and watch them offline for up to 48 hours. This function is not unlocked for all users, and is mainly enabled in countries where communications infrastructure is still developing. HLS has also put together a system for offline playback. It requires the creation of corresponding codes in an iOS app.

In both cases support is provided in some apps and systems, but when the technology begins to take off in earnest standardization will be necessary. It remains to be seen whether the industry can rise to the challenge and follow through with moves such as these.

Author:
Bunji Yamamoto
Mr. Yamamoto is a Senior Engineer in the Content Delivery and Media Business Department of the Corporate Planning Division, IIJ. He joined IIJ Media Communications in 1995 and has worked at IIJ since 2005. He is mainly involved with the development of streaming technology. Among his contributions to development of the market is the organization of the Streams-JP Mailing List, which discusses this technology.

Wikipedia as a Language Resource

3.1 Introduction

I imagine that many people make use of Wikipedia, the world's largest online encyclopedia. Although issues with its reliability have been pointed out, I use it to look things up on a daily basis due to the vast amount of information that can be obtained from the greatest encyclopedia of all time in terms of both in quality and quantity. Also, as I am involved in researching Wikipedia as social data, I am extremely interested in its pageview count statistics.

There seem to be many researchers who treat Wikipedia as a subject of their research. For example, DBpedia*[1] and its Japanese version*[2] are endeavors that release data they have extracted from Wikipedia and converted to LOD (Linked Open Data) format files, which are apparently used for research into the Semantic Web among other things. Wikipedia is also recognized in the natural language processing research field as a language resource storing vast quantities of sample sentences, and a variety of methods for utilizing it have been proposed. Consequently, in this report I would like to step away from my own research and discuss the theme of Wikipedia as a language resource.

3.2 Writing "Unix Archaeology"

One of the reasons I became interested in Wikipedia as a language resource goes back to when I published a book called "Unix Archaeology*[3]" in April of this year. This book introduces historical facts related to UNIX based on various documents, and if pressed I'd have to categorize it as a history book.

This book is based on a collection of 26 articles that were published in the monthly magazine "UNIX USER" between 2003 and 2005, now organized and re-edited in book form. One of its characteristics is that it was written purely using reference material collected through Google Search, without any interviews or trips to the library. I don't know about today, but back then only a few years had passed since Google's search engine has been released in 1998, so this was a rather reckless writing style.

That said, at the time I had faith in this approach (which thinking back now was based on extremely flimsy grounds). This relates to an occupation that appeared in the 1980s called a "database search engineer," also commonly called a "searcher." A TV program covering this job discussed the example of a novelist and searcher teaming up to write a new novel, in which the searcher told the novelist that they didn't need to gather any reference material at all to write the story, further boasting that if the novelist told them the information they needed, it could all be pulled out of a database. As a rookie who had just entered employment, at the time I was very skeptical that such a thing could be possible. When I was asked to write a series of articles more than ten years later, I remembered this episode and thought this approach may be possible for me as well, with Google's search engine now at my disposal.

In reality, there was a delay of about six months between the time I accepted the writing request and when the first article was published. The editorial department expected me to use this time to accumulate several articles worth of material, but I spent most of the time looking into keyword selection and sorting order, or in other words search patterns, to locate the material I wanted using Google. As a result, six months later I had only completed the first and second articles, and afterwards I had to pay fairly stiff reparations. This is how I ended up settling on a slightly eccentric writing style, adding to my manuscript under the topic of "delving as deep as possible" with search engine in hand.

3.3 Drilldown Searches

Now there is a term called "drilldown" that perfectly expresses "delving as deep as possible." According to Wikipedia's "Drill down"*[4] entry, this is defined as to move from one place to another, from information to detailed data, by focusing in on something."

*[1] DBpedia (http://wiki.dbpedia.org/).

*[2] DBpedia in Japanese (http://ja.dbpedia.org/).

*[3] Unix Archaeology (http://asciidwango.jp/post/142281038535/unix%E8%80%83%E5%8F%A4%E5%AD%A6-truth-of-the-legend) (in Japanese).

*[4] Drill down (https://en.wikipedia.org/wiki/Drill_down).

I imagine for me the illustrations given on this page regarding online users and web-surfers are the closest match, but my focus was on digging up documents that weren't well known regarding the history of computers, which I'd say was a very special case.

For example:

> In 7th Edition UNIX, the kernel was also completely rewritten in C. This is a relatively well-known fact. From a number of documents left by Dennis Ritchie (papers, lecture materials, and interviews), we can confirm that this task was undertaken by Ken Thompson, Dennis Ritchie, and Steve Johnson, and that Steve Johnson was involved because he was the developer of the Portable C Compiler (PCC). So, are there any documents in which Steve Johnson discusses the development task himself?

Alternatively:

> Regarding the development of BSD UNIX, when 4BSD was released there was criticism mainly centered around the fact that it performed worse than 3BSD, and to deal with this the UCB's Computer Systems Research Group released 4.1BSD with performance tuning. This is brought up in many documents discussing the history of UNIX. However, a document authored by Kirk McKusick revealed the fact that this tuning was done by Bill Joy, who was infuriated by an extremely combative critical article by a person called David Kashtan. So, are there any documents that discuss the content of this critical article by Kashtan, as well as the counterargument by Bill Joy?

When attempting drilldowns related to historical facts like these, searches are extremely difficult to carry out. Repeating the cycle of reading the relevant parts of documents you've found and selecting keywords from within these to search for new documents, and then writing a six to ten page article every month, is a task that requires a lot of energy and concentration. Rather than the literary work of a novelist, I'd liken it more to a journalist writing an article.

In fact, when it was decided the articles would be published in book form 12 years after serialization ended, it was necessary to do a considerable amount of rewrites and compose new text, so I tried to think back to that time and work in the same way again. However, drilling down to historical facts had become an unbearably hard task for someone of my age. I began to wonder whether at least part of this work could be handled using computers.

3.4 Named Entity Recognition (NER)

Since then, I have continued to ponder the computerization of historical fact drilldowns. This was initially more about making preliminary arrangements for my writing activity, rather than for research. Of course, there was also the ulterior motive of it lowering the hurdle for receiving writing requests (laughs).

To build software that computerizes this process, or in other words mimics the way I search for documents, it is first necessary to clarify how I had been locating them. The most important knowledge regarding bibliographic searches and the extraction of factual information using a search engine that I picked up through writing the aforementioned articles was to focus on proper nouns. It may seem conventional, but this corresponds to the names of people and organizations, as well as computer model numbers and nicknames in the case of historical facts regarding computers. By collecting as many of these proper nouns as possible and specifying combinations of them as search keywords, it was often possible to greatly narrow down search results in the way that I desired. In light of this, I set to work on finding a technique for extracting proper nouns from English text.

Not surprisingly, this involves the field of research into natural language processing, and straight after I began looking I realized two things.

(1) Statistical natural language processing is mainly used now
(2) The aspects of natural language processing being researched are considerably different between Japan and English-speaking countries.

Regarding the first point, this has completely different objectives and goals to my current research, but in both cases research is carried out using statistics-based techniques. In particular, the basic technology of data analysis is also a subject I research, and looking just at this basic technology we can see there are many common points.

Regarding the second point, until then I had in some respects assumed that when the subject of natural language processing came up, it implicitly referred to Japanese. I thought the main theme was areas of research closely associated with Japanese, such as morphological analysis and machine translation. But I now get the impression that English natural language processing (understandably) incorporates very different themes.

The technique of extracting proper nouns from English text I sought turned out to be one of the main themes of natural language processing in English--a process called Named Entity Recognition (NER).

3.5 NER in the Natural Language Toolkit (NLTK)

Today, NER is incorporated into the Natural Language Toolkit (NLTK), so as shown in Figure 1 you can attempt to extract named entities comparatively easily using Python script.

When given an English text file name, this script executes NER and produces the kind of output shown in Figure 2.

It displays a word, part of speech, and NE tag on each line. When the first letter of the NE tag is “B-” it means the word is the first part of a named entity, whereas “I-” is a continuation of one. PERSON and ORGANIZATION are self-explanatory, but GSP stands for “Geo-Socio-Political group.” We can see that “Hillary Clinton” and “Donald Trump” are recognized as people’s names.

Next, as an example of a slightly larger body of text, I tried extracting people’s names from the “Twenty Years of Berkeley Unix” book by Marshall Kirk McKusick, who also aided me in writing Unix Archaeology (Figure 3).

There is some misrecognition, but in addition to well-known names such as Ken Thompson, Dennis Ritchie, and Bill Joy, BSD UNIX development staff members other than Kirk McKusick, such as Ozalp Babaoglu, Sam Leffler, Mike Karels, and Keith Bostic, have also been correctly recognized. This demonstrates that recent research into statistical natural language processing has made it possible to extract named entities with a certain degree of accuracy, without any special tuning for a particular purpose (such as topics related to the history of UNIX).

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import nltk
import sys

if __name__ == "__main__":

    param = sys.argv

    args = param[1]

    with open(args, 'r') as f:
        sample = f.read()

    sentences = nltk.sent_tokenize(sample)
    tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in sentenc\es]
    tagged_sentences = [nltk.pos_tag(sentence) for sentence in tokenized_senten\ces]

# for NLTK 2.7
#   chunked_sentences = nltk.batch_ne_chunk(tagged_sentences)

# for NLTK 3.0
    chunked_sentences = nltk.ne_chunk_sents(tagged_sentences)

    entity_names = []
    for tree in chunked_sentences:
        print nltk.chunk.tree2conllstr(tree)
```

Figure 1: named_entity_recognition.py

3.6 Statistical Natural Language Processing and Corpora

It would probably be necessary to study research into statistical natural language processing to learn the inner workings of how it is possible to achieve decent extraction accuracy without any special tuning. However, “corpus” is a term that crops up frequently when reading reference material.

According to the English Wikipedia, in linguistics a corpus is “a large and structured set of texts,” which is “used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.” To an outsider like me this explanation is hard to comprehend by itself, but if you remember Zipf’s law many people are familiar with that states the percentage of the total accounted for by a word with an appearance frequency of rank k is proportional to 1/k, you may find it a bit easier to visualize the mysterious relationship between our writings and statistics. In other words, when people write compositions they tend to do it with an unconscious statistical bias. I imagine this is why statistical natural language processing works so well.

I actually experienced this behavior with a statistical bias viscerally through writing “Unix Archaeology.” For example, Dennis Ritchie published a large number of documents that discuss the development background of UNIX on his home

```
$ cat NHK-short.txt
A US poll shows that Democratic presidential nominee Hillary Clinton's lead
over her Republican rival, Donald Trump, has shrunk to 3 percentage points.
$ ./named_entity_recognition.py NHK-short.txt
A DT O
US NNP B-GSP
poll NN O
shows VBZ O
that IN O
Democratic JJ B-ORGANIZATION
presidential JJ O
nominee NN O
Hillary NNP B-PERSON
Clinton NNP I-PERSON
's POS O
lead NN O
over IN O
her PRP$ O
Republican JJ B-ORGANIZATION
rival NN O
, , O
Donald NNP B-PERSON
Trump NNP I-PERSON
, , O
has VBZ O
shrunk NN O
to TO O
3 CD O
percentage NN O
points NNS O
. . O
$
```

Figure 2: Results of Executing NLTK’s Named Entity Recognition on a Sample English Passage

```
Alan Nemeth
Babaoglu
Beranek
Berkeley
Berkeley Software Design
Berkeley Software Distribution
Berkeley Unix
Bert Halstead
Bill Jolitz
Bill Joy
Bob Baker
Bob Fabry
Bob Guffy
Bob Kridle
Bostic
Casey Leedom
Chuck Haley
DARPA
Dan Lynch
David
Dennis Ritchie
Dickinson R. Debevoise
District Judge
Domenico Ferrari
Duane Adams
Eugene Wong
Fabry
Fateman
Ferrari
Freely Redistributable Marshall Kirk McKusick Early History Ken Thompson
Haley
Hibler
Jeff Schriebman
Jerry Popek
Jim Kulp
John Reiser
Jolitz
Joy
Karels
Keith
Keith Bostic
Keith Lantz
Keith Standiford
Ken Thompson
Laura Tong
Leffler
Linux
Lucasfilm
Math
Michael Stonebraker
Mike
Mike Karels
Mike Muuse
Networking Release
Ozalp Babaoglu
Pascal
Pauline
Peter Kessler
Professor Domenico Ferrari
Professor Richard Fateman
Professors Michael Stonebraker
Ray Noorda
Rick Macklem
Rick Rashid
Rob Gurwitz
Robert Elz
Sam Leffler
Schriebman
Schwartz
Statistics
Support Meanwhile
Susan Graham
System
System III
System Manual
System V
Tahoe
Thompson
Tom London
Unix
Unix Early
Utah
```

Figure 3: Persons Mentioned in “Twenty Years of Berkeley Unix”

page*5 This came in handy while I was writing my book, but upon reading a number of sentences over and over I often noticed identical turns of phrase used in multiple different places. In short, it seems wording that readers may interpret as a habit of the writer can be recognized as a statistical bias through statistical text analysis.

Incidentally, Dennis Ritchie's most noticeable habit was to write the name of the developer of PCC as "Steve Johnson." Ritchie's writings all use this name, but the developer's real name is Stephen Johnson, and in his own papers and Wikipedia entry the spelling Stephen is used. I'm not sure whether this is a misunderstanding by Ritchie, or if the developer was known by the nickname Steve at Bell Labs, but either way it baffled me a lot when I was trying to follow up facts and found no documents at all no matter what criteria I searched for.

It seems that statistical natural language processing is a form of research that obtains a variety of new knowledge from the statistical bias in text written by people.

3.7 Creating a Named Entity Corpus from Wikipedia

Let us go back to discussing NLTK. NER in NLTK is outlined in "Extracting Information from Text" in the seventh chapter of "Natural Language Processing with Python," and its implementation is the "ACE Named Entity Chunker (Maximum entropy)" listed first in the NLTK Corpora*6.

This machine learning model called "Maximum entropy" (there is an explanation under "6.6 Maximum Entropy Classifiers" in the same book) uses an NE corpus with named entity tags. The creation of an NE corpus with named entity tags such as "PERSON" or "ORGANIZATION" is an extremely time consuming task that must be done manually, and for this reason almost no NE corpora are distributed for free. In fact, for NLTK as well only the chunker trained on the Automatic Content Extraction (ACE) corpus is distributed as a pickle file. The corpus itself is not included.

There have been efforts to automatically generate this NE corpus using Wikipedia page data. The paper entitled "Transforming Wikipedia into Named Entity Training Data*7" proposes the use of Wikipedia to create named entity tagged corpora.

The terminology and names found in Wikipedia articles are often linked to other related articles. The basic idea of this paper is to convert these links between articles into named entity tags. For example, in an article about the character "James Bond" in the novel "Thunderball" by "Ian Fleming," we can expect each named entity to have mutual links set up between them. The Ian Fleming article would no doubt indicate that this named entity is a person, and the Thunderball article would show that it is a novel. In other words, it is possible to tag the original article automatically by tracing the links (Figure 4).

The paper states it is possible to create a massive corpus by extracting millions of articles from Wikipedia for NER training. The following four steps are given for this process.

1. Classify all articles into entity classes
2. Split Wikipedia articles into sentences
3. Label named entities according to link targets
4. Select sentences for inclusion in a corpus

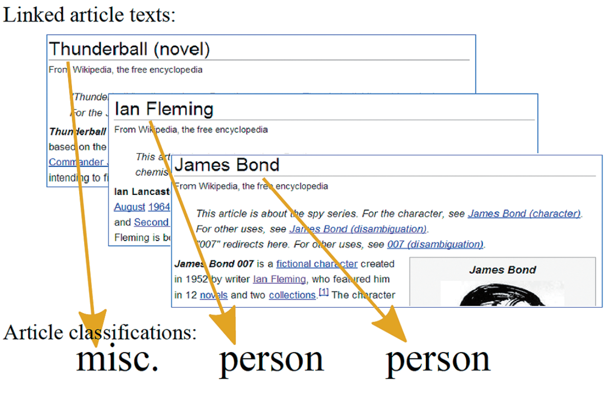

Linked article texts:
Thunderball (novel)
From Wikipedia, the free encyclopedia
Ian Fleming
From Wikipedia, the free encyclopedia
James Bond
From Wikipedia, the free encyclopedia
This article is about the spy series. For the character, see James Bond (character).
For other uses, see James Bond (disambiguation).
"007" redirects here. For other uses, see 007 (disambiguation).
James Bond 007 is a fictional character created in 1952 by writer Ian Fleming, who featured him in 12 novels and two collections.[1] The character
James Bond
Article classifications:
misc.
person
person

Figure 4: Deriving Training Sentences from Wikipedia Text

[*5] Dennis Ritchie (https://www.bell-labs.com/usr/dmr/www/).

[*6] NLTK Corpora (http://www.nltk.org/nltk_data/).

[*7] Transforming Wikipedia into Named Entity Training Data (https://www.aclweb.org/anthology/U/U08/U08-1016.pdf).

Using this procedure, in the paper an attempt is made to create an NE corpus based on the standard CoNLL categories of entity class (LOC, ORG, PER, and MISC).

Applying this procedure would clearly be a big data process in the case of Wikipedia, which currently has over five million articles in English and over a million articles in Japanese. In particular, the bootstrapping approach to classification indicated in the paper, which involves some manual work to identify classes, is a tricky issue to deal with when you want to create a corpus by machine alone.

3.8 Conclusion

Unexpectedly, I get the sense that the drilldown searches for historical documentation that I imagined would progress quite far if I applied the findings of the latest natural language processing research. I am also a little surprised that the basic technology developed for the analysis of Wikipedia that I am involved with in my research could also be used for this initiative.

The task of extracting the names of people and systems from documents could be achieved using NER, but to improve the accuracy of this it will be essential to obtain NE corpora and use these for training. It was good to learn that the Wikipedia we utilize day-to-day can be used as a resource for creating our own NE corpora. The remaining issue is figuring out how to pick up articles on people from the English version of Wikipedia, which has over five million entries.

I don’t see this issue as a big problem for the history book writing I am engaged in right now. This is because the writing of a history book is nothing but digging up information on eras, people, and events based on a theme the writer determines. The task of collecting articles from the English Wikipedia on people associated with the theme I am writing about is a common daily practice for me. Reporters that write non-fiction or news stories and social science researchers have thorough knowledge of named entities other than people, including their classification. I believe considering ways to gather and share their collective wisdom would be the quickest way to resolve problems.

That sums up how an unforeseen situation led to me learning about the current state of natural language processing research. I think it will be possible to apply the results of this to the analysis of social data that I am handling in my research right now as well. One example would be “factorial analysis.” Performing time series analysis of social data enables you to detect sudden fluctuations or bursts, but to investigate the root cause of these you need to trace the social trends that were in effect at the time. I had considered collecting new stories and other data using a search engine, and I think the methods I have introduced in this report could be used to determine the search criteria for achieving this.

Author:
Akito Fujita
Chief Architect, Business Strategy and Development Center, IIJ Innovation Institute Inc. (IIJ-II). Mr. Fujita joined IIJ in 2008.
He is engaged in the research and development of cloud computing technology utilizing knowledge gained through structured overlay research.

December 2016 Vol.33

Internet Initiative Japan

About Internet Initiative Japan Inc. (IIJ)

IIJ was established in 1992, mainly by a group of engineers who had been involved in research and development activities related to the Internet, under the concept of promoting the widespread use of the Internet in Japan.

IIJ currently operates one of the largest Internet backbones in Japan, manages Internet infrastructures, and provides comprehensive high-quality system environments (including Internet access, systems integration, and outsourcing services, etc.) to high-end business users including the government and other public offices and financial institutions.

In addition, IIJ actively shares knowledge accumulated through service development and Internet backbone operation, and is making efforts to expand the Internet used as a social infrastructure.

Internet Initiative Japan Inc.

Address: Iidabashi Grand Bloom, 2-10-2 Fujimi, Chiyoda-ku,
Tokyo 102-0071, Japan
Email: info@iij.ad.jp URL: http://www.iij.ad.jp/en/