IIR

Infrastructure Security

Machine Learning and Security

Broadband Traffic Report

Comparing Broadband and Mobile Traffic

Technology Trends

Social Big Data

IIJ

Internet Initiative Japan

Internet Infrastructure Review
August 2015 Vol.28

Executive Summary

My name is Yoshikazu Yamai, Director of the Service Operation Division, and from this report I will be taking over the Executive Summary from Mr. Asaba. I look forward to tackling this role.

The incident in which personal information leaked from the Japan Pension Service is still fresh in our minds. I believe this incident boosted recognition of information security within society, and further reinforced the perception that security measures are necessary for the protection of not only infrastructure, but also individuals and organizations when using the Internet. As a builder and operator of infrastructure, our aim is to provide it in the form of services, but it is clear that we are at risk of being outpaced by the threats we face. These days, my hope is that security operations involving coordination between providers and users will increase the momentum of moves to protect the Internet.

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, "Infrastructure Security," we discuss and analyze incidents occurring on the Internet, such as hacktivism-based attacks and unauthorized access related to social media. In our focused research, we examine machine learning and security, look at the Angler Exploit Kit that has been thriving lately, and shed light on the actual use of ID management technology.

In Chapter 2, "Broadband Traffic Report," we analyze the recent state of Internet traffic. We also discuss the mobile traffic that has been growing at a steady rate for the past few years, looking at traffic on IIJ's personal and enterprise mobile services, and comparing it with broadband traffic. Delving into mobile trends provides fascinating insight.

In Chapter 3, "Technology Trends," we cover big data. Concepts such as IoT and CPS have been in the limelight a lot lately, but in the world of social networking, these make up just one slice of the huge variety of data transmitted. The endless flow of tweets, and collective wisdom in social form found at Wikipedia, are also important sources of big data. This is where we showcase the research we are working on in our research lab.

It could be said that Internet infrastructure is changing from mere physical infrastructure to infrastructure that also serves as a social platform. In this report we will continue to review the situation from this broad perspective, through the lens of technology.

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
Upon joining IIJ in June 1999, Mr. Yamai 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.
Since then he has managed IIJ's Service Operation Division as Director.
He also heads IIJ's data center operations, and he played a key role in the establishment of the container-based "Matsue Data Center Park," which was the first in Japan to use outside-air cooling.

Machine Learning and Security

1.1 Introduction

This report summarizes incidents to which IIJ responded, based on general information obtained by IIJ itself related to the stable operation of the Internet, information from observations of incidents, information acquired through our services, and information obtained from companies and organizations with which IIJ has cooperative relationships. This volume covers the period of time from April 1 through June 30, 2015. In this period a number of hacktivism-based attacks were once again carried out by Anonymous and other groups, and there was a rash of attacks including SNS account hijackings and website defacements. There were also many incidents of malware infection through targeted attacks, and it was identified that the personal information of as many as 1.25 million people may have leaked in incidents affecting the Japan Pension Service. Information leaks caused by unauthorized access also continue to occur, including an incident in which the U.S. Office of Personnel Management was breached, leading to the leak of information on roughly 4 million employees. These examples show that many security-related incidents continue to occur on the Internet.

1.2 Incident Summary

Here, we discuss the IIJ handling and response to incidents that occurred between April 1 and June 30, 2015. Figure 1 shows the distribution of incidents handled during this period[1].

■ **The Activities of Anonymous and Other Hacktivists**
Attacks by hacktivists such as Anonymous continued during this period. DDoS attacks and information leaks occurred at government-related and corporate sites in a large number of countries stemming from a variety of incidents and causes.

Individuals and organizations thought to be associated with ISIL or sympathetic to its principles carried out website defacements and SNS account hijackings around the world. In April, a French TV station was targeted in a large-scale attack, interrupting the broadcast of programs. At the same time as the attack that compromised the station's internal system[2], its SNS accounts such as Facebook and Twitter were hijacked and unauthorized posts made. In opposition to this, Anonymous continued publishing lists of websites and SNS accounts thought to be connected to ISIL, and encouraging their suspension or deletion (OpISIS). There was also ongoing confusion, including a group that had been active as a Palestinian branch of Anonymous being targeted in attacks by other Anonymous factions, who accused them of supporting ISIL.

In protest against airstrikes in Yemen carried out by the government of Saudi Arabia, among other reasons, there were DDoS attacks on Saudi Arabian newspaper companies, and leaks of confidential information caused by server compromises targeting a number of government agencies (OpSaudi). Additionally, in opposition to the bombing of Gaza
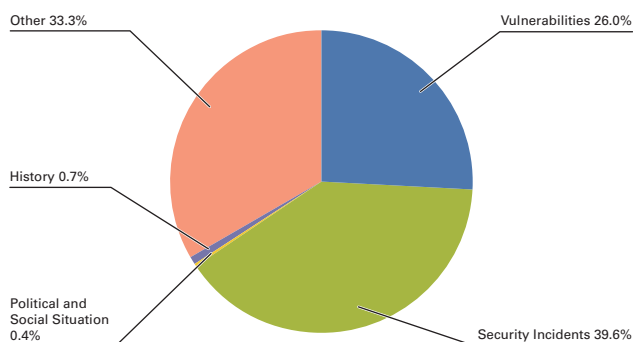

Figure 1: Incident Ratio by Category (April 1 to June 30, 2015)

Other 33.3%
Vulnerabilities 26.0%
History 0.7%
Political and Social Situation 0.4%
Security Incidents 39.6%

---

*1   Incidents discussed in this report are categorized as vulnerabilities, political and social situations, history, security incidents or other.
Vulnerabilities: Responses to vulnerabilities associated with network equipment, server equipment or software commonly used over the Internet or in user environments.
Political and Social Situations: Responses to incidents related to domestic and foreign circumstances and international events such as international conferences attended by VIPs and attacks originating in international disputes.
History: Historically significant dates; warning/alarms, detection of incidents, measures taken in response, etc., related to attacks in connection with a past historical fact.
Security Incidents: Unexpected incidents and related responses such as wide propagation of network worms and other malware; DDoS attacks against certain websites.
Other: Security-related information, and incidents not directly associated with security problems, including highly concentrated traffic associated with a notable event.

*2   It has been pointed out that a RAT may have been used in these attacks. Trend Micro, "Kjw0rm VBS Malware Tied To Attacks on French TV Station TV5Monde" (http://blog.trendmicro.com/trendlabs-security-intelligence/vbs-malware-tied-to-media-attacks/).

by Israel, there were ongoing attacks resulting in the leak of personal and credit card information due to incidents of unauthorized access at a number of Israeli companies (OpIsrael), etc. As this demonstrates, Internet-based attacks also continue to occur in Middle-Eastern countries due to conflicts and diplomatic circumstances.

In Canada, there were a number of DDoS attacks as well as website defacements and leaks of internal information due to server compromises targeting a number of government agencies (OpC51). These were related to the passing of an Anti-terrorism Act in June that significantly expands and strengthens the authority of the Canadian Security Intelligence Service, triggering privacy concerns. In India, there were DDoS attacks and also leaks of account information caused by unauthorized access targeting a number of government agencies, including the Telecom Regulatory Authority of India, in protest against government moves to bolster Internet regulations. In Japan, there were reports of incidents of unauthorized access in May thought to be carried out by Anonymous to protest against the commercial use of marine animals such as whales and dolphins (OpSeaWorld). These led to the leak of registrant information such as email addresses from the Japanese Association of Zoos and Aquariums (JAZA).

Unknown attackers claiming affiliation with the Syrian Electronic Army continued to hijack accounts and deface websites, with affected organizations including U.S. newspaper company the Washington Post, as well as the U.S. Army. Other attacks by hacktivists such as Anonymous continued on government and government-related sites around the world, such as in Italy, Brazil, and China. There were also ongoing hijacking incidents targeting prominent SNS accounts, such as those of government institutions.

■ **Vulnerabilities and their Handling**

During this period fixes were released for Microsoft's Windows[3][4][5][6][7], Internet Explorer[8][9][10], and Office[11]. Fixes were also released for Adobe Systems' Adobe Flash Player. A quarterly update was provided for Oracle's Java SE, fixing many vulnerabilities. Additionally, because support for Java 7 is scheduled to end as of this update, migration to Java 8 is recommended. Several of these vulnerabilities were exploited in the wild before patches were released.

Regarding server applications, a quarterly update was released for a number of Oracle products, including the Oracle database server, fixing many vulnerabilities. An XSS vulnerability in multiple versions of the WordPress CMS that could allow arbitrary code execution was discovered and fixed. An XSS vulnerability was also discovered and fixed in a WordPress plug-in[12]. A vulnerability that could allow remote code execution due to insufficient checking of entered values was discovered in fixed in the Movable Type CMS[13].

---

*3 "Microsoft Security Bulletin MS15-034 - Critical: Vulnerability in HTTP.sys Could Allow Remote Code Execution (3042553)" (https://technet.microsoft.com/en-us/library/security/ms15-034.aspx).

*4 "Microsoft Security Bulletin MS15-035 - Critical: Vulnerability in Microsoft Graphics Component Could Allow Remote Code Execution (3046306)" (https://technet.microsoft.com/en-us/library/security/ms15-035.aspx).

*5 "Microsoft Security Bulletin MS15-044 - Critical: Vulnerabilities in Microsoft Font Drivers Could Allow Remote Code Execution (3057110)" (https://technet.microsoft.com/en-us/library/security/ms15-044.aspx).

*6 "Microsoft Security Bulletin MS15-045 - Critical: Vulnerability in Windows Journal Could Allow Remote Code Execution (3046002)" (https://technet.microsoft.com/en-us/library/security/ms15-045.aspx).

*7 "Microsoft Security Bulletin MS15-057 - Critical: Vulnerability in Windows Media Player Could Allow Remote Code Execution (3033890)" (https://technet.microsoft.com/en-us/library/security/ms15-057.aspx).

*8 "Microsoft Security Bulletin MS15-032 - Critical: Cumulative Security Update for Internet Explorer (3038314)" (https://technet.microsoft.com/en-us/library/security/ms15-032.aspx).

*9 "Microsoft Security Bulletin MS15-043 - Critical: Cumulative Security Update for Internet Explorer (3049563)" (https://technet.microsoft.com/en-us/library/security/ms15-043.aspx).

*10 "Microsoft Security Bulletin MS15-056 - Critical: Cumulative Security Update for Internet Explorer (3058515)" (https://technet.microsoft.com/en-us/library/security/ms15-056.aspx).

*11 "Microsoft Security Bulletin MS15-033 - Critical: Vulnerabilities in Microsoft Office Could Allow Remote Code Execution (3048019)" (https://technet.microsoft.com/en-us/library/security/ms15-033.aspx).

*12 See the following blog post from U.S. security firm Sucuri for more information. "Security Advisory: Persistent XSS in WP-Super-Cache" (https://blog.sucuri.net/2015/04/security-advisory-persistent-xss-in-wp-super-cache.html).

*13 Six Apart, "MOVABLE TYPE 6.0.8 AND 5.2.13 RELEASED TO CLOSE SECURITY VULNERABILITY" (https://movabletype.org/news/2015/04/movable_type_608_and_5213_released_to_close_security_vulnera.html).

April Incidents

1	**S** **1st: A number of reports were published indicating that large-scale DDoS attacks on GitHub that occurred in March were related to transmissions to China.**
2	See the following NETRESEC announcement for more information on these attacks. "China's Man-on-the-Side Attack on GitHub" (http://www.netresec.com/?page=Blog&month=2015-03&post=China%27s-Man-on-the-Side-Attack-on-GitHub). Security researcher Robert Graham has also looked into where the MITM attacks were being carried out in the following report. "Pin-pointing China's attack against GitHub" (http://blog.erratasec.com/2015/04/pin-pointing-chinas-attack-against.html#.VahhVfntnK4).
3	
4	**S** **8th: It was announced that member information registered on a service site for stockholders used by a number of listed companies may have leaked.**
5	This incident was uncovered when investor solicitation was carried out by email and phone based on the information that had leaked. Upon later investigation it was announced that this was an inside job rather than being caused by unauthorized access.
6	**O** **8th: The FBI issued an alert due to a series of website defacements exploiting vulnerabilities in WordPress plug-ins that were carried out by someone claiming affiliation with ISIL.**
7	The Internet Crime Complaint Center (IC3), "ISIL DEFACEMENTS EXPLOITING WORDPRESS VULNERABILITIES" (http://www.ic3.gov/media/2015/150407-1.aspx).
8	
9	**S** **9th: French TV network the TV5MONDE group was accessed without authorization by someone claiming affiliation with ISIL, leading to damages including the hijacking of websites and SNS site accounts. Other impact from these attacks included the interruption of TV broadcasts for seven hours.**
10	"Press release from April 9th 2015" (https://asia.tv5monde.com/Resources/Articles/Events/2015/04-en/Press-release-from-April-9th-2015?lang=en-US).
11	**S** **10th: The European Police Office (Europol) announced they had taken down the Beebone botnet (also known as AAEH, etc.) together with investigative organizations and security firms in various countries, including the Federal Bureau of Investigations (FBI).**
12	"International Police Operation Targets Polymorphic BEEBONE Botnet" (https://www.europol.europa.eu/content/international-police-operation-targets-polymorphic-beebone-botnet).
13	**S** **10th: The Tokyo Metropolitan Police Department initiated a takedown of malware related to Internet banking in Japan (VAWTRAK), in collaboration with the Ministry of Internal Affairs and Communications, Telecom-ISAC Japan, and private-sector businesses.** "Operation to Disable a Net Banking Virus" (http://www.keishicho.metro.tokyo.jp/haiteku/haiteku/haiteku504.htm) (in Japanese).
14	
15	**S** **14th: INTERPOL announced they had taken down the SIMDA botnet together with investigative organizations and private-sector companies in various countries.**
16	This operation was carried out on April 9 and coordinated by the INTERPOL Global Complex for Innovation (IGCI), which INTERPOL established in Singapore to conduct research and development as well as investigative support regarding cybercrime countermeasures. "INTERPOL coordinates global operation to take down Simda botnet" (http://www.interpol.int/News-and-media/News/2015/N2015-038).
17	**S** **14th: An incident occurred in which the domains of well-known companies such as Google and Yahoo were hijacked due to unauthorized access to Malaysian .my domains by an unknown attacker.**
18	MYNIC, "GOOGLE & YAHOO MALAYSIA DOMAIN BACK TO NORMAL WITHIN 24 HOURS" (https://www.mynic.my/upload_media/press_release.pdf).
19	**V** **15th: Microsoft published their Security Bulletin Summary for April 2015, and released 11 updates, including four critical updates such as MS15-034, as well as seven important updates.**
20	"Microsoft Security Bulletin Summary for April 2015" (https://technet.microsoft.com/library/security/ms15-apr).
21	**V** **15th: A number of vulnerabilities in Adobe Flash Player that could allow arbitrary code execution were discovered and fixed.** "Security updates available for Adobe Flash Player" (http://helpx.adobe.com/security/products/flash-player/apsb15-06.html).
22	**V** **15th: Oracle released their quarterly scheduled update for a number of products including Oracle, fixing a total of 98 vulnerabilities, including 14 in Java SE. Support for Java 7 ended as of this update.** "Oracle Critical Patch Update Advisory - April 2015" (http://www.oracle.com/technetwork/topics/security/cpuapr2015-2365600.html).
23	
24	**O** **15th: The European Commission (EC) announced they had send a Statement of Objections to Google, due to suspicions Google had violated EU antitrust laws by abusing their dominant position as a search service to give their own shopping service preferential treatment in general search results. The commission also stated they would investigate the Android mobile operating system separately.**
25	See the following press release for more information, "Antitrust: Commission sends Statement of Objections to Google on comparison shopping service" (http://europa.eu/rapid/press-release_MEMO-15-4781_en.htm).
26	For information on the Android mobile operating system, see "Antitrust: Commission opens formal investigation against Google in relation to Android mobile operating system" (http://europa.eu/rapid/press-release_MEMO-15-4782_en.htm).
27	**V** **22nd: A number of vulnerabilities in WordPress, including an XSS vulnerability that could allow arbitrary code execution by unauthorized users, were discovered and fixed.**
28	"WordPress 4.1.2 Security Release" (https://wordpress.org/news/2015/04/wordpress-4-1-2/).
29	**V** **28th: An XSS vulnerability in WordPress that could allow arbitrary code execution on a server via comment posting was discovered and fixed.** "WordPress 4.2.1 Security Release" (https://wordpress.org/news/2015/04/wordpress-4-2-1/).
30	

*Dates are in Japan Standard Time

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

In May, a buffer overflow vulnerability (CVE-2015-3456) in QEMU's virtual Floppy Disk Controller was disclosed by U.S. security firm CrowdStrike and fixed. This vulnerability was named VIRTUALIZED ENVIRONMENT NEGLECTED OPERATIONS MANIPULATION (VENOM) by the discoverer. Its impact was widespread, and it garnered a lot of attention because it allowed DoS attacks and arbitrary code execution on a host from the guest side in virtualization software such as Xen or KVM, which are used for cloud services.

Also in May, a vulnerability in the TLS protocol DH key exchange (CVE-2015-4000) that could allow an attacker to intercept or alter communications by forcing a downgrade to a weaker cipher via a MITM attack was discovered and fixed. This vulnerability was given the name Logjam, and like the FREAK vulnerability in SSL/TLS disclosed in March, it generated a lot of news due to its potential to affect many servers and browsers.

■ **Malware Infections and Information Leaks Due to Targeted Attacks**
During the current survey period, there were frequent incidents such as malware infections on PCs within organizations, as well as resulting information leaks. In June, the personal information of 1,010,000 people was leaked due to infection by malware attached to emails at the Japan Pension Service*14. Due to this incident, it was later revealed that similar incidents had been occurring since September last year at a number of companies and organizations, as well as hospitals and universities. In each case, it is known the infection took place via so-called targeted emails that had malware attached, with the sender misrepresented using free email accounts. Upon infection, the malware sent files on the infected PC or shared servers to an external party.

Similarly, there have also been reports of malware infections through email that was misrepresented as being from a multi-function printer*15. Other than these cases, malware infections on internal PCs and resulting information leaks have also occurred at a number of local public bodies. It has been identified that these infections may have been caused by so-called watering hole attacks, when users accessed an altered website.

In May, the IPA published the "Initiative for Cyber Security Information sharing Partnership of Japan (J-CSIP) Annual Activity Report FY2014," which summarized the activities of the Cyber Security Information sharing Partnership of Japan (J-CSIP)*16. In this report, it was mentioned that ongoing attacks thought to have been carried out by the same attacker have been observed over the course of 31 months. As these facts demonstrate, attacks are becoming more diverse through the appearance of new techniques and methods, and the increasing sophistication of the malware used. It is becoming difficult to deal with these threats using a stand-alone countermeasure, such as the introduction of anti-virus software. Efforts are underway to examine countermeasures for attacks like this that cannot be handled using existing security measures. One example is the "System Design Guide for Thwarting Advanced Targeted Email Attacks"*17 published by the IPA in September 2014.

■ **Information Leaks Due to Unauthorized Access**
Information leaks caused by unauthorized access also continue to occur. In May, the website of a U.S. beauty-related company was accessed without authorization, leading to the leak of credit card information and other data. The internal database of a U.S. insurance company was also compromised, causing the leak of personal information for 1.1 million individuals. In June, data related to approximately 100,000 taxpayers leaked from the U.S. Internal Revenue Service (IRS). The U.S. Office of Personnel Management was also accessed without authorization, and data on roughly 4 million federal employees leaked. Also in June, a U.S. password management service was compromised, and users were prompted to change their master password due to the possibility that some authentication information had leaked.

*14 Japan Pension Service, "An apology to Japan Pension Service customers whose personal information was leaked" (http://www.nenkin.go.jp/n/data/service/0000028648uArRENS1eQ.pdf) (in Japanese).

*15 Emails misrepresented as being from a multi-function printer have been known for a long time, but Trend Micro's security blog reported that these incidents spiked sharply in June.
"Japan also affected by emails misrepresented as notifications from multi-function printers that distribute macro-type malware" (http://blog.trendmicro.co.jp/archives/11776) (in Japanese).

*16 An initiative established in 2011 for carrying out early responses and information sharing between member organizations, with a focus on critical infrastructure companies. "Initiative for Cyber Security Information sharing Partnership of Japan (J-CSIP)" (https://www.ipa.go.jp/security/J-CSIP/) (in Japanese).

*17 IPA, "System Design Guide for Thwarting Advanced Targeted Email Attacks" (http://www.ipa.go.jp/files/000035723.pdf).

May Incidents

1	**V** **7th:** A number of vulnerabilities in WordPress, including an XSS vulnerability that could allow arbitrary code execution by unauthorized users, were discovered and fixed. "WordPress 4.2.2 Security and Maintenance Release" (https://wordpress.org/news/2015/05/wordpress-4-2-2/).
2	
3	**O** **12th:** IPA published their SSL/TLS encryption configuration guidelines, to enable encryption settings that take into consideration appropriate security for SSL/TLS server builders and operators. "SSL/TLS Encryption Configuration Guidelines - For a Secure Website (Encryption Configuration Measures)" (http://www.ipa.go.jp/security/vuln/ssl_crypt_config.html) (in Japanese).
4	
5	**V** **13th:** Microsoft published their Security Bulletin Summary for May 2015, and released 13 updates, including three critical updates, MS15-043, MS15-044 and MS15-45, as well as 10 important updates. "Microsoft Security Bulletin Summary for May 2015" (https://technet.microsoft.com/library/security/ms15-may).
6	**V** **13th:** A number of vulnerabilities in Adobe Flash Player that could allow unauthorized termination or arbitrary code execution were discovered and fixed. "APSB15-09: Security updates available for Adobe Flash Player" (https://helpx.adobe.com/security/products/flash-player/apsb15-09.html).
7	
8	**V** **14th:** A vulnerability (CVE-2015-3456) in the QEMU Floppy Disk Controller (FDC) used in Xen and KVM that could allow abnormal server termination or arbitrary code execution was discovered and fixed. This vulnerability was named VENOM by the discoverer. See the following website of the discoverer, CrowdStrike, for more information. "VENOM VIRTUALIZED ENVIRONMENT NEGLECTED OPERATIONS MANIPULATION" (http://venom.crowdstrike.com/).
9	
10	
11	**O** **15th:** IETF published HTTP/2, which was approved in February, as an RFC. "Hypertext Transfer Protocol Version 2 (HTTP/2)" (http://www.rfc-editor.org/rfc/rfc7540.txt).
12	
13	**V** **21st:** A vulnerability in the TLS protocol DH key exchange that could allow communications to be intercepted or altered by forcing selection of weaker encryption through a MITM attack was discovered and fixed. See the following explanation by the discoverer for more details. "The Logjam Attack" (https://weakdh.org/).
14	
15	**S** **22nd:** DDoS attacks accompanied by threats were made on a number of financial institutions and online stores, resulting in systems being difficult to access, etc.
16	**O** **22nd:** The Ministry of Internal Affairs and Communications published its Recommendations on Promotion of Cyber Security Policy coordinated by the Information Security Advisory Board. "'Recommendations on Promotion of Cyber Security Policy' published" (http://www.soumu.go.jp/menu_news/s-news/01ryutsu03_02000093.html) (in Japanese).
17	**O** **22nd:** IPA and JPCERT/CC published the 2015 version of the "Information Security Early Warning Partnership Guidelines," for reference when reporting or receiving notification of information related to vulnerabilities. This relates to the Information Security Early Warning Partnership, an initiative for promoting the trouble-free distribution of vulnerability information and the spread of countermeasures. IPA, "Information Security Early Warning Partnership Guidelines" (http://www.ipa.go.jp/security/ciadr/partnership_guide.html) (in Japanese).
18	
19	
20	**S** **25th:** The National Police Agency issued an alert regarding an increase in scanning activity that targets a utility tool for an online game. This tool has a proxy function implemented, so this scanning activity is thought to be searching for so-called open proxies that can be used externally by a third party. See "Regarding an increase in proxy searches targeting specific ports" (http://www.npa.go.jp/cyberpolice/detect/pdf/20150525.pdf) (in Japanese) for more information.
21	**O** **25th:** The 2nd general meeting of the Cyber Security Strategic Headquarters was held, and a new Cyber Security Strategy (proposed) that determines the basic policy for cyber security measures was decided on. NISC, "Cybersecurity Strategy" (http://www.nisc.go.jp/eng/pdf/cs-strategy-en.pdf).
22	
23	
24	**S** **26th:** The National Police Agency issued an alert regarding an increase in scanning activity targeting vulnerabilities in certain PLC software used in industrial control systems. "Regarding the observation of access targeting vulnerabilities in the PLC used in industrial control systems" (http://www.npa.go.jp/cyberpolice/detect/pdf/20150526.pdf) (in Japanese).
25	**S** **26th:** The JPCERT Coordination Center issued an alert regarding the alteration of website content and resulting occurrence of ransomware infections through redirection to attack sites. "Alert regarding ransomware infections" (https://www.jpcert.or.jp/english/at/2015/at150015.txt).
26	
27	**S** **27th:** The U.S. Internal Revenue Service (IRS) was accessed without authorization, and information such as the social security numbers of around 100,000 taxpayers leaked. "IRS Statement on the 'Get Transcript' Application" (http://www.irs.gov/uac/Newsroom/IRS-Statement-on-the-Get-Transcript-Application).
28	
29	
30	
31	

*Dates are in Japan Standard Time

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

In Japan, there were also incidents of unauthorized access at one of the major ISPs in May, leading to the leak of FTP accounts, etc. In April, there was an incident in which internal documents from a university were published on the Internet due to flawed server configuration. In June, a server for operation verification used at the National Institute of Informatics was accessed without authorization and exploited as a stepping stone in a DDoS attack, due to a weak password being configured. In an incident in which registered information leaked from a stockholders service site, unauthorized access from outside was suspected in April when the leak was disclosed, but the final report published in May revealed that it was an inside job.

■ **Government Agency Initiatives**

Government agency activities with regard to security measures included the holding of the 2nd general meeting of the Cyber Security Strategic Headquarters in May, where a new Cyber Security Strategy (proposed) that determines basic policy for cyber security measures was decided on. This set forth cyber security policy for around the next three years, laying out a course of action for items such as the promotion of research and development and cultivation of human resources to meet objectives, in light of the 2020 Olympic and Paralympic Games to be held in Tokyo. It is next scheduled to undergo revisions through public comment, before a cabinet decision is made.

In April, a proposal was put together for revisions to the Guidelines for the Protection of Personal Information in the Telecommunications Business and Explanation, which were under consideration by the "Working Group on the Handling of Personal and User Information" held under the supervision of the Ministry of Internal Affairs and Communications' ICT Service Safety and Security Study Group. Following amendments based on public comment, the proposed revisions were made in June. The main revised points included the addition of specific examples for the appropriate acquisition of personal information, guidelines for the retention period for connection authentication logs with regards to communication history, and the deletion of conditions for acquiring location information during a criminal investigation.

■ **Other**

In April, the INTERPOL Global Complex for Innovation (IGCI), an organization specializing in cybercrime countermeasures that the International Criminal Police Organization (ICPO) established in Singapore, began full operation[18]. Ahead of this, the IGCI coordinated a takedown of the Simda botnet in collaboration with law enforcement agencies and private-sector businesses in various countries. In Japan, the Tokyo Metropolitan Police Department, the Ministry of Internal Affairs and Communications, Telecom-ISAC Japan, and private-sector companies worked together in the VAWTRAK takedown operation.

In May, it was discovered that the SourceForge open source software library site had bundled and distributed third-party software in software installers published on the site without permission. This resulted in them receiving complaints from the development project[19], and massive backlash from other developers and the community. Due to the overwhelmingly negative reaction they received, SourceForge has rescinded this policy[20].

Additionally, between May and June there were DDoS attacks accompanied by threats that targeted a number of financial institutions and online stores. In these attacks, the threats included demands for payment via bitcoin, with the implication that a DDoS attack of several hundred Gbps would be made if these demands were not met. The DDoS attacks and payment via bitcoin are points similar to attacks made since last year by a group calling themselves DD4BC, so it is thought they may be involved[21]. Attacks have also been made on financial institutions in Europe and Hong Kong, so ongoing vigilance is required.

In June, a minor was arrested on suspicion of involvement in an incident that occurred in December 2014 in which the website of a publisher was accessed without authorization, and altered to redirect visitors to another website. This minor was arrested again in July on suspicion of using someone else's credit card without permission.

---

*18 "International gathering marks inauguration of INTERPOL Global Complex for Innovation" (http://www.interpol.int/News-and-media/News/2015/N2015-039).

*19 GIMP PROJECT, "GIMP PROJECT'S OFFICIAL STATEMENT ON SOURCEFORGE'S ACTIONS" (http://www.gimp.org/).

*20 SourceForge Community Blog, "Third party offers will be presented with Opt-In projects only" (http://sourceforge.net/blog/third-party-offers-willbe-presented-with-opt-in-projects-only/).

*21 See the following Arbor Networks report for more information on DD4BC. "DD4BC DDoS Extortion Threat Activity" (https://asert.arbornetworks.com/dd4bc-ddos-extortion-threat-activity/).

June Incidents

1
2

S **1st:** The Japan Pension Service announced that the personal information of up to 1.25 million people may have leaked after it was accessed without authorization due to a malware infection.
"Regarding the leak of personal information from the Japan Pension Service"
(http://www.nenkin.go.jp/oshirase/topics/2015/20150721.files/0000150601ndjlleouli.pdf) (in Japanese).

3
4

O **3rd:** The USA Freedom Act, which includes provisions that reform the monitoring activities of the National Security Agency (NSA), was passed in the U.S. Senate, and signed into law by the President. This law limits the information gathering activities of the NSA, which had been criticized for its mass collection of phone records around the world, including those of U.S. citizens.
Congress.gov, "H.R.2048 - USA FREEDOM Act of 2015" (https://www.congress.gov/bill/114th-congress/house-bill/2048/text).

5
6

S **5th:** The U.S. Office of Personnel Management (OPM) was accessed without authorization, and data on roughly 4 million federal employees leaked.
"OPM to Notify Employees of Cybersecurity Incident"
(https://www.opm.gov/news/releases/2015/06/opm-to-notify-employees-of-cybersecurity-incident/).

7
8

S **5th:** The National Institute of Informatics announced that one of its servers for operation verification had been accessed without authorization, and used as a stepping stone in a DoS attack.
"Unauthorized access on operation verification server" (http://www.nii.ac.jp/news/2015/0605) (in Japanese).

9

O **5th:** A newly-merged company issued an alert after a domain it had used previously was obtained by a third party, and used to redirect users to other sites, including fraudulent ones.

10
11

V **10th:** Microsoft published their Security Bulletin Summary for June 2015, and released eight updates, including two critical updates MS15-056 and MS15-57, as well as six important updates.
"Microsoft Security Bulletin Summary for June 2015" (https://technet.microsoft.com/library/security/ms15-jun).

12
13

V **10th:** A number of vulnerabilities in Adobe Reader and Acrobat that could allow unauthorized termination and arbitrary code execution were discovered and fixed.
"APSB15-10: Security updates available for Adobe Reader and Acrobat" (https://helpx.adobe.com/security/products/reader/apsb15-10.html).

14
15

O **12th:** The Ministry of Internal Affairs and Communications issued an alert due to incidents in which telephone services such as IP phones were used by third parties without authorization, resulting in users being billed for expensive international telephone charges.
Causes included the exploitation of vulnerabilities in PBX or routers connected to IP phones, and the theft of IDs and passwords for use in international phone calls due to a weak password being assigned. "Alert regarding the unauthorized use of IP phones by third parties"
(http://www.soumu.go.jp/menu_kyotsuu/important/kinkyu02_000191.html) (in Japanese).

16
17

S **16th:** U.S. company LastPass, which provides a password management tool, announced it had been compromised and some authentication hashes and user account information had leaked.
"LastPass Security Notice" (https://blog.lastpass.com/2015/06/lastpass-security-notice.html/).

18
19

O **16th:** The Council of Anti-Phishing Japan issued an alert due to confirmation of techniques redirecting users to a banking phishing site using SMS (Short Message Service) at a number of banks from late May.
"[Alert] Beware of redirection to bank phishing sites by SMS (Short Message Service) (June 16, 2015)"
(https://www.antiphishing.jp/news/alert/_sms_20150616.html) (in Japanese).

20
21

V **17th:** A number of vulnerabilities that could allow arbitrary code execution were discovered and fixed in the keyboard function of the SwiftKey SDK used in Samsung brand mobile devices.
US-CERT, "Vulnerability Note VU#155412 Samsung Galaxy S phones fail to properly validate SwiftKey language pack updates"
(http://www.kb.cert.org/vuls/id/155412).

22
23

S **22nd:** The flight plan computers of LOT Polish Airlines were attacked, causing flights to be canceled or delayed.
LOT Polish Airlines, "TODAY AFTERNOON LOT ENCOUNTERED IT ATTACK, THAT AFFECTED OUR GROUND OPERATION SYSTEMS."
(http://corporate.lot.com/pl/en/press-news?article=772922).

24
25

V **24th:** A vulnerability in Adobe Flash Player that could allow arbitrary code execution was discovered and fixed. It was confirmed that this vulnerability had already been exploited in limited targeted attacks.
"Security updates available for Adobe Flash Player" (https://helpx.adobe.com/security/products/flash-player/apsb15-14.html).

26
27

O **24th:** The Ministry of Internal Affairs and Communications revised its Guidelines for the Protection of Personal Information in the Telecommunications Business and Explanation.
"Guidelines for the Protection of Personal Information in the Telecommunications Business"
(http://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/telecom_perinfo_guideline_intro.html) (in Japanese).

28
29
30

S **25th:** DDoS attacks accompanied by threats were made on the online systems of a certain bank, resulting in systems being difficult to access, etc.

*Dates are in Japan Standard Time

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

The Ministry of Internal Affairs and Communications issued an alert due to a rise in the number of incidents in which telephone services such as IP phones were used by a third party without authorization, resulting in users being billed for expensive international telephone charges. Subsequently, a security firm pointed out that the products of a certain vendor had caused significant damages because they were operated with a default password set, and allowed access from the Internet, but the vendor has denied this. Regarding this issue, based on discussions at the Workshop on the Appropriate Way to Handle Cyber Attacks in the Telecommunications Business, in July the Ministry of Internal Affairs and Communications sent a request for cooperation to telecommunications business organizations, asking them to make it known that telecommunications companies that have contracts with users to provide international telephone services should take appropriate measures to stop unauthorized use before it occurs, and prevent the spread of damages[22].

A series of several vulnerabilities were discovered on SSL/TLS servers. Meanwhile, there are also issues such as the fact that applying the latest countermeasures prevents connection from older devices, meaning that designs and operation that account for the tradeoff between security and required compatibility are necessary. In answer to this, in May the IPA published their SSL/TLS encryption configuration guidelines, to enable encryption settings that take into consideration appropriate security for SSL/TLS server builders and operators.

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 are not the type that utilizes advanced knowledge such as that of vulnerabilities, but rather cause large volumes of unnecessary traffic to overwhelm network bandwidth or server processes for the purpose of hindering services.

■ **Direct Observations**

Figure 2 shows the circumstances of DDoS attacks handled by the IIJ DDoS Protection Service between April 1 and June 30, 2015.

This information shows traffic anomalies judged to be attacks based on IIJ DDoS Protection Service standards. IIJ also responds to other DDoS attacks, but these incidents are excluded from the figure due to the difficulty in accurately ascertaining the facts of each situation.

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 categorizes DDoS attacks into three types: attacks
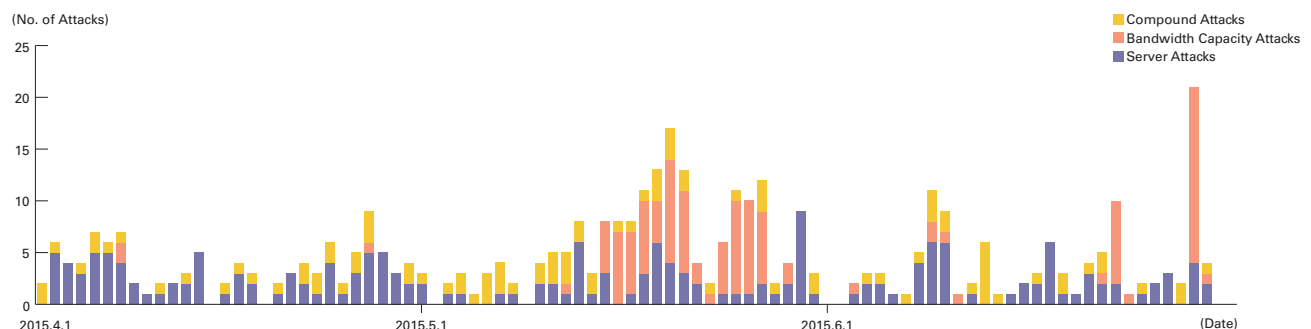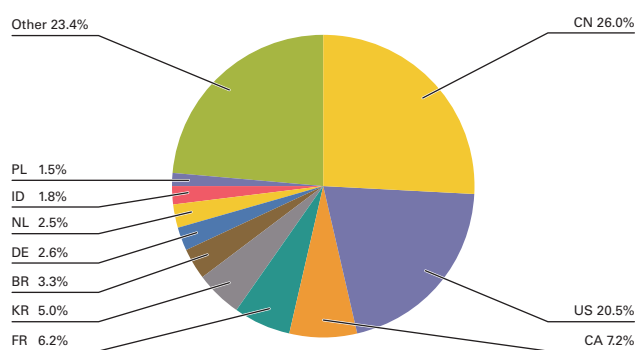

Figure 2: Trends in DDoS Attacks

---

*22 Ministry of Internal Affairs and Communications, "Regarding measures to prevent the unauthorized use of IP phones, etc., by third parties (request)" (http://www.soumu.go.jp/menu_news/s-news/01ryutsu03_02000096.html) (in Japanese).

on bandwidth capacity*23, attacks on servers*24, and compound attacks (several types of attacks on a single target conducted at the same time).

During the three months under study, IIJ dealt with 361 DDoS attacks. This averages to 3.97 attacks per day, indicating a decrease in the average daily number of attacks compared to our prior report. Server attacks accounted for 52.1% of all incidents, while compound attacks accounted for 14.9%, and bandwidth capacity attacks 33%. The largest attack observed during the period under study was classified as a compound attack, and resulted in 8.25 Gbps of bandwidth using up to 2,713,000 pps packets.

Of all attacks, 77.3% ended within 30 minutes of commencement, 22.7% lasted between 30 minutes and 24 hours, and none lasted over 24 hours. The longest sustained attack was a compound attack that lasted for 10 hours and 37 minutes.

In most cases, we observed an extremely large number of IP addresses, whether domestic or foreign. We believe this is accounted for by the use of IP spoofing*25 and botnet*26 usage as the method for conducting DDoS attacks.
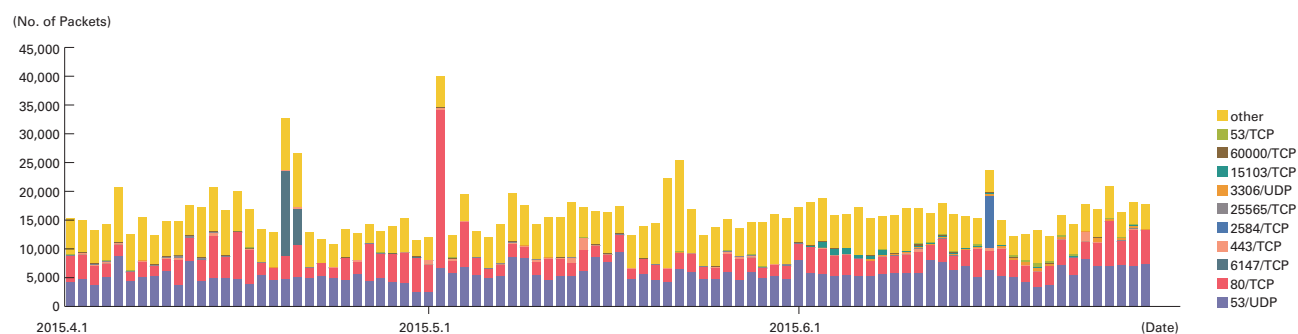

Figure 3: DDoS Attack Targets by Country According to Backscatter Observations

■ **Backscatter Observations**

Next we present our observations of DDoS attack backscatter using the honeypots*27 set up by the MITF, a malware activity observation project operated by IIJ*28. By monitoring backscatter it is possible to detect some of the DDoS attacks occurring on external networks as a third party without any interposition.

For the backscatter observed between April 1 and June 30, 2015, Figure 3 shows the sender's IP addresses classified by country, and Figure 4 shows trends in packet numbers by port.

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

*23   Attack that overwhelms the network bandwidth capacity of a target by sending massive volumes of larger-than-necessary IP packets and fragments. The use of UDP packets is called a UDP flood, while the use of ICMP packets is called an ICMP flood.

*24   TCP SYN flood, TCP connection flood, and HTTP GET flood attacks. TCP SYN flood attacks send mass volumes of SYN packets that signal the start of TCP connections, forcing the target to prepare for major incoming connections, causing the wastage of processing capacity and memory. TCP connection flood attacks establish mass volumes of actual TCP connections. HTTP GET flood attacks establish TCP connections on a Web server, and then send mass volumes of HTTP GET protocol commands, wasting processing capacity and memory.

*25   Misrepresentation of a sender's IP address. Creates and sends an attack packet that has been given an address other than the actual IP address of the attacker to make it appear as if the attack is coming from a different location, or from a large number of individuals.

*26   A "bot" is a type of malware that institutes an attack after receiving a command from an external C&C server. A network constructed of a large number of bots acting in concert is called a botnet.

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

*28   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."

The port most commonly targeted by the DDoS attacks observed was the 53/UDP port used for DNS, accounting for 34.5% of the total. This was followed by 80/TCP used for Web services at 22.3%, so the top two ports accounted for 56.8% of the total. Attacks were also observed on 443/TCP used for HTTPS, 53/TCP used for DNS, and 25565/TCP sometimes used for game servers, as well as 6147/TCP and 2584/TCP, which are not commonly used.

Examining the daily average number of packets for the 53/UDP communications observed often since February 2014, we can see that although it dropped from around 6,200 in the previous survey period to around 5,600, it remains high.

Looking at the origin of backscatter thought to indicate IP addresses targeted by DDoS by country in Figure 3, China accounted for the largest ratio at 26.0%. The United States and Canada followed at 20.5% and 7.2%, respectively.

Regarding particularly large numbers of backscatter packets observed by port, there were attacks on the Web servers (80/ TCP) of a U.S. hosting provider between April 26 and May 1, and on the server group of a hosting provider in Canada on May 2. Attacks observed on other ports included those targeting 6147/TCP between April 19 and 20, but because the source IP address for the backscatter was a private address, the attack target is not known. On June 17, attacks on the servers of a U.S. ISP that targeted 2584/TCP were observed.

Notable DDoS attacks during the current survey period that were detected via IIJ's observations of backscatter included attacks on a U.K. online casino between April 13 and April 15, and attacks on a French company involved in nuclear energy May 5. Attacks were also detected on a number of government agency sites in Canada on June 18 and June 28.
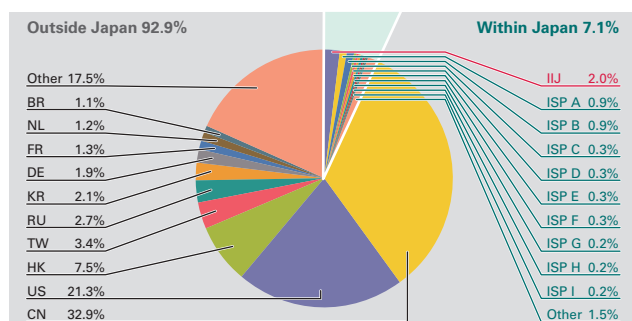

**Figure 5: Sender Distribution
(by Country, Entire Period under Study)**

1.3.2 Malware Activities

Here, we will discuss the results of the observations of the MITF[29], a malware activity observation project operated by IIJ. The MITF uses honeypots[30] connected to the Internet in a manner similar to general users in order to observe communications arriving over the Internet. Most appear to be communications by malware selecting a target at random, or scans attempting to locate a target for attack.

■ **Status of Random Communications**

Figure 5 shows the distribution of sender's IP addresses by country for communications coming into the honeypots

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

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

*30 A system designed to simulate damages from attacks by emulating vulnerabilities, recording the behavior of attackers, and the activities of malware.

between April 1 and June 30, 2015. Figure 6 shows trends in the total volumes (incoming packets). The MITF has set up numerous honeypots for the purpose of observation. We have taken the average per honeypot, showing the trends for incoming packet types (top ten) over the entire period subject to study. Additionally, in these observations we corrected data to count multiple TCP connections as a single attack when the attack involved multiple connections to a specific port, such as attacks on MSRPC.

Much of the communications arriving at the honeypots during the current survey period were DNS Water Torture attack attempts, communications searching for devices to carry out DoS attacks on by amplifying response packets, communications searching for proxy servers, or scanning behavior targeting TCP ports used by Microsoft operating systems. Between June 7 and June 30, a large volume of communications targeting 53/UDP, used by DNS servers, occurred on one of our honeypots. Although the source IP address distribution for these communications was centered on the United States and China, they were mostly unique and covered a wide range, so it is possible that either a bot was used, or the source addresses were spoofed. Upon investigating the content of these communications, we found they were repeated queries for A records that do not exist, such as "(random string). www.example.com." For this reason, we believe this was an example of DNS Water Torture[31]. Several of the domains targeted in these attacks were servers in China (.com domains).

From a number of IP addresses, we also observed scanning behavior for 5353/UDP used for multicast DNS, 123/UDP used for NTP, and 5351/UDP used for NAT-PMP along with DNS TXT record queries. We believe these were searches for devices that could be used in reflection attacks by sending packets spoofed to be from the site you want to attack, amplifying the response packets. With regard to the 5351/UDP communications, there have been reports[32] claiming it was possible to control NAT-PMP externally on some devices, so this could have been a search for vulnerable devices.

From April to early May, there was an increase in 8888/TCP communications. These communications were carried out repeatedly, targeting a wide range of IP addresses from two IP addresses allocated to China. At the same time as 8888/TCP, scanning behavior was also observed for ports such as 37564/TCP[33] used by the KanColleViewer utility tool for an online game, 8118/TCP[34] used by a proxy server called Privoxy, 3128/TCP that is set by default for a number of proxy servers such as Squid, and 8090/TCP used as a backup for 8123/TCP and 8080/TCP, which are used by a proxy server called Polipo. For this reason, we believe these communications were searches for open proxy servers.

■ **Malware Network Activity**
Figure 7 shows the distribution of the specimen acquisition source for malware during the period under study, while Figure 8 shows trends in the total number of malware specimens acquired. Figure 9 shows trends in the number of unique specimens. In Figure 8 and Figure 9, the number of acquired specimens show the total number of specimens acquired per day[35], while the number of unique specimens is the number of specimen variants categorized according to their digest of a hash function[36]. Specimens are also identified using anti-virus software, and a breakdown of the top 10 variants is displayed color coded by malware name. As

*31   Secure64 Software Corporation, "Water Torture: A Slow Drip DNS DDoS Attack" (https://blog.secure64.com/?p=377). For an explanation in Japanese, see the following document written by Mr. Yasuhiro Orange Morishita of Japan Registry Services. "DNS Water Torture Attacks" (http://2014.seccon.jp/dns/dns_water_torture.pdf) (in Japanese). The MITF honeypots do not query authoritative servers or cache servers when they receive DNS query packets, so they provide no aid to attacks.

*32   "Vulnerability Note VU#184540 Incorrect implementation of NAT-PMP in multiple devices" (http://www.kb.cert.org/vuls/id/184540). Scanning behavior targeting this port has also been reported in the past. "Regarding scanning behavior for devices that allow external control of NAT-PMP" (http://www.npa.go.jp/cyberpolice/detect/pdf/20141030.pdf) (in Japanese).

*33   Because the KanColleViewer was configured as an open proxy by default, JVN and others issued an alert. "VNVU#98282440 Issues with 'Teitoku Gyo Mo Isogashii!' (KanColleViewer) operating as an open proxy" (https://jvn.jp/vu/JVNVU98282440/) (in Japanese). The National Police Agency also reported an increase in scanning for this port. "Regarding an increase in proxy searches targeting specific ports" (http://www.npa.go.jp/cyberpolice/detect/pdf/20150525.pdf) (in Japanese).
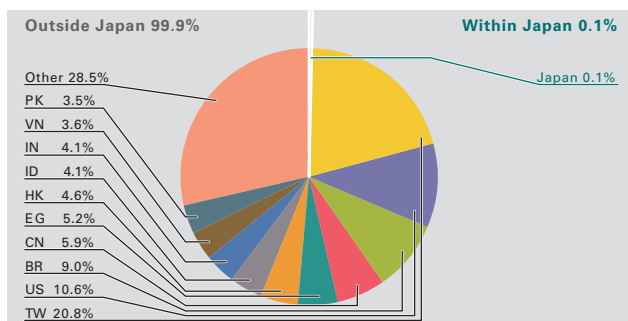
*34   An increase in access targeting this port has been reported in the past. "Internet observation results, etc. (January 2015)" (https://www.npa.go.jp/cyberpolice/detect/pdf/20150304.pdf) (in Japanese).

*35   This indicates the malware acquired by honeypots.

*36   This figure is derived by utilizing a one-way function (hash function) that outputs a fixed-length value for various input. The hash function is designed to produce as many different outputs as possible for different inputs. While we cannot guarantee the uniqueness of specimens by hash value, given that obfuscation and padding may result in specimens of the same malware having different hash values, the MITF has expended its best efforts to take this fact into consideration when using this methodology as a measurement index.
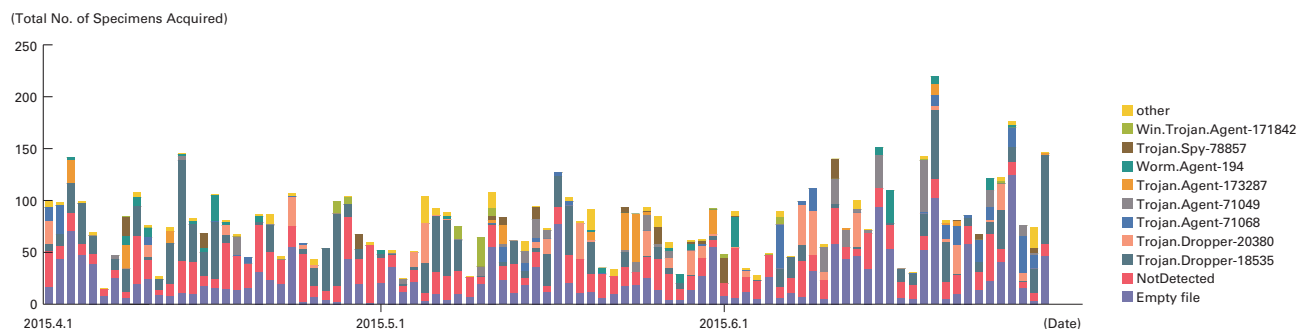
with our previous report, for Figure 8 and Figure 9 we have detected Conficker using multiple anti-virus software packages, and removed any Conficker results when totaling data.

On average, 91 specimens were acquired per day during the period under study, representing 17 different malware. After investigating the undetected specimens more closely, they included worms observed from IP addresses allocated to countries such as China, the United States, India, Taiwan, and Austria. Additionally, about 55% of undetected specimens were in text format. Because many of these text format specimens were HTML 404 or 403 error responses from Web servers, we believe this was due to infection behavior of malware such as old worms continuing despite the closure of download sites that newly-infected PCs access to download malware.
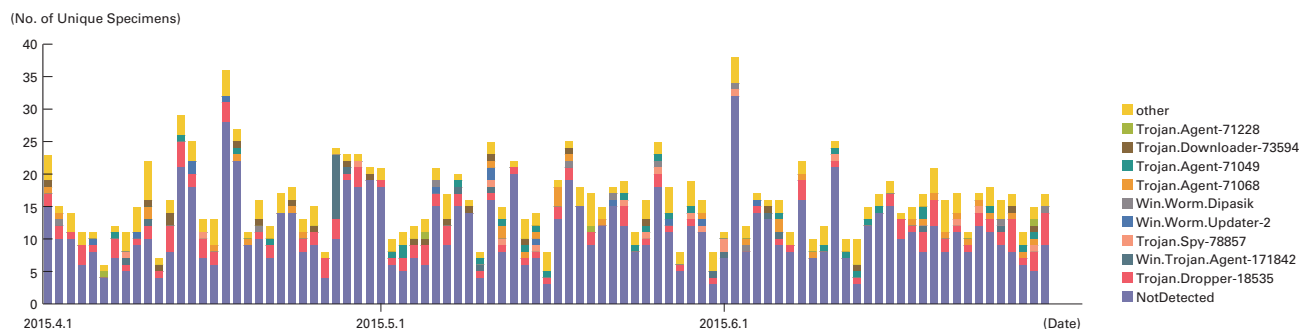
Under the MITF's independent analysis, during the current period under observation 90.0% of malware specimens acquired were worms, 1.8% were bots, and 8.2% were downloaders. In addition, the MITF confirmed the presence of 106 botnet C&C servers*37 and 9 malware distribution sites. The number of botnet C&C servers is higher than before, but this was due to the appearance of a specimen that used a DGA (Domain Generation Algorithm) during the current survey period.

Outside Japan 99.9%   **Within Japan 0.1%**

Japan 0.1%

Other 28.5%
PK   3.5%
VN   3.6%
IN   4.1%
ID   4.1%
HK   4.6%
E G   5.2%
CN   5.9%
BR   9.0%
US 10.6%
TW 20.8%

**Figure 7: Distribution of Acquired Specimens by Source
(by Country, Entire Period under Study, Excluding Conficker)**

(Total No. of Specimens Acquired)

- other
- Win.Trojan.Agent-171842
- Trojan.Spy-78857
- Worm.Agent-194
- Trojan.Agent-173287
- Trojan.Agent-71049
- Trojan.Agent-71068
- Trojan.Dropper-20380
- Trojan.Dropper-18535
- NotDetected
- Empty file

2015.4.1   2015.5.1   2015.6.1   (Date)

Figure 8: Trends in the Total Number of Malware Specimens Acquired (Excluding Conficker)

(No. of Unique Specimens)

- other
- Trojan.Agent-71228
- Trojan.Downloader-73594
- Trojan.Agent-71049
- Trojan.Agent-71068
- Win.Worm.Dipasik
- Win.Worm.Updater-2
- Trojan.Spy-78857
- Win.Trojan.Agent-171842
- Trojan.Dropper-18535
- NotDetected

2015.4.1   2015.5.1   2015.6.1   (Date)

Figure 9: Trends in the Number of Unique Specimens (Excluding Conficker)

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

■ **Conficker Activity**

Including Conficker, an average of 27,999 specimens were acquired per day during the period covered by this report, representing 549 different malware. While figures rise and fall over short periods, Conficker accounts for 99.7% of the total number of specimens acquired, and 97.0% of unique specimens. This demonstrates that Conficker remains the most prevalent malware by far, so we have omitted it from figures in this report. Compared to the previous survey period, the total number of specimens acquired increased by approximately 44% during the period covered by this report, and the number of unique specimens decreased by about 10%. The increase in the total number of specimens acquired was due to a spike in infection activity from IP addresses allocated to the United States during the current survey period. According to the observations of the Conficker Working Group*38, as of July 1, 2015, a total of 775,060 unique IP addresses are infected. This indicates a drop to about 24% of the 3.2 million PCs observed in November 2011, but it demonstrates that infections are still widespread.

1.3.3 SQL Injection Attacks

Of the types of different Web server attacks, IIJ conducts ongoing surveys related to SQL injection attacks*39. SQL injection attacks have flared up in frequency numerous times in the past, remaining one of the major topics in the Internet security. SQL injections are known to occur in one of three attack patterns: those that attempt to steal data, those that attempt to overload database servers, and those that attempt to rewrite Web content.

Figure 10 shows the distribution of SQL injection attacks against Web servers detected between April 1 and June 30, 2015. Figure 11 shows trends in the numbers of attacks. These are a summary of attacks detected by signatures on the IIJ Managed IPS Service. China was the source for 63.3% of attacks observed, while Japan and the United States accounted for 20.7% and 7.6%, respectively, with other countries following in order. There was a significant decrease in the number of SQL injection attacks against Web servers compared to the previous report. This is due to a drop in the number of large-scale attacks from China, although attacks from China are still continue to occur, and account for a large percentage of the total.
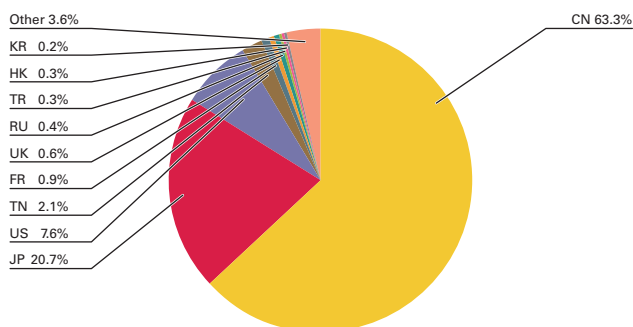
During this period, attacks from a specific attack source in China directed at specific targets took place between April 22 and April 23. Attacks from a different specific attack source in China targeting specific targets occurred on May 14. On June 21, attacks were made from a number of sources in the United States and Europe directed at specific targets. These attacks are thought to have been attempts to find vulnerabilities on Web servers.
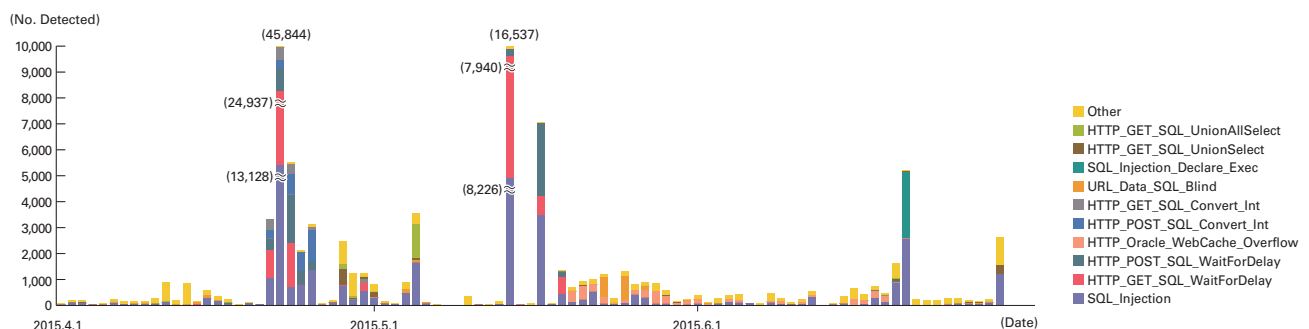

Figure 10: Distribution of SQL Injection Attacks by Source

Figure 11: Trends in SQL Injection Attacks (by Day, by Attack Type)

As previously shown, attacks of various types were properly detected and dealt with in the course of service. However, attack attempts continue, requiring ongoing attention.

1.3.4 Website Alterations

Here we indicate the status of website alterations as surveyed through the MITF Web crawler (client honeypot)[*40].

This Web crawler accesses hundreds of thousands of websites on a daily basis, with a focus on well-known and popular sites in Japan. We also add new target sites on a regular basis. 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 is easier to speculate on trends regarding fluctuations in the number of altered sites, as well as the vulnerabilities exploited and malware distributed.

The number of drive-by download attacks observed between April 1 and June 30, 2015, was more than double the number seen between January 1 and March 31, 2015. In the first half of this survey period many cases of Neutrino were observed, while in the latter half most attacks had shifted to Nuclear (Figure 12). For a period of time in mid-June, A new exploit kit called Sundown was detected. Sundown is used in attacks targeting users in Japan in particular[*41], and during observations by the MITF, we confirmed that the download of malware such as Kasidet ultimately took place, exploiting Flash vulnerabilities such as CVE-2015-0311 and CVE-2015-0313.

After updating the Web crawler system on July 1, 2015 to add functions and change the configuration, attacks were observed using Angler, which had not been detected by the MITF since January 2015. There were more Angler attacks than the total of other attacks combined, and since July 1 the total number of attacks detected has increased almost threefold. For this reason, we believe it highly likely that Angler was used in attacks targeting users in Japan before June 30 also. Also see "1.4.2 Angler Exploit Kit on the Rampage" for more information about our observations of Angler-based attacks and their details.

On the whole, attacks using drive-by downloads are occurring comparatively frequently. In early July 2015, a number of zero-day vulnerabilities that could be exploited by exploit kits were disclosed (CVE-2015-5119, CVE-2015-5122, etc.). Exploit kits such as the aforementioned Angler, Nuclear, and Neutrino added functions for exploiting these vulnerabilities in an extremely short period of time. We recommend browser users and administrators stay well aware of vulnerabilities in OSes and browser-related plug-ins, and carry out thorough countermeasures such as applying updates and enabling EMET. It is also important for website operators to implement countermeasures for Web content alteration and confirm the integrity of mashup content provided by third parties.

*Covers several hundreds of thousands of sites in Japan. In recent years, drive-by downloads have been configured to change attack details and whether or not attacks are made based on the client system environment or session information, source address attributes, and the quota achievement status of factors such as number of attacks. This means that results can vary wildly at times depending on the test environment and circumstances.

Figure 12: Rate of Drive-By Download Incidence When Viewing Websites (%) (by Exploit Kit)

*40 See "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 an explanation of Web crawler observation methods.

*41 Regarding Sundown, "Fast look at Sundown EK" (http://malware.dontneedcoffee.com/2015/06/fast-look-at-sundown-ek.html) touches on its functions and attack targets based on information gained from the Control Panel, etc.

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 perform independent surveys and analyses of prevalent incidents. Here, we present information from the surveys we have undertaken during this period covering three themes. First, we discuss machine learning and security. Next, we look at the Angler Exploit Kit that has been spreading like wildfire. Finally, we examine the actual use of ID management technology.

1.4.1 Machine Learning and Security

Machine learning is a branch of artificial intelligence, which has been studied for many years. This field involves the research and development of algorithms that improve their output automatically based on known data, much like human beings learn from experience. Together with the accumulation of a large amount of data on the Internet, and the availability of vast computational resources at a reasonable price due to cloud computing, etc., the academic achievements of machine learning and its applications in IT systems are being discussed more often these days.

In the past, when there was a topic for which the decision-making criteria could be made explicit, efforts were made to automate it using programming. Machine learning techniques can be used to learn from and imitate human decision making, even with topics for which criteria cannot easily be made explicit, and this is expected to be useful for promoting automation.

In the field of security, automatic decision making has been carried out using explicit criteria such as blacklists, whitelists, and signatures, but it is hoped that machine learning techniques will be effective for the many instances in which decision making is not easy using these methods alone. Additionally, as the difficulty of training security staff and the burden of their daily decision-making work are now recognized as issues, machine learning has attracted attention as a technology for automating or supporting decision-making tasks.

In this report, we give a broad overview of machine learning and examples of its application in the field of security, then examine whether it can be applied as a countermeasure to a number of threat types. We also consider the security risks of systems that incorporate machine learning.

■ About Machine Learning

First, we will discuss the main machine learning systems (Figure 13). When using machine learning, you first prepare training data consisting of pairs of input and expected output in advance[42]. Human decision making is "learned" beforehand using this training data, then an algorithm for decision making and the learned results are built in to the system you want to carry out automatic decision making on.

Decision-making algorithms generally incorporate many internal parameters, and are designed so that output changes in various ways according to the values set. The aforementioned "learning" is essentially the calculation of the appropriate setting values to comply with the training data. A range of techniques are being developed and researched using a decision making algorithm and the "learning algorithm" procedure for calculating its set values, such as neural networks, support vector machines (SVM), and decision tree learning.

When applying this to actual topics, it is necessary to analyze the topic in question and its data, select a suitable machine learning algorithm, and implement preprocessing (statistical processing, text processing, etc.) of input so it is easier for that algorithm to learn and make decisions.
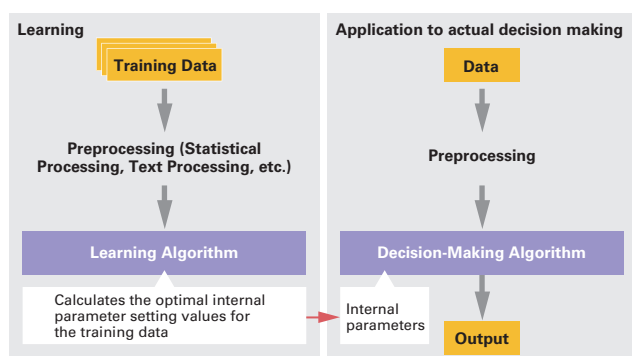

Figure 13: Supervised Machine Learning

---

*42　Here we will discuss the method called supervised learning. There is another method called unsupervised learning.

■ **Applications in the Field of Security**

Security products that incorporate machine learning techniques already exist. For example, the Bayesian filtering used in spam detection is technology that applies a machine learning technique known as a naïve Bayes classifier. Other examples include products that detect anomalies such as attacks and malware by learning communication logs, and products that detect malware communications by learning communication packets. In the field of malware detection, there are client device-oriented products that learn the characteristics of malware binaries and behavior, and incorporate these results into their detection engine.

Besides individual products, there are also cases in which machine learning is being applied to onsite incident response work. At Indonesia's Id-SIRTII, they are developing machine learning-based detection engines for each attack not compatible with signatures on their environment with many IDS implemented, and building attack detection systems to match their situation[43]. Information gathering work is also indispensable for security staff. One part of this is gathering and analyzing information on security incidents and accidents that occur at other companies or organizations on a daily basis. At NCSC-NL in the Netherlands, they are developing a system that applies machine learning to news articles they collect, automatically grouping articles about the same incidents and adding tags for each category (DDoS attacks, information leaks, etc.), to support the analysis work of staff[44].

Thus far, we have looked at individual systems that incorporate machine learning techniques. Next, we will examine examples of three actual threats: inside jobs, targeted attacks, and attacks on Web services. We will then consider whether machine learning can be applied to effectively counter them.

■ **Threat Countermeasures**

For the first example, we will examine whether applying machine learning to the threat of inside jobs is effective (Figure 14). Inside jobs include cases in which sensitive data is removed from the premises, as well as DoS attacks on important servers.

First, machine learning could be used to learn the normal communication profiles of each user based on traffic on an organization's network. Here, we will refer to information such as the destination, amount, and type (email, Internet, document transfer, etc.) of communications as a communication profile. It is extremely difficult to manually list all of the communication profiles that occur during normal work and follow future changes in detail, but this can be automated using machine learning.

Additionally, a system for detecting communications that deviate from the normal profile could be built, and traffic monitored. With this system in place, we could detect behavior suspected of being the unauthorized collection of sensitive information, such as when an insider obtains an unusually large amount of data from an important server, or obtains data from an important server they don't
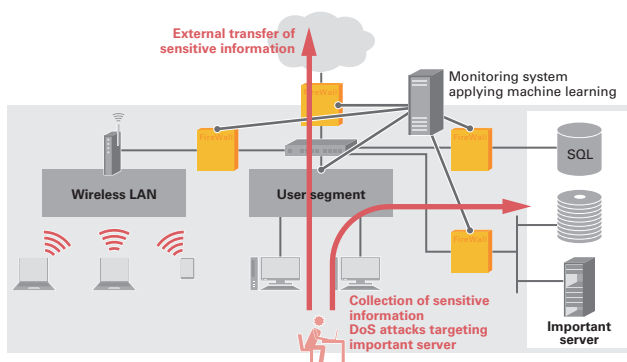

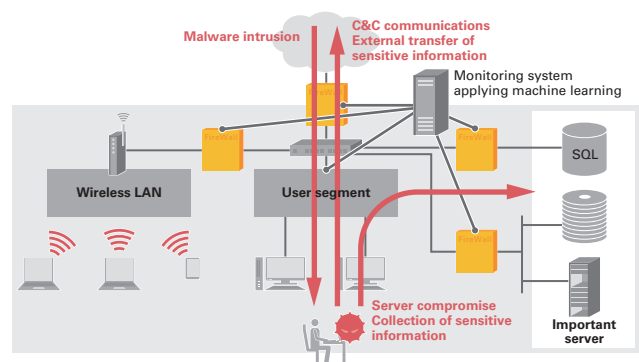**Figure 14: Applying Machine Learning to Inside Job Countermeasures**

Figure 15: Applying Machine Learning to Targeted Attack Countermeasures

*43 Bisyron MASDUKI (Id-SIRTII), Muhammad SALAHUDDIEN (Id-SIRTII), "Implementation of Machine Learning Methods for Improving Detection Accuracy on Intrusion Detection System (IDS)" (http://www.first.org/conference/2015/program#pimplementation-of-machine-learning-methods-forimproving-detection-accuracy-on-intrusion-detection-system-ids).

*44 Edwin TUMP (NCSC-NL), "Machine Learning for Cyber Security Intelligence" (http://www.first.org/conference/2015/program#pmachine-learningfor-cyber-security-intelligence).

normally use during the course of their work. Behavior suspected of being unauthorized removal of data, such as the transfer of a large amount of data outside the company, can also be detected as a departure from the normal communication profile. Attempts to launch a DoS attack on important servers by directing a large volume of traffic at them can be detected in the same way.

For the second example, we will examine the threat of targeted attacks (Figure 15). As with the inside job example, communication profiles for each user could be learned. The binary characteristics of targeted attack malware could also be learned. It may be useful to learn the characteristics of malware communications as well. Then, traffic could be monitored and binaries included in external communications, including email and Web traffic, could be inspected based on the learned results. This would make it possible to detect targeted attack malware intrusions through the monitoring point. Even if intrusions were allowed to happen and a client PC became infected with malware, the spread of infection, compromise of servers, and collection of sensitive information could be detected as deviations from the normal communication profiles for communications within the organization. External communications such as communications with C&C servers and the transfer of sensitive data can also be detected.

For the final example, we will look into the protection of Web services exposed to external parties (Figure 16). On public Web services, the issues of SQL injection attacks and unauthorized access to management interfaces are frequently addressed. First, machine learning could be used to learn the normal URL format, content and volume of data transferred and received, and page transitions. Then a system could be built to detect communications deviating from these. Web requests detected as having an abnormal URL format, or incoming data abnormal in content or size, may possibly be SQL injection attacks. When the amount of response data from a server is much larger than normal, the acquisition of data from a database would be suspected. Access to management pages that circumvents the management login page may indicate unauthorized access to the management interface.

As shown above, by learning appropriate details to match the threat, and monitoring the situation, we believe machine learning techniques function as effective countermeasures.

■ **The Security Risks of Machine Learning Systems**
Let us consider whether implementing machine learning techniques causes new risks to emerge.

The most conceivable threat is the compromise of the system incorporating machine learning itself. There is a risk of internal manipulation of the system, leading to it being altered or shut down to overlook attacks. That said, this risk is not only present in machine learning systems, but also monitoring systems as a whole.

There are also other risks on systems that learn criteria for discriminating between normal and abnormal behavior from the installed environment on an ongoing basis. When attacks are made with a volume and quality that gradually approach intended levels over the course of time, it is possible for them to be overlooked and detected as normal communications.

For example, imagine that a single reference file is extracted from a file server and sent to an external party. Because only a single file is involved, it may be learned as a normal external transmission instead of be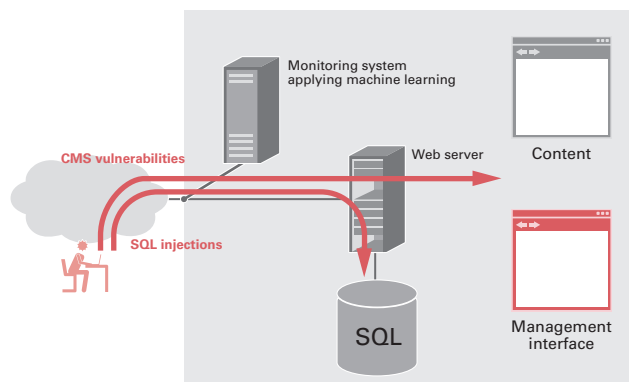ing detected as an attack. If two reference files were next extracted and sent externally, because a single file was considered normal, two files may also be evaluated and learned as normal as well. By repeating this process, it could ultimately be possible to transmit a large amount of materials to an external party.

This issue was already identified when anomaly detection technology first appeared. When an attack requires an extremely long time to carry out, it could be said that the hurdle to attacking is equally high. That means the more gradual the learning process is the better, but on the other hand, quickly conforming to changes in legitimate forms of work use is desirable. To resolve this dilemma, it is necessary to find a suitable balance for both.

Figure 16: Applying Machine Learning to Web Service Protection
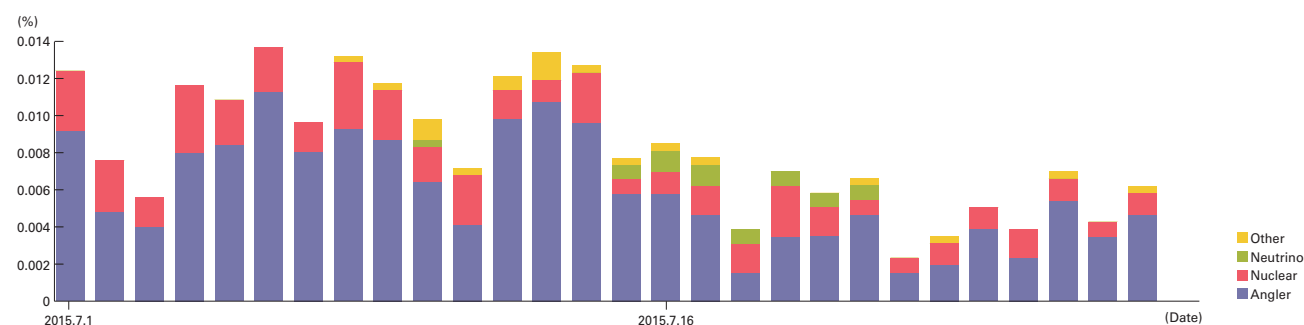
At this stage, the risks of implementing machine learning have not been discussed sufficiently. It is possible that new issues may be identified in the future, so we must continue to keep a close eye on developments.

■ **Summary**

One of the main goals of security-related technology is supporting the detection, analysis, and responses to security incidents by automating and speeding up the work involved. When using methodology that specifies distinct abnormal and normal states, such as blacklists and whitelists, a vast intermediate range that cannot realistically be enumerated remains. Machine learning shows promise as a technique with the potential to deal with this intermediate range automatically, and we believe that examples of applications for machine learning will become more common in the future.
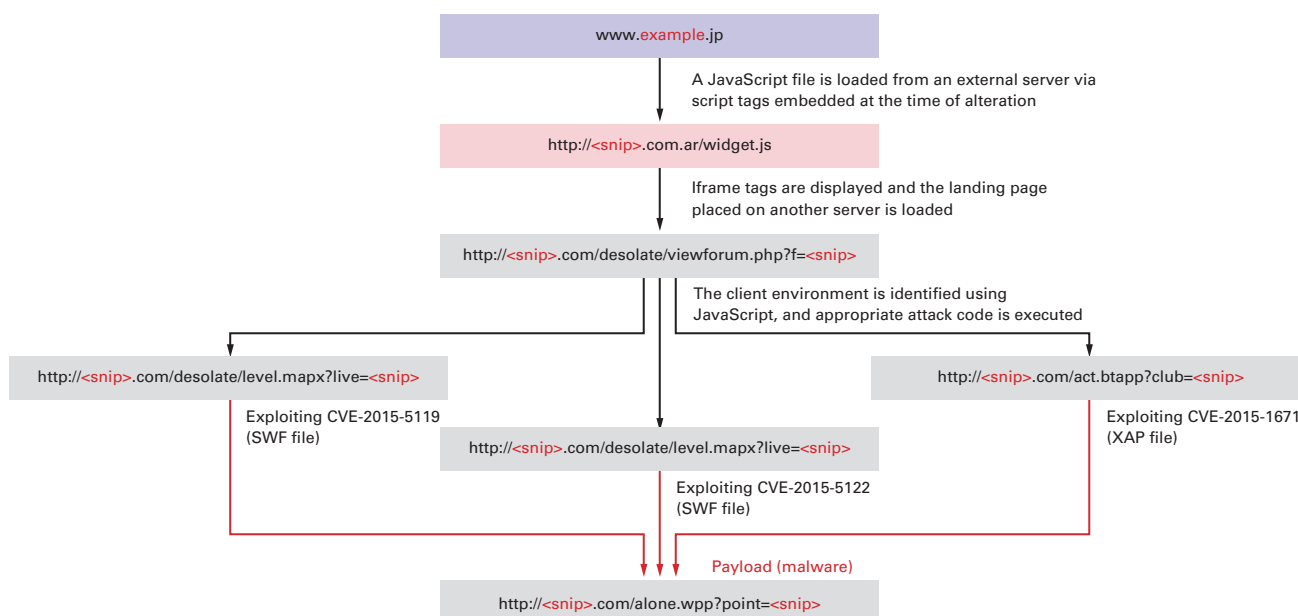
1.4.2 Angler Exploit Kit on the Rampage

At IIJ, we began to detect many Angler-based attacks after updating the MITF Web crawler system on July 1, 2015 (Figure 17)[45]. In this report, we give an overview of the functions and payload of the Angler Exploit Kit, which as of July 2015 is frequently used for drive-by downloads targeting Japan, and also examine trends in the attack environment.

*Note: Covers several hundreds of thousands of sites in Japan. In recent years, drive-by downloads have been configured to change attack details and whether or not attacks are made based on the client system environment or session information, source address attributes, and the quota achievement status of factors such as number of attacks. This means that results can vary wildly at times depending on the test environment and circumstances.

Figure 17: Drive-By Download Incidence (%) from July 1 to July 28, 2015 (by exploit kit)

*Note: The FQDN and path are both changed frequently. Additionally, because different strings are used for the directory, parameter names, and values each time, it is usually difficult to identify exploit kits by URL pattern alone.

Figure 18: Typical Angler URL Transition Examples (July 2015)

*45　See "1.3.4 Website Alterations" for more information on overall drive-by download trends.

■ **Attack Flow and Concealment Techniques**

Figure 18 shows examples of typical Angler URL transitions observed at the time of writing (late July, 2015). In these examples, script tags that load an external JavaScript file are inserted into HTML content on altered websites, and from these files an HTML file ("landing page") containing malicious JavaScript placed on another server (Infector) is loaded*46.

The landing page appears to be HTML content for an English document of around 150 KB, but it includes JavaScript with multiple layers of obfuscation (Figure 19).

This landing page is also equipped with mechanisms for detecting virtual environments or anti-virus software by exploiting an IE resource information disclosure vulnerability (CVE-2013-7331/MS14-052), etc.*47, and when these are found attacks are subsequently not carried out. If a targeted virtual environment or anti-virus software is not detected, attempts are made to exploit a memory corruption vulnerability in IE (CVE-2014-4130/MS14-056), execute SWF files that exploit Adobe Flash vulnerabilities (such as CVE-2015-5122, CVE-2015-5119, and CVE-2015-3113), or execute XAP files that exploit a Windows TrueType font parsing vulnerability (CVE-2015-1671) (Figure 20). This exploit content is also obfuscated, so it is difficult to identify its nature with superficial analysis (Figure 21, Figure 22).

Figure 19: Part of the Angler Landing Page (obfuscated)

Figure 21: Part of the SWF File Exploiting CVE-2015-5122 (obfuscated)

Figure 20: Part of the Angler Landing Page (CVE-2014-4130 exploit portion, deobfuscated)

Figure 22: Part of the DLL File Exploiting CVE-2015-1671 (obfuscated)

*46   Sometimes the infector landing page is loaded directly from the altered HTML file, without an external JavaScript file.

*47   For example, "CVE-2013-7331 and Exploit Kits" (http://malware.dontneedcoffee.com/2014/10/cve-2013-7331-and-exploit-kits.html) reported mechanisms in which the exploit kit checked the client environment by exploiting CVE-2013-7331 or using other techniques.

■ **Payload**

As of July 2015, the following two types of malware have been confirmed in the payload downloaded when an exploit is successful.

- **CryptoWall 3.0**
- **Necurs**

CryptoWall 3.0 is a form of ransomware frequently exploited both in Japan and overseas. It encrypts files saved on a client PC, then demands payment (via bitcoin) in exchange for the decryption key (Figure 23). CryptoWall 3.0 shares the key with the attacker's management server before performing encryption, but because it doesn't automatically acquire the proxy settings of the PC, encryption is not carried out in environments where external HTTP connections can only be made via proxy[48]. Meanwhile, Necurs has functions for disabling Windows Firewall and other anti-virus software, as well as functions that receive and execute external commands like a RAT. In many cases, it is also used as a downloader for introducing other malware[49].

■ **Infector Changes**

The servers (infectors) that host landing pages or payloads are discarded at short intervals. Between July 1 and July 28, 2015, IIJ confirmed 102 IP addresses as Angler infectors, but the average lifetime (the period they were used for the same attack) of each IP address was around 1.3 days. For Nuclear infectors the average was around 1.6 days, so it does not seem that this trend is limited to Angler. However, there is some bias in the AS[50] that manage Angler infector IP addresses, with the aforementioned 102 infectors being concentrated on 10 AS, and more than three quarters of the total belonging to the top two AS (Figure 24). The same trend also applies to Nuclear, but compared to Angler it is distributed over a larger number of AS (Figure 25). Furthermore, the FQDN that Angler uses for infectors is changed at even shorter intervals than the IP address. IIJ confirmed in our observations that the A records

Figure 23: A Threatening Message Displayed by CryptoWall 3.0

Figure 24: Angler Infector IP Address Numbers Observed Between July 1 and July 28, 2015 (by AS)

Figure 25: Nuclear Infector IP Address Numbers Observed Between July 1 and July 28, 2015 (by AS)

*48 Status as of the time of writing (late July, 2015). It is possible that a function for automatically obtaining the proxy settings of an OS may be added in a future version.

*49 See "Trojan:Win32/Necurs" (https://www.microsoft.com/security/portal/threat/encyclopedia/Entry.aspx?Name=Trojan:Win32/Necurs) for general information about Necurs.

*50 Autonomous System. Collections of IP networks under the same routing policy. Generally, each unique AP comprises an ISP or large-scale hosting provider.

for individual FQDN are deleted in 30 minutes, with these domains used and promptly discarded. For this reason, countermeasures that involve IP address or FQDN blacklists are less effective*51.

■ **Countermeasures**
Drive-by download countermeasures for the networks of organizations such as companies include suitable patch management and limiting the executable area for programs in client environments, enabling functions that deal with attacks on vulnerabilities such as EMET, and enabling "click-to-play" Web browser plug-ins*52. Compulsory use of HTTP proxies, and storage of access logs are also effective ways to mitigate damage and discover issues at an early stage. It is also possible to turn the aforementioned short lifetime of FQDN to one's own advantage. For example, by implementing an operation that extracts FQDN with comparatively small access numbers from the access log, and then checks the existence of A records for these after 20 or 30 minutes, it may be possible to extract infector nodes from the access log.

For website operators and administrators, it is crucial to strictly manage Web applications and content*53. Additionally, in view of the previously mentioned fact that infectors can be concentrated on certain AS, we recommend that checks be made to see whether the operator environment of a site (PaaS or IaaS, a hosting provider, etc.) is registered to an external blacklist, etc.

1.4.3 ID Management Technology: Online Authentication Methods Not Using Passwords

When a user with a certain ID (identifier) logs in to a server, the server authenticates the attribute information for that ID using a confidential token, thereby authorizing access to a range of resources. At this time, credentials associated with the ID are distributed as certificates that guarantee the attribute information and authorization information that handle the ID. These credentials are public information, but as they are distributed behind the scenes on systems, users may only see them directly on occasion.

In the previous report*54, we discussed ID management technologies such as these. In particular, with regard to user authentication, we covered the fact that progress has been made with the combined use of tokens categorized as "something you have" or "something you are" alongside "something you know," which is represented by password authentication*55. We are entering an age in which users can elect to use powerful personal authentication methods with their important communications, albeit at the expense of convenience to a certain degree. In this report, we introduce a number of technology trends that show the password authentication that has been widely used up until now is gradually being replaced. We also take a look at examples of actual services applying these.

■ **FIDO Alliance Activities**
The FIDO (Fast IDentity Online) Alliance*56 is a nonprofit organization established in June 2012 with the goal of drawing up open standards regarding login methods that are more convenient and secure than in the past. Under FIDO specifications, online

---

*51 Some of the CryptoWall 3.0 specimens confirmed by IIJ attempted to connect to the C&C servers listed in the following Sophos Threat Analysis, which are presumed to have been used on an ongoing basis since December 2014. "Troj/Ransom-BBD" (https://www.sophos.com/en-us/threat-center/threat-analyses/viruses-and-spyware/Troj~Ransom-BBD/detailed-analysis.aspx). Using a blacklist for malware C&C servers used as an exploit kit payload could reduce damages from drive-by downloads.

*52 See Vol.21 of this report (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol21_EN.pdf) at the end of "1.4.1 The PlugX RAT Used in Targeted Attacks" for more information about malware infection countermeasures in client environments.

*53 For more details regarding alteration countermeasures from the perspective of website operators and administrators, see Vol.24 of this report under "1.4.2 The Vawtrak Malware That Steals Authentication Information, etc. for Japanese Financial Institutions" (http://www.iij.ad.jp/en/company/development/iir/024.html), or the IIJ Security Diary at the end of "Follow-up on Alteration Incidents in Japan Exploiting BHEK2" (https://sect.iij.ad.jp/d/2013/03/225209.html) (in Japanese).

*54 See Vol.27 of this report under "1.4.2 ID Management Technology: From a Convenience and Security Perspective" (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol27_EN.pdf).

*55 Faith in authentication systems that only use passwords has been shaken, due to cases in which the PCs passwords are entered into no longer being trustworthy, such as when they have a keylogger installed, or are infected with malware. List-based attacks in which reused passwords have leaked from a database used on another service are also occurring frequently. One-time passwords, in which a different password is used every time, and multifactor authentication, in which a number of other tokens are used alongside conventional password authentication, are both being used more and more as solutions to this. See Vol.26 of this report under "1.4.3 ID Management Technology" (http://www.iij.ad.jp/en/company/development/iir/pdf/iir_vol26_EN.pdf) for more information on one-time passwords and multifactor authentication.

*56 "About the FIDO Alliance" (https://fidoalliance.org/about/overview/).

authentication is achieved through a combination of both offline and online processing. This is easier to understand if you consider that conventional password authentication is carried out using online processing only. With online processing, passwords are sent from the user (Prover) to the server (Verifier) along with IDs, etc. At this time, the password must be sent after ensuring a secure channel such as TLS, as in cases where a separate secure channel is not used, passwords can easily be intercepted by a third party. Under FIDO specifications, challenge-response authentication*57 using public key cryptography is performed via online processing so that communications related to authentication do not need to pass over a secure channel. Meanwhile, the offline processing system involves local authentication of users via a FIDO certified device. When local authentication is successful, a user is permitted to use the public key for the aforementioned public key cryptography authentication. When focusing on the online processing portion alone, because the simple protocol of challenge-response authentication via public key cryptography is used, it is easy to support a variety of local authentication such as biometrics. This highly expandable nature of FIDO is drawing a lot of interest and fuelling expectations.

Broadly speaking, two specifications are currently being drawn up by the FIDO Alliance. The UAF (Universal Authentication Framework) standard deals with biometrics authentication, while the U2F (Universal 2nd Factor) standard deals with multifactor authentication*58. Regarding the UAF standard, in May 2015 NTT DOCOMO became a board member of the FIDO Alliance*59, and at the same time two products implementing the UAF standard were FIDO certified*60. These smartphones feature iris authentication or fingerprint authentication*61, and have been touted as being convenient for the passwordless login methods they offer. Meanwhile, regarding the U2F standard concerning multifactor authentication, the Google Login Service*62 is one example on the FIDO certified list. It is possible to use a U2F standard USB device as one of the security keys*63 for the two-step authentication process, and these are also sold by a number of distributors in Japan. It currently only supports Google Chrome version 40 or later browsers, but Microsoft have stated they will support it on their next OS, Windows 10*64, and use is expected to spread in the future. It was also announced that the U2F standard will be expanded for use via NFC or Bluetooth*65. Additionally, both the NIST organization that decides security standard specifications for U.S. government systems and the United Kingdom Office of the Cabinet joined the FIDO Alliance in June*66, so we can assume that use at government agencies is under consideration. NIST published a draft of their Interagency Report regarding multifactor authentication using smart cards in July*67.

■ **Use of One-Time Passwords**
Due to a series of unauthorized login incidents targeting virtual currency and virtual items in online games, a shift from password authentication to the use of one-time passwords in conjunction with other devices such as smartphones was recommended*68. On this occasion, independent measures were taken to encourage use, including offering bonuses such as upgrades to items used in the game to one-time password users. Additionally, there are services that offer data recovery whenever possible if unauthorized login incidents cause damages.

---

*57 This consists of a user (Prover) responding to a challenge from the server (Verifier) with data created using a private key that only they could know (for example, a digital signature in response to a challenge).

*58 Combining multiple authentication methods to carry out a single authentication. For example, cases in which authentication is performed by entering a one-time password in addition to the regular password.

*59 "FIDO Alliance Welcomes NTT DOCOMO, INC. to Board of Directors" (https://fidoalliance.org/fido-alliance-welcomes-ntt-docomo-to-board/).

*60 "FIDO Certified" (https://fidoalliance.org/certification/fido-certified/).

*61 NTT DOCOMO, "DOCOMO Launches 10 New Mobile Devices for 2015 Summer Lineup" (https://www.nttdocomo.co.jp/english/info/media_center/pr/2015/0513_02.html). The press release states four products are FIDO certified, but this is thought to be because the SAMSUNG products are not listed as FIDO certified.

*62 https://accounts.google.com/

*63 Google Accounts Help, "Using Security Key for 2-Step Verification" (https://support.google.com/accounts/answer/6103523?hl=en).

*64 "Microsoft Announces FIDO Support Coming to Windows 10" (http://blogs.windows.com/business/2015/02/13/microsoft-announces-fido-support-coming-to-windows-10/).

*65 "FIDO Alliance Equips U2F Protocol for Mobile and Wireless Applications" (https://fidoalliance.org/fido-alliance-equips-u2f-for-mobile-and-wireless-applications/).
Bluetooth SIG, "Bluetooth SIG and FIDO Alliance Deliver Two-factor Authentication Via Bluetooth Smart" (http://www.bluetooth.com/Pages/Press-Releases-Detail.aspx?ItemID=233).

*66 "FIDO Alliance Announces Government Membership" (https://fidoalliance.org/fido-alliance-announces-government-membership/).

*67 NIST IR 8055, "Derived Personal Identity Verification (PIV) Credentials (DPC) Proof of Concept Research" (http://csrc.nist.gov/publications/drafts/nistir-8055/nistir_8055_draft.pdf). At the time of writing public comments are being accepted.

*68 For example, GungHo Games, Security Measures (http://www.gungho.jp/security/security.html) (in Japanese); NEXON Support, One-Time Passwords (http://www.nexon.co.jp/support/security/otp-guide.aspx) (in Japanese); GMO Gamepot, Using One-Time Passwords.

The use of one-time passwords is also growing in online banking. Up to now, online banking logins generally involved regular authentication using a password or PIN number. When performing important transactions such as bank transfers while logged in, user authentication was previously carried out by sending a card containing a table of random numbers by secure post such as registered letter, and prompting a different combination of these random numbers (password) to be entered for each transaction. There was an issue with the content of this random number table gradually leaking after many transactions if an attacker intercepted the details when the random number card was used. For this reason, attempts were made to bolster security by prompting users to enter the date the card was issued in addition to the password written on the random number table. However, this has limitations, so a switch was made to dedicated devices where the password automatically changes after a set period. Subsequently, dedicated devices with an input interface came to be used. With dedicated devices it is possible to link input and transactions, and there are moves to improve the security of authentication by switching to a system in which the destination account number has to be entered, for example. Currently this transaction authentication is only supported by some banks, but the same dedicated devices are used at other banks as well, so they are expected to support it in the future.

Furthermore, one-time passwords that do not require a device are increasingly used instead of those that utilize a dedicated device or smartphone app. While both use the same terminology, the concepts behind them are different. First, as a prerequisite the user must be able to receive email or SMS. When performing a regular password-based login, a token that can only be used for a short time (a one-time password) is generated randomly on the server, and sent via email or SMS. Under this system, the user sends the one-time password they received to the server to complete authentication. Separating the channel used to perform login and the channel used to receive the one-time password, for example, using both the Internet and a mobile phone network, has the benefit of improving security unless both are intercepted. Because dedicated devices are not used, systems can be operated with a comparatively low cost. In terms of assigning a password only for temporary use on each device such as smartphones to protect the original password, this could be considered a derivative password.

■ Social Login

More and more sites now offer a procedure called social login. This involves performing authentication using SNS or portal sites as an IdP, or in other words using them merely as a means of user authentication, instead of carrying out ID and password management independently on the servers providing given resources. This system makes it possible for users to log in without performing user registration on a new server, by granting permission to connect with a server they are already using when a new server is used. This has the benefit of not requiring servers and users to manage new passwords. However, when permitting connections to a new server using an ID that was previously used in a limited scope, such as within a certain SNS, it is necessary to take note of the transfer of information between the connected servers.

Some servers attempt to access an excessive amount of data for information gathering purposes, etc., and there have been incidents in which unwanted messages have been posted to SNS. Furthermore, there is a possibility that independent activities originally carried out under different IDs in different realms may be connected as the actions of the same person based on the information exchanged between linked servers. Consequently, in some cases prudence will be required on the part of the user, such as utilizing functions for switching between IDs so they act as separate entities (derivative IDs) while logged in, or carrying out procedures to ensure that links are removed after using a given server.

■ **Risk-Based Authentication**

Last of all, we will examine risk-based authentication. In July of this year, the Ministry of Internal Affairs and Communications published a field study*69 regarding password management on major services. In addition to the usable characters and length limit for configuring passwords, as well as whether passwords were hashed, the survey also covered items related to login attempts from the same IP address. These days there are many cases in which IDs are locked (temporarily disabling the IDs) or multifactor authentication carried out in addition to normal password authentication, when a difference is detected in the user environment, such as abnormal behavior during authentication, including out-of-the-ordinary logins from outside Japan. Rather than unilaterally spoiling the convenience of a service, such as always forcing users to perform multifactor authentication, intermediate authentication methods that take user convenience into account have begun to be used. These involve considering the "risk" of what could happen, and only shifting to a stronger authentication method when the server automatically detects a difference in user behavior.

In this way, the total cost of authentication is being optimized, taking into account user convenience and risks. Because the total cost changes based on the IT literacy of the user carrying out authentication, it will be necessary to find the optimal system for each organization. Users are also constantly fluctuating and not fixed, calling for reevaluation as the situation demands.

1.5 Conclusion

This report has provided a summary of security incidents to which IIJ has responded. In this report, we discussed machine learning and security, and the Angler Exploit Kit that is wreaking havoc. We also examined the actual use of ID management technology. IIJ makes every effort to inform the public about the dangers of Internet usage by identifying and publicizing incidents and associated responses in reports such as this.

Authors:
Mamoru Saito
Manager of the Office of Emergency Response and Clearinghouse for Security Information, Service Operation Division, IIJ. After working in security services development for enterprise customers, Mr. Saito became the representative of the IIJ Group emergency response team, IIJ-SECT in 2001, participating in FIRST, an international group of CSIRTs. Mr. Saito serves as a steering committee member of several industry groups, including Telecom-ISAC Japan, Nippon CSIRT Association, Information Security Operation providers Group Japan, and others.

Hirohide Tsuchiya (1.2 Incident Summary)
Hirohide Tsuchiya, Tadaaki Nagao, Hiroshi Suzuki, Hisao Nashiwa (1.3 Incident Survey)
Tadaaki Nagao (1.4.1 Machine Learning and Security)
Hisao Nashiwa, Hiroshi Suzuki (1.4.2 Angler Exploit Kit on the Rampage)
Yuji Suga (1.4.3 ID Management Technology: Online Authentication Methods Not Using Passwords)
Office of Emergency Response and Clearinghouse for Security Information, Service Operation Division, IIJ

Contributors:
Takahiro Haruyama, Minoru Kobayashi, Tadashi Kobayashi, Masahiko Kato, Masafumi Negishi, Yasunari Momoi, Hiroyuki Hiramatsu, Office of Emergency Response and Clearinghouse for Security Information, Service Operation Division, IIJ

*69　Ministry of Internal Affairs and Communications, "Results of our field study on the management/operation of IDs/passwords related to Web services" (http://www.soumu.go.jp/menu_news/s-news/01ryutsu03_02000099.html) (in Japanese).

Comparing Broadband and Mobile Traffic

2.1 Overview

In this report we analyze traffic over the broadband access services operated by IIJ every year and present the results[1][2][3][4][5][6]. We once again report on changes in traffic trends over the past year based on daily user traffic and usage by port. From this report, we will also report on traffic trends for mobile service traffic in the same way.

Figure 1 shows average monthly traffic trends for IIJ's overall broadband services and mobile services. The IN/OUT traffic indicates directions from an ISP's perspective. IN represents uploads from users, and OUT represents user downloads. Because we cannot disclose traffic numbers, the data for each is normalized to 1 for the maximum value.

For broadband, over the past year IN traffic has increased by 15%, while OUT traffic has increased by 38%. Over the last eight years download volumes have grown significantly more than upload volumes, with IN traffic increasing by factor of 1.6, and OUT traffic increasing by a factor of 3.8.

For mobile we only present figures for the past year, but over this period traffic grew rapidly, with IN traffic increasing by a factor of 3.5 and OUT traffic increasing by a factor of 4.2. However, the total volume of mobile traffic is still an order of magnitude lower than broadband.

2.2 About the Data

For broadband traffic, as with our previous reports, the survey data utilized here was collected using Sampled NetFlow from the routers accommodating fiber-optic and DSL broadband customers of our personal and enterprise broadband access services. For mobile traffic, access gateway billing information was used to determine usage amounts for personal and enterprise mobile services, while Sampled NetFlow data from the routers accommodating these services was employed to determine the ports used.

Because traffic trends differ between weekdays and weekends, we analyze a full week of traffic. In this case, we have used data for the week spanning June 1 to June 7, 2015. For broadband traffic we compare results with those for the week spanning May 26 to June 1, 2014, which we analyzed in the previous report.
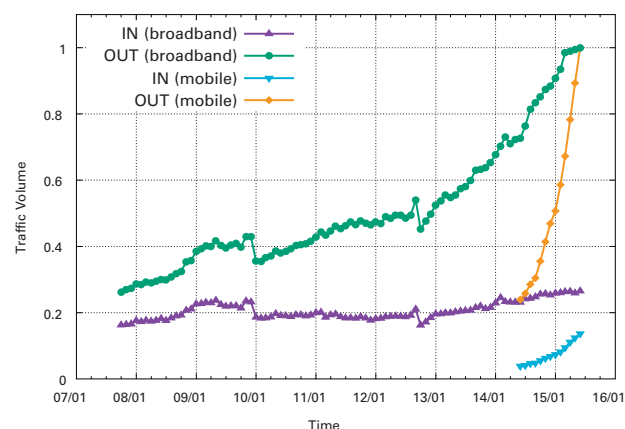

Figure 1: Trends in Monthly Traffic for Broadband and Mobile

Results are aggregated by subscription for broadband traffic, and by phone number for mobile traffic, as some subscriptions cover a number of phone numbers. The usage volume for each broadband user was obtained by matching the IP address assigned to users with the IP addresses observed. We collected statistical information by sampling packets using NetFlow. The sampling rate was set between 1/8192 and 1/16384, taking into account router performance and load. We estimated overall usage volumes by multiplying observed volumes by the reciprocal of the sampling rate. IIJ provides both fiber-optic and DSL access for its broadband services. However, fiber-

*1   Kenjiro Cho. Broadband Traffic Report: Traffic Volumes Rise Steadily Over the Past Year, and HTTPS Use Expands. Internet Infrastructure Review. Vol.24. pp28-33. August 2014.

*2   Kenjiro Cho. Broadband Traffic Report: The Impact of Criminalization of Illegal Downloads was Limited. Internet Infrastructure Review. Vol.20. pp32-37. August 2013.

*3   Kenjiro Cho. Broadband Traffic Report: Traffic Trends over the Past Year. Internet Infrastructure Review. Vol.16. pp33-37. August 2012.

*4   Kenjiro Cho. Broadband Traffic Report: Examining the Impact of the Earthquake on Traffic on a Macro Level. Internet Infrastructure Review. Vol.12. pp25-30. August 2011.

*5   Kenjiro Cho. Broadband Traffic Report: Traffic Shifting away from P2P File Sharing to Web Services. Internet Infrastructure Review. Vol.8. pp25-30. August 2010.

*6   Kenjiro Cho. Broadband Traffic: Increasing Traffic for General Users. Internet Infrastructure Review. Vol.4. pp 18-23. August 2009.

optic access now makes up the vast majority of use, with 95% of users observed in 2015 using fiber-optic connections, accounting for 98% of overall broadband traffic volumes.
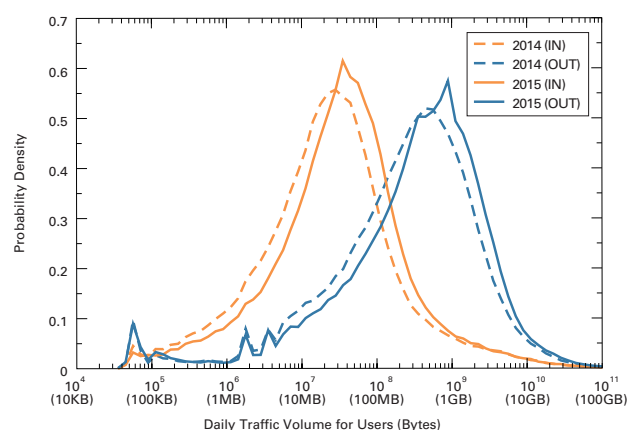
2.3 Daily Usage Levels for Users

First, we will examine the daily usage volumes for broadband and mobile users from several perspectives. Daily usage indicates the average daily usage calculated from a week's worth of data for each user.
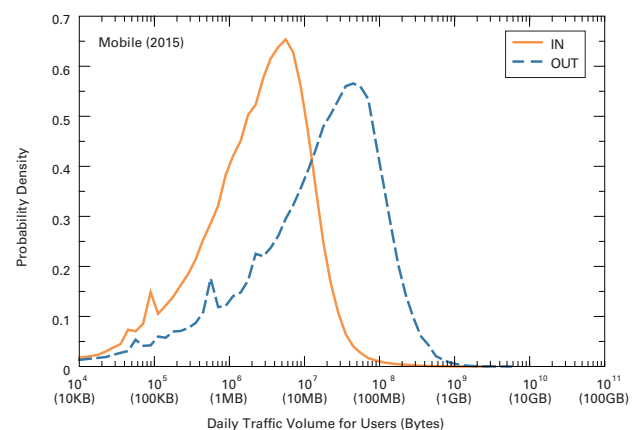
Figure 2 and Figure 3 show the average daily usage distribution (probability density function) per broadband and mobile user. They compare data for 2014 and 2015 divided into IN (upload) and OUT (download), with user traffic volume on the X axis, and user frequency on the Y axis.The X axis shows volumes between 10 KB (10^4) and 100 GB (10^{11}) using a logarithmic scale. Some users are outside the scope of the graph, but most fall within the 100 GB (10^{11}) range.

The IN and OUT distribution for broadband shows almost log-normal distribution, which looks like a normal distribution in a semi-log graph. A linear graph would show a long-tailed distribution, with the peak close to the left end and a slow decay towards the right. The OUT distribution is further to the right than the IN distribution, indicating that the download volume is more than an order of magnitude larger than the upload volume. Comparing 2014 and 2015, the peak distribution for both IN and OUT traffic has moved slightly to the right, demonstrating that overall user traffic volumes are increasing. Looking at OUT distribution, the peak has been steadily moving to the right over the past few years. However, the usage levels of heavy users on the right end have not increased much, and the distribution is beginning to lose its symmetry. Meanwhile, the tail of the IN distribution to the right has grown longer. Previously, both IN and OUT showed a clearer peak here, indicating heavy users with symmetrical IN/OUT volumes. For convenience, we labeled users with asymmetrical IN/OUT traffic distribution that make up the majority "client-type users," and the distribution of heavy users with symmetrical IN/OUT traffic that make up the minority on the right side "peer-type users." In this report we will continue to use these conventions. Over the past few years, the peak for peer-type users has shrunk to the point where it can hardly be distinguished. This indicates that the ratio of heavy users is decreasing.

The data for mobile traffic in Figure 3 indicates that usage volumes are significantly lower than broadband. Additionally, because there are limits on data usage, the ratio of heavy users to the right of the distribution is lower, creating left-right asymmetry. There are also no extremely heavy users. Due to those who use mobile only when going out and limits on data usage, there is greater variance in daily usage volumes for each user compared to broadband. For this reason, when you look at the daily average for a week worth of data, there is less variance between users than when examining individual days. Plotting distribution for individual days in the same way results in lower peaks, and raises the tails on either side, but the basic form and most frequent values of the distribution remain largely unchanged.

Figure 2: Daily Broadband User Traffic Volume Distribution Comparison of 2014 and 2015

Figure 3: Daily Mobile User Traffic Volume Distribution in 2015

Table 1 shows trends in the mean daily traffic value for broadband users as well as the modal value (the most frequent value that represents peak distribution). This year the peak position has shifted slightly from the center of the distribution so that the modal value was corrected towards the center of the distribution. Comparing the modal values in 2014 and in 2015, IN rose from 28 MB to 40 MB, and OUT rose from 447 MB to 708 MB. This demonstrates that, particularly in the case of downloads, the traffic volume for each user has increased. Meanwhile, because mean values are pulled up by the heavy users to the right of the graph, they are significantly higher than the modal values, with the mean IN value 467 MB and the mean OUT value 1,621 MB in 2015. In 2014, these were 437 MB and 1,287 MB, respectively. For mobile, as shown in Table 2, there is little difference between the mean and modal values due to the lack of extremely heavy users. The modal values were 5.5 MB for IN and 40 MB for OUT, with the mean values amounting to 6.0 MB for IN and 46.6 MB for OUT.

Figure 4 and Figure 5 plot the IN/OUT usage volumes for 5,000 randomly sampled users. The X axis shows OUT (download volume) and the Y axis shows IN (upload volume), with both using a logarithmic scale. Users with identical IN/OUT values are plotted on the diagonal line.
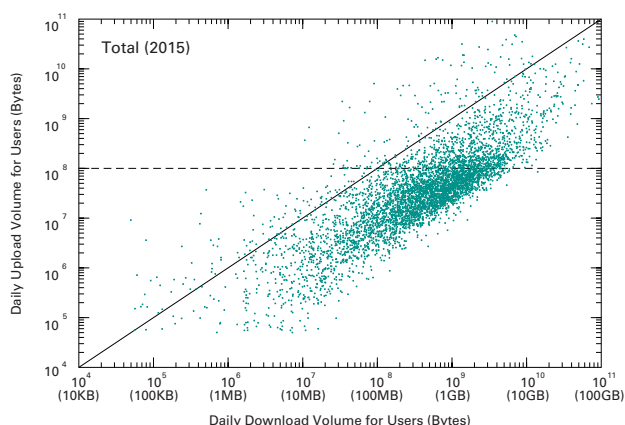
The cluster below the diagonal line and spread out parallel to it represents general client-type users with download volumes an order of magnitude higher than upload volumes. For broadband traffic, there was previously a clearly-recognizable cluster of peer-type heavy users spread out thinly on the upper right of the diagonal line, but this is now no longer discernible. Though we have separated client-type and peer-type users for convenience, in actual fact client-type general users also use peer-type applications such as Skype, and peer-type heavy users also use download-based applications on the Web, blurring the boundary between them. In other words, many users use both types of applications in varying ratios. There are also differences in the usage levels and IN/OUT ratio for each user, pointing to the existence of diverse forms of usage. Here, almost no difference can be discerned compared to 2014.

Year	IN (MB/day)		OUT (MB/day)	
	Mean	Mode	Mean	Mode
2005	430	3.5	447	32
2007	433	4	712	66
2008	483	5	797	94
2009	556	6	971	114
2010	469	7	910	145
2011	432	8.5	1,001	223
2012	410	14	1,026	282
2013	397	18	1,038	355
2014	437	28	1,287	447
2015	467	40	1,621	708

Table 1: Trends in Mean and Mode of Daily Traffic Volume for Each Broadband User

Year	IN (MB/day)		OUT (MB/day)	
	Mean	Mode	Mean	Mode
2015	6.0	5.5	46.6	40

Table 2: Trends in Mean and Mode of Daily Traffic Volume for Each Mobile User

Figure 4: IN/OUT Usage for Each Broadband User

Figure 5: IN/OUT Usage for Each Mobile User

The trend for OUT traffic to be an order of magnitude larger also applies to mobile, but usage volumes are lower than broadband, and there is less variance between IN/OUT.
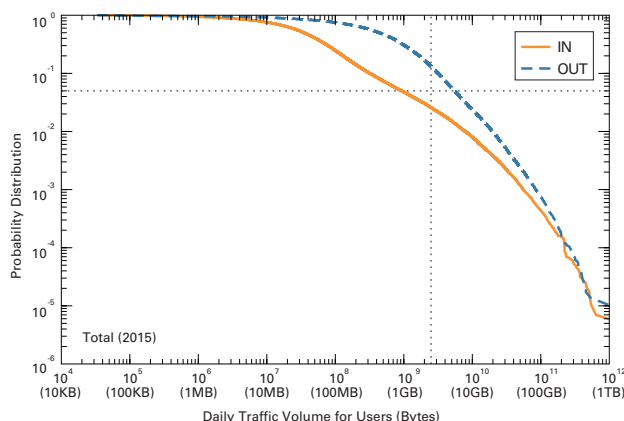
Figure 6 and Figure 7 show the complementary cumulative distribution of the daily traffic volume for users. This indicates the percentage of users with daily usage levels greater than the X axis value on the Y axis in a log-log scale, which is an effective way of examining the distribution of heavy users. The right side of the graph falls linearly, showing a long-tailed distribution close to power-law distribution. It can be said that heavy users are distributed statistically, and are by no means a special class of user. Heavy users also exhibit power-law distribution for mobile, but the ratio of heavy users is lower, and they also have OUT usage levels several times higher than their IN amounts.

There is a great deal of deviation in traffic usage levels between users, and as a result traffic volume for a small portion of users accounts for the majority of overall traffic. For example, the top 10% of broadband users make up 65% of the total OUT traffic, and 90% of the total IN traffic. Furthermore, the top 1% of users make up 29% of the total OUT traffic, and 63% of the total IN traffic. Along with the decrease in the ratio of heavy users over the past few years, the distribution bias is also dropping slightly.
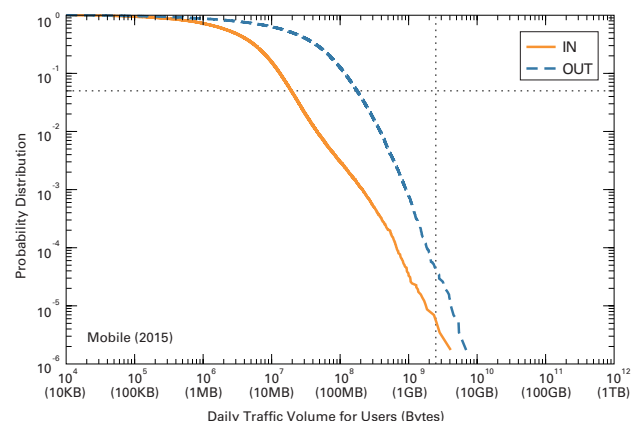
For mobile, the top 10% of users account for 49% of OUT traffic and 50% of IN traffic, while the top 1% make up 13% of OUT traffic and 19% of IN traffic. This also demonstrates the low ratio of heavy users among mobile users.

2.4 Usage by Port

Next, we will look at a breakdown of traffic and examine usage levels by port. Recently, it has been difficult to identify applications by port number. Many P2P applications use dynamic ports on both ends, and a large number of client/server applications utilize port 80 assigned to HTTP to avoid firewalls. To broadly categorize, when both parties use a dynamic port higher than port 1024, there is a high possibility of it being a P2P application, and when one party uses a well-known port lower than port 1024, it is likely to be a client/server application. In light of this, here we will look at usage levels for TCP and UDP connections by taking the lower port number of the source and destination ports.

Figure 6: Complementary Cumulative Distribution of the Daily Traffic Volume for Broadband Users

Figure 7: Complementary Cumulative Distribution of the Daily Traffic Volume for Mobile Users

As overall traffic is dominated by peer-type heavy user traffic, to examine trends for client-type general users, we have taken the rough approach of extracting data for users with a daily upload volume of less than 100 MB, and treating them as client-type users. This corresponds to users below the horizontal line at the IN=100 MB point in Figure 4.

Table 3 compares port usage ratios for broadband users in 2014 and 2015, showing total users and client-type users.

81% of overall traffic in 2015 is TCP based. The ratio of port 80 (HTTP) traffic was down from 45% in 2014 to 38% this year. The ratio of port 443 (HTTPS) traffic climbed from 9% to 23%. TCP dynamic port traffic, which has been on the decline, dropped from 24% in 2014 to 18% in 2015. The ratio of individual dynamic port numbers is tiny, with port 1935 used by Flash Player the highest at 2% of the total, and the next highest under 0.5%. Aside from TCP, there is port 443 (HTTPS) UDP traffic, which is thought to be Google's QUIC protocol. The rest is mostly VPN-related.

Looking exclusively at client-type users, port 80 traffic that accounted for 75% of the total in 2014 dropped 22 percentage points to 53% in 2015. Meanwhile, the next highest traffic ratio for port 443 (HTTPS) traffic climbed 21 percentage points, going from 14% in 2014 to 35% in 2015. We believe this was due to the shift of some HTTP traffic to HTTPS. The ratio of dynamic ports also decreased from 7% to 5%.

Table 4 shows port usage ratios for mobile users, which as a whole were close to the values for client-type users on broadband.

The growth in the use of HTTPS is due to more and more services making regular use of encrypted HTTPS communications since the existence of a U.S. National Security Agency (NSA) program for intercepting communications stirred up controversy in June 2013. Looking at HTTPS traffic volumes broken down by provider for 2015, about 80% was related to Google, demonstrating their efforts to proactively adopt HTTPS. We also believe YouTube traffic volumes are boosting HTTPS usage.

protocol port	2014		2015	
	total (%)	client type	total (%)	client type
TCP	80.2	97.4	80.8	94.7
(< 1024)	56.3	90.1	63.3	89.9
80 (http)	44.9	74.8	37.9	53.2
443 (https)	9.3	13.8	23.3	35.1
81	0.0	0.0	0.5	0.7
182	0.0	0.0	0.4	0.3
554 (rtmp)	0.4	0.3	0.2	0.1
22 (ssh)	0.3	0.0	0.2	0.0
(>= 1024)	23.8	7.3	17.5	4.8
1935 (rtmp)	2.5	4.0	1.8	2.4
8080	0.4	0.2	0.3	0.1
7144 (peercast)	0.3	0.0	0.2	0.0
UDP	12.5	1.8	11.4	2.6
443 (https)	0.0	0.0	0.9	0.9
1701 (12tp)	0.0	0.0	0.2	0.3
ESP	6.9	0.7	7.4	2.6
IP-ENCAP	0.2	0.0	0.2	0.0
GRE	0.2	0.0	0.2	0.0
ICMP	0.0	0.0	0.0	0.0

Table 3: Usage by Port for Broadband Users

protocol port	2015
	total (%)
TCP	93.8
80 (http)	52.5
443 (https)	37.4
81	0.5
993 (imaps)	0.5
1935 (rtmp)	0.5
UDP	5.2
1701 (12tp)	1.8
443 (https)	1.0
4500 (nat traveral)	0.3
80 (http)	0.2
53 (dns)	0.1
ESP	0.7
GRE	0.3
ICMP	0.0

Table 4: Usage by Port for Mobile Users

Figure 8 compares trends in TCP port usage over a week for overall broadband traffic in 2014 and 2015. Trends in TCP port usage are shown for four categories: port 80, port 443, other well-known ports, and dynamic ports. Traffic is normalized to 1 for the total peak traffic volume. Previously we included port 443 under other well-known ports, but due to the increase in its usage ratio, from this report we are plotting port 443 separately. Compared with 2014, we can see that the overall ratio of port 443 usage has increased further, and the use of dynamic ports is decreasing. The overall peak is between 21:00 and 1:00 the next day, and traffic also increases in the daytime on Saturday and Sunday, reflecting times when the Internet is used at home.

Figure 9 shows trends for port 80 and port 443 with regard to mobile, where traffic remains high from morning to night. On weekdays there are three peaks representing the morning commute hours, lunch break, and evening to night periods, demonstrating that usage times are different from broadband.

2.5 Conclusion

One trend we have seen in broadband traffic over the past year is significant growth in HTTPS usage, which we reported on in our last report. Other than this there were no major changes in overall trends, although this year download volumes increased 38%, and upload volumes climbed 15%. A year before these increases were 27% and 13%, so over the past year the rate of increase has also gone up.

Although still falling far short of broadband volumes, the mobile traffic we began reporting on in this volume has grown significantly in the past year. Due to bandwidth and data cap limitations, there are less heavy users for mobile compared to broadband. Usage ratios by port are close to light broadband users, suggesting the use of the same kind of services as broadband on mobile due to the increasing popularity of cloud-based services accessible from multiple devices. Meanwhile, some points were different to broadband, such as the standout usage times being the commuting hours and lunch breaks on weekdays.
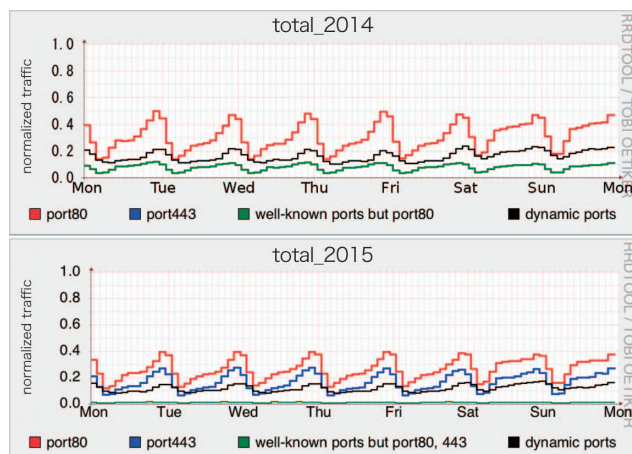

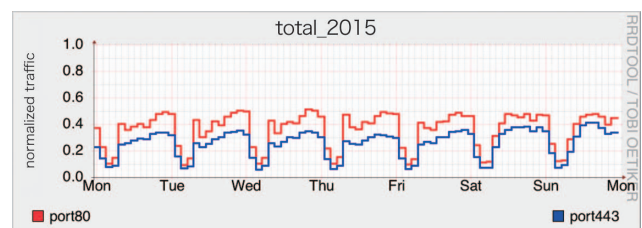**Figure 8: Weekly TCP Port Usage Trends for Broadband Users in 2014 (top) and 2015 (bottom)**

Figure 9: Weekly TCP Port Usage Trends for Mobile Users in 2015

Author:
Kenjiro Cho
Research Director, Research Laboratory, IIJ Innovation Institute Inc.
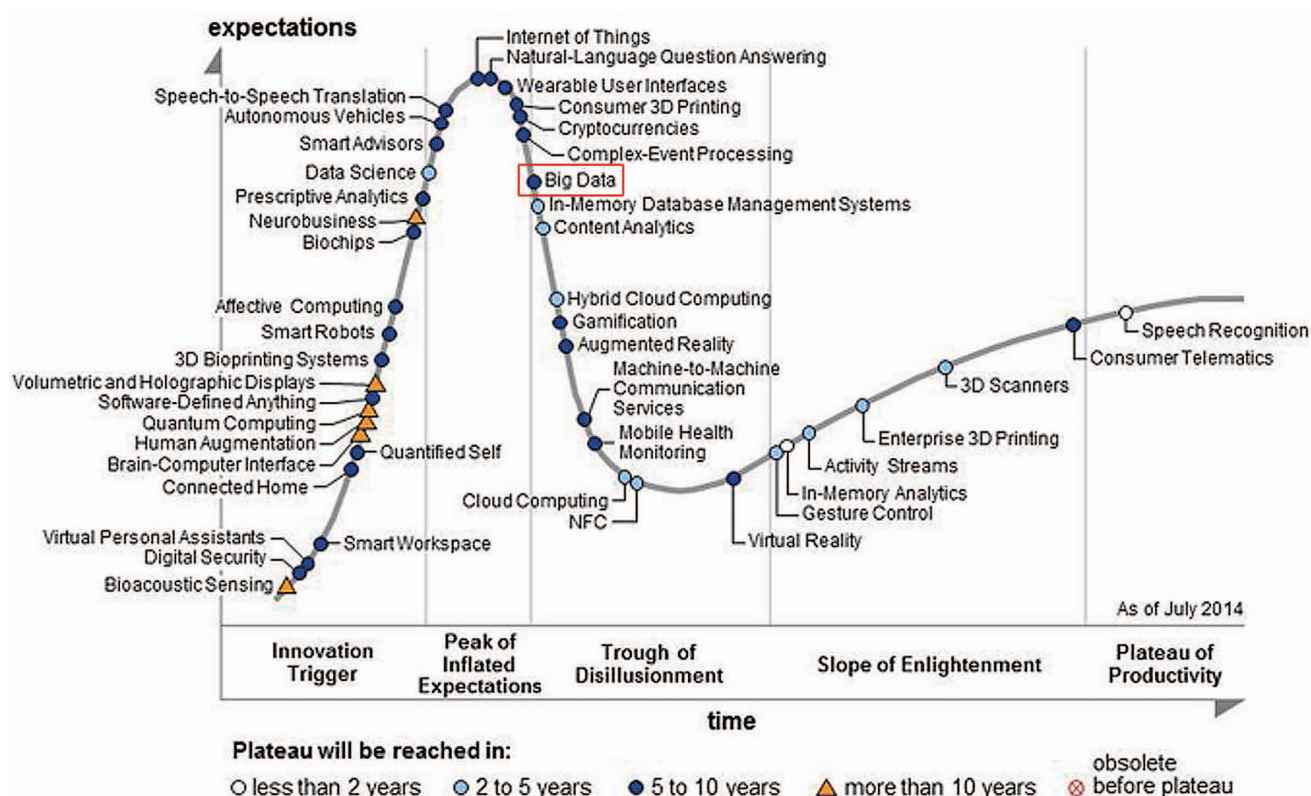
Social Big Data

3.1 The Current State of Big Data

Like "cloud computing," the term "big data" is extremely conceptual. Its definition remains vague, and what it means can change over the course of time, so it is often used in a very abstract way. For this reason I believe the term will survive for a long time. However, when it comes to "big data" in the constantly changing world of buzzwords, I get the feeling that as of 2015 it is coming to the end of its shelf life. This impression was also given in the Hype Cycle for Emerging Technologies published each year by Gartner, who stated in 2014 that big data was moving toward the Trough of Disillusionment, so there is likely some substance to this opinion (Figure 1).

However, because it is still towards the end of the Peak of Inflated Expectations in the Hype Cycle limited to Japan, it seems "big data" as a buzzword occupies an extremely tenuous position (Figure 2).

■ **Big Data Trends in Japan**

In any case, while "big data" has remained a buzzword that symbolized technology trends over the past few years, it has begun to show its age. I feel that other slightly more specific buzzwords, such as the "Internet of Things (IoT)", are now touted more often.

The Internet of Things (IoT) is a term put forward in 1999 by Kevin Ashton, who contributed to the creation of a global standard system for RFID. The concept involves networking a range of objects by assigning IDs to them, literally building an Internet of things. Another extremely similar concept is the "Cyber-Physical System (CPS)," which is defined as a system in which computational elements that control physical entities collaborate by sharing information. Both concepts involve bringing the physical world and cyber world together[*1].

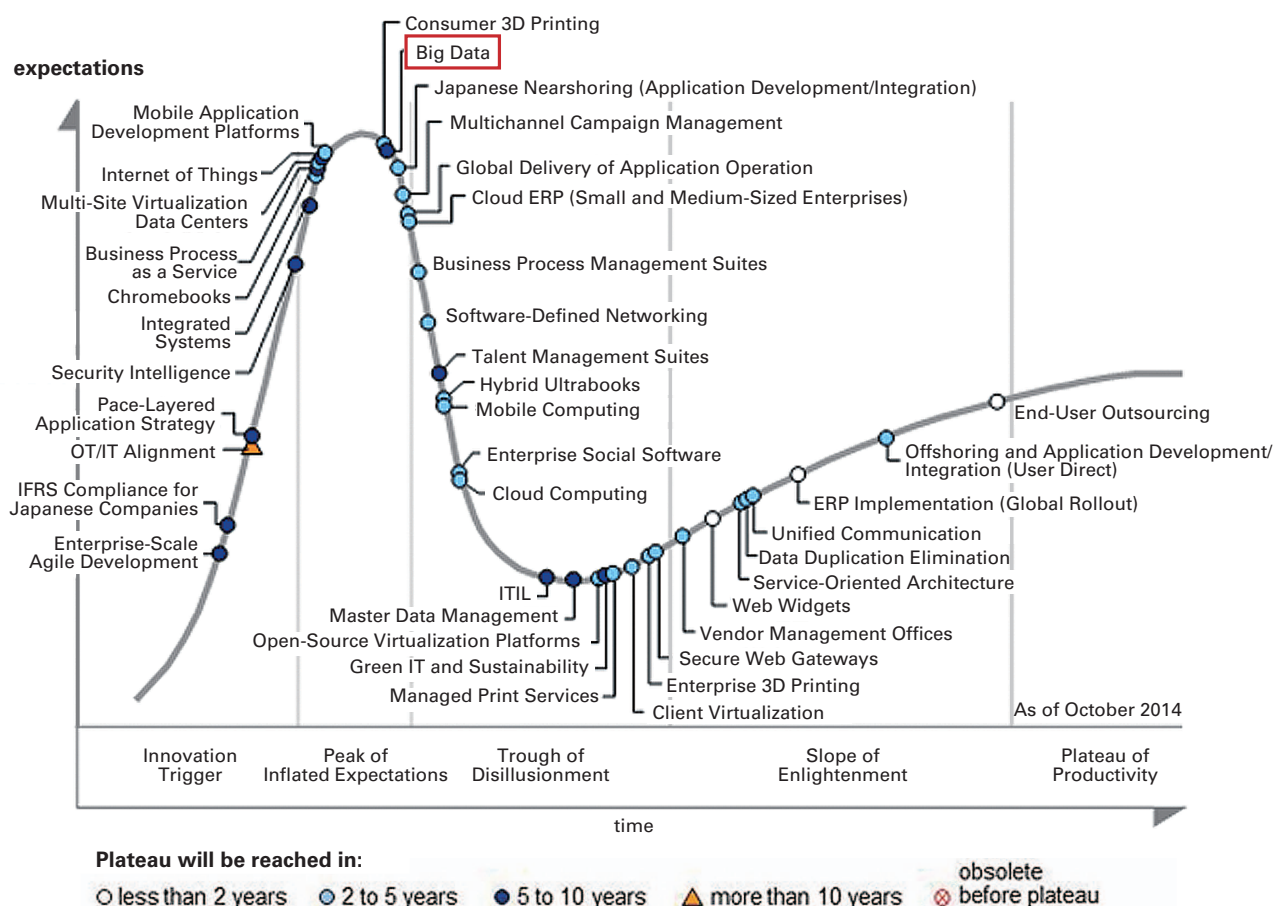**Figure 1: 2014 Hype Cycle for Emerging Technologies**

---

*1   "Cyber Physical Systems and Internet of Things: Connecting the real world with the cyber world" Kazuo Iwano, Yosuke Takashima, Journal of Information Processing and Management 2-15 vol.57 no.11 (https://www.jstage.jst.go.jp/browse/johokanri/57/11/_contents/-char/en/).

Today, IDs are already assigned to objects such as parcel delivery tags, contactless IC cards used as commuter passes, and smartphones, making it possible to track their movement, etc. Accordingly, the location information continually generated for each ID is big data in a literal sense, and by analyzing this it is possible to perform a range of optimizations. This is the "brighter future" that IoT and CPS are aimed at bringing about. For example, I would hazard a guess that many people have seen proposals for projects such as smart grids and smart cities.

In other words, IoT and CPS are massive paradigms aimed at revolutionizing and reforming social systems in a fundamental way. Research and development related to these technologies has the potential to directly impact many people living in society. One example is the issue of privacy for ordinary citizens[2]. For researchers interested in methods for analyzing big data such as myself, I believe that research into IoT and CPS is an area fraught with difficult challenges, considering the social aspects that must also be taken into account.

I think the manifestation of social issues such as these is evidence that progress has been made in big data research and development. In 2015, research into big data is currently in a temporary lull where we must carefully consider the direction in which we take it.

Source: Gartner (October 2014)

Figure 2: 2014 Technology Hype Cycle in Japan

---

*2 NICT (November 25, 2013, press release), "Proof-of-Concept Tests for Using ICT Technology at Large-Scale Complex Facilities Conducted at Osaka Station City" (http://www.nict.go.jp/press/2013/11/25-1.html) (in Japanese).
Nikkei Computer (March 11, 2014, Naoki Kiyoshima), "Facial Recognition Proof-of-Concept Tests at JR Osaka Station Building Postponed Due to Concerns About Invasion of Privacy" (http://itpro.nikkeibp.co.jp/article/NEWS/20140311/542723/) (in Japanese).
Nikkei Computer (November 7, 2014, Naoki Kiyoshima), "Facial Recognition Tests Using Security Cameras at JR Osaka Station Building to be Resumed with Limited Scope" (http://itpro.nikkeibp.co.jp/atcl/news/14/110701801/) (in Japanese).

I think the approach of linking the real world and cyberspace that IoT and CPS are oriented towards actually opens up new possibilities for big data analysis, in the sense that it creates data from facts. For example, if we could utilize data indicating facts such as "a certain person was in a set place at a given time," or "a certain location at a certain plant reached a set temperature at a given time" in analysis, we could treat at least these facts as conclusive information (logically unvarying). I surmise that this would contribute to raising the accuracy of analysis results. Furthermore, there are social demands requiring analysis and prediction to be carried out based on this kind of reliable data. For example, tasks such as the monitoring of nuclear power plant operation and the tracking of criminals have high urgency, and data that indicates facts is immensely valuable for analysis when certainty is regarded as important.

3.2 Social Big Data

Here I will refer to the data indicating facts that is created by IoT as "factual big data." Data with contrasting properties, for example the data used in analysis that takes into account individual impressions or preferences like "things that enrich your lifestyle" or "things to your taste," can be called "social big data." Messages circulated via social media such as Twitter and Facebook would be typical examples of this.

Social big data is generally interpreted as a hodgepodge of miscellaneous information. However, according to former Amazon Chief Scientist Andreas Weigend, the amount of social big data is growing rapidly, with the total amount of data generated in the year of 2009 exceeding the total amount generated up to 2008 since the dawn of history. He reports that this has brought about revolutionary change for the search and collection of customer product information[3].

Weigend points out that more and more social media users are sharing posts with explicit details on impressions of manufacturers and sites, or products they or their friends have purchased. This tends to be more effective than information provided by manufacturers as part of their advertising or public relations. Some manufacturers have also put in place systematic methods that encourage users to provide candid feedback, and reward them for valuable data.

Against this backdrop, the world of online marketing is currently shifting to a model in which specific data beneficial to third parties is created through collaborations between users (from e-business to me-business). Weigend calls this the "Social Data Revolution." The analysis of social big data may present answers to questions such as, "What can customer expectations bring about?" and "What can companies do to meet such expectations?" It is clear that the amount of big data on the Internet will continue to grow, and I believe that the spread of IoT and rising prominence of social media is accelerating the factors that are prompting the bipolarization of specific characteristics for the big data in circulation.

3.2.1 Properties of Social Big Data

Comparing social big data with factual big data from the perspective of big data analysis, its unique properties include the difficulty of machine generation, its multifactorial implications, the fact that ambiguity is acceptable, and the fact that conjecture is allowed (errors are tolerated). To use a slightly abstract expression, you could say that while factual big data is hard data generated by machines, social big data is soft data generated by people.

Normally, when people are involved in data generation, it becomes harder to analyze that data. For example, imagine you are surveying the most pleasant area in a large park. If you assume that "pleasant" means that it has a suitable temperature, you could place 10,000 devices with temperature sensors in the park, and build a system that collects this data around the clock. Analyzing the factual big data accumulated in this way would provide extremely clear results. Devices could be placed in every nook and cranny of the park, and once installed you could obtain temperature data for that location with absolute certainty. With 10,000 devices, a few are likely to fail, but as long as this can be confirmed you could simply replace them straight away. This would allow you to thoroughly investigate which areas have a temperature that people find pleasant.

---

*3   The Social Data Revolution(s) Andreas Weigend May 20, 2009 (https://hbr.org/2009/05/the-social-data-revolution.html).

In contrast, hiring 10,000 surveyors and asking them to measure the temperature at each point would be the social big data method of surveying and analyzing this data. Honest surveyors may use a thermometer to produce reports stating that "point XX was YY degrees," resulting in accuracy approaching that of devices fitted with temperature sensors. That said, there would likely also be surveyors that eschewed the use of a thermometer, and produced extremely vague reports stating things like "the area around the lake is cold." Others may check a Fahrenheit thermometer and answer "72 degrees," or just lie in their report and not make the trip to each point. There could even be surveyors who make irrelevant reports, such as "the wooded area is pleasant." This would make it very difficult to find points with a suitable temperature within the park.

However, if you asked the surveyors to find points that seem pleasant, the results of the survey would probably be very different. The surveyors would start by heading to places that look pleasant. They may run into other surveyors on the way, compare opinions, and seek likely places together, eventually settling on the place that seems the best when they find somewhere they both agree on. As a result, over time we could expect surveyors to gather at certain points in the park. Judging the right time, we could interview individual surveyors on their findings. By asking questions such as where they are, whether that place is pleasant, and why they think so, you could obtain information on pleasant points in the park that third parties can sympathize with.

The question of which survey result is more applicable depends on the purpose of the survey. For staff in charge of park management or environmental conservation, factual big data is probably all that is required. However, for the owner of a kiosk in the park, survey results based on social big data may actually be more desirable. The decisive difference between these two examples is that factual big data produces results comprised of objective data based on homogenous facts, while social big data is an aggregation of data based on the subjective view of surveyors (people). Both are big data, so consideration must be given to reliability to prevent erroneous data. Social big data also requires analysis techniques that take into account issues with trustworthiness, such as whether the data is correct or contains errors, and how much it can be trusted.

3.3 Wikipedia as Big Data

Many people may think of the messages circulated over Twitter when social big data is mentioned, but my focus is on Wikipedia.

As Wikipedia is the most popular electronic encyclopedia service, I doubt there is any need for me to give an overview of it here, but Wikipedia is a notable site when it comes to big data as well. On the Japanese[4] or English[5] "Wikipedia:Statistics" pages, you can view an up-to-date pageview count for articles on Wikipedia, etc.

As of July 20, 2015, Wikipedia has a total of 4,920,887 articles and 36,748,410 pages (including peripheral information). Even limiting results to Japanese, the total number comes to 974,894 articles and 2,785,007 pages. This data can also be modified or reproduced for secondary use based on the terms of the CC-BY-SA 3.0[6] license. In terms of licensing, you could say a site like this that can be used in the same way as open source software is the most secure big data.

Wikipedia data can be traced from a page called "Wikimedia Download"[7]. Wikipedia constantly has data backup tasks operating, and articles in each language are regularly backed up. To my knowledge, new backups are created around once a month.

*4   https://ja.wikipedia.org/wiki/Wikipedia:%E7%B5%B1%E8%A8%88
*5   https://en.wikipedia.org/wiki/Wikipedia:Statistics
*6   Creative Commons (https://en.wikipedia.org/wiki/Creative_Commons)
*7   http://dumps.wikimedia.org/

This Wikipedia data is already utilized in a variety of areas. One prominent example is DBpedia*8*9, a project that generates Linked Open Data (LOD) from Wikipedia data to create a database. This database is apparently being utilized in areas such as natural language processing and text mining.

3.3.1 Wikipedia as Social Big Data

Wikipedia maintains its electronic encyclopedia content via human wave tactics using a wiki system. So, is it social big data? In my opinion, Wikipedia article data should be classified as factual big data, as a lot of effort has been put into the system and its operation to improve its utility as an encyclopedia, and encourage writers to base their work on objective facts. Of course, because the data is written and edited manually using a wiki system, compared to data generated by machines there are more issues with data integrity, such as typos, factual errors, and inconsistencies between articles. On the other hand, it can also be thought of as extensive factual big data, which can cover matters for which the facts cannot be confirmed or circumstances without an established theory.

It must, however, be said that Wikipedia has a social big data aspect to it, too. Namely, the "Page view statistics for Wikimedia projects"*10, which have been published since 2013. This data is a tabulation of the page view count in hourly increments for all pages of the Wikimedia project. Going back you can obtain page view information from 2008 onward (or December 2007 to be exact).

In my research group we have kept a close eye on this data ever since we launched a service*11 based on it in June 2013, and it is fascinating to see how the rankings change, or in other words which hot topics show up in the top rankings (Figure 3).

3.3.2 Research by Tobias Preis

Tobias Preis, an Associate Professor of Behavioral Science & Finance at the University of Warwick, explains that individual Internet users seeking information is driving this kind of behavior. For example, after hearing about major news topics reported by the mass media, etc., individuals who use the Internet on a daily basis tend to gather information related to that topic by performing

Figure 3: Wikipedia Rankings

*8   http://wiki.dbpedia.org/

*9   http://ja.dbpedia.org/

*10   http://dumps.wikimedia.org/other/pagecounts-raw/

*11   http://www.gryfon.iij-ii.co.jp/ranking/

searches on search engines or viewing Wikipedia. It is thought that traces of this behavior are recorded as search engine query data or Wikipedia page views. Focusing on this phenomenon, Preis ascertained that search engine query data and stock market fluctuations were correlated in his 2010 paper, "Complex dynamics of our economic life on different scales: insights from search engine query data"[12].

Next, in his paper, "Quantifying Trading Behavior in Financial Markets Using Google Trends"[13], he suggested that an increase in searches for 98 terms related to finance included in query data obtained from Google Trends tended to precede major crashes in the financial market.

Furthermore, in his paper, "Quantifying Wikipedia Usage Patterns Before Stock Market Moves,"[14] he used knowledge gained through Google Trends analysis to identify that Wikipedia page view counts correlated with large-scale fluctuations in the stock market.

In these papers, Preis concluded that it was possible to gain new insight regarding the information gathering behavior of people pressed to make decisions from online data such as search engine query data and Wikipedia page views. For example, phenomena such as stock market crashes result from the decision-making of individual investors, but paying attention to online data makes it possible to discover signs of this at an early stage. This is a typical example of a solution that utilizes social big data.

3.4 Conclusion

In this report, I have focused on social big data and its characteristics. The Social Data Revolution advocated by Andreas Weigend suggests that in certain fields such as e-commerce, vast quantities of data related to human behavior for product purchases is beginning to accumulate, and there is a need for a new approach based on analysis of this data.

Because social big data is data that relates to people's behavior, development of analysis methods will fall under the domain of social science or behavioral science. However, given that we have never before been in a position to obtain such wide-ranging and detailed data regarding human behavior, it is likely to require a considerable amount of time to find effective techniques for extracting valuable knowledge. As far as I am aware, the approach of tracking and analyzing individual behavior based on data obtained from microblogs or SNS has not been very successful, despite the ability to produce analysis results of some kind, due to the difficulty of identifying meaningful knowledge in these results.

On the other hand, I believe Tobias Preis's approach of treating Internet user behavior as a macro phenomenon, and applying complex analysis techniques, may make it easier to obtain significant findings. Factors such as query data and Wikipedia PVC are used for this analysis, and Preis's papers seem to show that short-term forecasting will be possible through predicting group behavior by independent individuals. Additionally, by analyzing notable fluctuation phenomena identified through this analysis in further detail, I believe it should be possible to gain insight into root causes, etc. by tracking the behavior of individuals using data obtained from social media.

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.

*12 http://www.tobiaspreis.de/publications/prs_ptrsa_2010.pdf
*13 http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html
*14 http://www.nature.com/srep/2013/130508/srep01801/full/srep01801.html

IIJ
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.