社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－DPS－138 (31)
2009－CSEC－ 44 (31)
2009／3／6

公開ネットワークログデータセットの調査とワーム検知数の変遷調査

細井 琢朗† 松浦 幹太†

†東京大学生産技術研究所
153-8505 東京都目黒区駒場 4-6-1

あらまし インターネットはあまりに自由なため、危険性のある通信が常に飛び交っている。ネットワークログデータセット（以下、データセット）は、このようなインターネットを安全に利用するために無くてはならない各種技術の研究、開発、性能検証に不可欠である。これまで数多くの研究が様々なデータセットを用いて行われてきたが、残念ながらデータセットの多くは非公開となっており、研究成果の検証の壁となっている。

我々は、この壁を乗り越える基盤として期待されうる、公開データセットの調査を行った。また、その中の幾つかを用いてワーム検知数の変遷調査を試みた。

キーワード：ネットワークログデータセット、ネットワーク監視、ワーム

Survey on Public Network Log Dataset and Examination of Long-term Variation of Internet Worm Traffic

HOSOI Takurou† Kanta Matsuura†

†Institute of Industrial Science, The University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan

Abstract Communication in the Internet is so free (or unconfined) that there always flows hazardous traffic. A network log dataset (hereinafter simply called dataset) is an essential data for research, development, and evaluation of network technologies which are necessary to use the Internet securely. Many researches of this kind have been done with various datasets, but unfortunately most of these datasets are not publicly opened. It leads to an obstacle against verification of the results of these researches.

We conducted a survey on public datasets, which are expected to be a basis to overcome above obstacle. We also tried to examine the long-term variation of Internet worm traffic with some available public datasets.

Keywords: network log dataset, network monitoring, Internet worm

1 導入

インターネットは可用性を重視した管理・運営の下にあり、その中では基本的に自由な通信がなされている。このことにより、Web や電子メールのような様々な種類のサービスが容易に利用でき、また IP 電話のような新しいサービスの導入も困難なく導入できる。しかし一方でこの自由さは、ネットワークを介した攻撃先の

物色や、実際のネットワーク攻撃そのものなどの、危険につながる通信の存在も許してしまう。

上記のように安全とは言えないインターネットではあるが、現在多くの人々がこれを比較的危険なく利用している。その大きな理由の一つは、firewall やウィルス対策ソフトウェアといったセキュリティ技術が広く実装、利用されていることにある。このようなネットワークセキュリティ技術の研究、開発、性能検証では、その安全対策の対象の標本などとして、大抵の場合、ネットワークログデータセット（以下、データセット）を用いる必要がある。実際、これまでに数多くの研究が様々なデータセットを用いて行われてきている。残念ながら、これらの研究で用いられたデータセットの多くは、一般には非公開となっている。これは、通信の秘密やプライバシーの問題、または通信を開示したことによって発生する可能性のある賠償請求などを回避することが理由と思われる。すると上記の問題は避けられるが、代わりに、その研究の成果を第三者が検証しようとした際に、同じデータセットが使えないため、全く同じ追実験ができない、という問題が生じる。このことは、そもそも仕組みなどを公開する必要の無い製品開発などでは問題にならないが、科学的な手続きを重視した研究においてはその成果の有効性を担保し難くなるという問題を抱えることになり、好ましくない。

我々は、以上の問題意識から、この問題を解決する基盤として期待されうる、公開データセットの調査を行った。また、その利用方法の一つとして、その中の幾つかを用いてワーム検知数の変遷調査を試みた。

2 対象とするデータセット

ネットワークのログデータとしては、

(a) アクセスログのように、時刻情報と発信元、受信先のみ（または幾つかの付随情報）を全てのパケットについてネットワーク上のある点で記録したもの

(b) 時刻情報と各パケットヘッダ情報を全て記録したもの

(c) 時刻情報と各パケット全体を全て記録したもの

(d) 上記の記録の一部をある規則に従って抜粋したもの

などが考えられる。例えば、流量制御による負荷分散を考えているのであれば、多くの場合、(a) の記録で十分である。また、ある種類の通信のみを取り扱うのであれば、(d) の抜粋記録が扱いやすい。

ここでは、汎用性を考えて、(c) の全パケット記録を調査対象のデータセットとして考える。

3 調査方法

データセットを用いる研究領域は数多くある。また、データセットそのものやその抜粋手法を研究対象にしている研究（例：[1, 2]）もある。我々は、まずデータセットの研究論文の幾つかからデータセットを作成している研究機関を調べ、それらについて、データセットを公開し、Web 上で入手可能かどうかを確かめた。

また、侵入検知システム（IDS）の研究では、通常、その性能評価のために全パケット記録の大きなデータセットを使用する。また早くからそのデータセットを公開している例もある。そこで、侵入検知システムの研究についても、その幾つかの論文からデータセットを作成している研究機関を調べ、そのデータセットを Web 上で入手可能かどうかを確かめた。

4 調査結果

データセットの研究論文では、その統計的な特徴についての議論が中心になるためか、独自にデータセットを作成して研究を進めているものが多かった。そのため、公開されているデータセットは少数のみだった。また、ログデータが公開されていても、§2 の (a) のような、必要な統計量が得られるだけのデータである例もあった。

侵入検知システムの研究においても、特に近年は、独自のデータセットを作成して研究を行い、データセットを公開しない例が多く、少数のデータセットしか見つけることができなかった。

以下に、調査の結果得られた公開データセットを挙げる。

(1) DARPA データセット

(2) KDD Cup 1999 データセット

(3) CRAWDAD データセット

(4) WITZ データセット

以下、これらのデータセットの特徴を簡単に述べる。

4.1 DARPA データセット

侵入検知システムの研究のために MIT の Lincoln Laboratory で作成、公開されたデータセットであり、その研究分野で最も利用されてきた（最も有名な）ものである。大きく 1998 年、1999 年、2000 年の三つのセットに分かれており、それぞれ、

- 1998 年は約一日のデータセット（外部ネットワーク分のみ）が五日分（月曜日から金曜日）で一週間分のセットになり、これが、侵入検知システムの学習用として七週間分、学習後の検証用として二週間分、
- 1999 年は約一日のデータセット（外部ネットワーク分と内部ネットワーク分）が五日分（月曜日から金曜日）で一週間分のセットになり、これが、侵入検知システムの学習用として三週間分、学習後の検証用として二週間分、
- 2000 年は補足データセットであり、異なるシナリオの分散サービス妨害攻撃が行われている際のデータセット（数時間分）が二つと、1999 年のものよりも強い Microsoft Windows NT ホストへの攻撃の際のデータセット（数時間分）が一つ、

から成っている。

これらのデータセットは、一度収集した実パケットデータを、実験用に組まれた隔離されたネットワーク内で再度流している中に、人為的に攻撃パケットを混ぜた上で、それらのパケット全体を収集して作成されている。

4.2 KDD Cup 1999 データセット

国際会議（KDD 1999）で行われた、侵入検知システムの研究コンテスト（KDD'99 Classifier Learning Contest）のために作成され、用いられたデータセットである。このデータセットは DARPA データセットを元に、幾つかの統計的指標を維持するように抽出して作成された。このデータセットはコンテスト用のため、データファイルが一つしかない。

4.3 CRAWDAD データセット

Dartmouth College の Computer Science 専攻で作成、維持、公開されている、各種の無線通信のデータセットである。それぞれ異なるネットワーク環境で収集したデータセットを集めており、2002 年のものから現在のものまで、数多くのものがある。現在も新しいデータセットが追加されている。

このデータセットを入手するには、ユーザ登録が必要である。

4.4 WITZ データセット

Waikato University で行われている、KAREN (KIWI ADVANCED RESEARCH AND EDUCATION NETWORK) の研究の一環で作成されたデータセットである。1999 年、2001 年、2003 年、2007 年に収集したデータセット（長さ数十分）が全部で 11 個ある。収集時間は DARPA データセットに比べて短いが、ネットワーク容量が大きいためか、各データセットの大きさは DARPA のもののおおよそ十倍程度ある。

このデータセットについては、その収集の環境や方法などの情報はデータセットと一緒に配

布されてはいない。

5 ワーム検知数変遷調査の試み

§4 の結果、公開データセットを複数得ることができた。この章では、その利用例として、その中の DARPA データセットを用いて、ワーム検知数の変遷調査を試みた結果について述べる。

5.1 データの選定理由

今回は利用するデータセットとして、DARPA データセットを用いた。これは以下の消極的な理由により決定した。

- CRAWDAD データセットは、それぞれ異なるネットワーク環境で収集されたものであり、ワームの検知を行った結果得られる検知数の比較が一般には困難なため、今回は利用しなかった。
- KDD Cup 1999 データセットは、データファイルが一つしかないこと、また DARPA データセットを元に作成されているので、DARPA データセットを用いる今回は利用する必要がないと判断した。
- WITZ データセットは、複数年にわたることから、利用の第一候補に挙げていたが、今回検知のために使用するソフトウェア（Snort [7]）がこのデータセットを受け付けなかったため、今回は利用を諦めた。

5.2 ワームの検知方法

今回は試みとしてワームの検知を行うことから、高い検知性能を持つ特別な方法を使うのではなく、一般に広く用いられている、ネットワーク IDS（NIDS）ソフトウェアの Snort[1] [7] を使用してワームの検知を行った。この NIDS ソフトウェアはオープンソースであり、検知手段としてこれを採用することは、研究で行った実験に使用した重要な構成要素が公開されないという本調査の問題意識にも合致する。

Snort は侵入検知システムとして、各種の危険につながるパケットや実際の攻撃パケット、またウィルスやワームを検知するルールを実装している。今回は簡単のため、すぐに入手可能なルールの中で最も新しいもの[2]を導入し、その中のルールを全て用いた。この場合、ワームに関する全部で 74 件のシグネチャが検知ルールに含まれる。

侵入検知の結果はアラートログとして指定されたファイルに記録される。この記録を用いてワームの検知数を数える。但し、「あるワームが一件検知された」という数え方は、一般には明確に定義されていない。これは得られた検知数の利用方法とも関係してくる。例えば、どれだけのホストが危険に晒されたかを知りたい場合は、ワームの活動を熟知した上で、検知された警告からその活動状況を推測し、一つの感染活動の開始から終了（もしくは中断）までを一件として数える方法が適している。今回は簡単のため、アラートログファイルの中で、キーワード「worm」（大文字、小文字の区別は付けない）を含む警告一件をワーム検知一件として数えることにした。この数え方では、ワームの感染先探索、外部からの実際の感染パケット、内部の感染ホストからの感染活動の一つ一つが全て一件ずつに数えられてしまうが、内部に感染ホストが無く、外部からの感染も成功しないと思われる DARPA データセットの収集環境では、これである年、ある日にワームが活発に活動していたかどうかを十分に把握できると考える。

5.3 ワームの検知結果

まず、§5.2 の方法で、DARPA データセット全体（学習用、検証用合わせて 100 個のデータファイル）に対して Snort によりアラートログを作成した。各アラートログに含まれる警告は、数千件から数十万件と幅があった。これは元になるデータセットの中に含まれるパケット数に

[1] Snort ver.2.8.3.2

[2] VRT Certified Rule for Snort v2.4 (unregistered user release)

幅があるためと思われる。

これらのアラートログを§5.2の数え方で処理し、ワームの検知数を調べた。その結果、DARPAデータセット全体からはワームを一件も検知できなかった。その理由として、以下の三点が考えられる。

(i) DARPAデータセットを作成していた当時は、まだワームが現在ほど多くなく、また広く感染していなかったため、元々ワームの通信がこのデータセットに含まれていない可能性。

(ii) DARPAデータセットを作成する際に、公開するには危険なワームの通信を削除した可能性。

(iii) Snortのルールに含まれるワームのシグネチャに、当時存在したワームのものが含まれていない可能性。

特に(iii)については、Snortのルールを調べたところ、確認できたワームのシグネチャの中では、ワームが2000年に発見されたものが最も古かった。そのため、ワーム検知数の変遷が、1998年、1999年、2000年と、零（もしくは一日に一件未満）であったと主張することはこの調査からではできないという、残念な変遷調査結果となった。

6 まとめ

本調査では、研究で使用した重要なデータが公開されないという問題意識を元に、公開ネットワークログデータセットの調査を行い、四つの公開データセットを得ることができた。また、これらのデータセットについて特徴を調べ、利用の試みとしてDARPAデータセットを用いたワーム検知数の変遷調査を試みた。その結果はデータセットの選定と検知方法の不備から、残念ながら特に何らかの主張を明確にできるものにはならなかった。

公開データセットの利用の試みとしてDARPAデータセットを利用した実験を行った結果、このデータセットはデータ形式が使いやすく、データセットの作成方法などの解説文書も揃っているため、データセットを使った実験、研究の初歩として用いるのに向いていること、また、多くの研究に使われてきたという歴史から既存研究との比較を行う目的には向いていること、そしてデータセットの古さが問題になることがあることが分かった。

今後は、プログラム、もしくはデータセットの修正をすることで、WITSデータセットを用いて今回と同様のワーム検知数の変遷調査を行うことと、公開データセットの継続調査を計画している。

参考文献

[1] Sooyoung Chae, Hosub Lee, Jaeik Cho, Manhyun Jung, Jongin Lim, Jongsub Moon, “A Regression Method to compare Network Data and Modeling Data using Generalized Additive Model”, WISA 2008, 5A-2 (Sep. 2008)

[2] Nicolas Hohn and Darryl Veitch, “Inverting Sampled Traffic,” IEEE/ACM Transactions on Networking, Vol.14, No.1, pp.68–80 (Feb. 2006)

[3] MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation Data Sets, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html

[4] The UCI KDD Archive, KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[5] CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth, http://crawdad.cs.dartmouth.edu/

[6] WITS (Waikato Internet Traffic Storage), http://wiki.karen.net.nz/index.php/WITS

[7] Snort — the de facto standard for intrusion detection/prevention, http://www.snort.org/