情報セキュリティ研究向けネットワークデータの配布における技術的課題の現状調査

細 井 琢 朗[†1]　松 浦 幹 太[†1]

侵入検知システムの研究向けに作成された DARPA Datasets (1998, 1999, 2000) の公開から 10 年，日本ではマルウェア研究向けに CCC DATAset (2008, 2009, 2010, 2011) の配布が行われ，多くの研究に寄与した．この他にも近年，PREDICT，WOMBAT，DETER など，情報セキュリティ研究向けのネットワークデータを共通使用できる環境を整備し，研究に役立てる活動が行われている．本発表では，これらの活動について主に情報保護と安全性に関わる技術に注目し，現状を調査した結果を報告する．

A Survey of Present Situation on Technical Issues about Network Data Dissemination for Information Security Research

HOSOI Takurou[†1] and Kanta Matsuura[†1]

After 10 years from the public release of DARPA Datasets (1998, 1999, 2000) which were synthesized mainly for Intrusion Detection System research and development, CCC DATAset (2008, 2009, 2010, 2011) were distributed for anti-malware researches in Japan and contributed to many research activities. There are also other projects, like PREDICT, WOMBAT, DETER, which support network security researches by developing and maintaining environments to access to commonly usable network data. This report is a result of a survey focusing on present situation of technival issues about such projects.

†1 東京大学
The University of Tokyo

1. はじめに

ネットワークに関する各種のデータ，特にネットワーク通信を何らかの形で残したデータ（ログ）は，ネットワークセキュリティ技術の研究，開発，教育に役立っている．例えば，MIT Lincoln Laboratory で製作され Web 上で一般にも公開されている，DARPA Datasets[1)] と呼ばれているネットワーク通信データセットの登場により，侵入検知システム（IDS）の研究が非常に活発に進められた．また近年では，CCC DATAset 2008，2009，MWS 2010，2011 Datasets[2)]（CCC DATAset 2010，2011 を含む）というデータセットの作成と配布も行われている．こちらのデータセットは提供先を研究者向けに限定することで，主にマルウェア対策の研究に有用なものとなっている．

このように有用なデータセットが共有されることで，異なる研究，技術の性能評価が同じデータセットによって行われ，より客観的な比較が可能になることや，データの容易な入手による研究の促進などが期待される．しかしその共有にはいくつかの解決すべき問題があるため，多くは非公開となっている．まず，インターネット上では危険性のある通信も行われており，それを記録したデータセットは取り扱いに注意が必要になる．次に，ネットワークを流れる通信の記録を作ると，そのネットワークを使用していた利用者の個人的な情報を含んでしまい，プライバシーの観点からそれをそのまま配布することには問題がある．また，場合によってはネットワークの情報，特に各ネットワーク機器の接続情報を隠す必要があるために，データセットを共有できないこともある．

このようなデータの情報保護とデータ提供元の安全性の問題を，上記の二つの例（DARPA Dataset，CCC DATAset/MWS Datasets）ではそれぞれ異なる方法で解決している．まず DARPA Dataset は，実際に行われた通信そのものではなく，実験ネットワーク内の通信を収集して作られている．収集する際に流す実験ネットワーク内通信としては，実際の通信を一旦収集したものと，整えられた攻撃通信が同時に流されている．これにより，実際の背景通信に意図通りの攻撃通信が混ぜられた通信データが，通信内容の保護の必要無しに，またデータ収集元の情報を（実験ネットワークの情報で上書きすることで）隠した形で配布できている．一方 CCC DATAset（/MWS Datasets）では，これらのデータを使った研究・開発の手法上の制限をなるべく減らすために，実際に収集された通信そのものを「攻撃通信データ」として配布している．このデータにはマルウェアが実際に攻撃・感染の際に行った通信が含まれているため，データの漏洩などがあってはならない．そこでこのデータセットの場合，データに含まれる情報の保護とデータ提供元の安全を，データ提供者とデー

タの利用者の間で利用に関する契約を結ぶことで担保している．

DARPA Dataset や CCC DATAset（/MWS Datasets）の例以外にも，ネットワークデータを共通使用することで，研究や開発に貢献することを目的とした，PREDICT[3] や WOMBAT[4] といったプロジェクトがある．これらのプロジェクトでも，情報保護と安全性を何らかの形で維持した上でデータの使用が可能になっているはずである。もしこれらの点が技術的に克服されていた場合，その技術を他のデータセット配布にも応用できる可能性がある．そこで本稿では，主に通信データの提供の際に，これらの課題が技術的にどれだけ克服できているかに注目し，各プロジェクトのデータ提供の仕組みの現状を調査した結果を報告する．

2. 調 査 方 法

本調査は，文献調査の方法で行った．調査対象の文献には，各プロジェクトのＷｅｂサイトそのものと，そこで公開されているプロジェクトの各種報告書，またそこで成果として挙げられている学術論文を用いた．これらの中からデータの提供や配布に関する記述を抽出し，主に通信データの提供の際，情報保護と安全性がどのように保たれているかについて，特に技術的な解決手段に注目して調査した．

通信データ提供の際の情報保護と安全性の問題には，以下の三点があることが判っている．

(1) （データの危険性）
インターネット上では危険性のある通信も行われている．それを記録したデータは取り扱いに注意が必要になる．

(2) （個人情報の保護）
ネットワークを流れる通信の記録を作ると，そのネットワークを使用していた利用者の個人的な情報を含んでしまうことが多い．プライバシーの観点から，この記録をそのまま配布することには問題がある．

(3) （データ収集対象の安全性）
ネットワークの情報，特にデータの収集対象範囲の各ネットワーク機器の接続情報を隠す必要がある場合がある．データの提供元がこの安全性を必要とする場合は，これを担保してデータ提供が行われなくてはならない．

これらの問題の一部は，匿名化技術（例 :[7]）などの既存技術である程度解決できる．これらの点が実際のネットワークデータ提供でどのように解決されているかを調べることが，この調査の主眼の一つである．

調査は，既に詳細が分かっている DATPA Dataset と CCC DATAset（/MWS Datasets）以外の，PREDICT[3]，WOMBAT[4]，DETER[5] を対象に行った．なお，CAIDA[6] でもネットワークデータの配布を行っているが，調査時間の関係から本稿では取り扱わない．

3. 調 査 結 果

第 1 節，第 2 節で述べたように，ネットワークデータの配布には情報保護と安全性の問題がある．この点を，よく知られた二つの例（DARPA Dataset，CCC DATAset/MWS Datasets）ではそれぞれ異なる方法で解決している．

3.1 DARPA Dataset の場合

DARPA Dataset は，実際に行われた通信そのものではなく，実験ネットワーク内の通信を収集したものになっている．収集する際に流す実験ネットワーク内通信として，実際の通信を一旦収集したものと，整えられた攻撃通信が同時に流すことで，実際の背景通信に意図通りの攻撃通信が混ぜられた通信データを収集している．これにより，通信内容の保護の必要はなくなり，またデータ収集元の情報も，実験ネットワークの情報で上書きすることで隠される．即ち，実験ネットワーク内での通信の再送と再収集という技術的工夫により，情報保護と安全性の問題を解決している．ただし，DARPA Dataset には内容そのものが危険性を持つような通信は含まれないためこの方法で問題が解決しているが，一般にはこの限りではない．

3.2 CCC DATAset（/MWS Datasets）の場合

CCC DATAset（/MWS Datasets）では，これらのデータを使った研究・開発の手法上の制限をなるべく減らすために，実際に収集された通信そのものを「攻撃通信データ」として配布している．このデータにはマルウェアが実際に攻撃・感染の際に行った通信が含まれているため，データの漏洩などがあってはならない．そこでこのデータセットの場合，データに含まれる情報の保護とデータ提供元の安全を，データ提供者とデータの利用者の間で利用に関する契約を結ぶことで担保している．すなわち，技術的な手段は用いずに情報保護と安全性の確保を行っている．この方法では適切な組織運営が必要となる．

本調査の対象である PREDICT[3]，WOMBAT[4]，DETER[5] についての調査結果を以下の小節で示す．

3.3 PREDICT

PREDICT（the Protected Repository for the Defense of Infrastructure Against Cyber

Threats）[3] は，アメリカ合衆国国土安全保障省（Department of Homeland Security, DHS）の Science and Technology Directorate's Cyber Securiry R&D Group（S&T/CCI）の後援により運営されている，ネットワークデータの提供を進めるプロジェクトである．プロジェクト自体は DHS そのものではなく，非営利団体である PREDICT Coordinating Center（PCC）を中心に運営されている．PCC は現在，非営利研究組織である RTI International 内に置かれている．

PCC とそのいくつかの補助組織は，セキュリティ関連のネットワークデータの提供元を集め，契約を結び，研究者に配布できるデータを集めている．また PCC は，申請のあったネットワークやセキュリティの研究者を審査の上で登録する．そして PCC がデータ提供元とそのデータを利用したい研究者を結び付けることで，情報セキュリティ研究のためのネットワークデータの配布を実現している．現在，以下のデータが提供されている．

- Topology Measurement Data
- Blackhole Address Space Data
- Hi-Speed ISP Exchange Data from OC-48 operational network (packet traces)
- Full Packet Headers
- Domain Name Server (DNS) Root Server Data
- Internet Topology Data
- Address Space Allocation Data
- Enterprise Data （大企業の LAN 通信）
- BGP Routing Table Data
- BGP Update Messages & BGP Routing Table Dumps
- VOIP Statistical Data (End-to-End Quality)
- Firewall Logs
- Traffic Flows via Netflow
- Network Management Data （SNMP）
- Intrusion Detection System (IDS) Logs
- Anonymized Internet Witty Worm Data
- Code-Red Worm Data

これらのデータを提供するにあたって，情報保護と安全性を確保するため，PREDICT は以下の手段を講じている．

(1) （契約）
PCC とデータ提供元，PCC と研究者，また必要に応じてデータ提供元と研究者の間で，データ提供に関する契約を結ぶ．この契約（特に研究者が結ぶもの）には，データの適正な利用や，データの漏洩を防ぐ適切な対策を講じる義務，この契約内容の遵守を確認する監査の実施などが含まれている．これにより，データの目的外使用や漏洩などを防ぐ．

(2) （IP アドレスの匿名化）
PREDICT では，個人を特定できる情報（personally identifiable information, PII）の漏洩を特に警戒している．そのため，一部の例外を除き，ほとんどのデータでは IP アドレスを匿名化している．例えば，VOIP に関する統計データにおいては，通話に利用したネットワーク範囲の情報が通話品質の研究に必要なため，prefix を残して IP アドレスの匿名化がなされている．
唯一，「Topology Measurement Data」はネットワーク機器間の通信からできており，実際のネットワーク利用者は関与していないため，IP アドレスは匿名化されていない．ただし，契約によりこれらの IP アドレスの公表は行えない．

(3) （通信内容の削除）
上記で述べたように，PREDICT は個人を特定できる情報（PII）を厳しい規制の下で取り扱っている．そのため，電子メールの内容や Web ページといった通信内容は，全て削除されている．

(4) （データ利用者の審査）
データ利用の契約が適切に守られることを確約するための手段の一つとして，PREDICT ではデータの利用者は審査の上，登録された者に限られている．登録を申請できるのは，現在のところ，アメリカ国内で活動している研究者に限られている．
データ利用者はパスワードで認証され，アクセス制御により守られた機器を通じてデータを利用する．

(5) （研究成果の公表前の審査）
研究者がその研究成果などを PREDICT の契約に含まれる人物以外に知らせる場合（学会発表など），その内容がデータの秘密に触れていないかどうか，PCC（とその補助機関）の事前審査を受ける必要がある．

必要十分な漏洩対策などは，PCC とその補助機関により規定される．データ提供元やデータ利用者は，その規定に沿うことで契約を守ることになる．

以上をまとめると，IP アドレスの匿名化と，通信内容の削除が，情報保護と安全性のた

めに PREDICT が行っている技術的対策であることが分かった．また，契約とそれに基づく確認，監査により，それ以外の部分を手当てしていることも分かった．

3.4 WOMBAT

WOMBAT（Worldwide Observatory of Malicious Behaviors and Attack Threats）[4] は，European Community's Seventh Framework Programme (FP7/2007-2013) から資金を得て運営されている，セキュリティに関連するデータの提供と各種の脅威についての解析結果の提供を行うプロジェクトである．このプロジェクトも PREDICT と同様，データの提供元と研究者を契約により結び付け，研究者がデータを利用できる仕組みを提供している．ただし，PREDICT が主に通信データを提供しているのとは異なり，WOMBAT では主にマルウェアのサンプルとそれに付随する情報（シグネチャなど）の提供を行っている．データ提供元に無料のマルウェア検査 Web サービスの VirusTotal[8] などを含めることで，千個を超える最新のマルウェアサンプルを提供している．

WOMBAT はその提供データの情報保護と安全性を確保するため，以下の手段を講じている．

(1) （データへのアクセス）
WOMBAT が提供しているデータは，マルウェアサンプルのようにそれ自身の所持が法律上問題になるものや，各種のシグネチャのように更新頻度が早いものが多い．またその頻繁な更新のために，WOMBAT 自身はデータそのものは持っておらず，データ提供元にのみある場合もある．そのためデータそのものを配布するのではなく，データへアクセスできる WOMBAT API（WAPI）が配布されている．プログラムがこれを通してデータを利用することで，利用者は自身の研究を進めることができる．
このアクセス方法は利用者の権限に応じて制御されている．これにより，マルウェアサンプルの流出などがある程度防げる．

(2) （データアクセス許可）
公開データ以外のデータを利用するには，データの利用者として WOMBAT から許可を得る必要がある．この許可は WOMBAT との契約によって得られる．この契約は利用者に情報保護とデータの適正な利用の義務を課す．
2008 年 11 月 19 日付けの WOMBAT の Web サイトの情報によると，SGNet honeypot をインストールし，Leurre.com のプロジェクトに参加することで，WOMBAT へのデータ提供元の一員になり，それに応じて WOMBAT のデータの一部が使用できるようになるとのことである．また，公開データには特に許可なくアクセスできる．

(3) （匿名化）
データ提供元を安全に保つため，マルウェアサンプルの ID ，IP アドレス，無線通信で使われる各種のアドレスは必要に応じて匿名化される．例えば侵入検知システム（IDS）のアラートログの場合，送信先 IP アドレス（IDS の IP アドレス）は匿名化されるが，発信元 IP アドレスは匿名化されず，データに含まれるべき情報として利用者に提供される．

(4) （パケット削除）
前述した通り，WOMBAT は主にマルウェアサンプルとそれに関連するデータを提供している．そのため提供されている通信データとして確認できたものは，ネットワーク型 IDS である NEMU[9] で収集された通信データ（pcap trace）だけである．このような通信データの提供の際は，個人情報が漏れないよう，一部のパケットを削除して提供される．

以上をまとめると，WAPI を通したアクセス制御付きのデータ利用と，一部の IP アドレスの匿名化，問題になるパケットの削除が，情報保護と安全性のために WOMBAT が行っている技術的対策であることが分かった．また契約により，データ利用者には情報保護や正当なデータの利用の義務が課せられ，技術的対策以外の部分を手当てしていることも分かった．

3.5 その他のプロジェクト

これまで取り上げてきたネットワークデータの提供方法とは異なった方法で共通の研究用データを共有できる仕組みもある．その一例として，ここでは DETER[5] を紹介する．

DETER はアメリカ合衆国国土安全保障省（Department of Homeland Security）Science and Technology（DHS S&T）と全米科学財団（National Science Foundation, NSF）の資金援助で 2003 年に始められ，現在 DHS ，NSF ，アメリカ合衆国国防総省（Department of Defense, DoD）の資金援助を受けて運営されている，ネットワークセキュリティとコンピュータセキュリティの研究のためのテストベッド環境（DeterLab）を提供している，研究プロジェクトである．

ネットワークセキュリティ研究のテストベッドでは，多くの場合，検知や防御の対象とする通信に加え，インターネット内を常に流れている背景通信や，通常のユーザの通信などを利用する．DeterLab ではこれらの通信をパケットジェネレータを使って発生させており，ほぼ同じトラフィックを何度でも流すことができる．この性質を利用すると，研究の促進な

どの点で，共通のデータが利用できる状況と同じ効果を期待できる．また，データが実験環境内に留まったままであるため，実験結果の適切な引渡しにより，データの情報保護とデータ提供元の安全性を確保できる．

DeterLab の使用のための参加登録は，DETER の Web サイトから申請できる．

4. 考察：CCC DATAset（/MWS Datasets）の場合との比較

ここでは本稿で取り上げた二つのネットワークデータ利用提供プロジェクトと CCC DATAset（/MWS Datasets）[2] の配布について，情報保護と安全性の観点から，類似点と相違点を挙げる．

まず最大の類似点は，どのデータ提供の活動でも，データの提供元とデータ利用者の間でデータの利用に関する契約を結ぶ必要があることである．このことから，ネットワークセキュリティやコンピュータセキュリティの研究に有用なデータは，その有用性のために技術的な情報保護や安全性の確保が難しい，と現状では見なされていることが分かる．

その他の類似点としては，これらの三つの活動で提供されるデータの種別に，なんらかの通信データと，検知システムのログと，マルウェア情報があることが挙げられるが，これはデータの有用性から導かれる性質であり，情報保護と安全性から来たものではない．

相違点の中で最大のものは，CCC DATAset（/MWS Datasets）の配布では情報保護と安全性が契約のみで確保されている点である．このことにより，CCC DATAset（/MWS Datasets）の利用者は研究のために生の通信データを利用できる．一方 PREDICT や WOMBAT では，提供されるデータ内の IP アドレスは基本的には匿名化される．これは一見大きな違いだが，実は PREDICT でも実ユーザが介在しない「Topology Measurement Data」は IP アドレスの匿名化を行わずにデータ提供されている．つまり，提供されるデータの性質により匿名化の必要の有無が変わっているだけであることが分かる．CCC DATAset（/MWS Datasets）でも実ユーザが関与する通信が含まれる通信データやログが提供される際には，そのデータには匿名化処理が施されるであろう．

その他の相違点としては，各プロジェクトで提供される主なデータの種別が異なる点が挙げられる．どのプロジェクトでも通信データ，検知システムのログ，マルウェアに関する情報が提供されるが，PREDICT は基幹ネットワークなどの多種類の通信データが主に提供されるのに対し，WOMBAT と CCC DATAset（/MWS Datasets）ではマルウェアに関するデータが主に提供される．この違いは情報保護と安全性から来たものではなく，各プロジェクトの成り立ちや，各プロジェクトの行われている地域（PREDICT は米国，WOMBAT は欧州，CCC DATAset（/MWS Datasets）は日本）の違いが反映されたものと考えられる．

5. ま と め

情報セキュリティ研究向けのネットワークデータの配布に関して，そのデータ提供の際にデータの情報保護とデータ提供元の安全性の問題を技術的に解決できているかどうかに着目し，実際の例を文献により調査した．その結果，現在行われているデータ提供では，技術的な対策は IP アドレスの匿名化と個人が特定できる情報の削除がなされる程度であり，多くの部分がデータ提供元とデータ利用者の間の契約により手当てされていることが分かった．

今後は，今回は触れなかった CAIDA によるデータ提供の例なども調査し，情報保護と安全性に対する技術的対策が他にないか調査を続ける．特に，IP アドレスの匿名化だけではデータ提供元の位置特定を防ぐには不十分であることが判っており[7),10)]，この問題へも技術的対策が取れないか研究を進めたい．

参 考 文 献

1) DARPA Datasets (1998, 1999, 2000), MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation Data Sets, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html
2) 畑田充弘, 他, “マルウェア対策のための研究用データセット ～MWS 2011 Datasets～”, コンピュータセキュリティシンポジウム 2011 (CSS 2011), マルウェア対策研究人材育成ワークショップ 2011 (MWS 2011), （2011 年 10 月）
3) PREDICT, The Protected Repository for the Defense of Infrastructure Against Cyber Threats, https://www.predict.org/Default.aspx?tabid=40
4) Worldwide Observatory of Malicious Behaviors and Attack Threats (WOMBAT), http://wombat-project.eu/
5) The DETER Project, http://deter-project.org/
6) CAIDA Internet Data – Passive Data Sources, http://www.caida.org/data/passive/
7) Justin King, Kiran Lakkaraju, Adam Slagell, “A Taxonomy and Adversarial Model for Attacks against Network Log Anonymization”, Proceedings of the 2009 ACM symposium on Applied Computing, pp.1286-1293 (March 2009)
8) VirusTotal — Free Online Virus, Malware and URL Scanner, http://www.virustotal.com/
9) M. Polychronakis, K.G. Anagnostakis, E.P. Markatos, “Real-world Polymorphic Attack Detection ”, In Proceedings of the 4th International Annual Workshop on

Digital Forensics & Incident Analysis (WDFIA) (June 2009)
10) 細井琢朗, 松浦幹太, “情報セキュリティ研究用ハニーポット通信データの一般頒布に向けた技術的要件の調査”, マルウェア対策研究人材育成ワークショップ 2011 (MWS2011), 発表 2A2-5 （2011 年 10 月）