

修士論文 2011 年度 (平成 23 年度)

デバイスの通信情報を用いた
ユーザ識別プラットフォームの設計と実装

慶應義塾大学大学院 政策・メディア研究科
上原 雄貴

デバイスの通信情報を用いた ユーザ識別プラットフォームの設計と実装

近年、インターネットを介した各種サービスにおいて、ユーザごとの行動履歴や趣味趣向に応じて最適化されたサービスを適時に提供することで、新たな付加価値を生み出すための取組みが行われている。実際にこのようなサービスを利用するためには、利用者が専用のアカウント登録や専用の機器の設置といった特別な操作が必要である。しかし、これらはユーザに特定の操作を要求することや、運用コストが高いなどの問題点がある。

本論文では許諾を得たユーザのデバイスが発信する特徴的な通信を収集することで、ホストやユーザを識別するプラットフォームを提案する。普段デバイスが発信している情報を利用することで、ユーザの手間や管理コストを低減することができる。

本論文が提案する手法は大きく3つある。1つ目は本提案手法を用いた複数のクライアントを各対象ネットワークに配置し情報を収集、識別する方法である。この手法は一般ユーザが同じセグメントに接続するホストの情報を収集する。2つ目はネットワーク管理者がネットワーク上に発信しているパケットのヘッダ情報のみを収集、解析することでホストを識別する手法である。この手法は管理者がネットワークの上流で情報を収集するため、ネットワークセグメントに影響されることなく対象ネットワーク全体の情報を収集することができる。3つ目は、在席情報を用いたホストとユーザを結びつける手法である。

これら3つの手法を用いてプラットフォームを作成し、200名程度のユーザが参加するカンファレンスやアカデミックなネットワークで実験した。その結果、ホストやユーザを本手法によって特定可能であることを示した。これによって、サービス提供者はユーザに特別な操作や設定などを求めることやデバイスの制約等を受けることなくユーザごとに特化したサービスを提供できる可能性を示した。

キーワード

1. ネットワーク監視, 2. フォレンジック, 3. プライバシー,
4. インターネット

慶應義塾大学大学院 政策・メディア研究科

上原 雄貴

Design and Implementation of User Identification Platform using Device Information

As demand for Internet services specifically customized for each user increases, service providers make an effort to give extra value by providing useful services, based on personal action history and personal interest. However, in order to use these services, they require users to install applications to identify users and hosts.

In this paper, I propose a platform to identify and track users by collecting and analyzing network traffic, which enables users to receive customized services without installing special applications or using special devices.

I propose three methods to identify users and hosts. The first method is to identify hosts using the host's MAC address by monitoring the network. Several sensors, which are general-purpose computers running our monitoring software while also in normal use by users, passively collect host information on the network. The second method is host tracking by network communication information surveillance. It is performed by collecting information from packet headers in the upstream network infrastructure. This method can collect information across the entire network without separately monitoring each network segment. The third method is to identify users by tracking their presence in front of the computer.

The practicality of these methods has been verified through experiments, performed on a conference and an academic network. Through the experiments, the feasibility for service providers to offer users user specific services, without requiring users to perform particular actions or settings, nor limiting users to use specific devices, has been shown.

Keywords :

1. Tracking Users,hosts, 2. Forensic, 3. Privacy,4. Internet

Keio University Graduate School of Media and Governance

Yuki Uehara

目次

第1章 序論	1
1.1 背景	1
1.2 目的	1
1.3 アプローチ	1
1.4 本論文の構成	3
第2章 ホスト追跡とプライバシーを巡る動向	4
2.1 ユーザ，ホスト追跡の現状	4
2.1.1 ユーザ自身が任意のタイミングで能動的に情報を送信するモデル	4
2.1.2 ユーザが常時情報を能動的に発信し続けるモデル	5
2.1.3 サービス提供者が受動的に情報を収集するモデル	6
2.2 現状のホスト，ユーザ識別	7
2.2.1 ホスト，ユーザ識別の問題点	7
2.2.2 現状の問題点とその対応	9
2.2.3 サービス実現のための契約形態	10
2.3 新しいサービスの動向	11
第3章 関連研究	13
3.1 ネットワーク管理者がユーザ発信する情報を収集解析する手法	13
3.2 SNS上の情報とホストを結びつけるユーザ特定手法	13
3.3 ブラウザ情報を利用したホスト識別手法	13
3.4 Web上での情報収集	14
3.5 まとめ	14
第4章 ネットワークデバイス情報を用いたホスト追跡手法	15
4.1 手法概要	15
4.2 端末識別情報の選定	15
4.3 事前実験及び検証	20
4.3.1 行動履歴の取得	21
4.3.2 端末間の関係	21
4.3.3 登録情報の変化検証	23
4.4 提案手法	24
4.5 設計及び実装	24

4.5.1	要件定義	24
4.5.2	クライアント設計, 実装	25
4.5.3	サーバ設計, 実装	28
4.6	実験	30
4.6.1	実験環境	30
4.6.2	実験結果	31
4.7	考察	33
4.8	まとめ	33
第5章	パケットヘッダ情報を用いたホスト識別手法	34
5.1	手法概要	34
5.2	事前実験	35
5.3	判別手法	36
5.4	検証	40
5.4.1	カンファレンスにおける検証	41
5.4.2	研究室ネットワークにおける検証	43
5.5	考察	45
5.6	まとめ	46
第6章	ホストとユーザを結びつける手法	47
6.1	手法概要	47
6.2	判別手法	47
6.2.1	在席情報による識別	48
6.2.2	SNS を利用した結びつけ手法	51
6.3	ホスト利用者特定手法の検証	51
6.3.1	在席情報を利用したホスト利用者特定手法の検証	51
6.3.2	SNS を利用したホスト利用者特定手法の結果	53
6.4	考察	54
6.5	まとめ	54
第7章	プライバシーに関する考察	56
7.1	識別システムとプライバシー	56
7.2	個人情報取り扱い形態とその懸念	56
7.2.1	オプトインにおける運用の特徴	57
7.2.2	オプトアウトにおける運用	58
7.2.3	ノンプライバシーにおける運用	60
7.3	サービス応用例とその運用形態	60
7.4	まとめ	62

第 8 章 結論	63
8.1 本論文のまとめ	63
8.2 今後の課題と展望	64
8.2.1 情報にアクセスするユーザの認証手法の確立	64
8.2.2 情報のオプトアウト手法の確立	64
8.2.3 パケットヘッダ情報を用いた識別手法における識別要素の追加	65
8.2.4 ユーザ特定におけるプライバシーの数理モデル化	65
謝辞	66

目 次

1.1	本提案手法の概要図	2
2.1	クライアントがサーバから付与された識別要素を管理	5
2.2	ユーザがサーバに情報を直接送信	6
2.3	ユーザが全体に情報を発信し、受信側がデータを処理	7
2.4	プライバシーマークサンプル	9
2.5	TRUSTe マークサンプル	9
4.1	MAC アドレスによるホストの識別	17
4.2	Bluetooth によるユーザとホストの識別	17
4.3	カメラを用いたユーザ識別の実験	19
4.4	カメラを用いたユーザ識別結果	20
4.5	ネットワーク構成図	21
4.6	ホスト識別を用いたサービス	22
4.7	端末間の関係の表示	23
4.8	システム概要図	25
4.9	Windows 版クライアント	27
4.10	MacOSX 版クライアント	27
4.11	MAC アドレス登録インターフェース	27
4.12	アクセス制限機構	28
4.13	収集状況の表示	29
4.14	ユーザの位置情報表示画面	30
4.15	クライアントのネットワーク遷移	31
4.16	実際のクライアントのネットワーク遷移履歴	32
5.1	パケットヘッダ情報によるプロファイル手法	37
5.2	起動時と復旧時におけるパケットの発信タイミングの比較	39
5.3	パケットヘッダ情報によるホスト識別結果	42
6.1	ホスト情報を推測する手法	48
6.2	ユーザの絞り込み概要図	48
6.3	ホスト所有者の絞り込み手法	50
6.4	人のタイミング情報の登録	51
6.5	ホスト識別手法結果	52

6.6 SNS を利用したホスト利用者特定手法の試行	53
6.7 SNS を利用したホスト利用者特定手法の結果	54

表 目 次

4.1	ホスト識別情報の選定	16
4.2	クライアント開発環境	26
4.3	サーバ開発環境	28
4.4	実験環境	30
4.5	ホスト識別の実験結果	33
5.1	事前に作成したホストのプロファイルデータ	35
5.2	本手法の検証環境	40
5.3	パケットヘッダ情報によるホスト識別結果	41
5.4	事前に作成したホストのプロファイルデータ	44
7.1	OECD ガイドラインと各サービス形態の対応	57

第1章 序論

1.1 背景

近年、情報技術の発展によってユーザに関する様々な情報がデジタル通信上で送受信され、記録として残せるようになった。これによって、ユーザに関する情報を収集、解析することにより、ユーザの要求に応じたサービスの提供が実現した。特にユーザやホストを識別することによって、レコメンドサービスなどユーザに応じて様々なサービスや取り組みが行われている。しかし、そのようなサービスはユーザ自身にアカウント登録などの特別な操作が必要となる。また、ホストを識別する際には、従来IP アドレス Cookie などを用いるが、IP アドレスは可変であり、Cookie などの識別要素は容易に変更、詐称可能である。そこで、本研究は特定のユーザが利用するデバイスを保持するユーザの利用形態によって変化する特徴的なトラフィックパターンや、機器情報を利用する。これによって、ホストとユーザを識別し、必要な場合は共有するプラットフォームを提案する。

1.2 目的

本提案手法ではユーザの発信するパケット情報の特徴を収集し、ユーザやホストを識別するプラットフォームの作成を目的とする。これによって、ユーザに特別な操作や専用機器を強制することなく、ユーザやホストの識別が可能となる。情報の利用ポリシーによって取得できる情報は変化するが、このプラットフォームを作成することによって、ネットワークの状態把握や、企業内におけるフォレンジック、ユーザ間の人間関係の把握、デジタルサイネージなどのユーザに応じたコンテンツを表示など様々な応用ができる。そして、そのプラットフォームの範囲を広げることで、より低コストで様々な場所で情報の利用が実現可能となる。

1.3 アプローチ

本研究はプラットフォーム作成にあたり、デバイスの情報を用いたホストを識別する2つの手法と、ホストとユーザを結びつける1つの手法の合計3つを提案する。図1.1の左上部分が、デバイスの機器情報を利用したホスト識別手法、左下がパケットヘッダ情報を利用したホスト識別手法であり、右部分がユーザとホストを結びつける手法を示す。

1つ目は複数のホストがクライアントとなることで、様々なネットワークにおいてユーザの識別をする手法である。この手法は、本手法を用いたプログラムを各ホストがインス

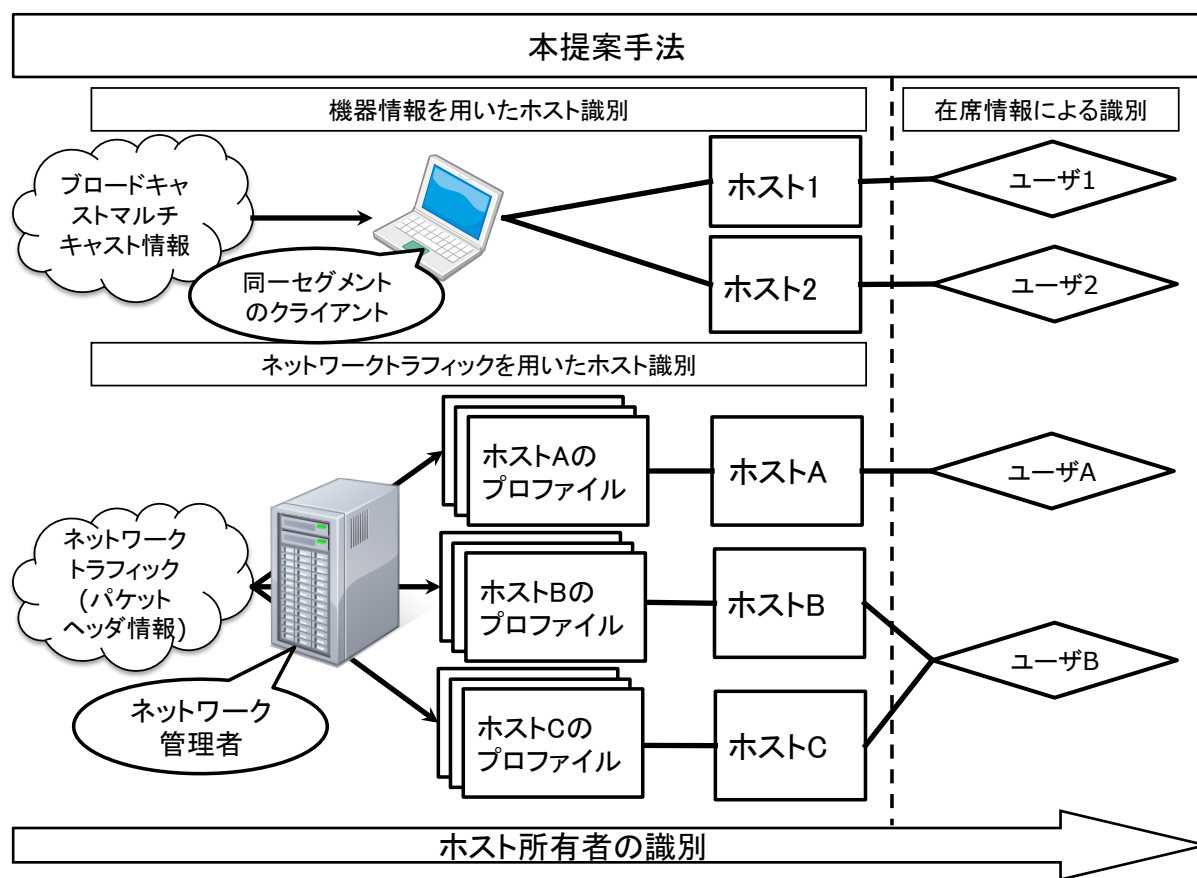


図 1.1: 本提案手法の概要図

トールし、クライアントがセンサとなる。クライアントはネットワークに流れるブロードキャスト、及びマルチキャストを収集することでホストの識別や情報収集が可能となる。そして、本プログラムを利用している同士のコミュニケーションをすることによって、情報の共有や、クライアントの判別が可能となる。この手法は事前にユーザの利用者の同意を得ていることを前提とする。

2つ目のモデルは、ネットワーク管理者などネットワーク上のトラフィックを取得できるユーザが、法的観点やネットワークの制約上ペイロードを見ることができない状況でパケットのヘッダ情報のみから、ホストを識別する手法である。1つ目の手法は複数のネットワークやセグメントにクライアントがいなければ情報を収集できない。しかし、この手法はネットワークの上流にセンサを設置することによって、すべてのセグメントを対象とすることができる。また、ペイロードは取得せずにパケットヘッダ情報のみ取得することで、ペイロードの取得に比べてユーザの同意を得ることが容易となる。本論文で述べるパケットヘッダ情報とは、IP、TCP、UDP、ICMPヘッダの事を指す。MACアドレスはスイッチを経由した場合変更され、取得できないため収集対象から除外する。これによってMACアドレスが詐称された場合でも1つ目の手法とは異なりホストの識別が可能である。

本論文では2つの手法を提案するが、これらはホストのみ識別可能であり、ホストと

ユーザを結びつけるユーザ識別まではできない。そのため、ホスト識別後、ユーザとホストを結びつける手法が必要である。そこで、ユーザとホストの在席情報を利用することによって結びつける手法を提案する。これによって、ユーザがどのホストを保持しているのか、ユーザがデバイスを常に持ち続けているかが判明するため、ホストとユーザの識別ができる。

1.4 本論文の構成

本論文は全 8 章から構成される。第 1 章では本論文の全体の概要及び構成を述べる。第 2 章では本研究において、ホスト追跡とプライバシーを巡る動向を述べた上で、技術的、ポリシ的な課題を述べる。第 3 章では本研究の関連研究として、既存のユーザ、ホスト識別手法について言及し、これらの問題点を明らかにする。第 4 章では機器情報を用いたホスト識別手法の提案と、実装、及び検証について述べる。第 5 章ではネットワークトラフィックを用いたホスト識別手法提案し、実験及び評価について述べる。第 6 章ではホストとユーザを結びつける手法と実装及び評価について述べる。第 7 章では提案したプラットフォームのプライバシーに関わる問題を整理、考察する。第 8 章では本システムのまとめと今後の課題、展望について述べる。

第2章 ホスト追跡とプライバシーを巡る動向

2.1 ユーザ、ホスト追跡の現状

近年、デジタル情報を発信する媒体は、パソコン、携帯電話など多岐に渡る。特にインターネットの発展は凄まじく、日本におけるインターネット人口は1997年の572万人から、10年後の2007年には8227万人[1]と爆発的に増加し、更に増え続けている。このことから、多くのユーザが端末を通じて情報を頻繁に発信していると言える。そして、ユーザがデバイスを持ち歩くようになり、ユーザに特化した情報を提供するサービスの需要が高まっている。本章では、ユーザやホストを追跡するモデルを3つに分けて具体例を挙げると共に、ユーザやホスト識別に対するプライバシーの動向について述べる。近年、特にユーザやホストの識別によるサービスが増加している。これらのサービスの代表的なものには、ユーザにアカウント登録をさせて、ユーザがサーバに向けて情報を発信する2つのモデルと、サービス提供側が周囲の情報を収集するモデルの大きく3つに分類できる。

2.1.1 ユーザ自身が任意のタイミングで能動的に情報を送信するモデル

ユーザ、ホスト識別において最も代表的なものが、ユーザ自身が任意のタイミングで情報を送信するモデルである。図2.1にそのモデルを示す。

まず、ユーザはアカウント登録やアプリケーションをインストールし、サーバにホスト、ユーザを識別する識別要素を付与される。この識別要素に含まれている情報はユーザを識別するためのIDや、クライアントが設定した情報の保存、コンテンツの情報やユーザ情報などが挙げられる。識別要素を保持しているユーザが再度そのサーバを利用する際に、自身が保有する識別要素をサーバに送信する。これによって、サーバは識別要素をもとにユーザを認証する。そして、ユーザの意思で位置情報や情報をサーバに対して送信し、フィードバックを得たい場合に、蓄積した情報の解析結果を得ることができる。これらを利用することによって、ユーザに応じた個別のサービスが可能となる。

ユーザが同意した上で、アプリケーションのインストールを行い、サービスにログインした状態でサイトの閲覧などを行うことで、サーバに情報を送信する。品質向上のためにバグレポートや利用履歴を送信するなどの機能を備えたアプリケーションもその一例である。また、Webサイト閲覧時にログインをした状態でWebページにアクセスすると、おすすめの情報を提供するレコメンドサービスなどもAmazon[2]をはじめとして多く存在する。

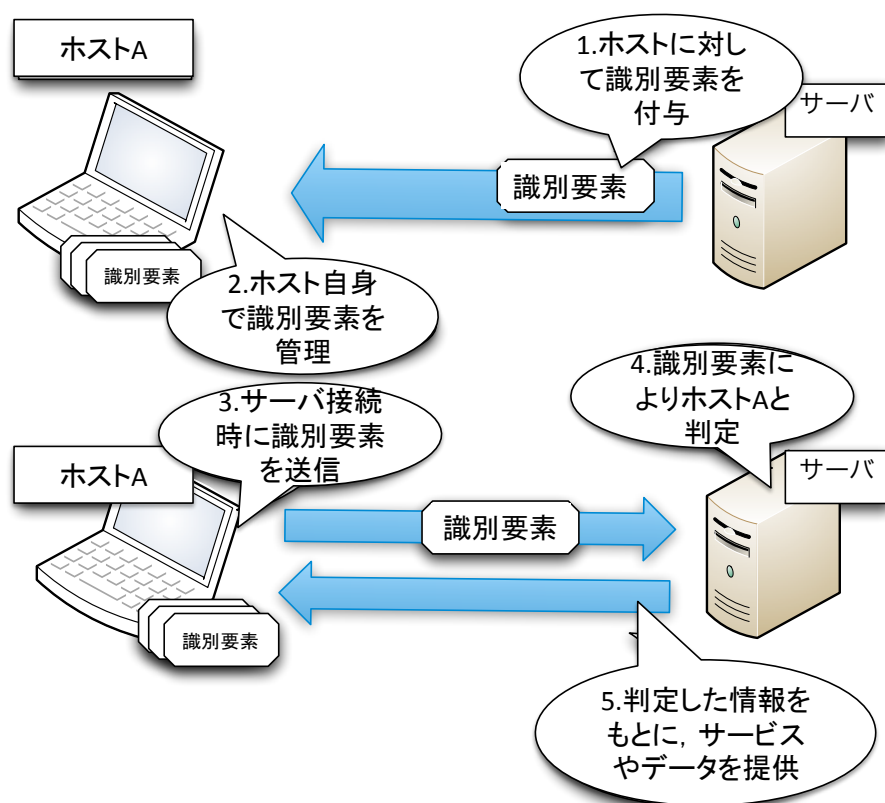


図 2.1: クライアントがサーバから付与された識別要素を管理

このモデルを利用する代表的なプロジェクトに、Microsoft の研究である MyLifeBits Project[3] が挙げられる。この研究は、人の行為をデジタル化して記録に残し、ユーザが過去を分析しながら新たな行動につなげるようになるのが目的である。他にもユーザの位置情報を共有するサービスである foursquare[4] などがある。これらのモデルの特徴は、ユーザは何らかの登録やアプリケーションをインストールし、許可されたデータのみを送信することである。またライフログとは別にユーザ別のレコメンデーションも例として挙げられる。これはユーザの嗜好や行動を分析することによって、ユーザに必要な情報を提供することを目的としている。これによって、ユーザへの効果的な広告などが実現可能となる。

2.1.2 ユーザが常時情報を能動的に発信し続けるモデル

ユーザが携帯電話や位置情報サービスなどを利用し、ユーザ識別情報とコンテンツを同時に発信し続けるモデルである。図 2.2 にそのモデルを示す。

常に情報を発信するものとして、端末にインストールするエージェントが挙げられる。このアプリケーションによる研究は、高度なパーソナライズ実現のためのユーザプロフィール統合サービスエージェントの設計 [5] をはじめとして広く研究されている。他にも、インターネット上でのトレンドを知る調査にもこのような手法は利用されている。ネット

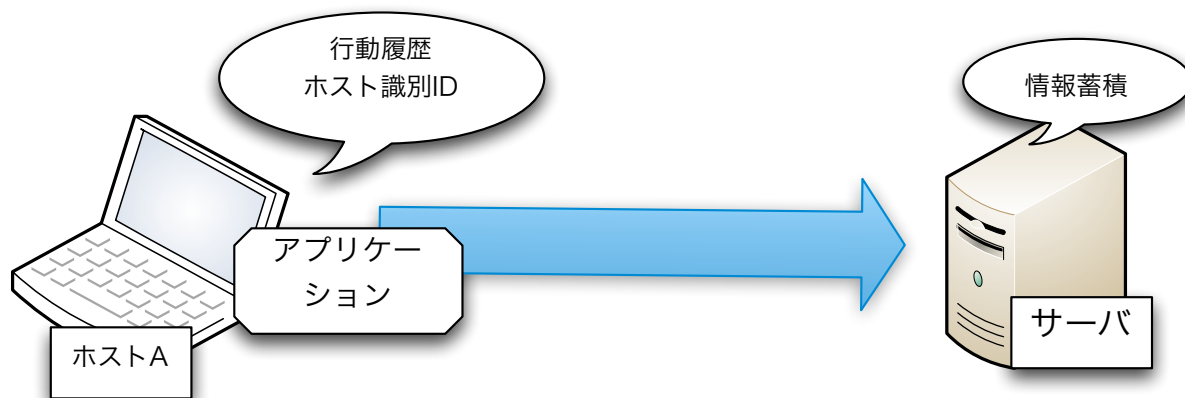


図 2.2: ユーザがサーバに情報を直接送信

視聴率白書 2008-2009[6] でもこの手法が採用されている。しかし、これらの手法はクライアントすべてにアプリケーションを導入しなければならないため、情報を取得した側のコストが高い。常に情報を発信し続けるものとしては RFID, Zigbee といった機器、携帯電話が利用している基地局情報、カーナビなどの位置情報サービスなどが挙げられる。これらは、専用の機器が必要となり、サービス提供側はユーザに配布や、専用ソフトのインストールをさせる必要がある。そして、常時情報を提供するため、ユーザの同意が不可欠となる。

2.1.3 サービス提供者が受動的に情報を収集するモデル

ユーザが通信をする際にネットワーク全体や周囲に対して情報をブロードキャストやマルチキャストによって通知する場合がある。また、Bluetooth や RFID など機器を利用して情報を発信する方法もある。それらの状態を取得して、サービスを提供や、トラフィック情報を収集することができる。また、ネットワークにおいて通信のやりとりからアプリケーションを推測するなど、データ収集側は受動的に個人情報を収集する。そのモデルを図 2.3 に示す。

Deep Packet Inspection(DPI) と呼ばれる技術がある。この技術はリアルタイムにフローパターンやパケットの特徴や振る舞い、パケットの制御情報などからトラフィックを解析する技術である。この技術を利用して行動分析型ターゲット広告をする企業に NebuAd[7] が挙げられる。NebuAd はユーザの同意を得てサービスを提供する代わりに、ユーザの行動を分析して広告するというサービスを提供している。取得した情報をデータベース化し、その統計からアプリケーションの特定する。日本国内でも、ISP が P2P プロトコルを利用しているホストのトラフィック制御にこの技術を利用している。しかし、この技術は通信の秘密に抵触する可能性があるというデメリットを保持している。このように、行動ターゲットに関する議論はオプトアウトに関しては様々な議論や解釈がある [8]。

受動的にトラフィックを収集、解析する研究は多く行われている。例えば、ネットワー

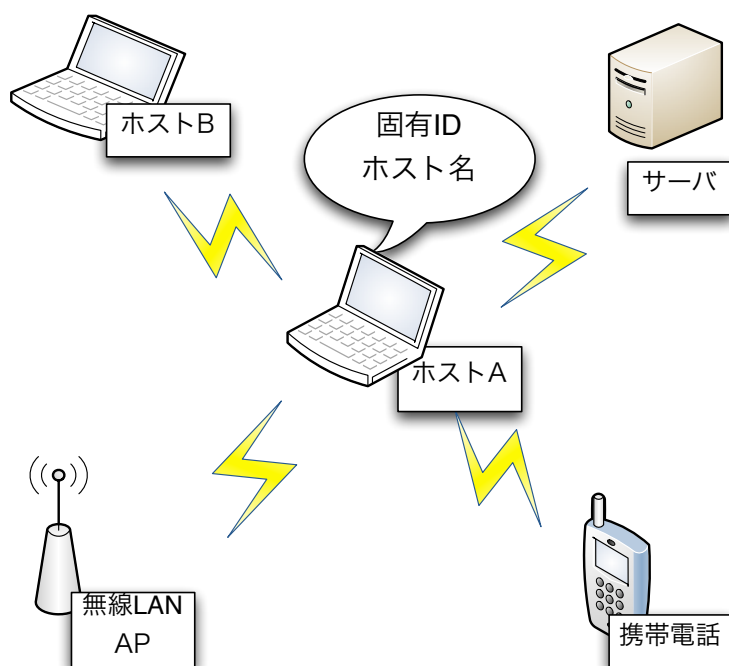


図 2.3: ユーザが全体に情報を発信し, 受信側がデータを処理

クトラフィックに着目することでセキュリティインシデントをリアルタイムで発見する研究に Passive Network Discovery for Real Time Situation Awareness[9] などがある。この研究は様々な Passive finger printing を利用することによって、ネットワークに負荷をかけることなく情報を取得し、ユーザを特定する。この研究で利用する Passive finger printing によって取得できる情報は稼働中のホストや OS 情報、ホストの役割、提供サービス、プロトコル、ネットワークの IP アドレス設定である。他にも、BLINC : Multilevel traffic classification in the dark[10] などが挙げられるようにパケットのヘッダ情報を利用してネットワークに流れるプロトコルやアプリケーションを把握する研究など様々である。

このように、デジタル情報を収集する上で様々な技術がある。その反面、これらの技術を利用することで、ユーザのプライバシーを脅かされる可能性がある。

2.2 現状のホスト, ユーザ識別

2.2.1 ホスト, ユーザ識別の問題点

スマートフォンなどのモバイルデバイスやパーソナルコンピュータの小型化によって、端末を保持、持ち歩くユーザが増加している。一方で、ユーザが意図しない間にも端末が様々な情報を発信し、それらの情報からユーザが追跡される可能性がある。例えば、サービスの提供を受けるつもりがない場合にも周囲にデータを発信している場合や、ユーザが知らない間にユーザ情報を含む通信を発信している、もしくは利用されている場合も

ある。

ユーザが知らない場合にも情報を収集されることが問題になったケースの 1 つとして、Android OS における Carria IQ によるデバイスの記憶の事例が挙げられる。これは、対象機器の保持する操作の履歴や位置情報、通信をすべて記録するという問題があった。セキュリティチームの解析によると、キャリア側の問題であることが明らかになった [11] が、情報を無断で利用されることに対して社会的に懸念される問題であることが示された。他にも、ユーザは識別要素を自身で管理するが、識別要素を利用していることを理解した上で容易に管理できる場合と、本人が知り得ぬところで識別要素を付与され管理が難しい場合がある。例えば、HTTP Cookie は多くのサイトで利用され、Web ブラウザの設定で Cookie の制限は容易に可能である。最近では、Firefox [12] などのブラウザは設定によって、HTTP Cookie や履歴など痕跡を残さずに Web アクセスが可能となるため、HTTP Cookie の利用に関してユーザのプライバシーに配慮がされていると言える。HTTP Cookie とは別に、本人が知り得ぬところで識別要素を付与され、管理が難しいという識別要素の例に Flash Cookie が挙げられる。この Flash Cookie は、Web ブラウザだけでは削除することはできない。保存できる情報量は HTTP Cookie が 4KB に対して、100KB であり、より多くの情報が記録できる。Flash Cookie を確認・管理する方法も、Adobe [13] の Flash Cookie のページで設定しなければならない。また、HTTP Cookie に比べて認知度が低いため、存在を知らないユーザも多い。Flash Cookie の危険性についての調査 [14] において、QuantCast [15] に掲載されているユーザ利用率が高い上位 100 サイトのうち、54 のサイトで Flash Cookie が利用されていることが示されている。そのため、ユーザが管理しにくい、もしくは知りえぬ識別要素がネットワーク上でやり取りされていることを示している。このように Flash Cookie はユーザが知り得ぬところで識別要素が付与され、かつユーザ自身が識別要素の管理が困難である。そのため、プライバシーを脅かす可能性がある。

位置情報を保護する問題として、米国議会上院で、セキュリティ専門家の Soltani 氏は、わたしの情報をそれほど持っていない人物が、わたしを追跡することが可能になっている。以前どこに住んでいたのか、どこに引っ越したのかが分かってしまう [16] との証言をしたように、近年 MAC アドレスに基づいた情報に対して、議論されることが増えている。

i モード ID [17] といった機器独自の ID を識別要素にした、かんたんログインというサービスがある。これは、機器固有の ID を利用することで認証サービスを提供するものであるが、固有 ID は詐称が可能であり対象ユーザになりすますことができるという問題がある。そのため、どの情報が識別要素になりそのリスクは何かを明確に議論されなければならないということが社会に示された一件である。

次に、普段我々が発信している情報を利用されることで、プライバシーが脅かされてしまうことがある。その問題の 1 つとして挙げられているのが、Google による AP 情報を収集する行為が挙げられる。Google が街中で配置されている AP 情報を収集することで、無線 LAN アクセスポイントの所有者の住所などに結び付けられる懸念があるという指摘がされた。他にも Google が公開する WiFi 対応デバイスの位置情報サービスなど、ユーザがサービスを利用する上で、どのように対応するかが焦点とされている。

もし、これらの情報がプライバシーを侵害される場合、これらの情報を第三者が取得、分

析することで、ホストとユーザを結び付けて、追跡される危険性がある。これによって、利用者の行動履歴などや、ホストに関する情報を取得されプライバシーを脅かされるリスクが発生する可能性がある。本論文で述べるリスクとは、ホストとユーザが結び付けられることによって、ユーザの追跡が可能となることと定義する。しかし、これらの情報はよりよいサービスを提供する上では非常に重要なデータとなるため、ユーザの利用形態やプライバシーの対応をすることで、ユーザの恩恵も大きくなる。そのため、如何にこのような状況に対応するかが求められている。

2.2.2 現状の問題点とその対応

第 2.2.1 節で述べたように、近年、個人情報に関する事件が多発した理由からサービス、コンテンツ提供者もユーザのプライバシーについて対策する必要がある。しかし、個人に関わる情報を利用することによって、効率的なサービスが可能となる。プライバシーの起源は、Samuel D. Warren, Louis D. Brandeis らが 1980 年に発表した The Right to Privacy[18]にある the right to be let alone であると言われている。そこからプライバシーは個人に関する情報が本人に同意を得ることなく利用されないことを指す。近年は情報化に伴い様々なプライバシーを守る法律や、技術が確立されている。

このような情報化の背景により、ユーザも利用する企業やサイトがユーザプライバシーを守るかという点に注意を払うようになった。サービスやサイトがプライバシー対策を施しているかどうかの判断基準の一つにプライバシーマーク制度[19]がある。これは日本工業規格 JIS Q 15001 個人情報保護マネジメントシステムの要求事項に適合して、個人情報について適切な保護措置が可能な体制を整備している事業者等を認定する制度であり、プライバシーマークをサイトに記載されている。図 2.4 にプライバシーマークのサンプルを示す。



図 2.4: プライバシマークサンプル



図 2.5: TRUSTe マークサンプル

このような認証マークを付与する事例は数多くあり、TRUSTe もその一つである。TRUSTe は情報の完全公開と利用者の同意を基本理念としており、1980 年 9 月 23 日に経済協力開発機構の理事会で採択された「プライバシー保護と個人データの国際流通についての勧告」の中に記述されており、8つの原則をもとにしている。図 2.5 にそのサンプルマークを示す。これらのマークを付与しているサイトでは、個人情報の利用規約を記載し、ユーザに同意を得るケースが多い。これらのマークをサイトに載せることによって、利用者はそのサイトを信頼できるかの基準とすることができる。

企業だけではなく、通信事業者にもユーザのプライバシーを保護する法律が存在する。電気通信事業法第 4 条第 1 項で通信の秘密において”電気通信事業者の取扱中に係る通信の秘密は、侵してはならない”として定めている。通信の秘密が焦点となった例に、NTT のぷらら [20] の Winny 規制サービス提供が通信の秘密に抵触する恐れがあることが総務省から通達された事例が挙げられる。これは、NTT のぷららが Winny の利用を禁止するために、ユーザの通信を閲覧し、規制するというものである。ここで焦点になったのはユーザの同意を得ていないという点である。この通達を受け、ユーザの同意を得るという条件で Winny の規制サービス [21] に修正した。日本国外でもこのような事例が多くある。前述した米 NenbuAd では、同意を得ないユーザがサービスから外れるオプトアウト形式でユーザの通信を取得して広告を出すというサービスを展開していたが、多くの提携先やユーザが難色を示し、最終的には CEO が辞任しアメリカから撤退するという事例があった。これらの事例から、個人情報収集・解析をする際には、ユーザの同意や理解を得ることが必要と言える。

また、個人情報を所得管理する側が利用する技術は多岐に渡る。例えば、ユーザの通信をすべて暗号化することや、データを集計して統計にする際にランダムに変数を変化させる k-匿名性を用いた技術 [22] などがある。これによって、ユーザのプライバシーを侵害しないまま、統計を取ることが可能になった。

他にもユーザのプライバシーを保護する法律に電波法が挙げられる。電波法にも秘密の保護として第五十九条に”何人も法律に別段の定めがある場合を除くほか、特定の相手方に対して行われる無線通信（電気通信事業法第四条第一項又は第九十条第二項の通信たるものを除く。第百九条において同じ。）を傍受してその存在若しくは内容を漏らし、又はこれを窃用してはならない。”[23] と明記されている。つまり、管理者や同一セグメントに存在するユーザが、他のユーザのパケットを取得した場合、同意を得た場合を除き、そこから判明した情報を漏らしてはいけない。そのため、管理者はトラフィックから得た情報を秘匿する必要がある。

また、これらのサービスを利用するためにはユーザアカウント登録や、専用デバイスの利用などの制約が存在する。そのため、サービス提供者はユーザに特別な操作や設定などを求めることなくサービスを提供することが求められている。

2.2.3 サービス実現のための契約形態

個人に関わる情報やユーザ追跡に関するサービスを行う上で必要となる点は、どのようにユーザの許可を得た上で、サービスを提供するかを決める運用、契約形態である。運用、契約形態は大きくオプトイン、オプトアウト、ノンプライバシーの 3 つの形態に分類できる。ユーザのプライバシーに関する情報をサービスとして利用する場合は、どの形態が良いのかを検証しなければならない。

オプトインは様々なサービスやアプリケーションが利用されている方式である。利用用途を明確にした上で、ユーザに同意を得てサービスを提供する。これは、現在最もメジャーな方法であり、事前に契約をすることでプライバシーによる懸念を払拭する。しかし、

許可したユーザのみしか情報を収集することができず、利用者が少なくなるという問題点がある。

オプトアウトは、オプトインとは異なり、拒否しないユーザはすべてこのサービスを許可しているという前提でデータ収集やサービス提供をする方式である。一つの例として挙げられるものとして、Google の AP 情報収集が挙げられる。これは、2.2.1 節でも述べている。この対応として、Google は AP の SSID に特定の文字列 ‘_nomap’ をつけると、収集されないオプトアウトの形式を保持している。これによって、特定文字列がある AP は収集対象から外すという行為をしている。このように、拒否するユーザが意思を表示することで、サービスや収集対象から外れることをオプトアウトと呼ばれる。この方式は、すべての機器ユーザを対象とすることができるが、サービスや収集されていることを知らないユーザの情報も収集することとなるという問題が挙げられる。

また、個人情報の保護に関する法律における第三者提供におけるオプトアウトとは、あらかじめ以下の 4 項目を本人に通知する、もしくは本人が容易に知りえる状態に置いておくと共に、本人の求めに応じて第三者への提供を停止することである。ここでの 4 項目とは第三者への提供を利用目的とすること、第三者に提供される個人データの項目、第三者への提供の手段又は方法、本人の求めに応じて第三者への提供を停止することを指す。

ノンプライバシーは利用者のすべての情報を収集する方式である。これには、テロ対策や犯罪捜査と言った国主体で行われることが多い。一つの例としては、2001 年に設立されたアメリカの公立法 107-56, Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001[24] が挙げられる。また、国だけではなく企業においても勤務中の社員のネットワークトラフィックの監視や管理をことも多いと言える。この手法によってペイロードや個人情報を収集できる反面、明確な法律や契約がない場合は収集が犯罪と見なされる可能性がある。

2.3 新しいサービスの動向

第 2.2.3 節で述べた課題をクリアすることによって、新しいサービスや、今まで実現が難しいとされていたサービスが可能となる。

例えば、2010 年において Google[25] 社はデバイスの MAC アドレスを用いたデバイス追跡プラットフォームを提供した。これによって、スマートフォンやデバイスホストの接続している AP 情報と事前に登録した場所情報、及び GPS 情報をもとにその端末の位置情報や、行動履歴が閲覧できる。これによって、デバイスを紛失した際の対応が容易になる。

Connect Free 社の提供するサービスも新しいサービス形態として着目できる。このモデルケースは、AP を各所に設置してユーザに無料に提供する代わりに、ユーザに応じた広告を提供する。その際に、ホストの情報からユーザの趣味趣向を知ることができれば、より効率的な情報提供が可能となる。また、ホスト情報を同じネットワークで共有することで、ネットワーク上でのコミュニケーション促進や、ネットワークの状況把握が可能となる。このように、新しいサービスの創発ロケーションベースのサービスに応用可能で

ある.

このように, ホストやユーザの識別する手法は様々なプライバシーの問題を含んでいるが, 十分に議論することによって, 今後のサービスの発展が期待できる.

第3章 関連研究

ユーザが発信する情報を利用して、ホストやユーザに関する情報を取得する様々な研究が存在する。ここではユーザが発信する情報をもとにホストやユーザを特定する研究を挙げる。

3.1 ネットワーク管理者がユーザ発信する情報を収集解析する手法

受動的にトラフィックを収集、解析する研究において、ネットワークトラフィックに着目することでネットワークの状態やプロトコルを把握する研究に KARAGIANNIS らの研究 [10] や、Montigny-Leboeuf らの研究 [9] が挙げられる。また、和泉らはパケットのヘッダ情報を利用してネットワーク上に存在するホストが利用するアプリケーションを識別する手法 [26] を提案している。しかし、これらの研究の目的はネットワーク状態の把握であり、ホストやユーザを特定する本論文の目的とは異なる。

3.2 SNS 上の情報とホストを結びつけるユーザ特定手法

ソーシャルネットワーキングサービス (SNS) を利用したプライバシーの脅威を調査した研究に Krishnamurthy らによる研究 [27] が挙げられる。この研究はユーザが SNS に登録する情報と SNS から発行される Third-party Cookie を解析することで、個人と IP アドレスを結びつける手法とそのリスクを提唱している。また、複数の SNS から情報を収集することによって、個人情報を複数取得し、ユーザのプロファイル作成を可能にする。SNS には、E-mail アドレス、住所に関する情報、本人の写真などを記載する場合があります。これらの情報を識別要素とすることで、個人情報を得る。しかし、Web ページや SNS を利用しているため、対象ユーザがこれらの情報を利用していない場合は対応できない。

3.3 ブラウザ情報を利用したホスト識別手法

ブラウザの情報を利用することでユーザの識別が可能か検証する研究に Electronic Frontier Foundation による研究 [28] が挙げられる。この研究では HTTP における User Agent string、プラグインのバージョンやフォントの設定などのデータを総合して、ホストを識別する。そして、ユーザのブラウザ情報を収集したデータベースをもとに、識別するプロ

グラムを公開することで、ユーザに Web 閲覧などの情報を利用したトラッキングや広告に対する脅威を周知することを目的としている。ブラウザ情報といった普段意識されない情報を利用してホストを識別する方法は、本研究と類似するものがあるが、特定のブラウザしか情報を取得できないことや、HTTP を利用しないホストの識別ができないという制約がある。

3.4 Web 上での情報収集

インターネット上での検索エンジンを利用することで、対象とするユーザの人間関係や、社会的な立場が明らかになる場合がある。松尾らによる Web 上の情報からの人間関係ネットワークの抽出 [29] では、検索エンジンを用いてターゲットとなるユーザの人間関係を抽出している。これによって、ユーザの実社会における人間関係や研究分野などを知ることが可能となる。推測に利用している情報はすべて Web 上で公開している情報のみであるが、適合率は 8 割を超えるため非常に有用であると言える。他にも、本村らによるネットワーク上の情報を利用して、プライバシーを侵害や対策 [30] が提案されている。しかし、これらの研究は対象ユーザが Web 上に情報を公開していない場合はこの手法は利用できず、汎用性に欠ける。

3.5 まとめ

本章では、既存のアプリケーションや過去に提案された枠組みから本研究に関連の深いものについて述べ、現在のホスト、ユーザ識別手法において、汎用性が低い点などの問題点を明らかにした。次章では、本章で述べた問題点を解決するシステムの提案する。

第4章 ネットワークデバイス情報を用いたホスト追跡手法

4.1 手法概要

本研究ではユーザが普段発信している情報を収集することで、ユーザやホストの追跡できる可能性に着目した。本研究で取り扱う情報は、デバイスが発信するブロードキャスト、マルチキャスト通信であり、各デバイス特有の情報をもとにプロファイルを作成する。そして、ホスト固有の情報をもとにホストを識別し、ユーザを追跡できるかを検証する事前実験を行った。事前実験で想定する環境はホットスポットや、公衆無線 LAN など不特定多数のユーザがネットワークに参加し、特定の場所にセンサを設置できる環境である。次に、収集するセンサプログラムを数十台のクライアントにインストールし、事前実験と同様に不特定多数のユーザが存在する複数箇所のネットワークにおいて、ホストを追跡できるかを検証する。

4.2 端末識別情報の選定

ホストを一意に識別する情報を選定する。表 4.1 に各情報をまとめる。

- IP アドレス

ホストの識別という点において、ネットワーク上で利用されているものに IP アドレスが挙げられる。ホストは IP アドレスを付与されることによって、ローカルなネットワークを越えて通信が可能となるため、識別子の汎用性は高く取得も容易と言える。また、識別子の詐称も特定の場合でしかできないため、詐称は困難である。しかし、IP アドレスは可変であり、DHCP サーバの設定によっては常に同じ IP アドレスが振られ続けるという保証が無い場合、必ずしも常に同じ機器を識別できるわけではない。

- ホスト名

ホスト識別において利用できる情報に、ホスト名が挙げられる。これは samba や mDNS といったプロトコル利用時に、対象を決定する場合に利用することが多い。取得も容易であり、同じネットワークセグメントにいただけで取得できることや、DHCP サーバのブロードキャスト情報を取得することでホスト名を取得できる。しかし、スマートフォンなどはデフォルトのホスト名や機器名が利用されていること

表 4.1: ホスト識別情報の選定

識別情報	一意性	詐称の容易さ	取得, 識別の容易さ
IP アドレス	一意	△	○
ホスト名	環境に依存	◎	△
MAC アドレス	一意	○	◎
Bluetooth ID	一意	△	○
アカウントの登録情報	一意	△	◎
Cookie	一意	○	○
パケットのペイロード情報	取得情報に依存	○	△
写真, 映像による識別情報	△	○	△
回線保有会社から付与された ID	一意	×	△

◎：容易, ○:可能, △：条件次第, ×：困難

も多く、複数の端末で同じホスト名が確認されている。また、ホスト名は変更が容易であり、識別要素になりにくい点からホストの識別子としては利用が困難である。

● MAC アドレス

機器のネットワークカード毎に割り振られている MAC アドレスはホストの識別要素としてなり得る可能性が高い。MAC アドレスは IP アドレスとは異なり、基本的にユーザが意図的に変更しない限りは、変化することはない。また、パケットの Ether ヘッダとして必ず付与されているため、同じネットワークで、対象ホストのブロードキャスト情報を取得するだけで容易に取得可能である。そのためホストの識別は容易であり、MAC アドレスを利用した認証機構 [31] は多く存在する。しかし、ホストの詐称は比較的簡単に変更可能である。これらの識別要素の特徴を踏まえて、どの程度ホストの出現を検知できるか、筆者の所属する研究室で対象のホストを決定し、約 3 日間 MAC アドレス収集、識別をする実験した。その結果を図 4.1 に示す。

X 軸はユーザが出現した時間であり、Y 軸は対象ホストのラベル。各グラフのラベルは事前実験を実施した日付である。この図から対象ホストは初日、2 日目の 13 時まで滞在し、3 日目にまた出現したことを示す。この結果は、実際のホストの挙動と一致しており、ホストの識別をできていることを示した。

このことから、MAC アドレスはホスト識別要素として利用できることが分かった。

● Bluetooth ID

Bluetooth のペアリング時にお互いの Bluetooth ID を利用するため、機器の識別という点では十分に利用可能である。詐称に関しては、ID 変更は MAC アドレスよりも困難である。この Bluetooth 情報の特性をもとにユーザ毎のライフタイムを収集した。その結果を図 4.2 に示す。この実験も、同じく筆者が所属する研究室で約

4.2. 端末識別情報の選定 第4章 ネットワークデバイス情報を用いたホスト追跡手法

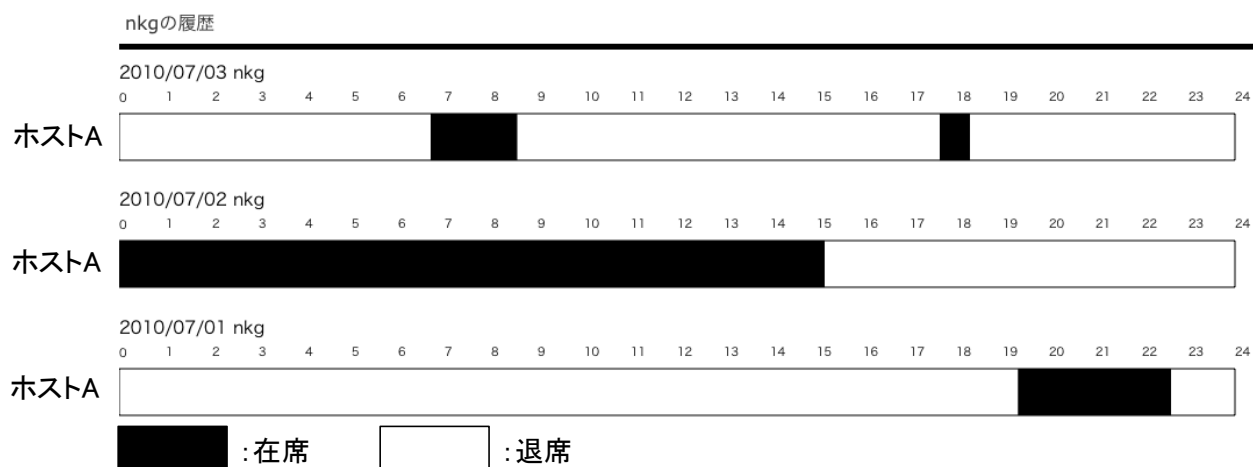


図 4.1: MAC アドレスによるホストの識別

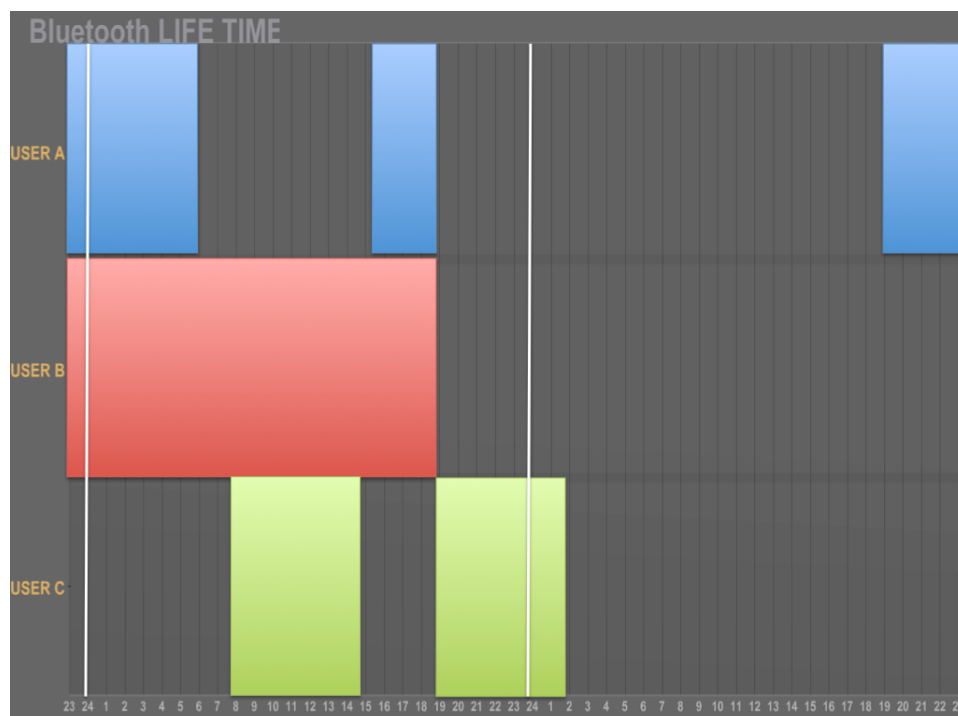


図 4.2: Bluetooth によるユーザとホストの識別

2日行った。X軸が取得した3日間の情報であり、Y軸が各ユーザのホストのラベルである。この検証では、一般的に持ち歩きしているユーザのホスト3種類を対象にした。3台のホストの収集することによって、各ホストにも特徴が出るのがわかった。ユーザAのホスト、ユーザBのホストは比較的夜晚に出現する傾向があることに対して、ユーザCのホストは昼に出現が確認されている。また、この検証は木曜日から土曜日にかけて行っているため、どのユーザも土曜日の夕方まで研究室を

訪れてないことが分かる。

この情報から、Bluetooth がユーザや機器の識別がある程度できる事がわかった。しかし、すべての機器に Bluetooth が搭載されているわけではないため、対象範囲がかなり狭まってしまう問題がある。2010 年度におけるインターネットコム株式会社と株式会社マーシュによる Bluetooth に関する調査 [32] においては 300 人中 76 人、約 36.5% のユーザが利用しているとの結果を発表した。このことから、識別要素として利用するためには、対象ユーザが十分に多いとは言えない。また、常に Bluetooth をオンにしていないと取得できない問題もある。そのため、必ずしも正確に取得できるとは限らない。

● アカウントの登録情報

ユーザやホストを識別する一般的な手法にアカウントを発行して、ユーザにログインさせる方法がある。ユーザがパスワードなどを入力し認証されることでホストやユーザを識別することができる。基本的にはアカウント作成時に、利用規約で同意を求めるため、プライバシーや個人情報利用に関して、サービス利用範囲において問題になりにくいという利点がある。また、ユーザが指定したアカウント名、パスワードの入力を求めて認証をするため一意性は高い。しかし、ユーザが最初にアカウントを作成する手間や、サービス利用時にログインという動作が必要になるためユーザにとってコストが高い。そのため、アカウントを作成しないホストや、アカウント登録できない機器はこの認証を利用できないため識別できないという問題がある。

● パケットのペイロード情報

ホストが発信するパケットのペイロード情報を利用して、ホストの識別が可能である。この手法はペイロードの中に含まれているアプリケーションやサービスの ID を利用してホストやユーザの識別が可能である。しかし、法的観点や取得環境などによって制約が多く、この手法は利用できないことが多い。例えば、ホストが発信するパケットを収集するために、ネットワーク管理者などがネットワークの中継地点に収集機器を設置する必要がある。また、ペイロードを閲覧する場合はユーザの個人情報と直結するため、すべてのユーザに同意を得る必要がある。そのため、犯罪捜査や企業における監視など、利用ケースが限られる。

● Flash, HTTP Cookie 情報

サービス提供者側はインターネットを経由したホストを識別するために Cookie と呼ばれる識別要素をホストに付与する手法を利用している。Cookie には HTTP Cookie と Flash Cookie と呼ばれるものがあり、両者ともホストのブラウザに対して付与するものである。その各 Cookie に識別要素が含まれているため、Cookie を取得することで各ホストのブラウザの識別ができる。しかし、ユーザが Cookie を消去した場合はホストを識別することができず、Cookie を利用していないホストに対しては識別ができない。詐称に関しては Cookie の中身を暗号化することで ID を取得されないようにできるが、MAN-IN-THE-MIDDLE ATTACK によって詐称もでき



図 4.3: カメラを用いたユーザ識別の実験

る場合がある。また、暗号化されていない場合はユーザが自由に Cookie を編集やなりすましが可能である点が挙げられる。

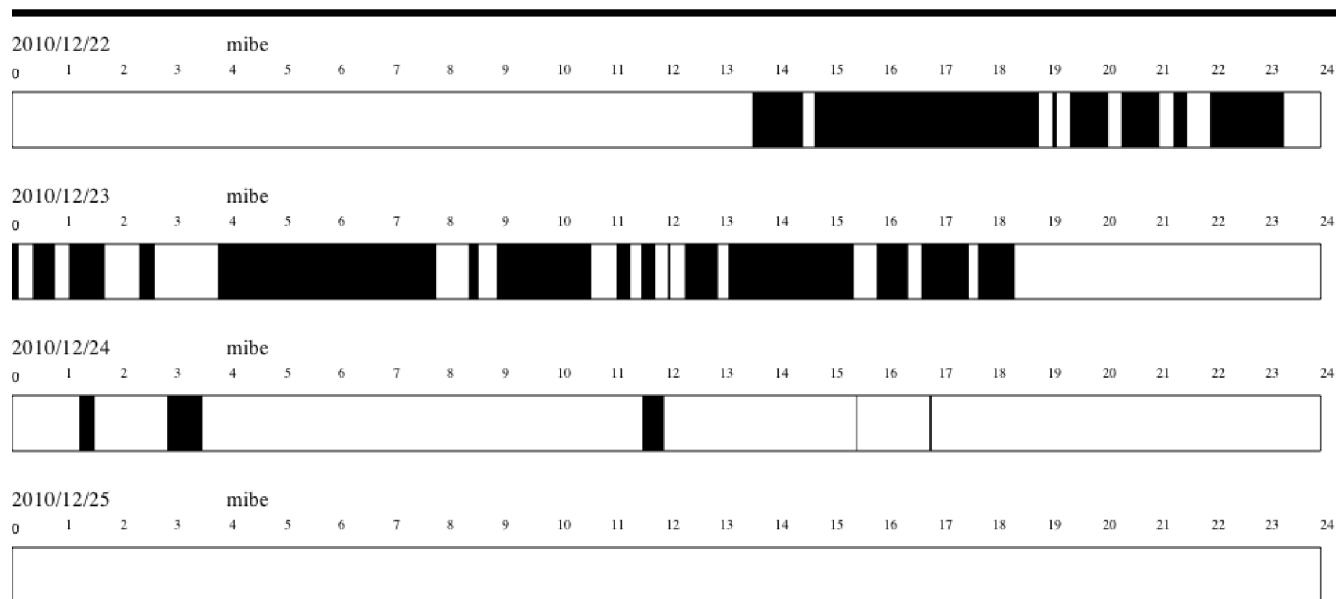
- 写真、映像による情報

監視カメラや、写真、視認による情報を利用してユーザの識別できる可能性がある。例えば、人の顔認識を利用してユーザを識別することができるが、ユーザの顔が分るように撮影する必要や、部屋の明暗など撮影条件が多くあるため、簡単に取得することができない。また、各ユーザのサンプルを事前取得し、機器に学習させる必要があり、運用コストは高いと言える。

次に、対象ユーザの席などユーザしか利用しない場所がある場合は、いるかいないかの存在確認は、顔認識に比べて容易である。そこで、対象ユーザの席にマークを設置し、カメラでそのマークを監視することで対象ユーザの存在確認が可能かを調査した。図 4.3 にその実験の様子を示した。ディスプレイに写っている画面がカメラで取り込み、ARToolKid マークを検出する。このマークが映る状態でカメラを固定し、人が座っている間はマークを認識できないため在席と見なす。この手法で取得した結果をその結果を図 4.4 に示す。この実験は、ユーザが席に座るとマークがカメラで認識できなくなり、ユーザがいたと判定し図を黒く塗る。この施行を約 4 日間実験し、対象ユーザのライフタイムを収集できるか実験した。その結果、3 日間はユーザの挙動と同じ内容の結果を出したが、最終日はユーザのマークが破損したため、情報を収集することができなくなった。この結果から、対象ユーザ毎にカメラを設置するなど、運用コストが高く情報が取れなくなる可能性が高いことがわかった。

- 回線保有会社から付与された ID

携帯会社などの回線保有会社による利用端末の ID を用いたホスト識別が挙げら



Copyright © nakajima All Rights Reserved.

図 4.4: カメラを用いたユーザ識別結果

れる。この手法は、携帯電話など、各回線を利用する機器に各会社の ID が付与される。機器はその ID をもとに通信を行うため、確実に機器の識別ができる。また、詐称防止の設計もされているためなりすましもされにくい。そして、利用端末の基地局を収集することによって機器や保有者の行動履歴を取得することができる。しかし、契約した機器のみしか情報を収集することができず、行動履歴も個人情報となるため、利用できるケースは限定される。

これらの情報を踏まえて、多くのネットワークで情報を収集できる汎用性、取得が簡単に出来る容易さ、ホスト識別ができる正確性を考慮した結果、MAC アドレスが最も妥当であることが判明した。

4.3 事前実験及び検証

第 4.2 節で示した結果を踏まえ、MAC アドレスをもとにホスト識別がどの程度変更されるかの検証と、ホスト識別が長期的に可能かを観測した。想定するネットワーク環境を図 4.5 に示す。対象ネットワーク上に存在するすべてのホストのブロードキャストを必ず受け取れることを前提とする。そして、そのネットワークにおいて、2009 年 3 月から 2011 年 12 月の約二年間、登録した 23 人のホストの MAC アドレスが、日常利用において変更されるかどうかとホストの識別がどこまで可能か計測する実験を行った。

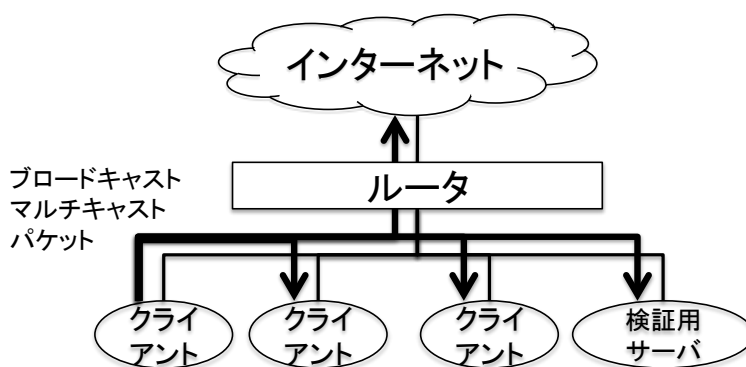


図 4.5: ネットワーク構成図

4.3.1 行動履歴の取得

端末を持ち歩く機会が増加した結果、端末がネットワークに接続、切断する情報を収集することによって、端末所有者の行動履歴を取得することができる。例えば、大学にいるユーザの場合、ネットワークに接続した時間帯を出席時間、切断した時間を退席時間とすることで、ライフタイムを取得することができる。また、本システムを各場所に配置することで場所情報を追加し、より詳細なユーザ情報を得ることができる。これによって、時間、場所情報を収集することができる。場所情報の取得は、Aircap[33]などの無線の電波を取得するツールを利用することやネットワークの位置を知ることによって収集できる。このツールを利用することで、無線情報を利用しているホストの MAC アドレスが判明し、物理的な位置の把握が可能となる。

ユーザの行動履歴を取得する際に重要となるのは、ネットワークに接続、切断時間と場所に関する情報である。ネットワークにおいて、取得した情報が検出された場合は、ホストの所有者が存在する可能性が高い。そこで、場所情報、時間から、ホスト所有者がどのような移動をしたのか推測することができる。

これらの情報を踏まえて、筆者が所属する研究室で、ユーザの同意を事前に取得し、同意を得たユーザのホストを登録することで、誰が対象のネットワークに参加しているかを検証した。図 4.6 にそのサービス表示例を示す。

これによって、各ホストの在籍時間帯を取得することができる。第 4.2 節の図 4.1 に示すように、対象のホスト利用者の行動を表すことができる。これによって、ユーザがいつ研究室にいたのか、いつ退席したのか、授業の時間帯を知ることができ、各ユーザの入退室を管理することが可能である。

4.3.2 端末間の関係

各端末間のネットワーク接続時間帯と接続場所の情報を収集し、組み合わせることによって、端末間の関係が推測できる。例えば、同じ時間帯にネットワーク接続、切断を繰り返した場合、同じユーザが保持する端末の可能性が高い。同じ端末でない場合も、ホ

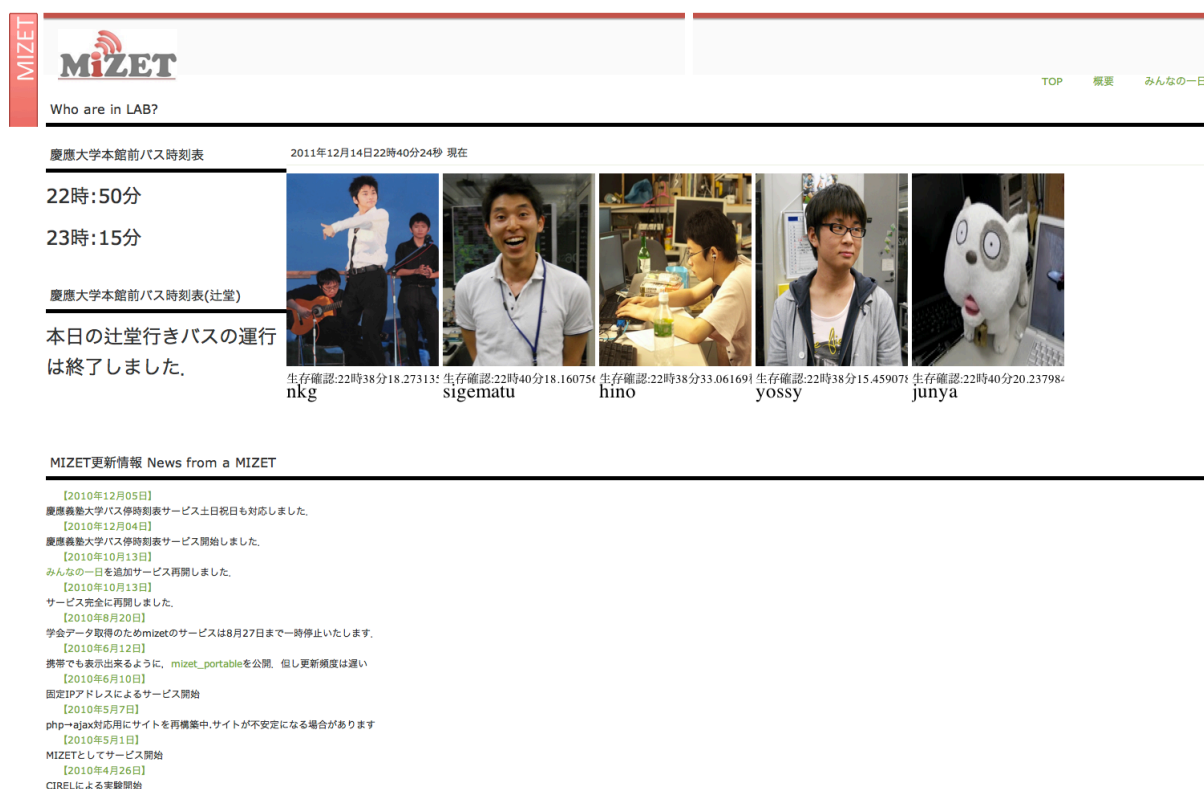


図 4.6: ホスト識別を用いたサービス

スト所有者の関係を推測することができる。また、第 4.3.1 節のように行動履歴を取得し、各ユーザの時間情報と場所情報を組み合わせることによって、ユーザの人間関係などを推測できる。

ユーザは一人で複数のホストを保持することが多くなった。例えば、スマートフォンとホストなど複数の端末を保持している場合同じ時間にネットワークに接続し、同じ時間に退室していることが考えられる。また、異なる所有者同士であっても、同じ時間帯にネットワークに接続、切断を繰り返す場合、人間関係は強固であるといえる。つまり、ネットワークへの接続時間帯、切断時間、場所を取得することによって、端末間の関係を推測できる。

端末間関係を表示する機能について述べる。端末間関係は移動履歴で取得したデータをもとに、同じ結果や、類似した結果をもとに、端末間の関係を表示する。これによって、一台で複数のホストを保持しているユーザが分かる可能性がある。図 4.7 に例を挙げる。これは、二つの端末が全く同じ時間繰り返し検出されている。完全に一致していないが、ほぼ同じである。そして、このように同じよう履歴が数日間に渡って検出した。このことから、この二つの端末は関連していると言える。実際に、これら二つの端末所有者を調査したところ、同じユーザが複数所有していたため、端末間の関係を取得することは有用であることが分かる。これによって、200 台以上のホストが存在するネットワークにおいて、

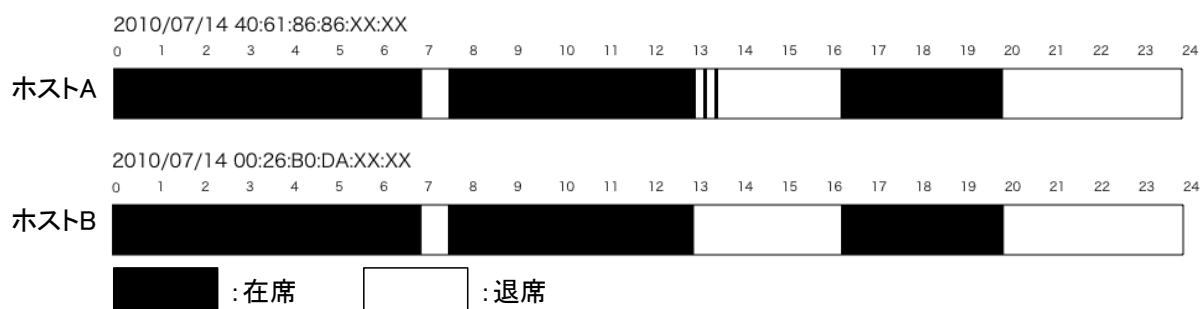


図 4.7: 端末間の関係の表示

一人のユーザが複数のホストを保持していると推測できるホスト 3 組を発見できた。

事前実験によって、今まで単体では個人情報と認識されていない情報と外部情報を組み合わせることで個人が所有するホストや、所有者自身のデータを取得することができた。特に、ホスト情報を事前に取得していない場合でも、対象のユーザがネットワーク上にいたという情報からホストの推測できる可能性がある。また、他にも、一人が保持している複数のホストを特定できることを示した。一度ホストに関する情報を取得された場合、今後永続的に追跡される可能性がある。そのため、MAC アドレスや、サービス探索情報などを取得され、組み合わせることで個人情報となり得る。

しかし、本システムでは、常駐ホストや、ホスト所有者がホストを放置されたまま退席した場合、研究室にユーザがいるにも関わらず実験ネットワーク以外のネットワークにアクセスした場合は検出することができず、ホストを識別できていない。また、端末間関係において、一人で複数のホストを保持しているユーザは多く確認されているが、3 組のみしか推測できなかった。その原因としては、筆者が所属するネットワークで、複数台のホストを保持するユーザが極端に少ないため、常に一方のホストを稼働しているケースや、モバイルデバイスを Wi-Fi に必ず接続するわけではない事が挙げられる。そのため、ある程度は似たような挙動になる場合があるが、必ずしも同じように連動する限りではないことが分かる。

4.3.3 登録情報の変化検証

約二年間事前実験を行い、対象 25 名のユーザの MAC アドレスが変更されたかを検証した。実験の結果、自ら MAC アドレスを変更したユーザはいなかったが、途中からユーザが在席するにもかかわらず検知できないケースがあった。これは、ホストを買い換えなどの結果、利用ホストの MAC アドレスが変化したためである。実験期間中に MAC アドレスが変化したユーザは 25 名中 7 名であった。このことから、MAC アドレスは長期的に識別要素として利用できるが、必ずしも常に同じものは無いが、他の 18 名のユーザは 1 年間常に同じホストを利用し続けていることがわかった。

4.4 提案手法

第 4.3 で述べたように MAC アドレスから多くの情報を収集することができることを示した。そこで、本研究は機器情報を用いたホスト識別手法を提案する。また、事前実験の対象は一つのネットワークであったため、複数のセグメントやネットワーク環境に対応する必要がある。

本提案手法は、ネットワークに参加するホストに本提案手法のプログラムをインストールし、センサとなることで、広範囲に MAC アドレス情報を収集、統合することを目的とする。ホストが発信するブロードキャスト情報、マルチキャスト情報に着目し、同じネットワークに接続するユーザの MAC アドレスを収集する。収集するクライアントは本提案手法によるプログラムをインストールしたホストであり、そのホストの数が増加するほど、情報収集ネットワークの範囲は広域となる。この MAC アドレス情報にユーザ情報や SNS といったユーザの利用するサービス情報、場所情報や接続 AP 情報を組み合わせることによって、移動するデバイスと利用者を含む関連する情報を包括的かつリアルタイムに管理するプラットフォームの構築が可能となる。集約された情報を各クライアントが共有することで、任意のネットワークにどのようなユーザがいるかを、その場にいたくともユーザは取得可能となる。図 4.8 に本研究の概要図を示す。

まず、ネットワーク α において本提案手法を用いたユーザ A が、周囲の情報を収集する。その結果、デバイス Z がネットワーク α にいることが分る。この情報をサーバに送信する。同じように各ネットワーク β , γ でも同じように各センサが収集した情報を統合し、参照できるようにすることで、どのホストがどこにいるのかを把握することができる。

4.5 設計及び実装

4.5.1 要件定義

必要になる要件と、それを満たすための機能を以下にまとめる。まずはクライアントの要件について述べる。

- **パケットの取得**

取得する情報は、ネットワークセグメント全体に流れているブロードキャスト情報とマルチキャスト情報である。

- **アプリケーションの汎用性**

本手法はネットワークに接続できるあらゆるホストがセンサとなり設置されることが望ましい。様々な OS 上で動作が可能である必要がある。

- **データの送信**

収集した情報を、サーバに送信する。サーバとの通信ができない場合は、情報を蓄積して、通信が可能となった場合にデータを送信できる機構が必要である。

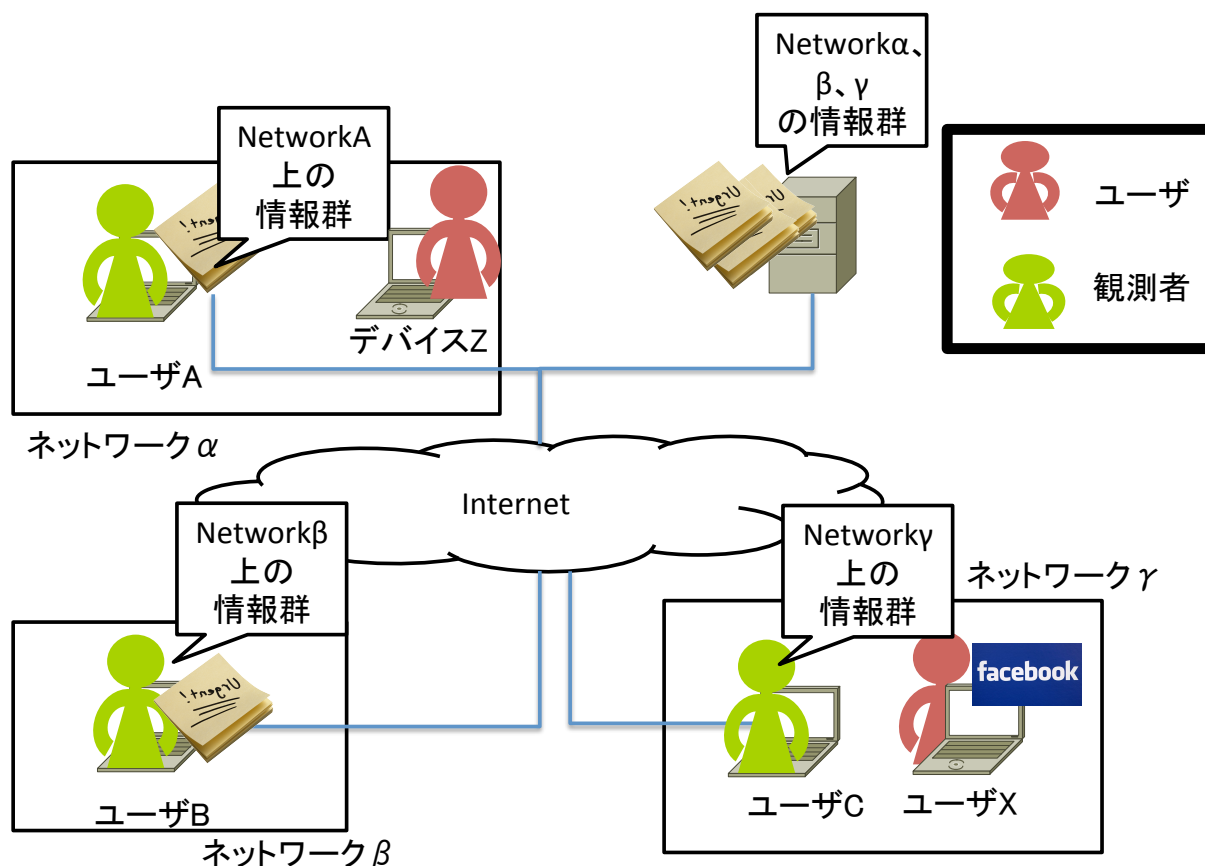


図 4.8: システム概要図

- ホストの認証

取得した情報を閲覧する場合、ホストの認証をした上で、ホストに応じた情報を提示する機構が求められる。そのため、ホストの認証機能が必要である。

次にサーバの要件について述べる。

- データ共有

収集した情報をユーザ側に共有する。収集時は各 MAC アドレスからどの情報が送られてきたかも同時に記録する。

- ユーザ権限

収集したデータをユーザに向けて開示する場合に、すべてのユーザ情報を開示するのではなく、お互いに許可されたユーザ同士で情報を共有できるようにする。

4.5.2 クライアント設計, 実装

パケットを取得する際にネットワークに流れるブロードキャスト情報及びマルチキャスト情報を収集する。同時に、ブロードキャスト情報には MAC アドレスだけではなくホス

表 4.2: クライアント開発環境

開発言語	対応 OS
Visual C#	Windows Vista, 7
Python2.7	MacOSX10.5, 10.6, Ubuntu OS10.10, 11.04, Debian6

ト名が収集可能である。例えば、DHCP サーバがネットワークにブロードキャストする特定の情報に、ネットワークに参加しているホストのホスト名が含まれることがある。そのため、DHCP サーバから情報を収集する際に、ホスト名も同時取得する。他にも、Samba, mDNS プロトコルからもホスト名の取得が可能であるため取得する。

公共無線 LAN の一つである BB モバイルポイント [34] においては、セグメントのブロードキャスト情報やマルチキャスト情報はプライベート VLAN がネットワークで利用されている。そのため、ブロードキャスト、マルチキャストは接続ホストには届かないように設定されている。しかし、ホストが接続時に DHCP サーバとの通信の過程で、DHCP サーバに接続を試みた MAC アドレス情報をネットワーク全体のホストに送信する性質がある。この性質を利用することによって、BB モバイルポイントに接続しているホストの情報を収集することが可能である。

クライアントがどのネットワークに接続をしているか、ネットワークの位置及び、場所の名前は重要な識別要素となる。現在クライアントが収集している情報はどこなのかを調べるにあたって、BSSID 及びネットワークの gateway の MAC アドレスを収集する。そして、ユーザは可能であれば場所に関しての情報を登録する。

MAC アドレスが誰の持ち主かが判明した場合、ユーザはこの MAC アドレスがだれのものか登録することができる機構が必要である。実際に、利用しているホストが Windows なのか、MacOSX なのか OS をホストから大まかな判断は可能であり、それをもとにホストを識別できる場合がある。

クライアント開発環境を表 4.3 に示す。クライアントは、第 4.5.2 節で述べたように、様々な OS でインストールされることが望ましいため、Windows XP, 7, MacOSX10.5, 10.6, Ubuntu OS10, 11, Debian6 の 8 種類の OS に対応した。Windows 版では C# で記載し、MacOSX や Linux は python で記載した。これらの開発環境で実装したクライアントを Windows 版は図 4.9 に、MacOSX 版は図 4.10 に示す。

表示する情報は、取得時間、ホスト名、ユーザ名、MAC アドレス、ベンダーコードである。また、同じネットワークセグメントに接続しているクライアントが判別できることで第 4.5.3 節述べる情報共有が容易になるため、クライアント同士がお互いに判別、コミュニケーションできる実装にした。

次に、MAC アドレスを登録するインターフェースを図 4.11 に示す。このインターフェースを利用して、対象の MAC アドレスの所有者が判明すれば、登録できるように設計した。これによって、ユーザの名前や、SNS 情報を結びつけることができる。

このクライアントを複数のホストにインストールし、情報を収集することで、クライアントが接続するネットワークの状況を知ることができる。これらのデータはサーバに送信さ

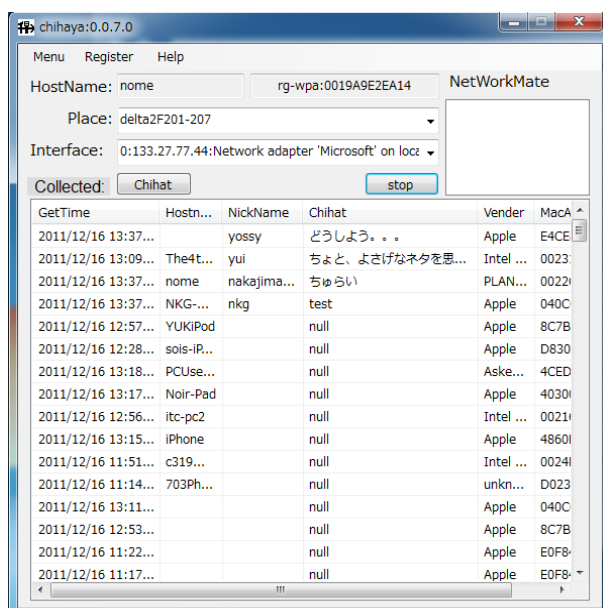


図 4.9: Windows 版クライアント

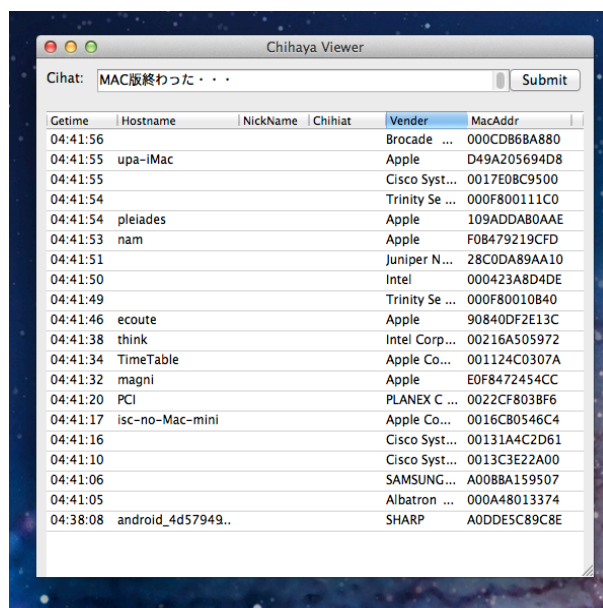


図 4.10: MacOSX 版クライアント



図 4.11: MAC アドレス登録インターフェース

れて解析し、その後クライアントに解析情報をフィードバックされる。

ホストの認証について述べる。ホストの認証に MAC アドレスを用いるのは、なりすましの可能性があるためふさわしくない。そこで、配布アプリケーションにシリアルキーを付与する。このシリアルキーとホストの登録した MAC アドレスを用いることでホストの認証を行う。このように、本提案手法では、アプリケーションインストール時にシリアルキーを保持し、認証で利用できるように設定される機構を保持している。

表 4.3: サーバ開発環境

OS	開発データベース	開発言語
Ubuntu10	PostgreSQL9	PHP5

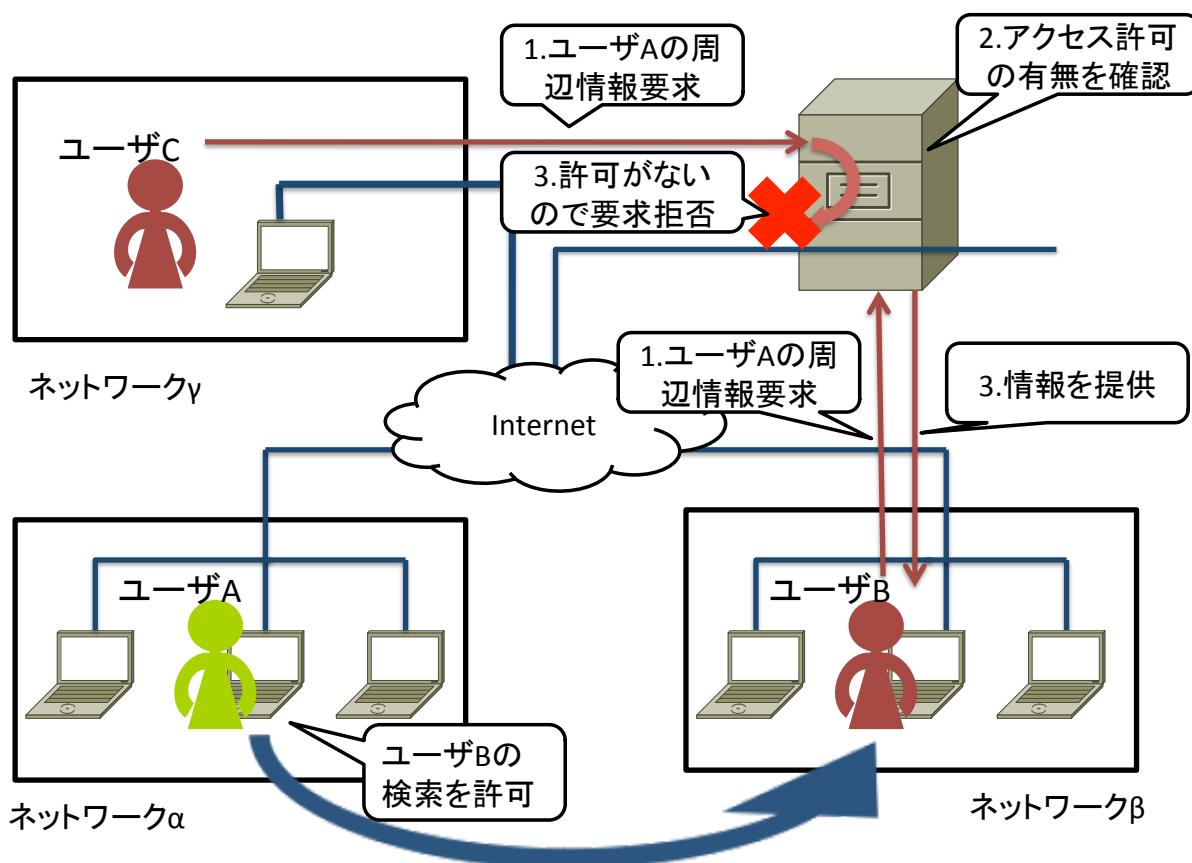


図 4.12: アクセス制限機構

4.5.3 サーバ設計, 実装

サーバ側の設計と実装について述べる。

サーバのデータ共有はクライアントが情報を収集し、サーバがクライアントに対して情報をフィードバックするサーバクライアントモデルとする。

サーバに送られた情報をクライアントが参照する際において、すべてのユーザが収集した情報を閲覧することは個人のプライバシーが脅かされる懸念がある。そこで、ユーザごとにアクセス制御の機構が必要である。図 4.12 に情報提供のアクセス制御についての概要を示す。

ユーザの情報制御のアクセス制限について述べる。まず、ユーザ B はネットワーク α の情報を取得したい場合、ユーザ A に検索の許可を得る。それに対してユーザ A はユー

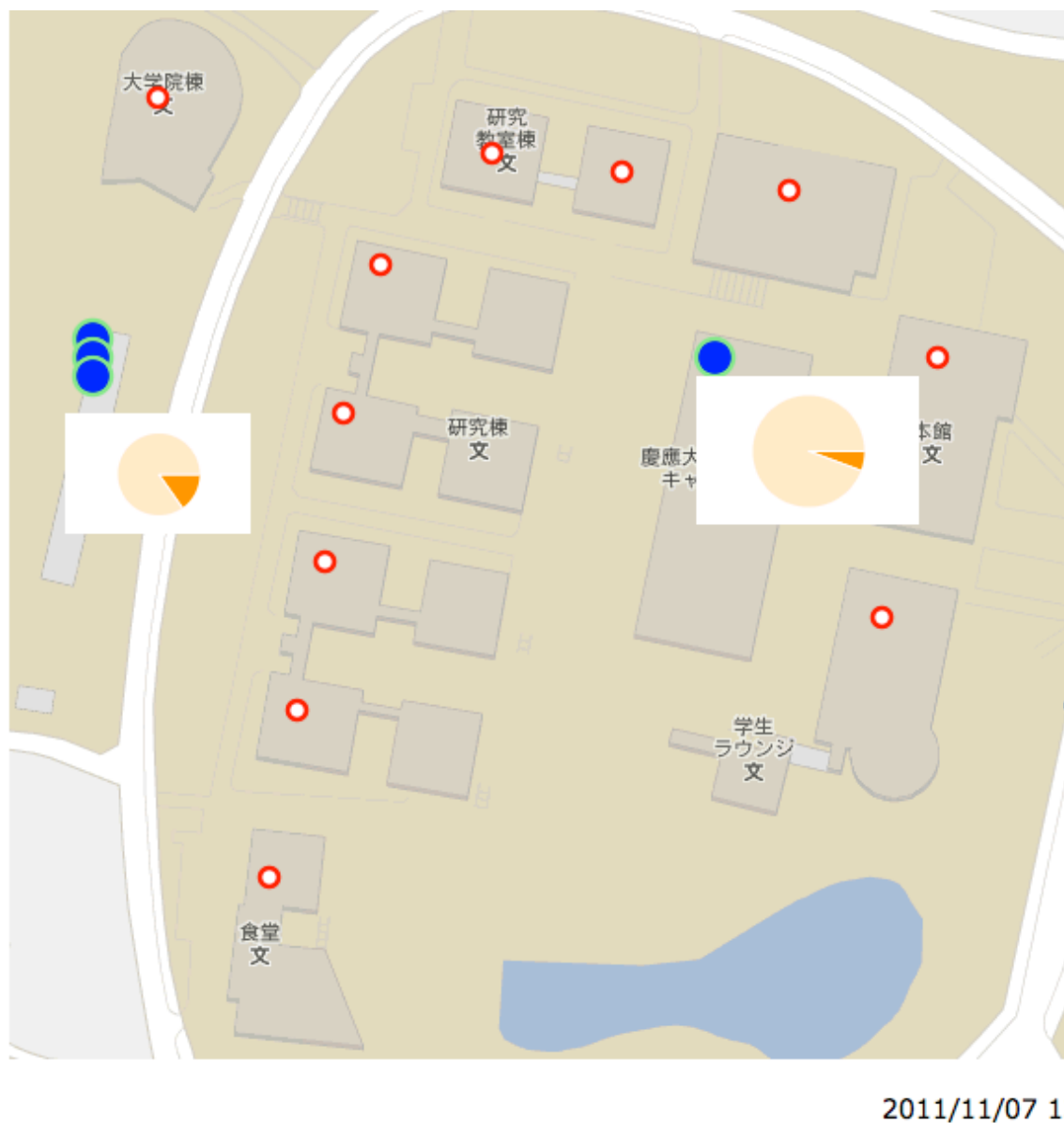


図 4.13: 収集状況の表示

ザ B に対して検索の期間や対象情報の許可を出す。その後、ユーザ B はサーバに情報を要求しネットワーク α のデータを取得する。情報を提供するユーザ A は提供時にどの情報をどこまで提供できる公開範囲を選ぶ事ができると共に、許可に対して有効期限及び、許可のユーザを設定できるようにする。これによって、ユーザ A の許可を得ていないユーザは対象ネットワーク情報を収集することができない。

サーバで収集した情報の表示例を図 4.14 に示す。この図は、実験期間のある一期間を表示したものであり、青い丸はセンサとなるクライアントである。そして、オレンジ色の円は取得した MAC アドレスである。その円グラフにおいて濃いオレンジはユーザ名が



図 4.14: ユーザの位置情報表示画面

表 4.4: 実験環境

実験開始	実験終了	実験参加人数	実験ネットワーク数
2011 年 12 月 2 日	2011 年 12 月 19 日	32 名	16

わかっている MAC アドレスであり、薄いオレンジはユーザ名がわからない MAC アドレスの割合である。この円グラフは取得する MAC アドレスの数に応じて大きさが変わる。ユーザは自身が収集した情報、あるいは許諾を得たユーザの情報を閲覧することができる。そして、どのホストがどこにいるのかを知ることができる。また、ユーザは対象の場所を表示することで図 4.14 に示すように誰がいるかを知ることができる。

4.6 実験

本提案手法によって目的が実現可能かを検証するため実験する。

4.6.1 実験環境

本提案手法の実験環境を表 4.4 に示す。

本提案手法を検証するためには、様々なネットワークでホストを識別できる必要がある。そこで、約数千人のユーザが利用するネットワークを実験の対象とする。実験対象ネットワークは 16 のセグメントに別れており、対象ユーザの識別ができるかを検証する。どのセグメントを利用しているかを取得することで、対象ホストのおおよその位置が把握できる。実験対象ネットワークは様々な場所に応用可能である。例えば、企業のネットワーク

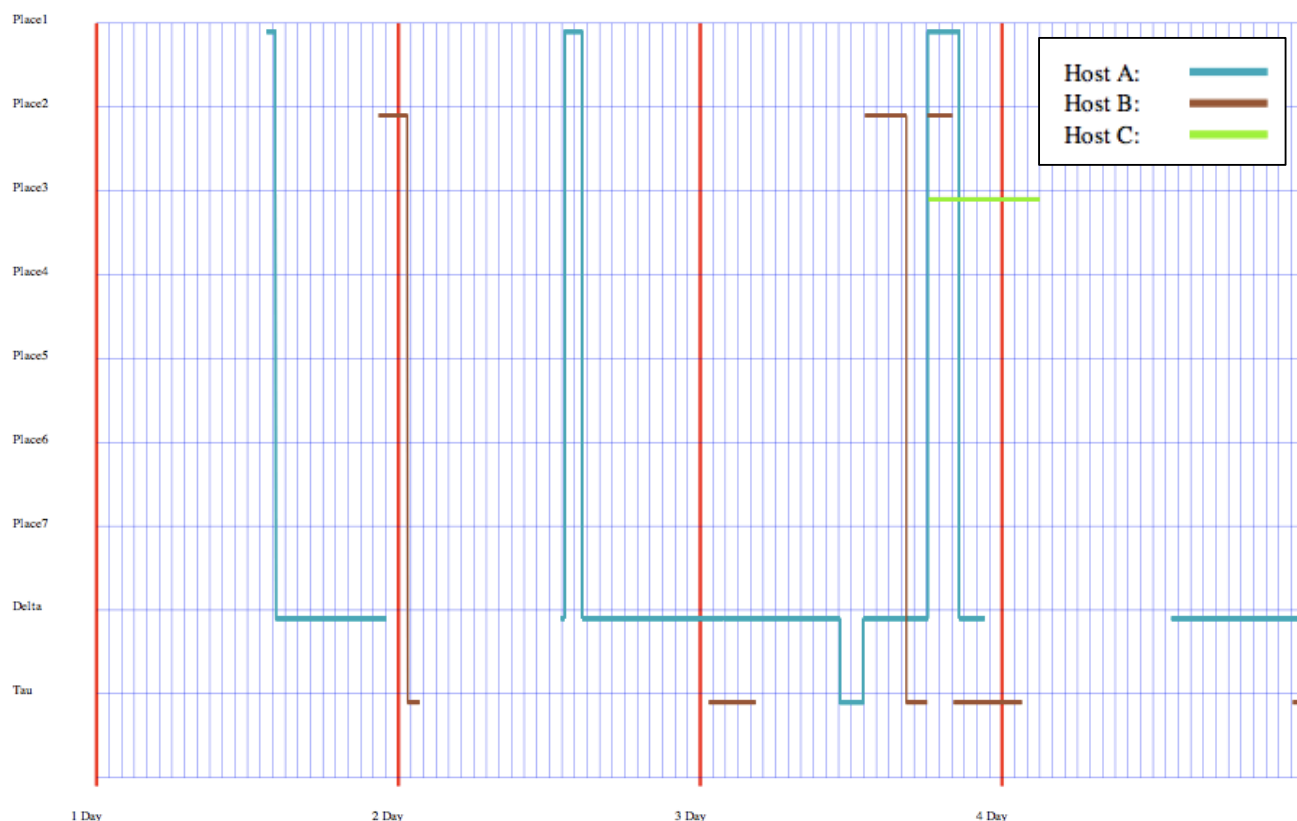


図 4.15: クライアントのネットワーク遷移

において部署ごとにセグメントが分かれていることが多くある。そのため、この実験ネットワークでホストの識別が可能であれば、他の場合にも転用可能である。

本手法をインストールした実験参加者のホストは 32 名である。検証期間は 2011 年 12 月 2 日から 2011 年 12 月 19 日の 17 日である。本手法によって、対象ユーザをどの程度追跡できるかを検証する。

追跡対象者は対象のネットワーク郡を日々利用しているユーザを 3 人選び、実験協力の同意を得た上で実験した。そして、実験協力者にどの AP やネットワーク接続しているかを調査し、記録するツールをインストールしてもらい実験した。

4.6.2 実験結果

本提案手法の結果について述べる。まず、本実験が想定しているネットワークにおいて、合計 9 つのセグメントが存在する。そのセグメントにセンサを配置し、3 台のホストがどこにいたのか検知した結果を図 4.13 に示す。

Y 軸のラベルは場所のラベルである。Place は該当ネットワークが同一の場所であるが、セグメントが 7 つに分かれているため、Place1 から Place7 とセグメントごとに番号を振っ

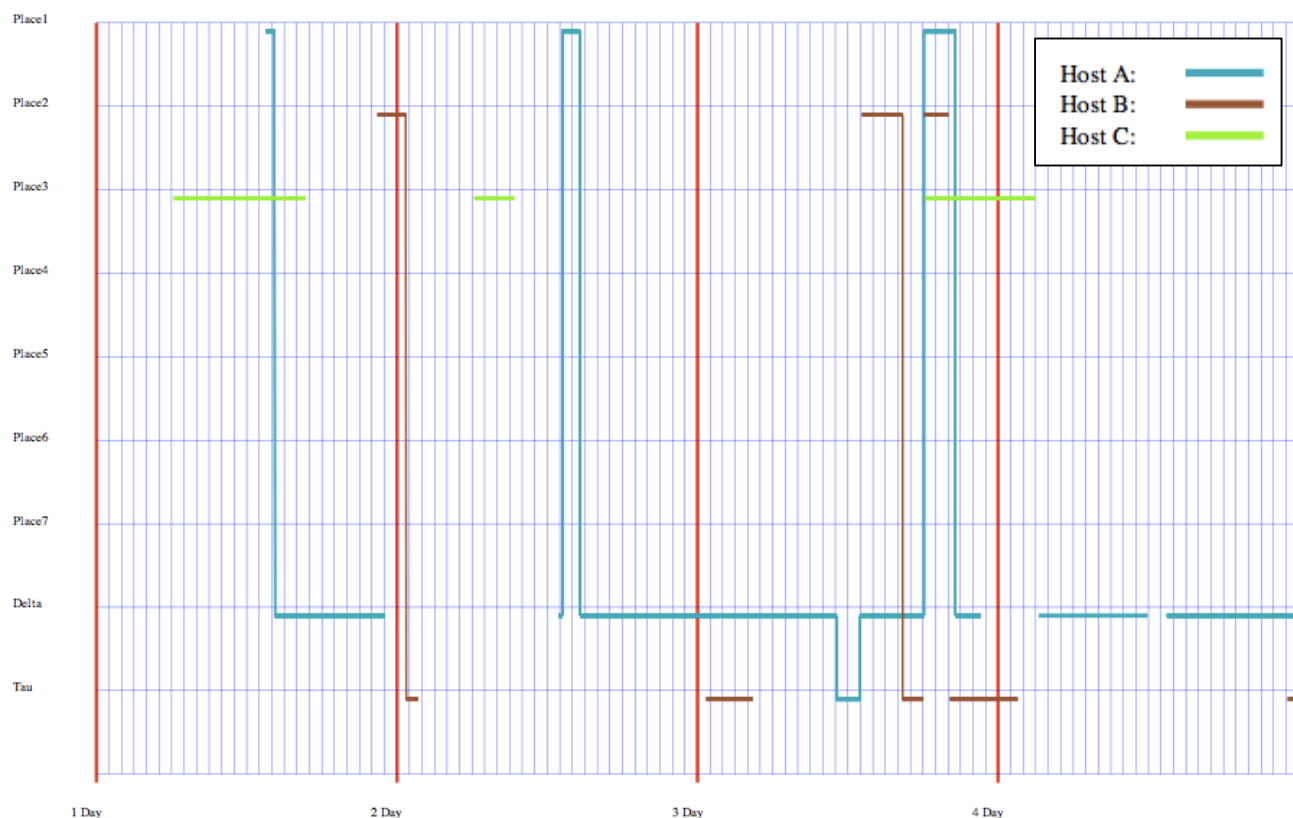


図 4.16: 実際のクライアントのネットワーク遷移履歴

ている。そして、別の 2 ヶ所のネットワークの計 9 つのネットワークで実験をした。X 軸は実験開始日の 2011 年 12 月 9 日からの経過日数を示している。また、各ホストがネットワークを移動する際に、15 分以内に他のネットワークに移動した場合は縦線が付与される。この図からホスト A、ホスト B に関してはかなり高い割合でホストの識別と追跡ができていたことが分かった。同時に、ホストの利用ネットワーク推移まで把握することができた。しかし、ホスト C に関して多くの情報を収集できなかった。

収集した情報の正しさを把握して比較する必要がある。そこで、識別ホスト A、B、C にどのネットワークに接続したかを記録するアプリケーションをインストールし情報を収集した。図 4.16 に各ホストの実際の動きを示す。

図 4.16 の各ホストのネットワークの利用履歴と、図 4.15 の本提案手法の結果と比較する。比較した結果、ホストの挙動と本提案手法で取得した結果がどの程度合致していたかを表 4.5 に示す。比較した結果、ホスト B に関しては完全に実際の動きと実験結果が同じであり、対象ネットワークにおいて、ホスト B を正確に検知できることがわかった。次に、ホスト A に関しては完全に情報を収集することができなかった。これは 4 日目にクライアントが存在しなかったため、検知できなかったと考えられる。同時に他のネットワークにおいても、3 日目の真夜中から朝にかけては検知率が低下していた。ホスト C に関しては情報があまり収集できていない結果となった。原因としては、ホスト C が存在する Place3

表 4.5: ホスト識別の実験結果

ホスト	一致率
ホスト A	86%
ホスト B	100%
ホスト C	37%

は本提案手法を用いたクライアントが存在しないことが原因であり、実際にクライアントは Place3 に関しては 1 台のみしか接続していなかったためである。

また、日中は 6 割以上のネットワークセグメントの情報を収集可能であったが、夜間は 3 割程度しか情報を収集することができなかった。この原因はセンサとなるクライアントは夜間や利用していない場合は電源を切るためであった。他にも、曜日による収集率の低下や、対象ネットワークのイベントなどによっても大きく変化することが分かった。

4.7 考察

本手法の情報の収集は本手法をインストールしたクライアントに依存しているため、ホスト識別が可能な場合がネットワークや状況に応じて変化する問題がある。しかし、クライアントが多く集まる場所は該当するネットワークのホスト収集の精度が上がるため、人の移動を把握することができることが分かった。この特性を利用することによって、イベント会場におけるホットスポットな場所や人が多く集まる場所をクライアントの数や精度の正確さを取得し解析することで場の雰囲気や人の流れを把握することが期待できる。

本手法はクライアントが本提案手法を用いて情報を収集することが利点である反面、対象のネットワークにホストが存在しない場合は情報が収集できないという問題が存在する。そのため、企業においてホストの監視など常にネットワークの状況を取得する場合は、各セグメントにセンサを設置するなどの対応する必要がある。また、管理者が利用する用途の場合は、クライアントが対象のセグメントに何台接続しているか常に管理する手法を検討しなければならない。

4.8 まとめ

本章ではホストの識別要素に MAC アドレスに着目した。そして、各ホストがクライアントとなりネットワークに存在する MAC アドレスを収集することによって、ホストを識別する手法を提案し実装した。実験の結果、識別対象のホスト 3 台のうち、1 台のネットワークの利用推移を完全に把握でき、1 台は 8 割以上の利用推移を把握することができた。しかし、残り 1 台は十分に利用推移を把握することができなかったという結果になった。このことから、情報の取得はセンサとなるクライアントの設置条件に左右されるが、目的に応じてセンサを設置することで、情報の収集が可能であることを示した。

第5章 パケットヘッダ情報を用いたホスト識別手法

5.1 手法概要

従来、機器の識別には IP アドレス、MAC アドレスや Cookie などの識別子が利用されている。しかし、IP アドレスは可変であり、Cookie などを利用したホストの識別においても、ホストが Cookie を受け入れない場合や、ユーザが Cookie を削除することができるため、継続的にホストを追跡することが難しい。また、第 4 節で述べた手法は、各セグメントにセンサを配置する必要があるが、対象セグメントやネットワークが増加するたびにセンサを配置しなければならないという問題点がある。他にも機器の識別子として利用されている MAC アドレスは詐称が容易に可能である。さらに、近年普及しているノート PC などは有線や無線といった一つの機器に複数の MAC アドレスが付与されている場合が多い。そのため、有線と無線のインターフェースを持つ 1 台の機器を複数の機器として判断してしまう問題がある。これらの問題に対応するためには、IP アドレスや MAC アドレス、Cookie などの識別子以外で機器を特定できる方法が必要である。

本論文ではユーザが普段インターネット上に発信しているパケット情報を、ネットワーク管理者が観測し、組み合わせることでホストやユーザを特定する手法を提案する。本提案手法で利用する情報は、普段ホストが送信しているパケットヘッダ情報である。2010 年度総務省において Deep Packet Inspection によるターゲッティング広告の技術が取り上げられた際 [35] に、ペイロードを取得することについて議論が巻き起こっている。そのため、ペイロードは取得せずにパケットヘッダ情報のみ取得することで、この制約を外すことや、ペイロードの取得に比べてユーザの同意が得ることが容易となる。本論文で述べるパケットヘッダ情報とは、IP、TCP、UDP、ICMP ヘッダの事を指す。そして、ネットワーク管理者は、これらの情報を対象の組織内 LAN 内などの管理ネットワークで取得し、本手法を用いる事を想定している。これによってユーザは意識することなくサービスを受けることができ、サービス提供側やネットワーク管理者はユーザに操作を要求することなく、より正確に管理ネットワークに参加している機器やユーザ情報を得ることができる。

ネットワーク管理者は、管理しているネットワークにおけるユーザの発信するパケットの情報を容易に取得できる。一般的なネットワークでは、セキュリティ対策やネットワークの運用の関係上、管理ネットワークと外部のネットワークとの中継点でトラフィックを監視できる場合が多い。しかし、ペイロードを取得する場合は電気通信事業者法などの観点からユーザ全員の同意を得なければならない。そこで、ペイロードを除いたパケットのヘッダ情報からホストを特定する手法を提案する。

表 5.1: 事前に作成したホストのプロファイルデータ

項目	ホスト A 候補数	ホスト B 候補数	ホスト C 候補
接続したメールサーバ群	62	2	32
接続した SSH サーバ	13	0	16
OS	20	22	18
特徴的な挙動	0	23	0
接続した IRC	0	0	0
起動時のアプリケーション挙動	18	0	18
アプリケーションアップデートの通信	2	0	0

5.2 事前実験

特定のユーザが情報を利用するデバイスにはユーザの利用形態によって発信する情報に特徴的なトラフィックパターンが現れる可能性が高い。そのトラフィックパターンを取得、解析することによって、デバイスの追跡が可能かどうかの事前実験を行った。ユーザが情報を利用するデバイスの特徴は、対象のネットワークによって異なることが推測できる。そのため、今回の事前実験においては、200 名程度が在籍する筆者の所属する研究室の学術ネットワークを対象とする。このネットワークは外部と通信を行う際に、必ず本手法を適用した機器を通過する構成であり、IP アドレスは DHCP によって割り振られた IP アドレスと固定 IP アドレスが混在し、ローカル IP アドレスの重複や NAT 構成がされていないネットワークである。その環境下で、無作為に選出したホスト 3 台の通信を観測した。そして、実験で取得した各項目を利用することによって、どの程度ホストを絞り込むことができるかを検証した。結果を表 5.1 に示す。

本ネットワークにおける事前実験において、発生頻度が高く取得が容易な通信と、発生頻度が低い取得すると絞り込みが容易になる通信の二種類に分類できることがわかる。まず、発生頻度の高い識別要素にメールサーバの接続先情報が挙げられる。メールサービスは多くのユーザが利用し、複数メールサーバに接続しているホストは容易に絞り込みが可能であることが分かった。ホスト候補 B、C は複数のメールサーバにアクセスしており、メールサービスの利用及び接続先サーバを識別要素とすることで、ホスト B は 2 台まで、ホスト C は 32 台まで絞り込むことができた。そのため、取得が容易かつ判別能力は高いと言える。次に、リモートログインもコミュニティによっては比較的使われやすいサービスである。取得の容易さはメールサーバ接続情報と比較した場合には劣るが、容易に取得でき、ホスト毎に差異が出やすい事が分かった。ホスト A、C は SSH を利用しており、メールと同じく接続先サーバを識別要素とすることで、ホスト A は 13 台、ホスト C は 16 台まで絞ることができた。このことから、メールよりも発生頻度は低い、ホストの分類は容易にできるため、識別要素として利用できることが判明した。そして、OS 情報は Windows, MacOSX, Ubuntu であれば取得可能であるが、OS 情報はホストの絞り

込みを十分にできなかった。そのため、メールやリモートログインと組み合わせて利用することでホスト判別基準の一つとして利用することが望ましいことが分かった。一方で、発生頻度が低い、取得できると判別が容易な情報にホストのソフトウェアアップデートや特徴的な挙動が挙げられる。これらは取得することができれば絞り込みを容易にするが、アップデートであればポーリング、特徴的な挙動であればユーザがアプリケーションを動かす必要があるなど、発生頻度が低い。そのため、各項目における優先順位は低く設定する。同様に、現在利用人口が減っている IRC も優先順位を低く設定する。しかし、予備実験において、3 台の対象ホストは IRC を利用していなかったが、ネットワークにおいて IRC を利用するホストは 6 台存在したため、もしホストが IRC 利用している場合は強力な識別要素となりえることが判明した。また、起動時のアプリケーション挙動はアップデートと比べて取得は容易であるが、対象ネットワークに電源を切った状態で起動をする必要があるため、発生頻度は高いとは言えない。

事前実験において、前述した識別要素を利用し、学術ネットワークでホストを識別した場合、200 名が参加するネットワークから、各 3 台のホストを 1 台に絞り込むことができた。このことから、事前実験を行ったネットワークにおいて、特定のユーザが情報を利用するデバイスは、ユーザの利用形態の特徴によって発信する情報に特徴的なトラフィックパターンが現れる可能性が高いことが分かった。同時に、事前実験で利用した情報の組み合わせの順番を、取得容易かつホストを判別できる要素が高い順に並べると、1. メールサーバ接続先情報、2.SSH サーバ接続先情報、3.OS 情報、4. 起動時のアプリケーション挙動、接続した IRC サーバ、アップデート情報、特徴的な挙動であった。これらの順番でホストを絞り込むことでホストの特定が可能である。また、ネットワークによって特定のユーザが情報を利用するデバイスは、ユーザの利用形態によって発信する特徴的なトラフィックパターンが変化する可能性が高いため、ISP などのユーザが異なるネットワークで利用する場合は、識別要素とする情報を再度検討する必要がある。

5.3 判別手法

第 5.2 節において、各ホストは利用状況に応じて、それぞれ特徴的なパケットを送受信していることが分かった。そこで、パケットのヘッダ情報からは、送信先、発信元 IP アドレス、ポート番号、発信タイミング、利用サービスなどの情報を取得、組み合わせることで、プロファイルを作成しホストを識別する手法を提案する。また、識別要素として取得する情報は、ネットワークやホストによって異なるが、今回は事前検証で挙げた大学や企業に所属するユーザが多く参加するネットワークでの識別要素について述べる。

図 5.1 にパケットヘッダ情報からホストを特定する手法の概要を示す。ネットワークトラフィックから、各ホストの振る舞いに関するデータを収集、蓄積する。本手法では、ホストの振る舞いにより蓄積された情報をプロファイルと定義する。プロファイルとして蓄積される情報は、事前検証の情報と、ホストを継続して追跡するためにネットワーク接続時の状態情報を加えた以下の 5 つで識別する。

- OS の種類:発信元ポート番号

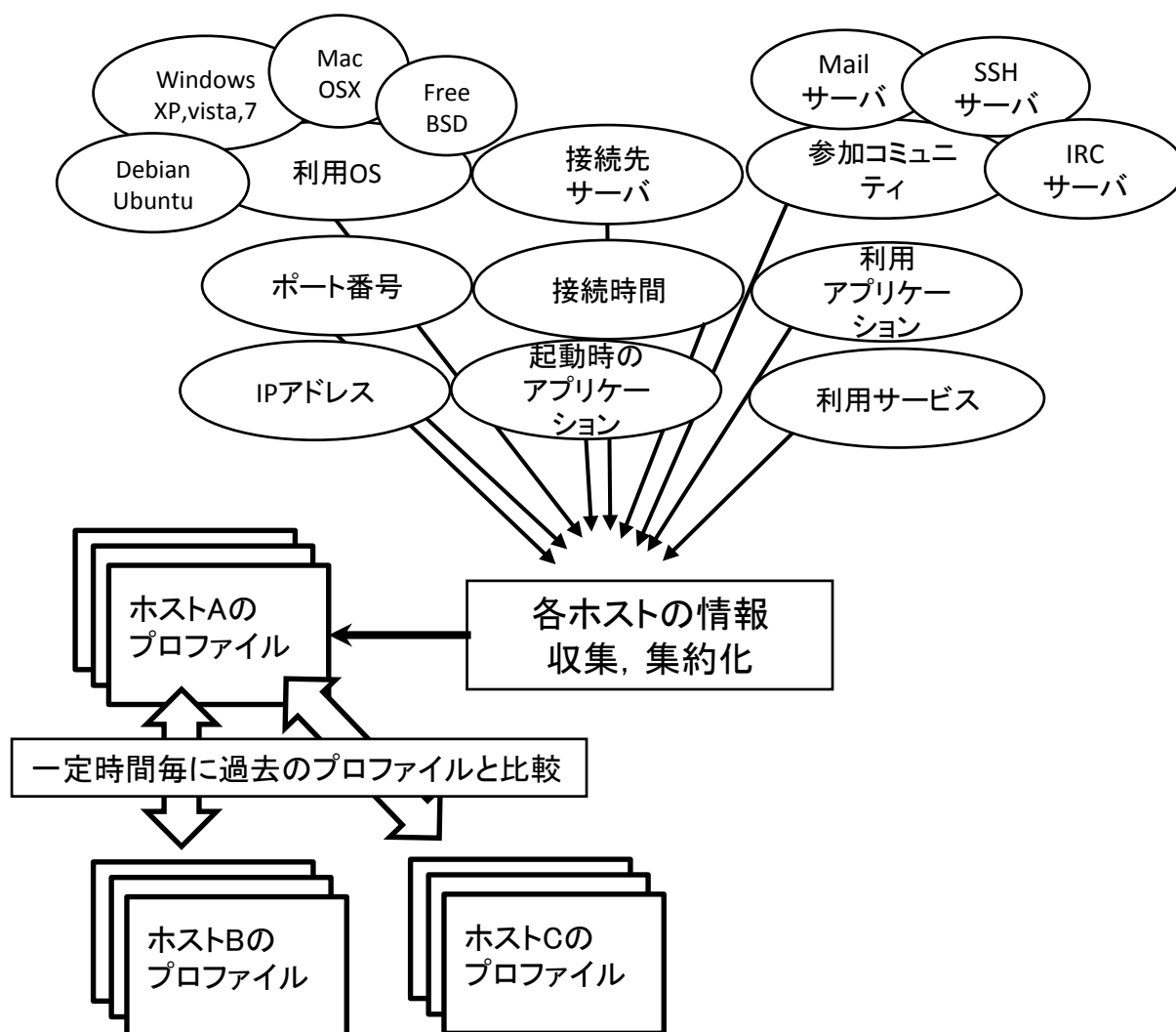


図 5.1: パケットヘッダ情報によるプロフィール手法

発信元ポート番号から、OS の推測や一度ネットワークを離脱したホストを検索することができ、これらはユーザ毎に異なるという特徴があるため識別要素として取得する。サスペンドから復旧したホストは、サスペンド前に利用されていた発信元ポートの数個前のポート番号を利用する傾向があり、ホストの利用している最後の発信元ポート番号を保存することによって、再度追跡が可能である。また、発信元ポート番号から OS 情報を取得することもできる。利用する発信元ポート番号は OS ごとに異なり、Windows XP や Ubuntu は特徴があり容易に識別できる。この技術は Passive Finger Printing[36]でも利用されている。また、OS アップデートの際の通信を取得することによって、MacOSX や CentOS などの他 OS を判別することも可能である。

● 利用サービス、アプリケーション

ユーザが利用するサービスやプロトコルも重要な識別要素となる。前提として、

サービスを提供するサーバのアドレスブロックを取得しているとし、ホストが通信したサーバの IP アドレスからホストの利用しているサービスを特定する。そして、特定のプロトコル利用時の送信先 IP アドレスとポート番号の組み合わせも利用する。今回のネットワークにおいて、識別要素とするプロトコルは SSH, IMAP, POP, VPN, IRC とする。他にも定期的な通信や、普段一般ユーザにとって利用されない通信ポートも十分な識別要素になり得る。

● ホストの所属するコミュニティ

利用サービス、アプリケーション利用時の接続先サーバも識別要素として利用できる。企業や大学に所属するユーザは、その組織が提供するサーバを利用するため、ユーザが利用したリモートログインサーバやメールサーバを基準とすることで、ホスト利用者の所属する組織、団体ごとにホスト群が作成できる。本論文では接続先サーバ毎に分類したホスト群をホストの所属するコミュニティと定義する。それらの情報を取得し、ホストのコミュニティを識別要素とすることでホストを判別する。また、接続先 IP アドレスから位置情報や所属組織を類推するサービス [37] などと併用することでより正確にホスト利用者の所属組織を分類できる。

● 自動起動に登録されているアプリケーション

ホスト起動時にアプリケーションが立ち上がる動作もホストの識別基準となる。自動起動に設定されている場合、OS の起動と共にアプリケーションが特定の通信を行うため判断することができる。実際にホストが起動する際に発信するタイミングを計測したところ、IP アドレスが付与されてから約 60 秒後に事前に登録してあるアプリケーションが起動していたため、60 秒後の挙動に着目した。60 秒から 120 秒の間にパケットの発信があった場合、なんらかのアプリケーションが自動起動に設定されている可能性がある。その際の接続先サーバから自動起動に設定されているアプリケーションを推測し、ホストの識別要素とする。

● ネットワーク接続時の状態情報

ネットワークに接続したホストが電源を切った状態からの起動か、サスペンドからの復旧かを判定することはホストの特定において重要である。前述した自動起動に登録されているアプリケーションや、発信元ポート番号の情報を識別要素として利用する場合、ホストが起動か、サスペンドからの復旧かによって、利用情報が変化するためである。起動か復旧かの判別に IP アドレスが付与された時間からパケット発信のタイミングを計測する手法が挙げられる。図 5.2 はクリーンインストールした MacOSX の IP アドレスが付与されてからのホストの起動時と復旧時のパケットの送信タイミングの比較である。X 軸は IP アドレスが付与されてからの経過時間であり、Y 軸は起動時及び復旧時のホストの挙動のラベルである。この図から復旧時は、電源が落ちている状態からの起動時に比べて、IP アドレスが付与されてからすぐに TCP を始めとしたパケットを送信している。つまり、サスペンドはアプリケーションや OS に必要な機能を立ち上げている状態で行われているため、起動して数秒で異なる挙動をした場合、サスペンドからの復旧と判断することができる。

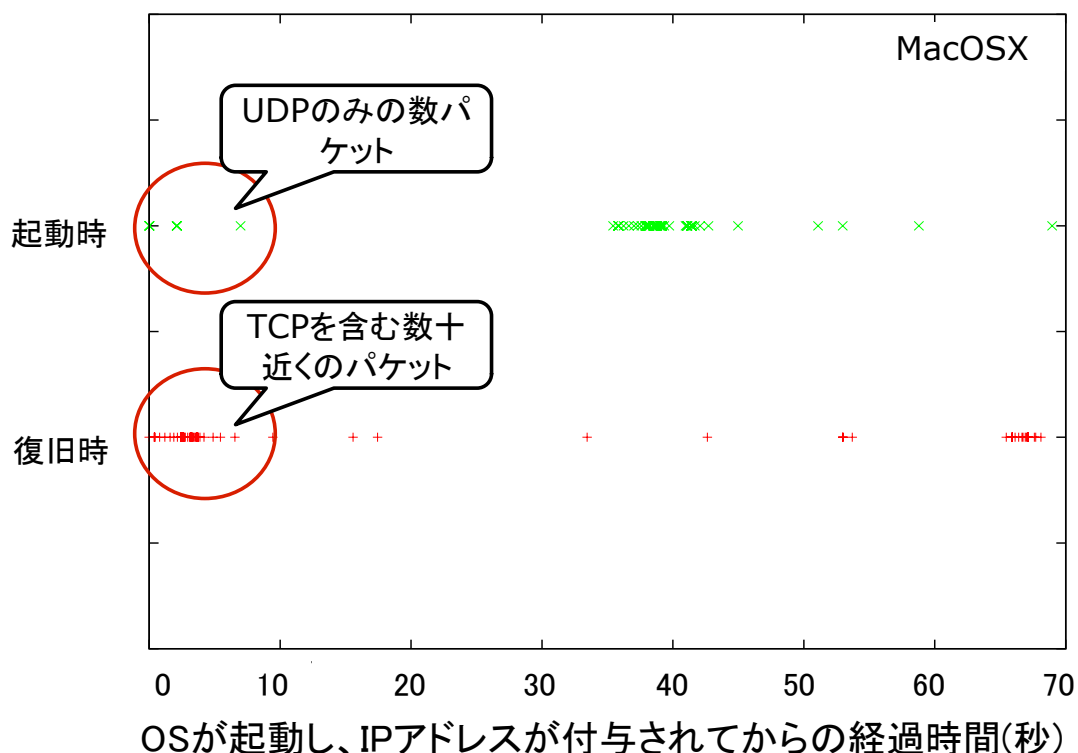


図 5.2: 起動時と復旧時におけるパケットの発信タイミングの比較

また、Windows, Ubuntu, CentOS も同様に検証したところ、接続時の復旧情報はこれら 3 つの OS にも利用可能であることが判明した。

これら 5 つの情報をもとに、ホスト毎にプロファイルを作成する。ホストの特定は蓄積されたプロファイルをもとに行うため、すべてのホストは最初の 30 分間は識別せずに前述した利用情報を収集する。そして、30 分経過毎に過去のプロファイルと一致するかを比較する。プロファイルが一致した場合は、該当ホストは過去接続したことがあるホストであるとして、プロファイルデータを統合する。このように、30 分間プロファイルを作成し、その後過去のプロファイルと比較することでホストを特定する。ホスト特定手法を式 (5.1) に示す。

$$H = M \cap S \cap I \cap A \cap F \cap O \quad (5.1)$$

H は特定対象ホストが含まれたホスト群の集合である。そして、 M は特定対象のホストが利用したメールサーバすべてに接続したホストの集合である。ただし、特定対象ホストがどのメールサーバにも接続しなかった場合は、メールサーバを利用しないホストの集合となる。同様に S は特定対象のホストが利用した SSH サーバを利用したホスト群、 I は特定対象ホストが利用した IRC サーバと通信したホストの集合、 A は特定対象ホストが利用したすべてのアプリケーションを利用しているホストの集合、 F は特定対象ホストが発信した特徴的な通信を発信したホストの集合、 O は特定対象ホスト OS を保持している

表 5.2: 本手法の検証環境

検証環境\要素	検証開始時刻	検証期間	ユーザ数	観測ホスト数	トラフィック量
カンファレンス	2009/9/07/17:44	約 47 時間	約 200 名	246 台	約 160GB
研究室 ネットワーク	2010/7/14/18:33 2010/7/20/03:08	約 26 時間 約 24 時間	約 200 名	192 台	約 62GB

ホストの集合を指す。本論文では、式 (5.1) の結果、 H が 1 台のみとなれば対象ホストを特定できたものとして扱う。 H が 0 台の場合は新規ホストのプロファイルとして扱う。

既にプロファイルが作成されていた場合について述べる。図 5.1 に示すように、ネットワークに新しいホストが参加した場合、30 分間ユーザの挙動を取得して新規作成したプロファイルと既存のプロファイルと比較する。比較する情報がひとつも取れず、10 分以内にネットワークを離れたホストのプロファイルは破棄する。そして、新規作成したプロファイルと既存プロファイルが同一であった場合はプロファイルを統合する。統合する場合、プロファイルは完全に上書きせずに差分情報をプロファイルに追加する。また、比較の際は基本的には、一番初めに作成された既存ファイルと比較対象データとし、差分情報は比較対象としない。この処置はプロファイルの増分の情報が肥大化し、すべての新規プロファイルが該当する、もしくは該当なしという結果が出るのを防止するためである。しかし、同一の内容のプロファイルが複数作成された場合や、新規作成したプロファイル情報がどの既存のプロファイルにも当てはまらない場合、候補のプロファイルが複数存在する場合には差分情報も比較対象とする。

5.4 検証

パケットヘッダ情報によるホストを識別手法の検証を 2 つの環境で実施した。表 5.2 に検証環境を示す。

まず、一つ目の検証は、200 人以上の様々な企業や大学といった組織のユーザが参加するカンファレンスで行った。カンファレンスでユーザへのネットワークはすべて無線で提供されており、ユーザが保持するホストが発信したパケットヘッダ情報を収集して、特定を試みた。本手法の有効性を確認する実験をした期間は 2009 年 9 月 7 日 17 時 44 分から 9 月 9 日 16 時 8 分までである。次に、接続先サーバなどが類似し、本論文で定義するホストの所属するコミュニティが同じホストが多く存在する筆者の研究室ネットワークで検証する。このネットワーク環境は、メールや Web ブラウザしか利用しないユーザも多数存在する。まず、2010 年 7 月 14 日 18 時 33 分から、2010 年 7 月 15 日 20 時 17 分のデータから特定のホストのプロファイルを作成し、2010 年 7 月 20 日 3 時 8 分から 7 月 21 日 3 時 30 分のデータを用いて、ホストの特定をした。これら 2 つのネットワークに参加するユーザは事前実験のネットワーク環境に参加するユーザと類似するため、第 5.3 節で述べた判別情報を利用してホストの特定を試みる。

表 5.3: パケットヘッダ情報によるホスト識別結果

識別要素\ホスト	H_1	H_2	H_3	H_4	H_5	H_6
メール サーバ	M_1, M_2 M_3	M_1, M_2 M_4, M_5	M_6	M_1, M_2	M_2	M_1, M_2
SSH サーバ	S_1, S_3 S_4, S_5	S_1, S_6 S_7, S_8 S_9	S_1			
OS	MacOSX	MacOSX	Windows XP	MacOSX	Windows Vista	MacOSX
IRC Server			I_1, I_2 I_3, I_4			
特徴的な挙動	定期的な HTTP 通信				特殊な 発信元 ポート利用	VPN, 定期的に HTTP 通信
起動時の挙動	HTTP 通信 IMAP		HTTP	MSG	HTTP 通信	
アップデート	Evernote MacOSX	MacOSX	Windows	MacOSX	Firefox Quick Time	MacOSX Safari

H:ホスト S:サーバ MSG:Windows Live メッセージ

5.4.1 カンファレンスにおける検証

200 人以上が参加するカンファレンスにおいてパケットヘッダ情報のみを利用した手法で、筆者の所属する研究会のメンバー 6 人が利用しているホストを特定できるか検証をした。対象のユーザのホスト情報は事前に収集し、カンファレンスで観測した 246 台のホストから、該当ホストを発見できたかどうかで検証する。表 5.3 に特定対象としたホストの情報を示す。 H_1 や H_2 は特定ホストを指し、同じく S_1 は特定のサーバを指す。

まず、SSH とメールの接続先サーバを識別要素として、特定のコミュニティに分類後、ユーザのホストを特定する。SSH やメールは自身と関連するサーバを利用するため、宛先ホストからコミュニティの分類が可能である。コミュニティによってホストを分類した結果、特定対象ホストと推測できるホストは 13 台あり、246 台から 13 台まで絞ることができた。ホスト識別結果を図 5.3 に示す。図 5.3 の Y 軸はホストの候補数である。246 台のホストからコミュニティによる分類をした後の 13 台の絞り込み結果を示している。X 軸の各項目は、識別要素をもとにホストを絞り込むことを示している。CO はコミュニティによる分類、M は利用メールサーバ、I は利用 IRC サーバ、O は OS、SA は OS 起動時のアプリケーション、TA は利用アプリケーション、F は特徴的な通信、UP は update 情報を識別要素とした分類を指す。

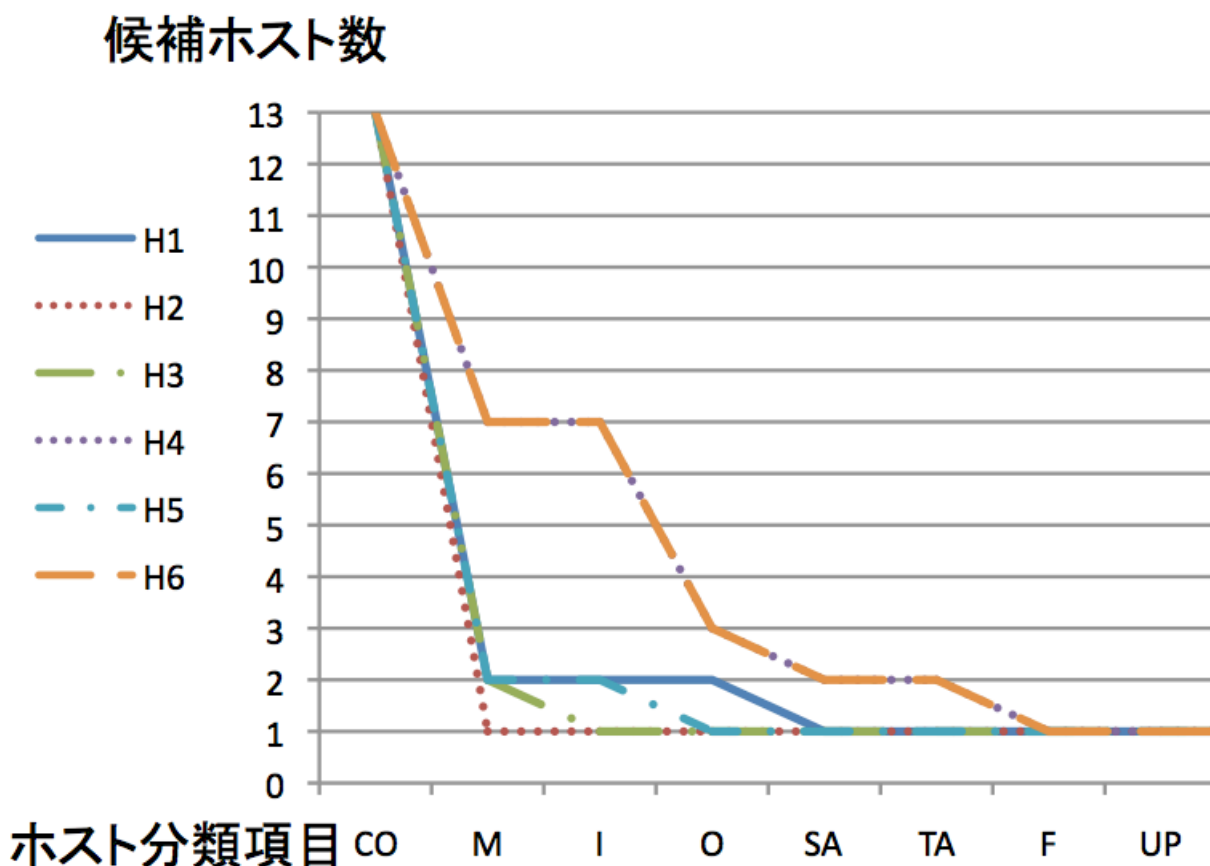


図 5.3: パケットヘッダ情報によるホスト識別結果

次に、高頻度で利用するメールの接続先サーバに焦点を当てる。 H_1 , H_2 は多くのメールサーバとやり取りする傾向があるのに対して、 H_3 は複数のアカウントを外部メールサーバで統合している可能性が高いことがわかる。 M_1 に接続したホストは 8 台、 M_2 に接続したホストは 9 台で、 M_1 と M_2 の 2 台のみに接続したホストは 5 台、 M_3 も加えた 3 台に接続したホストは 2 台あった。 M_6 に接続したホストは 3 台あり、メールクライアント起動時に M_1 , M_2 , M_4 , M_5 の 4 台に接続を試みるホストは 1 台のみであった。メールによる識別で H_2 の候補は 13 台から 1 台に絞り込むことができた。次に IRC を利用しているホストは 13 台中 1 台しか存在しなかった。OS による識別要素においては、13 台中 8 台が MacOSX, 3 台が Windows であり、1 台が WindowsXP, 残りは Windows Vista であった。最後に残った 2 台は Linux であることが分かった。そして、ネットワーク接続時に OS やソフトウェアのアップデート通信を取得し、Firefox を利用しているホストは 4 台であることが分かった。また、 H_1 は、起動時から定期的に同じサーバに対して HTTP でデータを送信する傾向がある。今回は、アプリケーションサーバのアドレスブロックを保持しているため、このアプリケーションが Evernote[38] であり、起動時に設定しているホストが 2 台存在することが判明した。また、 H_6 も起動時から同じサーバにアクセスし続けているため、常駐アプリを自動起動に設定していると言える。 H_5 は発信元ポート番号とし

て 1444 番を常に利用している。このため、特殊なサービスを利用していると推測できる。そして、起動時に毎回 HTTP を利用したダウンロードをしているのが観測された。\$H_5\$ は起動時のパケットのタイミングから Windows であることが判明している。Windows はアンチウイルスソフトが導入されているため、アンチウイルスソフトのシグネチャの更新である可能性が高い。実際に接続先 IP アドレスから Avast と呼ばれるアンチウイルスソフトを利用していることが分かった。\$H_6\$ では、VPN を利用するとともに起動時から定期的な HTTP 通信をしている。VPN を利用しているホストは 13 台中 2 台しか観測できず、かつ MacOSX は 1 台のみであった。これにより \$H_6\$ が判明した。このように、ホストごとの特徴を抽出し分類することで、6 台中 6 台すべてを一意に特定することができた。また、機器特定の際にかかった時間は 30 分から 1 時間以内であった。機器の挙動によって特定できる時間は変化するが、本実験において 4 台のホストは 30 分間のプロファイルを作成直後に自動で判定することができた。

5.4.2 研究室ネットワークにおける検証

作成したプロファイルで、後日ネットワークに接続したホストを特定できれば、今後特定のホストを継続して追跡できるためホストを持つ個人の追跡が可能である。そこで、筆者の所属する研究室において事前にプロファイルを作成し、ホストを分類できるか検証した。プロファイル作成期間は 2010 年 7 月 14 日 18 時 33 分から、2010 年 7 月 15 日 20 時 17 分のデータを、識別検証用のデータは 2010 年 7 月 20 日 3 時 8 分から、2010 年 7 月 21 日 3 時 30 分のデータを利用する。事前に取得したプロファイルを表 5.4 に示す。このプロファイルは、機器に接続して 30 分間のデータをもとに作成した。

対象ホスト \$Ha_1\$ について言及する。まず、事前に作成したプロファイルデータでは接続したメールサーバは 4 台あり、IMAP によって取得しているホストは 3 台、POP によって取得しているホストは 1 台ある。また、利用している SSH サーバは 3 台あった。利用アプリケーションは Evernote と Dropbox[39] であった。このプロファイルをもとにホストを特定する。

$$Ma_{mail1} = Ma_1 \cap Ma_2 \cap Ma_3 \cap Ma_4 \quad (5.2)$$

$$Sa_{ssh1} = Sa_1 \cap Sa_2 \cap Sa_3 \quad (5.3)$$

$$\begin{aligned} Ha_1 &= Ma_{mail1} \cap Sa_{ssh1} \cap A_1 \cap O \\ &= Sa_{ssh1} \cap Ma_{mail1} \cap A_1 \cap O \\ &= Sa_1 \cap Sa_2 \cap Sa_3 \cap Ma_1 \cap Ma_2 \cap Ma_3 \cap Ma_4 \cap A_1 \cap O \end{aligned} \quad (5.4)$$

まず、メールサーバへのアクセスに着目してホストを分類した。検証期間中に、メールサーバ 1 に接続したホストの集合 \$Ma_1\$ は 26 台観測された。メールサーバ 2 に接続したホストの集合 \$Ma_2\$ は 35 台あり、\$Ma_1\$、\$Ma_2\$ 両方に接続したホストは、21 台あった。また、メールサーバ 3 に接続したホスト群 \$Ma_3\$ は、4 台あった。\$Ma_1\$、\$Ma_2\$、\$Ma_3\$ の集合すべてに含まれるホストは 3 台であった。そして、POP を利用してメールサーバ 4 に接続し

表 5.4: 事前に作成したホストのプロファイルデータ

識別要素\ホスト	Ha_1	Ha_2	Ha_3
メール サーバ	Ma_1, Ma_2 Ma_3, Ma_4	Ma_1, Ma_2 Ma_5	Ma_1, Ma_2, Ma_5
SSH サーバ	Sa_1, Sa_2, Sa_3	Sa_1, Sa_4, Sa_5	
IRC サーバ	Ia_1		
OS	MacOSX	MacOSX	Windows XP
特徴的な挙動	Evernote の利用 VPN 接続 Dropbox		Dropbox
起動時の挙動			Dropbox
アップデート	Evernote MacOSX	MacOSX Firefox	

H:ホスト S:サーバ

たホスト群 Ma_4 は 4 台であった。式 (5.2) に挙げられるように、 Ma_1 - Ma_4 の集合すべてに含まれるホスト群 Ma_{mail1} は 3 台あった。次に、メールと同様に SSH サーバを利用したホストの分類に式 (5.3) を用いる。SSH サーバ 1 に接続したホストの集合 Sa_1 には 11 台のホストが、SSH サーバ 2 に接続したホストの集合 Sa_2 には 6 台、SSH サーバ 3 に接続したホストの集合 Sa_3 には 1 台のホストが観測された。そして、 Sa_1, Sa_2, Sa_3 のすべてに含まれるホストの集合 Sa_{ssh1} は 1 台だけであった。また、SSH と IMAP サーバに接続したホストの論理積をとったところ、1 台のみであることが判明した。そして、パケットの傾向から MacOSX であることや、Evernote を利用していることが分かった。MacOSX を利用しているホストの集合 O には 12 台、Evernote を利用しているホストの集合 A_1 には 8 台観測された。これらを式 (5.4) にまとめる。その結果、事前に作成した Ha_1 のプロファイルと、検証データから分類できた Ha_1 は同じホストであると推測でき、実際に Ha_1 が H_1 であることを確認した。

$$Ma_{mail2} = Ma_1 \cap Ma_2 \cap Ma_5 \quad (5.5)$$

$$Sa_{ssh2} = Sa_1 \quad (5.6)$$

$$\begin{aligned} Ha_{2-1}, Ha_{2-2} &= Ma_{mail2} \cap Sa_{ssh2} \cap A_{stapp} \cap O \\ &= Sa_1 \cap Ma_1 \cap Ma_2 \cap Ma_5 \cap A_{stapp} \cap O \end{aligned} \quad (5.7)$$

Ha_2 について述べる。事前に作成したプロファイルデータによると、 Ha_2 はメールサーバ Ma_1, Ma_2 と Ma_5 の各集合に含まれる。また、SSH サーバでは Sa_1, Sa_4, Sa_5 の集合に含まれていることが分かった。このプロファイルをもとにホストを特定する。まず、

式 (5.5) を用いて、メールサーバによる分類をする。 Ma_1 , Ma_2 は前述した通りであるが、メールサーバ 5 に接続したホストの集合 Ma_5 には 25 台のホストがあった。 Ma_1 , Ma_2 , Ma_5 の集合すべてに含まれるホスト群 Ma_{mail2} は 7 台観測された。 SSH サーバにおいて、 Sa_1 の集合には 11 台のホストが確認されたが、 Sa_4 , Sa_5 はどのホストも含まれなかった。そこで、式 (5.6) に挙げるように、 Sa_{ssh2} と Sa_1 は同じ集合とする。そして、 Ma_{mail2} と Sa_{ssh2} の両方に含まれるホストの集合には 2 台ホストがあった。また、起動時に Dropbox を利用しているホスト A_{stapp} は 3 台あった。そして、ホストが MacOSX である集合 O を付け加えて、式 (5.7) を用いたが、2 台のホスト Ha_{2-1} , Ha_{2-2} のどちらが Ha_2 であるか特定できなかった。その原因を調査したところ、 Ha_{2-1} , Ha_{2-2} は 1 台のホストの無線と有線のインターフェースであり、1 台のホストからパケットが発信していることが判明した。そのため、ホストの特定はできたと言える。

$$H_3 = Ma_{mail2} \cap A_2 \cap O_{winxp} \quad (5.8)$$

H_3 について述べる。 H_3 は SSH を利用せず、メールや HTTP 通信を利用するホストである。また、発信元ポート番号から Windows XP であり、Dropbox を利用している。これらのプロファイルデータを踏まえて、検証データから H_3 を特定する。メールサーバによる識別においては、 H_2 の場合と同じく式 (5.5) を用いて 7 台のホストが該当することが分かった。OS に関しては Windows XP を利用しているホストの集合 O_{winxp} には 5 台のホストが含まれる。また、Dropbox を利用しているホストの集合 A_2 には 2 台のホストが該当した。これらを式 (5.8) にまとめる。Windows XP を利用しており、かつ Dropbox を利用し、 Ma_{mail2} に含まれるホストは 1 台のみであり、これが対象ホスト H_3 であることを確認した。これにより、複数の同じコミュニティのユーザが多く存在する環境において、3 台の特定対象ホストすべてを特定できた。

また、機器特定の際にかかった時間はカンファレンスにおける検証と同じく 30 分から 1 時間以内であった。本検証において 2 台のホストは 30 分間のプロファイルを作成直後に判定することができた。

5.5 考察

本論文では SSH やメールの接続サーバを識別要素として挙げた。これは対象のネットワークの参加者が大学や企業に所属するユーザが多く存在するカンファレンスであり、SSH やメールが多く利用されている環境であることが事前検証で判明したためである。特定のユーザが情報を利用するデバイスは、ユーザの利用形態の特徴によって発信する情報に特徴的なトラフィックパターンが現れるため、ユーザが参加するネットワークによっては取得する情報や識別要素も異なる。そのため、識別要素として利用できるプロトコルは企業やカンファレンス、一般的な家庭ネットワークなど環境によって変化する。例えば、日本の 2008 年度のインターネットのトラフィックの傾向 [40] によると SSH は全トラフィックの 0.1% 程度しか利用されておらず、一般的なネットワークでは SSH を識別要素として

利用は困難である。しかし、インターネット白書 2009[41]によるとリモートログインを許可している企業は約 40%以上にのぼるため、企業のネットワークにおいては SSH の割合は識別要素として利用できる可能性が高いと推測できる。また、ISP などが提供する一般的なネットワークにおいて本手法を適応する場合は、ホストのインストール済みのアプリケーションに着目することで識別が可能であると推測できる。例えば、スタートアップ時のホストの挙動、メッセンジャ、スタートアップに登録されているアプリケーション、OS や、メールクライアントの使用の有無などを識別要素として利用できる。このように、ユーザの振る舞いによって取得可能な情報は変化するため、ネットワークに応じて、識別要素を利用することで様々なネットワークでホストの識別が可能である。

5.6 まとめ

パケットのヘッダ情報をもとにホストの識別をする手法を検証した結果、メールサーバや SSH サーバに接続したという情報は、大学に所属するユーザや企業に所属するユーザが多く存在する環境において重要な識別要素となることが判明した。この識別要素を利用することでホストをある程度、もしくは完全に特定することができる。特にメールサーバの接続に関しては、メールクライアント利用時に事前に設定した複数のメールサーバ群との通信が必ず発生するため、ホストの特定では重要な識別要素であると言える。その後、アプリケーションや OS 情報を利用してホストの特定をする。これによって、パケットのヘッダ情報のみからホストを絞り込むことや、複数インターフェースからパケットを発信する 1 台のホストを特定できることを示した。実際に事前にプロファイルを作成した 3 台のホストに関して、本人拒否率は 100%であり、他人受入率は 0%という検証結果が出た。本提案手法によって、ユーザアカウント登録や専用デバイスの利用をユーザに強いることなく、ユーザのホストを検出することが可能であることを示した。しかし、重要な識別要素となった SSH や IRC は、メールや Web ブラウジングのみのユーザも存在するため、SSH サーバを利用しないようなユーザの特定も必要である。そのような一般的なユーザは、カンファレンスでの検証では H_3 , H_4 , 研究室ネットワークの検証では H_3 を特定したことで可能であることを示した。これによって、本提案手法はホストを識別することができることが分かった。

第6章 ホストとユーザを結びつける手法

6.1 手法概要

ホスト利用者を判別するにあたり、誰がどの機器を保有しているかといったユーザ情報はネットワーク管理者や、ユーザに応じたサービスを提供する側にとっては重要な情報である。特に、近年スマートフォンやノートブックの台頭によって、一人で複数の機器を持つユーザが増加しているため、誰がどのホストを何台持っているかの情報を取得することによって、より効率的なサービス提供やネットワーク管理が可能である。しかし、前述した機器情報収集によるホスト識別や、パケットヘッダ情報による手法では、ホストの特定、追跡が可能であるが、ユーザの特定までは不可能である。ホストとホスト利用者を結びつける方法に、ユーザにホストの機器情報と本人の名前といった識別IDを登録するアカウント登録がある。しかし、アカウント登録はユーザに操作を要求するため不特定多数のユーザが利用する場合や、管理の緩いネットワークにおいて、ユーザにアカウント登録を義務付けられない場合がある。また、アカウント登録は容易に詐称が可能であり、完全にユーザを特定する場合には適さないことが多いため、ホストユーザを結びつける方法が必要である。

そこで、ユーザ在席情報などを用いて対象ホストとホスト利用者を結びつける手法を提案する。本手法はGPSなどの専用デバイスの利用といった制限をなくすとともに、対象ホストの利用者及び、一人で複数の機器を持つユーザを特定することを目的とする。本手法の前提として第4章と第5章で述べたホストの識別手法によって、ホストを一意に特定できた場合を想定し、特定できたホストとユーザを結びつける。

6.2 判別手法

本提案手法の概要図を図6.1に示す。本手法はターゲットとするホスト所有者が特定の時間、ネットワークを利用している情報、もしくは利用していない情報をホストの情報と組み合わせてホスト所有者を特定する。ターゲットとするホスト所有者の在席情報を確認する方法は複数存在する。例えば、情報収集者が、直接ターゲットのホスト利用者がネットワークを利用しているかの情報を視認や、会議の出欠名簿などのリストから判断する、同じ場所でターゲットのホスト利用者と一緒にいたユーザがその情報を伝達、もしくは監視カメラなどにターゲットが映っている情報などを利用することが挙げられる。今回は、前述したホスト所有者がネットワークの利用の有無の確認に、情報収集者による視認も用いた手法を利用する。対象のユーザが、在席した時間、退席した時間とパケットのヘッ

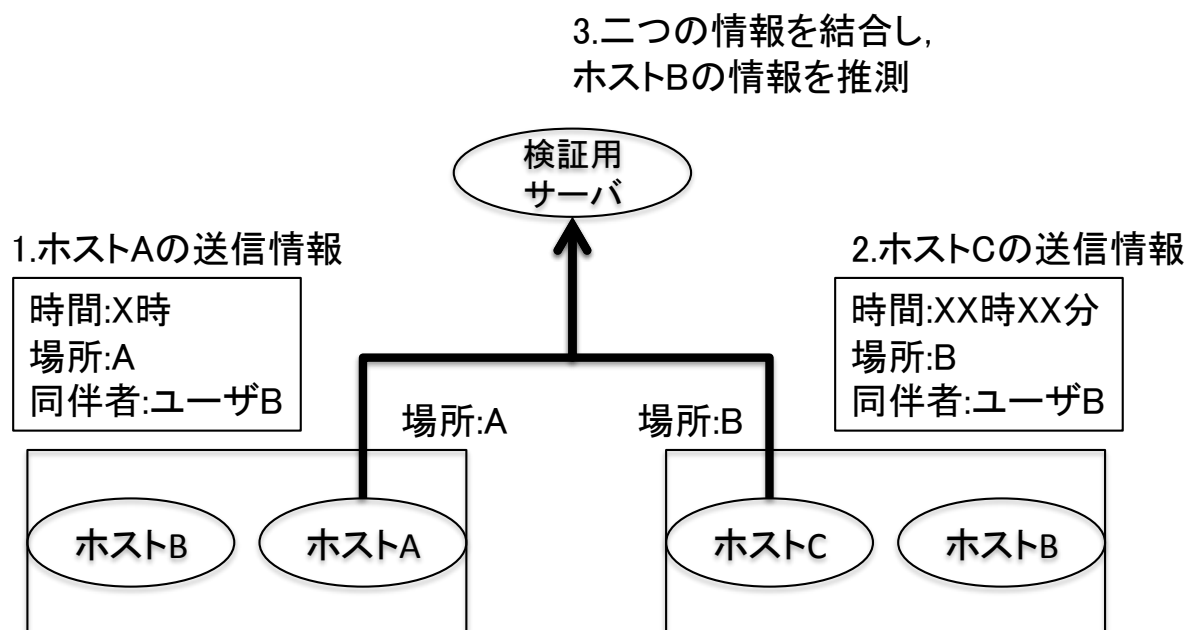


図 6.1: ホスト情報を推測する手法

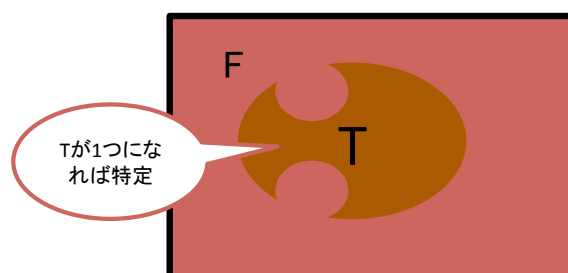


図 6.2: ユーザの絞り込み概要図

ダ情報によるホスト識別情報の論理積を取ることで、ホスト利用者の特定が可能である。これによって、アカウント登録と比べて、ユーザがMACアドレスだけでなくホスト自体も変更した場合や、新しくネットワークに参加したデバイスをネットワークに持ち込んだ場合にも発見することが可能である。しかし、筆者が提案している視認によるユーザの識別手法 [42] では、所有者の在席時間のみでは確実に絞りきることができなかった。そこで、本論文では対象ユーザの不在情報を追加することにより、ホストとホスト利用者の結びつけの手法を改良し、精度を向上した手法を提案する。

6.2.1 在席情報による識別

図 6.2 に本手法のアルゴリズム概要を示す。対象ユーザが在席したという集合 T に対し、対象ユーザが見つからなかった集合 F の情報を追加していき、 T が 1 になればユーザ

を識別できたと言える。

そして、実際にユーザが存在しない場合や、存在しなかったケースが考慮される。その場合、対象ユーザの MAC アドレスが含まれた集合を $Target^{set}$ とし、サンプリング時に対象ユーザの存在を確認する集合を S_T 、ターゲットを確認しなかった集合を S_F とし、 $Target^{set}$ は式 6.1 で表す事ができる。

$$Target^{set} = (S_T \cap (\cap \bar{S}_F)) \quad (6.1)$$

この $Target^{set}$ の MAC アドレス群を 1 台に絞り込むことが、本手法の目的である。具体的な手法として、まず、観測者 1 がユーザ X のホストと同じ場所で行動をし、観測者 1 はホスト A を利用してユーザ X が場所 A にいるという情報を本システムに通知する。通知する内容は、物理的な場所、対象ユーザの名前、時間の 3 つが含まれる。本システムはその情報を収集し、入力があったタイミングの前後 5 分間に観測されたホストの中に、ユーザ X のホストが含まれているとする。次に、後日観測者 2 がユーザ X のホストと同じ場所で行動した場合も同様にホスト C を利用して本システムに通知することで、ユーザ X のホストを過去のデータと照らし合わせて絞り込むことができる。この動作をホストが 1 つになるまで繰り返せば、ユーザ X の持つホスト B を特定できる。絞り込みによるホストを特定する方法を式 (6.2) に示す。H が対象とするホストの集合、 TH_n は n 回目のターゲットユーザが存在する視認情報を入力した時間、前後 5 分間に検出された H の候補となるホストの集合である。また、本人が存在しなかったという情報も有用な識別要素となる。ターゲットユーザが存在しないホスト群を NH とし、m 回目のターゲットユーザが存在しないホスト群を NH_m とする。これによって、前述した常時稼動している機器があった場合でも、正確に情報を取得できる。これら、 TH と NH_m の論理積により H が 1 台しか存在しなければ、ホストとホスト利用者を結びつけることができたと言える。

$$H = TH_1 \cap \dots \cap TH_n \cap \overline{NH_1} \cap \dots \cap \overline{NH_m} \quad (6.2)$$

本手法はホスト利用者がネットワークを利用している情報をもとに、ユーザを特定するが、常時稼動する機器を除外することでより精度を向上させることができる。例えば、ホスト利用者が長時間同じネットワークを利用しているケースが挙げられる。その場合、ユーザがネットワークに存在しない場合の情報を取得できず、ホスト利用者がネットワークを利用している情報のみであった場合、式 (6.2) の手法では、必ず常時稼動するホストも候補として組み込まれてしまう。そのため、可能であれば常時稼動しているホストやサーバを、事前に本手法で収集し、除外することで精度を向上することが期待できる。

次に、ユーザのデバイスが 1 台に絞り込めていない場合やユーザが複数台のデバイスを保持している場合に、クライアントが収集した MAX 個の MAC アドレスのうち、ターゲットユーザが保持するデバイスの MAC アドレス (M_i) である可能性は式 6.3 で表すことができる。

$$P(M_i | Target^{set}) = \frac{\sum_{y=1}^{MAX-1} MAX-1 C_y}{\sum_{k=1}^{MAX} MAX C_y} \quad (6.3)$$

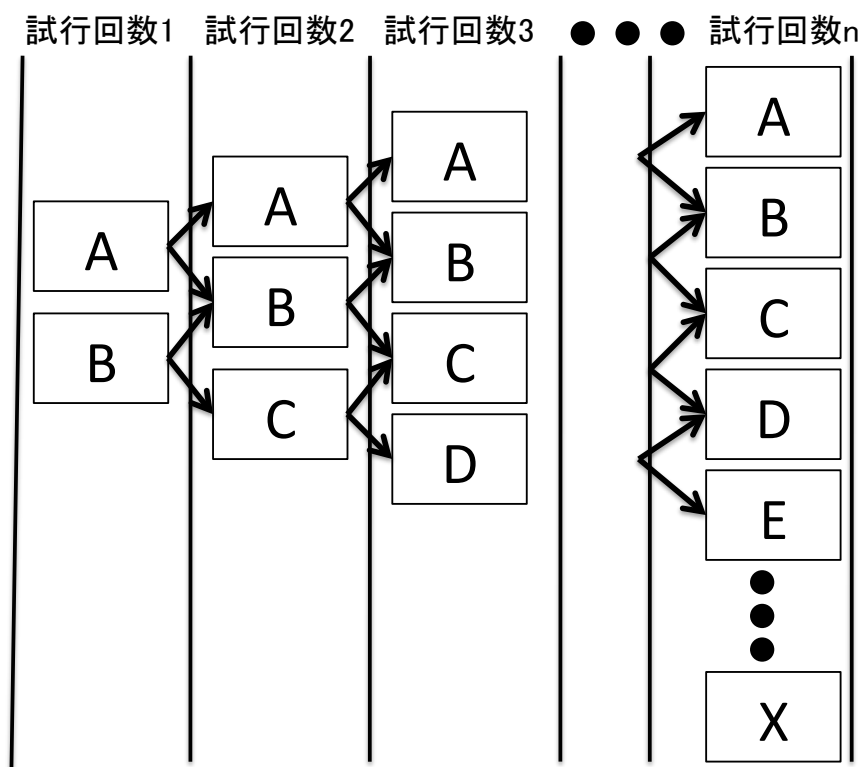


図 6.3: ホスト所有者の絞り込み手法

これによって、多くの候補を絞り込むことができる。この絞り込み手法を図 6.3 に示す。まず、1 回目の試行によって発見した MAC アドレスは GroupA に、発見できなかったホストは GroupB に組み込まれる。2 回目の試行でまた発見した場合は GroupA、発見できなかった場合は GroupB となる。1 回目の試行で発見できなかったが、2 回目で発見できた場合は GroupB に組み込まれる。この Group の数は試行回数 n に 1 を加えた値の数が存在する。

例えば、5 回の試行で 4 回発見され 1 回は発見できなかった場合は対象の MAC アドレスは GroupB となる。基本的に対象ユーザの MAC アドレスの候補は GroupA が検出されるが、GroupA の MAC アドレスがすべてない場合は、GroupB の MAC アドレスが候補として表示される。これによって、式 6.2 とは異なり継続して MAC アドレスの探索が可能となる。

本手法を第 4.4 節のクライアントに実装した。タイミング情報を登録するインターフェースの実装を図 6.4 に示す。上の図が人の情報を入力するインターフェースであり、下の図は第 6.2.2 節で述べる SNS を利用したユーザ結びつけ手法を利用するためのインターフェースである。

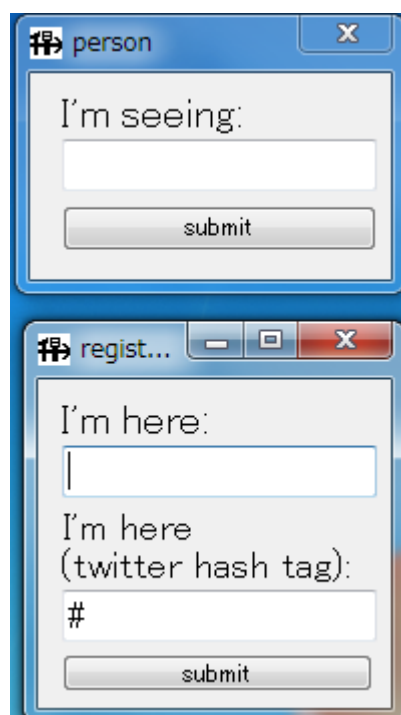


図 6.4: 人のタイミング情報の登録

6.2.2 SNS を利用した結びつけ手法

存在情報だけではなく SNS 情報も利用可能である。例えば，twitter サービスにおいては，イベントなどのハッシュタグや GPS 情報を利用できる。対象ユーザのデバイスの位置情報やイベントに参加していることを発信した際にセンサとなるクライアントがいた場合，存在情報と同じようにユーザを特定することが可能となる。これによって，ネットワークの場所推定や，個人とソーシャル ID の結びつけが可能となり，ホスト情報と対象ユーザの人間関係を始めた SNS に関する項目を取得することができる。

6.3 ホスト利用者特定手法の検証

6.3.1 在席情報を利用したホスト利用者特定手法の検証

実際に本手法でホスト所有者を特定できるか，筆者の所属する研究室で検証した。筆者の所属する研究室は 100 名程度のユーザが在籍し，ネットワーク上では 200 台を越えるホストが観測できる。また，研究室には誰か 1 人は在席しており，数日間連続して宿泊するユーザも存在する。このような環境において，無作為に選出した 3 台のホスト情報から，ホスト所有者が特定できるか検証した。検証方法はそのユーザが所属するネットワークに出現しているのを筆者が視認した上で情報を入力する。

対象ホストの候補数/観測されたホストの総数

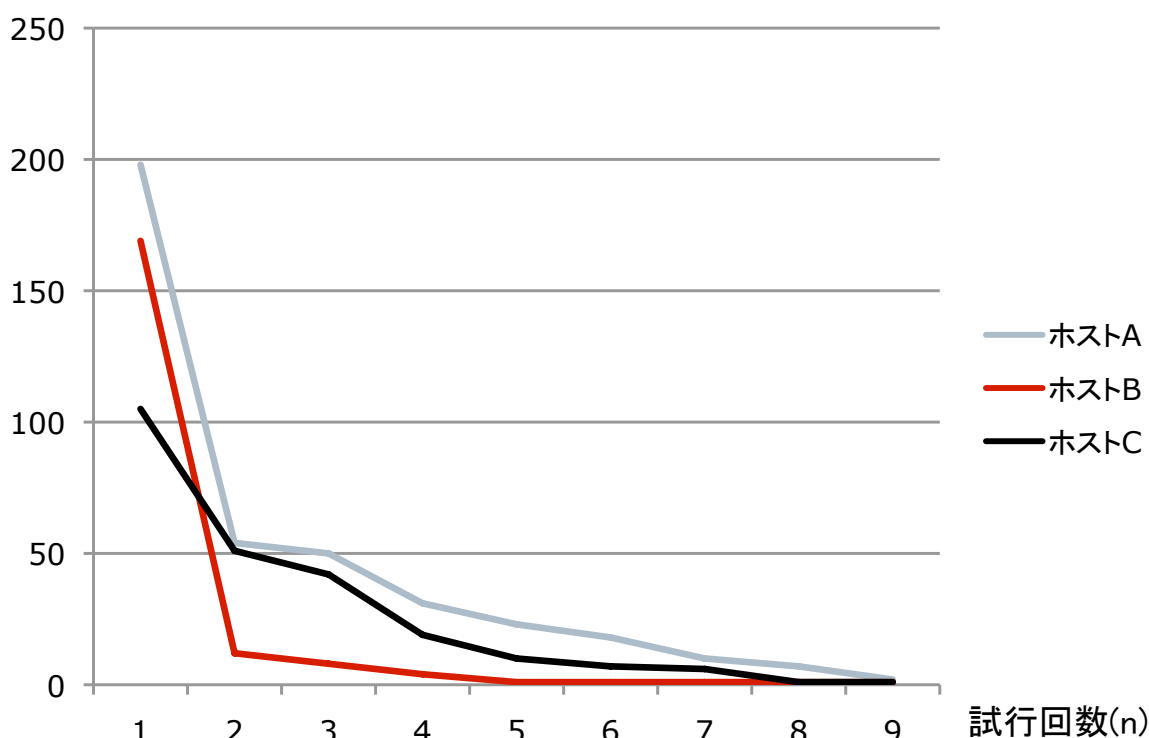


図 6.5: ホスト識別手法結果

しかし、検証環境において、長期間泊まりこむホスト利用者が存在するという特殊な場合も考えられるため、第 6.2 説で述べたように、常時稼動する機器と判別が困難となる例外が発生する可能性がある。その例外に対応するため、常時稼動しているホストをリストから除外する必要がある。そこで、3 日前からデータを収集し、取得した 363 台のホストから常時稼動しているホスト 66 台のリストを作成、除外した。そして、各対象ユーザの在席情報と退席情報を交互に入力した。結果を図 6.5 に常時稼動ホストを除外した結果を示す。

ホスト A は 9 回で 2 台、ホスト B は 9 回で 1 台、ホスト C は 5 回の試行で 1 台に絞り込むことができた。ホスト A が 1 台まで絞り込むことが困難であった理由は、1 人でノートパソコンとスマートフォンの 2 つのデバイスを所持するユーザであったため、一意に特定ができなかったのが原因であった。他の実験でもこのような例は多く挙げられた。このように複数回以上の施行で 1 台にまで絞りきれなかった場合は、対象の 2 台の機器がネットワークに出現したすべての時間を参照し、その時間が一致するかを確かめる。今回の場合は 2 つの機器がネットワークに出現した時間が全く同一であったため、ユーザが複数所持しているホストであると判定した。これらの結果から在席情報と退席情報を取得した視認によるホスト利用者を特定する手法は有効であると言える。これらのことから、本提案手法においては、5 回から 9 回程度の絞り込みでユーザを特定できたとと言える。

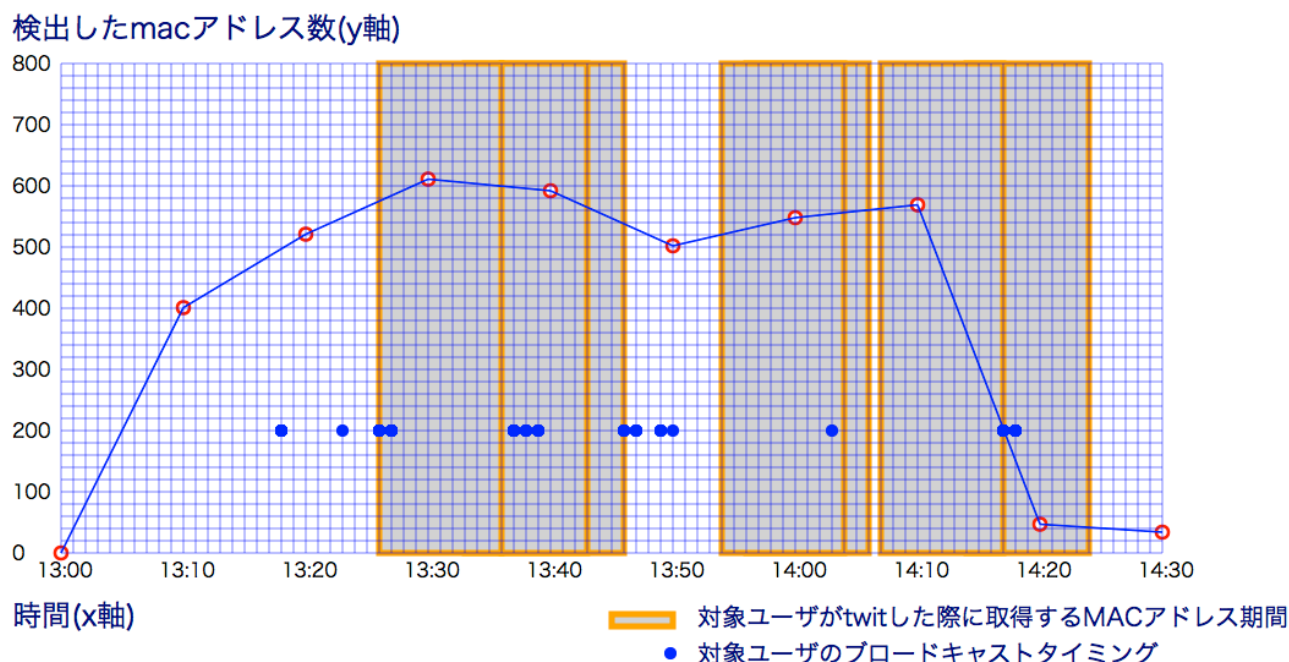


図 6.6: SNS を利用したホスト利用者特定手法の試行

また、ベンダ情報を用いた手法では、2回から6回程度の施行で1台のホスト絞り込めることができた。このことから、ベンダ情報はホストの絞り込みに対して有効であることが分かった。

6.3.2 SNS を利用したホスト利用者特定手法の結果

図 6.6 に SNS と取得した MAC アドレス情報を絞り込んだ結果を示す。

Y 軸が検出した MAC アドレス数であり、X 軸が時刻である。この折れ線グラフはセンサとなるクライアントが収集した MAC アドレスの総数であり、青い点は対象ユーザがブロードキャストしたタイミングであり、オレンジ色の枠は対象ユーザが Twitter で情報を送信した前後 5 分間の記録である。この図においては、90 分間において 7 回ユーザが情報を発信しているため、7 回ホストを絞り込む。その結果、この試行において 430 台のホストに絞り込むことができた。

図 6.6 の試行を 5 回繰り返して、どの程度ユーザを絞り込むことが可能かを実験した結果を図 6.7 に示す。1 回目の試行で候補であるホストは 430 台であり、2 回目の試行では 122 台、3 回目の試行では 81 台、4 回目は 23 台まで絞り込むことができた。しかし、5 回目の試行では 21 台と 4 回目と比べて 2 台のみ減らすだけの結果となった。イベントの回数上 5 回しか実験できなかったため、絞り込むことができなかった。また、24 時間常に稼動しているホストを除外した結果、5 台まで絞り込むことができた。

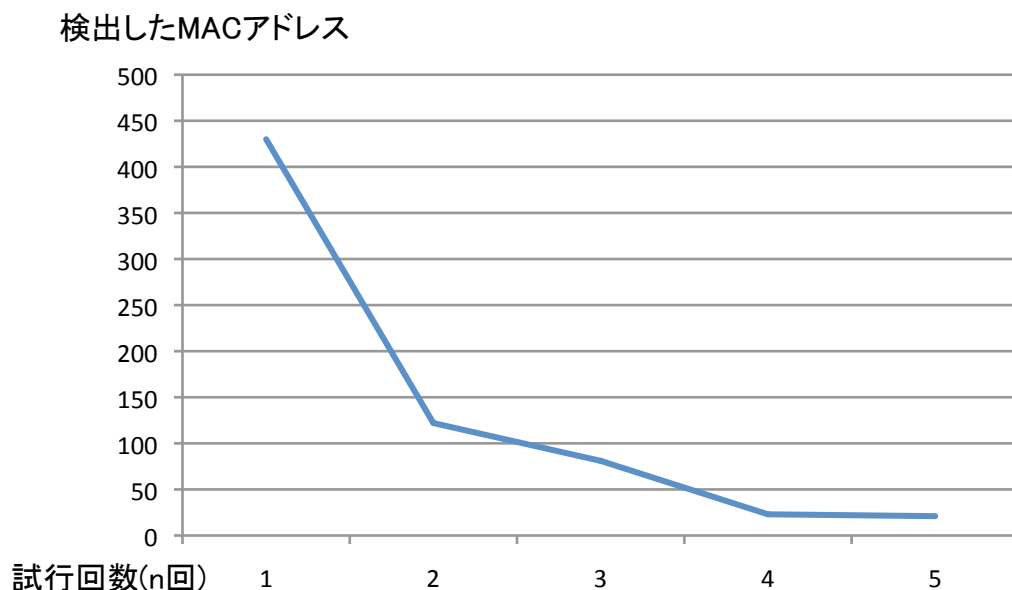


図 6.7: SNS を利用したホスト利用者特定手法の結果

6.4 考察

本手法は以下の点を踏まえて、改善する必要がある。1点目は、ホストマッチングシステムにおける常駐ホストを除外する必要がある。本論文において、常駐ホストは一定の範囲で除外されているが、確実に除外できるわけではない。そのため、常駐ホストを検出後、リストに記録するなどの対応が必要である。2点目は、ホスト放置による誤検出の対応が必要である。ホストをネットワークに接続したまま放置した場合、本システムは誤検知する可能性がある。そのため、放置時のパケットのタイミングを調査することで、検出するか実験する必要がある。

SNS によるホスト利用者特定手法では5回では十分に絞り込むことができなかった。その原因として、イベントの特性上、同じユーザが複数回出現するような特性であることが考えられる。しかし、常時稼働しているホストを除外した結果5台に絞り込めた。また、本実験においては1つのイベントで実験したが、複数のイベントで本実験をした場合は、かなり絞り込めることが予想できる。

6.5 まとめ

ユーザのネットワーク在席情報を利用した識別手法において、ホスト所有者の特定が可能であることを示した。今回は情報収集者の在席情報による識別情報を利用したが、それ以外の情報も利用可能である。例えば、管理者自身がリアルタイムに情報を入力することなく、一般ユーザから対象ユーザの情報を得ることや、会議の参加者名簿や会議のログなどと連動することにより、管理コストを低減することができる。対象ユーザの情報を得る

手段の一つとして SNS の情報を挙げた。そして、SNS 情報を用いたユーザ識別手法を提案し、実装した。SNS を用いた施行は、5 回の試行回数で 20 台程度に絞り込めた。本実験は一つのイベントのみを対象としたため、複数の会場で実験した場合、さらに絞り込めることが期待できる。

しかし、アカウント登録手法と比較すると、ユーザ全員にアカウント登録の義務付けが困難なネットワークなどでの利用や、MAC アドレスなどを詐称された場合でもホストやホスト利用者を特定できる反面、ネットワーク管理者の管理コストはネットワーク参加者の数に比例する。そのため、ネットワークの環境や目的によってアカウント登録手法や本手法を組み合わせることで、より効率的にホストとホスト利用者を結びつけることができる。

本検証環境は、必ず一人のユーザが存在する環境で行った。そのため、多くのユーザが同時に在席、退席する企業のような組織のネットワークとは異なる。しかし、企業や他のネットワークにおいても、ユーザの在席情報、退席情報は用事や仕事などで異なるため、検出できると推測できるが、実際に検証する必要がある。今後は、管理者以外のユーザが管理者に対象ユーザの在席情報を伝えられる基盤を提供する。

第7章 プライバシに関する考察

本提案手法によって、ホストやユーザの識別が可能であることを示した。そこで、本章では各手法の識別システムを実際に導入した場合の社会的影響についてプライバシーの考察をすると共に、今後の課題について述べる。

7.1 識別システムとプライバシー

近年、個人情報の流出やプライバシーの侵害が年々増加しており、プライバシーに関する懸念が強まっている [43]。しかし、端末を所有しているユーザは、あまり対策をしていない。(独) 情報処理推進機構セキュリティセンター調査 [44] では、日本人はプライバシー侵害のリスクを認識し強い関心を得ている一方、対策に関してはEU 諸国より低いという結果となった。つまり、プライバシーについて関心はあるが、リスク対策などの行動が伴っていないというのが現状である。その原因の一つとして、個人情報に関わる情報が多く煩雑であり、何を守ればいいのか不明瞭であるということが挙げられる。これは、普段から意識せずに発信している情報が個人情報となりうるなど、計算技術の発展によって、個人情報とされる情報が変化しているのが一つの要因となっている。このような情報は、ホットスポットや公衆無線 LAN など不特定多数のユーザが接続する機会が増えているため、何の情報を発信しているのかを知ることがより重要となっている。このように、公共の場で個人に関する情報を発信している場合、端末に関する情報、端末所有者の行動履歴などを取得されプライバシーを脅かされる可能性がある。しかし、その反面、どのような情報を提供してよいかユーザが知ることによって、サービスの幅をより広げる事ができる。

近年では個人情報を扱うにあたり、プライバシーの議論を踏まえたシステム設計など十分な配慮が求められる [45]。このプライバシーの保護とサービスの発展のバランスが非常に重要であるため、十分な検討が必要である。

7.2 個人情報取り扱い形態とその懸念

第 2.2.3 節で述べたように、サービス提供時に情報を利用できる範囲は、サービスの形態によって異なる。形態には、オプトイン、オプトアウト及びノンプライバシーの3種類存在するが、どのサービスがどこまで利用できるかを以下にまとめる。

取得した情報がどのように使われているか明確にされていない場合、更に問題が生じる事となる。個人情報を利用する場合、OECD 原則 [46] に準じてサービスを提供することが望ましいとされているため、本提案手法と OECD8 原則について考察する。

表 7.1: OECD ガイドラインと各サービス形態の対応

OECD8 原則の項目	オプトイン	オプトアウト	ノンプライバシー
収集制限の原則	○	×	×
データ内容の原則	○	○	×
目的明確化の原則	○	○	×
利用制限の原則	○	○	×
安全保護の原則	○	○	○
公開の原則	○	○	×
個人参加の原則	○	×	×
責任の原則	○	○	×

○：満たしている，満たすことが可能

×：満たしていない，満たす必要がない

7.2.1 オプトインにおける運用の特徴

オプトインによるサービス形態は，許可を得たユーザのホストを登録し，ホスト及びユーザの識別をする．ネットワークデバイス情報を用いたホスト追跡手法においては，クライアント同士の通信のみを収集しその情報以外はすべて破棄するという状態である．オプトインの場合，ユーザはサーバに情報を登録するため，パケットヘッダ情報を用いたホスト識別手法や，ホストとユーザを結びつける手法は必要ない．つまり，本提案手法をオプトイン形式で運用した場合，ユーザはアカウント登録など予めホストとユーザ情報を収集し，お互いに登録したユーザ同士で情報を交換する方法となる．また，周囲のブロードキャスト情報や，マルチキャスト情報の収集する必要なく，サーバと直接データをやり取りするだけで実現可能である．この形態は一般的な Web サービスなどと同じく，一般的なサーバクライアントモデルと同様になる．利点としては，ユーザの同意を得ているため法的観点からの問題が少ない点があげられる．反対に欠点としては，同意したユーザの情報しか取れないため，十分なサービスを提供できないという問題点がある．

本提案手法をオプトインで利用した場合の OECD8 原則について述べる．オプトインの運用が OECD8 原則を満たしているか検討した場合，本提案手法はほぼすべての基準を満たすことが可能である．順番に検討する．

● 収集制限の原則

収集の原則であるが個人データは同意を得て収集される点については，オプトインで有るため利用規約に同意したユーザのみ利用する．そのため，すべての利用者は同意して本システムを利用しているため，満たしていると言える．

● データ内容の原則

利用規約に目的を明記し，管理者が目的通りに運用することによってこの項目は満たすことができる．また，情報は最新の内容に保つように努める必要がある．

- **目的明確化の原則**

データ内容の原則と同じく、目的を明確化し遵守することで目的明確化の原則を満たすことができる。また、目的が変更時された場合、収集制限の原則におけるユーザの同意も再度必要となる場合があるため、変更が必要な場合とそうでない場合の範囲を明確にする必要がある。

- **利用制限の原則**

収集データを目的以外に利用しないことで利用制限の原則を満たすことができる。しかし、第三者に譲渡する場合は利用目的として明確化する必要がある。MAC アドレスやそれに紐づいたホスト、ユーザ情報は個人情報と扱われるため、公開範囲や譲渡先について明確に規定されなければならない。

- **安全保護の原則**

個人データの紛失、破壊、私用、修正、開示などから保護される必要がある。この原則に関しては、サーバ管理者など対象データを触れる管理者に対して監督を行い、漏洩防止の処置を講じることで満たすことができる。

- **公開の原則**

データの存在、利用目的、管理者の公開する原則に関しては、利用規約で明記することで満たすことができる。また、訂正や利用停止に対して遅滞なく対処できるようにする必要がある。

- **個人参加の原則**

本提案システムはサーバクライアントモデルであるため、利用規約に同意したユーザが本人に対する情報の開示要求に容易に応えることができる。そのためこの原則を満たしていると言える。

- **責任の原則**

利用規約に責任者を明記し、データ紛失時や改ざん時の補償を明記することで責任の原則を満たすことができる。

オプトインの運用が OECD8 原則満たしているかを比較した結果、運用方法や利用規約の整備をすることで原則を満たすことが可能であることが分かった。

7.2.2 オプトアウトにおける運用

オプトアウトにおける運用について述べる。オプトアウトによるサービス形態は、サービスを許諾しないユーザが特定の行動を起こすことによって、サービスから除外する方法である。この形態で本手法を運用した場合、特定の行動を取らないユーザはすべてサービスに同意したとみなし、ホストやユーザを識別する。本論文ではこのオプトアウトの形態

を想定している。オプトアウト形式のメリットは、ネットワークに存在するすべてのホストを収集できることである。そして、それらの情報を共有することによって、ホストの識別やユーザの識別が容易になる。反対にデメリットとしては、ユーザが知らないうちにホストやユーザが識別されてしまうため、プライバシーの侵害となってしまう可能性がある。そのため、サービス開始時には十分に告知して周知を図ると共に、識別されたくないユーザを除外する機構が重要である。

本提案手法をオプトアウトで利用した場合の OECD8 原則について述べる。オプトアウトで運用した場合、OECD8 原則を満たしているか検討した。

- **収集制限の原則**

オプトアウトによる運用の場合、収集の原則を満たすことができない。オプトアウトの場合、ユーザの同意を得たものとして扱うが周知が徹底されていない場合において、ユーザの同意を得ずに情報を収集してしまうことが想定される。そのため、ユーザが知らないうちに個人情報を収集される脅威が発生する。

- **データ内容の原則**

利用規約に目的を明記し、管理者が目的通りに運用することによってこの項目は満たすことができる。また、情報は最新の内容に保つように努める必要がある。

- **目的明確化の原則**

データ内容の原則と同じく、目的を明確化し遵守することで目的明確化の原則を満たすことができる。また、目的が変更された場合、収集制限の原則におけるユーザの同意も再度必要となる場合があるため、変更が必要な場合とそうでない場合の範囲を明確にする必要がある。個人情報保護法におけるオプトアウトの原則として、本人に通知もしくは知り得る状況に置いていなくてはならない。そのため、目的明確化においてはオプトアウトでの運用にあたり満たしていなければならない項目である。

- **利用制限の原則**

利用制限の原則はオプトインによる原則と同じく、収集データを目的以外に利用しないことで利用制限の原則を満たすことができる。しかし、第三者に譲渡する場合は利用目的として明確化する必要がある。MAC アドレスやそれに紐づいたホスト、ユーザ情報は個人情報と扱われるため、公開範囲や譲渡先について明確に規定されなければならない。しかし、オプトインとは異なり Web サイトなどのユーザの利用規約を変更されるだけで、目的の変更は可能と判断される場合が多い。一方、ユーザは知らない間に利用規約が変更となり問題となる場合も多いため慎重な運用が求められる。

- **安全保護の原則**

オプトインの運用と同じく、個人データの紛失、破壊、私用、修正、開示などから保護される必要がある。この原則に関しては、サーバ管理者など対象データを触れる管理者に対して監督を行い、漏洩防止の処置を講じることで満たすことができる。

- **公開の原則**

データの存在，利用目的，管理者の公開する原則に関しては，利用規約で明記することで満たすことができる。しかし，オプトインでの運用とは異なり，ユーザがサービスの存在を知らないまま情報を収集されているケースが想定できる。その場合，訂正や利用停止の申請をできない。このことから，公開の原則を十分に満たすことはできないと言える。

- **個人参加の原則**

本提案システムはサーバクライアントモデルであるため，利用規約に同意したユーザが本人に対する情報の開示要求に容易に応えることができる。そのためこの原則を満たしていると言える。

- **責任の原則**

利用規約に責任者を明記し，データ紛失時や改ざん時の補償を明記することで責任の原則を満たすことができる。

オプアウトの運用が OECD8 原則満たしているかを比較した結果，収集制限の原則と，個人参加の原則を満たすことができないことがわかった。MAC アドレスを用いてユーザ情報を結びつける行為が個人情報の脅威となると将来なった場合にオプアウトの運用は困難である。そのため，どこまでが許されるのかなどプライバシーに関しての議論が必要である。

7.2.3 ノンプライバシーにおける運用

ノンプライバシーによる本提案手法の利用は主に犯罪捜査や企業におけるフォレンジックや企業ネットワークにおけるホスト及び，ユーザの監視に利用する。犯罪捜査においては，法的措置によってプライバシーに関する法律や規制が適用されないことがある。そのため，OECD ガイドラインやユーザの同意などに関わらず情報の収集解析が可能となる。

また，企業における情報の監視において，監視をする企業の社員は企業に属する際にプライバシーに関わる契約に同意している事が多い。そのため，分類としてはノンプライバシーではなくオプトインとなるが，契約の内容は広く情報収集に同意するものであることが多いため，広く解釈が可能である。その意味からノンプライバシーと言える。

ノンプライバシーによる運用において，本提案手法はホストやユーザの識別を容易に行うことができる。また，すべてのユーザの取得情報を見ることができるため，一般的な手法よりも容易にホストやユーザの特定が可能となる。

7.3 サービス応用例とその運用形態

将来的に本手法によって用いられるサービスの例と適切な運用形態について述べる。

● 広告への利用

本提案プラットフォームはターゲティング広告に利用する事ができる。スマートフォンや小型 PC を持ち歩くユーザは非常に増加しており、ホストを識別することでユーザも識別することができる。同時に、識別したホストに興味や趣向だけでなく人間関係の情報が結びついていた場合は、はじめて検出したホストで人間関係をもとに、同じく広告を出す事ができる。

サービス利用形態に関しては、オプトインが望ましい。特にユーザの趣味や趣向とホスト情報が結び付けられる場合、プライバシーが脅かされる場合があるので、オプトインによる運用が有効的である。

● ネットワークの状態把握

本提案プラットフォームを用いてユーザが容易にネットワークの状態を把握する事ができる。例えば、初めて行った場所にどのようなユーザが在席しているのかの情報や、どのような人々がネットワークを利用しているのか、どの程度のユーザが参加しているのかなど、ネットワークの状態を把握することができる。

サービス利用形態に関しては、オプトインが望ましい。特にどのユーザが同じネットワークに存在するかを知られることは、プライバシーが脅かされる可能性が考慮されるため、オプトインによる運用が有効的である。

● ユーザ間の交流支援

本提案プラットフォームを利用することで、同じネットワークに存在するユーザが認識できる。それに加えて、情報を共有することで、ユーザは対象ユーザがどのネットワークにいるか知ることができる。そのため、Google latitude[47]などのサービスのよう位置情報の共有などにも利用可能である。また、本提案プラットフォームはサーバクライアントモデルであるが、P2P を利用して情報を共有することで、サーバと接続できない場合でもサービスを継続して受けることが可能である。

サービス利用形態に関しては、オプトインが望ましい。特にユーザの位置情報を共有するとなると、ユーザトラッキングが常時可能となってしまうため、ユーザ自身が共有の許可、不許可を選択できる形態でなくてはならない。そのため、オプトインによる運用が有効的である。

● イベントにおける人の移動分析

イベントにおいて、人がどのように移動するのかといった分析用ツールにも本提案プラットフォームは利用可能である。第 4.7 節で述べたように、機器情報を用いたホスト識別手法はセンサの設置状況に応じて精度が変化する。しかし、ホストや人が多く集まる場所では精度が向上するため、人がどこに多く集まるかを把握する事ができる。これをイベント会場に応用することで、従来必要であった RFID など専用機器の配布や、人数を数える監視員の設置が不要となる。そして、より容易にどの場所で人が多くいたか、どのように移動したかといった人の遷移情報を知ることができる。

サービス利用形態に関しては、オプトインが望ましい。特にユーザの位置情報を

共有するとなると、ユーザトラッキングが常時可能となってしまうため、ユーザ自身が共有の許可、不許可を選択できる形態でなくてはならない。そのため、オプトインによる運用が有効的である。

- **出席管理への利用**

教育機関や企業において、対象ユーザの出欠及び、出退勤を本提案プラットフォームで管理することができる。特に今まで紙や IC カードなどを用いてユーザの出退勤などを管理しているが、このプラットフォームによって、対象者のホストや機器を登録するだけで、ユーザに負担を強いることなく可能となる。しかし、なりすましが容易であるために、システム導入には十分に検討する必要があるが、管理者やユーザの手間を低減することができる。

サービス利用形態に関しては、オプトインが望ましい。ユーザの行動履歴を記録共有してしまうため、対象のユーザ以外の情報を保持することは望ましくない。そのため、事前にユーザに許可を得るオプトイン形式をとることで、ユーザの行動履歴から出欠や出退勤の管理が可能となる。

- **既存サービスとの併用**

第 2.3 節において述べているが、新しいサービスと併用することで、サービスをより発展させることが可能となる。現状ではアカウント登録などユーザが自身の情報を入力することでホストやユーザを識別しなければサービス受けることができない。しかし、本提案プラットフォームを利用することで、ユーザは登録やログインといった作業をせずともサービスを容易に受けることが可能となる。

このように既存のサービスとプラットフォームを併用する場合においては、利用形態はオプトインが望ましい。細やかな形態はサービスの内容によるが、ホストの情報をもとに識別をするため、識別するためのプロフィールが必要となる。このプロフィールが人によってはプライバシーの脅威となるため、オプトインにより事前に同意を得てサービスを運用することが重要である。

7.4 まとめ

本章では、本論文が提案するプラットフォームが実社会において利用された場合のプライバシーの影響や懸念について述べた。利用形態には、オプトイン、オプトアウト、ノンプライバシーの 3 形態があることを定義した上で、どのような利用形態が望ましいかを OECD ガイドラインと比較した結果、オプトインが望ましいことが分かった。また、本提案プラットフォームが将来的にどのようなサービスに応用されるかを述べ、あるべき利用形態についてまとめた。

第8章 結論

8.1 本論文のまとめ

本論文の目的は、許諾を得たユーザのホストが普段ネットワークに発信する情報を収集することによって、ホストやユーザを識別するプラットフォームを提案し、設計実装した。これによって、サービス提供者はユーザに合わせたサービスの提供ができ、ユーザも意識せずにサービス提供が可能となるプラットフォームを実現した。

既存のユーザ、ホストの識別手法において識別が容易にできない理由として大きく2つ挙げられる。1つ目は汎用性の欠如の問題である。Webサービスによるアカウント登録手法や、専用機器によるホスト識別手法では、ほとんどの場合で他のサービスに利用できず、対応サービスごとにアカウントや機器が必要となるため、汎用性が高いとは言えない。2つ目は運用コストの増加である。ホストやユーザを識別するために、GPSやRFIDなどの専用機器を付与することがあるが、これらはホストやユーザごとに配布する必要があり、コストが高くなる。また、サービスを受けるユーザはアカウント登録や、常に専用機器を持ち歩く、もしくはホストにインストールする必要があるという問題点がある。

本論文ではこれらの問題を解決するために、普段ホストが発信している情報をもとにホストとユーザを識別するユーザ識別プラットフォームを提案した。プラットフォームは3つの手法で構成される。その3つの手法は**機器情報を用いたホスト識別手法**、**ネットワークトラフィックを用いたホスト識別手法**と、**在席情報によるユーザ識別手法**である。

機器情報を用いたホスト識別手法は、一般ユーザのホストがセンサとなることでネットワーク上のブロードキャスト、マルチキャスト情報を収集し解析することでホストを識別する。実際に数千人以上が参加するネットワークにおいて、実験をした結果、ネットワークの構成やセンサの設置場所によって異なるが十分にホストが識別可能であることを示した。その反面、センサがネットワーク上に存在しなければ情報を取得することができず、常に監視する必要がある場合はセンサをネットワークセグメントごとに設置する必要がある。

ネットワークトラフィックを用いたホスト識別手法では、ネットワーク管理者が管理対象のネットワークにおいて、ユーザが発信するパケットヘッダ情報を用いてホストを識別する。その際に、法的観点からペイロードを取得できない場合や、ルータを経由してMACアドレスが変化する場合があるため、この2つの情報は除外してホストを識別する。この手法で識別可能かを様々な所属のユーザが200名以上参加するアカデミックなカンファレンスで実験した。その結果、対象6名のホストを全台発見することができたため、この手法は有効であると言える。しかし、ネットワークによって識別要素は異なるため、必要な識別要素は検証する必要がある。

在席情報によるユーザ識別手法はホストとユーザを結びつけることで、ユーザを識別する。在席情報とホストがネットワークに発信する情報を組み合わせて、ホストの利用者を絞り込む。この手法で実験した所 4 回から 9 回程度の試行でユーザを特定することができた。そして、ベンダ情報を組み合わせることで、2 回から 6 回程度と試行回数を短縮できることが分かった。また、SNS などでユーザが発信した情報とホスト情報を組み合わせて絞り込む方法においては、確実に 1 台まで絞り込むことはできなかったが、条件に応じて識別できることを示した。

本論文では、ユーザ識別プラットフォーム作成のために、機器情報を用いたホスト識別手法、ネットワークトラフィックを用いたホスト識別手法、在席情報によるユーザ識別手法の 3 つを提案し、それらがホストとユーザを識別できることを示した。そして、このプラットフォームを用いることで、サービス提供者はユーザに意識させることなくホストやユーザの識別が可能であることを示した。

本手法によってサービスを容易に提供できるようになった反面、第 7 章で述べたように、ユーザプライバシーに配慮する必要がある。本手法により、パケットヘッダ情報のみでホストの追跡が可能であるため、トラフィックデータの収集や譲渡の際にはユーザの同意を得る、オプトイン形式にするなどサービス提供者はユーザプライバシーの配慮などが求められるため、今後利用形態においては議論を続ける必要がある。

8.2 今後の課題と展望

8.2.1 情報にアクセスするユーザの認証手法の確立

デバイス情報を用いたホストの識別手法において、センサとなるクライアントの情報を保護する必要がある。情報を収集する上で、センサとなるクライアントはかなり正確な情報が取得可能であるため、取得されるホストと比べてプライバシーの影響を受けやすい。現状においては、センサとなるホストとその他のホストの取得情報に差異はないが、長期的な収集、解析においてセンサとなるクライアントは精密なデータが利用者のプライバシーを脅かす可能性がある。そのため、情報の収集解時にプライバシー問題の解決する必要がある。解決手法の 1 つとして、匿名化する手法が挙げられる。これは、情報の発信者を特定できない、もしくは MAC アドレスを表示せずに情報を収集、解析する手法であり、token の利用例としては、図書館における匿名での本を貸出システムや、電子投票システムで利用されているブラインド署名 [48] 及びゼロ知識証明 [49] を利用することによって、プライバシーの問題を解決、軽減できる。また、運用形態などプライバシーに関しての問題は残るため、十分な議論が必要である。

8.2.2 情報のオプトアウト手法の確立

本手法をオプトアウトのサービス形態で運用場合などを想定した場合、サービスに同意をしないユーザを考慮しなければならない。情報の収集を拒否するユーザの対応とし

て、本プログラムをインストールしたホスト以外はブロードキャストを取らない方法以外に、特定の MAC アドレスは取得しないという方法がある。例えば、Google 社の提供するサービスで AP 情報を収集している。しかし、オプトアウトを望むユーザに関しては AP の SSID に特定の文字列を入れることで、対象は情報の収集対象とならない。この方法と同じく MAC アドレスを特定の文字列に変更することで収集対象としない手法が考えられる。しかし、MAC アドレスの変更は比較的容易であるがユーザにとって負担となってしまうため、十分なオプトアウトの機構を検討する必要がある。

8.2.3 パケットヘッダ情報を用いた識別手法における識別要素の追加

本手法が対象とするアプリケーションの追加や、特徴の抽出を更に細かく検討する必要がある。また、サービスを特定するために特定サービスのアドレスブロックを保持しているが、アドレス範囲が不明な場合に、本システムの識別要素として利用している情報が取得できない場合が考えられる。そのため、今後サービスごとにトラフィック量や通信パターンからサービスを推測する必要がある。このように今後、様々なホストの特定ができるように、より多くの識別要素を検討する。

8.2.4 ユーザ特定におけるプライバシーの数理モデル化

本提案プラットフォームにおいて、どこまで個人に関する情報が収集されればプライバシーを脅かされるのかという指針が必要である。例えば、第 6 章で述べたホストとユーザを結びつける手法において、誤認情報の扱いについて検討し、ケースごとに特定される確率を提示することで、サービス提供側に置いてはホスト識別の正確さが判明する。また、匿名を希望するユーザはどの程度まで取得されたらプライバシーが脅かされるのかを知ること、どこまで情報を提供してよいのかという指針となる。そのためにも、ホストやユーザを結びつけるための数理モデル化が必要である。

謝辞

本研究を進めるにあたり、主査である慶應義塾大学環境情報学部教授村井純博士に感謝致します。また、副査である慶應義塾大学環境情報学部教授中村修博士、同学部教授武田圭史博士に感謝致します。

モバイル広域ネットワークプロジェクトにおいて、ご指導とご助言を頂きました、慶應義塾大学環境情報学部准教授楠本博之博士、同学部准教授三次仁博士、同学部准教授植原啓介博士、同学部准教授 Rodney D. Van Meter III 博士、同大学政策・メディア研究科特任講師斉藤賢爾博士、同大学政策・メディア研究科特別研究講師吉藤英明博士、同大学政策・メディア研究科特別研究講師佐藤雅明博士、同大学 SFC 研究所上席所員水谷正慶博士に感謝致します。

特に、水谷正慶博士と武田圭史博士に感謝致します。水谷正慶博士は私が学部生の頃から私の研究を見ていただき、大学を離れてからも私の研究相談にのって頂きました。そして、武田圭史博士は非常に多忙の中、多くの時間を割いていただきました。この方々なくしては論文誌採録や修士論文を執筆することはできませんでした。感謝致します。

慶應義塾大学大学院政策・メディア研究科後期博士課程岡田耕司氏、堀場勝広氏、空閑洋平氏、松谷健史氏、鈴木詩織氏、久松剛氏、松園和久氏、三島和宏氏に感謝致します。同研究科修士課程佐藤弘崇氏、三部剛義氏、米村茂氏、宮崎圭太氏、澤井暖氏に感謝致します。同大学 SFC 研究所所員 (訪問) の梅田昂翔氏、福岡英哲氏、村井研究室 OB である金井瑛氏、中村友一氏、黒宮佑介氏、峯木厳氏、六田佳祐氏に感謝致します。特に OB の 5 名と博士過程の方々には学部の頃から非常にお世話になりました。重ねて感謝致します。

実験に協力いただいた同大学総合政策学部及び環境情報学部にも所属、鴻野弘明氏、相見眞男、碓井利宣、Vu Xuan Duong 氏、佐藤文啓氏、関根冬輝氏、Pham Van Hung 氏、山本知典氏、吉原洋樹氏、有馬怜文氏、大矢崇央氏、鴻野弘明氏、小松真氏、高岡賢二氏、中島明日香女史、由井卓哉氏、吉原大道氏、露木航平氏、中安恒樹氏、深谷哲史氏、吉田裕氏、小澤麗女史、川本卓弥氏、廣田一貴氏に感謝致します。そして、研究室で充実した時間を共にした、同環境情報学部中村遼氏、澁田拓也氏、横石雄大氏、佐藤友紀氏、徳田・村井・楠本・中村・高汐・重近・バンミーター・植原・三次・中澤・武田合同研究プロジェクトの皆様に感謝致します。

私と共に修士論文を執筆した、慶應義塾大学大学院政策・メディア研究科永山翔太氏、波多野敏明氏、重松邦彦氏、山口修平氏に感謝致します。彼らと一緒に歩めたからこそ最後まで諦めず執筆をできたと確信しています。

私とフラメンコを引き合わせてくれた SFC スペイン舞踊部の面々に深く感謝致します。フラメンコによって多くのことを学び、ここで得た経験や出会いが日常だけでなく修士論文最終発表の場などでも、私を支える原動力となりました。また、社会の厳しさと人の面

白さ，付き合い方を教えてくれた劇団ひまわりに所属する面々にも同じく感謝致します。
最後に，大学入学から大学院修了の 6 年間に渡る研究生活で，私を支え続けてくれた家族に心から感謝致します。以上を持って謝辞と致します。

参考文献

- [1] 財団法人インターネット協会. **インターネット白書 2008**. 株式会社インプレス R&D, 6 2008.
- [2] Inc. Amazon.com. Amazon.co.jp. <http://www.amazon.co.jp/>.
- [3] Jim Gemmell. Clean living: Eliminating near-duplicates in lifetime personal storage. *MSR-TR-2006-30*, 2006.
- [4] foursquare. foursquare. <https://foursquare.com/>, 1 2012.
- [5] 山崎賢児 and 勅使河原海. 高度なパーソナライズ実現のための統合サービスエージェントの設計. *IPSJ SIG Technical Report*, pages 105–110, 3 2005.
- [6] 衣袋宏美. ネット視聴率白書 2008-2009. 翔泳社, 2008.
- [7] NebAD Inc. *NebAD Inc.* <http://www.nebuad.com>, 6 2008.
- [8] 生貝直人. **情報社会と共同規制**. 勁草書房, 2011.
- [9] Annie De Montigny-Leboeuf. Passive network discovery for real time situation awareness. In *In Proceedings of the The RTO Information Systems Technology Panel (IST) Symposium on Adaptive Defence in Unclassified Networks*, pages 288–300, 4 2004.
- [10] Thomas Karagiannis, Konstantina Papagiannaki, and Michalis Faloutsos. Blinc: Multilevel traffic classification in the dark. In *In Proceedings of ACM SIGCOMM*, pages 229–240, 2005.
- [11] Dan Rosenberg. It’s bugs all the way down. <http://vulnfactory.org/blog/2011/12/05/carrieriq-the-real-story/>, 12 2011.
- [12] Mozilla Japan. 次世代ブラウザ firefox. <http://mozilla.jp/firefox/>, 1 2012.
- [13] Adobe Systems Incorporated. Adobe. <http://www.adobe.com/>, 12 2011.
- [14] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris J. Hoofnagle. Flash cookies and privacy. *SSRN eLibrary*, 2009.
- [15] Quantcast Corporation. Quantcast - home. <http://www.quantcast.com/>, 12 2011.

- [16] CBS Interactive. Exclusive: Google's web mapping can track your phone. http://news.cnet.com/8301-31921_3-20070742-281/exclusive-googles-web-mapping-can-track-your-phone/, 2011.
- [17] INC NTT DOCOMO. 『i モード id』の提供開始について. http://www.nttdocomo.co.jp/info/notice/page/080228_00.html, 12 2011.
- [18] Samuel D. vWarren and Louis D. Brandeis. The right to privacy. *Harward Law Review*, 4(5):193–220, December 1890.
- [19] JIPDEC. プライバシマーク制度. <http://privacymark.jp/index.html>, 12 2009.
- [20] Pala Inc. ぶららホームページ. <http://www.plala.or.jp>, 11 2009.
- [21] NTTPlala. ニュースリリース (2006.6.13). http://www.plala.or.jp/member/option_service/secuplus/nbb/start_wfs.html, 12 2009.
- [22] 村本俊祐, 上土井陽子, and 若林真一. k-匿名性を利用したデータ一般化によるプライバシー保護. *DEWS2007*, 2007.
- [23] 総務省. **電波法 第 5 章**, 12 2011.
- [24] Uniting and strengthening america by providing appropriate tools required to intercept and obstruct terrorism (usa patriot act) act of 2001. http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ056.107, 10 2001.
- [25] Google. Google. <http://www.google.com>, 12 2010.
- [26] 和泉 勇治, 佐藤 剛, and 田中 和之. パケットペイロード長の発生確率に基づいたネットワークアプリケーション識別 (bci/bmi その周辺, 一般). **電子情報通信学会技術研究報告. NC, ニューロコンピューティング**, 110(295):7–12, 2010-11-11.
- [27] B. Krishnamurthy and C.E. Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 7–12. ACM, 2009.
- [28] Electronic Frontier Foundation. How unique is your browser? *Privacy Enhancing Technologies Symposium*, 2010.
- [29] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, and 石塚 満. Web 上の情報からの人間関係ネットワークの抽出. **人工知能学会論文誌 = Transactions of the Japanese Society for Artificial Intelligence : AI**, 20:46–56, 20051101.

- [30] 本村 憲史 and 金田 重郎. ネットワーク上での情報統合によるプライバシー侵害とその対策. **電子情報通信学会技術研究報告**. *OFS, オフィスシステム*, 98(75):29–36, 5 1998.
- [31] 大規模キャンパスネットワークにおける mac アドレス認証の管理手法 (パラレル, インターネットと情報倫理教育, 一般). **電子情報通信学会技術研究報告**. *IA, インターネットアーキテクチャ*, 108(460):265–270, 2009-02-26.
- [32] japan.internet.com. Bluetooth に関する調査. <http://japan.internet.com/research/member/>, 6 2010.
- [33] CACE Technologies Inc. Cace technologies - aircap. <http://www.cacetech.com/products/airpcap.html>, 12 2011.
- [34] SOFT BANK TEKECOM. Bb モバイルポイント 外出先でブロードバンドするなら bb モバイルポイント. <http://tm.softbank.jp/wlan/index.html>, 12 2011.
- [35] 総務省. 利用者視点を踏まえた ict サービスに係る諸問題に関する研究会第二次提言. http://www.soumu.go.jp/main_content/000062003.pdf, 5 2010.
- [36] M. Zalewski and OS Passive. Fingerprinting tool. <http://lcamtuf.coredump.cx/p0f.shtml>, 1 2010.
- [37] dokodoko.jp. dokodoko.jp. <http://dokodoko.jp/>, 5 2011.
- [38] Evernote Corporation. Welcome to your notable world — evernote corporation. <http://www.evernote.com/>, 1 2010.
- [39] Dropbox. Dropbox. <https://www.dropbox.com/>, 7 2010.
- [40] Kenjiro Cho, Kensuke Fukuda, Hiroshi Esaki, and Akira Kato. Observing slow crustal movement in residential user traffic. In *Proceedings of the 2008 ACM CoNEXT Conference*, CoNEXT '08, pages 12:1–12:12, New York, NY, USA, 2008. ACM.
- [41] 財団法人インターネット協会. **インターネット白書 2009**. 株式会社インプレス R&D, 6 2009.
- [42] 上原雄貴, 水谷正慶, 武田圭史, and 村井純. デバイス情報収集によるユーザ追跡システムの試作. **コンピュータセキュリティシンポジウム 2010**, 2010:837–842, 10 2010.
- [43] NPO 日本ネットワークセキュリティ協会. **2009 年情報セキュリティインシデントに関する調査報告書**, 7 2010.
- [44] (独) 情報処理推進機構セキュリティセンター. eid に対するセキュリティとプライバシーに関するリスク認知と受容の調査報告, 8 2010.

- [45] 佐藤 雅明. **インターネット上での自動車情報基盤の構築**. PhD thesis, 慶應義塾大学 政策・メディア研究科, 2008.
- [46] OECD RECOMMENDATION CONCERNING AND GUIDELINES GOVERNING THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA, 10 1980. O.E.C.D. Document C(80)58(Final).
- [47] Google. Google latitude. <http://www.google.co.jp/mobile/latitude/>, 1 2012.
- [48] David Chaum. Blind signatures for untraceable payments. In *CRYPTO*, pages 199–203, 1982.
- [49] S Goldwasser, S Micali, and C Rackoff. The knowledge complexity of interactive proof-systems. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, STOC '85, pages 291–304. ACM, 1985.