

All rights are reserved and copyright of this manuscript belongs to the authors. This manuscript have been published without reviewing and editing as received from the authors: posting the manuscript to SCIS 2007 does not prevent future submissions to any journals or conferences with proceedings.

SCIS 2007 The 2007 Symposium on
Cryptography and Information Security
Sasebo, Japan, Jan. 23-26, 2007
The Institute of Electronics,
Information and Communication Engineers

P2P ファイル交換ソフトウェア環境を対象とした観測に関する一考察 Feasibility Study for the observation of P2P file exchange environment

寺田真敏*
Masato Terada

鵜飼裕司†
Yuji Ukai

金居良治†
Ryoji Kanai

土居範久**
Norihisa Doi

あらまし P2P (Peer to Peer) ファイル交換ソフトウェア利用が広がる中, その利用実態に関する調査報告は数少ない. 本調査研究の目的は, P2P ファイル交換ソフトウェア環境の利用実態に関する具体的な調査方法を示すこと, その方法に基づき調査した結果を示すことで, 調査手法の有効性を示すと共に, P2P ファイル交換ソフトウェア環境の利用実態に関する調査活動を支援することにある. 本稿では, P2P ファイル交換ソフトウェア環境の調査研究事例として, Winnybot と呼ぶツールを用いて収集したデータを元に Winny 稼動ノード数, ファイルのプロパティが格納されている Winny のキー流通量について報告する.

キーワード P2P, ファイル交換ソフトウェア, 観測, Winny

1 はじめに

P2P (Peer to Peer) ファイル交換ソフトウェア利用が広がる中, その利用実態に関する調査報告は数少ない. この背景には, P2P ファイル交換ソフトウェアが水平型の広域ネットワーク構成をとること, 固定ポート番号を利用した通信ではないこと, 実態把握に必要な手法が確立していないことなど, 調査活動を阻む壁が多いことにも起因していると思われる.

本調査研究の目的は, P2P ファイル交換ソフトウェア環境の利用実態に関する具体的な調査方法を示すこと, その方法に基づき調査した結果を示すことで, 調査手法の有効性を示すと共に, P2P ファイル交換ソフトウェア環境の利用実態に関する調査活動を支援することにある. 本稿では, 調査研究の事例として, Winnybot [1] と呼ぶツールを用いて収集したデータを元に Winny 稼動ノード数, ファイルのプロパティが格納されている Winny のキー流通量について報告する.

なお, 本調査事例は, 社団法人コンピュータソフトウェア著作権協会から「Winnybot を用いた Winny ネットワークの調査」の依頼を受け, 提供された観測データを元に中央大学において調査作業を実施した.

2 関連研究

本章では, P2P ファイル交換ソフトウェア環境の観測対象と, 観測対象に関して, これまでに報告されている関連研究ならびに調査報告について述べる.

(1) トラフィック

NetFlow を用いて一日当たりのユーザの送受信トラフィック量を測定した文献 2) (調査時期: 2004 年 4, 5, 10 月, 2005 年 2, 7 月)によれば, 1 日当たり 2.5GB 以上のトラフィックを発生させているヘビーユーザは一定割合で存在し, 少数のヘビーユーザが帯域を使用している. また, データセンタからユーザへのトラフィックは 16.7%, ユーザ間トラフィックは 62.2% であり, ポート番号 1024 以上のトラフィックが 82% を占めていることから, P2P によるトラフィックが帯域を占めている可能性を示唆している. しかし, NetFlow はポート番号レベルのトラフィック量測定に留まり, P2P によるトラフィックの断定まではできない. 文献 3) 4) では, トラフィックが P2P であることを判定するため, トラフィックの特徴量やピア間通信のサーバ/クライアント関係を用いたトラフィックの特定方法を提案している.

(2) ノード数

全ノード数把握が困難な Winny ネットワークの規模を推定するため, 文献 5) では, 実測定によって得られた Winny の通信データをもとにコンピュータシミュレーション実験を行い, Winny ネットワークの規模推定する方法を提案している. また, 文献 6) (調査時期: 2006 年 8 月)では, 独自の手法を実装した 11 台の観測装置によって得られた Winny ノード情報から一意なノードを

* 中央大学研究開発機構, 〒112-8551 東京都文京区春日 1-13-27, Research and Development Initiative Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551 Japan.

** 中央大学大学院 理工学研究科, 〒112-8551 東京都文京区春日 1-13-27, Graduate School of Science Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551 Japan.

† eEye Digital Security, 1 Columbia, Aliso Viejo, California 92656, United States

算出し、“平日で 39 万から多い日では 41 万、土日になると 43 万から 44 万以上のノード数”を Winny ノード数の推移として報告している。

(3) コンテンツの流通状況

文献 7) (調査時期：2006 年 6 月)では、約 18,000 名からのアンケート調査結果に基づき、ファイル交換ソフトウェアの利用者推計値は 176 万人で、音楽、映像などのファイル交換に利用されていると報告している [8]。また、文献 9) (調査時期：2003 年 4 月)では、WinMX, Gnutella, Winny を対象としたコンテンツ分析を行っており、Winny では、全ファイル数が 23 万、ファイルサイズが約 63MB, avi や mpg などの動画像, zip や rar といった圧縮ファイルが流通していることを報告している。

3 Winny ネットワークの観測環境

本章では、社団法人コンピュータソフトウェア著作権協会が調査にあたり稼働させた Winny ネットワークの観測環境について述べる。

3.1 用語定義

観測環境ならびに、調査結果の報告に先立ち、本稿で使用する用語を説明する。

- ノード
IP アドレスを用いた Winny ノードの識別子であり、Winny ノード数算出に使用する。
- プライマリキー
“IP アドレス+ポート番号+ファイルのハッシュ値”から構成したキーの識別子であり、本稿では Winny ノードを加味したファイルの識別子としても使用する。キーならびにファイルの流通量算出に使用する。
- ハッシュキー
ファイルのハッシュ値を用いた識別子であり、本稿では重複のないファイル数、すなわち、一意なオリジナルファイル数の算出に使用する。
- 仮想キー
仮想キーは Winny で定義された用語であり、キーの拡散や検索クエリによって取得したキーを Winny ノード内に持っているが、対応するキャッシュファイル（キャッシュブロック保有率が 0%）を持たないキーのことである。

3.2 観測ツール Winnybot

Winny ネットワークの観測に利用された観測ツール Winnybot のモジュール構成を図 1 に示す。観測エンジンは、Winny ネットワーク上のいずれかのノードに接続した後、拡散クエリ送信要求（コマンド 0x0A）を送信し、クエリ送信（コマンド 0x0D）を受信する操作を繰り返すことで、Winny ネットワーク上で交換されている

ファイルのプロパティが含まれているキー情報を網羅的に収集する。また、データベース格納スクリプトは、観測エンジンが収集したキー情報を定期的にデータベースに格納する。

表 1 にキー情報として観測可能な項目を示す。なお、観測エンジンは、キー情報以外に Winny ノードの参照関係のプロパティが含まれているノード情報の収集機能も備えているが、本調査では、キー情報のみを使用した。

3.3 調査に用いた観測環境

本調査に用いた Winny ネットワークの観測環境は、観測ツール Winnybot が搭載された 9 台の観測装置（装置名 M1 から M9）から構成されている。各観測装置のデータ収集性能については、観測期間中の 1 時間あたりのプライマリキー平均収集件数を用いて示す（図 2）。また、図 2 には観測エンジンが収集するデータ重複度を示すため、1 時間あたりに収集した一意なプライマリキーの平均収集件数を併記した。

4 Winny ネットワークの調査

本章では、前述の観測環境を用いて調査した Winny 稼働ノード数、ファイル数、ファイルサイズとキー流通量について述べる。

4.1 調査方法

(1) 観測期間

2006 年 10 月 6 日(金)~10 月 15 日(日)の 10 日間

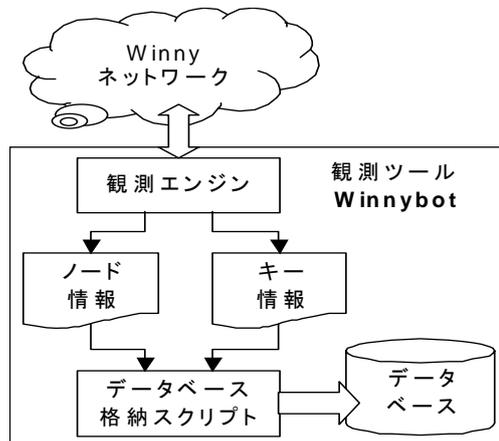


図 1 観測ツール Winnybot のモジュール構成

表 1 キー情報として観測可能な項目

項目
キーの IP アドレス
キーのポート番号
BBS スレッド管理ノードの IP アドレス
BBS スレッド管理ノードのポート番号
ファイルサイズ
キー更新日時
被参照ブロック数
キー消滅判定タイマー
ファイルのハッシュ値
トリップ情報

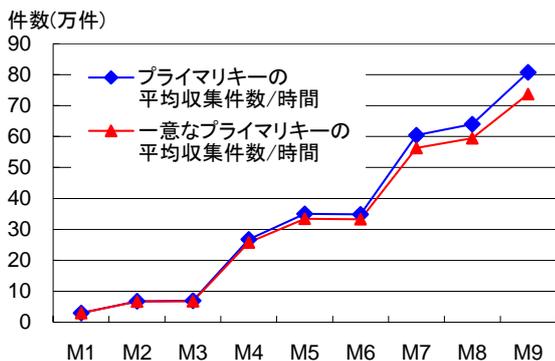


図 2 各観測装置のデータ収集性能

(2) 観測装置

観測期間中、4.3 節に示す 9 台の観測装置が常時稼動し、Winy ネットワーク上で交換されているファイルのプロパティが含まれているキー情報の収集が行われた。提供された観測データは、平均 7,100 万件/日、計約 7 億件のキー情報である (図 3)。

(3) 推定方法

Winy ノードの稼動数、キーならびにファイルの流通量算出にあたっては、単位時間に、観測装置台数の増加に伴って発生する一意な観測件数の収束状況を用いて算出した。ここで単位時間とは、観測件数の集約期間である。例えば、単位時間が 24 時間の場合、24 時間の間に観測したデータ全体を算出対象とすることを意味する。単位時間 24 時間とした Winy ノードの稼動数算出を例にとると、24 時間の間に観測したデータ全体を対象に、観測装置毎に観測した一意なノード数を足し合わせた総計 (以降、単純総計と呼ぶ) と観測装置全体で観測した一意なノード数の累積 (以降、累積総計と呼ぶ) とを組み合わせることで、ノードの一意な観測件数の収束状況を得て、その収束点を Winy ノードの稼動数とする。すなわち、単純総計に対して累積総数が収束した件数が Winy ノードの稼動数、キーならびにファイルの流通量となる (図 4)。

4.2 調査結果

(1) Winy ノードの稼動数

Winy ノードの稼動数については、表 1 の“キーの IP アドレス”を観測項目とし、24 時間を単位時間として、単純総計と累積総計との観測件数の収束状況から算出した。なお、単純総計のプロットにあたっては、各観測装置のデータ収集性能を考慮し、収集能力の高い装置 M9 から M1 の順に一意なノード数の加算を行なっている。

収束判定の結果を図 5 に示す。これによれば、Winy ノードの稼動数は、1 日あたり約 35~40 万ノードである。また、単位時間を 3, 6, 12, 24 時間に変えた場合、ノード数の収束件数は、ほぼ線形に増加していることがわかる (図 6)。

(2) 一意なオリジナルファイル数

Winy ネットワーク上に存在する一意なオリジナルファイル数については、表 1 の“ファイルのハッシュ値”を観測項目とし、24 時間を単位時間として、単純総計と累積総計との観測件数の収束状況から算出した。

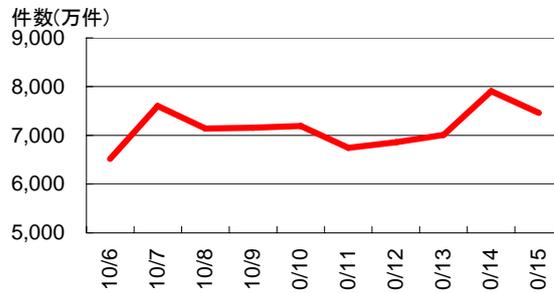


図 3 観測期間中のキー観測総数

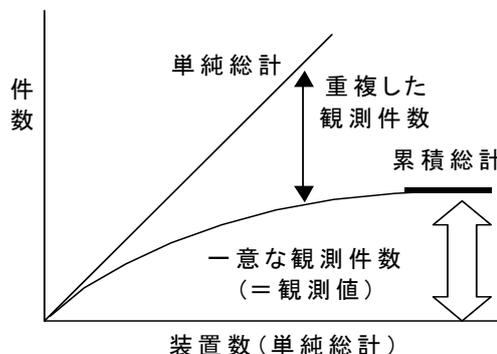


図 4 単純総計と累積総計とを用いた収束判定

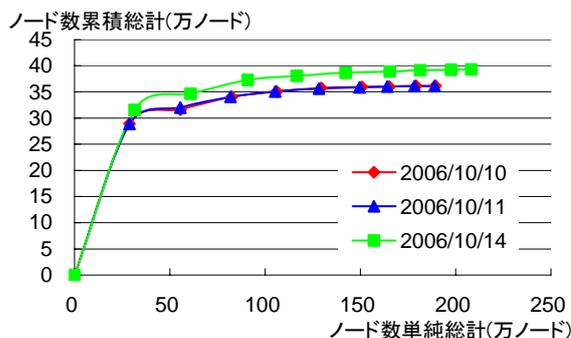


図 5 キー情報から算出した Winy ノード数

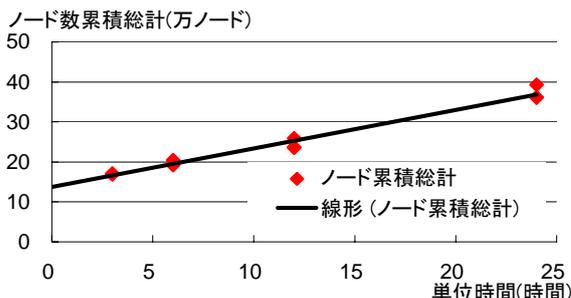


図 6 単位時間毎の Winy ノードの累積総計

収束判定の結果を図 7 に示す。これによれば、Winny ネットワーク上に存在する一意なオリジナルファイル数は 1 日あたり約 450 万～500 万ファイル (平均すると約 470 万ファイル) である。

(3) 総ファイルサイズ

Winny ネットワーク上に存在する一意なオリジナルファイルの総ファイルサイズについては、表 1 の“ファイルのハッシュ値”とファイルのハッシュ値に対応する“ファイルサイズ”を観測項目とし、24 時間を単位時間として、ハッシュキーによる単純総計と、ファイルサイズによる累積総計の収束状況から算出した。なお、表 1 で得られるファイルサイズがマイナスの場合には除外し、それ以外については正しい値であると仮定し算出している。収束判定の結果を図 8 に示す。これによれば、Winny ネットワーク上に存在する一意なオリジナルファイルの総ファイルサイズは、1 日あたり平均すると約 290 テラバイトとなる。また、オリジナルファイル数が 1 日あたり約 470 万ファイルであると仮定すると、平均ファイルサイズは、290 テラバイト÷470 万ファイル=約 61 メガバイトとなる。なお、ファイルサイズの分布は、図 9 に示す通り、1M バイト以上 10M バイト以下のファイル数が多い。

(4) プライマリキーの流通量

プライマリキーの流通量は、Winny ネットワーク上のファイル交換に關与するキー情報の流通量であり、仮想キーと Winny ノードがキャッシュブロックを保有するキーの総計である。ここで、キャッシュブロックを保有するキーには、キャッシュブロックが完全には揃っていない状態のファイルである部分キャッシュ、キャッシュブロックが全て揃ったファイルである完全キャッシュとアップロードフォルダに格納されているファイルに関するキーが存在する。

Winny ノードを加味したファイル識別子であるプライマリキーの流通量については、表 1 の“キーの IP アドレス+キーのポート番号+ファイルのハッシュ値”を観測項目とし、6 (18 時～24 時)、24 時間を単位時間として、単純総計と累積総計との観測件数の収束状況から算出した。収束判定の結果を図 10、図 11 に示す。

この収束結果によれば、1 日あたりの一意なプライマリキー (=仮想キー+Winny ノードがキャッシュブロックを保有するキー) の流通量は、平均すると約 3,400 万件である。また、単位時間を 3, 6, 12, 24 時間に変えた場合、プライマリキーの流通量は、ほぼ線形に増加しており (図 12)、単位時間 12 時間以降は、約 40%が重複観測データであることがわかる (図 13)。

4.3 考察

(1) 観測装置の観測網羅性

Winny ノードの稼働数 (図 5)、Winny ネットワーク上に存在する一意なオリジナルファイル数 (図 7) とプライマリキーの流通量 (図 11) のいずれも、単純総計と累積総計の近似曲線が装置台数増加と共に収束してい

ることから、複数の観測装置による観測網羅性は高いと判断できる。

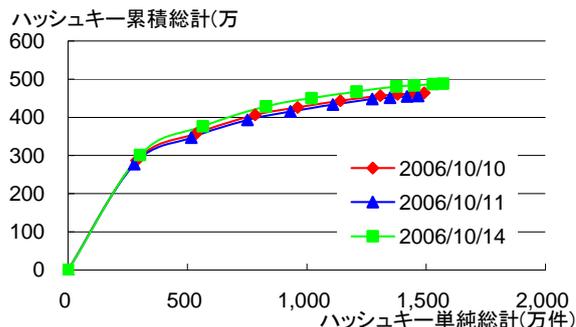


図 7 ハッシュキーの累積総計 (24 時間)

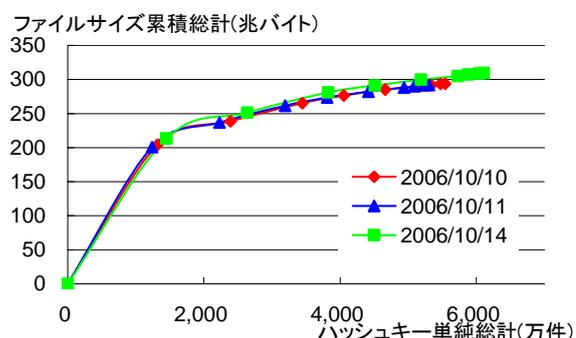


図 8 ファイルサイズの累積総計 (24 時間)

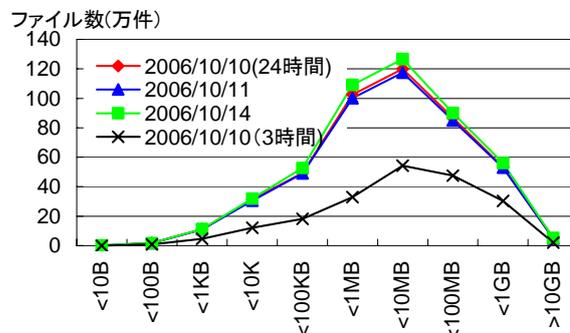


図 9 ファイルサイズの分布 (24 時間)

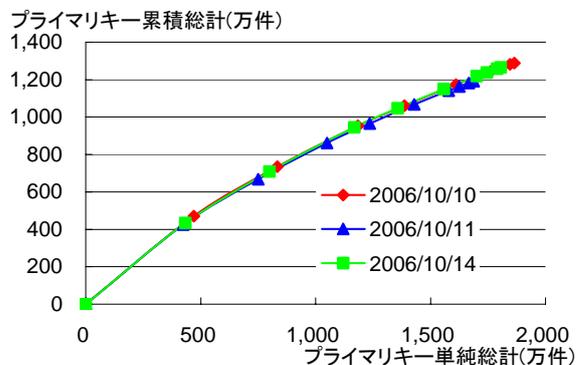


図 10 プライマリキーの累積総計 (6 時間)

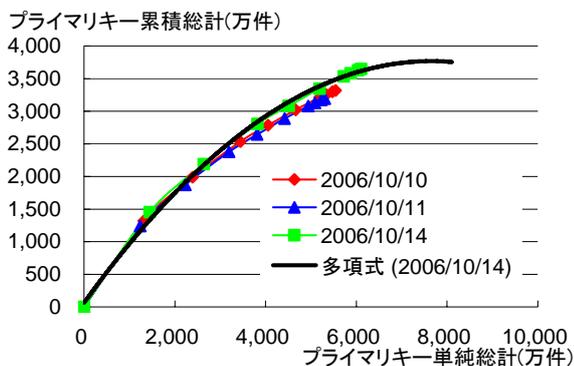


図 11 プライマリキーの累積総計 (24 時間)

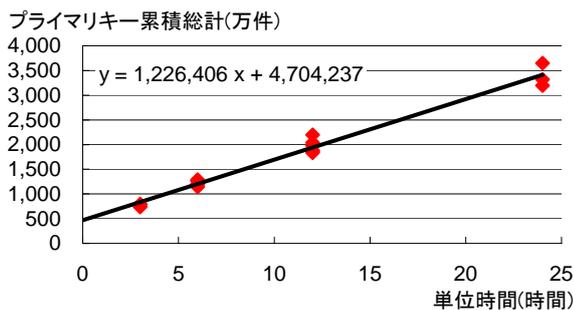


図 12 単位時間毎のプライマリキーの累積総計

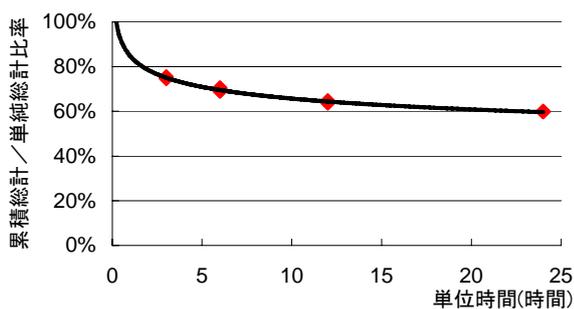


図 13 単位時間毎のプライマリキーの累積総計 / 単純総計の比率

また、この調査結果から、プライマリキーの流通量については、毎時 60 万キーを収集する観測装置 (例えば、M7, M8, M9) を 3 台用意すれば全流通量の約 70% (= 2500 万 ÷ 3400 万) をカバーでき、ファイル交換の状況を十分に把握できると思われる。

(2) Winny ノードが保有するファイル数

調査結果に基づき、Winny ノードが保有するファイル数を推定する。Winny ノードが保有するファイル数の推定にあたっては、上限 (単位時間あたりの一意なプライマリキーの流通量) と下限 (単位時間あたりの一意なハッシュキーの流通量) を設定できる。すなわち、実際に保有されているキャッシュブロックに関わるプライマリキーの流通量は上限と下限の間に位置し、プライマリキーの全流通量から仮想キーの流通量を差し引けば推定できる (図 14)。そこで、本調査では、参照ノードとして用意した Winny ノードの情報を保有する仮想キーの出現頻度から仮想キーの流通量を推定した。

調査期間中に、参照ノードの情報を保有する仮想キーは 1,047 件 [*] の観測され、そのうち 97% が単位時間 24 時間の出現頻度は 1 回であった (図 15)。隣接時間帯とは、4 時ならびに 5 時という隣り合った時間帯に観測されたことを意味し、この観測件数も含めると 99% が単位時間 24 時間の出現頻度は 1 回である。これに対し、キャッシュブロックが保有されているプライマリキーは単位時間 24 時間に複数回出現する傾向が見られた。図 15 の結果と仮想キーの生存期間 [**] とを考慮すると、単位時間 24 時間に 1 回しか観測できなかったプライマリキーは、キャッシュブロックが保有されたプライマリキーの可能性は低く、キー転送のみに関与している仮想キーであると推定できる (表 2)。

単位時間あたりのプライマリキー出現回数 (図 16, 図 17) に示す。出現頻度 1 回の割合は、単位時間が長くなるにつれ減るものの、プライマリキー全流通量の約 70% 前後とほぼ一定している。

表 2 に基づき推定すると、出現頻度 2 回以上のプライマリキー、すなわち、キャッシュブロックが保有されているプライマリキーは約 1,100 万件となり (図 18)、Winny ネットワーク上には約 1,100 万ファイルが存在していると推定される。平均ファイルサイズを 61 メガバイトとすると、61 メガバイト × 1,100 万ファイル = 約 670 テラバイトが Winny ネットワーク上に存在していることになる。

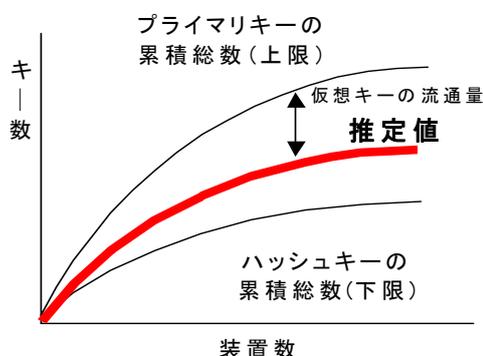


図 14 流通量の上限と下限

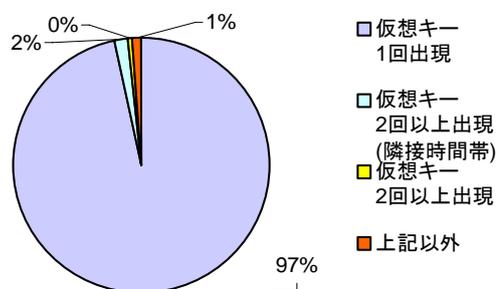


図 15 参照ノード情報を持つ仮想キーの出現頻度

*) 同一時間帯 (一時間以内) に同一の仮想キーが複数観測された場合には、一件とカウントしている。

**) 表 1 のキー消滅判定タイマーフィールドに格納された値であり、初期値は 1,500 秒。

表 2 仮想キーの判定条件

判定条件	推定結果
単位時間 24 時間にプライマリーキーを 1 回しか観測できなかった場合	左記出現頻度のプライマリーキーは、キャッシュブロックが保有されたプライマリーキーである可能性は低く、単にキー転送のみに関与している仮想キーであると推定される。
上記以外	キャッシュブロックが保有されたプライマリーキーの可能性があると推定される。

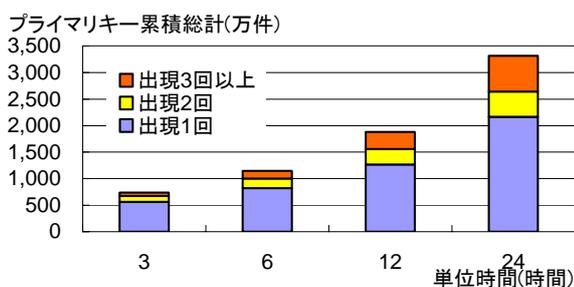


図 16 プライマリーキーの出現回数

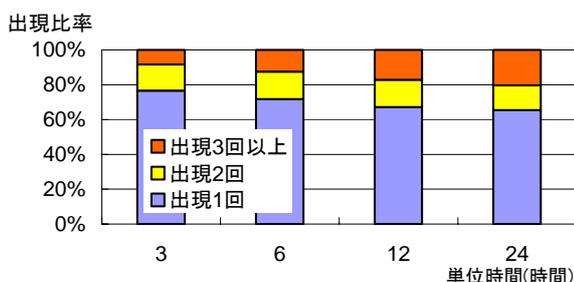


図 17 プライマリーキーの出現比率

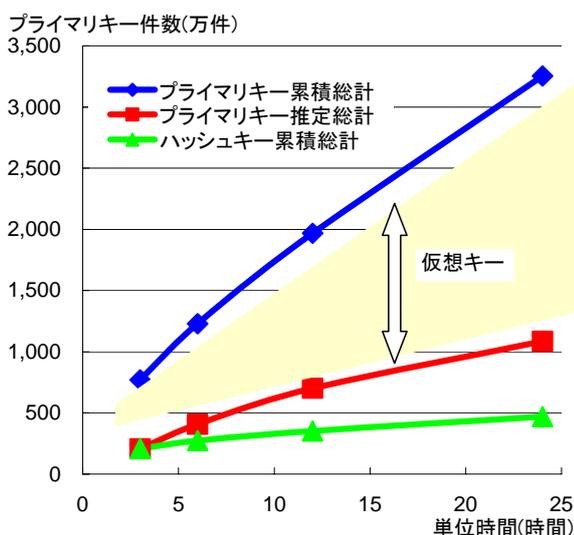


図 18 プライマリーキーの累積総計と推定総計

5 おわりに

本稿では、調査研究の事例として、Winnybot を用いて収集したデータを元に Winny 稼動ノード数、ファイ

ルのプロパティが格納されている Winny のキー流通量を推定した。

- Winny ノードの稼動数は、1 日あたり約 35~40 万ノードである。
- Winny ネットワーク上に存在する一意なオリジナルファイル数は1日あたり約 450 万~500 万ファイルである。

本稿が P2P ファイル交換ソフトウェア環境の利用実態に関する調査活動の参考になれば幸いである。今後の課題は、P2P ファイル交換ソフトウェア環境を対象とした定常的な観測手法の確立と、ノード数や流通量推定に関する手法の改善などが挙げられる。

謝辞

本調査研究は、社団法人コンピュータソフトウェア著作権協会の支援を受け実施した。本調査を進めるにあたり、有益な助言と協力を頂いた社団法人コンピュータソフトウェア著作権協会ならびに、関係者各位に深く感謝致します。

参考文献

- [1] 鶴飼裕司, “Inside Winny ~ Winny の解析とそのセキュリティ脅威分析”, <https://sec.scs.co.jp/eeeye/shiryo.html>
- [2] 福田健介, “ブロードバンドトラフィック分析”, <http://www.ndrc.kyutech.ac.jp/interop06/>
- [3] 大坐畠智 他, “パッシブ/アクティブ検知を用いた P2P トラフィック特定法”, 情報処理学会研究報告 分散システム/インターネット運用技術 Vol.2005 No.31
- [4] 松田崇 他, “P2P 弁別のためのトラフィック特徴量の提案”, 電子情報通信学会技術研究報告 Vol.105, No.12 (2005)
- [5] 蜂須賀大紀 他, “ピュア P2P ネットワーク構成ピア数推定法の一検討”, 電子情報通信学会技術研究報告 Vol.105, No.12 (2005)
- [6] NetAgent, “Winny ノード数の推移”, <http://www.onepointwall.jp/winny/winny-node.html>
- [7] 社団法人コンピュータソフトウェア著作権協会, “「ファイル交換ソフト利用実態調査」結果発表”(2006/7/25), <http://www2.accsjp.or.jp/topics/release1.html>
- [8] 社団法人コンピュータソフトウェア著作権協会, “「Winny」による被害相当額”(2006/11/28), <http://www2.accsjp.or.jp/news/release061128.html>
- [9] 大井恵太 他, “P2P ファイル共有におけるコンテンツ分析”, 情報処理学会研究報告 マルチメディア通信と分散処理 Vol.2003 No.87