2005－DSM－39（2）
2005／10／14

トラフィック特徴強化と可視化による P2P ファイル共有通信検出支援システムの構築

戸川 聡 *　　金西計英 **　　矢野米雄 ***

* 徳島大学大学院工学研究科
** 徳島大学高度情報化基盤センター
*** 徳島大学工学部

概要：Peer-to-Peer（P2P）型通信によるファイル共有が問題となっている．著作権法で保護された著作物をデータファイル化し共有することも当然ながら，ファイル共有による機密情報流出も問題となっている．これらの状況から，大学や企業のキャンパスネットワークでは P2P ファイル共有を禁止している．しかし現実には P2P ファイル共有が行われている場合がある．また，既存のフィルタリング技術では P2P ファイル共有を制限することは難しい．従って管理者はトラフィックを常時監視し，P2P ファイル共有通信の存在を認識しなければならない．本稿では，ネットワーク管理者が行う P2P ファイル共有通信の検出作業を支援するシステムを構築した．そして，実際に P2P ファイル共有プログラムが発するトラフィックを含むトラフィックを可視化し，有効性を検証した．

Peer-to-Peer File Sharing Detection Assistance System Using the Traffic Activity Extraction and Visualization

Satoshi Togawa*, Kazuhide Kanenishi** and Yoneo Yano***

*Graduate School of Engineering, University of Tokushima
**Center for Advanced Information Technology, University of Tokushima
***Faculty of Engineering, University of Tokushima

Abstract: In this research, we have proposed the assistance system for peer-to-peer traffic detection. Recently, an illegal file has been exchanged with peer-to-peer file exchange software. These files are extracted from music CD and DVD. Most files do not obtain the copyright person's approval and are open to the public. Neither enterprise nor the Campus Network user of the university must acquire these files from the problem in morality. However, the illegal file is actually acquired via Campus Network. The network administrator should observe the users' peer-to-peer communication. In this paper, first of all, We explain a problem of peer-to-peer file sharing system. Next, we explain the assistance system for peer-to-peer file sharing traffic detection. Finally, we conclude it.

1　はじめに

Peer-to-Peer（P2P）通信によるファイル共有が問題となっている．これを実現するソフトウエアとして，WinMX[1] や Winny[2]，BitTorrent[3]，Share などがある．インターネット上にはこれらのファイル共有ソフトウエアを用いた P2P ファイル共有コミュニティが形成されている．

ファイルを P2P コミュニティに公開する行為そのものに違法性はない．しかし，著作権法保護下の音楽 CD や DVD から抽出した楽曲データや動画データをファイル化し，著作権者の許諾を得ず公開することはできない．

また最近では，P2P コミュニティへの機密情報漏洩も発生している．Antinny などのワームは，システム上の電子メールデータや文書データ，表計算データをアーカイブし，P2P コミュニティに流出させる．これが原因となり，地方自治体から個人情報が流出した事例も報告されている [4].

これらの背景から，特に企業や大学のキャンパスネットワークでは，P2P ファイル共有ソフトウエアの使用を禁止する場合が多い．しかし現状は，キャンパスネットワークから一部利用者による P2P ファイル共有が行われている．ネットワーク管理者は，キャンパスネットワークを介して行われる違法行為や機密情報漏洩の兆候をできる限り把握し，適切な対策を講じなければならない．

管理者が P2P ファイル共有の制限を試みる場合，

パケットフィルタの適用を検討できる．しかしインターネット上の P2P ノードは，不特定多数かつ可変であるため，IP アドレスによるフィルタリングは困難である．さらに Winny や Share など，自立分散ノードの集合でコミュニティを構成するものは，標準的な待機ポート番号を持たないことが多い．この結果，P2P ノードはランダムな TCP ポート番号で接続を待ち受けるため，ポート番号によるフィルタリングも困難である．管理者が利用者による P2P ファイル共有を制限しようとするならば，キャンパスネットワーク内の P2P トラフィックに気づき，個別に対応しなければならない．

そこで本稿では，トラフィック特徴強化と可視化による P2P ファイル共有通信検出支援システムを提案する．本システムは，キャンパスネットワーク内から行われる P2P ファイル共有の検出を支援する．特に，自立分散ノードで構成される Pure 型 P2P ファイル共有の検出支援を目的とする．

全クライアントから送出されるトラフィックを対象に，特定の特徴を持つトラフィックを強調する．その後，全体傾向を俯瞰可能な特徴マップを生成し，管理者に提示する．特徴マップにより異変に気づいた管理者は，対象を絞った調査を行うなど，次段階の作業に取り掛かることができる．

以下本稿では，2 章で P2P ファイル共有通信の現状について述べ，3 章でこれらファイル共有通信の検出支援モデルについて述べる．4 章で本研究で使用するトラフィックモデルの構成と可視化手法を述べ，5 章で試作システムの概要を述べ，その後実証実験の概要と考察を述べる．

2 P2P ファイル共有通信の現状

2.1 P2P ファイル共有の通信モデル

P2P ファイル共有の通信モデルは，次の 2 つに分類できる [5].

Hybrid 型： 図 1 に Hybrid 型通信モデルを示す．これは，ファイル所在情報である索引を保持するインデックスサーバと，実体ファイルを保持するノード群から構成される．あるノードがファイル入手を試みる場合，目的ファイルの所在をインデックスサーバに問い合わせる．インデックスサーバは当該ファイルの所在情報を探索元ノードに返す．探索元ノードはファイルを所有するノードとコネクションを確立し，目的ファイルを入手する．

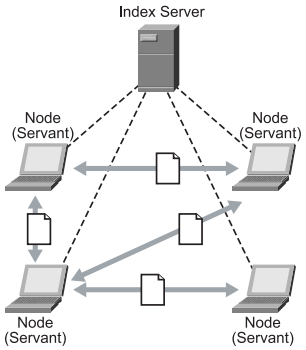

図 1: Hybrid 型 P2P ファイル共有モデル

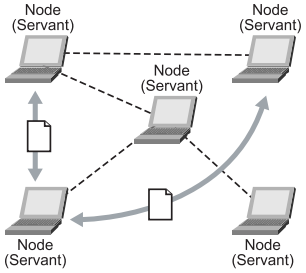

図 2: Pure 型 P2P ファイル共有モデル

Pure 型： 図 2 に Pure 型通信モデルを示す．Pure 型は索引情報を保持するインデックスサーバを持たない．ファイルやノードの探索機能はノード自体に実装される．ノードがファイルを探索する場合，近接するノードに探索要求を発行する．これを繰り返すことで探索要求は P2P コミュニティ内に伝搬する．あるノードが要求に合致するファイルを所有していた場合，そのノードはファイル所在情報を返す．所在情報を受信した探索元ノードは，ファイルを所有するノードとコネクションを確立し，実体ファイルを入手する．

Hybrid 型アプリケーション例として WinMX, BitTorrent があり，Pure 型アプリケーション例として Winny，Share を挙げることができる．

2.2 Pure 型 P2P ファイル共有の通信特性

Pure 型 P2P ファイル共有の通信特性を明らかにするため，予備実験にてトラフィック解析を行った．本実験の目的は P2P ファイル共有プロトコルを明らかにすることではない．送信元から見た Pure 型ファイル共有トラフィックの表層的振る舞いを明らかにする．

実験機に Share を導入しファイル共有トラフィックを発生させた．比較対象として，人手による Web

表 1: 予備実験での計測結果

IP パケット送信数（Share）	16,009
TCP PUSH フラグ付与件数（Share）	5,056
宛先 IP アドレス数（Share）	299
宛先 TCP ポート数（Share）	254
IP パケット送信数（Web 閲覧）	5,322
TCP PUSH フラグ付与件数（Web 閲覧）	469
宛先 IP アドレス数（Web 閲覧）	34
宛先 TCP ポート数（Web 閲覧）	2

閲覧を行った．実験時間はそれぞれ 15 分間とし，実験機から送信される IP パケットを収集し解析した．表 1 に計測結果を示す．

まず，Share の IP パケット送信数は 16,009 件であり，Web 閲覧の約 3 倍である．IP パケット送信数に占める TCP PUSH フラグ付与率は，Share が約 31.6&，Web 閲覧が約 8.8%である．Share から見た宛先 TCP ポート番号は 1 番から 65535 番まで一様分布していた．なお，Web 閲覧の宛先 TCP ポート番号は 80 番と 443 番のみであった．

これらから，Share は通常の Web 閲覧に比べ大量の IP パケットを送信し，TCP PUSH フラグの付与率が高い．これは，広範囲な宛先 IP アドレス，および一様分布する宛先 TCP ポート番号に対しコネクションを確立すると言える．加えて Share は，コネクション生成時に相手ノード IP アドレスを直接指定する．これは，相手ノード接続時にノードからの名前解決が発生しないと言える．

2.3 フィルタリングによる利用制限の検討

P2P ファイル共有通信制限のため，フィルタリングによる制限を検討する．

Hybrid 型 P2P ファイル共有ではインデックスサーバが単一障害点となる．インデックスサーバへの経路を遮断すれば，理論上容易にリソース検索機能を遮断できる．このため，既存のフィルタリング技術は，Hybrid 型 P2P ファイル共有通信の制限に関しては一定の効果が期待できる．

一方，Pure 型 P2P ファイル共有の制限は困難を伴う．前節で述べたように，Pure 型 P2P ファイル共有の通信では，ランダムかつ広範囲な宛先 IP アドレスかつ宛先 TCP ポートに対しコネクションを生成する．このため既存のフィルタリング技術による利用制限は困難と言える．

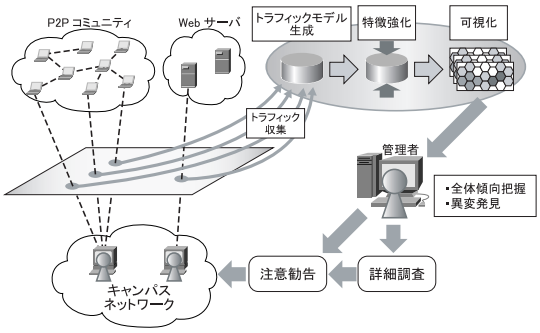

図 3: 検出支援モデル

3 P2P ファイル共有のための検出支援

3.1 特徴強化と可視化による検出支援モデル

管理者による P2P ファイル共有の検出支援を実現するため，図 3 に示す検出支援モデルを定義する．このモデルは「トラフィック生成」「特徴強化」「可視化」機能から構成される．

3.1.1 トラフィックモデル生成

キャンパスネットワークからの送信トラフィックを収集しモデル化する．

P2P ファイル共有検出のため把握すべきことは，広範囲なコネクションを持ち，TCP PUSH フラグ付与率が高いトラフィックが存在するか否かと言える．さらに，宛先 IP アドレスが名前解決の結果得られたものではなく，直接得られたものが多ければ，そのトラフィックを発生しているクライアントは P2P ファイル共有を行っている可能性が高い．

このため本研究では，ネットワーク内部から外部に対するコネクション生成状況が把握可能な情報抽出を行う．また，コネクションごとの TCP PUSH フラグ付与状況，および宛先 IP アドレスにおける名前解決の試行状況を抽出する．抽出した特徴量を送信元 IP アドレスを要素とし，単位時間ごとにモデル化する．本研究ではこれをトラフィックモデルと呼ぶ．

3.1.2 特徴強化

トラフィックモデルの特徴量を強化する．分散したコネクション状態を持ち，TCP PUSH フラグが付与された要素に重み付けを行う．さらに宛先 IP アドレスに関する名前解決が試みられていない要素についても重み付けを行う．これにより，P2P ファ

イル共有を行っている可能性を持つ要素を強調できる．この結果，他のトラフィックに埋没する P2P トラフィックを浮上させ，その存在を管理者に気づかせることができる．

3.1.3 可視化

単位時間で集積されたトラフィックモデルを可視化し，管理者に提示する．本研究で扱う監視は，全体傾向把握とその変化による異変発見の支援である．このため管理者への情報提示は一目で全体状況が把握できることが望ましい．単位時間の状況が把握できればトラフィック全体の俯瞰が可能となり，変化の追跡も容易となる．

4 モデル構成と可視化

4.1 トラフィックモデルの構成

トラフィックモデルは単位時間におけるトラフィック特性を定量的に集積しなければならない．このため，モデル生成にはベクトル空間モデル（Vector Space Model:VSM）を適用する．モデルを構成する特徴ベクトルには送信元 IP アドレスが対応し，特徴量として宛先 IP アドレスとその出現量を集積する．各要素の特徴量は，TCP PUSH フラグの出現量，および DNS 名前解決の有無により重みを付け，その特徴を強化する．

特徴ベクトルを $\boldsymbol{x}$，宛先 IP アドレスごとの出現量を a_1〜a_n とすると，特徴ベクトルは次式で表わされる．

$$\boldsymbol{x} = \{a_1, a_2, \ldots, a_n\}$$

トラフィックモデルは，生成されたすべての特徴ベクトルを集めたものである．トラフィックモデルを $\boldsymbol{D}$, 特徴ベクトルを $\boldsymbol{x}_1$〜$\boldsymbol{x}_m$ とするとトラフィックモデルは次式で表わされる．

$$\boldsymbol{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}^T$$

これにより，ネットワーク内ノードから送信されるトラフィック特性を，特徴ベクトル $\boldsymbol{x}$ のベクトル集合で表現できる．結果，ノード間類似度を特徴ベクトル間の余弦類似尺度のみで距離関係を算出でき，コネクション特性の類似性をベクトル間類似度で置き換えることができる．

4.2 自己組織化マップによる可視化

生成されたトラフィックモデルは多次元ベクトル集合として構成されている．これは送信元 IP アドレスと宛先 IP アドレスの関係が，多次元空間上の分布として表現できることを意味する．人間は基本的に三次元までの空間は直感的に把握可能だが，それ以上の多次元空間の把握には困難を伴う．

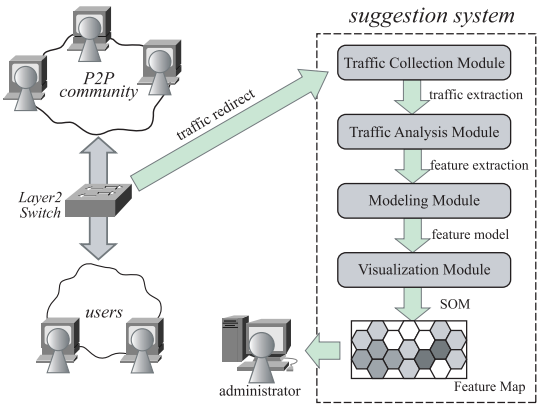

図 4: システム構成

自己組織化マップ (Self-Organizing Map:SOM) は，2 層のニューラルネットワークで構成される教師なし競合学習モデルである．SOM はデータ間の幾何学的構造を可能な限り保った状態で二次元平面に写像する．同時にクラスタリングをおこなう．この結果，管理者は平易な二次元平面にて管理対象組織のトラフィック傾向の俯瞰が可能となる．

5 試作システムの概要

本章では，実証実験のために構築した試作システムについて述べる．図 4 に試作システムの構成を示す．本システムは「トラフィック収集部」「トラフィック解析部」「モデル化部」「可視化部」から構成される．以下，各部の概要を述べる．

5.1 トラフィック収集部

トラフィック収集部では，監視対象ネットワークが受発信するすべての IP パケットを収集・蓄積する．L2 スイッチのポートミラーリング機能により，獲得する IP パケットを本システムにリダイレクトする．トラフィック収集部は，導入システムの Ethernet カードを promiscuous mode に設定し，リダイレクトされた IP パケットを収集する．

5.2 トラフィック解析部

収集された IP パケット群を解析し，送信元 IP アドレス，送信元ポート番号，宛先 IP アドレス，宛先ポート番号，パケットサイズ，フラグを抽出する．

表 2: 実験環境

CPU	Intel Pentium4 2.4GHz
Memory	640 Mbytes
HD	40 Gbytes
OS	Linux (kernel 2.4.18)

5.3 モデル化部

トラフィックモデルを生成する．送信元 IP1 つに対し，宛先 IP・ポート番号数が次元となる多次元ベクトルを生成する．モデル全体では特徴ベクトルが全送信元 IP 数分集積されたベクトル集合となる．ベクトルの各要素には宛先 IP アドレス・宛先ポート番号別にパケット出現回数とパケットサイズを集積する．

また，モデル化部では重み付けを行う．Share などの P2P プログラムやストリーミングプログラムではパケット送信時に PUSH フラグが設定される．このため，PUSH フラグが設定されたパケットは P2P やストリーミングによるトラフィックである可能性が高い．さらに P2P アプリケーションは，接続する相手ノードの IP アドレスを直接指定しコネクションを生成する．このため DNS による名前解決が発生しない．これらの特徴に合致する特徴ベクトルに重み付けし，特徴を強化する．

5.4 可視化部

得られたトラフィックモデルを SOM アルゴリズムを用いて可視化する．SOM アルゴリズムにより抽出されたパケット群が自己組織化され，似た特性を有する特徴ベクトルが集約された特徴マップが生成される．PUSH フラグが設定されたパケットなど，特に特徴を持つ特徴ベクトルはクラスタとして表出する．このため管理者に対し，管理対象ネットワークに発生した特異トラフィックへの気づきを支援できる．

6 実験と考察

6.1 実験環境

試作システムに実験データを入力し特徴マップ生成を行った．表 2 に実験環境を示す．

ある組織に許可を得て，2005 年 9 月 14 日にその組織内の端末が受発信したすべての IP パケットを収集し，実験データとした．なお，実験期間中 1 台の端末にて意図的に「Share」を動作させ，適当なデータファイルをダウンロードした．表 3 に実験データ件数および処理過程で生成された特徴ベクトル数を示す．

表 3: 実験データ件数

種別	件数
実験データ件数	1,073,614 件
特徴ベクトル生成数	12,683 件

表 4: 各ホスト処理状況

ホスト	処理状況
A	P2P ファイル共有（Share）
B	P2P Phone（Skype）
C	Windows Update 実行
D	Web 閲覧
E	組織内ファイアウォール（NATBOX）
F	ストリーミング受信
G	組織内ファイアウォール（NATBOX）

6.2 考察

図 5〜7 に実験で生成した特徴マップを示す．また表 4 に，実験時間における各ホストでの処理状況を示す．

1 つの特徴マップは 20 × 16 の 320 要素を持つ．それぞれの要素には比較的多く出現した特徴ベクトルが表出する．今回の実験で生成された特徴ベクトル総数は 12,683 件であるため，約 2%の大規模通信が表出することになる．特に広範囲の宛先 IP および宛先ポートに対して通信を行っている送信元 IP

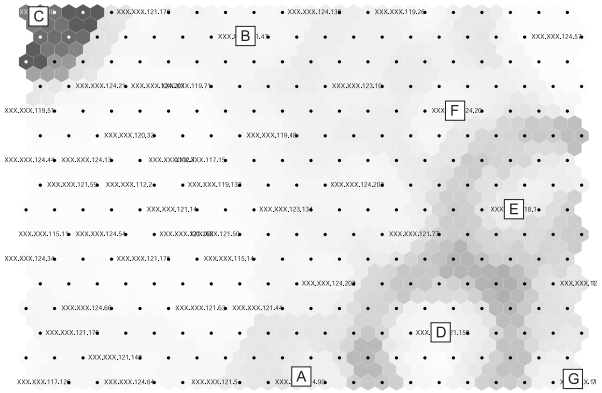

図 5: 特徴マップ（重み付けなし）

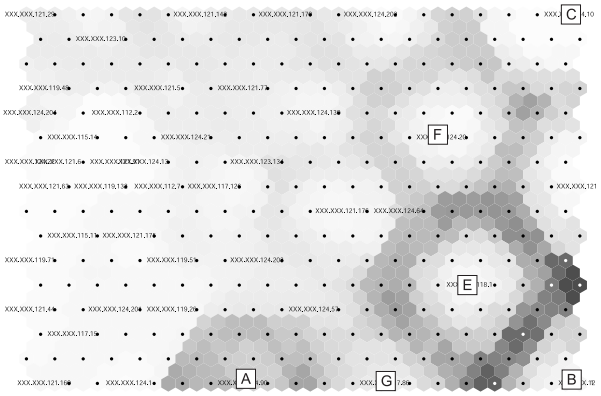

図 6: 特徴マップ（PUSH フラグ重み）

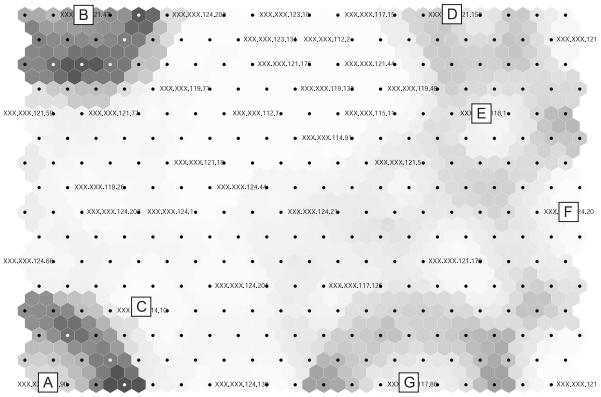

図 7: 特徴マップ（PUSH フラグ重み+DNS 重み）

のベクトルは自己組織化されクラスタとして表出している．

図 5 は重み付けをせず，宛先 IP アドレスとその通信量のみでトラフィックモデルを生成し可視化したものである．この特徴マップではホスト C が特徴的に表出している．これはインストール直後の Windows Update により大量のファイル転送が発生したものである．しかし，表出してほしいホストである A および B が認識しやすいとは言えない．

図 6 は，前述のトラフィックモデルに PUSH フラグ出現に応じた重み付けを行ったものである．ここではホスト A が若干強調されていることが分かる．しかし，ホスト E や F のような組織内ファイアウォールからのトラフィックや，ストリーミングを受信中のホストなども強調されている．

図 7 は，さらに DNS による名前解決情報を重みとして加えたものである．P2P クライアントは相手ノードに対し直接 IP アドレスを指定してコネクションを生成するため，DNS による名前解決が発生しない．この特徴を収集し，重み付けに利用することでホスト A および B を明確なクラスタとして表出させることができた．

7 まとめ

本稿では，企業や大学のキャンパスネットワークで行われる P2P ファイル共有通信の問題について述べ，これらの P2P トラフィックを既存のフィルタリング技術で制限することの困難性について述べた．その上でキャンパスネットワーク内から受発信される P2P トラフィックを検出する手法を検討し，管理者がおこなう P2P トラフィック検出のための支援モデルを提案した．さらに支援モデルを実現するために必要なトラフィックのモデル化手法について述べ，多次元モデルの認識限界を下げトラフィック傾向の俯瞰を可能にするために行う可視化手法について述べた．また，本提案の有効性を検証するために実装した試作システムについて述べ，実証実験の結果である特徴マップを示し考察をおこなった．

今後は重み付け手法の改良などにより，特徴マップ上での P2P トラフィックのより明確な提示を試みる．

参考文献

[1] WinMX Web Site, http://www.winmx.com/

[2] Winny Web Site, http://www.geocities.co.jp/SiliconValley/2949/

[3] BitTorrent Web Site, http://bittorrent.com/

[4] ITMedia, “秋田県湯沢市住民 1 万人分の個人情報を Winny で流出”, http://www.itmedia.co.jp/enterprise/articles/0504/15/news070.html

[5] 石川 博, “次世代データベースとデータマイニング”, CQ 出版社, 2005.

[6] 藤井聖, 中尾嘉宏, 中村豊, 藤川和利, 砂原秀樹, “フローを用いた特定トラフィック検出システムの運用”, 第 31 回分散システム/インターネット運用技術研究会, 2003.