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

2003－DPS－114 (3)
2003－EIP－ 20 (3)
2003／8／28

P2P ファイル共有におけるコンテンツ分析

大井恵太[†] 亀井 聡[††] 森 達哉[††]

インターネットへのアクセス環境の向上により，P2P アプリケーションの普及が著しい．特に，ファイル共有アプリケーションの普及は，著作権ビジネスに関わる者達をはじめとした様々な領域にその影響をおよぼしつつある．一方で，ファイル転送時にはサーバを介さない P2P アプリケーションの特性から，大規模な情報収集は困難であった．本稿では，P2P ファイル共有の規模，共有されるファイルの実態を明らかにするため，WinMX, Gnutella, Winny について，ヒューリスティックな測定手法に基づき，測定とコンテンツ分析を実施した．

Analysis of Peer-to-Peer File Sharing Applications

KEITA OOI[†] ,SATOSHI KAMEI[††] and TATUYA MORI[††]

As Internet access line bandwidth has increased, peer-to-peer applications have been increasing and they have had a great impact on networks. In particular, the spread of peer-to-peer file sharing applications raises concerns about copyrights. However, it is difficult to gather much information because files are not transmitted via a server. In this paper, we measure and analyze peer-to-peer file sharing applications, WinMX, Gnutella, and Winny by heuristic methods in order to clarify the scale of peer-to-peer file sharing and details of shared files.

1. はじめに

ADSL や FTTH といった定額の広帯域常時接続環境がインターネットの一般的な利用者の間に急速に普及してきている中，インターネットの新たな利用形態として，P2P (*Peer-to-Peer*) と呼ばれるアプリケーションの普及がめざましい [1]．特に，P2P の代名詞ともいえる Napster [2] や Gnutella [3] に代表される P2P ファイル共有アプリケーションでは大容量のマルチメディアファイルが共有されることが多く，広範囲にその影響を及ぼしつつある．

ネットワーク運用面で最も大きなインパクトとなるのは，ファイル転送量の爆発的な増加である．P2P ネットワーク上では共有されるファイルサイズが大容量化することもさることながら，P2P の特性のひとつである，スケーラビリティの向上も転送量増加の重要な一因となっている．これらの要因により，ネットワークにおける P2P トラヒックの占める割合は近年急激に増加し続けている．

一方，P2P ネットワーク上で共有されるファイルには，現行の著作権法上，共有が違法とされるものが多い．そのため，著作権ビジネスに関わる団体など，様々な分野から，共有規模や共有内容にも関心が集まっている．しかし，ファイル転送時にはサーバを介さない P2P の特性から，情報収集は容易ではなかった．

本稿では P2P アプリケーションにおけるファイル共有の現状を把握するため，国内で広くユーザを獲得している主要な P2P アプリケーションである，WinMX [4] と Winny [5]，さらにプロトコルが広く公開されていて，P2P レイヤ上での測定が容易な Gnutella を中心に実ネットワークでの測定を行い，実際に P2P ネットワーク上に流通しているファイル，コンテンツについての分析を行った．

以下，第 2 節で P2P ファイル共有とネットワークの利用形態の変遷について俯瞰した後，第 3 節でアプリケーションレイヤからの測定について様々な手法を概説する．最後に第 5 節でまとめと今後の課題について述べる．

† 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories. NTT Corporation
†† 日本電信電話株式会社 NTT サービスインテグレーション基盤研究所
NTT Service Integration Laboratories, NTT Corporation

2. P2P ファイル共有

P2P ファイル共有と呼ばれる分野は P2P の代名詞ともなった Napster をその起源とし，参加ユーザのディスク上にあるファイルを検索するメカニズムと，検索によって各ユーザが発見したファイルを直接，あるいはキャッシュや他ノードを介して間接的に自分のノードへと転送する機能をその特徴とする．

Napster では，ファイル検索のためのインデックスサーバを必要とするため，検索トラヒックはウェブと同様のパターンを示す．しかし，トラヒックの大部分を占める音楽ファイルの転送トラヒックは，各ノード(ピア)間でサーバを介さずに直接発生する．Napster 社は現在はそのサービスを停止しているが，OpenNap [6] プロトコルを利用した互換サーバ/互換クライアントは健在である．特に，互換クライアントのひとつである WinMX は日本国内において広く普及しており，国内ネットワークにおいて顕在化している P2P トラヒックの中では最大のものとされている．

Napster の後に登場した Gnutella においては，ファイル転送時だけでなく検索時も各ピア間で自律分散的に通信することで中央にサーバを必要としないアーキテクチャを用いており，中央サーバを持たないことから，サービスの停止や規制を行うことも困難となっている．

その後も様々な P2P ファイル共有アプリケーションが登場しており，現在世界で最大規模の P2P ネットワークを構成しているファイル共有アプリケーションは FastTrack 技術を用いた KaZaA [7] である．しかし，ファイル名が多言語に対応していないことなどから，日本国内において主流を占めているのは WinMX だとされている．

トラヒック面では，Freenet [8] の技術を元に改良を加えた Winny と呼ばれる国産アプリケーションのトラヒック量が急激に延びており，測定場所，特に ISP 内部においては，WinMX をも上回っている，という測定報告もある [1].

3. アプリケーションレイヤ測定および分析

P2P アプリケーションでは，IP ネットワークの上に P2P アプリケーション動作ノード間でオーバーレイした P2P ネットワークを構築する．このアプリケーションレイヤネットワークは，IP レイヤとは独立に構築される．このため，IP レイヤの測定だけでは利用者の規模を測定することも困難であり，P2P アプリケーションの全体像を把握するためには，アプリケーションレイヤでの測定を合わせて実施することが不可欠である．一方で，プロトコル仕様が公開されていないアプリケーションについてはネットワークレイヤの測定で利用状況を推定することすら困難である．しかも，アプリケーションによって収集可能な情報は異なるため，現状ではアプリケーション毎にヒューリスティックな測定手法を用いる必要がある．

本節では，今回の測定で実施した各種 P2P アプリケーションにおけるアプリケーションレイヤでの測定手法の概要を述べる．

3.1 各種アプリケーションの解説および測定手法

3.1.1 WinMX (Napster / OpenNap)

WinMX は Napster のプロトコル仕様をオープン化した OpenNap 仕様に従った Napster Client の一種である．OpenNap はネットワーク参加時に，各クライアントが自分の保持するファイル情報をインデックスサーバに登録する．

測定は 以下の方法を用いた (図 1).

- ユーザーリストの作成
  「Client Search Request」をインデックスサーバ (OpenNap サーバ) に発行し，キーワードに対する，応答メッセージである「Search Response」からユーザーリストを得る．
- 各ユーザーの共有ファイルの閲覧
  ユーザーリスト内のユーザーに対し，「Direct Browse」「Browse」命令を発行し，ユーザーが共有しているファイルのリストを得る．

以上の操作を繰り返すことで，ユーザの共有ファイルのリストを取得した．

3.1.2 Gnutella

Gnutella はファイル共有，検索，転送を目的とした P2P アプリケーションであり，サーバントと呼ば

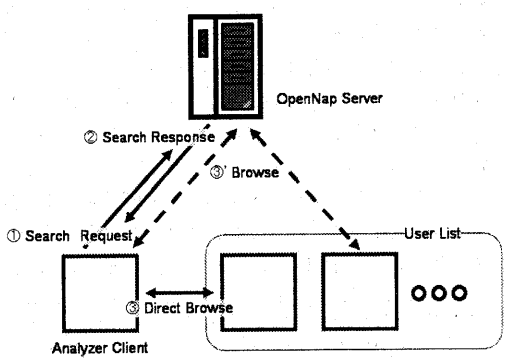


図 1 WinMX: 測定イメージ

れる自律動作するノードによって P2P ネットワーク，GnutellaNet を構成する．各サーバントはアプリケーションレイヤでのブロードキャストによってファイル検索を行うため，GnutellaNet には中央サーバは存在しない．

Gnutella については，プロトコルやパケット仕様がドキュメント化 [9] されているため，データ収集は比較的容易であり，様々な測定や評価が成されている [10–12]．

本稿においてはオープンソースで配布されている gtk-gnutella に改造を施し GnutellaNet 上の制御パケットをダンプする機能を追加したものを用いてネットワーク上の 1 点 (GnutellaNet 的には 3 サーバント) を通過する制御パケットを全て取得する形で測定を行った．

測定イメージは図 2 に示す．

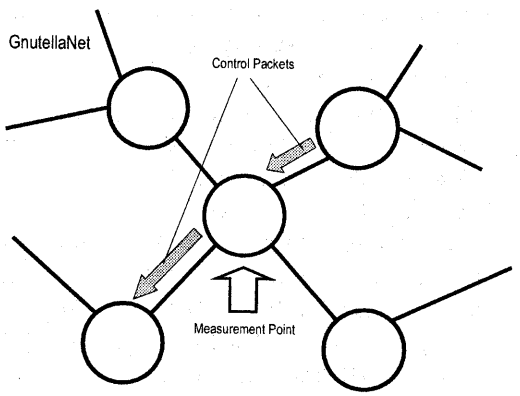


図 2 Gnutella: 測定イメージ

本測定は GnutellaNet 上での測定となるため，制御パケットの収集に留まり，実際のトラヒックの大部分を占める直接 2 点間で行われるファイル転送は測定の対象とはなっていない．このため，以下の測定結果は GnutellaNet で流通しているファイルのうちで検索対象となったファイルの情報に限定されていることに留意する必要がある．

また，同時にネットワークレイヤにおいて tcpdump を用いて P2P ネットワーク上の隣接ノードのアドレス，P2P 通信ポート番号の収集を行った．

3.1.3 Winny

Winny は匿名巨大掲示板 2 ちゃんねる [13] において，WinMX の後継を狙って開発が始まった．おおまかな挙動の仕様については作者の発言録という形で公開されており，匿名性の高い P2P アプリケーションである Freenet を元にしていると言われている．パケットフォーマットや内部構造は非公開であり，ソフトの配布自体もバイナリで行われている．

プロトコルを非公開にすることによって匿名性の確保や DOM ☆の増加を防いでいるため，内部構造には不明な点が数多くある．しかしながら，公表された情報から基本的挙動を分析したドキュメント [15] が公開されている．

測定に際しては，Winny の命令を上書きすることによって検索要求を定期的に発行するフリーソフトツールである，WinnyFileLister (WFL) を用いた．この際，定期的に検索語を切り換える (.a～.z) ことによってキーワード空間を探索し，ファイル情報を収集した．

また，同時にネットワークレイヤにおいて tcpdump を用い，P2P 上の隣接ノードにあたる，Winny の通信相手の IP アドレス収集も行った．測定イメージを図 3 に示す．

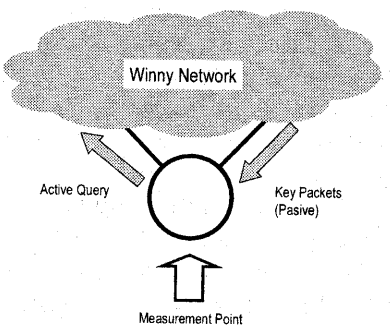


図 3 Winny: 測定イメージ

4. 測定結果の分析

4.1 ネットワーク規模

ネットワーク規模を示すデータとして，今回取得したデータにおける固有 IP アドレス，固有ファイル数，合計，平均ファイル容量を表 1 に示す．WinMX については 2003/02/14, 2003/02/18 にサンプル測定，Gnutella と Winny については，2003/04/18 から週末の 3 日間 (68 時間) 同時に測定した．

Gnutella については，文献 [16,17] では，測定方法は不明であるものの，ユーザは，10 万を越えるとされている．本稿の結果でも，P2P レイヤではそれを超える IP 数 (135,291) が収集できている．一方で，IP

☆ Download OnlyMembers; ファイルの取得だけを行い，コミュニティへの貢献，この場合ファイルのアップロードを行わない利用者のことを指す，Free Rider [14] とも呼ばれる．

表 1 測定データ (P2P ネットワークの規模)

種別	固有 IP 数	固有ファイル数	合計ファイル容量	平均ファイル容量
WinMX	350	165,933	25TB	150MB
Gnutella	30,052 135,291 (P2P)	3,883,752	1.3PB	330MB
Winny	39,123	235,423	27TB 6.4PB (share)	63MB

レイヤでは 3 万ノード程度の収集である.

Winny については, P2P ネットワークの構成法として, Gnutella よりも隣接ノードとの結びつきがゆるやかであり, 検索リンクの張り替えも頻繁に行われる一方, 実際のファイル転送量に応じたリンク数の増減等の機能があるため, 両者の比較は難しい. しかしながら, Gnutella での IP レイヤと P2P レイヤでのアドレス収集数の比から, 単純に Winny のユーザ数を推定すると, Gnutella に比肩する 10 万程度に達する可能性がある.

4.2 ファイル統計

4.2.1 サ イ ズ

図 4 はアプリケーション測定によって得られた WinMX, Gnutella, Winny のファイルサイズの累積分布 (補分布) を両対数プロットしたものである.

図 4(a) は WinMX による中央サーバ/クライアントへの問い合わせにより得たファイルのサイズ, 図 4(b) は GnutellaNet 上での QueryHit パケットから抽出したファイルのサイズ, 図 4(c) は WFL によって得られた Winny のネットワーク上に存在するファイルのサイズである. データの抽出方法が異なるため, これらの統計は単純に比較できないことに留意する必要があるが, 各 P2P ネットワークで流通しているファイルのサイズ分布に関するおおよその傾向を捉えていると考えられる.

各 P2P ネットワークにおけるファイルの平均サイズはそれぞれ点線部で示したとおりであり, WinMX が 153.64 Mbyte, Gnutella が 326.48 Mbyte, Winny が 62.82Mbyte である. また, いずれの P2P ネットワークにおいても, ファイルサイズ分布は 10^9 byte (およそ 1 Gbyte) 前後で分布の裾野部において急激な落ちこみを示す. これは各種メディアの特性や, ファイルシステム上の限界, およびハードディスクの物理的な限界に起因するカットオフが存在しているためと考えられる.

図 5 は, P2P ネットワークとウェブのファイルサイズ分布を比較したものである. ウェブのファイル統計データとしては, IRCache Project [18] にて公開されているプロクシサーバの履歴を用いた. 前述の裾野部分を除けば, およそ $10^4 \sim 10^8$ byte の広い範囲において, ウェブに比べても非常に裾野部分が厚い (heavy-tail な) 分布にしたがっていることが見て取れる.

4.2.2 拡張子分布

図 6, 図 7, 図 8 は流通コンテンツ種別の分析として, 得られたファイルを拡張子毎に分類したときの分布である. Gnutella はファイル数の半分程度が mp3 による音楽ファイルによって占められている一方で, WinMX と Winny は avi や mpg といった動画像が大きな割合を占めている. Winny ではそれらに加えて zip や rar といった圧縮ファイルもかなりのサイズのものが流通しており, 各アプリケーションで、共有されているファイル種別に特性がある.

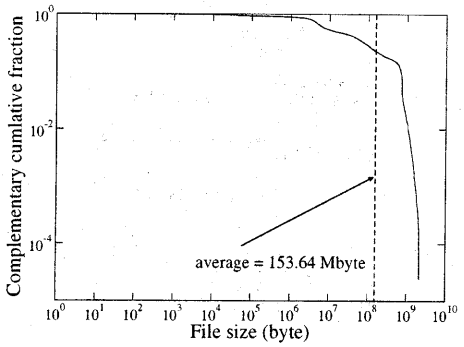
4.3 コンテンツの分析

図 10 は WinMX, Gnutella, Winny における共有ファイルについて, それぞれ x 軸: コンテンツの被共有数 (n), y 軸: n 人に共有されているコンテンツの数とした, 両対数グラフである.

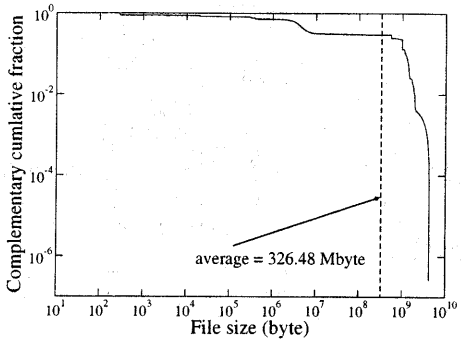
これらの図 10 に見て取れる直線性は, Lotoka の法則 (または Zipf の第二法則) と呼ばれるものである. 一般に, 個々の利用者が何かを選んだときに発生する記録では, このような, Lotoka の法則に則った形状 (Lotoka 型分布) を示すことが多い.

Lotoka 型分布では, 総被共有数に比べて総コンテンツ数が少ない場合に, コンテンツの被共有数が少ない領域 (左上の領域) において, 直線性を保たず, 下方に曲がることが知られている [19, 20].

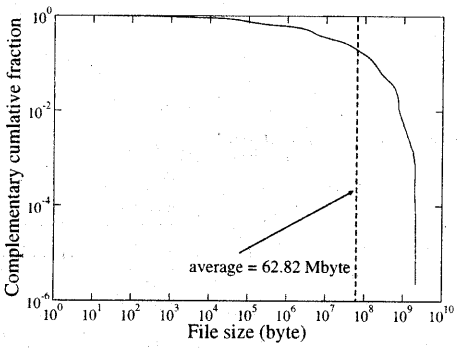
これらの図 10 では, 左上の領域まで直線性を保っていることから, Winny の場合は, 全ファイル数が 23 万, 利用者数 3.9 万の規模に対し, 十分な種類のコンテンツが供給されていることが予想できる.



(a) WinMX: ファイルサイズ分布



(b) Gnutella: ファイルサイズ分布



(c) Winny: ファイルサイズ分布

図 4 ファイルサイズ累積分布 (補分布)

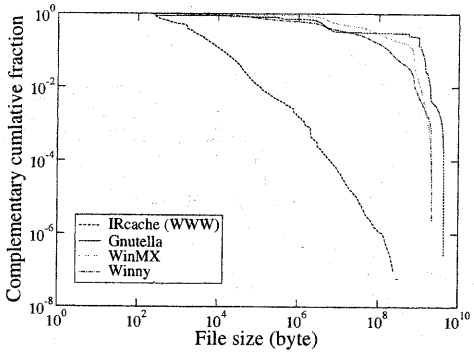
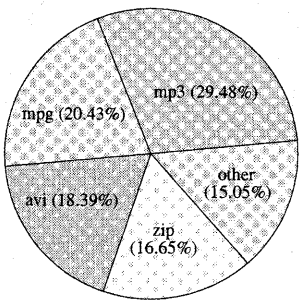
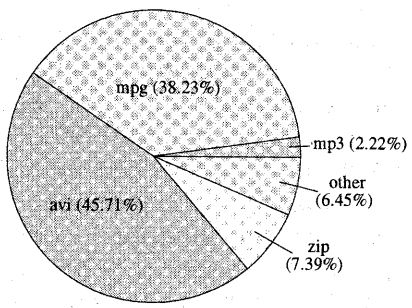


図 5 WinMX / Gnutella / Winny / Web のファイルサイズ累積分布 (補分布)

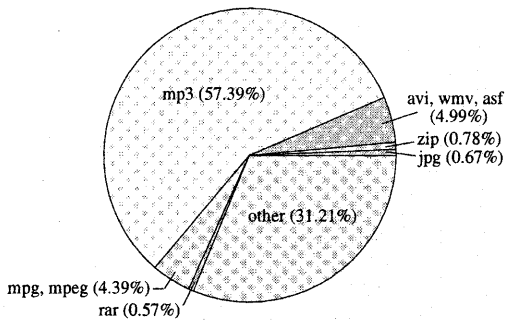


(a) ファイル数

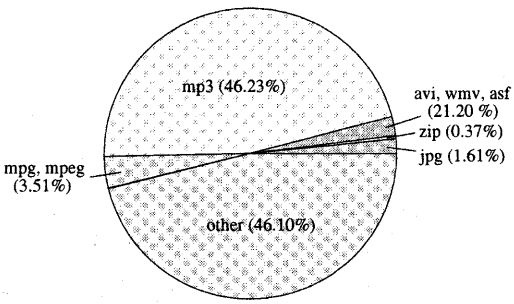


(b) ファイル容量

図 6 WinMX: 拡張子別ファイル分布

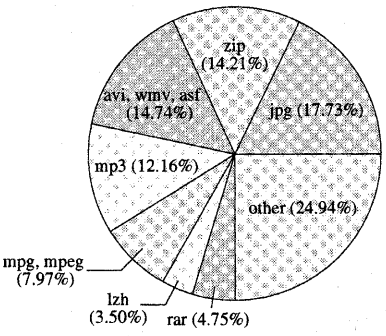


(a) ファイル数

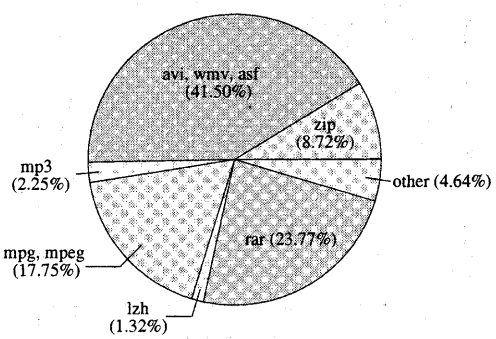


(b) ファイル容量

図 7 Gnutella: 拡張子別ファイル分布



(a) ファイル数



(b) ファイル容量

図 8 Winny: 拡張子別ファイル分布

5. ま と め

本稿では普及がめざましい P2P アプリケーションについて，その規模と共有されているファイルを明らかにするため，WinMX，Gnutella，Winny の 3 アプリケーションについて，測定と分析を実施した．その結果，国内だけで数万人以上のユーザ規模の P2P ファイル共有ネットワークが複数存在していること，P2P におけるファイル共有数の分布も Lotoka 分布に従っていること，が分かった．今後はこれらのファイル共有上の構造を用いて，より詳細な分析を行う予定である．

アプリケーションレイヤでの分析が容易ではないことからも分かるようにコンテンツに関してはいわゆる違法なファイル共有を止める手法を開発することは容

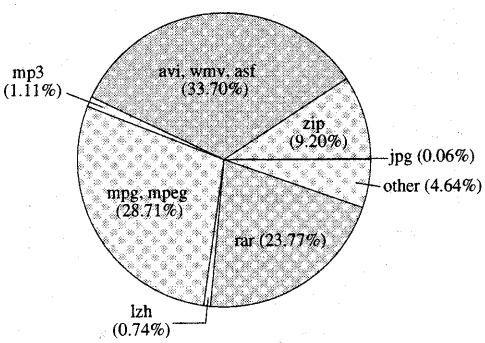
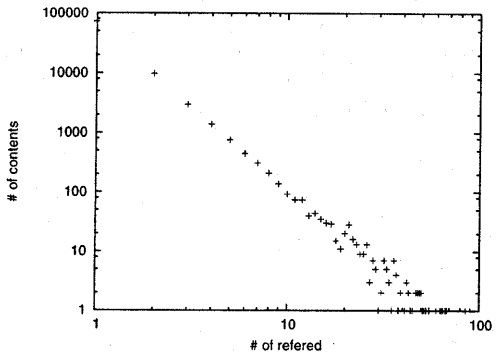
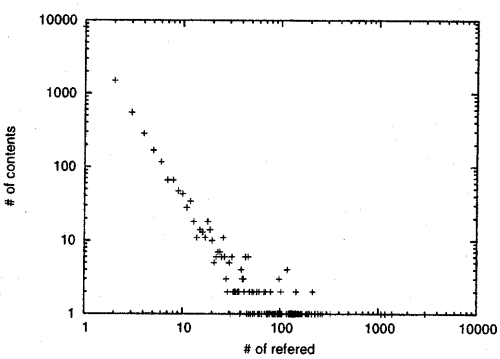


図 9 Winny: 拡張子別ファイルサイズ総量 (被参照量)

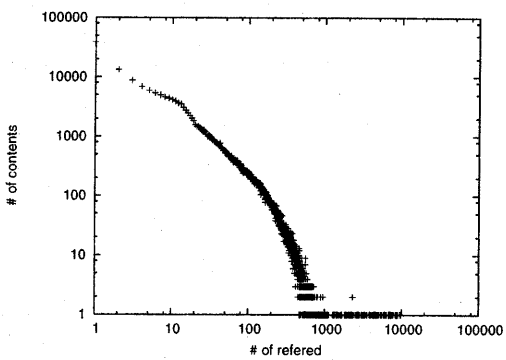
易ではない．このため，現行法の厳密な適用や法改正による厳罰化により，これらのファイル共有を防ごうとする試みもある．一方で，Richard M. Stallman [21] や，Lessig [22] に代表される現行の著作権法に対する改革論もあり，関心を集めている．こういった著作権に関する議論を行う際も，本稿の様に客観的な現状把握を行うことは，有意義であろう．



(a) WinMX: コンテンツ分布



(b) Gnutella: コンテンツ分布



(c) Winny: コンテンツ分布

図 10 共有コンテンツ分布

参 考 文 献

1) 亀井聡, 森達哉, 大井恵太, "P2P ファイル共有の実態と課題," 信学技報 *CQ2003-40*, 2003.
2) Napster Inc. `http://www.napster.com`.
3) "Gnutella." `http://www.gnutella.wego.com`, 2000.
4) Frontcode Technologies, "WinMX." `http://www.winmx.com`.
5) "Winny. " `http://www.geocities.co.jp/SiliconValley/2949/`.
6) "OpenNap: Open Source Napster Server." `http://opennap.sourceforge.net`.
7) "KaZaA." `http://www.kazaa.com`.
8) I.Clarke, O.Sandberg, and B.Wiley, "Freenet: A distributed anonymous information storage and retrieval system," *Designing Privacy Enhancing Technologies: International Workshop on Design Issues in Anonymity and Unobservability, LNCS2000*, Dec. 2000.
9) Clip2, "The gnutella protocol specification v0.4." `http://www9.limewire.com/developer/gnutella_protocol_0.4.pdf`.
10) Subhabrata Sen and Jia Wang, "Analyzing peer-to-peer traffic across large networks," *ACM SIGCOMM Internet Measurement Workshop 2002*, 2002.
11) Stefan Saroiu, P. Krishna Gummadi and Steven D. Gribble, "A measurement study of peer-to-peer file sharing systems," *Multimedia Computing and Networking (MMCN)*, Jan. 2002.
12) K.Sripanidkulchai, "The popularity of gnutella queruies and its implications on scalability," *O'Reilly Peer-to-Peer and Web Services Confierence*, Sept. 2001.
13) "2 ちゃんねる." `http://www.2ch.net`.
14) E.Adar and B.Huberman, "Free riding on gnutella," *Technical report, Xerox Parc*, Aug. 2000.
15) aki, "winny.info." `http://winny.info`.
16) "Rolling Host Count." `http://www.limewire.com/index.jsp/size`.
17) "SLYCK - File Sharing News and Info." `http://www.slyck.com`.
18) "IRCache Project." `http://www.ircache.net`.
19) 市川祐介, 佐藤基, "Web 検索サーバのログに基づいたアクセス傾向の分析," 信学技報 *IN99-65*, 1999.
20) 村中かほり, 松田三千代, 会田雅樹, 本橋健, 佐藤基, "有限なアドレス空間に対する internet アクセスパターンの特性分析," 信学技報 *IN2001-56*, 2001.
21) Richard M. Stallman, *Free Software, Free Society: Selected Essays of Richard M. Stallman.* GNU Press, 2002.
22) Lawrence Lessig, *The Future of Ideas: The Fate of the Commons in a Connected World.* Random House, 2001.