ファイル共有ソフトの利用に関する調査

～クローリング調査～

報 告 書 （概 要 版）

２００９年１２月

社団法人コンピュータソフトウェア著作権協会

社団法人日本レコード協会

日本国際映画著作権協会

目　次

Ⅰ．調査概要

１．調査の前提

調査は 2009 年 10 月 2 日 17:00 から 2009 年 10 月 3 日 17:00 の 24 時間､以下の P2P ネットワークに対応した手法を用いてネットワークを巡回（クローリング）し、実際にネットワーク上を流通している情報を自動収集、分析する形で実施した。

（１）Winny2

Winny プロトコルを利用したクローラを用いて、特にキーワードを設定することなく Winny ネットワーク上に流通するキー情報（ノード情報、ファイル情報）の自動収集を行った。複数のクローラを用いる事で、24 時間でほぼネットワークの全域をクローリングできる性能を確保している。

■基礎情報

・利用したソフトウェア　　P2P FINDER(Winny) 2009 年 9 月 Version
・設定情報　　総スレッド数 1,700

（２）Share EX2

Share プロトコルを利用したクローラを用いて、特にキーワードを設定することなく Share ネットワーク上に流通するキー情報（ノード情報、ファイル情報）の自動収集を行った。複数のクローラを用いる事で、24 時間でほぼネットワークの全域をクローリングできる性能を確保している。

■基礎情報

・利用したソフトウェア　　P2P FINDER(Share) 2009 年 9 月 Version
・設定情報　　総スレッド数 3,500

（３）Gnutella

Gnutella バージョン 0.6 プロトコルを利用したクローラを用いて、Gnutella ネットワーク上に流通するキー情報（ノード情報、ファイル情報）の自動収集を行った。Gnutella は全世界にノードが広がっており全域はクローリングしていない。

クローラは全ファイルを意味するキーワード（半角スペース 4 つ）を指定してクローリングを行う「キー情報クローラ」を用いている。

■基礎情報

・利用したソフトウェア　　P2P FINDER(Gnutella) 2009 年 9 月 Version
・設定情報　　キー情報クローラ　　総スレッド数 150

２．データの抽出

（１）フィルタリング

ファイル共有ソフトネットワーク上で、権利者に無許諾で送信可能な状態におかれ、流通しているファイルの調査を行った。

調査を行うにあたり、総取得件数からノード(IP とポート)およびファイル名が同一なデータを取り除いた後、(調査対象データ)、２万件をランダムに抽出した。調査対象データに対する抽出データ（２万件）の割合はそれぞれ Winny:0.025%、 Share 0.177%、 Gnutella 0.105%となった。調査対象データから、アダルト系キーワード、共通除外キーワードを含むデータを除外し、データを目視にて確認し、各ファイルについて推定されるジャンル、権利の対象および許諾の有無について調査した。

	工程
①総取得件数	クローラにより IP、ポート、ファイル名、時間を取得
②重複件数の削除	①で取得したデータのうち、IP、ポート番号、ファイル名が重複したデータを削除
③間引き後件数	②で取得したデータを 20000 件になるようにランダムに抽出
④アダルトキーワード除去	③で抽出したデータのうちファイル名にアダルトコンテンツと想定されるキーワードがあるデータを除外
⑤共通除外キーワード除去	④のデータのファイル名に共通除外キーワードがあるデータを除外
⑥合法ファイル抽出	④のデータのファイル名に「合法」のキーワードがあるデータを抽出
⑦キーワード抽出	⑤のデータを各ジャンルのキーワードで抽出

（２）権利の対象性算出方法

２．（１）で抽出したデータを目視にて以下のジャンルに分類を行った。

・著作物と推測されるもの
・アダルト
・同人
・不明ファイル
・危険ファイル
・合法ファイル

分類の際には、抽出データ（２万件）の目視確認を行い、２．（１）のキーワードフィルタリングで漏れたコンテンツのジャンルの修正も行った。

Ⅱ．調査結果

１．Winny2

（１）無許諾コンテンツの流通状況

流通コンテンツのうちおよそ半数のコンテンツが著作物と推測される。著作物と推測されるコンテンツの内訳は図２の通りである。

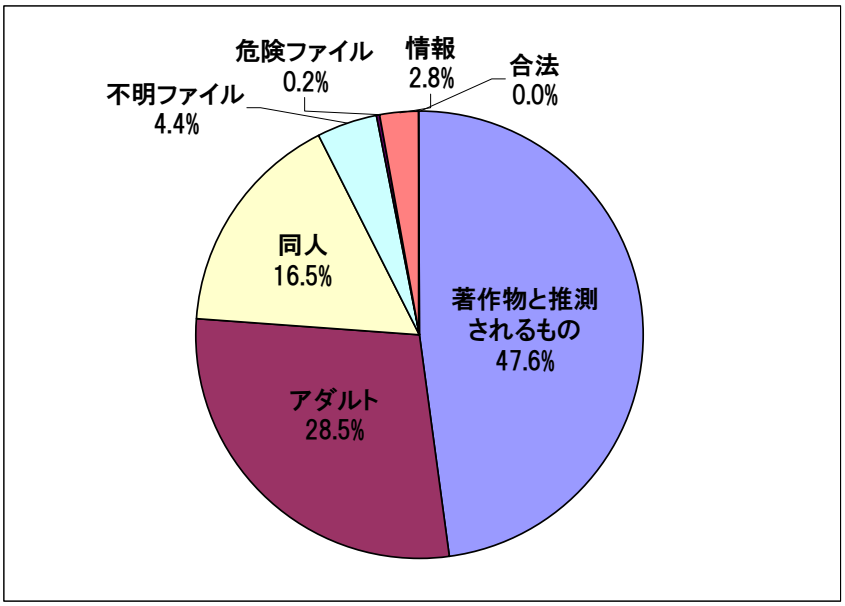


図 1 Winny コンテンツ流通状況(n=20,000)

※「著作物と推測されるもの」とは本調査で権利の所在が推定できるもの

※「アダルト」、「同人」とは本調査で権利の所在が判別できないため、権利の対象に関しての調査は見送ったもの

※「不明ファイル」とはタイトルからコンテンツの内容が判別できないもの

※「危険ファイル」とはタイトル、拡張子からウィルスなどと推定されるもの

※「情報」とはウィルス感染などで流出した個人・組織等の情報だと推定されるもの

※「その他」とはコンテンツの分類が音楽、映像関連、プログラム、書籍関連に含まれないもの

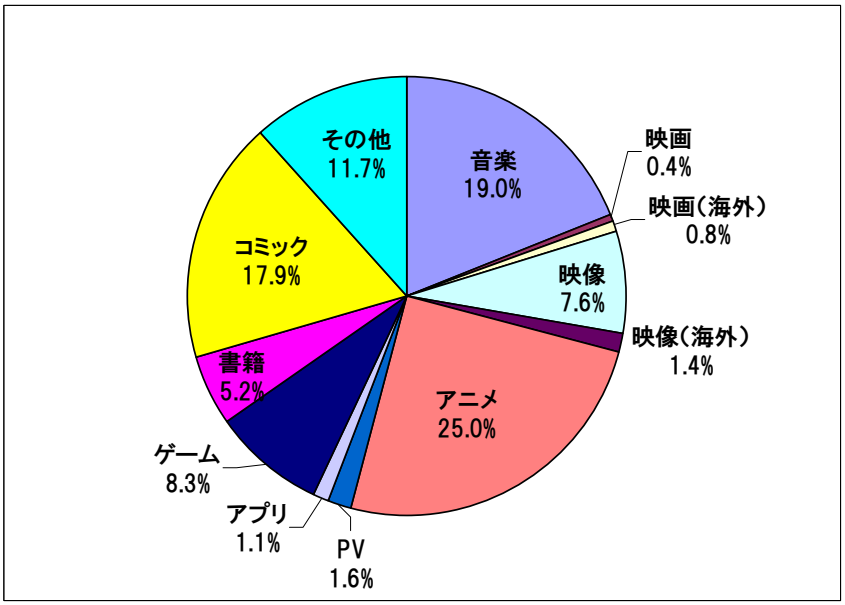


図 2 Winny 著作物と推測されるコンテンツの内訳(n=9,529)

（２）権利の対象性について

著作物と推測されるもののうち、98%に権利があり、かつ許諾がないものだと推定される。

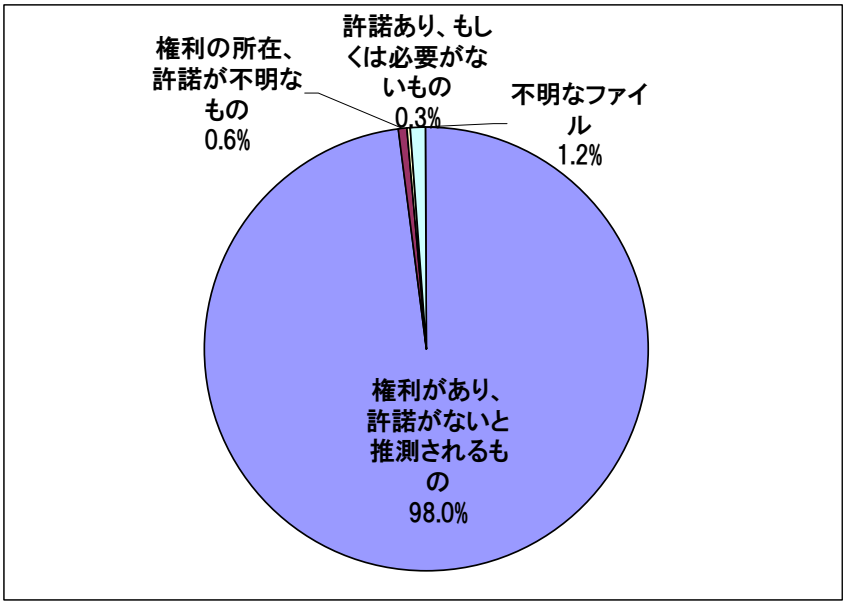


図 3 Winny 権利の対象性（全体）(n=9,529)

（３）ファイル量について

Winny のネットワークではファイルの情報はファイル本体から計算で算出されるハッシュ値を用いて管理されているため、一意なハッシュ値の件数を算出する事で Winny ネットワーク上に流通しているファイルの量を推定できる。本調査では、一日で 5,135,851 件の一意なハッシュ値が収集され、全数としてはおよそ 600 万件程度と推定される。

（４）ノード量について

本調査では IP アドレスとポート番号の一意な組み合わせをノードの量として算出した。本調査では、一日で 394,624 件の一意なノード情報が収集された。ただし、このノード数には偽キー（インデックスポイゾニングされたランダムな IP アドレス／P.11　補足参照）が含まれるため全数を把握するのが非常に困難となっている。

（５）検出ノードの国・地域について

一日で検出した 394,624 件のノードの国・地域を調査した。ただし、この結果はランダム IP を用いたインデックスポイゾニングの影響をうけているため、実態を表していないと考えられる。

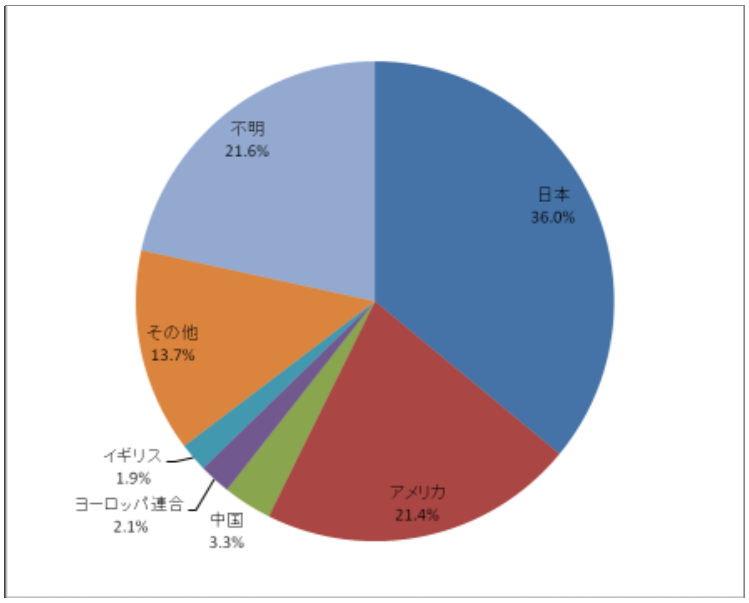


図 4　Winny 検出ノード 国・地域別分布(n=394,624)

インデックスポイゾニングの影響を排除するため Winny プロトコルを用いてクローラが接続した、Winny が確実に動作していると考えられるノードについて国・地域を調査した。

その結果、Winny が確実に動作していると考えられるノードの国・地域分布については昨年の検出ノード 国・地域分布とほぼ同じ傾向になった。

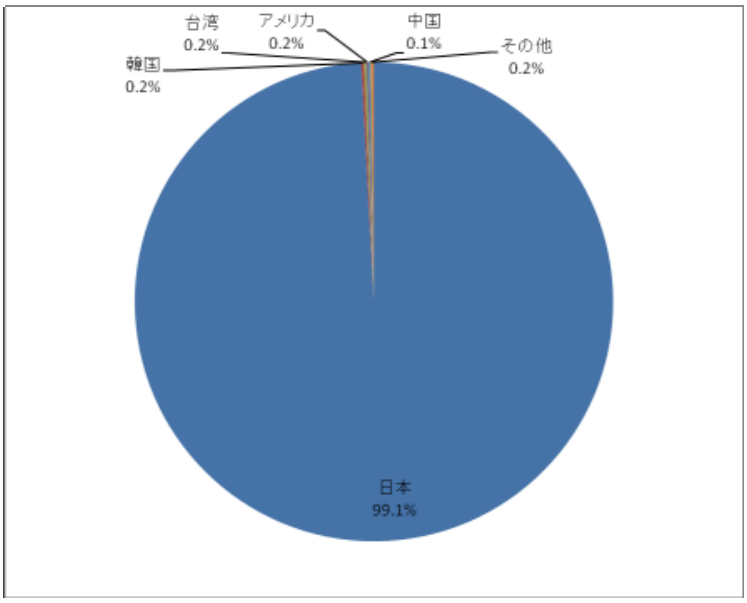


図 5　Winny 存在確認ノード 国・地域別分布(n=92,033)

２．Share EX2

（１）無許諾コンテンツの流通状況

流通コンテンツのうちおよそ半数のコンテンツが著作物と推測される。著作物と推測されるコンテンツの内訳は図７の通りである。

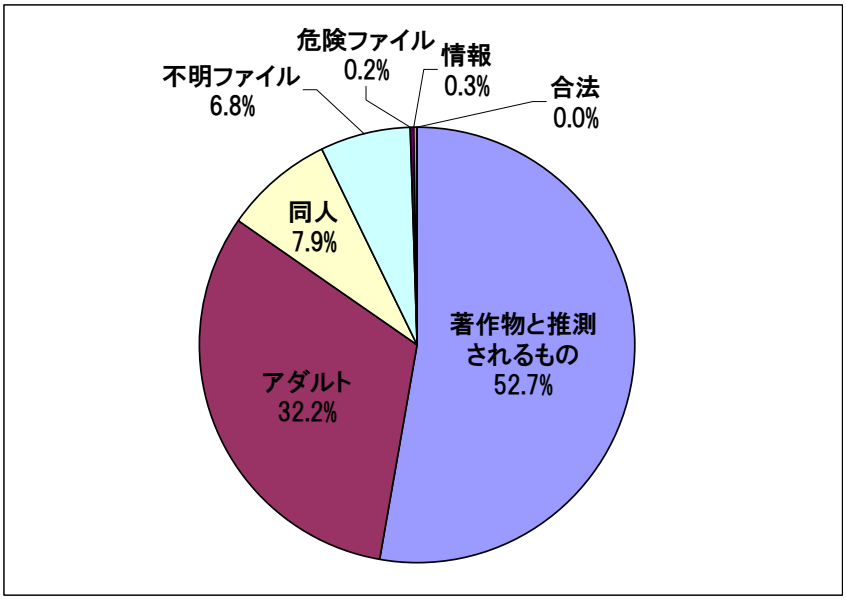


図 6 Share コンテンツ流通状況(n=20,000)

※「著作物と推測されるもの」とは本調査で権利の所在が推定できるもの

※「アダルト」、「同人」とは本調査で権利の所在が判別できないため、権利の対象に関しての調査は見送ったもの

※「不明ファイル」とはタイトルからコンテンツの内容が判別できないもの

※「危険ファイル」とはタイトル、拡張子からウィルスなどと推定されるもの

※「情報」とはウィルス感染などで流出した個人・組織等の情報だと推定されるもの

※「その他」とはコンテンツの分類が音楽、映像関連、プログラム、書籍関連に含まれないもの

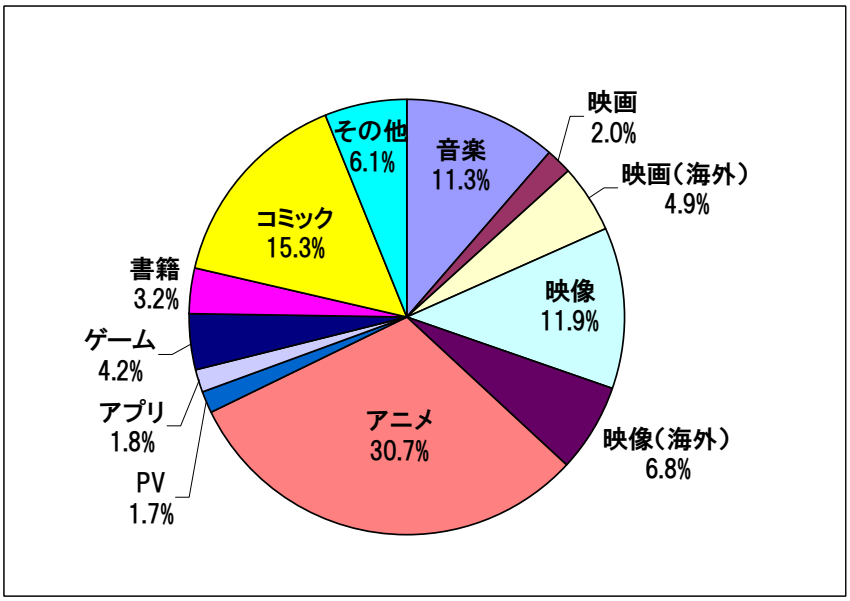


図 7　Share 著作物と推測されるコンテンツの内訳(n=10,533)

（２）権利の対象性について

著作物と推測されるもののうち、約 98%に権利があり、かつ許諾がないものだと推定される。

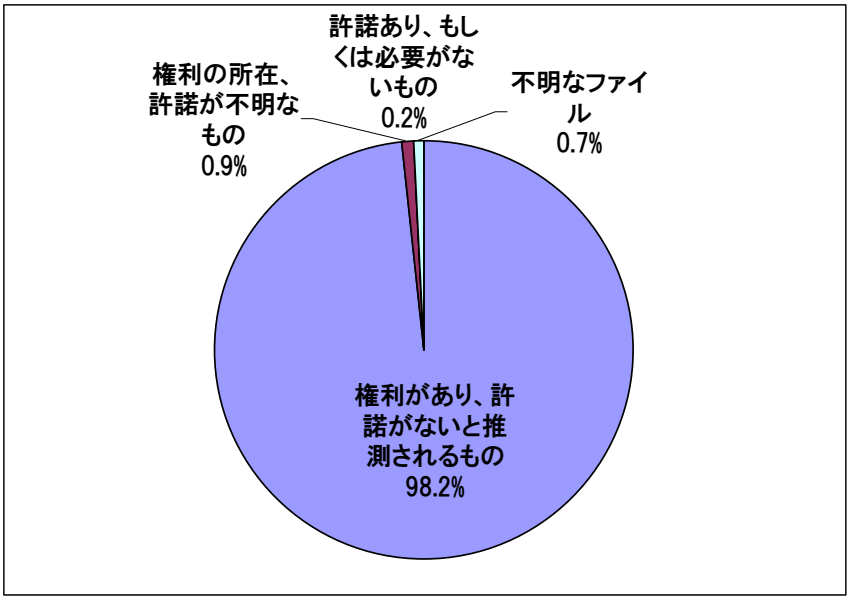


図 8　Share 権利の対象性（全体）(n=10,533)

（３）ファイル量について

Share 上ではファイルの情報はファイル本体から計算で算出されるハッシュ値を用いて管理されているため、一意なハッシュ値の数を算出する事で Share 上の流通しているファイルの量を推定できる。本調査では、一日で 697,575 件の一意なハッシュ値が収集された。全数としてはおよそ 75 万～80 万件程度と推定される。

（４）ノード量について

本調査では IP アドレスとポート番号の一意な組み合わせをノードの量として算出した。本調査では、一日で 208,825 件の一意なノード情報が収集された。全数としてはおよそ 21 万～22 万ノード程度と推定される。

（５）検出ノードの国・地域について

一日で検出した 208,825 件のノードの国・地域を調査した結果、94%が日本国内 IP での利用であった。

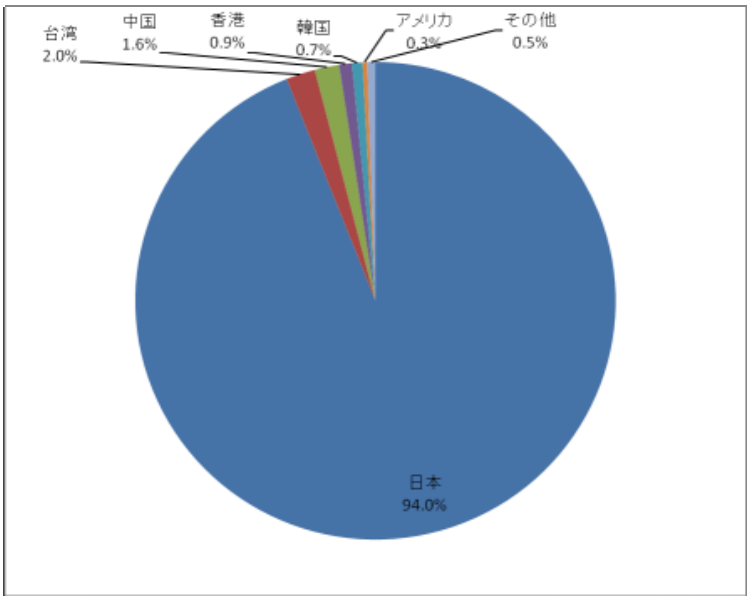


図 9 Share 検出ノード国・地域別分布(n=208,825)

３．Gnutella

（１）無許諾コンテンツの流通状況

流通コンテンツのうち 90%以上のコンテンツが著作物と推測される。著作物と推測されるコンテンツの内訳は図 11 の通りである。

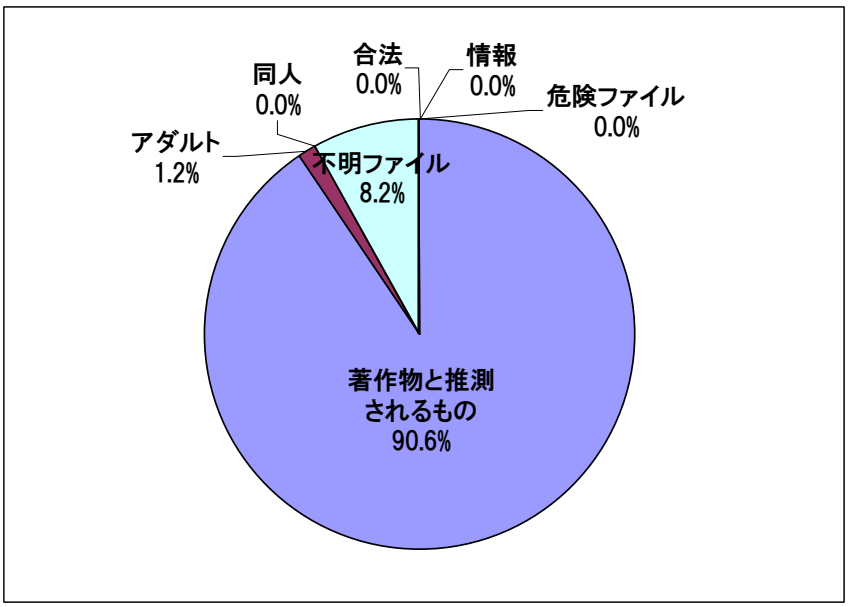


図 10 Gnutella コンテンツ流通状況(n=20,000)

※「著作物と推測されるもの」とは本調査で権利の所在が推定できるもの

※「アダルト」、「同人」とは本調査で権利の所在が判別できないため、権利の対象に関しての調査は見送ったもの

※「不明ファイル」とはタイトルからコンテンツの内容が判別できないもの

※「危険ファイル」とはタイトル、拡張子からウィルスなどと推定されるもの

※「情報」とはウィルス感染などで流出した個人・組織等の情報だと推定されるもの

※「その他」とはコンテンツの分類が音楽、映像関連、プログラム、書籍関連に含まれないもの

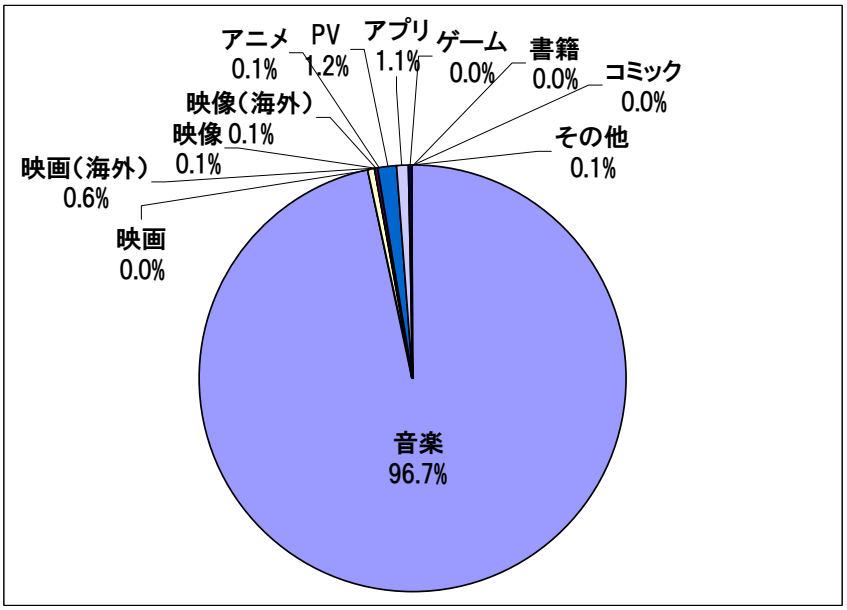


図 11　Gnutella 著作物と推測されるコンテンツの内訳(n=18,122)

（２）権利の対象性について

著作物と推測されるもののうち、99%に権利があり、かつ許諾がないものだと推定される。

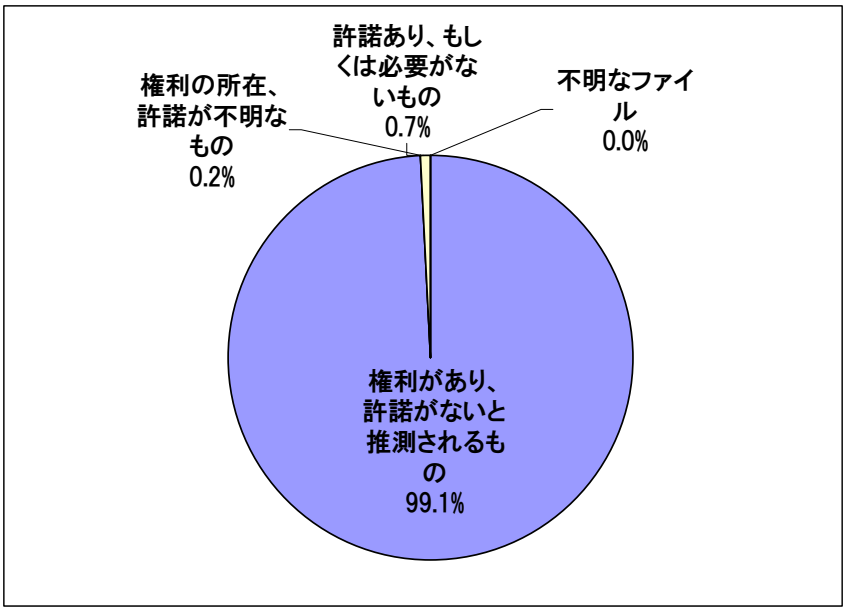


図 12　Gnutella（Limewire、Cabos）権利の対象性（全体）(n=18,122)

（３）ファイル量について

Gnutella 上では Winny,Share 同様ファイルの情報はファイル本体から計算で算出されるハッシュ値を用いて管理されている。ただし、Winny、Share と比較して、ネットワークが世界中に広がり、規模が大きいため、スレッド数と取得件数が比例している。

このことから、本調査ではキー情報クローラを用いて Gnutella ネットワークの全量を取得するのではなく、ネットワークに流れる一部のキー情報を抽出した。

（４）ノード量について

本調査では IP アドレスとポート番号の一意な組み合わせをノードの量として算出した。前項と同様にGnutellaネットワークに流れる一部のノード情報用いてGnutellaネットワークの全量を取得するのではなく、ネットワークに流れる一部のキー情報を抽出した。

（５）検出ノードの国・地域について

Winny や Share と異なり、日本以外の利用が約 96%を占めた。特にアメリカが約 57%以上と半数以上のノードが検出された。昨年度の調査では日本の割合は 2.26%で全体の 8 位だったが、今年度の割合は 3.67%で 4 位となり、利用者が増加していると考えられる。

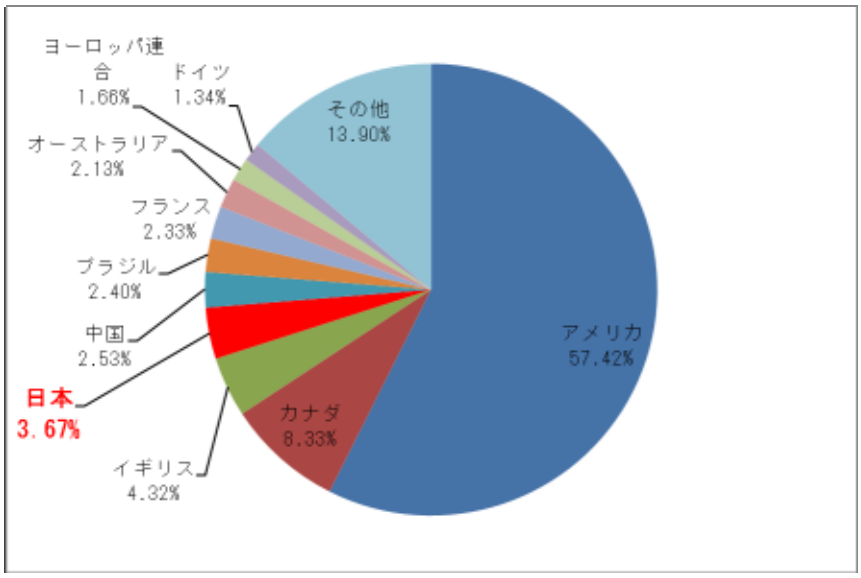


図 13 Gnutella 検出ノード国・地域別分布(n=669,197)

（６）ユーザーエージェントの分布について

各クローラが接続した際、ノードから通知されたユーザーエージェント(利用ソフト、バージョン)の分布を調査した。LimeWire 系のソフトウェアの割合が高いことが分かる。

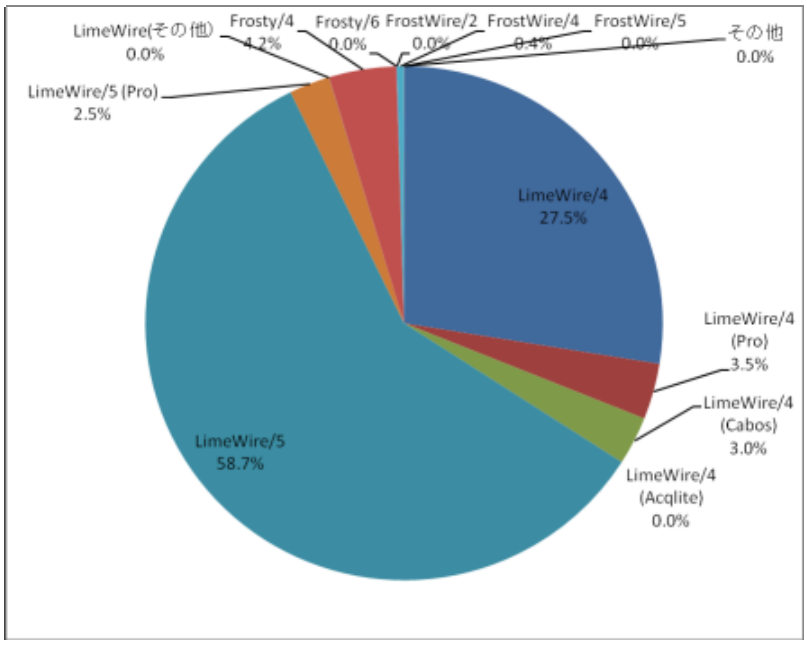


図 14 Gnutella ユーザーエージェント分布(n=34,624)

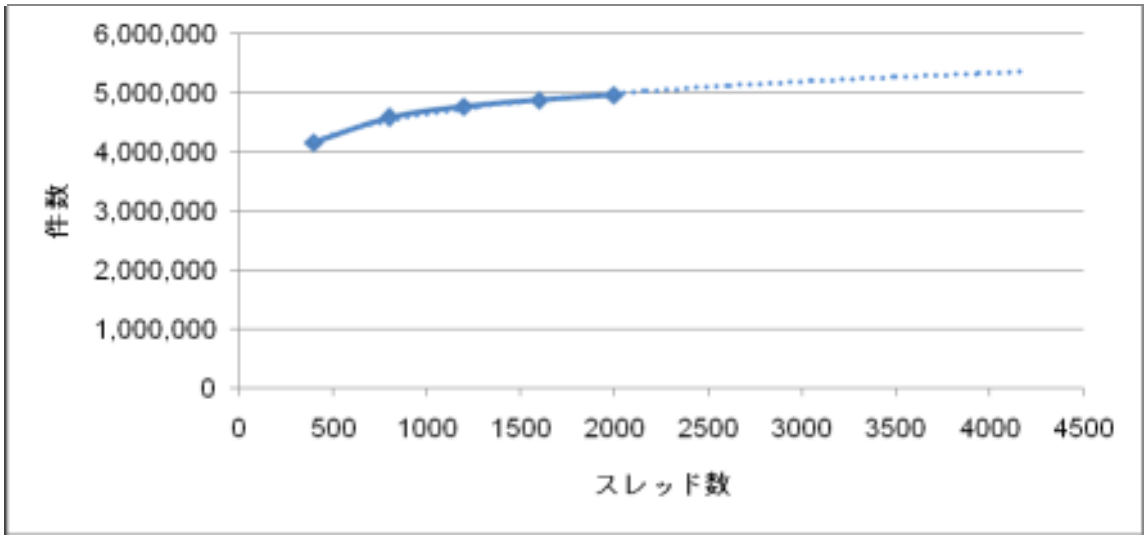
参考：Winny 追加調査

Winny で数ヶ月間にわたり行われていたランダム IP を利用したインデックスポイゾニングは 2009 年 10 月 30 日で終了したと思われる。今回の調査では IP アドレスが要素となるデータはインデックスポイゾニングの影響を受けたため、実数を把握するのが困難であった。

そのためインデックスポイゾニングの影響がなくなった 2009 年 11 月 13 日 17:00～2009 年 11 月 14 日 17:00 の 24 時間で Winny に関して同様の調査を行った。

ファイル量について

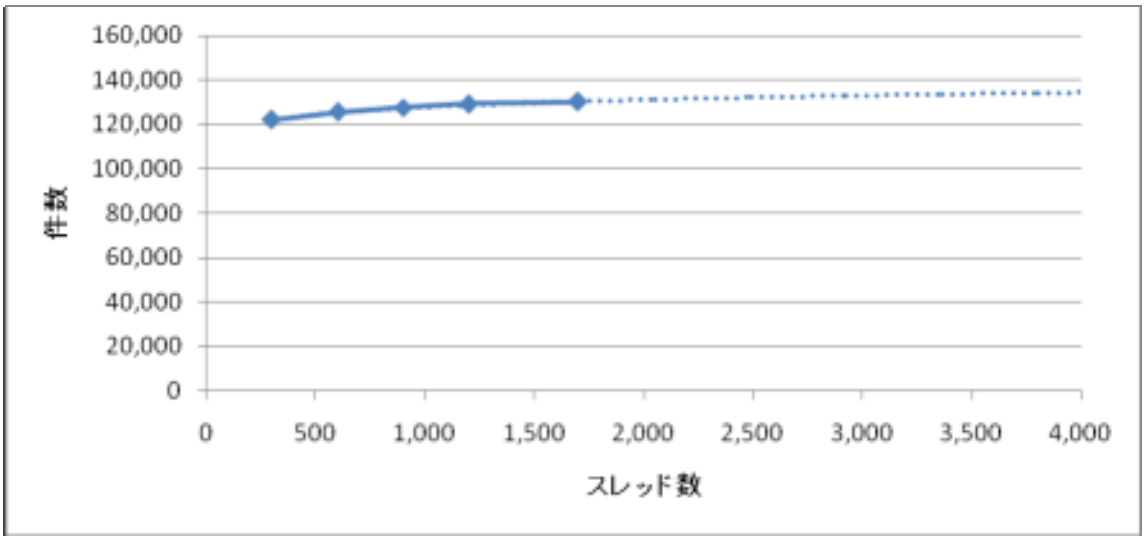
本調査では、一日で 4,966,833 件の一意なハッシュ値が収集された。全数としてはおよそ 550 万件程度と推定される。



参考図 1 Winny 一意なハッシュのスレッド数増加に伴う収束

ノード量について

本調査では、一日で 130,085 件の一意なノード情報が収集された。全数としては約 14 万ノード程度と推定される。



参考図　Winny 一意なノードのスレッド数増加に伴うグラフ

補足

調査期間について

今回の調査は 2009 年 10 月 2 日 17:00 から 2009 年 10 月 3 日 17:00 までの 24 時間で実施した。10 月 3 日が土曜日だったため、日中の Winny、Share の稼働ノードが平日よりも多めに観測されている。週末の日中の稼働ノードが多い傾向は継続的に観測されている。

Winny のインデックスポイゾニングについて

調査期間中、第三者によって、偽装キーの Winny ネットワークの注入（インデックスポイゾニング）が行われていたことが観測結果より推測される。このインデックスポイゾニングにより、Winny の検出結果が昨年までと大きく変化している。

インデックスポイゾニングに用いられた IP アドレスはランダムに生成された IP アドレスと考えられ、通常では存在しない IP アドレスも多く検出されている（例：0.1.2.3 など 0 から始まる IP アドレスなど）。

そのため、ノードの検出数が見た目に多く検出され、昨年までの Winny のノード数から増加したように見えるが、単純に増加したとはいえない。また、偽装されたキーと正常なキーを判別することは非常に困難である（異常な IP アドレスを削除することは可能であるが、正常に見えるキーの中にも偽装されているキーは含まれているため）。

Winny のキー情報について

Winny のキー情報は流通する過程で、稀に一部が変更されてしまう現象が確認されている。そのため、ファイル名やハッシュ、ファイルサイズなどが本来のものではないキーが観測されることがある。

Winny の検出ノード数について

Winny の検出ノード数には Port0 設定で利用しているノードは含まれていない。Port0 設定を行っている Winny のキー情報は中継している Winny ノードのキーとして流通するため、Port0 のノード（IP,PORT の組み合わせ）はキー情報では検知されないためである。

Gnutella の調査について

Gnutella ネットワークは全世界に広がっており、Winny、Share に比べてはるかに多いノードが検出されている。そのため、ネットワーク全体の総量ではなくサンプリングした割合で算出している。

参考文献

1)「Winny の技術」　金子勇　2005 年

2)「クローリング手法を用いた P2P ネットワークの観測」　寺田真敏ほか　2007 年

3)「Winny ネットワークにおけるインデックスポイゾニングの適用と評価」　吉田雅裕ほか　2008 年