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）

２００８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
１．調査の前提 1
（１）Winny2 1
（２）Share EX2 1
（３）Gnutella（Limewire、Cabos 等） 1
２．データの抽出 2
（１）フィルタリング 2
（２）権利の対象性算出方法 2
Ⅱ．調査結果 3
１．Winny2 3
（１）無許諾コンテンツの流通状況 3
（２）権利の対象性について 4
（３）ファイル量について 4
（４）ノード量について 4
（５）検出ノードの国・地域について 4
２．Share EX2 5
（１）無許諾コンテンツの流通状況 5
（２）権利の対象について 6
（３）ファイル量について 6
（４）ノード量について 6
（５）検出ノードの国・地域について 6
３．Gnutella（Limewire、Cabos 等） 7
（１）無許諾コンテンツの流通状況 7
（２）権利の対象性について 8
（３）検出ノードの国・地域について 8
補足 9
調査期間について 9
Winny の検出ノード数について 9
Gnutella（Limewire、Cabos 等）の調査について 9

Ⅰ．調査概要

１．調査の前提

調査は 2008 年 9 月 19 日 17:00 から 2008 年 9 月 20 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実際のネットワークをクローリングし、実際に流通している情報を取得、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） 2008 年 9 月 Version
・設定情報	総スレッド数 1600

（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） 2008 年 9 月 Version
・設定情報	総スレッド数 4300

（3）Gnutella（Limewire、Cabos 等）

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（Limewire、Cabos 等）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 2 種類あり、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」と、ハンドシェイク時に Crawler ヘッダを指定し、ノード情報を取得する「ノード情報クローラ」を用いている。

■基礎情報

・利用したソフトウェア	P2P FINDER（Gnutella） 2008 年 9 月 Version	
・設定情報	キー情報クローラ	総スレッド数 150
	ノード情報クローラ	総スレッド数 1800

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）、２万件のデータ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、Winny：0.0323%、 Share： 0.131% 、Gnutella（Limewire、Cabos 等）：4.98%となった。その後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、データを目視において確認し、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 20,000 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アダルトキーワード（「18 禁」「無修正」など）を除外**
⑤共通除外キーワード除去	**④のデータから共通除外キーワード（「同人」「ハッシュリスト」など）を除外**
⑥合法ファイル抽出	**④のデータから「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

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約 50%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。

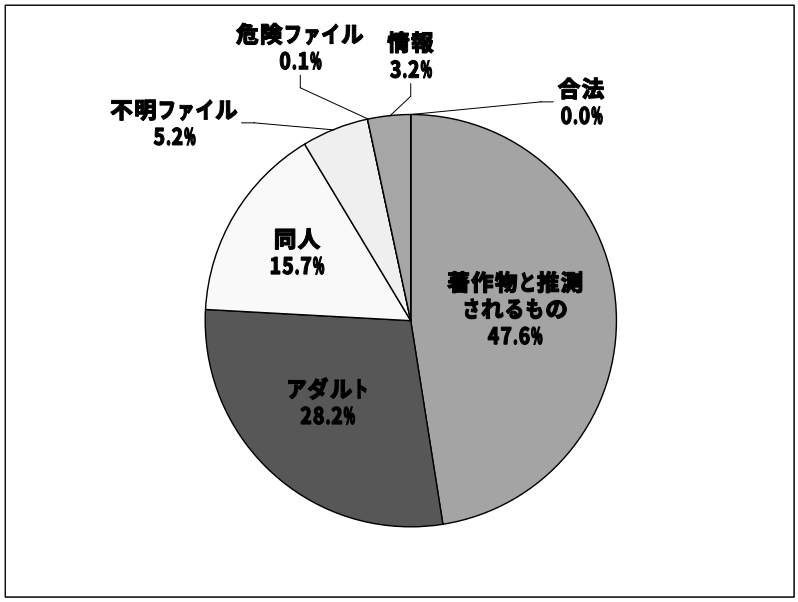

図 1 Winny コンテンツ流通状況（n=20000）

※「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

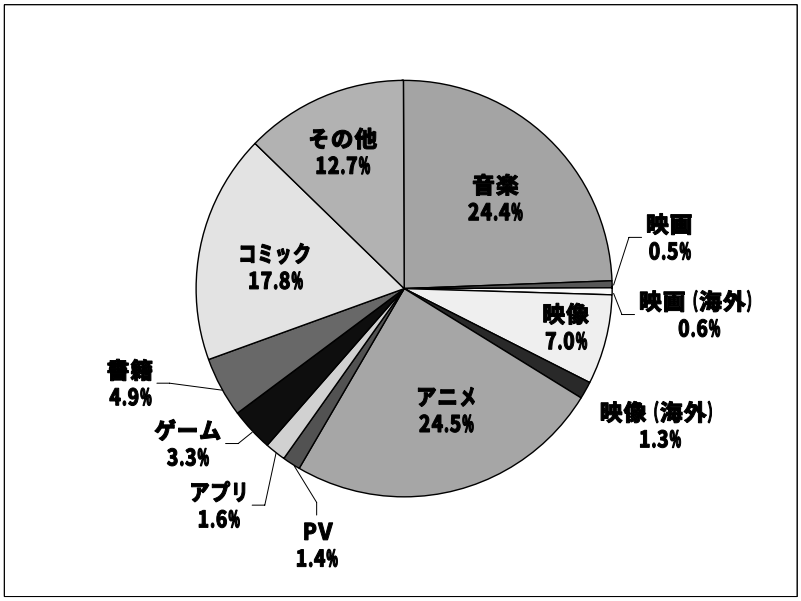

図 2 Winny 著作物と推測されるコンテンツの内訳（n=9518）

（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 97%に権利があり、かつ許諾がないものだと推定される。

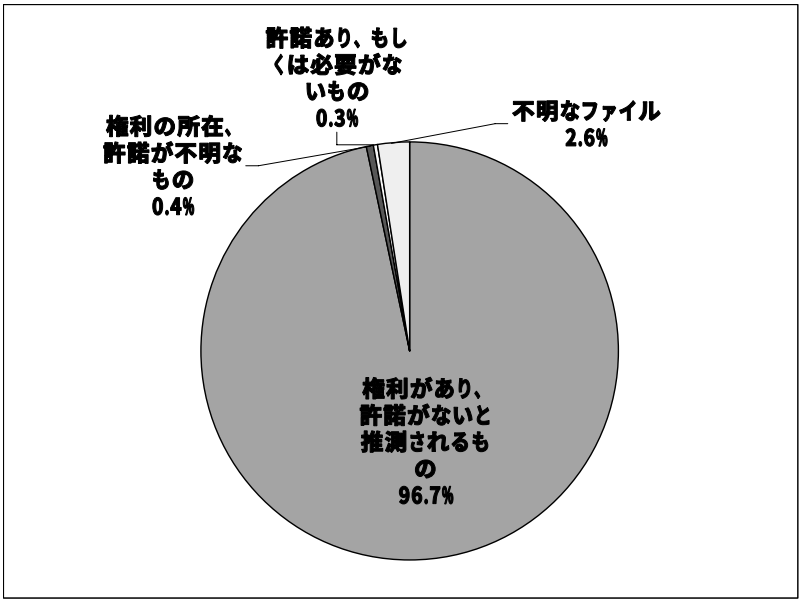

図 3 Winny 権利の対象性（全体）（n=20000）

（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 5,316,576 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。その結果、一日で 181,487 件の一意なノード情報が収集され、全数としてはおよそ 19 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 181,487 件のノードについて国・地域を調べた結果、約 98%が日本国内 IP の利用であった。

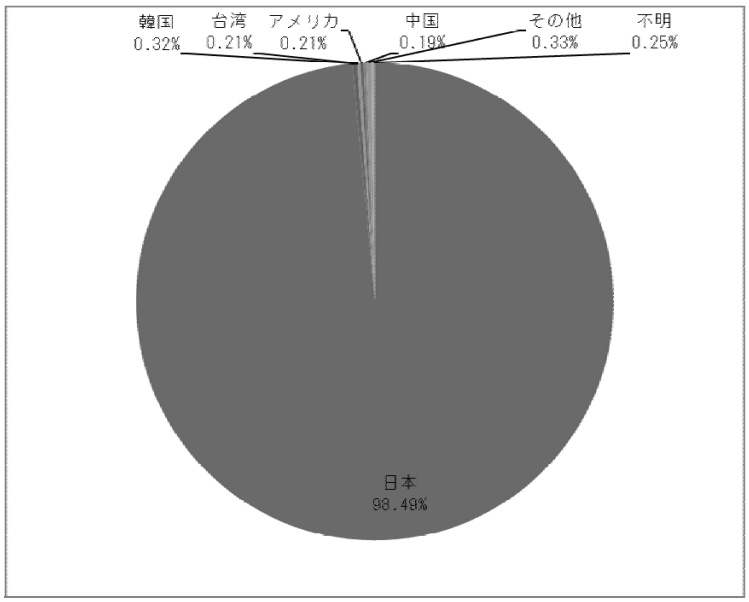

図 4 Winny 検出ノード国・地域別分布（n=181487）

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約 56%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。

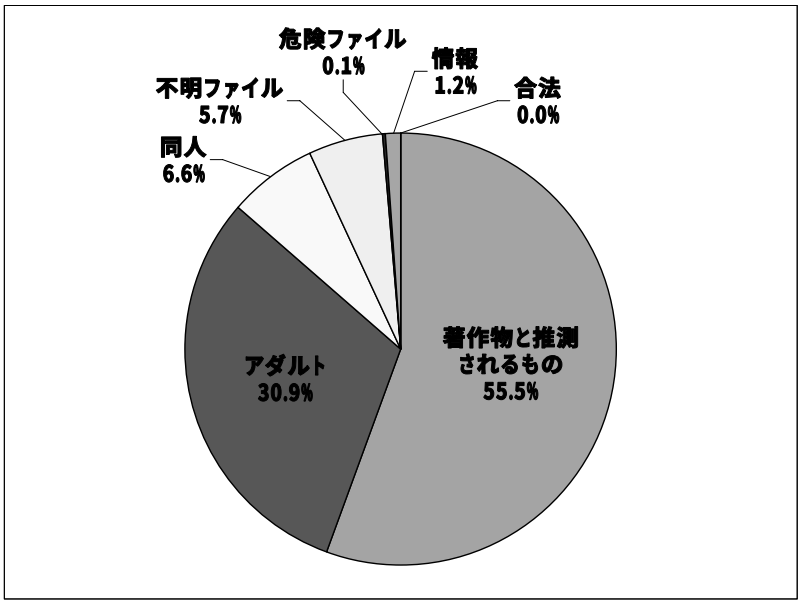

図 5 Share コンテンツ流通状況（n=20000）

※「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

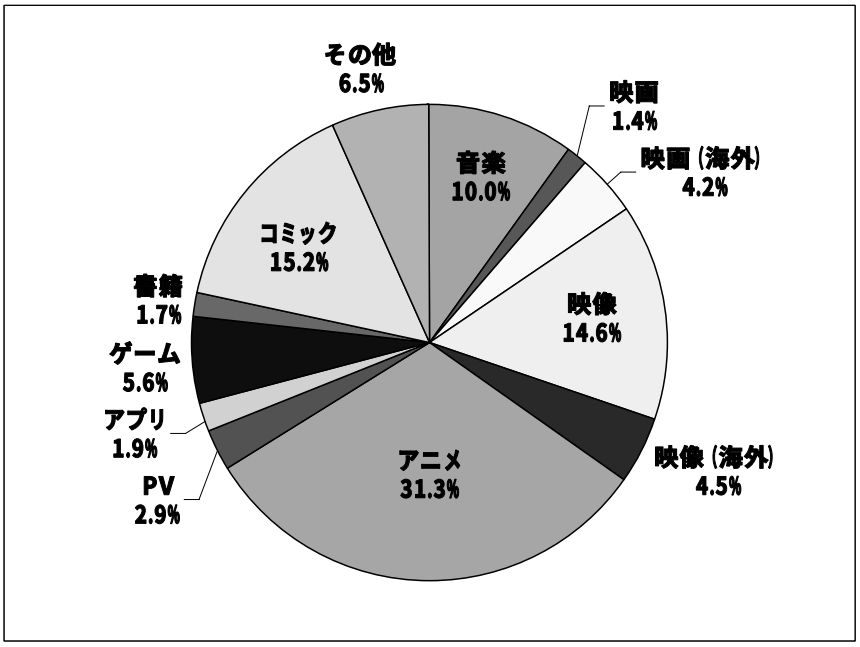

図 6　Share 著作物と推測されるコンテンツの内訳（n=11103）

（2）権利の対象について

著作物と推測されるもののうち、97%に権利があり、かつ許諾がないものだと推定される。

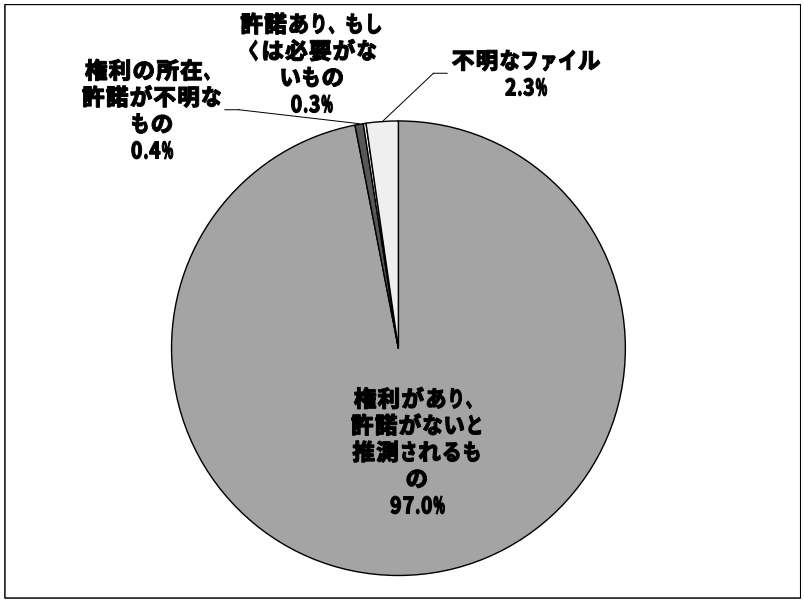

図 7 Share 権利の対象性（全体）（n=20000）

（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 712,144 件の一意なハッシュ値が収集され、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。その結果、一日で 209,367 件の一意なノード情報が収集され、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

一日で検出された 209,367 件のノードについて国・地域を調べた結果、約 94%が日本国内 IP の利用であった。

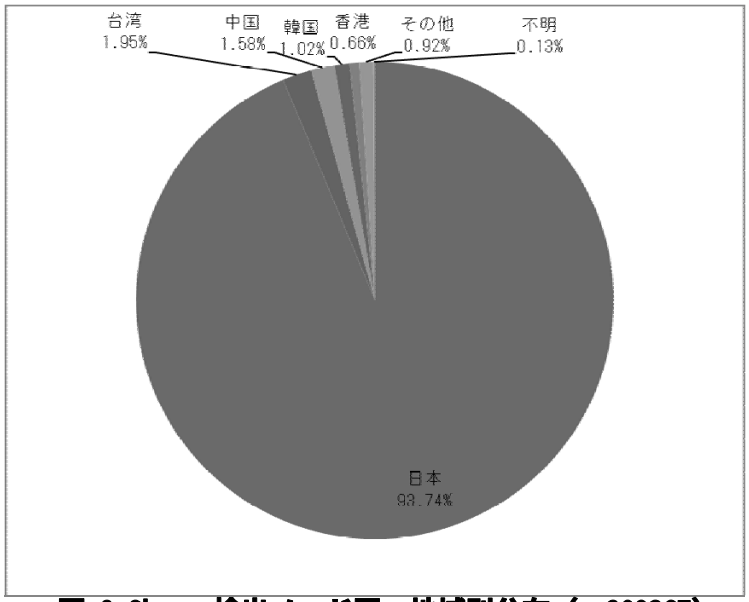

図 8 Share 検出ノード国・地域別分布（n=209367）

３．Gnutella（Limewire、Cabos 等）

（1）無許諾コンテンツの流通状況

流通コンテンツのうち約 80%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。

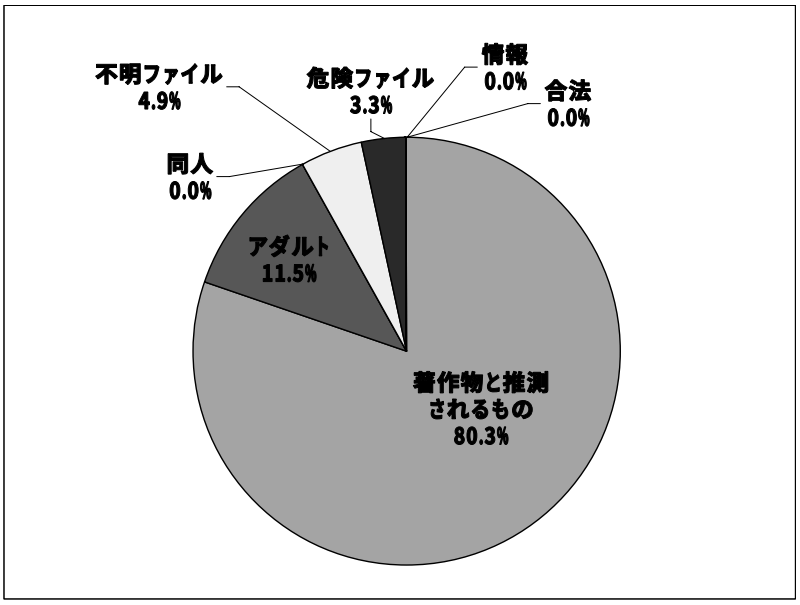

図 9 Gnutella（Limewire、Cabos 等）コンテンツ流通状況（n=20000）

※「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

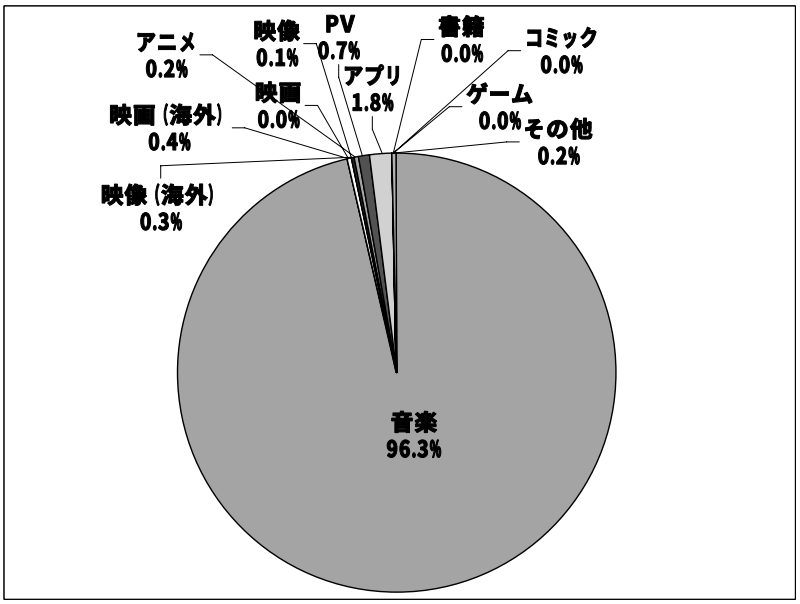

図 10　Gnutella（Limewire、Cabos 等）著作物と推測されるコンテンツの内訳（n=16063）

（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。

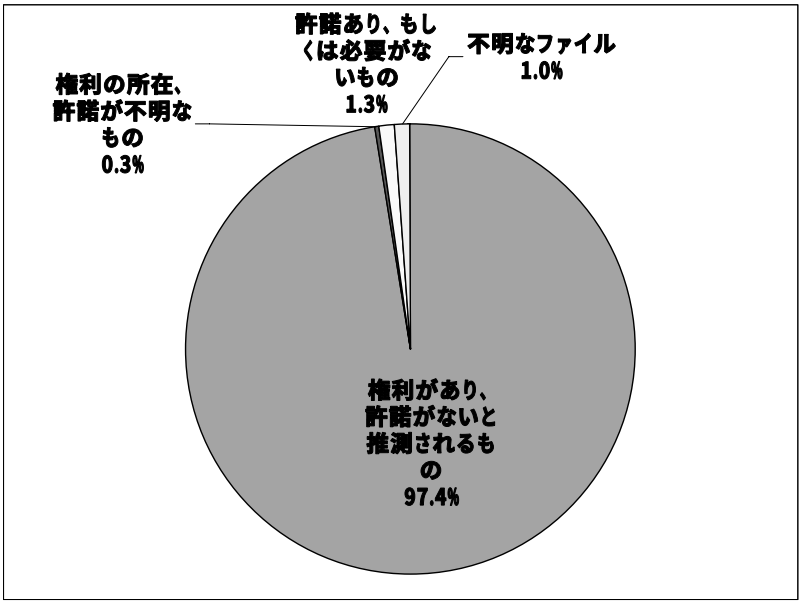

図 11　Gnutella（Limewire、Cabos 等）権利の対象性（全体）（n=20000）

（3）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 IP の利用が約 98%であった。特にアメリカが約 55%と半数以上のノードが検出された。

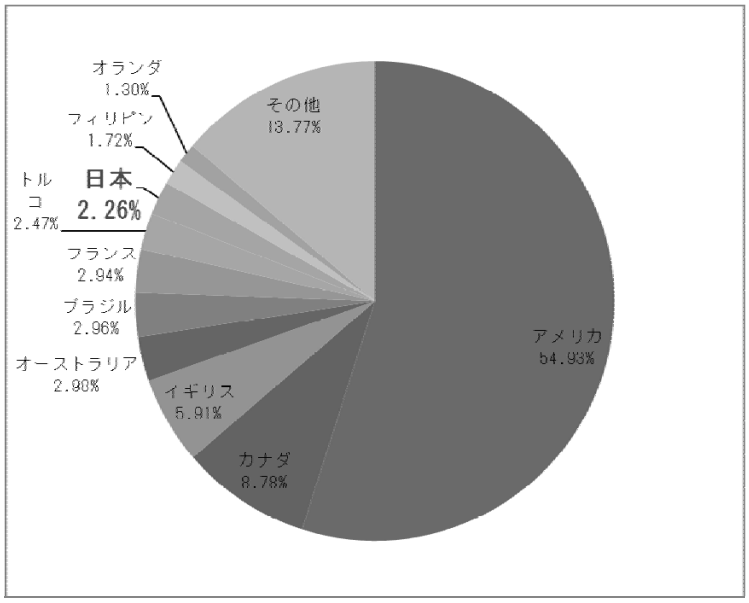

図 12 Gnutella（Limewire、Cabos 等）検出ノード国・地域別分布（n=771848）

補足

調査期間について

今回の調査は 2008 年 9 月 19 日 17:00 から 2008 年 9 月 20 日 17:00 までの 24 時間を調査した。2008 年 9 月 29 日が土曜日であったため、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（Limewire、Cabos 等）の調査について

収集効率を上げるため、ノード情報のみ収集するノード情報クローラと、キーワードその他の情報を収集するキー情報クローラを使い Gnutella ネットワークの情報を取得した。

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。