CREM
UMR CNRS 6211

Centre de Recherche en Économie et Management
Center for Research in Economics and Management

University of Caen
University of Rennes 1

Olson's Paradox Revisited: An Empirical Analysis of Incentives to contribute in P2P File-sharing Communities

Thierry Pénard
Raphaël Suire

University of Rennes 1, CREM-CNRS, M@rsouin

Sylvain Dejean

M@rsouin Telecom Bretagne, CREM

August 2011 - WP 2011-05

Working Paper

UNICAEN université de Caen Basse-Normandie
cnrs
UNIVERSITÉ DE RENNES 1

Olson's Paradox Revisited: An Empirical Analysis of Incentives to Contribute in P2P File-sharing Communities.

Sylvain Dejean[1]
M@rsouin, Telecom Bretagne
sylvain.dejean@telecom-bretagne.eu +33 (0) 229 001 445
GIS-M@rsouin, Technopole Brest Iroise
CS 83818 - 29238 BREST Cedex 3
Thierry Penard
CREM, Université de Rennes1, M@rsouin
thierry.penard@univ-rennes1.fr +33 (0) 223 233 520
Faculté des Sciences Economiques, University of Rennes 1, 7 place Hoche
CS 86514, F-35065 Rennes cedex
Raphaël Suire
CREM, Université de Rennes1, M@rsouin
raphael.suire@univ-rennes1.fr +33 (0) 223 233 336
Faculté des Sciences Economiques, University of Rennes 1, 7 place Hoche
CS 86514, F-35065 Rennes cedex

Janvier 2011

[1] This research received financial support from the Brittany Regional Council, as part of the P2Pimage research project. We also thank the participants at IIOC 2009 (Boston) and at the research seminars of Economix, CREM and Telecom Paris, as well as M. Arnold and M.D. Smith for their comments and suggestions.

Abstract:

This article aims to examine how the size of file-sharing communities affects their functioning and performance (i.e. their capacity to share content). Olson (1965) argued that small communities are more able to provide collective goods. Using an original database on BitTorrent file-sharing communities, our article finds a positive relationship between the size of a community and the amount of collective goods provided. But, the individual incentives to contribute slightly decrease with community size. These results seem to indicate that Peer to Peer file-sharing communities provide a pure (non rival) public good. We also show that specialized communities are more efficient than general communities to promote cooperative behavior. Finally, the rules designed by the administrators of these communities play an active role to manage voluntary contributions and improve file-sharing performance.

Classification code: H41; L86; K42

Keywords: Olson's paradox, collective goods, Peer-to-Peer, File-sharing, community.

Introduction

Olson (1965) developed a theory of collective action to explain why the existence of a common interest among a group of persons is not sufficient to act cooperatively. The outcome of collective action has the features of a public good because it benefits all people of the group regardless of their contribution (whether they have contributed a lot or have been free-rider). Olson argued that large groups are less able to promote their common interest than small ones because the individual incentives to contribute should diminish with group size. Olson's theory has influenced a large body of research in economics and politics. In particular, the presumed negative relationship between group size and the ability to provide collective good has been debated. Chamberlin (1974) considered that this depends on the nature of collective goods produced by the community. Rival goods are more likely to be provided by small groups, whereas inclusive or non rival goods are more efficiently produced in large groups. In the presence of rivalry, the portion of collective good appropriated by each member decreases as the size of the group rises. Hence, a large group reduces individual incentives to contribute and is less likely to succeed in providing collective goods. With inclusive goods, each additional member does not reduce the share of collective goods consumed by existing members. They will slightly reduce their contribution, but this will be largely compensated by the additional contributions of new members. This implies that the amount of voluntary contributions to a non rival good should increase with group size, contrary to the Olson conjecture (Mc Guire, 1974).

More recently, Esteban and Ray (2001) revisited Olson's "group size paradox". They showed that with increasing marginal costs on collective action, large groups are more efficient than small groups to provide collective goods, even in the case of rival goods. But, Pecorino and Temimi (2008) found that the Olson conjecture is satisfied when group members bear a fixed cost of participation and the collective good exhibits a high degree of rivalry (see also Palfrey and Rosenthal, 1984; Pecorino, 1999; Bergstrom, Varian and Blume, 1986; Andreoni, 1988; Gaube, 2001).

The impact of group size on voluntary contributions and free-riding has been examined in several experimental studies. Except for the article of Isaac, Walker and Williams (1994), most of them found a negative impact of group size on voluntary contributions in the context of public good experiments (Chamberlin, 1978; Isaac and Walker, 1988, Marvell and Ames, 1979). More recently, Zhang and Zhu (2010) examined the relationship between group size and incentives to contribute to Chinese Wikipedia. Using a

natural experience (the blocking of Wikipedia by Chinese authorities that prevent mainland Chinese to use and contribute to this open encyclopedia), they found that the non-blocked contributors decreased their contributions by 42,8% in reaction to the shrinking community size.

In this article, we want to revisit Olson's paradox in the context of Peer to Peer (P2P hereafter) communities. These virtual communities have some specific characteristics that are particularly interesting to test Olson's conjecture. First, these communities are more exposed to free-riding than physical communities because they gather anonymous and distant users (Adar and Huberman, 2000; DangNguyen and Penard, 2007; Krishnan et al., 2007). Secondly, these communities can be extremely volatile because the cost of entry and exit is low. They can attract thousands of new members in a few days, but their size can also rapidly decrease (Krishnan et al., 2003). Thirdly, data from P2P communities can be easily collected and it is possible to permanently keep track of the daily file-sharing activity in these communities.

Our article aims to empirically examine whether the size of P2P communities affects the provision of collective goods (i.e. the files shared within community). In other terms, does the size of a file-sharing community (measured by the number of active members) increase or reduce the incentives to contribute? In P2P file-sharing communities, voluntary contributions can take two different forms. First, members can feed the community with new content or files; i.e. they can upload and share files that will expand the catalog of content offered. Secondly members can share content that they have downloaded from other peers; i.e. after having downloaded a file, they can let this file available or accessible to the rest of the community (instead of removing it from the hard drive of their computer). In this case, they provide an additional source to download this file, and improve the speed and robustness of file-sharing.

This article is related to Asvanund et al. (2004) who analyzed how the size of music file-sharing communities may affect the availability and downloading quality of music files. The authors collected data on several public P2P networks (OpenNap) and found evidence of both negative and positive network effects in file-sharing communities. They estimated that the marginal benefit from an additional member decreases and the marginal cost increases with the size of the community. They concluded that the optimal size for an OpenNap community is bounded.

Our article seeks to extend Asvanund et al. (2004) analysis to P2P communities that share any kind of content (not only music) and use a supposedly more efficient protocol (the BitTorrent protocol). Contrary to the Napster protocol (studied in Asvanund et al., 2004), the

BitTorrent protocol prevents congestion by forcing users to share files during the time they are downloading them. In this case, users are contributing in an involuntary way to the community while downloading. But they can also voluntarily contribute by uploading new files or by letting accessible the files they have completely downloaded. How does community size influence the incentives to contribute and how does it impact the amount of collective good (the number of unique files available in the community)? Moreover, the administrators of BitTorrent communities have the possibility to set some specific organizational rules, aimed at screening new members, monitoring behavior, or filtering content for instance. How does the design of file-sharing communities affect the incentives to contribute and community performance?

Our article uses data collected on 42 private BitTorrent communities during two months (Dec. 2007-Feb. 2008). Our communities require members to be registered and sometimes to be co-opted by a member of the community. For each community and twice a day, we gathered information on the number of members, the number of files available, and the number of seeders and leechers. A seeder is a peer who lets an entire file available for download and the leecher is a peer who is currently downloading a file. From the data, we compute the ratio of seeders to leechers that indicates the quality of downloading and is an indirect means of measuring the incentives to voluntarily contribute within a community. Our findings show a positive relationship between the total amount of collective goods provided and the size of a P2P community. But, the number of members has a negative impact on the quality of access to these collective goods (i.e on the ratio of seeders to leechers). According to Chamberlin (1974), these results suggest that the outcome of P2P communities is a pure public good (inclusive or non rival).

The rules designed by the administrators of these communities have also a significant impact on voluntary contributions. Our results highlight the fact that these communities are very innovative to regulate and manage indirect social interaction between anonymous and distant peers. Moreover, communities that share specialized content and filter their members are more efficient in the provision of collective goods.

The article is organized as follows. In the next section, we describe the BitTorrent file sharing protocol and the individual contribution mechanisms. The dataset is described in the section 3. Section 4 presents the econometric model and comments the results. Section 5 concludes.

2. BitTorrent file sharing system

BitTorrent is now the most popular P2P file-sharing protocol in the world. Originally, in 2001, Bram Cohen designed this protocol to improve file-sharing for large size files. BitTorrent is a “non pure” peer-to-peer system in which a central server, called the “tracker” collects information on the resources peers want to share (meta-data on the size, name and description of the shared files) and coordinates the transfer of files among users.

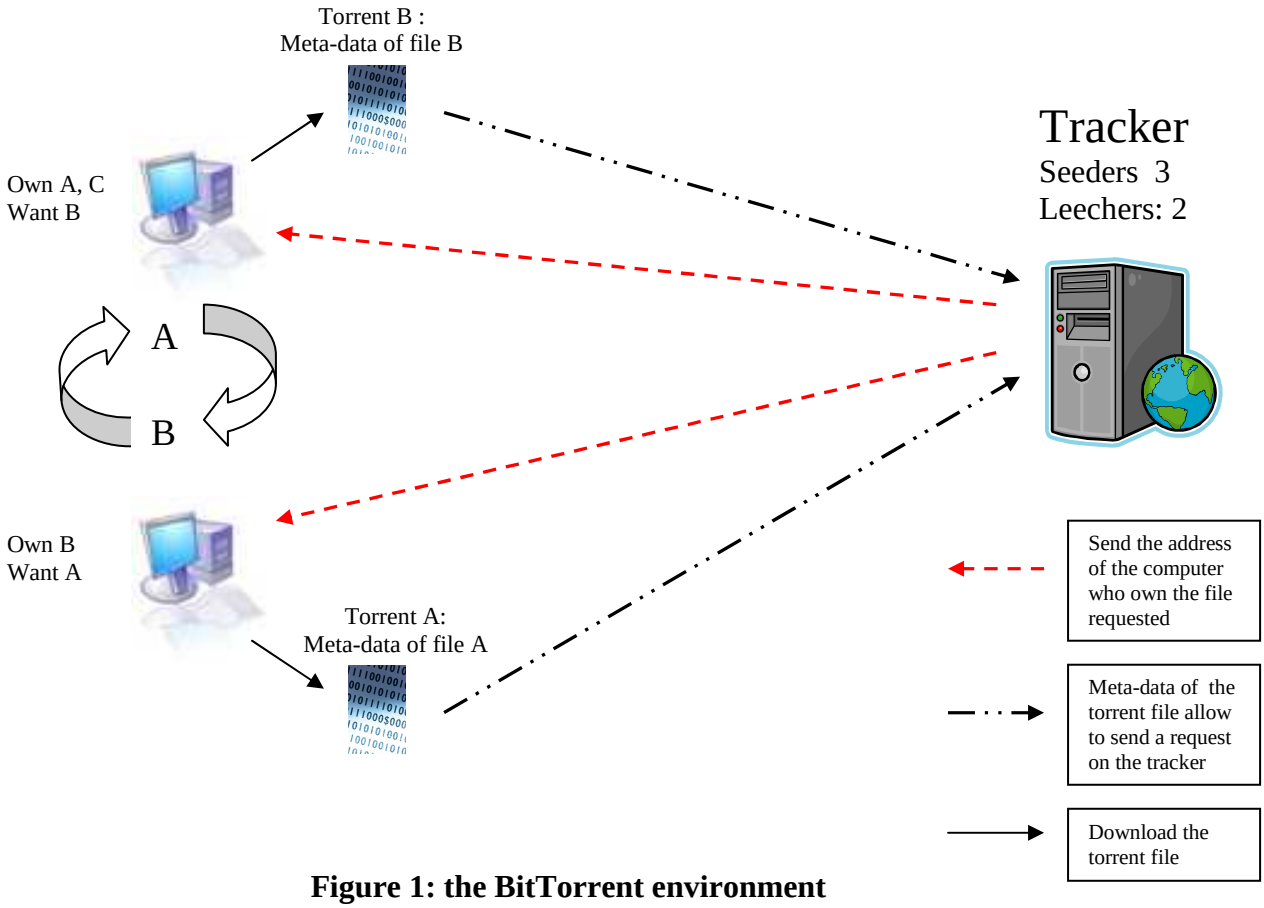


Figure 1: the BitTorrent environment

To download a file, the user has first to install a BitTorrent client (Azureus, Bit comet, µtorrent). Then, the user has to connect to a tracker that will send the address of the torrent that contains the desired file. To optimize the bandwidth allocation, files are divided into identically sized pieces called “schunk” and can be reconstituted only with “hashing information” contained in the torrent file. Once connected to the tracker, the uploaders and downloaders of a file are automatically in contact with each other via their BitTorrent client and are exchanging pieces of files (figure 1). By helping users to find each other, the tracker also gathers statistical information about downloads and uploads. A user who is downloading a file indexed in the tracker, is called a “leecher” and a user who lets the entire file accessible

to other users is called a “seeder”. The sum of leechers and seeders corresponds to the number of “peers” (a peer refers to a user who owns at least a piece of this file)[2].

Opportunistic behavior and congestion were strong issues in the first generation of P2P file sharing systems (Adar and Huberman, 2000; Krishnan et al., 2007). The BitTorrent protocol was designed by Bram Cohen to overcome this issue. It is based on tit-for-tat mechanism of file-sharing that imposes a minimum of cooperation (Cohen, 2003). Each peer is modeled as an intelligent automaton that maximizes its own interest (i.e. the downloading rate), rewarding peers who cooperate and punishing those who do not share. The more pieces of files a leecher is uploading towards another peer, the more pieces of files he can download from that peer.

For the user, the BitTorrent protocol is transparent and the process described in figure 1 is automated by the P2P client software. Unlike Napster or Gnutella, the user is automatically sharing the pieces of files that she is currently downloading. So she can never be a pure free rider and this reduces the risk of congestion (the higher the number of peers downloading the same file, the larger the number of sources to download pieces of this file in the meantime). However, forced sharing (while downloading) is not sufficient to guarantee the long-term viability of a community. Voluntary contributions are also important to feed the community with new content or files (by expanding the quantity and diversity of content) and to preserve the existing contents (by replicating the files). Voluntary sharing increases the quantity of files available for downloading and the quality of downloading.

What are the costs and the benefits for a member to contribute voluntarily (i.e. to upload new files or to keep a downloaded file accessible to other members)? The benefit of sharing is to increase one’s ratio of uploading to downloading. A better ratio can provide some privileges or priority in many BitTorrent communities (possibility to download more files and more rapidly). Interestingly this individual benefit may increase with the number of members in the community, because more members mean more potential downloaders (leechers) for the files that are shared by a seeder (*network effects*). A contributor can increase more rapidly her individual ratio of uploading/downloading in a large community than in a small one, and can be more rapidly eligible to the privileges reserved to the active contributors. But with a larger community, a seeder may compete with more users who offer the same files (*competition effects*). Consequently, the benefits of sharing a file could decrease with community size if competition effects are larger than network effects.

[2] In the terminology of bitTorrent protocol, peers who share the same torrent constitute a swarm.

The cost of contributing in a file-sharing community depends on the nature of the voluntary contribution. The cost of sharing an existing file (after having downloaded it) is mainly the opportunity cost of the resources that are used (the hard disk space used for the storage of the shared files and the bandwidth shared with downloaders). The cost also increases with the perceived risk of being caught and fined for illegal file-sharing increases. These costs tend to be independent of the size of the community[3]. A second way to contribute more actively is to feed the community with new content. This form of contribution is more costly that sharing existing files. Uploading a new file in a P2P community is a long and complex process that requires some skills. The seeder must check that the file is not already available and that the quality fits well with the standards of the community. After having converted the file in the appropriate format and uploaded it on the server, the submission has to be approved by the community moderators before being available for download. The costs of feeding the community are also size-independent and comprise both cognitive and material costs.

Consequently, the individual incentives to contribute should increase (*decrease*) with community size only if the individual benefits of sharing increase (*decrease*) with the number of active members within the community (as the cost is presumed to be size-independent). In other words, if network effects dominate competition effects, contributors will be more incite to share files as new members enter into the community.

The following table summarizes the forms of contributions and the related costs and benefits depending on the role played within the community.

[3] This is obvious for the hard disk space used to store files. For the shared bandwidth, one can object that this cost could increase with the size of the community as the number of potential downloaders rises. But seeders can always control the bandwidth that they want to share. By setting an upper limit to the shared bandwidth, they make sharing costs less sensitive to community size.

Table 1 - Characteristics of contribution behavior in a BitTorrent community

	Contribution	Nature of sharing	Costs	Benefits
Leecher	Involuntary	Pieces of the downloaded file during the downloading process		
Seeder	Voluntary	100% of a downloaded file	Opportunity cost of used resources (Hard disk and bandwidth) + the risk of being caught and fined for illegal file-sharing	An increase of the individual upload ratio with the associated privileges
Feeder and seeder	Voluntary	100% of a new file	Opportunity cost of used resources and time (Hard disk and bandwidth) + the risk of being caught and fined for illegal file-sharing + Learning costs	An increase of the individual upload ratio with the associated privileges

The different forms of contribution in BitTorrent communities (involuntary or voluntary) are of particular interest to study the relationship between incentives to contribute and group size. How does community size affect feeding and sharing behavior? How does it impact the size of the catalog (i.e. the number of unique files uploaded in the community). By measuring the ratio seeder over leecher we have a good proxy for the share of voluntary contribution in P2P communities. In the next section, we present data and the methodology used to test the Olson conjecture.

3. Data

Summary statistics

Our sample is composed of 42 P2P file-sharing communities that can be either general or specialized in a type of content (music, movies, sport, adult, video games, and e-learning)[4].

[4] A description of the 42 trackers is given in Annex 1 (location, category).

All of them are "private" and "semi-private" trackers which contrary to "public trackers" (or open P2P communities) require every user to be registered. These communities were randomly selected on the directory TorrentKing that listed several hundred communities at this time. Between December 17, 2007, and February 17, 2008, and twice per day (at 10 am and 10 pm GMT[5]), we collected the number of unique files available[6], the number of users registered as well as the number of seeders and leechers for each community. The panel gathers 5,097 observations (42 communities observed during 125 periods with 153 missing values).

We computed the ratio of seeders to leechers to obtain a measure of the propensity of members to contribute voluntarily. The higher the individual incentives to contribute, the larger the ratio of seeders to leechers. This ratio and the number of unique files are two complementary measures of the provision of collective good in file-sharing communities: the latter indicate the quantity of collective goods and the former the quality of access to these goods (a higher ratio ensures better download quality).

Table 2 shows that the mean of community size is of 101,721 members, and the mean of seeders and leechers (at a given time) is respectively 28,600 and 12,967. Significant size differences in terms of seeders, leechers and registered users exist among the 42 Peer-to-Peer communities (from 556 registered members for the smallest community to more than 1.8 millions of members for the largest, with a median of 10,496 members). Free-riding (i.e. being only a leecher without contributing as a seeder) seems to be limited in these 42 private communities. The ratio of seeders to leechers is 14.1 on average with a median of 6.41, a maximum of 242 and a minimum of 0.46. Finally, the mean number of unique files is 6,229 with a median of 1,652 files.

[5] We collected data at 11 am and 11 pm in France that is one hour ahead of Greenwich Mean Time.
[6] Files are called the "torrents" in the BitTorrent terminology.

Table 2: Summary statistics

	Observation	Mean	Standard Deviation	Minimum	Maximum	Quartiles		
						25%	50%	75%
Seeders	5097	28600	72967	20	406838	1587	4933	19092
Leechers	5097	12967	42621	1	337372	121	673	3535
Ratio seeders to leechers	5097	14.1	27.4	0.46	282	3.73	6.41	10.7
Unique files	5097	6299	13310	33	74635	562	1652	4626.5
Registered members	5097	101721	343722	556	1804581	4635	10496	31789

The heterogeneity in our sample of P2P communities seems to be related to the nature of shared content. Some communities are "general" and provide various contents, like movies, TV series, music, video games or software. Others are specialized in a category of content and only accept the sharing of files belonging to this category.

Table 3 displays the features of P2P communities per type of content shared. Seven categories of communities have been considered (Generalist, Music, Adult, Movies, Video Game, E-learning, Sport). The Kruskal-Wallis test shows that these groups of communities are significantly different in terms of size and behavior. The comparison of ratio suggests that free-riding is more widespread in our adult content communities than in music or e-learning communities. The sample of adult content communities is also characterized by a larger number of registered users.

Table 3: Descriptive statistics (mean) by category of communities

	Generalist	Music	Adult	Cinema	Video Game	E-learning	Sport	Kruskal-Wallis test
Number of communities	18	8	4	3	3	1	5	
Seeders	18646.5	12063	164712	12105.5	8464.5	3491.8	6169	***
Leechers	16118.7	1622	55970	2536	1291.6	42.46	1154.8	***
Ratio seeders to leechers	8.37	30.81	2.37	5.59	8.61	103.75	6.51	***
Unique files available	5178.2	5985	21488	7337	2076.8	2089.72	1282.96	***
Registered members	21698.5	31919.5	833116	19739.9	17876.8	11643.1	14310.4	***

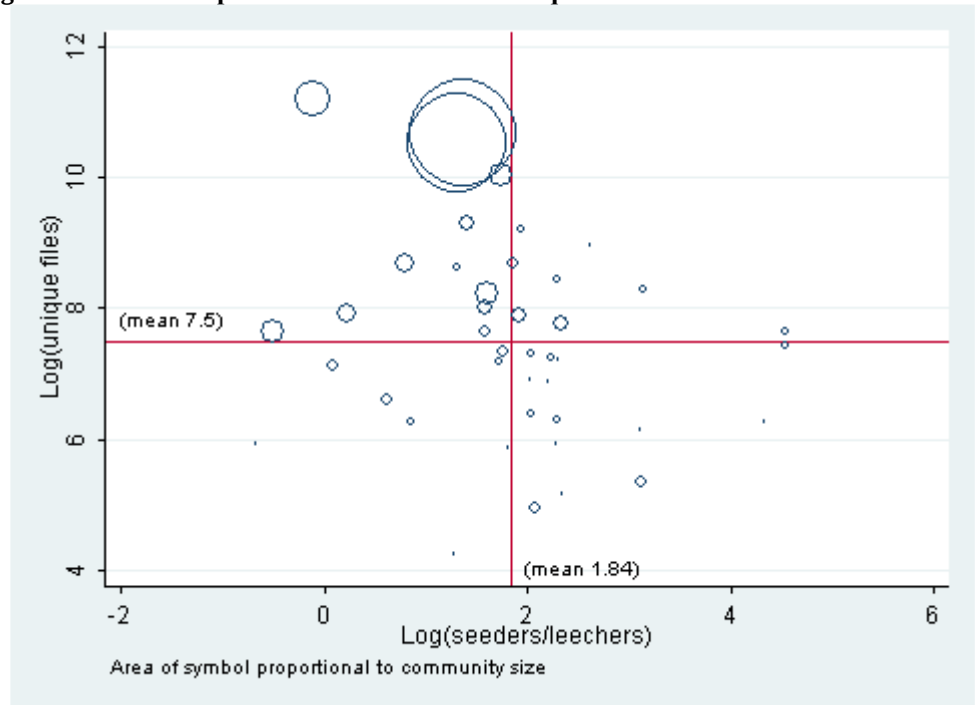
Note: ***, **, * mean significant at the level of 1%, 5%, and 10% respectively

Some evidence on the relationship between the size of a community and the amount of voluntary contributions

Olson's conjecture presumes a negative causal relationship between group size and the provision of collective goods. In the case of P2P file-sharing communities, we have two dimensions: a quantitative dimension (the number of unique files) and a qualitative dimension (the ratio of seeders to leechers).

Figure 2 illustrates the correlation between these two dimensions. This figure was built with the mean number of unique files and the mean ratio of seeders to leechers within each community over the 125 periods. Moreover, each circle is proportional to the mean size of each community. We observe that the largest circles are concentrated in the upper left quarter. The biggest communities tend to provide a larger library of files , but exhibit a smaller ratio of seeders to leechers.

Figure 2: Relationship between the number of unique files and the ratio seeders to leechers



Some evidence on the relationship between the design of P2P communities and the amount of voluntary contributions

Our sample is composed of non-public BitTorrent communities. These communities require their users to be registered before having access to the catalog of content. Tracker administrators can also set other rules to constrain or control members' behavior. Our 42 communities present some differences in terms of organizational rules. These distinct features are taken into consideration through several variables.

First, we distinguish private and semi-private communities. A community is **"private"** when new users must be invited by a member of this community. This filtering device should encourage cooperative behavior and reciprocity among co-opted members. Consequently, the amount of voluntary contributions should be higher in private communities than in semi-private communities. We can also presume that private communities gather individuals that have more similar tastes and preferences, which is an enhancing-cooperation factor (homophily effect).

We also have a dummy variable called **"control**" when the administrators of the tracker enforce a "sharing ratio" rule. It means that members that do not achieve a given ratio of uploading to downloading, cannot download any file or can be excluded temporarily or definitely from the community. The enforced sharing ratio varies across communities, but is usually around 1 (the members must share at least as much as they download). This coercive rule should prevent individual voluntary contributions from shrinking whatever the size of the community, enhancing the stability of large communities. But by providing external incentives this rule could crowd-out intrinsic motivations to contribute (Benabou and Tirole, 2003) and undermine the quality of content shared by the peers. If the impact is clearly positive on the ratio of seeders to leechers, this rule has more ambiguous effects on the number of files. It depends whether the dominant strategy to increase one's sharing ratio is to upload new files or to replicate existing files. The second strategy doesn't expand the amount of collective goods, but only improve the access to the pool of common resources.

We also control for the nature of content exchanged. The community is "**specialized**" (versus generalist) when file sharing is restricted to a specific category of content (for example, video games, music or adult video). A specialized community should generate more reciprocal attention and more cooperation than a generalist community.

Finally, we measure the visibility of our sample of communities in the BitTorrent universe by searching each tracker's name on mininova.org[7]. If the search engine replies by listing several files that belong to this tracker, we consider that this tracker is "**advertised**". For the administrators of a community, the interest of promoting their tracker on public search engine like mininova is explained by Curly Fries the founder of *TorrentFries*[8]: "*Dump sites are great promotional methods. Sites such as MiniNova and Demonoid allow you to upload torrents tracked elsewhere, so configure your new tracker to accept unregistered IP addresses*

[7] Mininova had been the largest torrent search engine with more than 3 billions of visitors per day. Under the pressure of legal authorities, it was closed at the end of 2009.
[8] Torrent Fries is one of a rare site dedicated to the running of a tracker.

(temporarily if you intend to go private) and upload your torrents to a bunch of dump sites like that. In the torrents' descriptions, include a comment such as "find more great torrents like this at www.example.com". You can even throw a text file inside the torrent to the same effect. You'd be amazed by how well it works". By enhancing the visibility of their community in BitTorrent meta-search engines, the administrators can attract new members that will help disseminate and replicate the catalog of content within the community. This advertising strategy aims to stimulate network effects and hence to increase the individual benefits of contributing. But that can also lead to more heterogeneity in members' preferences and tastes within community and the new members attracted by public search engines can behave more opportunistically (i.e. being more leechers than seeders). Advertising can be an interesting strategy to launch a community, but it is more risky for a mature community if it decreases the ratio seeders to leechers.

We perform Mann-Whitney tests to identify some links between the activity of a community (measured by the number of seeders, leechers, registered members, and unique files) and its governance form (private, control, advertised, specialized). Table 3 suggests that communities with stricter rules for admission (private) and for downloading (control) have less registered members and a smaller catalog, but exhibit a higher degree of voluntary contribution (i.e. higher ratio seeders to leechers). As expected, communities who advertise their content on public search engine have a lower ratio of seeders to leechers. They are also smaller in terms of catalog and members than non advertized ones suggesting that these communities use advertisement in their early stage of development. Finally, specialized communities tend to have more voluntary contributions (more unique files and a higher ratio of seeders to leechers) than generalized communities. To summarize the individual incentives to contribute seem to be higher in private and specialized communities that regulate downloading behavior.

Table 4: Mann-Whitney test for the features of P2P communities

	# communities	Seeders	Leechers	Ratios seeders to leechers[9]	Unique files available	Registered members
Private	4	8973	1029	31.9	2780	9581
Semi private	38	30552	14154	23.94	6650	110883
Mann-Whitney test		****	ns	***	***	***
		(-4.64)	(1.59)	(-3.88)	(-4.6)	(7.57)
control	8	21225	11117	15.74	5147	67356
No control	34	59234	20651	7.27	11088	560994
Mann-Whitney test		***	***	***	***	***
		(16.2)	(19.5)	(-12.7)	(9.9)	(7.3)
Specialized	24	35775	10695	18.22	7108	159401
Generalist	18	18646	16118	8.37	5178	21698
Mann-Whitney test		***	***	**	***	***
		(-14.7)	(-4.6)	(-2.5)	(-14.3)	(-19.9)
Advertised	19	8112	1803	9.31	3237	17399
Non advertised	23	46435	22684	18.26	8966	175120
Mann-Whitney test		***	***	***	***	***
		(25.2)	(21.2)	(-12.8)	(15.8)	(21.5)

Note: ***, **, * mean significant at the level of 1%, 5%, and 10% respectively

4. Econometric models and results

In this section, we present the specifications and the results of the econometric models used to analyze the impact of community size on the incentives to contribute and the amount of collective goods.

Econometric models

We estimate two complementary models using both the ratio of seeders to leechers (model M1) and the number of unique files (model M2) as dependent variable. For each model, we adopt a log-log specification using the log of registered users[10]. The two models enable us to examine how the size and the design of a P2P community influence its

[9] The statistical "mean for the ratio seeders to leechers" is obtained by calculating the mean ratio of seeders to leechers in each community and then deriving the average of mean ratios in each category of communities. This is a better measure than the ratio of mean seeders over mean leechers in each category of communities .

[10] We have also estimated the two models using linear and quadratic specifications (for registered members). Our results remain robust to these alternative specifications. But, the log specification gives the better goodness of fit.

performance, measured by the amount of collective goods provided (M2) and the quality of access to these collective goods (M1). The estimated models can be formulated as follows:

$$\log(ratio_{it}) = \beta_0 + \beta_1 \log(registered_{it}) + \beta_2 (control_i) + \beta_3 (private_i) + \beta_4 (advertised_i) + \beta_5 (specialized_i) + \varepsilon_{it} \quad \text{(M1)}$$

$$\log(unique\ files_{it}) = \beta_0 + \beta_1 \log(registered_{it}) + \beta_2 (control_i) + \beta_3 (private_i) + \beta_4 (advertised_i) + \beta_5 (specialized_i) + \varepsilon_{it} \quad \text{(M2)}$$

Except for the number of registered members, the explanatory variables that control for the features of each community are time invariant dummies. As we have to deal with time-series cross-sectional (TSCS) data with a number of periods superior to the number of communities, potential problems in the error structure have to be addressed. First, the Breush-Pagan/Cook-Weisberg test for constant variance fell within the confidence interval of 10 percent for (M1) and (M2). Secondly, the strong heterogeneity in the size of our 42 communities is likely to cause a problem of groupwise heteroscedasticity. The modified Wald test for groupwise heteroscedasticity confirms that the variance of error process differs across units for (M1) and (M2). Because our data exhibits a large temporal dimension and that observations at 10 am are correlated with observations at 10 pm, we suspect the presence of residuals serial correlation. This is confirmed by a test for autocorrelation in panel-data (Woodridge, 2002)[11], but only for equation (M2). For all these reasons, the feasible general least square (FGLS) is the most appropriate estimator in presence of panel-level heteroscedasticity and autocorrelation. The FGLS is similar to generalized least squares except that it uses an estimated variance-covariance matrix since the true matrix is not known directly. The covariance matrix is estimated by iteration, using the OLS estimators in the first step[12].

We also conduct fixed effects regressions to control for unobserved time-invariant characteristics of the 42 communities. Fixed effects models allow us to focus on the impact of within-community size variations on the provision of collective goods. However, the drawback of using fixed effect estimator is that time-invariant variables cannot be estimated. Plumper and Troeger (2007) propose to use the fixed effect vector decomposition (FEVD)

[11] Drukker (2003) provides a simple program to perform this test in Stata.
[12] Using the Panel Corrected Standard Errors (PCSE) estimators proposed by Beck and Katz (1995) would have been a possibility. However this is less efficient for panel data when temporal dimension exceeds individual dimension (Chen et al. 2006).

method to deal with time invariant (or slow moving) variables and fixed effect estimation. The three stage procedure proceeds as follows: the first stage performs a unit fixed effect estimation, the second stage regresses the unit effect on the time invariant variables (which allows to distinguish between the explained and unexplained part of the unit effect) and the third stage performs a pooled-OLS regression on time-variant, time-invariant variables and the unexplained part of the unit effect. According to Plumper and Troeger (2007), the most important condition to ensure the reliability of the FEVD estimator is that between-variation have to be larger than within-variation which is a strong property of our dataset.

Table 5 and Table 6 display the estimates of the two models (M1) and (M2). The columns (1a) and (1b) report respectively the OLS (ordinary Least Square) estimates (the robust standard errors in brackets are calculated using the Hubber and White sandwich estimator) and the FGLS estimates, (controlling for heteroscedasticity and serial autocorrelation with a first order auto regressive coefficient) using community size as the only explanatory variable. Columns (2a) and (2b) display the OLS and FGLS estimates when we control for the governance rules within community. Finally, the fixed effect (FE) and the fixed effect vector decomposition (FEVD) estimates are reported in column (3a) and (3b). .

Table 5: Estimations of the impact of community size on the incentives to contribute

	Dep. Var= log(ratio seeders to leechers)					
	OLS (1a)	FGLS (1b)	OLS (2a)	FGLS (2b)	FE (3a)	FEVD (3b)
Log(registered	-0.26	-0.18	-0.29	-0.14	-0.05	-0.05
	(27.10)***	(18.07)***	(27.62)***	(9.76)***	(2.27)**	(19.77)***
Control			0.36297	0.05656		0.49072
			(6.16)***	(0.76)		(36.27)***
Private			0.59	0.46		0.67
			(13.43)***	(7.39)***		(67.78)***
Advertised			0.06	0.36		0.25
			(1.98)**	(7.73)***		(30.29)***
Specialized			0.70	0.48		0.56
			(22.67)***	(9.84)***		(66.17)***
η						1.00
						(259.27)***
Constant	4.37101	3.44990	3.67	2.14	2.35803	1.33
	(44.10)***	(34.40)***	(28.60)***	(12.26)***	(10.37)***	(45.15)***
Observations	5097	5097	5097	5097	5097	5097
Communities		42		42	42	
R2	0.13		0.24			

Table 6: Estimations of the impact of community size on the amount of unique files shared within the community

	Dep. Var= log(unique files)					
	OLS (1a)	FGLS (1b)	OLS (2a)	FGLS (2b)	FE (3a)	FEVD (3b)
Log(registered)	0.66	0.07	0.67	0.05	0.12	0.12
	(89.37)***	(12.55)***	(73.51)***	(8.66)***	(8.62)***	(70.16)***
Control			0.72489	0.62205		0.42289
			(22.97)***	(11.62)***		(61.22)***
Private			-0.37	-0.56		-0.56
			(7.13)***	(11.46)***		(110.13)***
Advertised			-0.13	-0.28		-0.58
			(3.44)***	(6.50)***		(133.87)***
Specialized			-0.25	0.27		0.09
			(6.97)***	(6.51)***		(20.77)***
ρ						1.00
						(549.18)***
Constant	1.13613	6.90405	1.47	7.59	6.38834	7.02
	(15.19)***	(118.19)***	(13.22)***	(93.15)***	(49.50)***	(401.72)***
Observations	5097	5097	5097	5097	5097	5097
Communities		42		42		42
R2	0.49		0.51		0.6	

The impact of community size

We find a negative impact of community size on the ratio of seeders to leechers. The results suggest that the individual incentives to contribute voluntarily decrease with the number of members within a community, but at a decreasing rate (the coefficient of log(registered) is negative but between -1 and 0). When the size of a community increases by 100%, the ratio of seeders to leechers decreases by 14% on average (FGLS model in Table 5).This finding supports the idea that the incentives to contribute would never shrink to zero even in large communities. People tend to be less cooperative in larger community, but there is always a core of contributors who preserve the stability and quality of the file-sharing community (Krishnan et al., 2004; DangNguyen and Penard, 2007). If we control for unobserved community fixed effects, the negative impact of group size is lower, but still significant: a 100% increase in community size will reduce the ratio seeders to leechers by only 5%. This result suggests that the incentives to contribute are quite robust to within-community variation in registered members.

The results of table 6 show that community size has a positive impact on the amount of collective good provided by the file-sharing community. The important difference between

the estimated effect with the FGLS method and the OLS approach legitimate our choices to control for serial correlation. The estimates suggest that the quantity of unique files tend to increase with the number of members, but at a decreasing rate (the coefficient of log(registered) is positive but significantly below one). When the size of a community increases by 10%, the size of the catalog increases by 0.5% (by 1.2% if we control for community fixed effects). This weak effect is probably explained by the fact that original contents are hardly provided by recent members. The catalog of content is mostly expanded by core members who are generally more experienced. If a community integrates new members, the effect on the amount of unique files will be observed with a lag because it takes time for a new member to move from the periphery to the core of the community. This mechanism is observed in other Internet-mediated communities who provide collective goods like open source software communities (Masmoudi et al, 2009).

To summarize, even if the individual incentives to contribute voluntary tend to decrease as community size rises, the aggregate collective contributions slightly increase. In other words, even if each member is sharing less on average, the size of the catalog increases with the number of members within community. This seems to indicate that the provision of a file-sharing community is an inclusive or non rival good according to Chamberlin (1974).

Robustness checks

We are concerned with several potential problems related to the dataset and the specification of our models. We can suspect causality problem as the relationship between group size and the ratio or the number of unique files can be reversed. Indeed, users can decide to register in a community only if the number of files and the level of cooperation are sufficiently high. Columns (1a) and (1b) in tables 7 and 8 estimate the model using FGLS and FEVD estimators with the lagged community size (the lag corresponds to two periods, meaning one day) to tackle with causality problems. The estimates with lagged members show similar impact of community size on the individual incentives to contribute and the aggregate amount of contribution.

As suggested by table 2 and figure 2, our 42 communities are very heterogeneous. Taking a closer look at the dataset reveals that two communities specialized in adult content have more than 1.5 millions of users, while the third largest community only has 100,000 subscribers. The weight of these two large communities can produce heteroscedasticity in the variance of the error term. Columns (2a) and (2b) in table 7 and 8 estimate the two models

(M1) and (M2) without these two largest communities. Once again we obtain similar results for the effects of community size.

Table 7 Robustness tests for the estimations of the impact of community size on the incentives to contribute

	Dep. Var= log(ratio seeders to leecher)s			
	with lagged variables		without the two largest communities	
	FGLS (1a)	FEVD (1b)	FGLS (2a)	FEVD (2b)
Log(registered)			-0.38	-0.06
			(18.79)***	(15.09)***
Log(registered) T-2	-0.12	-0.04		
	(8.54)***	(12.93)***		
Control	0.06667	0.50740	-0.08497	0.47168
	(0.89)	(37.24)***	(1.22)	(34.08)***
Private	0.46	0.68	0.78	0.73
	(7.38)***	(67.95)***	(12.30)***	(68.67)***
Advertised	0.37	0.27	0.35	0.24
	(7.89)***	(31.70)***	(8.55)***	(28.32)***
Specialized	0.48	0.55	0.49	0.57
	(9.52)***	(64.51)***	(11.07)***	(66.43)***
η_i		1.00		1.00
		(257.70)***		(244.71)***
Constant	1.94	1.15	4.14	1.30
	(11.20)***	(38.54)***	(19.42)***	(35.42)***
Observations	5011	5011	4847	4847
Communities		42		40
R2				

Table 8: Robustness tests for the estimations of the impact of community size on the amount of unique files shared within the community

	Dep. Var= log(unique files)			
	with lagged variables		Without the two largest communities	
	FGLS (1a)	FEVD (1b)	FGLS (2a)	FEVD (2b)
Log(registered)		0.11		0.11
		(66.78)***		(57.17)***
Log(registered) T-2	0.04		0.02	
	(8.02)***		(4.99)***	
Control	0.61458	0.41700	0.73602	0.51582
	(11.72)***	(60.59)***	(11.97)***	(73.18)***
Private	-0.55	-0.56	-0.24	-0.40
	(11.28)***	(111.53)***	(5.91)***	(74.10)***
Advertised	-0.30	-0.59	-0.20	-0.41
	(6.87)***	(135.51)***	(4.36)***	(93.67)***
Specialized	0.28	0.09	0.11	-0.04
	(6.75)***	(21.32)***	(2.42)**	(7.99)***
ρ		1.00		1.00
		(550.58)***		(537.66)***
Constant	7.61	7.08	7.47	6.77
	(92.96)***	(405.31)***	(109.39)***	(339.67)***
Observations	5011	5011	4847	4847
Communities		42		40
R2				

The role played by the rules designed by communities' administrators

Olson (1965) stated that large groups could overcome free-riding by providing private incentives or exclusive services to the active members. Some of the dummies used to control for the features of our P2P communities can be analyzed as private incentives (here **private** and **control**).

Tables 5 and 6 show that private communities provide a higher ratio of seeders to leechers but a lower quantity of collective goods. This can be explained by the fact that private communities are more selective and can handpick their members based on their ability to contribute to the collective good. Entry regulation enables not only to prevent opportunistic behavior, but also to better segment users' needs and interests.

Enforcing a minimum ratio of uploading to downloading is expected to rule out opportunistic behavior (free-riding). Table 5 confirms that monitoring behavior increases the incentives to contribute leading to a higher ratio of seeders to leechers. Our results also show that communities that enforce a minimum ratio of uploading to downloading also provide a

larger catalog. Strategic behavior can explain this finding. It is known that the easiest way to reach the required sharing ratio is to share popular files (that are frequently requested). But if too many people share the same files, they will experience difficulties to increase their "sharing ratio" (few leechers for many seeders). An alternative (and more effective) strategy for restoring your sharing ratio is to upload new files and thus become the only seeder of these files.

Specialized communities seem to encourage voluntary contribution. Probably, members of specialized communities are more strongly involved and incited to cooperate with each other (Asvanund *et al.*, 2006). Table 5 confirms the idea that the propensity to contribute is higher in a topic-oriented community. Moreover, the catalog of content tends to be larger in specialized communities.

Finally, a community that relies on public search engine to promote its catalog (an *advertised* community) has a higher proportion of contributors, but a more limited catalog of content.

These findings highlight the fact that a community must design efficient organizational rules using a mix of incentive and coercive tools, to prevent free riding behavior and provide a high quantity and quality of collective goods that match members' preferences. The decrease of searching cost as well as the enhancement of individual capabilities to share is not a sufficient condition to ensure a sustainable model of file sharing.

5. Conclusion

This paper has investigated the relationship between the size of file-sharing communities and their ability to provide collective goods (measured by the quantity and availability of content in the community). During two months between December 2007 and February 2008, we collected data on the activity of 42 private and semi-private bitTorrent communities. Our results suggest that the collective provision in these communities can be analyzed as a pure public good. The amount of collective good increases with the number of registered users whereas the individual propensity to contribute decreases with community enlargement. We also show that the rules designed by the administrators of these communities have a significant impact on their performance and their sustainable size. We find that stricter monitoring schemes have a positive impact on the incentives to contribute. However, the amount of unique files shared is lower in a private community. In other words, the provision

of a large catalog (or a long tail) of contents that match individual preferences cannot be disconnected from the design and management of these virtual communities. This challenging issue deserves further investigation. It would be interesting to compare the centralized model of online merchants and the decentralized model of P2P community to manage and promote the long tail. Which model is the more efficient and sustainable to connect the supply and demand of rare content? How do you articulate market and non market incentives, external and intrinsic motivations to provide and distribute niche and popular content?

A limitation of this study is the absence of individual data to analyze members' behavior within community. A future avenue of research is to collect individual-level data in several communities in order to examine the dynamics of individual contributions. This will enable us to compare behavior of new members and early members, and to analyze how they react to a change in community size. It would be also interesting to understand how members move from the periphery to the core of a community over time and to identify the different strategies of voluntary contributions within a community.

References:

Adar, E., Huberman, B., 2000. Free Riding on Gnutella. First Monday **5**.

Andreoni, J., 1988. "Privately Provided Public Goods in a Large Economy: The Limits of Altruism," Journal of Public Economics, 1988, 35 (1), 57–73.

Andreoni, J., 2007. "Giving Gifts to Groups: How Altruism Depends On the Number of Recipients," Journal of Public Economics, 91 (9), 1731–1749.

Asvanund, A., Clay, K., Krishnan, R., Smith, M.D., 2004. An Empirical Analysis of Network Externalities in Peer-to-Peer Music-Sharing Networks. Information Systems Research **15**, 155-174.

Asvanund A., Krishnan, R., Smith M.D., Telang, R., 2006. Interest-Based Self-Organizing Peer-to-Peer Networks: A Club Economic Approach. Working paper available at SSRN http://ssrn.com/abstract=585345.

Beck, N., Katz, J.N., 1995. What to Do and Not to Do With Time-Series Cross Section Data. The American Political Science 89, 634-647.

Benabou, R., Tirole, J., 2003. Intrinsic and Extrinsic Motivation. Review of Economic Studies 70(3), 489-520.

Bergstrom, T., Blume, L., Varian, H., 1986. On the Private Provision of Public Goods. Journal of Public Economics 29, 25–49.

Bharambe, A.R., Herley, C., Padmanabhan, V.N., 2005. Analyzing and Improving BitTorrent. Technical Report MSR-TR-2005-03, Microsoft Research.

Chamberlin, J., 1974. Provision of Collective Goods As a Function of Group Size. American Political Science Review 68, 707-716.

Chamberlin, J., 1978. The Logic of Collective Action: Some Experimental Results. Behavioural Science 26, 441-45.

Chen, X., Lin, S., Reed, W.R., 2006. A Monte Carlo Estimation of the Efficiency of the PCSE Estimator. Working paper, College of Business and Economics University of Canterbury.

Cohen, B., 2003. Incentives Build Robustness in BitTorrent. Proceedings of International Workshop on peer-to-peer Systems.

Dang-Nguyen, G., Pénard, T., 2007. Network Cooperation and Incentives Inside Online Communities. In : Brousseau, E., Curien, N., (Eds.), Internet and Digital Economy, Cambridge University Press.

Drukker, D.M., 2003. Testing For Serial-Correlation in Linear Panel-data Models. The Stata Journal 3, 168-177.

Esteban, J., Ray, D., 2001. Collective Action and the Group Size Paradox. The American Political Science Review 95, 663-672.

Gaube, T., 2001. Group size and free riding when private and public goods are gross substitutes. Economics Letters 70, 127 –132.

Huang, K., Wang, L., Zhang, D., Liu, Y., 2007. A Dynamic Quota-Based Peers Selection in Bit Torrent. Proceedings of the Sixth International Conference on Grid and Cooperative Computing.

Isaac, R.M., Walker, J.M., 1988. Group Size Effect in Public Good Provision: The Voluntary Contributions Mechanism. The Quaterly Journal of Economics 103, 179-99.

Isaac, R.M., Walker, J.M., Williams, A.W., 1994. Group Size and the Voluntary Provision of Public Goods Experimental Evidence Utilizing Large Group. Journal of Public Economics 54, 1-36.

Krishnan, R., Smith, M.D., Tang, Z., Telang, R., 2004. The Virtual Commons: Why Free Riding Can Be Tolerate in File Sharing Network. Working paper, available at SSRN http://papers.ssrn.com/sol3/papers.cfm?abstract_id=450241.

Krishnan, R., Smith, M.D., Tang, Z., Telang, R., 2007. Digital Business Models for Peer-to-Peer Networks: Analysis and Economic Issue. Review of Network Economics, 6, 194-213.

Krishnan, R, Smith, M.D., Telang, R., 2003. The Economics of Peer-to-Peer Networks. Journal of Information Technology Theory and Applications, 5(3), 31-44.

Legout A., Urvoy-Keller G., Michiardi, P., 2005. Rare First and Chock Algorithms are Enough. Proceding of the 6th ACM SIGCOMM conference on Internet measurement.

McGuire, M.C., 1974. Group size, group homogeneity, and the aggregate provision of a pure public good under Cournot behavior. Public Choice 18, 107–126.

Marwell, G., Ames, R.E, 1979. Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size and the Free-Rider Problem. The American Journal of Sociology 84, 1335-1360.

Masmoudi H., Den Besten M., De loupy C., Dalle J.M., 2009. “Peeling the onion : the words and actions that distinguish core from periphery in bug reports and how core and periphery interact together”, Fifth International Conference on Open Source Systems, Sweden.

Olson, M., 1965. The Logic of Collective Action. Harvard University Press.

Palfrey, T. R., Rosenthal, H., 1984.“Participation and the Provision of Discrete Public Goods: A Strategic Analysis,” Journal of Public Economics, 24 (2), 171–193.

Pecorino, P., 1999. The Effect of Group Size on Public Good Provision in a Repeated Game Setting. Journal of Public Economics 72, 121-134.

Pecorino, P., Temimi, A., 2008. The Group Size Paradox Revisited. Journal of Public Economic Theory 10, 785-799.

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. Cambridge, MIT press, Cambridge

Yue Y., Lin C., Tan, Z., 2006. Analysing the Performance and fairness of Bit-Torrent Like Protocol Using a general Fluid Model. Computer Communication, Vol. 29, 3946-3956.

Zhang, M., Zhu, F., 2010. Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. American Economic Review, Forthcoming.

Annex 1

Site of the tracker	N° tracker	content	private	specialized	control	advertised
http://www.captain-tracker.fr/index.php	1	general	0	0	1	0
http://www.sharing-torrents.com/	2	general	0	0	1	1
http://leparrain.mine.nu/torrents.php	3	general	0	0	0	0
http://www.unlimited-tracker.net/	4	general	0	0	1	1
http://www.nhltorrents.co.uk/	5	sport	0	1	1	1
http://xtremewrestlingtorrents.net/static.php	6	sport	0	1	1	0
http://www.dimeadozen.org/index.php	7	music	0	1	0	1
http://www.indietorrents.com/index.php	8	music	1	1	1	0
http://shnflac.net/index.php	9	music	0	1	0	0
http://jamtothis.com/	10	music	0	1	1	0
http://www.browntracker.net/browse.php	11	music	0	1	0	0
http://anvilofsound.com/	12	music	0	1	0	1
http://mixes.dfx.at/index.php	13	music	0	1	1	0
http://asiandvdclub.org/	14	cinéma	0	1	1	0
http://alt.bitworld.to/browse.php	15	general	0	0	1	1
http://www.araditracker.com/	16	general	0	0	1	0
http://www.titaniumtorrents.net/	17	general	1	0	1	1
http://dididave.com/	18	general	0	0	1	0
http://www.quebectorrent.com/	19	general	0	0	1	1
http://cinemageddon.org/	20	cinéma	0	1	1	1
http://www.blades-heaven.com/index.php	21	general	0	0	1	1
http://www.puretna.com/	22	adult	0	1	0	0
http://www.kingdomxxx.com/	23	adult	0	1	0	0
http://www.empornium.us/	24	adult	0	1	1	0
http://www.pornevo.com/	25	adult	0	1	1	0
http://www.underground-gamer.com/	27	vidéo game	0	1	1	1
http://www.pleasuredome.org.uk/	28	vidéo game	0	1	1	1
http://my-gamebox.com/	29	vidéo game	0	1	1	1
http://thepeerhub.com/	30	general	0	0	1	1
http://bitnation.com/index.php	31	general	1	0	1	1
http://p2pworld.ulmb.com/	32	general	0	0	1	1
http://torrent-hackers.co.uk/	33	general	0	0	1	1
http://www.sport-scene.net/	35	sport	0	1	1	0
http://www.sportbit.org/	36	sport	0	1	1	0

http://www.prosporttorrents.net	37	sport	0	1	1	1
http://www.mamietracker.com/index.php	38	general	0	0	0	1
http://zombtracker.the-zomb.com/	39	music	0	1	1	0
http://cinematik.net/	40	cinéma	1	1	1	0
http://www.zinebytes.org/	41	e-learning	0	1	1	0
http://www.mytracker.ru/index.php	42	general	0	0	1	0
http://linuxmafia.net/	43	general	0	0	1	1
http://zerotracker.com/index.php	44	general	0	0	1	0