DRAFT
COMMENTS WELCOME

An Empirical Analysis of Network Externalities in P2P Music-Sharing Networks

Atip Asvanund, Karen Clay, Ramayya Krishnan, Michael D. Smith

{atip, kclay, rk2x, mds}@andrew.cmu.edu

H. John Heinz III School of Public Policy and Management
Carnegie Mellon University, Pittsburgh, PA 15213

This Draft: July 2002

http://www.heinz.cmu.edu/~mds/p2pe.pdf

Acknowledgements: We thank participants at the Carnegie Mellon University Graduate School for Industrial Administration, the University of California at Berkeley SIMS and Haas Schools, and the 2001 Telecommunications Policy Research Conference and schools for valuable comments on this research. Financial support was provided by the National Science Foundation through grant IIS-0118767.

An Empirical Analysis of Network Externalities in P2P Music-Sharing Networks

Abstract

Peer-to-peer (P2P) networks are becoming an important medium for the distribution of consumer information goods. However, there is little academic research into the behavior of these networks. We analyze the impact of positive and negative network externalities on the optimal size of P2P networks. Using data collected from the six most popular OpenNap P2P music-sharing networks between December 19, 2000 and April 22, 2001 we find that additional users contribute value in terms of additional network content at a diminishing rate, while they impose costs in terms of congestion on shared resources at an increasing rate.

Using an analytic model, we explore technical solutions to the congestion problem, for example by increasing network capacity. This model suggests that although increasing capacity will allow more users to participate on the network, there may be little incentive for network operators to do so. This is because diminishing positive network externalities imply decreasing content benefits to adding more users. Together these results suggest that the optimal size of a P2P network may be bounded in many common implementations. We conclude by discussing various options to improve network performance including network membership rules and usage-based pricing.

1. Introduction

Peer-to-peer (P2P) networks are an important social phenomenon and an emerging medium for the distribution of consumer information goods. They may also become important tools for the management of data and services within the enterprise in the near future. The most well known application for the P2P technology is music sharing. After its launch in May 1999, Napster enabled millions of individual users to share MP3 music tracks. Napster and subsequent P2P file sharing networks demonstrate the potential for P2P networks to facilitate collective usage of resources shared by autonomous peers.

In spite of this potential, there is little academic research into the behavior of these networks in terms of the value users bring to the network, the costs they impose on the network, or the impact of these factors on network behavior. Systematic research to address these questions is important for a variety of constituencies including engineers designing protocols to support P2P networks, entrepreneurs developing P2P-based businesses, and intellectual property holders seeking to develop their own networks and to minimize the use of non-complying networks.

In this paper, we study network externalities that arise in P2P music-sharing networks. A network externality is the marginal effect that an additional user of a network has on existing users. In contrast to network externalities that arise in telecommunication networks, which are driven by user membership (e.g., Metcalfe's law), network externalities in P2P networks are driven by network content. Because of this, P2P networks have many of the properties of club goods and public goods. In the economics literature, public goods are goods that are non-excludable in supply and non-rivalrous in demand (Samuelson 1954). Non-excludability means the good must be supplied to everyone or not at all and non-rivalrous means that the

consumption of the good by one individual does not reduce the utility of consumption by other individuals. Examples of public goods include radio broadcasts or sunlight. Club goods, in contrast are goods that are excludable in supply and have some degree of rivalry in demand (i.e., between completely rivalrous and non-rivalrous). A discrete unit of apple is completely rivalrous because when you consume it, its utility is completely depleted. On the other hand, a radio broadcast is non-rivalrous. The quality does not degrade when you increase listeners. A swimming pool, however, is somewhere in between. One person may not make the pool unusable by others, but too many people can make it crowded and reduce utility. The focus of club goods is in determining the right member size for a finite collective who have the exclusive right to use a club good (e.g., swimming pool) (Buchanan 1965).

The services provided over P2P networks have some of the characteristics of public goods. Membership is typically not controlled and once a user gains access to the network they can access all the services provided by the network (i.e., non-excludability). Further, as we discuss in more detail below, in the absence of free-riding song replication should scale with network size, creating a situation where the consumption of network resources is non-rivalrous. However, these characteristics may also be relaxed in some environments to create an environment more similar to club goods. For example, P2P networks can impose some degree of excludability through membership rules or limits on access to network services (e.g., through subscription services and digital rights management systems as employed by Napster in early 2002). Likewise, increasing free-riding can create rivalry among users for scarce network content.[1]

[1] Note that the economic characteristics of the P2P network services we refer to are independent from the legal and ethical characteristics of the information provided over these services.

Two results from the public and club economics literature are significant in the context of P2P networks. First, Hardin (1968) argues that the self-interested consumption of public goods may deplete the overall public utility, a.k.a. the "tragedy of the commons." Second, the propensity to "free-ride" (enjoying the public good provided by others while not supplying the good yourself) worsens as group size increases as argued by Olson (1968) and shown analytically by Palfrey and Rosenthal (1984) and Hindriks and Pancs (2001) among others. In applying these results we draw on both the economics literature and the information systems (IS) literature with regard to the technical characteristics of P2P networks.

To empirically analyze these questions in the context of P2P networks, we develop a reduced form utility model of user behavior and use this model to empirically analyze the characteristics of free-riding and network utility as a function of network size. Our empirical analysis uses a data set gathered from the six most popular OpenNap (a.k.a. Open Source Napster) networks in December 2001. Our data were collected from December 19, 2000 to April 22, 2001 and include information on network congestion, and song availability and replication (number of copies of the song available for sharing on the network) for 170 randomly selected songs in 17 musical genres. Using ordinary least squares, logit, Poisson, and zero-inflated Poisson regression models, we find that additional users contribute value in terms of additional network content at a diminishing rate, while they impose costs in terms of congestion on shared resources at an increasing rate. This points to a potential inefficiency with larger P2P networks.

To explore this further, we apply an analytical model used in telecommunications analysis to assess the benefit of increasing network capacity in reducing congestion. This model shows that although increasing capacity may allow more users to participate on a network, there may be little incentive for network operators to provision this capacity. This is because diminishing

positive network externalities imply decreasing benefits to adding more capacity (and therefore users).

The remainder of this paper proceeds as follows. Section 2 provides background on P2P networks and reviews the relevant IS and economics literature. Section 3 presents a model of positive and negative externalities in P2P networks and discusses hypotheses for how these externalities should vary with network size. Section 4 discusses the empirical data we collect to address these hypotheses and section 5 presents our empirical results. Section 6 concludes and identifies areas for future research.

2. Background

At its core, P2P networking enables resource sharing directly between network users. These resources are most commonly in the form of information, such as files or digital content, but can also include storage capacity or computing power. Thus, P2P networking is different from a traditional client-server environment where all network resources are contained in and managed by a central location.

Internet Relay Chat (IRC), which was developed in the late 1980s, was one of the first P2P network services. IRC allowed for the transmission of text messages, and later digital content, directly between groups of network users. Subsequent P2P file sharing networks, such as Napster, OpenNap, Gnutella, Kazaa, and Morpheus, achieved much higher levels of adoption (and publicity) by enhancing the ability to locate and download content from the network. More recently, P2P networks are gaining popularity for applications such as distributed computing (e.g., SETI@Home), collaboration (e.g., Groove Networks), and enterprise information sharing (e.g., Bad Blue).

In addition to the variety of application environments, P2P networks have also been deployed with a variety of network architectures. These architectures can be summarized along two axes: the degree of decentralization of the network content and the degree of decentralization of the catalog of this content. The degree of decentralization of network content pertains to whether the content is stored in a central location (increasing direct management of the content), or is stored in a distributed manner separately by the individual peers (speeding download time, caching content within the network infrastructure, offloading bandwidth burden to the edge of the network and eliminating a single point of failure for content distribution). The degree of decentralization of the catalog of content pertains to whether this catalog is stored in a central location (increasing the accuracy and reliability of the catalog), or is stored in a distributed manner separately by the individual peers (eliminating a single point of failure for directory services).[2]

P2P networks have adopted a variety of designs with regard to content and catalog decentralization. For example, the Napster protocol, which was implemented in the Napster and OpenNap networks, has fully decentralized content, but a centralized file catalog. The Gnutella protocol decentralizes both the content and file catalog. Gnutella users connect to a few (3-5) other peers, who maintain recursive connections to subsequent peers down the path. A user broadcasts requests for content along these paths until they have reached a predetermined depth (typically 7 hops). The more recent Morpheus and Kazaa networks use Supernodes —

[2] There are various realizations of distributed catalogs. These range from "virtual catalogs" that are generated by querying individual nodes in a network in response to a request for content as is done in Gnutella networks to maintaining catalogs about the content available in a small neighborhood of nodes as is done in Kazaa and Morpheus. The ability to evade law enforcement through the use of distributed catalogs is cited as an "advantage" by opponents of the intellectual property regime prevailing in the music industry.

distinguished peers that host catalogs of their neighboring peers — making them quasi-centralized in catalog and decentralized in content.

In this paper, we focus our analysis on the Napster protocol as implemented in the OpenNap networks. After its launch in May of 1999, Napster became the first mainstream P2P music-sharing network. Shortly thereafter, an open source group reversed engineered its protocol and built the non-commercial OpenNap networks. The OpenNap networks consist of multiple disjointed centralized networks, while Napster consists of a single monolithic network.

As defined by the Napster protocol, users perform the following steps to use the OpenNap network. First, a user must login to a central OpenNap server. Once logged in, the user computer stays connected to the central server for the duration of their presence on the network. The user computer that is connected in this manner is referred to as a peer. Servers have limited capacity for maintaining such simultaneous connections with peers. All login requests from peers when the server is above capacity are rejected. After a successful login, a list of the files the user is sharing and associated information is uploaded to a central catalog at the server. Any change in the content available on the user computer is immediately uploaded resulting in a catalog that is always current. The catalog contains file information and peer location (i.e., IP address) for all content on the network. To locate a file the user places a keyword query against this catalog database and the database returns a list of any matching results. This list includes the name, length, encoding speed, and provider for each file. The client program issues a ping request to each provider and sorts the list in ascending order by ping time (i.e., a measure of the congestion at the peer). At this point the user chooses an entry in the list to initiate a download from the provider. This request may be accepted or queued by the provider. Providers typically accept a limited number of simultaneous downloads and queue any additional download requests. Once

the download request is accepted, the requesting peer computer downloads the content from the provider. Unless otherwise specified, the requesting peer now becomes a provider of the content. This is referred to as autoreplication and is manner by which content replicas are created on the network. In this manner content replicates on the network.

Few papers have studied the behavior of P2P networks of this sort. Our work builds on the few papers that do and on a larger body of economics research that addresses public and club goods.[3] With respect to P2P research, Adar and Huberman (2000) found that 1% of Gnutella users provided 50% of query results and 70% of users provided no results (i.e., were free-riding). These results highlight the sub-optimal sharing that can obtain in the absence of external incentives on user behavior. In addressing these problems, Golle, Leyton-Brown and Mironov (2001) use a game theoretic model to study the benefit of introducing micro-payment systems to P2P networks. Our research extends this work by analyzing network externalities in P2P networks and the drivers of free-riding behavior in the context of optimal network size.

3. Model

Initially, users chose among competing networks, which are characterized by positive and negative network externalities. A network externality is the marginal effect that an additional user has on existing users. Positive externalities arise because the number of users is positively related to the range of content available and the number of copies of each track, which *ceteris paribus* will increase variety and reduce the expected download time. Negative externalities arise because more users increase the expected login, query, and download times.

[3] For example, the large body of experimental research that addresses free-riding in a typical public goods setting. See Davis and Holt (1993) or Ledyard (1995) for reviews of this literature.

Formally, let each user on the network have utility given by the sum of the utility from the availability and replication of a vector of content F and the (dis)utility of a vector of congestion effects C:

$$U(F(N), C(N)) = U_F(F(N)) + U_C(C(N)) \quad (1)$$

Consistent with these definitions, let users be (weakly) better off when more variety or more replicas of content are provided by the network and (weakly) worse off when network congestion increases:

$$\partial U/\partial f \geq 0 \quad (2)$$

$$\partial U/\partial c \leq 0 \quad (3)$$

where f is an element of the vector F and c is an element of the vector C, and let content and congestion (weakly) increase in N:

$$\partial f/\partial N \geq 0 \quad (4)$$

$$\partial c/\partial N \geq 0 \quad (5)$$

Finally, assume that U is concave in both f and c:

$$\partial^2 U/\partial f^2 \leq 0 \quad (6)$$

$$\partial^2 U/\partial c^2 \leq 0 \quad (7)$$

such that users have a diminishing marginal utility from more content and more replicas of content, and the marginal impact of congestion on utility is either constant or declining (i.e., an additional minute of wait time has a larger impact on utility when wait time is 1 minute than when it is 100 minutes).

Using this model, we wish to characterize how positive network externalities from content and negative network externalities from congestion vary with the number of network users (i.e., $\partial U_F/\partial N$, $\partial^2 U_F/\partial N^2$, $\partial U_C/\partial N$, $\partial^2 U_C/\partial N^2$). From this, we also wish to analyze how aggregate network utility varies with the number of network users (i.e., $\partial U/\partial N$, $\partial^2 U/\partial N^2$) as a way to understand the impact of network externalities on optimal network size. Using (1) the former values are given by

$$\partial U_F/\partial N=\partial U/\partial f\bullet\partial f/\partial N \tag{8}$$

$$\partial^2 U_F/\partial N^2=\partial^2 U/\partial f^2\bullet(\partial f/\partial N)^2+\partial U/\partial f\bullet\partial^2 f/\partial N^2 \tag{9}$$

$$\partial U_C/\partial N=\partial U/\partial c\bullet\partial c/\partial N \tag{10}$$

$$\partial^2 U_C/\partial N^2=\partial^2 U/\partial c^2\bullet(\partial c/\partial N)^2+\partial U/\partial c\bullet\partial^2 c/\partial N^2 \tag{11}$$

and the latter values are given by $\partial U/\partial N=\partial U_F/\partial N+\partial U_C/\partial N$ and $\partial^2 U_F/\partial N^2=\partial^2 U_F/\partial N^2+\partial^2 U_F/\partial N^2$.

By (2) and (4) $\partial U_F/\partial N\geq 0$ and similarly by (3) and (5) $\partial U_C/\partial N\leq 0$. Thus, more users increase both the value of the content on the network and the cost of congestion for network users. In addition, while it is true that for networks to form at all $\partial U/\partial N$ must be positive for sufficiently small N, over the full spectrum of network sizes $\partial U/\partial N$ may be either positive or negative based on the relative magnitudes of the content and congestion effects.

Thus, the optimal number of users is bounded if the second derivative of utility with respect to the number of users is negative ($\partial^2 U/\partial N^2<0$). By (2), (3), (6), and (7) this will be the case if both of the following hypotheses are true:

Hypothesis 1: $\partial^2 f/\partial N^2<0$

Hypothesis 2: $\partial^2 c/\partial N^2>0$

Hypothesis 1 is consistent with analytic models in the club goods literature that the propensity to free-ride increases in network size (e.g., Palfrey and Rosenthal 1984, Hindriks and Pancs 2001). If free-riding increases with network size, additional users will be less likely to provide either new content or replicas of existing content on the network as network size increases. Hypothesis 2 is consistent with the technical characteristics of the network in terms of capacity constraints on network logins, user-defined constraints on the number of simultaneous downloads, and technological constraints on the number of simultaneous queries that can be processed by the centralized catalog.

To test hypothesis 1, we measure the collective content on the network in terms of availability and replication. Availability measures the number of unique songs that are provided on the network. Replication measures number of copies of each song and may be a particularly important measurement of network behavior. As noted, a default property of the Napster software is that consumers of a song also become providers for the song, auto-replicating the song for the network. Autoreplication allows a P2P network to efficiently meet download demand from users. A more popular song will have more providers than a less popular song. The value of replication is that it helps distribute the load on the providers if multiple users choose to download songs simultaneously. It is important for replication to scale consistently with network size in order for download performance to scale well.

In an ideal case, the replication of a song will always scale consistently with network size. That is, the replication per user of a song will always remain constant and equal the fraction of all users desiring the song. However, this will not be the case to the extent that users choose to free ride by consuming network resources but disabling sharing. This problem may be exacerbated in larger networks because, as noted above, it is well established in the economics literature that the

private provision of public goods tends to diminish as group size increases. In applying this result to P2P networks it is important to note that in a typical case of private provision of public goods, individuals have to take action to contribute resources. In the case of P2P networks, individuals have to take action *not* to contribute resources. However, despite the necessity of taking action to disable file sharing, users may do so if they are concerned about legal risks or the congestion other users will impose on their connection to the network.

To test hypothesis 2, we measure the cost of accessing content on the network in terms of login congestion, query congestion, download attempt congestion, and download speed congestion. These measures of the negative network externalities reflect the steps in user interaction with centralized P2P networks where the congestion or delays may take place. Login congestion measures the difficulty of logging on to the network. We expect login congestion to be low initially and to quickly rise as network size approaches server capacity. Query congestion measures the delay in waiting for a search query result. When users perform search queries for a file, they place traffic demand on the centralized servers that perform database lookups, potentially degrading network performance for other users. This may happen in two ways: having more users may increase the size of the database that contains the listing of the files provided by the users; and having more users may generate more simultaneous search queries that the centralized servers must process. Download attempt congestion measures the number of attempts that a peer must make before they find a provider that does not queue their download request. As noted above, P2P nodes can define a maximum number of simultaneous downloads they are willing to serve. Requests above this value are then queued for subsequent processing. Download speed measures the amount of time to download content over the network. If some degree of free riding is found in the analysis of replication, we expect that fewer providers will

be handling more concurrent downloads in larger networks, increasing the likelihood that download attempts are queued and decreasing download speed as more users share bandwidth limited connections.

4. Data

To empirically test these hypotheses, we collected data from six OpenNap networks on network congestion characteristics and content availability for 170 songs. The six OpenNap networks used in the data collection were the most popular networks listed by Napigator.com at the beginning of our collection period. The 170 songs were selected at random from the full repertoire of all popular artists in 17 separate genres listed at Amazon.com.[4] Our data were collected every 18 hours from December 2000 to April 2001 and include user count, server count, login congestion (the number of login retries before a successful login), query congestion, and song availability, song replication (total number of copies on the network), and broadband song replication (total number over broadband connections) (Table 1). These data were collected using an automated software agent written for this purpose. The agent implemented an open source documentation of the Napster protocol and was specifically designed to mimic the actions of typical Napster users.[5]

We choose popular artists because song availability was very low for a random selection of songs from all artists. The main drawback of this approach is that some tracks may become less popular over our data collection. However, the content was selected from the full repertoire of the artist (not just their most recent album) meaning that the only a few tracks were recent

[4] We used Amazon.com's listings after determining that it had one of the most comprehensive publicly available databases of music content available on the Internet.
[5] This agent was also designed to have a negligible impact on network performance by spreading out content queries over time and by only downloading a small portion of songs when determining download speeds.

releases.[6] We further checked the sensitivity of our results to changes in popularity by referencing the 36 most popular album charts tracked by Billboard at the beginning and end of our sample period. We found 5 songs that were contained in albums that moved off the Billboard album charts during our sample period. Eliminating these songs from our analysis would not change any of our results

Table 1: Summary Statistics

Variable	*Obs.*	*Mean*	*St. Dev.*	*Min*	*Max*
Login Congestion, Query Congestion, and Song Availability					
User Count	323	3,118	2,283	68	8,618
Server Count	323	7	3.40	1	15
Song Availability	83,640	0.54	0.50	0	1
Song Availability (Broadband Connection)	83,640	0.45	0.50	0	1
Number of Songs	83,640	11	30	0	555
Number of Songs (Broadband Connections)	83,640	6	21	0	460
Login Congestion (Seconds)	323	3	8	0	71
Query Congestion (Seconds)	323	10	17	0.13	90
Download Attempts and Speed					
User Count	13	2,620	687	1,458	3,588
Download Attempts	582	2.85	4.37	1	45
Download Speed (kbps)	582	32	33	0	200

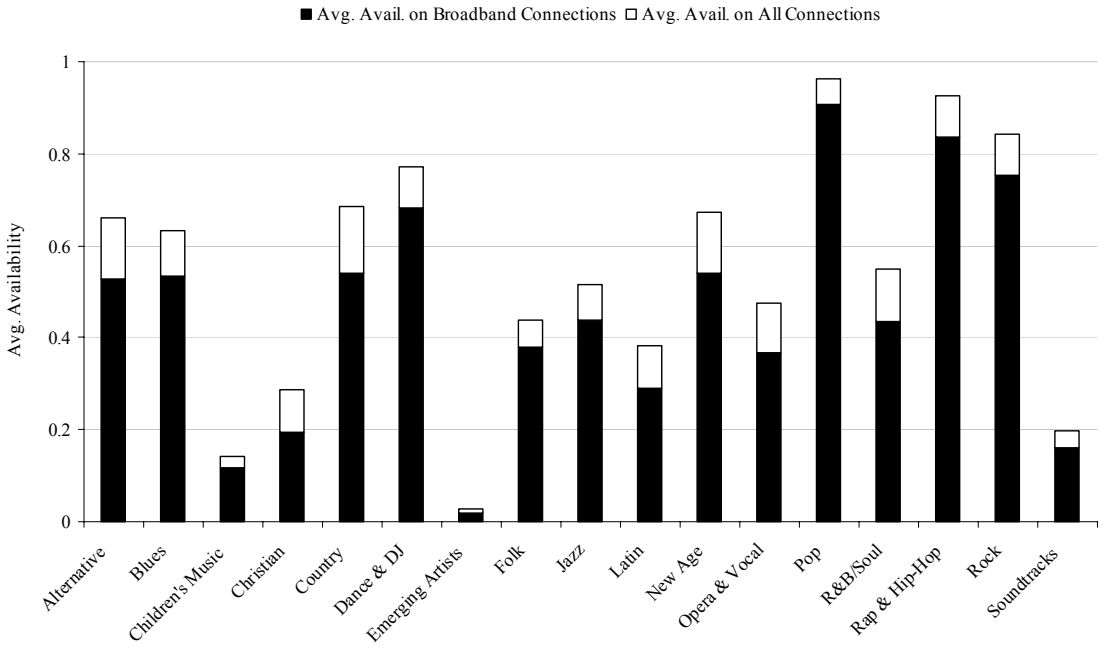
Our data exhibit substantial variation in music track availability across genres and connection types. Figure 1 shows that on average OpenNap networks had 95 percent of the tracks in Pop, the most popular genre, and 2.7 percent of the tracks in Emerging Artists, the least popular genre. If a user restricted preference to the music available through a broadband connection, the numbers fell to 90 percent for Pop and 1.7 percent for Emerging Artists.

To further explore how congestion varies with network size, we collected an additional dataset on download congestion and speed in March 28, 2001 to April 19, 2001 (Table 1). This dataset

[6] We note that for all genres except emerging artists the list of best selling artists did not change over the data

includes information on the size of the network and two measures of the congestion a user would face when trying to download a song. The first measure, download attempts, is the number of download attempts our agent had to make (starting with the listing with the lowest ping time) before finding a peer that did not queue the download request. The second measure, download speed, is the download speed our agent observed when downloading the song.

Figure 1: Average Availability on OpenNap Networks



5. Empirical Analysis of the Network Externalities

5.1. Positive Network Externalities

Hypothesis 1 states that the marginal value that a user brings to the network, measured in terms of availability and replication, will decline in the size of the network. We use two models to test this hypothesis. For availability, we use a logit model to regress song availability (0/1) onto user count. For replication, we use an OLS regression of the number of song replicas onto user count. In each case, we compare the fit across three different specifications for user count: the log of

collection period

user count, a third degree polynomial of centered user count, and user count alone for comparison. Centering is used to control for multicollinearity among the polynomial terms (Aiken and West (1991), Bryk and Raudenbush (1992)).[7] We also use dummy variables for connnection speed, time period,[8] music genre, and network to control for other, potentially confounding, sources of variation.

Results for both regressions are presented in Table 2. Consistent with hypothesis 1, specifications 1, 2, 4, and 5 show that while user count has a strong positive effect on the probability that a music track is available and on the number of replicas, marginal value of additional users diminishes as networks increase in size. Specifications 1 and 4 provide the best fit for the two separate regressions, but their results are similar to specifications 2 and 5. Furthermore, each of these specifications of user count has better fit than the linear specifications (specifications 3 and 6).

These results are shown graphically in Figures 2 and 3, which use the estimated coefficients for specifications 1 and 4 to graph availability and replication as functions of user count for the Pop, Jazz, and Emerging Artist genres. These graphs show that the availability and replication results vary significantly across genres.[9] However, in each case the marginal value an additional user brings to the network declines with the number of users, confirming hypothesis 1. As noted above, in the absence of free riding, we would expect replication to scale linearly with network

[7] Standard diagnostics suggest that centering in this way reduces collinearity in our polynomial terms.
[8] Time period I ranged from the beginning of data collection to when Napster announced its subscription plan on January 29, 2001. Time period II ranged from the subscription plan announcement to when Napster started filtering copyrighted tracks on March 2, 2001. Time period III ranged from the filtering of music tracks on Napster to the end of data collection. These time periods mark significant changes in OpenNap usage as users migrated from the Napster network to OpenNap.
[9] For example, the 4,000th user to join a network is 37% as likely to provide new pop content, 48% as likely to provide new jazz content, and 68% as likely to provide new emerging artist content as the 2,000th user to join the network.

size. The results shown are consistent with an increase in free riding as network size increases. We discuss this possibility in more detail below.

Table 2: Regression Results for Positive Network Externalities

Specification	*1*	*2*	*3*	*4*	*5*	*6*
Regression Model	*Logit- Log*	*Logit-Polynomial*	*Logit-Linear*	*OLS-Polynomial*	*OLS-Natural Log*	*OLS-Linear*
Dep. Var.	Avail.	Avail.	Avail.	Replication	Replication	Replication
Ln(user_count)	0.467				4.63	
	(0.008)				(0.07)	
user_count			1.82e-04			1.74e-03
			(4.05e-06)			(3.27e-05)
user_count'		4.72e-05		0.003		
		(1.27-e05)		(4.17e-05)		
user_count'^{2}		-1.51e-07		-2.77E-07		
		(6.47e-09)		(1.64e-08)		
user_count'^{3}		9.03e-12		*-1.71e-12*		
		(5.04e-13)		*(1.27e-12)*		
broadband	-0.486	-0.485	-0.481	-4.74	-4.73	-4.74
	(0.012)	(.012)	(.012)	(0.11)	(0.11)	(0.11)
time_II	-0.223	-0.219	-.275	-2.93	-3.12	-3.67
	(0.013)	(0.013)	(0.013)	(0.13)	(0.13)	(0.13)
time_III	0.080	0.063	0.167	-6.06	-7.86	-7.96
	(0.025)	(0.025)	(0.025)	(0.21)	(0.20)	(0.20)
Genre [16]	Yes	Yes	Yes	Yes	Yes	Yes
Network [5]	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	166,770	166,770	166,770	166,770	166,770	166,770
(pseudo) R^2	0.253	0.252	0.247	0.199	0.191	0.186

Notes: Centering is accomplished as user_count' = avg(user_count) - user_count. Standard errors are in parentheses. Values in brackets denote the number of fixed effect variables for genre and network type. Italicized coefficients are insignificant (P=.05).

Figure 2: Availability Regression Result

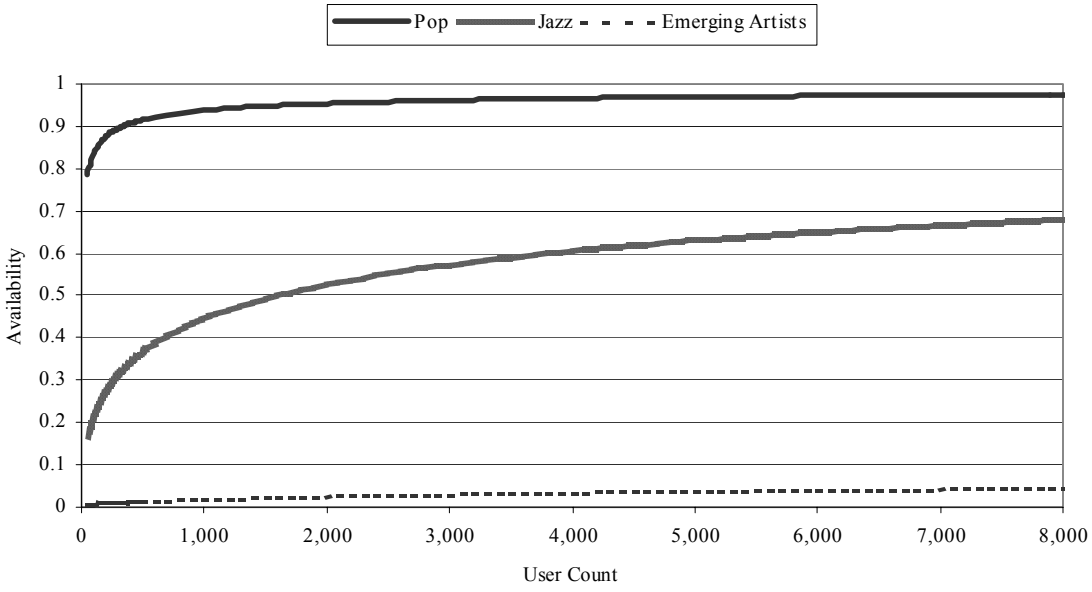
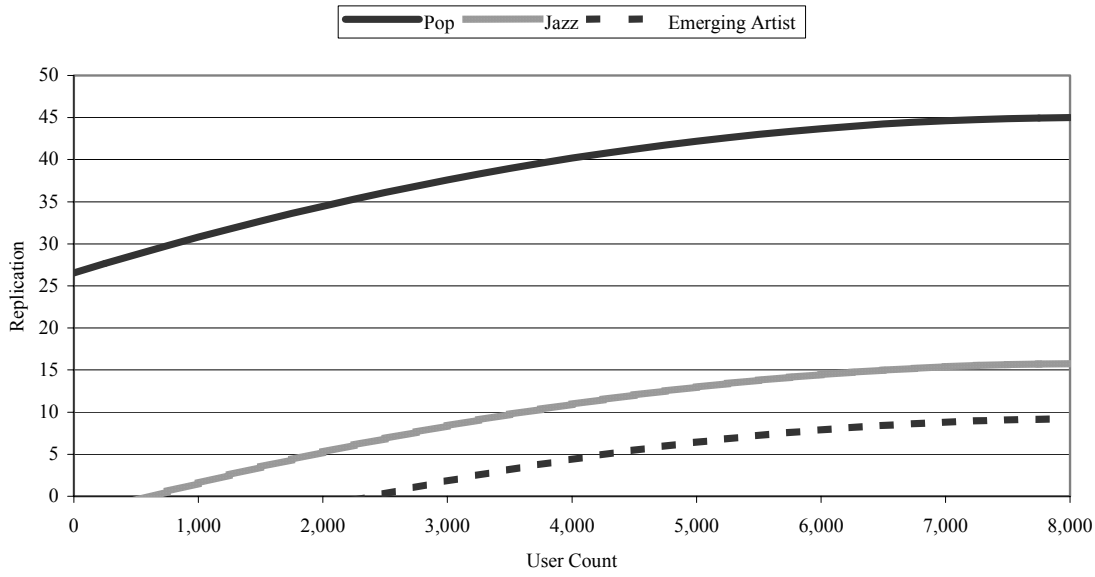


Figure 3: Replication Regression Result



5.2. Negative Network Externalities

As noted above, the negative network externalities are reflected in four measures: an increase in the number of login retries necessary to access the network, longer query times, an increase in

the number of queued download attempts, and longer download times. We use four separate regressions to analyze how these measures change with network size.

For login congestion, we model the number of login retries necessary to gain access to the network as a function of user count and server count (a proxy for network capacity). We use a zero-inflated Poisson regression model (Lambert 1992), and control for the fixed effects for time periods and networks. The Poisson model captures the behavior of a count dependent variable with a long right tail. Zero-inflation controls for the fact that below network capacity no retries are necessary. For query congestion, we regress the log of query congestion onto user count, server count, and the fixed effects for networks and time periods. Download attempts are analyzed using a Poisson model of download attempts onto user count and genre and network fixed effects.[10] We analyze download speed congestion using an OLS model of download speed onto user count and the connection type fixed effects (network fixed effects are collinear with user count in the download dataset).

Table 3 presents the results of the four regressions. Consistent with hypothesis 2, congestion increases in user count at an increasing rate for all types of congestion analyzed. The relationship between user count and congestion is shown graphically in Figure 4, which projects our results in terms of length in seconds. We assume each login retry and download attempt to take 12 and 15 seconds. We estimate download speed for downloading a 5MB file from a cable modem. These assumptions reflect the average values in our empirical analysis.

[10] As noted above, the supplemental dataset on download congestion was collected over a 3 week time period (Mar 28 to April 19) and therefore we do not use time period fixed effects in these regressions.

Table 3: Regression Results for Negative Network Externalities

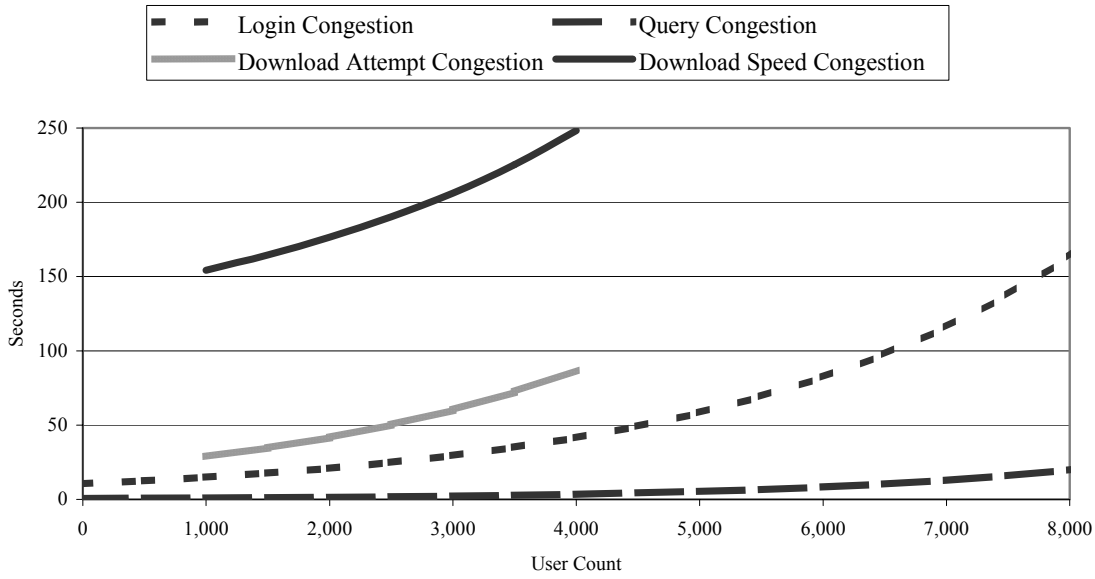
Regression	***Login Poisson***	***Login Inflated***	***Query Time***	***Download Attempt***	***Download Speed***
Method	*Zero-Inflated Poisson*		*OLS*	*Poisson*	*OLS*
Dep. Var.	# of Login Retries	Prob. No Congestion	Log of Query Time	# of Download Attempts	Download Speed (kbps)
user_count	2.8e-04	-7.1e-04	4.7e-04	4.17e-04	-4.1
	(4.9e-05)	(1.7e-04)	(-7.3e-05)	(-6.9e-05)	(1.98)
server_count	*0.0085*	*-0.056*	-0.079		
	(0.148)	*(0.064)*	(0.029)		
time [2]	yes	yes	yes		
Network [5]	yes	yes	yes	yes	
genre [16]				yes	Yes
Connection [10]					Yes
observations	323	323	323	582	582
(pseudo) R2	0.28		0.45	0.12	0.18

Notes: Standard errors are in parentheses. Values in brackets denote the number of fixed effect variables. Fixed effects are suppressed for simplicity. Italicized coefficients are insignificant at p=.05.

5.3. The Impact of Increasing Server Capacity

Thus, our empirical results seem to confirm hypotheses 1 and 2, which in turn suggest that network utility is concave in the number of users and that the optimal network size is bounded in the number of users. One obvious question arising from this analysis is how will these bounds change as capacity is added to the network. In this section we use standard telecommunications traffic models to analyze this question. We find that additional capacity will allow more users to access the network before the network returns to its previous levels of login congestion. However, additional users still add benefits at a decreasing rate (i.e., $\partial^2 U_F/\partial N^2 \leq 0$). Further, additional capacity does not solve the primary user-level problems demonstrated above: increasing free riding, increasing download attempt congestion, and decreasing download speed with larger networks. In sum, increased capacity has limited power to increase optimal network size.

Figure 4: Congestion Summary



Our formal analysis of this question is as follows. We first use the Erlang B equation (Frankel 1976) to estimate the effect of changing capacity on login congestion.

$$P(login\ congestion) = \frac{\frac{\rho^C}{c!}}{\sum_{i=0}^{c} \frac{\rho^i}{i!}} \tag{12}$$

The Erlang B equation models the probability of congestion in a telecommunications switch. The centralized server in our setting is analogous to the switch since peer computers maintain stateful connections to the login server as telephones do to a switch for the duration of a call.

In this formulation $\rho = \lambda/\mu$, where λ is a Poisson random variable for the average number of users who arrive at a network each day; μ, also a Poisson random variable, is the service rate for each connection; and c is the capacity of the network. We calibrate these parameters as follows. Our empirical data indicate that on average users hold a connection for 12 hours, therefore $\mu=2$ connections per day. We use two capacity sizes: c=4,000 (approximately the mean network size

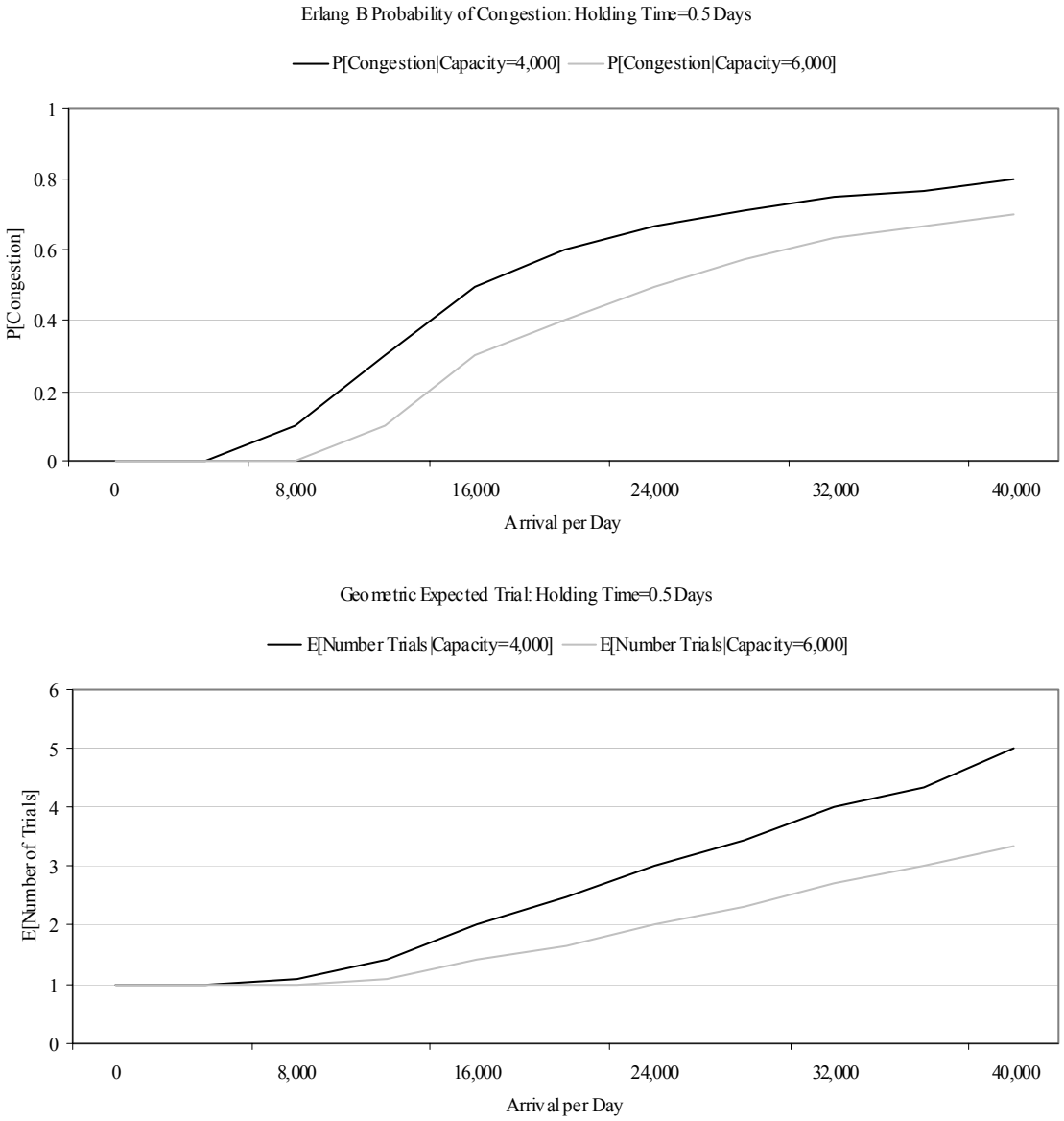
in our data), and c=6,000 (a larger network in our data). To model increases in arrival rate resulting from increasing capacities, we allow λ to vary between 0 and 40,000 users per day. Given the probability of congestion from the Erlang B model, we model the number of retries before a successful login as a geometric random variable, which is the standard model for access attempts above capacity. The average number of retries is given by the mean of the geometric random variable (equation 2):

$$E(login\ retries) = \frac{1}{1 - P(login\ congestion)} \tag{13}$$

This formulation may not exactly match the values found in the empirical analysis because, in the data collected, the agent repeatedly tried to login without any wait time. Thus, the retry attempts may not be independent as assumed by the geometric model. Nevertheless, the model should yield generally consistent results when compared to our data.

Using equations 12 and 13, Figure 5 demonstrates the rate of change of the probability of congestion and expected number of retries at two different capacity levels as a function of the arrival rate of users to the network. For any given arrival rate, it is clear that as capacity increases, both measures of congestion decrease. However, as the arrival rate increases, the same levels of congestion recur in the higher capacity network. Further, as noted above, the additional users attracted by the additional capacity provide value in terms of availability and replication at a diminishing rate. This suggests that there may be little incentive for network operators to increase capacity because the congestion may rise to the same level with little gain in availability or replication.

Figure 5: Illustration of Increasing Capacity on Login Congestion



6. Discussion and Areas for Future Research

P2P networks have had a significant impact on the distribution of information goods and may play a significant role in knowledge sharing and knowledge management within the enterprise in the future. However, despite their importance and potential, there is little academic research analyzing the their value and performance characteristics as a function of network size. This research addresses this gap by analyzing the positive and negative externalities that additional

users impose on centralized P2P networks, and how these externalities impact optimal network size and scalability.

Using data collected from the six most popular OpenNap networks we find that additional users provide positive network externalities based on the quantity and uniqueness of the files they provide and negative network externalities in terms of login, query, download attempt and download speed congestion. However, the marginal value of an additional user declines with the number of existing users while the marginal cost of congestion imposed by users increases with network size. Using a reduced form utility model we show that these findings imply that optimal network size is bounded for these centralized P2P networks.

We also find that the number of replicas of content per user decreases with network size. This is consistent with findings in the public economics literature that free riding worsens with group size and suggests that increased free-riding in larger P2P network is inherent to the collective action of users in the private provision of public goods. While it is impossible to isolate this explanation from a decrease in replicas due to unobserved customer heterogeneity, an increase in free riding with increasing network size would imply that we must treat the content on P2P networks as rivalrous goods in which autoreplication fails to scale supply to meet demand and over-consumption can lead to congestion.

We use traffic models to explore the impact of increasing network capacity on optimal network size. We find that increased capacity has a limited impact on optimal network size and that the value of increased capacity decreases with network size. We also note that, while login and query congestion are driven by capacity considerations, download attempt and download speed congestion are driven by free riding which is worsened by increased network size in our data. In

short, network performance may only be optimized through the use of managerial rules or policies that align user incentives with the desired outcome of the collective network. Analysis of such rules using both priced and non-priced incentives is a fruitful area for future research.

A business implication of this study is that the value of the P2P network does not scale in the way that traditional networks, such as telecommunication networks, do. The value present in telecommunications networks is a function of the number of users where marginal value is increasing in network size. In contrast, the value in P2P networks is based on collective content and the marginal value of collective content is decreasing in the number of users. Because of this, P2P networks are unlikely to be "winner-take-all markets." This suggests that network operators should adopt niche strategies based on features or content to maximize the value provided to their share of network users.

The policy implication of this study is that P2P networks, in their current stage, follow the economic theory of private provision of public goods. Free riding exists and can decrease network scalability. Unless appropriate private incentives are implemented through managerial rules or pricing policies, the degree of free riding will eventually outweigh the benefit of having more users in the network. From a technological perspective, these observations stress the importance of incorporating economic-based managerial rules or pricing policies into protocol designs to align private user incentives with the goals of the collective network.

It is important to note that our empirical results only apply to the centralized peer to peer architectures used in the OpenNap networks to provide consumer information goods. Future work should focus on extending our results to other context domains such as peer-to-peer

networks for information sharing within corporations or in other architectures such as Gnutella, Aimster, or Kazaa/Morpheus.

References

Adar, E., B. A. Huberman. 2000. Free Riding on Gnutella. *First Monday: Peer-Reviewed Journal on the Internet.* **5**(10). Retrieved August 16, 2001 from www.firstmonday.dk/issues/issue5_10/adar/index.html.

Aiken, L.S. and West, S.G. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, Newburg Park, California.

Bryk, A.S. and Raudenbush, S.W. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newburg Park, California.

Buchanan, J. M. 1965. An Economic Theory of Clubs. *Economica* **32**(125) 1-14.

Cameron, A. C., P.K. Trivedi. 1998. *Regression analysis of count data*. Cambridge University Press, New York.

Davis, Douglas D. Charles A. Holt. 1993. *Experimental Economics*. Princeton University Press, Princeton, NJ.

Frankel, Theodor. 1976. *Tables for Traffic Management & Design*. ABC Teletraining, Dallas, Texas.

Golle, P., K. Leyton-Brown, I. Mironov. 2001. Incentive for Sharing in Peer-to-Peer Networks. Working Paper, Computer Science Department, Stanford University, Palo Alto, CA.

Hindriks, J. and Pancs, R. 2001, Free Riding on Altruism and group size. Working paper, University of London.

Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1-14.

Ledyard, John O. 1995. Public Goods: A Survey of Experimental Research, Kagel and Roth, eds. *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ.

Olson, M. 1968. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA.

Palfrey, T., H. Rosenthal. 1984. Participation and the provision of discrete public goods: a strategic analysis. *Journal of Public Economics* **24** 171-93.

Samuelson, P. 1954. The Pure Theory of Public Goods. *The Review of Economics and Statistics* **36**(4) 387-389.