

Diffusion Models for Peer-to-Peer (P2P) Media Distribution : On the Impact of Decentralized, Constrained Supply

Kartik Hosanagar • Peng Han • Yong Tan

Operations & Information Management, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104
aQuantive, Inc. (Microsoft), 821 2nd Avenue, Seattle, WA 98104
Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195

kartikh@wharton.upenn.edu • realbug@gmail.com • ytan@u.washington.edu

Abstract

In Peer-to-Peer (P2P) media distribution, users obtain content from other users who already have it. This form of decentralized product distribution demonstrates several unique features. Only a small fraction of users in the network are queried when a potential adopter seeks a file and many of these users may even free-ride i.e. not distribute the content to others. As a result, generated demand may not always be fulfilled immediately. We present mixing models for product diffusion in P2P networks that capture decentralized product distribution by current adopters, incomplete demand fulfillment and other unique aspects of P2P product diffusion. The models serve to demonstrate the important role that P2P search process and distribution referrals – payments made to users that distribute files – play in efficient P2P media distribution. We demonstrate the ability of our diffusion models to derive normative insights for P2P media distributors by studying the effectiveness of distribution referrals in speeding product diffusion and determining optimal referral policies for fully decentralized and hierarchical P2P networks.

Keywords: Peer to Peer file diffusion, P2P, supply-constrained diffusion, free-riding, mixing model of diffusion, distributed systems.

Acknowledgements: The authors would like to thank Sumit Sarkar, an Associate Editor and two anonymous referees for their valuable feedback. All errors remain our own.

1. Introduction

Peer-to-Peer (P2P) networks are distributed networks in which the participants share their own resources in addition to consuming them from others. In P2P-based media distribution, users download content from other users who have the content and in turn redistribute it to future adopters. P2P allows a content provider to efficiently distribute content at a relatively low cost and is also effective in handling flash crowds in content distribution (Padmanabhan and Sripanidkulchai 2002). A number of encryption technologies have also emerged to prevent piracy in P2P networks. Thus, although the early use of the technology was for illegal file sharing, P2P is increasingly being adopted for legitimate media distribution on the Internet. P2P is being used for online radio (e.g. Social.fm), Internet TV (e.g. Joost) and software distribution (e.g. distribution of the RedHat Linux OS on BitTorrent). Recently, NBC Universal and AOL announced Internet TV initiatives based on P2P technologies. In addition, Altnet, iMesh, Grooveshark, rVibe, We7 and several other firms use a P2P distribution platform to sell music and other digital media licensed from the music labels. In 2004, there were over 50 million legal downloads per month on Kazaa for over 10,000 titles in Altnet's library (Currah 2004). In 2007, over 9.5M music files were uploaded by 10,000 beta users of Grooveshark's P2P music download service.¹

Nodes in a P2P network are potential consumers of a digital product and, by virtue of the design of P2P networks, can redistribute the product upon buying it. P2P distribution is unique relative to centralized media distribution in that decentralized supply often imposes a constraint on demand fulfillment. In most P2P architectures, only a fraction of the nodes in the network are queried in response to a request in order to prevent flooding the network with queries. Thus, even if a file exists in a network, nodes containing the file may not always be queried. Further, even if a node containing the requested file is queried, the node may not distribute the file. Free riders – users that consume content from the network, without sharing or redistributing it to other users – are known to be pervasive in P2P networks (Adar and Huberman 2000). As a result of these two factors, generated demand may not always be satisfied. Thus,

¹ Source: personal conversations with Sam Tarantino, CEO of Grooveshark, January 2008

“supply-side” factors related to incomplete search of the network and the redistribution incentives have a crucial impact on file diffusion within the network.

Commonly used redistribution incentives include penalties to users who free-ride and rewards to users who contribute. Penalties include intentionally slowing the download of free-riders as in the BitTorrent protocol. Penalties are used sparingly in commercial P2P systems where users pay to obtain content. Commercial P2P networks often provide payments, known as distribution referrals, to users who distribute content to others in the network. Distribution referrals have been advocated in various studies (e.g., Golle et al 2001; Arora et al. 2003) and are used in several commercial P2P networks. For example, Altnet pays users on the Kazaa network who agree to join Altnet as distribution points (New York Times 2003). Grooveshark and rVibe also compensate users for distributing content. In these networks, whenever a new user purchases a track, a small payment is made to the user that distributes the content to the buyer. rVibe currently pays the distributing user \$0.05 on a \$0.99 sale and Grooveshark pays the user \$0.25 per sale (Techcrunch 2007). Hereafter, we refer to these payments as referrals.²

In this paper, we propose a model for diffusion of digital products in P2P networks that explicitly captures the supply-side factors - file search and redistribution incentives - described above and demonstrate the applicability of our model by studying optimal payments to users who distribute content. Modeling product diffusion is of considerable interest to managers. Diffusion models can be used for demand forecasting and for planning a variety of pre-launch and post-launch strategic decisions such as optimal level of product sampling, optimal pricing and optimal timing of successive generations of a product (Mahajan et al. 2000). As a result, product diffusion models have been actively studied in marketing for over forty years. These models primarily focus on the demand generation process. They generally assume that the generated demand is always fulfilled and do not model the important supply-side constraint in P2P networks. The few papers in the product diffusion literature that have modeled

² However, it is useful to distinguish the referrals in P2P from traditional referrals. Traditional referral fees are paid to existing customers for bringing new customers to the firm. In contrast, the referral in P2P encourages existing customers to distribute content, which increases the file availability and in turn sales. Thus, the primary impact is on the supply-side.

supply constraints have done so in centralized settings where supply does not depend on the number of current adopters or actions taken by them. In contrast, current adopters generate the supply in P2P diffusion. These supply-side factors, tied to the file search process and free-riding in the network, significantly influence product diffusion in P2P networks and are in fact of most interest to P2P network managers. It is thus important to incorporate these into diffusion models.

Our study of product diffusion in P2P networks makes three main contributions. First, we develop a diffusion model that incorporates supply-side constraints unique to P2P networks and derive analytical results regarding the sales dynamics. Specifically, we develop a mixing model of file diffusion that incorporates both incomplete search and free-riding in P2P networks. Second, we present an application on optimal referrals that endogenizes the impact of distribution referrals on file availability and overall profits. We derive expressions for the optimal referral and show that a referral policy that accounts for diffusion dynamics is far more profitable than a myopic policy that ignores these dynamics. Finally, we demonstrate that the file search architecture exerts a strong influence on diffusion. We find that a hierarchical architecture with a few groups demonstrates faster diffusion than flat P2P networks. However, there are diminishing returns from increased centralization.

The rest of this paper is organized as follows. In Section 2, we review related literature. In Section 3, starting with existing epidemiological models, we develop analytical models to capture the dynamics of diffusion in P2P networks. Section 3.1 introduces the notation. Section 3.2 develops diffusion models for flat P2P networks and includes a study of the optimal referral policy. Section 3.3 focuses on product diffusion and optimal referrals in hierarchical P2P networks and examines the impact of the search architecture on the rate of diffusion. Section 3.4 employs simulations to test the robustness of the models under more general demand processes. Section 4 concludes this study, and discusses future work.

2. Literature Review

There are three streams of work highly relevant to our study of P2P content diffusion. The first two relate to the literature on new product diffusion and epidemiological diffusion respectively. The third relates to studies of free riding in P2P networks and the use of referral payments to address the same.

Product Diffusion: Direct work on P2P content diffusion is limited. A notable exception is the work by Izal et al. (2003) on an empirical study of file diffusion in the BitTorrent network. Developing an analytical model of file diffusion within networks is not the focal point of the paper. However, there is a vast body of work on new product diffusion models, dating back to 1960s. Fourt and Woodlock (1960) propose a product diffusion model in which a fixed fraction of the consumers who have not yet bought the product do so every period. Bass (1969) proposed an extension that additionally incorporates word of mouth (WOM) communication between current adopters and potential adopters. Building on this model, work has also been done to incorporate effects of advertising and promotion (Horsky and Simon 1983), competition (Krishnan et al. 2000) and pricing (Bass 1980).

These diffusion models focus on demand generation and assume that demand is always fulfilled. In P2P networks, a potential adopter is only able to conduct an incomplete search of the network and even the queried nodes may free-ride. As a result, demand is often not fulfilled immediately. In this sense, the supply-constrained diffusion models of Ho et al. (2002), Kumar and Swaminathan (2003), and Jain et al. (1991) are more relevant to our diffusion model. These papers study situations where demand is not met due to production capacity constraints. However, they focus on centralized settings wherein the managerial intervention is tied to capacity sizing. In contrast, the distribution infrastructure in P2P networks is decentralized and the product supply involves a social process. The more the number of current adopters and higher their willingness to distribute a product, greater is the file availability in the network. Thus, product adoption directly increases product supply. Further, the relevant managerial interventions relate to the design of the search process and incentives to encourage nodes to distribute content rather than an increase in the centralized capacity. Our diffusion model uniquely captures these variables and the dependence between product diffusion and social supply.

Epidemiological Diffusion: Infectious diseases spread from infected nodes to susceptible nodes and are examples of decentralized diffusion processes. Epidemic diffusion models have been used in computer networking research, including in studies of information diffusion in mobile ad hoc networks (Khelil and Becker 2002) and spread of computer viruses (Kephart and White 1991). The diffusion

mechanism in P2P networks is in many ways similar to the spread of diseases. Broadly speaking, when a node seeks content in a P2P network, a request is sent out to other nodes in the network. File transfer is completed once a node that shares the desired resource is found. Similarly, in the spread of diseases, an infected individual makes contact with people around her and disease transmission occurs once a susceptible individual is contacted. Thus, the susceptible in epidemic diffusion is analogous to the node seeking content in P2P and the infected is analogous to the node that distributes content.

Despite these similarities, diffusion in a P2P network is unique in several ways. One major difference is that the susceptible agent typically receives a disease passively whereas the node seeking content initiates the contact in P2P diffusion. This implies that epidemic contact occurs in a semi-random manner, while the search process in P2P network is not. P2P search can be architected and the design choices have a notable impact on diffusion. For example, although the content in P2P networks is always distributed, the search process can be completely centralized, completely decentralized/flat or hierarchical (Asvanund et al 2004). Networks like Napster maintained a single central catalog and search requests were forwarded to the central server. Gnutella version 0.4 is completely decentralized with each node maintaining its own catalog and responding to search queries. Gnutella version 0.6 and Kazaa use a hierarchical architecture wherein nodes connect to supernodes which are in turn connected to each other. The supernodes index the content for their nodes and respond to search requests. An additional difference is that the reproductive capacity of a virus usually grows proportionately with the diffusion of a disease. In contrast, file availability and free-riding impose a constraint on product diffusion in P2P networks. Finally, epidemiologists are interested in slowing the diffusion through vaccinations or by quarantining the infected. P2P managers are interested in speeding the diffusion by appropriately architecting P2P search process or using referrals and other incentives to encourage users to distribute content.

Free Riding in P2P and Payment-based Incentives: Free riding, which has been widely documented in P2P networks (Adar and Huberman 2000, Asvanund et al 2004), can slow the diffusion of products within the network. Several approaches have been proposed to alleviate this problem, for example, offering higher Quality of Service (QoS) to users that share their resources (Kamvar et al 2003).

In commercial P2P networks, where users pay for content, content providers can use payments to encourage users to distribute content to others in the network (Arora et al. 2003, Lang and Vragov 2005). Golle et al (2001) discuss the use of micropayments to reward peers for distributing content. A number of commercial P2P systems such as Grooveshark and rVibe use payment-based incentives to encourage users to share and distribute content.

3. P2P Diffusion Model

The diffusion models we develop are homogeneous mixing models. In homogeneous mixing models, there is no spatial structure, i.e. specific neighbors of a node are not modeled, and all nodes within a compartment (i.e. of the same type) run the same risk of being infected. One can employ the equivalent of mean field analysis for large populations ($N \rightarrow \infty$) to develop differential equations that capture the dynamics.³ Popular examples of homogeneous mixing models include the SIR/SIS models in epidemiology (Diekmann and Heesterbeek 2000), Bass model in marketing (Bass 1969) and the Lotka-Volterra model in population ecology (Brauer and Castillo-Chavez 2000). The implication of the homogeneous mixing setup is that we do not model the specific topological connections in the P2P network. The diffusion dynamics depend only on the number of nodes seeking or distributing content rather than which particular nodes seek or distribute content. We model the search process in P2P networks within the framework of the homogeneous mixing model. Specifically, groups in hierarchical networks are modeled as additional compartments in the diffusion model.

We begin by introducing our notation in Section 3.1. In Sections 3.2 and 3.3, we focus on diffusion in completely decentralized (flat) networks and hierarchical networks respectively.

3.1 Notation, Definitions and Assumptions

Consider a digital product such as a music file being distributed in a P2P network. Nodes in the network are potential consumers of the product and can obtain it from nodes that already have the product.

³ In mean field techniques, the distributions of quantities over their randomness are represented instead by their average values (Newman et al. 2000). By ignoring certain dependencies, a closed set of equations for the expected values of variables can be derived (Oppen and Saad 2001). Mean field approximations are often very accurate in the limit for large population sizes (e.g., see Newman et al. 2000; Anderson and Britton 2000). We thank an anonymous referee for pointing us to this literature.

Whenever a node distributes a copy of the digital product to another node, it gets a referral payment from the firm. In the discussion below, we assume that the firm paying the referral fee is the copyright holder.⁴

Let N denote the total number of nodes in the network. We assume that there is no intermediary, i.e. a node that buys and sells the product while not seeking it, and that a node is used by one individual only, so no node consumes a file more than once. Thus N is also the market potential, or the total number of nodes that may eventually buy the product. The nodes that have the product already are called *satisfied nodes*. The number of satisfied nodes in the network at time t is denoted by $Q(t)$ and the fraction of nodes that are satisfied is $q(t) = Q(t) / N$. The nodes that do not yet have the product are called *seeking nodes*. The number of seeking nodes at t is $N - Q(t)$ and the fraction of seeking nodes in the network is $(1 - q(t))$. β denotes the rate at which seeking nodes attempt to seek the product. Seeking nodes can obtain the product from satisfied nodes. Note however that not all satisfied nodes distribute the product. Satisfied nodes that are willing to distribute the product are called *seeds*. The number of seeds and fraction of nodes that are seeds are denoted by $S(t)$ and $s(t) = S(t) / N$ respectively. In the rest of the model development, we simplify notation by denoting the number of seeking nodes, satisfied nodes and seeds by $(N - Q)$, Q and S respectively (and the corresponding fractions by $(1 - q)$, q and s). It is implied that these are dynamic variables. Note that a satisfied node does not have to be a seed, but a seed must be a satisfied node first. Thus, $S \leq Q$.

We assume that the network contains some altruistic nodes that are always willing to be seeds even if there is no compensation for doing so and that the remaining nodes are “strategic” i.e. they will not be a seed unless the distribution referral is greater than their cost of being a seed. The fraction of nodes that are altruistic is denoted by α and the number of altruistic seeds is αQ . The assumption of altruistic nodes is without loss of generality as α may be set equal to zero. However, some users are known to distribute content in P2P networks in which no distribution referral is offered. This willingness to share may be fueled by a sense of community, reciprocity or other such factors. Let c denote a node’s cost of

⁴ In reality, the firm can be the copyright holder of the content, the online retailer (i.e., the P2P firm) or both. Cooperative promotions between retailers and labels (copyright holders) are quite common in the music industry.

distributing a file and r denote the referral payment made to the seed that distributes the file. A strategic satisfied node will be a seed if and only if $c \leq r$. Assuming that c is Uniformly distributed in $[0, C]$ across the strategic nodes, the number of strategic satisfied nodes that act as seeds is $(1 - \alpha)Q(r/C)$. Therefore,

$$S = \begin{cases} \left(\alpha + \frac{r}{C}(1 - \alpha) \right) Q, & \text{if } r \leq C; \\ Q, & \text{otherwise.} \end{cases} \quad (1)$$

When $r = 0$, file distribution relies solely on the goodwill of the altruistic nodes. We also assume that the firm is unable to price discriminate, i.e., altruistic nodes also get the referral even though they would share regardless.

Finally, we describe the search process in the network and associated notation. We consider both a completely decentralized as well as a hierarchical network in this paper. In a completely decentralized or flat network, whenever a seeking node seeks a file, i neighbors of the seeking node are randomly selected and queried. Each of the queried nodes again forwards the request to another i unique nodes. This process continues until the maximum number of hops, denoted by j , is reached. Thus, the request will be sent to a total of $k = i + i^2 + \dots + i^j = i(i^j - 1)/(i - 1)$ nodes. The search fails if none of these k nodes are seeds.⁵ In Gnutella 0.4, a seeking node will query 7 of its neighbors. If these nodes do not have the file, they each contact 7 of their neighbors, and so on until the maximum hop count of 10 is reached (Ross and Rubenstein 2003).

Now consider a 2 level hierarchical P2P network in which N nodes are organized into M groups. We assume that all the groups are homogenous, except that one of them labeled Group-I has the initial satisfied node. Since the assignment of nodes to groups in current P2P networks is independent of both the request rate of the nodes and the content supplied by them (Garces-Erice et al. 2003), the assumption of homogeneity across groups is reasonable. The number of nodes in any group is given by $n = N/M$.

⁵ We do not model the network topology. We only assume that a node's neighbors are randomly assigned and uncorrelated with whether that neighbor is a seed. Thus, each of the k queried nodes is just as likely to be a seed as any other node in the network. This implicitly models the interaction as though the network is fully connected. The assumption is for analytical tractability but is not as restrictive as it might appear for two reasons. First, P2P networks are regularly rewired and the neighbors are constantly changing as a result. Second, P2P networks have a small diameter (Ianmitchi et al. 2004) so most nodes are reachable in a few hops.

The number of satisfied nodes and fraction of nodes that are satisfied in Group-I are denoted Q_i and q_i respectively. Similarly, Q and q denote the corresponding variables for the other groups. Each group in the network has a super-node that provides indexing services to all the nodes within its group as shown in Figure 1. Any request from a seeking node will be satisfied if there is a seed within the group. This is because the super-node indexes all nodes in the group and can forward the request to the seed directly. If there is no seed in the group, the super node forwards the request to $(l-1)$ randomly selected groups.⁶ Thus, a total of l groups in the network are queried. The probability that a randomly queried satisfied node distributes the file is given by $\phi = \alpha + (1-\alpha)r/C$. Thus, given nq satisfied nodes in a group, the probability that there is no seed in the group is $p = (1-\phi)^{nq}$. Similarly, the probability that there is no seed in Group-I is $p_i = (1-\phi)^{nq_i}$. Table 1 summarizes the notation that will be used in the rest of this paper.

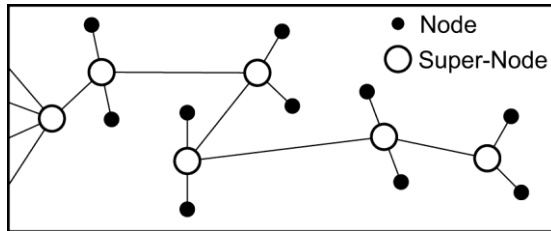


Figure 1: Hierarchical P2P Network

Table 1: Glossary of Terms

t	Time
N	Total number of nodes in the network (also the market potential)
$Q(t)$	Number of satisfied nodes at time t , $Q(0) > 0$; $q = Q / N$
$S(t)$	Number of seeds at time t , $S(0) > 0$; $s = S / N$
α	Fraction of nodes that are altruistic
β	Average rate at which a seeking node seeks content
r	Referral per download offered to a node that distributes content
c	The cost per unit time of being a seed. $c \in U[0, C]$
Notation for a Flat P2P Network	
i	Number of requests generated by seeking nodes in a flat network
j	Maximum number of hops for requests in a flat network

⁶ Again, we assume that neighbors of supernodes are assigned randomly, uncorrelated with the density of seeds in the groups. This models the interaction as though the supernodes are fully connected.

k	Maximum number of nodes queried. $k = i(i^j - 1)/(i - 1)$
Notation for a Hierarchical P2P Network	
M	Total number of groups in a hierarchical network
n	Number of nodes in a single group ($n = N/M$)
l	Number of groups queried in a hierarchical network
Q_l, q_l	Number and fraction of satisfied nodes in Group-I
Q, q	Number and fraction of satisfied nodes in a group other than Group-I
ϕ	Probability that a randomly queried satisfied node distributes the file
$p = (1 - \phi)^{nq}$	Probability that there is no seed in a group other than Group-I
$p_l = (1 - \phi)^{nq_l}$	Probability that there is no seed in Group-I

3.2 Diffusion Model for Flat Networks

We begin by considering a completely decentralized P2P network. Whenever a seeking node seeks content, requests are sent to other nodes and forwarded within the network. A maximum of k nodes are queried as described in Section 3.1. If a request reaches a seed, the file is transferred to the seeking node, which then becomes a satisfied node. Otherwise, the request fails and the status of the seeking node remains unchanged. The node can return in a later period to seek the content. Since β is the rate at which a seeking node seeks content, $\beta(N - Q)$ is the total number of seeking nodes seeking content at an instant when there are Q satisfied nodes. If each of these nodes sends out k requests, $\left(1 - \left(1 - (S/N)\right)^k\right)$ is the fraction of the nodes that reach at least one seed.⁷ Thus, the mean field diffusion equation for Q , i.e. the average rate at which seeking nodes become satisfied nodes, is

$$\frac{dQ}{dt} = \beta(N - Q) \left(1 - \left(1 - \frac{S}{N}\right)^k\right), \text{ or } \frac{dq}{dt} = \beta(1 - q) \left(1 - (1 - s)^k\right), \quad (2)$$

We can substitute the expression for S from (1). A referral equal to C ensures that every satisfied node is a seed. Thus a referral greater than C is unnecessary. Substituting (1) into (2) and assuming $r \leq C$,

$$\frac{dq}{dt} = \beta(1 - q) \left(1 - (1 - \phi q)^k\right). \quad (3)$$

⁷ Because neighbor assignment is uncorrelated with whether a node is a seed, the probability that a neighbor is a seed is just S/N .

where $\phi = (\alpha + (1 - \alpha)r/C)$. Solving the above differential equation with the initial condition of $Q(0) = 1$ (or, $q(0) = q_0 = 1/N$), we obtain the following solution (proof in online supplement)

$$\frac{1}{a^k - 1} \ln(x - a) - \frac{1}{k(a - 1)} \ln(x - 1) - \frac{1}{k} \sum_{j=1}^{k-1} \frac{x_j}{a - x_j} \ln(x - x_j) = \beta t + C_1, \quad (4)$$

where $a = (1 - \alpha)(1 - r/C)$, $x = 1 - \phi q$, C_1 is a constant of integration computed in the online supplement and $x_j = e^{i2\pi j/k}$ is the j^{th} root of $x^k = 1$, not including $x = 1$. In the supplement, we show that (4) has no imaginary parts and that the solution is unique. We also show that sales (dq/dt) is unimodal.

Proposition 1: The product penetration $q(t_c)$ at a given time instant t_c increases with k in a convex fashion early in the file's diffusion (small q) and in a concave fashion late in the diffusion (large q). The time $t(q_c)$ to achieve a given level of product penetration q_c decreases with k in a convex fashion.

The proof is in the Appendix. Figure 2 illustrates proposition 1 for sample parameter values ($N = 10,000$, $Q(0) = 1$, $C = 1$, $\alpha = 0.1$, $\beta = 0.5$, and $r = 0.2$). The time to achieve 50% network penetration decreases with k in a convex manner, and product penetration after 8 time units ($q(8)$) is S-shaped in k . Proposition 1 indicates that the marginal benefit of querying more nodes can be increasing early in the diffusion. But once there are sufficient satisfied nodes, querying more nodes no longer provides the significant returns that it did before. Considering the negative effects of request flooding, there should be a threshold beyond which any further increase in k is not desirable. One potential impact of request flooding is that nodes have to respond to many more requests which in turn may reduce their likelihood of serving as seeds. Our model does not explicitly incorporate these costs of request flooding. However, additional analysis reveals that if the number of seeds decreases with k in a concave manner, there may be a threshold value of k beyond which the gains identified in Proposition 1 are offset by the reduction in the seeds. The analysis is omitted from this paper but available upon request from the authors.

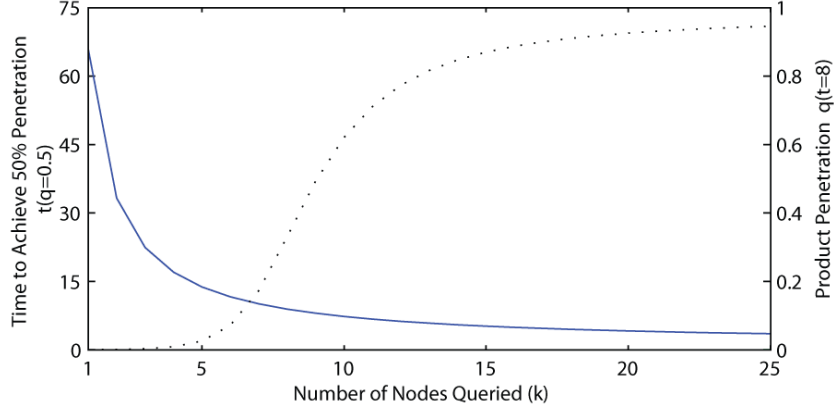


Figure 2: Impact of querying more nodes on diffusion speed

Proposition 2: The product penetration $q(t_c)$ at a given time instant t_c increases with the referral r . Simultaneously, the time $t(q_c)$ to achieve a given level of product penetration q_c decreases with r in a convex fashion.

The proof is in the Appendix. An increase in the referral increases the willingness of satisfied nodes to serve as seeds. This in turn increases file availability and helps speed product diffusion. There is a tradeoff between offering a high referral to speed diffusion and reducing the referral to increase margins. We evaluate this tradeoff in Section 3.2.1 to compute the optimal referral r^* .

Consider a special case in which $k = 1$, i.e. only one node in the network is queried whenever a seeking node attempts to obtain the content. (3) can now be solved to obtain the following closed-form expressions:

$$q(t, r) = \frac{1}{1 + (N - 1)e^{-A(r)t}}, \quad (5)$$

and

$$t(q, r) = \frac{1}{A(r)} \left(\ln \frac{q}{1 - q} + \ln(N - 1) \right), \quad (6)$$

where $A(r) = \beta(\alpha + r(1 - \alpha)/C)$. (5) and (6) specify an S-shaped P2P diffusion curve. These equations can be used to derive the optimal referral that the firm should offer. We turn to that question now.

3.2.1. Optimal Referral

We demonstrate the applicability of the diffusion model by studying optimal referral strategies in P2P networks. Diffusion modeling can be particularly useful here relative to myopic policies that determine referrals based on the immediate impact on sales without accounting for the impact on future sales.

Denote the length of the planning horizon by T . Assume the referral r is constant throughout the planning horizon, and the unit price of the product is constant and normalized to 1. The firm's profit from each sale is $(1 - r)$. Thus, the referral optimization problem is

$$\max_r \left(\int_0^T (1-r) \frac{dq}{dt} dt \right) \approx (1-r)q(T). \quad (7)$$

In the equation above⁸, we do not discount future sales for analytical tractability. The referral optimization problem for arbitrary values of k is relatively easy to solve numerically but closed-form results are hard to come by. In the following analysis, we focus on the case in which $k = 1$ due to the tractability it affords. Numerical analysis suggests that although the magnitude of the referrals changes with k , the results are qualitatively similar to the ones highlighted below.

Substituting $q(t)$ from (5) into (7) and computing the first order condition with respect to r gives:

$$(N-1) \exp\left(-\beta T \left(\alpha + (1-\alpha) \frac{r}{C}\right)\right) \left(\beta T (1-\alpha) \frac{1-r}{C} - 1\right) - 1 = 0. \quad (8)$$

Solving (8) and imposing the bounds on r , the result follows (proof in the appendix):

Proposition 3: The optimal referral for the flat P2P network with diffusion specified by (5) is

$$r^* = \max \left[0, \min \left[C, 1 - \frac{C}{\beta T (1-\alpha)} \left(1 + W \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right) \right) \right] \right], \quad (9)$$

where $W(x)$ is the Lambert W-function (solution of $W(x) \exp(W(x)) = x$).

Using the properties of the Lambert W-function, the following properties of the optimal referral are derived in the appendix.

⁸ Strictly speaking, the objective function is $(1-r)(q(T) - 1/N)$. Since N is large, we approximate this to $(1-r)q(T)$.

Proposition 4: The optimal referral is (a) non-increasing with the fraction of nodes in the network that are altruistic (α), (b) non-decreasing with the P2P network's size (N), and (c) non-decreasing with request rate β for $\beta < \beta_{th}$ and non-increasing with β for $\beta > \beta_{th}$ where

$$\beta_{th} = \frac{1}{T} \left(2 + W \left((N-1)e^{-1} \right) + \frac{e}{N-1} \exp \left(W \left((N-1)e^{-1} \right) \right) \right) \left(\alpha + \frac{1-\alpha}{C} \right)^{-1}. \quad (10)$$

As expected, an increase in the number of altruistic nodes reduces the need for a high referral. Similarly, product penetration occurs at a slower rate in a large network (see (5)). As a result, a higher referral is needed to speed the diffusion in a large network.

Figure 3 illustrates the impact of β for sample parameters ($N = 10000$, $Q(0) = 1$, $C = 1$, $\alpha = 0.1$). When β is extremely small, i.e. request rate among seeking nodes is low, the best strategy for the firm is to offer no referral. This is because the bottleneck is not free riding but low rate of demand generation. Once β reaches a certain threshold value, the referral starts having an impact and thus increases. Finally, a very high request rate helps speed up diffusion considerably. This helps generate altruistic satisfied nodes at a faster rate, which in turn reduces the need for a very high referral. As a result, we observe the switch from a non-decreasing relationship to a non-increasing relationship between r^* and β .

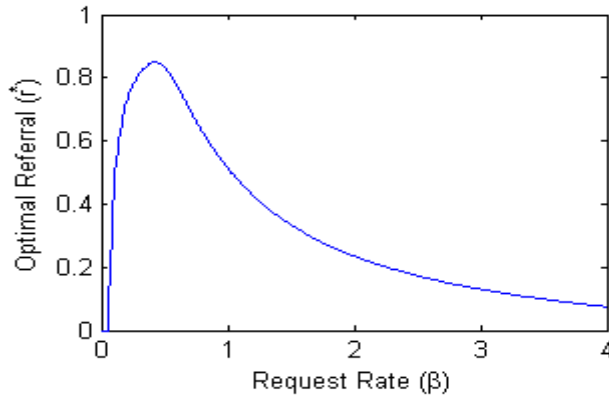


Figure 3: Optimal Referral Vs Request Rate

Now consider a myopic referral policy that does not account for the impact of referrals on future sales. The myopic referral maximizes the instantaneous profit given by

$$\pi(t) = (1-r) \frac{dq}{dt} = (1-r) \beta \left(\alpha + \frac{r(1-\alpha)}{C} \right) (1-q)q. \quad (11)$$

Setting $d\pi/dr = 0$, the optimal myopic referral is

$$r_M^* = \frac{1}{2} - \frac{C\alpha}{2(1-\alpha)}. \quad (12)$$

Proposition 5: The myopic referral is non-increasing with the fraction of nodes that are altruistic (α) and independent of both the request arrival rate β and the network size N .

The proof follows from (12) and imposing the bounds on the referral. Interestingly the myopic referral, unlike the optimal referral, is independent of both the request arrival rate β and the network size N . This is because the instantaneous profit depends on the instantaneous demand, the fraction that is fulfilled and the margin per sale. The referral only impacts the fraction of demand that is fulfilled and the margin per sale but not the instantaneous demand which is fixed for the purposes of computing the instantaneous sales. As a result, factors such as β and N that influence only the instantaneous demand but not demand fulfillment or margins are irrelevant to the computation of the myopic referral. In contrast, the optimal policy accounts for the fact that the referral impacts the demand generated in future periods and thus the optimal referral interacts with the demand terms.

In Figure 4, we plot the percentage increase in profit achieved by the optimal referral relative to the myopic referral against the fraction of altruistic nodes in the network (α) and the request rate (β). The remaining parameters are $N = 10000$, $T = 20$ and $C = 1$. When α is high or β is either too low or too high, the need to offer referrals diminishes (see Propositions 4 and 5). Hence it does not matter whether a myopic or optimal referral is offered. However for intermediate request rates and low levels of altruism, as observed in reality (Izal et al. 2004; Adar and Huberman 2000), the optimal referral significantly outperforms the myopic referral.

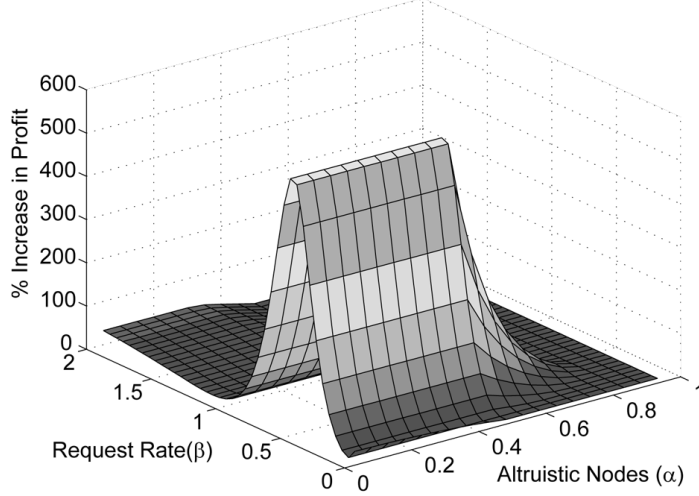


Figure 4: Profit Differential between Optimal and Myopic Referral

3.3 Diffusion in Hierarchical P2P Networks

A number of P2P networks are hierarchical, rather than completely decentralized, in order to reduce query flooding. Kazaa, a popular P2P network, has a two level structure where leaf nodes are organized into groups. A rough estimate from 2003 indicates that there were 10,000 groups in the Kazaa network with the super-node in each group handling 200 to 500 nodes (Ross and Rubenstein 2003).

Consider a 2 level hierarchical P2P network with M groups. As described in Section 3.1, the super-node of the seeking node's group will query all nodes within the group and also forward the request to $(l-1)$ neighboring groups. Let us first consider diffusion in Group-I:

$$\frac{dq_l}{dt} = \beta(1 - q_l)(1 - p_l p^{l-1}) = \beta(1 - q_l)(1 - (1 - \phi)^{n(q_l + (l-1)q)}). \quad (13)$$

In (13), $\beta(1 - q_l)$ is the generated demand and $(1 - p_l p^{l-1})$ is the fraction of this demand that is fulfilled (p_l and p represent the probability that there is no seed in Group-I and another group respectively).

Now consider diffusion in the remaining $(M - 1)$ groups. The generated demand in any group is $\beta(1 - q)$. The probability that there is a seed within a group is $(1 - p)$. When there is no seed within the group, two cases arise. The first case is one in which all $(l-1)$ requests are forwarded to groups other than Group-I. The probability of this case (i.e. of not querying Group-I) is

$$\binom{M-2}{l-1} \cdot \binom{M-1}{l-1}^{-1} = \frac{M-l}{M-1}. \quad (14)$$

The probability that one of these $(l-1)$ groups contains a seed is $(1-p^{l-1})$. The second case is one in which one of the $(l-1)$ requests goes to Group-I and the remaining $(l-2)$ requests are sent to other groups. The probability of querying Group-I is $1-(M-l)/(M-1)=(l-1)/(M-1)$. The probability that one of the groups contains a seed is $(1-p_l p^{l-2})$. Hence, the probability that generated demand is fulfilled is

$$(1-p) + p \left(\frac{M-l}{M-1} (1-p^{l-1}) + \frac{l-1}{M-1} (1-p_l p^{l-2}) \right). \quad (15)$$

Substituting the expression for p and p_l , we get the diffusion equation

$$\frac{dq}{dt} = \beta(1-q) \left(1 - (1-\phi)^{n(l-1)q} \left(\frac{M-l}{M-1} (1-\phi)^{nq} + \frac{l-1}{M-1} (1-\phi)^{nq_l} \right) \right). \quad (16)$$

We can jointly solve (16) and (13) to obtain the diffusion trajectory in each of the groups. The diffusion curve for the network as a whole is then obtained by aggregating the diffusion curves across all groups.⁹ While (13) and (16) can be solved numerically, there is no closed-form solution to these differential equations. The numerical solutions are discussed below.

Figure 5 plots the diffusion curves based on (13) and (16) for different values of M . The remaining parameters are $N = 10000$, $C = 1$, $\alpha = 0.1$, $\beta = 0.5$, $l = 2$, and $r = 0.2$. When the number of groups M is large the P2P network is similar to the decentralized network of Section 3.2. As M decreases, the fraction of requests reaching a seed increases significantly relative to a completely flat network and product diffusion occurs faster. Figure 6 plots the time to achieve 50% network penetration for the same parameters as in Figure 5. Initially, as we move from a flat structure to hierarchical structure, a small decrease in M will result in a significant reduction in diffusion time. However, as M reaches a sufficiently small value, further centralization will not necessarily help speed up the diffusion noticeably. The bottleneck is no longer the incomplete search of the network; rather, the bottlenecks are the rate of

⁹ Overall product penetration in the network is $((M-1)q + q_l)/M$. For large M , this can be approximated to q .

demand generation and the willingness of satisfied nodes to distribute the file. Given that the primary motivation for reducing the number of groups is to speed diffusion, there is little reason to prefer a completely (or nearly) centralized P2P architecture given other drawbacks of centralized structure such as the increased load on super-nodes. For example, Figures 5 and 6 suggest that there is little to gain by reducing groups below 50 for the given network configuration.

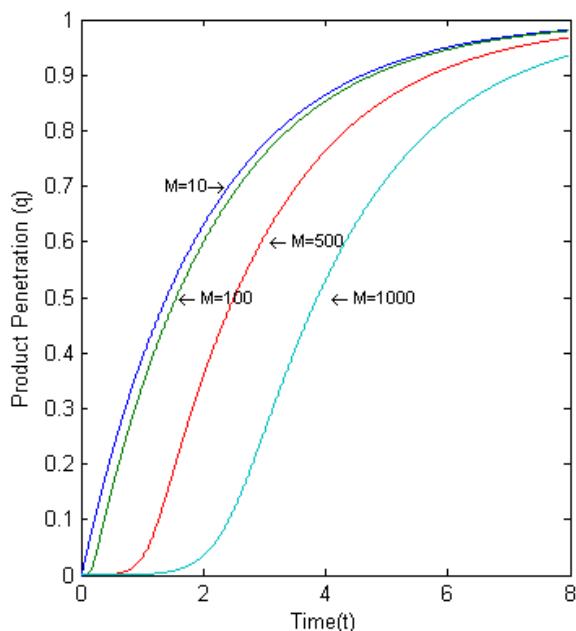


Figure 5: Diffusion process under varying levels of network centralization

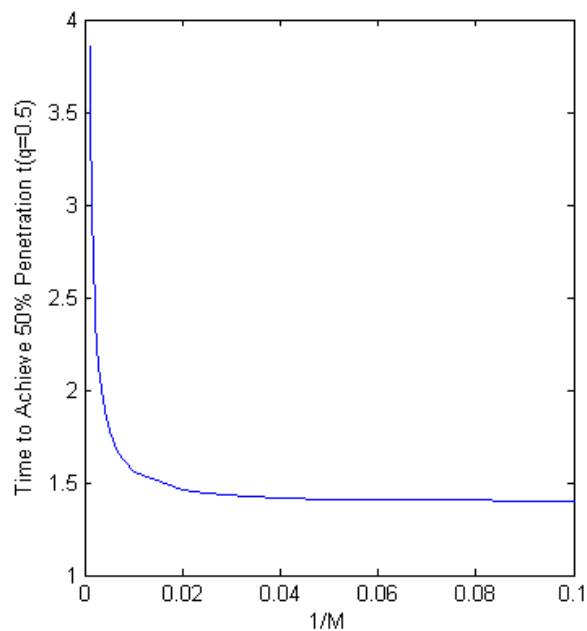


Figure 6: Impact of network centralization on time to achieve 50% network penetration

To illustrate the joint impact of incomplete search and free-riding, Figure 7 plots the time to achieve 50% network penetration against the number of groups in the network M and the distribution referral r . Other parameters are the same as in Figure 5. When both aspects of supply are active constraints, there is much to gain by increasing the referral and/or reducing the number of groups in the network. However if the referral is very high then almost all nodes distribute the file and the gains from reducing M are somewhat modest. Similarly, if the network is highly centralized, the referral may not be as crucial. This of course assumes that there are sufficient altruistic nodes in the network ($\alpha = 0.1$ in Figure 7). If most nodes in the network are not altruistic, then a referral will still be needed even with considerable centralization of the network. In summary, there are a number of interesting design parameters (M , r , l) available to managers and network designers that can be carefully tuned based on the diffusion model.

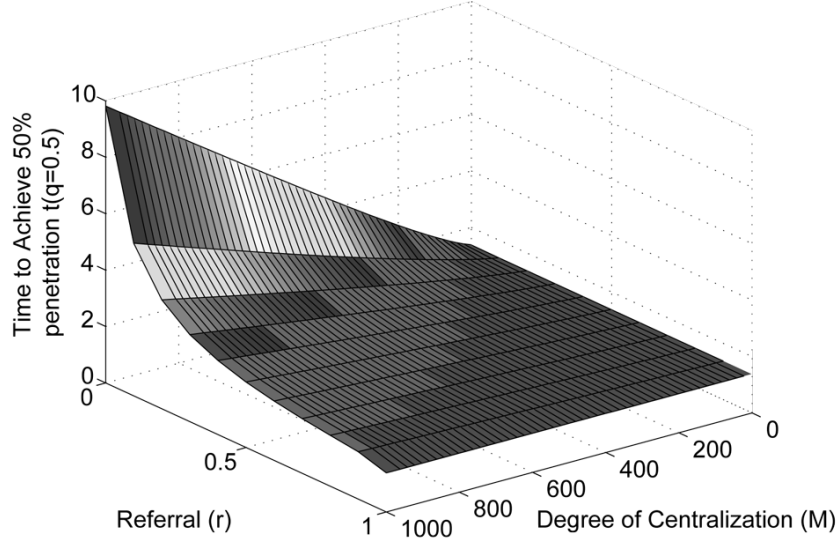


Figure 7: Impact of Incomplete Search and Free-riding on diffusion speed

3.3.1 Optimal Referral in Hierarchical Networks

In this Section, we investigate the optimal referral in the hierarchical network. Due to the complexity of the diffusion equation for a hierarchical network, we cannot derive a closed-form analytic expression for the optimal referral. We evaluate it numerically using (7) as the objective function and (13) and (16) as the diffusion equations. Figure 8 plots the optimal referral r^* against the number of groups in the network (M) and the number of requests forwarded (l). The other parameters are $N = 10000$, $C = 1$, $\alpha = 0.1$, $\beta = 0.5$. Increasing the number of requests helps reduce the magnitude of the referral. Further, increasing the degree of centralization of the P2P network (i.e. reducing M) reduces the need to offer high referrals (provided there are a reasonable number of altruistic nodes). At the same time, if there are very few altruistic nodes in the network then increasing the degree of centralization cannot entirely replace the need for referrals (Figure 9).

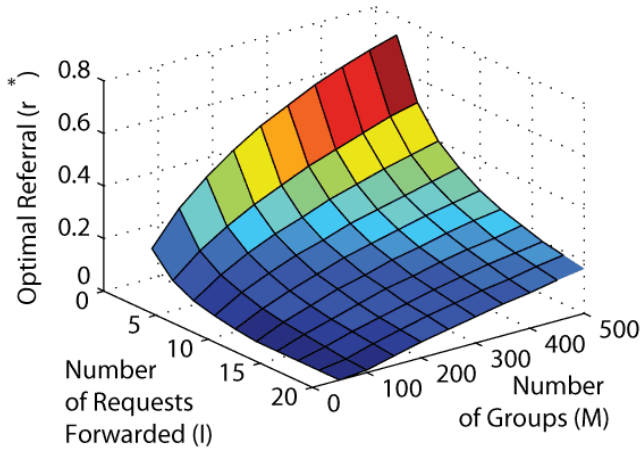


Figure 8: Impact of Search Process on Optimal Referral

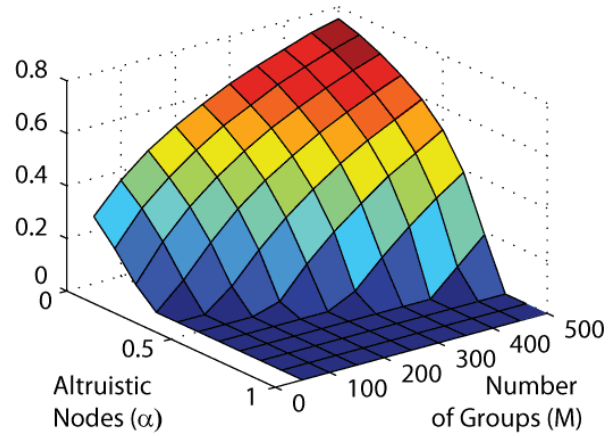


Figure 9: Impact of Altruism and Network Centralization on Optimal Referral

3.4 Simulations

Our diffusion models focus on supply-side factors and model a simple demand process based on mean field analysis. We now employ simulations to study the robustness of the models under more general demand processes. Specifically, the simulations incorporate heterogeneity by allowing β to vary across nodes. Further, given the mean request rate β for any node, the simulation models a Poisson request generation process for each node (recollect that the model assumed a deterministic process). Thus, the timing of requests varies across simulation runs. We then compare the average outcome over these runs with the predictions from the models.

The input to a simulation is the parameter set $\{N, \alpha, \beta, C, r\}$. Further $\{k\}$ or $\{l, M\}$ are additionally specified based on the type of network. First, N nodes are generated. A fraction (α) of nodes are randomly selected to be altruistic nodes. For the remaining nodes, the cost of being a seed (c) is drawn from a Uniform distribution in $[0, C]$. Before the diffusion begins, one of the nodes is randomly chosen as the initial satisfied node. To allow heterogeneity, the content seeking rate is uniformly distributed from 0.5β to 1.5β across the nodes. Given the mean seeking rate of a node, the request process is modeled as a Poisson distribution. That is, the time between successive attempts by a specific seeking node to locate content is drawn from an exponential distribution. Hierarchical networks are simulated by randomly forming M groups before the diffusion begins. Each group has a randomly selected super-node.

The diffusion process is simulated through a series of discrete steps. In the first step, the time until the first request is drawn for each node from the respective exponential distributions. The node with the lowest value is chosen as the first seeking node. The node is assigned several random neighbors at the time it is ready to request content. Requests are then forwarded to k randomly selected neighbors. If k exceeds the number of neighbors then these neighbors forward the request to their randomly assigned neighbors until k unique nodes are queried. If none of the requests reach a seed, the status of the seeking node is unchanged and the time before the node's next attempt to locate the content is drawn. If the queried node is a seed, the status of the querying node is changed to a satisfied node. This newly satisfied node will be a seed if it is altruistic or if its cost of distributing content is less than the referral. Next, the node with the lowest time to a request is selected. The process repeats until all nodes are satisfied or the time reaches an upper threshold. The simulation is repeated 100 times with the same parameters and the average value of q is recorded as the simulation result. The simulation for the hierarchical network is similar except that nodes are organized into groups and requests are always handled by super-nodes.

The predicted diffusion curve and the curve observed in the simulations (averaged over 100 runs) for a flat P2P network are plotted in Figure 10. The parameter configuration is $N = 10000$, $Q(0) = 1$, $C = 1$, $\alpha = 0.1$, $\beta = 0.5$, $k = 8$ and $r = 0.2$. The mean field model reasonably approximates the mean diffusion curve. Figure 11 indicates that the model fits the simulations for a hierarchical P2P network as well. The parameter settings are $N = 10000$, $M = 100$, $C = 1$, $\alpha = 0.1$, $\beta = 1$, and $r = 0.001$.

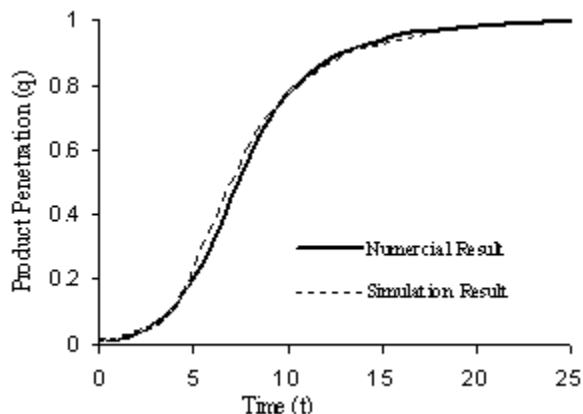


Figure 10: Simulation result for flat P2P with Multiple requests

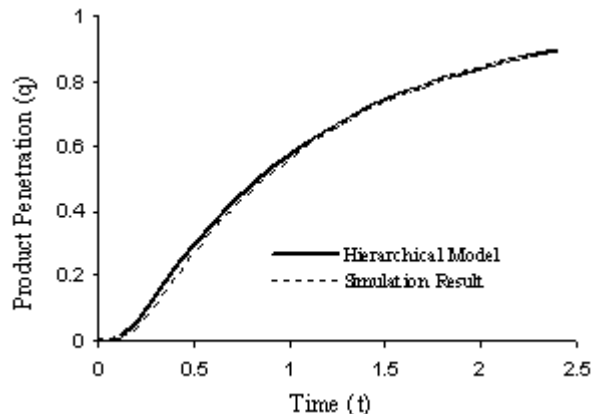


Figure 11: Simulation result for Hierarchical P2P Network

4. Conclusions and Future Directions

With the increasing use of P2P networks for distribution of digital products, modeling P2P product diffusion is of considerable interest. However, P2P diffusion demonstrates several unique characteristics not captured by traditional models. For example, generated demand is often not fulfilled immediately because of the decentralized distribution in P2P coupled with the incomplete search and free-riding by nodes. P2P media distributors are often most interested in understanding and designing these P2P-specific attributes such as P2P search architecture or use of distribution referrals to encourage file sharing. We developed models to capture the influence of the decentralized supply on product diffusion in P2P networks and demonstrated its application in determining optimal distribution referrals. Simulations suggest that the models are robust to heterogeneity across nodes and stochasticity in the request process.

There exist a number of interesting avenues for future work. On the modeling front, we develop a mixing model and do not model detailed spatial structure in a P2P network. An interesting direction to pursue is that of modeling network topology and the entry and exit of nodes in the network. The network models of Anderson (1998), Durett (1999), Eubank et al. (2004) and Ganesh et al. (2005) are highly applicable in this regard. Models that incorporate network structure can sometimes generate novel insights on the diffusion process (e.g. see Durett 1999). Given the spatial structure, it is interesting to ask what types of network topologies and search strategies are effective in locating seeds in the network. In this context, recent work on decentralized search in complex networks is relevant (Kleinberg 2006; Liben-Nowell et al. 2005). Vega-Redondo (2007) provides an excellent overview of the topic.

We assumed a monopoly setup in which unfulfilled demand returns in a future period. In reality, unfulfilled demand may be permanently lost to a competitor especially in the presence of a centralized media distributor such as iTunes. In addition, we assumed that the referral does not impact the demand process itself whereas it can impact the timing of purchases when participants are forward-looking. Modeling the loss of sales to competitors and the impact of distribution referrals on the demand process are likely to be particularly relevant to practice.

An interesting direction for future study would be in applying the models to address managerial questions tied to P2P media distribution. For example, how do prior results on optimal dynamic pricing, promotion, timing of product release, etc change under decentralized product supply in P2P? How should a firm "seed" a new product i.e., use free samples to support the distribution of a product whose availability is limited early in its diffusion? Similarly, a more detailed investigation of dynamic referral policies can be particularly useful for P2P managers. We are currently investigating this issue in ongoing work (Hosanagar et al. 2008). It would also be interesting to incorporate congestion costs to determine the optimal number of nodes to query in a decentralized network. Finally, a highly valuable extension would be the estimation of diffusion parameters using data from real P2P networks. Although it is ideal to estimate diffusion parameters from the closed-form solutions of diffusion equations, it may be necessary in our case to estimate parameters using discrete-time difference equations due to the complexity of the P2P diffusion models. Mahajan et al. (2000) provide an excellent overview of Maximum Likelihood Estimation (MLE) and Nonlinear Least Squares (NLS) based estimation techniques for diffusion models.

All of the above comments suggest that there are a number of open problems and further study on P2P product diffusion modeling can prove useful. Our research is only a first step in this direction.

References

- Adar, E. and B. A. Huberman. "Free Riding on Gnutella." *First Monday: Peer-Reviewed Journal on the Internet* 5-10 (2000). Available at http://www.firstmonday.dk/issues/issue5_10/adar/index.html.
- Andersson, H. Limit Theorems for a random Graph Epidemic Model, *The Annals of Applied Probability*, 1998.
- Andersson, H. and T. Britton. *Lecture Notes in Statistics: Stochastic Epidemic Models and Their statistical Analysis*. New York: Springer-Verlag, 2000.
- Arora, G., M. Hanneghan, and M. Merabti. "CasPaCE: A Framework for Cascading Payments in Peer-to-Peer Digital Content Exchange." *Proc. 4th PostGraduate Networking Conference*, Liverpool, UK, June 16-17, 2003. Available at <http://www.cms.livjm.ac.uk/pgnet2003/submissions/Paper-26.pdf>.
- Asvanund, A., K. Clay, R. Krishnan, and M. Smith. "An Empirical Analysis of Network Externalities in Peer-To-Peer Music Sharing Networks." *Information System Research* (15:2), June 2004.
- Bass, F. "A New Product Growth Model for Consumer Durables," *Management Science*, Vol. 15, January 1969, pp. 215-227.

- Bass, F. "The Relationship between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durable Technological Innovations." *Journal of Business*, 53 July 1980, 551-67.
- Brauer, F. and Castillo-Chavez. *Mathematical Models in Population Biology and Epidemiology*. New York: Springer, 2001.
- Currah, A. The Digital Storm: The Strategic Challenge of Internet Distribution to the Hollywood Studio System. Executive Report, University of Oxford, November 2004.
- Diekmann, O. and J.A.P. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases*. Wiley series in mathematical and computational biology. ISBN 0 471 98682 9, 2000.
- Durrett, R. Stochastic Spatial Models, *SIAM Review*, 1999.
- Eubank, S., H. Guclu, V.S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai and N. Wang, Modeling Disease Outbreaks in Realistic Urban Social Networks, *Nature*, 429, pp. May 2004.
- Fourt, L.A, Woodlock, J.W, "Early prediction of market success for grocery products", *Journal of Marketing*, Vol. 25 pp.31-38, 1960.
- Ganesh, A., L. Massoulie, D. Towsley, The Effect of Network Topology on the Spread of Epidemics, *IEEE Infocom 2005*, Miami, FL, 2005.
- Golle, P., K. Leyton-Brown, I. Mironov, and M. Lillibridge. "Incentives for sharing in peer-to-peer networks," *Proceedings of the 3rd ACM conference on Electronic Commerce*, Tampa, Florida, 2001.
- Ho, T., S. Savin, and C. Terwiesch, "Managing Demand and Sales Dynamics in Constrained New Product Diffusion under Supply Constraint", *Management Science*, Vol. 48, Number 2, 2002.
- Hosanagar, K., P. Han, and Y. Tan. "Optimal Dynamic Referrals in P2P Networks," Working paper, University of Pennsylvania, 2007.
- Horsky, D. and L. S. Simon. "Advertising and the diffusion of new products." *Marketing Sci.* 2 (1), 1983.
- A. Iamnitchi, M. Ripeanu, and I. Foster. Small-world filesharing communities. *Proc. Of IEEE Infocom*, 2004.
- Izal, M., G. Urvoy-Keller, E.W. Biersack, P.A. Felber, A. Al Hamra, and L. Garc'es-Erice. "Dissecting BitTorrent: Five Months in a Torrent's Lifetime." *Proceedings of Passive & Active Measurement Workshop 2004*, Antibes Juan-les-Pins, France, April 19-20, 2004.
- Jain, D., V. Mahajan, E. Muller. Innovation diffusion in presence of supply restrictions. *Marketing Science*, 10(1) 100-113, 1991.
- Kamvar, S. D., M. T. Schlosser, and H. Garcia-Molina. "Incentives for Combatting Freeriding on P2P Networks", In *Euro-Par 2003*. June, 2003.
- Kephart, J. O. and S. R. White. "Directed-Graph Epidemiological Models of Computer Viruses." *IEEE Computer Society Symposium on Research in Security and Privacy*, Oakland, California, May 20-22, 1991, pp. 343-359.

- Khelil, A., C. Becker, J. Tian, and K. Rothermel. "An Epidemic Model for Information Diffusion in MANETs." *The Fifth ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Atlanta, GA, September 28, 2002, pp. 54-60.
- Kleinberg, J. Complex Networks and Decentralized Search Algorithms. *Proceedings of the International Congress of Mathematicians*, 2006.
- Krishnan, T.V., F.M. Bass, V. Kumar. "Impact of a late entrant on the diffusion of a new product/service." *J. Marketing Res.* 37, May 2000, 269–278.
- Kumar, S., and J. Swaminathan: "Diffusion of Innovations under Supply Constraints," *Operations Research*, 51(6), pp. 866-879, 2003.
- Lang, K., and R. Vragov, "A Pricing Mechanism for Digital Content Distribution over Peer-to-Peer Networks", *Journal of Management Information Systems*, 22 (2), Fall 2005.
- Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic Routing in Social Networks, *Proceedings of the National Academy of Sciences*, Vol. 102, No. 33, 2005.
- Mahajan, V., E. Muller and Y. Wind. *New-Product Diffusion Models*. Kluwer Academic Press, 2000.
- Newman, M. E. J., C. Moore, and D. J. Watts. "Mean-field Solution of the Small-world Network Model." *Physical Review Letters*, vol. 84, no. 14, pp. 3201-3204, 2000.
- New York Times. E-Commerce Report: Incentive marketing spreads on the Internet, with offers of discounts or credit toward gifts. June 2003.
- Opper, M. and D. Saad. *Advanced Mean Field Methods - Theory and Practice*. MIT Press, 2001.
- Padmanabhan, V. N. and K. Sripanidkulchai. "The Case for Cooperative Networking." Peter Druschel, M. Frans Kaashoek, Antony I. T. Rowstron (Eds.): *Peer-to-Peer Systems*, First International Workshop, IPTPS 2002, Cambridge, MA, March 7-8, 2002, pp. 178-190.
- Ross, K.W. and D. Rubenstein. "Tutorial on P2P Systems." *The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*, San Francisco, CA, March 30-April 3, 2003.
- Techcrunch. P2P Music Sharing Service Grooveshark Ups Compensation. October 2007.
- Vega-Redondo, F. *Complex Social Networks*. Econometric Society Monograph Series, Cambridge University Press, 2007.

Appendix A: Diffusion in Flat P2P Network

Proposition 1: The product penetration $q(t_c)$ at a given time instant t_c increases with k in a convex fashion early in the file's diffusion (small q) and in a concave fashion late in the diffusion (large q). The time $t(q_c)$ to achieve a given level of product penetration q_c decreases with k in a convex fashion.

Proof: Suppose we fix $t = t_c$. Given the number of requests forwarded (k), let the value of q at this instant

be denoted $q = q(k)$. The diffusion equation can be rewritten as $\int_{q_0}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^k)} = \beta t_c$. Now

suppose we instead had $(k+1)$ requests forwarded, then $\int_{q_0}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^{k+1})} = \beta t_c$. Therefore,

$$\int_{q_0}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^k)} - \int_{q_0}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^{k+1})} = 0. \text{ Rewriting it,}$$

$$\int_{q(k)}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^k)} = \int_{q_0}^{q(k+1)} \frac{(1-\phi x)^k \phi x dx}{(1-x)(1-(1-\phi x)^{k+1})(1-(1-\phi x)^k)}. \quad (\text{A1})$$

Since the second term is greater than zero, we can conclude that $q(k+1) > q(k)$. Thus, an increase in the number of nodes queried serves to increase the product penetration.

Now let us evaluate the nature of the increasing relationship. Just as with (A1), we can derive,

$$\int_{q(k-1)}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^k)} = \int_{q_0}^{q(k-1)} \frac{(1-\phi x)^{k-1} \phi x dx}{(1-x)(1-(1-\phi x)^k)(1-(1-\phi x)^{k-1})}. \quad (\text{A2})$$

Taking the difference between (A1) and (A2), we have

$$\int_{q(k)}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^k)} - \int_{q(k-1)}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^k)}$$

$$= \int_{q(k-1)}^{q(k+1)} \frac{(1-\phi x)^k \phi x dx}{(1-x)(1-(1-\phi x)^{k+1})(1-(1-\phi x)^k)} + \int_{q_0}^{q(k-1)} \frac{(1-\phi x)^{k-1} (\phi x)^2 dx}{(1-x)(1-(1-\phi x)^{k+1})(1-(1-\phi x)^{k-1})} > 0$$

Therefore:

$$\int_{q(k)}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^k)} > \int_{q(k-1)}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^k)}. \quad (\text{A3})$$

The relationship between $q(k+1)$, $q(k)$, and $q(k-1)$ depends on the function

$\psi_k(q) = \int_q \frac{dx}{(1-x)(1-(1-\phi x)^k)}$, which is increasing in q . If $\psi_k(q)$ is concave, then for inequality (A3) to

hold, we require that $q(k+1) + q(k-1) - 2q(k) > 0$. That is, we require $q(k)$ to be convex. The second derivative of $\psi_k(q)$ is negative when q is small. Thus, $q(k)$ is convex with k for small q .

To prove that $q(k)$ is concave for large q , we rewrite (A1) and (A2) as:

$$\int_{q(k)}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^{k+1})} = \int_{q_0}^{q(k)} \frac{(1-\phi x)^k \phi x dx}{(1-x)(1-(1-\phi x)^{k+1})(1-(1-\phi x)^k)};$$

$$\int_{q(k-1)}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^{k-1})} = \int_{q_0}^{q(k)} \frac{(1-\phi x)^{k-1} \phi x dx}{(1-x)(1-(1-\phi x)^k)(1-(1-\phi x)^{k-1})}.$$

Taking the difference between the above two expressions just as we did with (A1) and (A2) to get (A3),

$$\int_{q(k)}^{q(k+1)} \frac{dx}{(1-x)(1-(1-\phi x)^{k-1})} < \int_{q(k-1)}^{q(k)} \frac{dx}{(1-x)(1-(1-\phi x)^{k-1})}. \quad (\text{A4})$$

Similarly, $\psi_{k-1}(q) = \int_q \frac{dx}{(1-x)(1-(1-\phi x)^{k-1})}$ is increasing in q . If $\psi_{k-1}(q)$ is convex, we require

$q(k+1) + q(k-1) - 2q(k) < 0$ to make inequality (A4) hold. That is, we require $q(k)$ to be concave. The second derivative of $\psi_{k-1}(q)$ is positive for large q . It thus follows that $q(k)$ is concave with k for large q .

Now, let us evaluate the impact of k on the time to achieve a given level of product penetration.

Suppose we fix $q = q_c$. Denote the time taken to achieve this level of product penetration by $t(k)$. Then,

$$t(k+1) - t(k) = -\frac{1}{\beta} \int_{q_0}^{q_c} \frac{\phi x (1-\phi x)^k dx}{(1-x)(1-(1-\phi x)^k)(1-(1-\phi x)^{k+1})} < 0.$$

The second difference is:

$$t(k+1) + t(k-1) - 2t(k) = \frac{1}{\beta} \int_{q_0}^{q_c} \frac{(\phi x)^2 (1-\phi x)^{k-1} (1+(1-\phi x)^k) dx}{(1-x)(1-(1-\phi x)^{k-1})(1-(1-\phi x)^k)(1-(1-\phi x)^{k+1})} > 0.$$

Proposition 2: The product penetration $q(t_c)$ at a given time instant t_c increases with the referral r . Simultaneously, the time $t(q_c)$ to achieve a given level of product penetration q_c decreases with r in a convex fashion.

Proof: Recall that $\int_{q_0}^q \frac{dx}{(1-x)(1-(1-\phi x)^k)} = \beta t$. Suppose we fix $t = t_c$. We examine how q varies with

respect to r , we have

$$\frac{dq}{dr} \frac{1}{(1-q)(1-(1-\phi q)^k)} - \frac{k(1-\alpha)}{C} \int_{q_0}^q \frac{x(1-\phi x)^{k-1} dx}{(1-x)(1-(1-\phi x)^k)^2} = 0,$$

or,
$$\frac{dq}{dr} = (1-q)(1-(1-\phi q)^k) \frac{k(1-\alpha)}{C} \int_{q_0}^q \frac{x(1-\phi x)^{k-1} dx}{(1-x)(1-(1-\phi x)^k)^2} > 0.$$

Now, let us evaluate the impact of r on the time to achieve a given level of product penetration.

Suppose we fix $q = q_c$. Taking the derivative of t with respect to r , we get

$$\frac{dt}{dr} = -\frac{k(1-\alpha)}{\beta C} \int_{q_0}^{q_c} \frac{x(1-\phi x)^{k-1} dx}{(1-x)(1-(1-\phi x)^k)^2} < 0, \text{ and}$$

$$\frac{d^2 t}{dr^2} = \frac{k(1-\alpha)^2}{\beta C^2} \int_{q_0}^{q_c} \frac{x^2(1-\phi x)^{k-2} (k-1+(k+1)(1-\phi x)^k) dx}{(1-x)(1-(1-\phi x)^k)^3} > 0.$$

Proposition 3: The optimal referral for the flat P2P network with diffusion specified by (5) is

$$r^* = \max \left[0, \min \left[C, 1 - \frac{C}{\beta T(1-\alpha)} \left(1 + W \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right) \right) \right] \right],$$

where $W(x)$ is the Lambert W-function (solution of $W(x)\exp(W(x)) = x$).

Proof: Setting the first order condition specified in (8) equal to zero, we get

$$(N-1) \exp \left(-\beta T \left(\alpha + (1-\alpha) \frac{r}{C} \right) \right) \left(\beta T(1-\alpha) \frac{1-r}{C} - 1 \right) = 1. \text{ Rearranging this condition, we have}$$

$$\left(\frac{(1-r)\beta T(1-\alpha)}{C} - 1 \right) \exp \left(\frac{(1-r)\beta T(1-\alpha)}{C} - 1 \right) = \frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right).$$

This equation is of the form $W(x)\exp(W(x)) = x$, where

$$W \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right) = \frac{(1-r)\beta T(1-\alpha)}{C} - 1. \quad (\text{A5})$$

Given the Lambert W-function $W()$, the solution to (A5) is:

$$r^* = 1 - \frac{C}{\beta T(1-\alpha)} \left(1 + W \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right) \right).$$

The second derivative at $r = r^*$ is

$$-(N-1) \exp \left(-\beta T \left(\alpha + \frac{1-\alpha}{C} r^* \right) \right) (1-r^*) \left(\beta T \frac{1-\alpha}{C} \right)^2 \left/ \left(1 + (N-1) \exp \left(-\beta T \left(\alpha + \frac{1-\alpha}{C} r^* \right) \right) \right)^2 < 0,$$

since $r^* < 1$. Therefore, the function is locally concave and r^* is a local maximum. However, the objective function is convex for small r and concave for large r . We can nonetheless show that the objective function is unimodal, hence, $r = r^*$ is a global maximum. To see this, consider the first derivative with respect to r . The denominator is positive, so the sign is determined by the numerator:

$$Z(r) = -1 + (N-1) \exp \left(-\beta T \left(\alpha + \frac{1-\alpha}{C} r \right) \right) \left((1-r) \beta T \frac{1-\alpha}{C} - 1 \right).$$

Suppose that $Z(0) > 0$. This is the condition under which a non-zero optimal r exists. Then

$$Z'(r) = -(N-1) \exp \left(-\beta T \left(\alpha + \frac{1-\alpha}{C} r \right) \right) (1-r) \left(\beta T \frac{1-\alpha}{C} \right)^2$$

which is negative if $r < 1$ and positive if $r > 1$. Therefore $Z(r)$, starting from a positive value at $r = 0$, decreases with r , crosses 0 at $r = r^* < 1$, and eventually increases with r for $r > 1$. However, for $r > 1$, $Z(r) < 0$. So, $Z(r) \geq 0$ if $r \leq r^*$, and $Z(r) < 0$ if $r > r^*$. This shows that the objective function is unimodal.

Proposition 4: The optimal referral is (a) non-increasing with the fraction of nodes in the network that are altruistic (α), (b) non-decreasing with the P2P network's size (N), and (c) non-decreasing with request rate β for $\beta < \beta_{th}$ and non-increasing with β for $\beta > \beta_{th}$ where

$$\beta_{th} = \frac{1}{T} \left(2 + W \left((N-1)e^{-1} \right) + \frac{e}{N-1} \exp \left(W \left((N-1)e^{-1} \right) \right) \right) \left(\alpha + \frac{1-\alpha}{C} \right)^{-1}.$$

Proof: From (8), we have $\beta T(1-\alpha) \frac{1-r}{C} - 1 = \frac{1}{(N-1) \exp(-\beta T(\alpha + (1-\alpha)r/C))} > 0$. Thus,

$\beta T(1-\alpha)(1-r^*)/C > 1$. It then follows that $1-r^* > 0$. We will use this result below.

$$\begin{aligned}\frac{\partial r^*}{\partial \alpha} &= -\frac{C}{\beta T(1-\alpha)^2} \left(1 + W \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right) \right) \\ &\quad - \frac{C}{\beta T(1-\alpha)} W' \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right) \frac{\beta T(1-1/C)}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right).\end{aligned}$$

Using (8) and the equation $W'(x) = W(x)/((1+W(x))x)$, we can rewrite the above expression:

$$\frac{\partial r^*}{\partial \alpha} = -\frac{C}{1-\alpha} \left(\frac{1-r^*}{C} + 1 - \frac{1}{C} - \frac{C(1-1/C)}{\beta T(1-r^*)(1-\alpha)} \right) < -\frac{C}{1-\alpha} \left(\frac{1-r^*}{C} + 1 - \frac{1}{C} - 1 + \frac{1}{C} \right) = -\frac{1-r^*}{1-\alpha} < 0.$$

(b): W is an increasing function of N . Because r^* is increasing in W , it follows that r^* increases with N .

(c): We have
$$\frac{\partial r^*}{\partial(\beta T)} = \frac{C}{(\beta T)^2(1-\alpha)}(1+w) - \frac{C}{\beta T(1-\alpha)} w' \frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \left(\alpha + \frac{1-\alpha}{C} \right),$$

where $w = W \left(\frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right) \right)$ and $w' = dW(x)/dx$. We apply the following two

properties of the Lambert-W function:

$$we^w = \frac{1}{N-1} \exp \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) - 1 \right), \quad (\text{A6})$$

and $(1+w)e^w w' = 1$. Using these properties, the derivative can be rewritten as follows:

$$\frac{\partial r^*}{\partial(\beta T)} = \frac{C}{(\beta T)^2(1-\alpha)}(1+w) - \frac{C}{\beta T(1-\alpha)} \frac{w}{1+w} \left(\alpha + \frac{1-\alpha}{C} \right).$$

Hence, the sign of the above expression is determined by the expression:

$$(1+w)^2 - w \left(\beta T \left(\alpha + \frac{1-\alpha}{C} \right) \right). \quad (\text{A7})$$

Using (A6), (A7) can be rewritten as $1+w - w \ln((N-1)w)$, which is positive for small w and negative

for $w > w_{th}$, where $w_{th} = \frac{e}{N-1} \exp(W((N-1)e^{-1}))$. It follows that r^* increases when $\beta < \beta_{th}$ and

decreases for $\beta > \beta_{th}$, where $\beta_{th} = \frac{1}{T} \left(2 + W((N-1)e^{-1}) + \frac{e}{N-1} \exp(W((N-1)e^{-1})) \right) \left(\alpha + \frac{1-\alpha}{C} \right)^{-1}$.

Imposing $r \in [0, C]$, it follows that r^* is non-decreasing for $\beta < \beta_{th}$ and non-increasing for $\beta > \beta_{th}$.