Statistical Modelling of Information Sharing: Community, Membership and Content

W.-Y. Ng, W.K. Lin, D.M. Chiu
Department of Information Engineering
The Chinese University of Hong Kong
{wyng,wklin3,dmchiu}@ie.cuhk.edu.hk

June 29, 2005

Abstract

File-sharing systems, like many online and traditional information sharing communities (e.g. newsgroups, BBS, forums, interest clubs), are dynamical systems in nature. As peers get in and out of the system, the information content made available by the prevailing membership varies continually in amount as well as composition, which in turn affects all peers' join/leave decisions. As a result, the dynamics of membership and information content are strongly coupled, suggesting interesting issues about growth, sustenance and stability.

In this paper, we propose to study such communities with a simple statistical model of an *information sharing club*. Carrying their private payloads of information goods as potential supply to the club, peers join or leave on the basis of whether the information they demand is currently available. Information goods are chunked and *typed*, as in a file sharing system where peers contribute different files, or a forum where messages are grouped by topics or threads. Peers' demand and supply are then characterized by statistical distributions over the type domain.

This model reveals interesting critical behaviour with multiple equilibria. A sharp growth threshold is derived: the club may grow towards a sustainable equilibrium only if the value of an *control parameter* is above the threshold, or shrink to emptiness otherwise. The control parameter is composite and comprises the peer population size, the level of their contributed supply, the club's efficiency in information search, the spread of supply and demand over the type domain, as well as the goodness of match between them.

1 Introduction

The notion of a peer-to-peer system means different things to different people. To some, it is a way to use the commodity personal computers to do the job of large and expensive servers [1], [2], [3]. Others build application layer multicast systems out of it [4], [5]. But there is one thing in common for almost all peer-to-peer systems, that is the coming-together of such a system depends on the number of peers (large or small) wanting to participate.

The formation of such cooperation without central management has its inherent advantages: it saves the cost of central management; and more importantly, it automatically adapts to the need (for example, in terms of time and scope) of the peers who collectively form the club. But what are the forces that attract peers together? What would cause a peer-to-peer system to grow, sustain itself, or fall apart? Are there some fundamental reasons that apply to different peer-to-peer systems?

The purpose of the peer-to-peer systems is invariably to share some resources or information. Economists differentiate between two kinds of goods that are shared: rivalrous and non-rivalrous goods. The former diminishes when shared. Compute power, storage and communication bandwidth are examples of rivalrous goods. Many information goods, however, are inherently non-rivalrous. In other words, they can be readily replicated many times with little or no cost.

Motivated by the above questions, we formulate a model for a club where members share non-rivalrous information goods. Conventional resources such as computers and bandwidth are assumed to be abundant. In such a setting, the strength of a club is determined by the amount as well as the *composition* of the information content made available by the club's prevailing membership, and how that content fits the potential members demands. Based on this simple model, it is then possible to derive some very basic conditions for a club to form and sustain. The model predicts a critical population size, from which enough peers will find matching interest and form a club. Furthermore, the model can be used to understand the dynamics of the content and membership, and whether and how it leads to an equilibrium.

Since the model is simple, it is also general enough to be applied to many other information sharing paradigms. Examples include web-based collaborative environments, newsgroups where peers contribute their opinions about different topics, or other forums or communities for information sharing.

The rest of the paper is organized as follows. Section 2 is devoted to modeling the peers in terms of their demand and contribution, which then leads to a model of a club in terms of its content. Section 3 models peers' decisions of joining or leaving a club, and consequently the conditions for club formation, and other equilibrium properties of it. Section 4 illustrates the properties of the model through numerical examples. Section 5 discusses the contribution of this model, the interpretation of various results, as well as the limitations of our model. Finally, we conclude and discuss future directions.

1.1 Related Works

Many other papers tried to model incentives in peer-to-peer systems and the resulting club dynamics. [6] discusses private versus public goods, and argues that messages shared in web forums are private goods, thus suggesting sharing is not simply an altruistic behavior. Several papers focus on how to relieve the cost/congestion of some rivalrous resources, such as bandwidth and other resources that a peer has to consume. For example [7] suggests a possible rationale for peers' contributions is to relieve the bandwidth stress when they share, their actions thereby benefit the peers themselves. [8], [9], [10] use game theoretic approaches to model and understand the sharing incentives in peer-to-peer networks. These works also discuss incentive-compatible solutions to peer-to-peer systems. In comparison, our model brings out a new angle that is complementary and somewhat orthogonal to the above works.

Our work is in part motivated by [11] in which a general model is used to explain the vitality of a peer-to-peer network when different types of peers are involved. The type of a peer is characterized by the peer's generosity, which is used as a threshold to determine when a peer would contribute to the club rather than free-ride. Their model does not explicitly capture different types of information goods themselves, therefore the motivation for sharing remains rather abstract.

Our work attempts to explain the motivation of the peers by characterizing the different types of peers based on their contribution and demand of different types of information goods. A peer's decision to join a club can then be related to the extent the club can satisfy the peer's interest (demand). This sheds more (at least different) insights to what brings peers together in the first place.

2 The Information Sharing Club (ISC) Model

The Information Sharing Club (ISC) model has three basic components. First, a population of N *peers*, denoted by $\mathcal{N}$, may freely join or leave the club any time at their own will. Each peer carries a payload of information goods which are shared with other current members only when he joins the club.

Second, information goods are chunked and *typed*, the same way that versions of different files are served in a file sharing system, or messages of various topics are hosted in a forum. Information chunks of the same type are not differentiated: an instance of information demand specifies the chunk type only and is satisfied by *any* chunk of that type, as when request for a file is satisfied with any copy of it, or when information query returns any piece of information of the specified class (e.g. as implied by the query criteria, for instance).

Third, the club maintains a platform on which information chunks shared by members are maintained and searched. A perfect membership system makes sure that only requests by current members are processed. A request may comprise one or more instances of demand, and is successfully served when all instances are satisfied. However, the search may not be perfect and is conducted with efficiency $\rho \in (0, 1]$, defined as the probability that any shared chunk is actually found in time by the platform in response to a request.

We make probabilistic assumptions about both demand and supply: peer i's demand instances as well as the content of his private payload, in terms of chunk types, are drawn from statistical distributions. Specifically, we assume peer i's private payload comprises $K_i \geq 0$ chunks drawn from distribution $g_i(s)$, $s \in \mathcal{S} \triangleq \{1, 2, \ldots\}$ where $\mathcal{S}$ is the set of all types. The total payload of any group of members (*membership*) $\mathcal{G} \subset \mathcal{N}$ is then given by

$$g_{\mathcal{G}}(s) \triangleq \frac{\sum_{i \in \mathcal{G}} K_i \; g_i(s)}{\sum_{i \in \mathcal{G}} K_i}, \quad \mathcal{G} \subset \mathcal{N}$$

Without loss of generality, we assume the *aggregate supply function* $g(s) \triangleq g_{\mathcal{N}}(s)$ to be monotonically non-increasing. The type variable s may then be interpreted as a *supply rank (s-rank)*. In other words, $s = 1$ and $s = |\mathcal{S}|$ denote the most and least supplied chunk types respectively.

Likewise, we define the *aggregate demand function* $h(s) \triangleq h_{\mathcal{N}}(s)$ where

$$h_{\mathcal{G}}(s) \triangleq \frac{\sum_{i \in \mathcal{G}} M_i \; h_i(s)}{\sum_{i \in \mathcal{G}} M_i}, \;\; \mathcal{G} \subset \mathcal{N}$$

as peer i generates demand instances at a rate of M_i chunks per unit time, drawn from distribution $h_i(s)$, $s \in \mathcal{S}$[1].

For current club membership $\mathcal{C}$, the expected number of chunks of type s being shared would be given by $\mu_{\mathcal{C}}(s) \triangleq n \; k_{\mathcal{C}} \; g_{\mathcal{C}}(s)$ where $n \triangleq |\mathcal{C}|$ is the membership size and $k_{\mathcal{C}} \triangleq \sum_{i \in \mathcal{C}} K_i/|\mathcal{C}|$ is the payload size averaged over the current club membership. Conditioning on the membership size, we have

$$\mu_n(s) = n \; k \; g(s)$$

where $k \triangleq \sum_{i=1}^{N} K_i/N \; > \; 0$ is the payload size averaged over all peers. We assume further that members' contents are drawn independently, which implies a Poisson distribution for the actual total number of type s chunks being shared. Subsequently demand instances for chunk type s have an average failure rate of $e^{-\mu_n(s) \; \rho} \; = \; e^{-n \; k \; g(s) \; \rho}$. The average success rate of peer i's demand being satisfied in a club of size n is therefore

$$p_i(n) \;\; \triangleq E_{h_i(s)}[1 - e^{-n \; k \; g(s) \; \rho}] \tag{1}$$

where $E[\cdot]$ is the expectation operator. This is compatible with the non-rivalrous assumption as it is independent of the level of demand for this chunk type.

2.1 An example: music information sharing club

Tables 1 and 2 depict an example of six peers sharing music information of five different types. For simplicity, we assume identical payload sizes (identical K_i's) and demand rates (identical M_i's) so

[1] Another possible ranking of the types is *popularity rank (p-rank)*, which ranks the types according to the aggregate demand instead. In cases when the p-rank is more natural to work with, such as when supply is being driven by demand and p-ranks are more readily known, we may derive the requisite demand functions in s-rank as

$$h_i(s) \triangleq \sum_r \frac{\phi(r, s)}{f(r)} \; f_i(r)$$

where $f_i(r)$ is peer i's demand distribution over the p-rank domain and $\phi(r, s)$ is the joint distribution of the two rank measures that captures how well supply follows demand. (Perfect following would imply $\phi(r, s) = 0 \; \forall r \neq s$.)

that the aggregate distributions are simple unweighted averages of the peers' distributions. Table 3 gives the resulting s-ranks and p-ranks of the five music types. The information may be news and messages about the different music types when the club is a discussion forum in nature, or musical audio files when it is a file sharing platform.

Table 1: Distributions of peers' private payloads, $g_i(s)$

	Pop	Classical	Oldies	World	Alternative
Alfred	0.4	0.3	0.1	0.1	0.1
Bob	0.4	0.2	0.2	0.15	0.05
Connie	0.3	0.3	0.2	0.1	0.1
David	0.2	0.3	0.3	0.15	0.05
Eric	0.5	0.05	0.2	0.15	0.1
Florence	0.1	0.4	0.1	0.1	0.3
aggregate supply, g(s)	0.317	0.258	0.18	0.125	0.12

Table 2: Distributions of peers demand, $h_i(s)$

	Pop	Classical	Oldies	World	Alternative
Alfred	0.1	0.4	0.3	0.1	0.1
Bob	0.05	0.5	0.1	0.3	0.05
Connie	0.1	0.2	0.3	0.2	0.2
David	0.1	0.4	0.3	0.15	0.05
Eric	0.1	0.4	0.2	0.2	0.1
Florence	0.2	0.3	0.1	0.2	0.2
aggregate demand, h(s)	0.108	0.367	0.217	0.192	0.117

Table 3: The supply and the popularity rank

	1	2	3	4	5
Supply rank (s)	Pop	Classical	Oldies	World	Alternative
Popularity rank (r)	Classical	Oldies	World	Alternative	Pop

A peer's success rate would depend on the types of goods he demands on one hand, viz. $h_i(s)$, and the aggregate supply $g(s)$ on the other. For instance, Alfred's average success rate is given by:

$$p_{\text{Alfred}} = 1 - (0.1\,(e^{-6\,(0.317)}) + 0.4\,(e^{-6\,(0.258)}) + \ldots + 0.1\,(e^{-6\,(0.12)})) = 0.69$$

3 Dynamic equilibrium of membership and content

Generally speaking, peers would join the club as members and share their private payloads as long as their requests are sufficiently met. We make two simplifying assumptions here: (1) a peer would join as long as a single current request is met, and leave otherwise; and (2) any request comprises $d \geq 1$ instances of demand. The probability that peer i would join when membership is $\mathcal{C}$ is then $P_{\mathcal{C},i} \triangleq p_{\mathcal{C},i}^d$ where $p_{\mathcal{C},i}$ is the probability that an instance of peer i's demand is satisfied when membership is $\mathcal{C}$. Conditioning on the membership size n, the expected joining probability of peer i is

$$P_i(n) \triangleq P_i(n)^d \tag{2}$$

Membership dynamics and content dynamics are closely coupled: as peers join and leave, they alter the total shared content, inducing others to revise their join/leave decisions. The membership size changes always unless the two-way flows between members and non-members are balanced.

Consequently, we may define a *statistical equilibrium membership size* n_{eq} as the solution of the balance condition

$$\begin{array}{rrcl} & (N - n_{eq})\bar{P}(n_{eq}) & = & n_{eq}(1 - \bar{P}(n_{eq})) \\ \Leftrightarrow & \bar{P}(n_{eq}) & = & \dfrac{n_{eq}}{N} \end{array} \tag{3}$$

where $\bar{P}(n) = \frac{1}{N}\sum_{i=1}^{N} P_i(n)$ is the joining probability averaged over all peers and all possible memberships of size n. Note that equation (3) is in the form of a fixed point equation which is indicative of the coupled dynamics of membership and content. Further, the stability condition for a fixed point n_{eq} is simply

$$\left.\frac{\partial \bar{P}(n)}{\partial n}\right|_{n=n_{eq}} < \frac{1}{N} \tag{4}$$

Note that an empty membership is always a fixed point, and would always be stable for sufficiently small N, in which case autonomous growth from an empty or small membership is very difficult if not impossible.

Theorem 3.1 (Empty Membership Instability). *Empty membership is unstable if and only if requests are simple, viz.* $d = 1$, *and*

$$\pi \triangleq N\, k\, \rho \sum_s h(s)\; g(s) \geq 1 \quad . \tag{5}$$

Proof. Consider:

$$\bar{P}(n) \;=\; \frac{1}{N}\sum_{i=1}^{N} P_i(n) \;=\; \frac{1}{N}\sum_{i=1}^{N} p_i(n)^d \qquad d \geq 1 \;\; .$$

Differentiating with respect to n:

$$\begin{array}{rrcl} & \dfrac{N}{d}\dfrac{\partial\bar{P}(n)}{\partial n} & = & \displaystyle\sum_{i=1}^{N} p_i(n)^{d-1}\frac{\partial p_i(n)}{\partial n} \\ \Leftrightarrow & \dfrac{N}{dk\rho}\dfrac{\partial\bar{P}(n)}{\partial n} & = & \displaystyle\sum_{i=1}^{N} p_i(n)^{d-1} E_{h_i(s)}[e^{-nk\rho g(s)}g(s)] \end{array}$$

Since $p_i(0) = 0$, it follows that $\partial\bar{P}(n)/\partial n\,\Big|_{n=0} = 0$ for $d > 1$, in which case an empty membership is always stable. When $d = 1$,

$$\begin{array}{rrcl} & \dfrac{N}{k\rho}\dfrac{\partial\bar{P}(n)}{\partial n} & = & \displaystyle\sum_{i=1}^{N} E_{h(s)}\left[g(s)\right] = N\sum_{s} h(s)g(s) \\ \Leftrightarrow & \dfrac{\partial\bar{P}(n)}{\partial n} & = & \displaystyle k\rho\sum_{s}\; h(s)g(s) \end{array}$$

whence (5) follows from the stability condition (4) for the empty membership fixed point $n_{eq} = 0$. **Q.E.D.**

In our model, we regard empty membership instability as a necessary condition for autonomous growth from an empty or small club membership. The above theorem implies that favourable conditions are large k (contribution from members), large ρ (search efficiency) and a large value of $\sum_s h(s)g(s)$, an inner product of $h(s)$ and $g(s)$. Note that

$$\sum_{s} h(s)g(s) \equiv \|h\|\,\|g\| \cdot \langle h(s), g(s)\rangle$$

where $\|h\|$ and $\|g\|$ are the 2-norms of $h(s)$ and $g(s)$ respectively, and $\langle h(s), g(s)\rangle$ is their normalized inner product which measures their similarity, or goodness of match. Other favourable conditions are therefore a good match between aggregate demand and supply, and *skewness* – or small spread – of their distributions over the chunk types.

3.1 Music information sharing club example with simple requests

Figure (1) shows $\bar{P}(n)$ for the music information sharing club example for four $k\rho$ values for the simple request case, viz. $d = 1$.

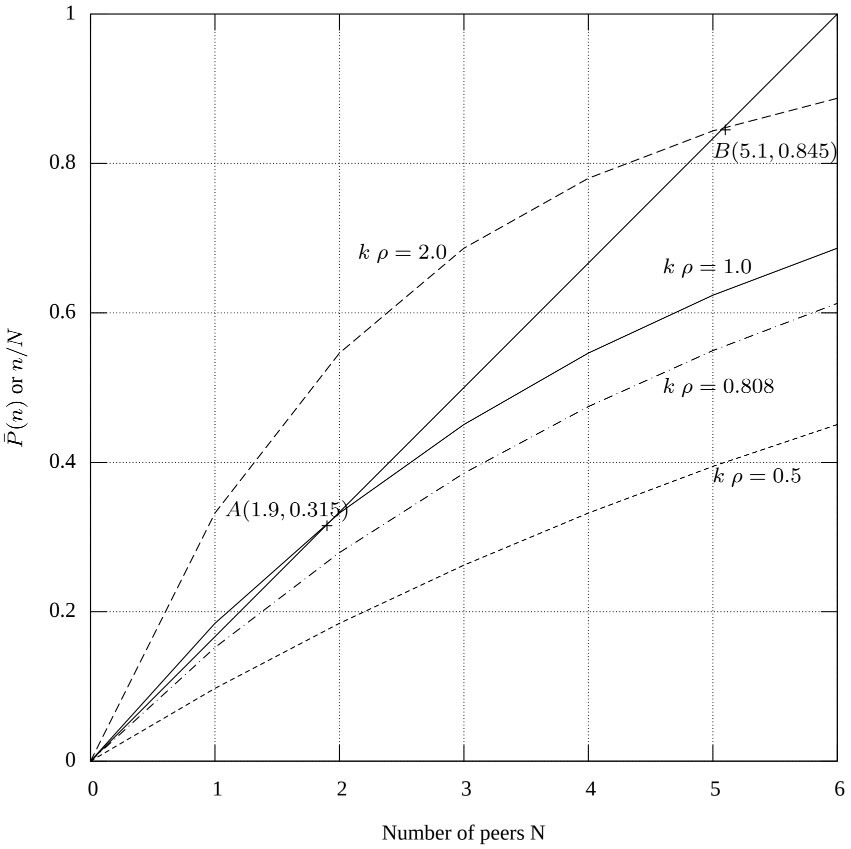


Figure 1: The music information sharing club example

For $k\rho = 2$, the model predicts that an empty club is unstable. Any disturbance, e.g. voluntary sharing or contribution, would trigger it to grow. The club would stagger rapidly towards the fixed point $n = 5.1$ — where $\bar{P}(x) = 5.1/6 = 0.85$ and and sustain itself around there. The peers are active members over 80% of the time on average. For $k\rho = 1$, an empty club is again unstable but the club sustains itself at a smaller average size of $n = 1.9$. With less supply and/or less efficient search function, peers are active only around 30% of the time on average. For $k\rho = 0.5$, an empty club now becomes stable. Joining peers are always more than offset by leaving members such that a positive membership is always transient. Peers are almost always inactive. Finally $k\rho = (N \sum_s h(s)g(s))^{-1} = 0.808$ is the critical case when an empty club is just stable/unstable.

It is important to note that the above analysis is of the average case. The actual dynamics of a realization of the club membership over time as $\mathcal{C}(t) \subset \mathcal{N}$ would sketch a sample path $(|\mathcal{C}(t)|, P_{\mathcal{C}(t)}(n))$ that staggers around the corresponding $\bar{P}(n)$ curve[2]. However, the family of $\bar{P}(n)$ curves for all π values define a direction field of average directions of the forces that act upon any sample path. The average direction is towards growth above the n/N diagonal, and towards shrinkage below, as shown in figure (2). In other words, the n/N diagonal is a boundary between two phases of the club dynamics, a growth phase for the club states above it and a shrinkage phase for those below. This is a powerful way to visualize the club dynamics, especially when π may vary over time in more complex cases.

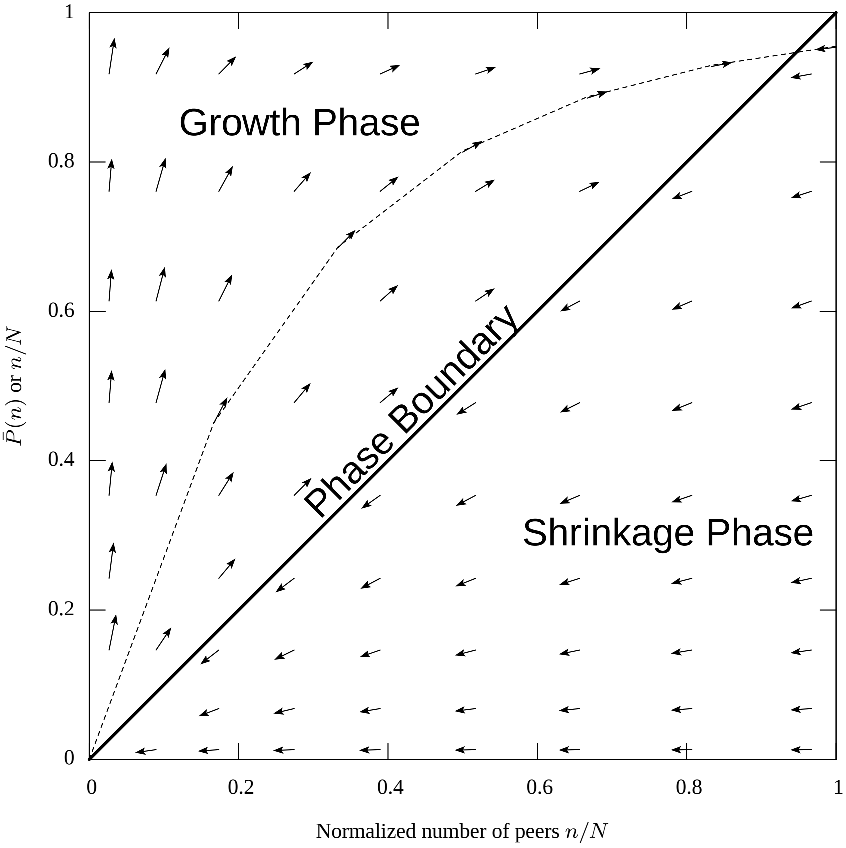


Figure 2: Phase diagram of club dynamics with direction field

[2] The staggering, or departure from the average case, would depend on the extent and rate of mixing, viz., the stochasticity of the club membership. Generally speaking, a large number of active peers with strong flows both in and out of the club would stay close to the average case with less staggering. Otherwise a sample path may actually get stuck with a niche self-sufficient club that sees neither peers joining nor members leaving.

3.2 Critical behaviour and multiple equilibria

Note that $p_i(0) = 0$ and $p_i(n)$ is bounded and concave increasing in n. When $d = 1$, $\bar{P}(n)$ is bounded and concave increasing in n also. Subsequently, there is at most one stable positive fixed point. Theorem 3.1 establishes a sharp threshold for π, a composite *control parameter* of the club as a dynamical system. The club would stabilize at an empty membership when $\pi < 1$, or the unique stable positive fixed point of equation (3) otherwise. In cases when π varies across the threshold of unity, the club would undergo critical change, and move towards either of the two stable fixed points.

When $d > 1$, an empty membership is always stable according to Theorem 3.1. For peer population above some minimum level $N_{crit} > 0$ such that n/N_{crit} is first tangential to $\bar{P}(n)$ as in figure (3), at least two positive fixed points exist. The smaller one would be unstable while the larger is always stable (see figure (3)). The smaller fixed point signifies a lower threshold, a "critical mass" of membership needed for autonomous growth thereafter. The club would be in danger of collapse whenever its membership falls below this level, even when such fall is transient to begin with.

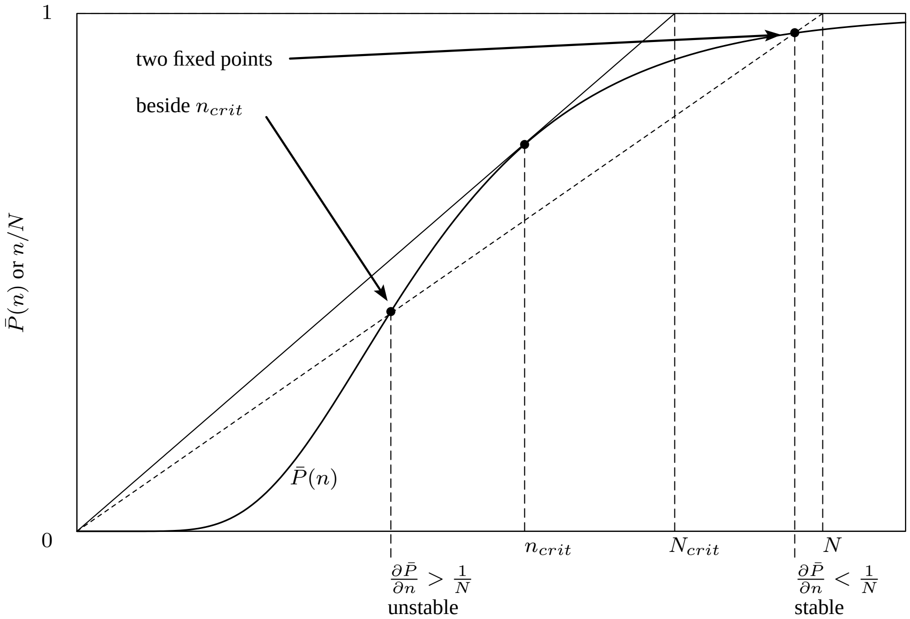


Figure 3: Critical population and bifurcation of fixed points, for $d > 1$

Proposition 3.1 (Critical Population and Bifurcation). *N_{crit} is the smallest solution to the simultaneous equations*

$$\left.\frac{\partial \bar{P}(n)}{\partial n}\right|_{n_{crit}} = \frac{\bar{P}(n_{crit})}{n_{crit}} = \frac{1}{N_{crit}}$$

where n_{crit} is a bifurcation point: once N increases above N_{crit}, two fixed points appear on either sides of n_{crit} and move away from it.

This proposition follows simply from the fact that $\bar{P}(n)$ is smooth, increasing and upper bounded at 1 (see figure (3)). The membership level n_{crit} is metastable as it is exactly marginal to the stability condition (4). In the special case when the peers are not differentiated in that $h_i(s) = h(s)$, the increase in $\bar{P}(n)$ concentrates around an inflection point just before n_{crit}. However, when the $h_i(s)$'s are spread out so that $\sum_s h(s)g(s)$ is highly variable, $\bar{P}(n)$ would increase more gradually. As a result, the bifurcation may occur more sharply with a wider spread between the two resulting fixed points.

4 A numerical example with truncated Zipfian aggregate demand

Consider a population of N peers with a truncated Zipfian aggregate supply, viz.:

$$g(s) = cs^{-\beta} \quad 1 \leq s \leq s_{max} \tag{6}$$

where $c = (\sum_{s=1}^{s_{max}} s^{-\beta})^{-1}$. This rank-frequency distribution is widely observed in Web and peer-to-peer file popularity measurement studies [12], [13] . The exponent β is often around and below 1. Its skewness as measured by its norm is

$$\|g\| = c\sqrt{\sum_{s=1}^{s_{max}} s^{-2\beta}}$$

which is determined by two key parameters, viz. the *peakedness* of the Zipfian distribution as governed by the exponent β, and the *variety* of chunk types as governed by s_{max}.

Generally speaking, $g(s)$ may match the aggregate demand $h(s)$ to different degrees. Below we analyze two cases, viz. the perfect match case when $h(s) = g(s)$ and the imperfect match case due to a simple shift between $h(s)$ and $g(s)$. Also, we consider simple requests ($d = 1$) throughout.

4.1 Perfect match case: $h(s) = g(s)$

According to Theorem 3.1, the autonomous growth condition is

$$N_{crit} \quad \geq \quad \frac{1}{k_{crit}\ \rho_{crit}} \frac{1}{c^2 \sum_s s^{-2\beta}} \tag{7}$$

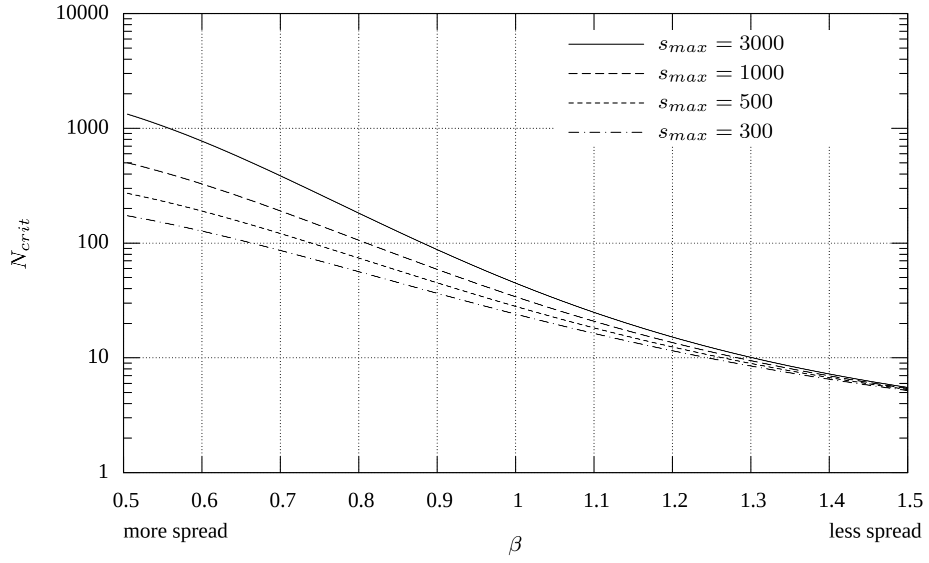


Figure 4: N_{crit} vs (β, s_{max}) for perfect match case ($k\,\rho = 1$)

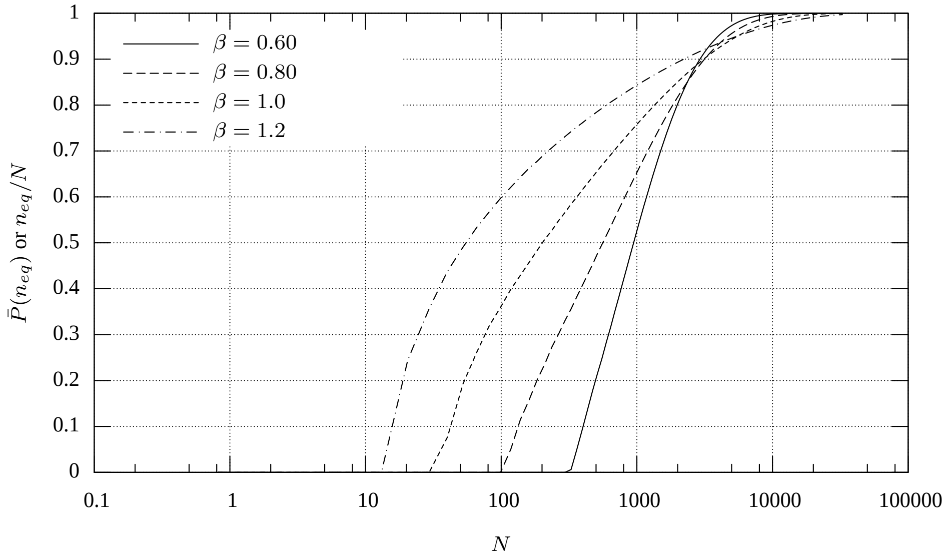


Figure 5: Expected equilibrium membership and success rate vs $N\,k\,\rho$ ($s_{max} = 1000$)

since $\langle g, h\rangle = 1$ and $\|g\| = \|h\| = c\sqrt{\sum_s s^{-2\beta}}$. The dependence of N_{crit} on β and s_{max} is shown in figure (4). Autonomous growth is favoured by large β (peakedness) and low s_{max} (variety).

Once the control parameter of the club is above the growth threshold, it would sustain around an equilibrium membership size n_{eq} as the unique stable fixed point of equation (3). Solving for different values of $Nk\rho$ and β gives figure (5). This figure shows the proportion n_{eq}/N, which is also the performance level of the club in terms of $\bar{P}(n_{eq})$, the average success rate of information search in the club.

4.2 Imperfect match due to simple shift

Supply and demand distributions would seldom match perfectly. In fact, one would often be considered leading the other. For example, a *demand lead* case would demand a wider variety of goods than the aggregate supply distribution offers, while a *supply lead* case sees more variety in the supplied goods. The "excess" in demanded types (or supplied types) reflects the types of goods that the supply (or the demand) cannot follow at a particular moment. Here we consider all such excess being concentrated in either the lowest or the highest ranks for simple illustrations.

In the supply lead case, the supply $g(s)$ is the same as defined in equation (6), and the demand distribution is :

$$h(s) = \begin{cases} 0 & \text{if } s \leq \delta \\ c'(s-\delta)^{-\beta} & \text{if } \delta < s \leq s_{max} \end{cases}$$

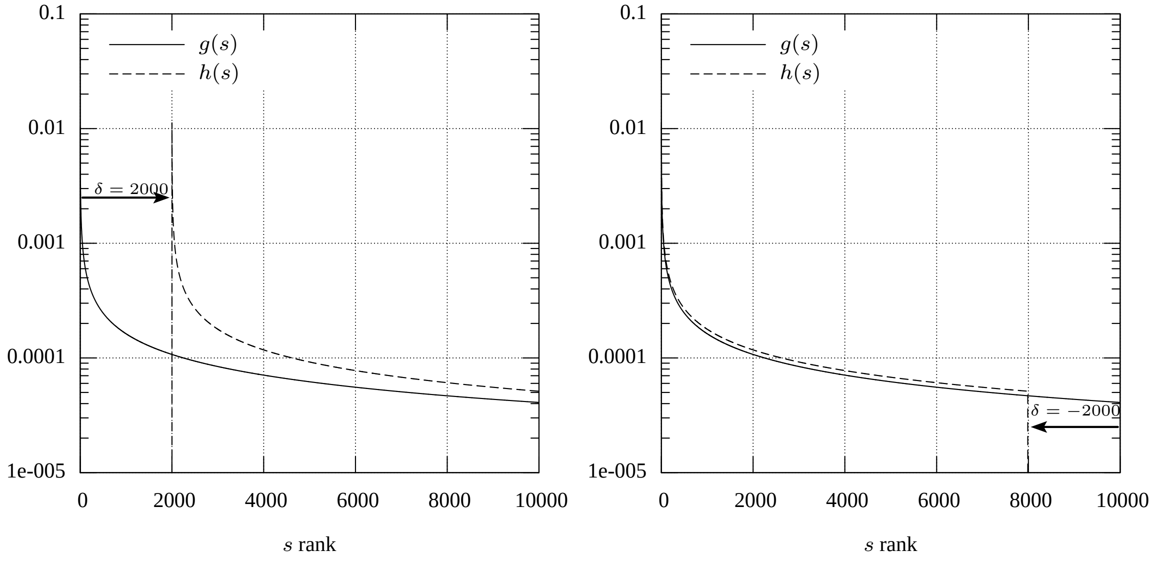
for $\delta \geq 0$, and

$$h(s) = \begin{cases} c''s^{-\beta} & \text{if } s \leq s_{max} + \delta \\ 0 & \text{if } s_{max} + \delta < s \leq s_{max} \end{cases}$$

for $\delta < 0$. c' and c'' are normalizing constants such that $\sum_s h(s) = 1$. See figure (6) for an illustrated example. A positive shift $\delta > 0$ means the excess types occupy the highest ranks, while a negative shift means they occupy the lowest ranks. The supply lead case would simply have the expressions of $g(s)$ and $h(s)$ exchanged.

Figure (7) shows that excess in the highest ranks are very demanding and would require a very large increase in N_{crit} for autonomous growth. However, excess in the lowest ranks actually decreases N_{crit} and autonomous growth becomes easier. This suggests that focussing of supply on chunk types of the highest ranks would trigger autonomous growth more readily.

In summary, the distinction between supply lead and demand lead cases is immaterial to the autonomous growth threshold, though it may be important to modelling supply and demand dynamics. What matters is where they differ — in the higher or lower ranks.



(a)

(b)

Figure 6: A supply lead case example ($s_{smax} = 10000$) with (a) positive δ and (b) negative δ

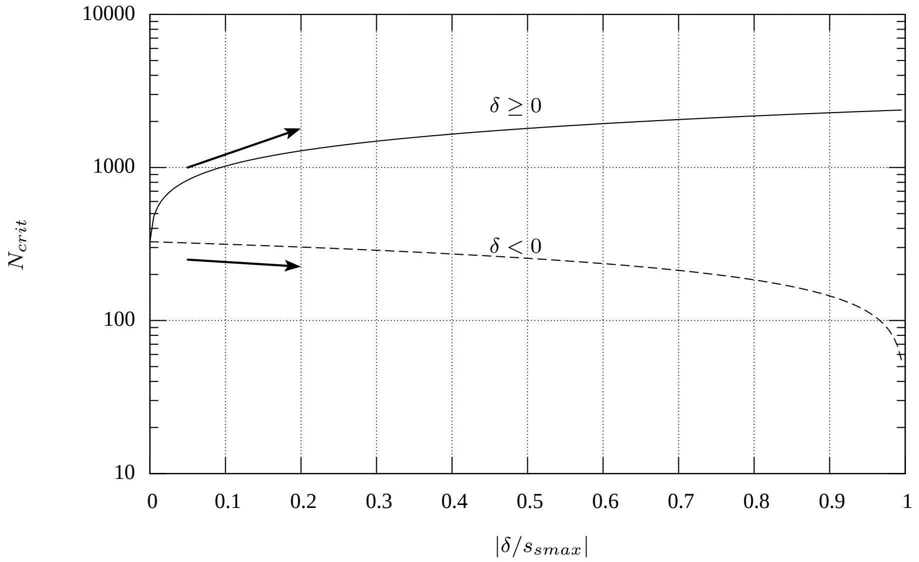


Figure 7: Effect of mismatch due to simple shift ($s_{max} = 1000$, $k\rho = 1$, $\beta = 0.6$)

5 Discussion

The simple ISC model displays interesting behaviour resulting from coupling of the membership and content dynamics. As a dynamical system, it exhibits phase transition in a composite control parameter π. Information sharing is sustainable with a good membership only when π is above a threshold of 1. It implies no effort in increasing π is worthwhile (e.g. by improving efficiency and supply, or increasing peer population) unless π goes above the threshold as a result.

Many researchers study the problem of excessive free-riding [11], [14], [15] that corresponds to the existence of many peers with $K_i = 0$ in our model. Incentive mechanisms are devised which reward contribution and/or penalize free-riding. One way to analyze such mechanisms is by extending the ISC model so that the search efficiency a peer sees becomes an increasing function of the contribution he decides on a rational basis. In this case, our ongoing study indicates that empty membership may become stable always, even for $d = 1$, as long as the club relies solely on members for content. An empty club and a rational (frugal) peer population are in a dead lock situation. The dead lock may be broken only when either the club has sufficient initial content to attract the peers, or some peers are generous enough to contribute without expecting extra benefit (therefore not rational in the conventional sense).

However, free-riding is not a problem *per se* under the non-rivalrous assumption. As free-riders are those with $K_i = 0$, we may redefine N to exclude them so that only contributing peers (those with positive payloads) are counted. As a result, N is reduced, and the average payload size k is increased (while the club's average content Nk remains the same as before). All results in this paper would continue to hold, albeit with the notion of a peer redefined. What matters is the contributing peer population: the club would grow as long as they are joining to share enough content and attract sufficiently many of themselves. The existence of free-riders is phantom to the system. They would be no nuisance as long as provisioning of extra copies carries no sharing cost.

Incentives would help not by reducing free-riding but by increasing the contributing peer population, viz. N. The distinction may seem frivolous as reduced free-riding is often regarded as increased contribution. However this may not be true always. One may imagine a negative incentive scheme which merely causes free-riders to demand less, or turn away altogether, without turning them into contributing peers. The club's content is not benefited. As incentive schemes are often costly to maintain in practice, negative schemes as such should be saved for positive ones that aim to increase

N directly. A reasonable principle in economizing the use of incentive schemes would be to focus on those peers who are bordering on free-riding, to coerce them into contributing.

In fact, the club's well being may actually be harmed in two possible ways when free-riding is discouraged. First, free-riders may develop into contributors if only they stay long enough for the club to become sufficiently important to them. Second, they may in fact be useful audience to the members, e.g. in newsgroups, BBS and forums, where wider circulation of the shared information often improves *all* due to network effects. (This would be diametrically *opposite* to the rivalrous assumption, and could actually be more appropriate than the non-rivalrous assumption if it more than offsets any sharing costs due to rivalrous consumption of other club resources.)

In cases where the non-rivalrous assumption is not appropriate due to significant sharing costs, e.g. in processing, storage and/or network bandwidth, penalizing free-riding would be more necessary to reduce their loading on the system and the contributing peers. A possible corresponding extension of the ISC model is to incorporate the natural reduction in availability of information goods as their demand increases. For instance, the failure rate of demand for chunk type s may become an increasing function of its total demand $nh(s)$ one way or the other. However, the choice of functional relation between availability and demand should depend on the nature of the sharing cost.

Apart from extensions needed in rivalrous situations, the ISC model has two intrinsic limitations. First, it captures only the average case behaviour of a nonlinear and stochastic dynamical system. Transient and lock-in, especially when the club is small and peers act with large delay, may render the average case view totally useless. Second, the join/leave decisions are often more heterogeneous than assumed here. Requests may comprise variable numbers of demand instances and peers may deliberate their decisions and behave differently.

6 Conclusion and further works

We have analyzed information sharing in a very general setting, by means of a statistical model (ISC) with peers of different demand and supply of information. As a dynamical system, the model exhibits interesting critical behaviour with multiple equilibria. A unique feature of the ISC model is that information is chunked and typed, as we believe modelling the composition of the information content is crucial in many situations of interest. Subsequently, it displays a sharp growth threshold that depends on the goodness of match in the types of information being demanded and supplied by the sharing members. While being rich in behaviour, this model is simple enough for detail analysis of the equilibrium states. In particular, we analyzed a truncated Zipfian distribution of information types and derived the growth threshold for the existence of any sustainable equilibrium, as well as the corresponding membership size and performance level.

Much simplicity of the ISC model stems from the non-rivalrous assumption made. Real situations are more complicated in that peers may be sharing both rivalrous resources and non-rivalrous information at the same time. However, Benkler [16] points out that overcapacity is a growing trend in distributed systems such as the Internet, so much so that even rivalrous resources are increasingly being shared like non-rivalrous goods. On the other hand, free-riding would work in the opposite direction if the community in question is prosperous and attracts so much free-riding that contention for some shared rivalrous resources begins to happen. The challenge is then to identify the major sources of social cost of sharing [17] and properly account for them. A natural extension of our work would be to study the interplay between an information sharing community and different types of host networks. Another extension is certainly the incentive issue: how incentive schemes should be devised in response to different mixes of rivalrous and non-rivalrous resources.

References

[1] "SETI@home." [Online]. Available: http://setiathome.ssl.berkeley.edu/

[2] J. Y. B. Lee and W. T. Leung, "Design and analysis of a fault-tolerant mechanism for a server-less video-on-demand system," in *Proceedings of International Conference on Parallel and Distributed Systems*, 2002.

[3] M. Hefeeda, A. Habib, B. Botev, D. Xu, and B. Bhargava, "PROMISE: Peer-to-peer media streaming using collectcast," in *Proceedings of ACM Multimedia*, 2003.

[4] "Kazaa." [Online]. Available: http://www.kazaa.com/

[5] "Bittorrent." [Online]. Available: http://bitconjurer.org/BitTorrent/

[6] B. Gu and S. Jarvenpaa, "Are contributions to P2P technical forums private or public goods? - an empirical investigation," in *Proceedings of Workshop on Economics of P2P Systems*, June 2003.

[7] R. Krishnan, M. D. Smith, Z. Tang, and R. Telang, "The virtual commons: Why free-riding can be tolerated in file sharing networks," in *Proceedings of International Conference on Information Systems*, 2002.

[8] C. Buragohain, D. Agrawal, and S. Suri, "A game theoretic framework for incentives in P2P systems," in *Proceedings of the Third International Conference on Peer-to-Peer Computing*, 2003.

[9] P. Golle, K. Leyton-Brown, I. Mironov, and M. Lillibridge, “Incentives for sharing in peer-to-peer networks,” in *Proceedings of the 2001 ACM Conference on Electronic Commerce*, 2001.

[10] K. Ranganathan, M. Ripeanu, A. Sarin, and I. Foster, “To share or not to share: An analysis of incentives to contribute in collaborative file sharing environments,” in *Proceedings of Workshop on Economics of P2P Systems*, June 2003.

[11] M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica, “Free-riding and whitewashing in peer-to-peer systems,” in *Proceedings of ACM SIGCOMM Workshop on Practice and Theory of Incentives in Networked Systems*, 2004.

[12] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, “Measurement, modeling, and analysis of a peer-to-peer file-sharing workload,” in *Proceedings of Symposium on Operating Systems Principles*, 2003.

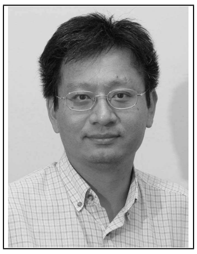
[13] S. Saroiu, P. K. Gummadi, and S. D. Gribble, “A measurement study of peer-to-peer file sharing systems,” in *Proceedings of Multimedia Computing and Networking*, 2002.

[14] J. Hindriks and R. Pancs, “Free riding on altruism and group size,” Queen Mary College, University of London, Department of Economics, Tech. Rep. wp-436, 2001.

[15] E. Adar and B. Huberman, “Free riding on gnutella,” *First Monday*, vol. 5, Sep 2000.

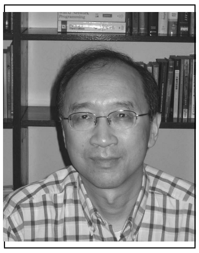
[16] Y. Benkler, “Sharing nicely: on shareable goods and the emergence of sharing as a modality of economic production,” *The Yale Law Journal*, vol. 114, pp. 273–358, 2004.

[17] H. R. Varian, “The social cost of sharing,” in *Proceedings of Workshop on Economics of P2P Systems*, June 2003.



[W.-Y. Ng] Wai-Yin Ng received his BA in 1985 (specializing in control and operational research) and PhD in control engineering in 1989, both from the University of Cambridge, U.K. and is currently associate professor in information engineering in The Chinese University of Hong Kong. His current research focus is in complex networks, a young vibrant science concerned with connectivity, complexity and emergent phenomena in both natural and artificial systems.

[W.K. Lin] Wing Kai Lin received his B.Eng degree in 2001 and is currently completing his Master degree in information engineering, both from the Chinese Univeristy of Hong Kong. His research interest includes replication in peer to peer systems and economics issues in incentive mechanisms.



[D.M. Chiu] Dah Ming Chiu received his B.Sc. degree from Imperial College London in 1975, and Ph.D. degree from Harvard University in 1980. He worked for Bell Labs, Digital Equipment Corporation and Sun Microsystems Laboratories. Currently, he is a professor in the Department of Information Engineering at the Chinese University of Hong Kong. He is on the editorial board of the International Journal of Communication Systems.