Stability of Peer-to-Peer Swarming Systems*

Daniel S. Menasché[1], Antônio A. Rocha[2], Edmundo de Souza e Silva[1], Rosa M. M. Leão[1], Don Towsley[3]

[1]Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro – Brazil

[2]Universidade Federal Fluminense (UFF), Niterói – Brazil

[3]University of Massachusetts at Amherst, Amherst – USA

{sadoc, arocha, edmundo, rosam}@land.ufrj.br, towsley@cs.umass.edu

Abstract. *Peer-to-peer swarming is one of the* de facto *solutions for content dissemination in today's Internet. By leveraging resources provided by users, peer-to-peer swarming is a simple and efficient mechanism for content distribution. Nonetheless, peer-to-peer systems are not always scalable. The goal of this paper is to study some limits on the scalability of peer-to-peer systems. To this aim, we propose a novel approach to derive insights on the stability of peer-to-peer systems. Then, we study the impact of different system parameters, such as peer and publisher policies for neighbor and piece selection, on the system's ability to scale. Using analytical models, we also propose novel strategies to improve the throughput of the system.*

1. Introduction

Peer-to-peer swarming, as used by BitTorrent, is one of the *de facto* solutions for content dissemination in today's Internet. By leveraging resources provided by users, peer-to-peer swarming is a simple and efficient mechanism for content distribution [Rocha et al. 2009]. Although peer-to-peer swarming has been widely studied for a decade, prior work has focused primarily on the positive aspects of peer-to-peer swarming. This paper focuses on the fundamental limits of scalability of peer-to-peer swarming systems.

In peer-to-peer swarming systems, as peers join a swarm to download a content they bring resources such as bandwidth and memory to the system. That way, the capacity of the system increases with the arrival rate of peers. Furthermore, increasing the arrival rate of peers can also increase content availability [Menasche et al. 2009b]. In the presence of publishers that are always present to serve peers and have enough service capacity for peers to smoothly complete their download, increasing the arrival rate of peers decreases the probability that a piece will be unavailable among peers [Menasche et al. 2010, Menasche et al. 2009a].

A system is said to be scalable if the system's throughput, $i.e.$, the rate at which users complete their downloads, increases linearly with increasing user population. The capacity of peer-to-peer systems is expected to increase with the arrival rate of peers once each peer acquires useful data to share with other peers. However, the increase in capacity might not be translated in a corresponding increase in system throughput, and peer-to-peer swarming systems are not always scalable. The system is said to be unstable

*This work is supported in part by grants from CNPq and FAPERJ.

(and non-scalable) when the population grows unboundedly with time, which happens, for instance, if the system load surpasses a certain threshold, $i.e.$, when the arrival rate of peers is beyond the "stability limit".

In essence, instability occurs in swarming systems because publishers can become bottlenecks. In this case, when two peers meet, they might not have useful data to share. Let the stability region of a system be the set of parameter values for which the system is known to be stable. The stability region of peer-to-peer swarming systems has recently gained attention from the research community [Mathieu and Reynier 2006, Hajek and Zhu 2010, Zhu and Hajek 2011, Menasche et al. 2011]. The problem has been observed in practical scenarios [Murai 2011], and is the object of study of this work.

In a peer-to-peer swarming system, each peer has to make two decisions before transmitting each piece: 1) which piece to transmit and 2) to whom to transmit it. Although the former decision has received some attention in previous works (for instance, it has been shown that rarest-first piece selection and random useful piece selection yield the same stability region [Hajek and Zhu 2010]), the implications of the peer selection strategy have not been thoroughly discussed (notable exceptions being [Mathieu and Reynier 2006, Menasche et al. 2011]). Previous works assumed peers choose their neighbors using random peer selection [Nunez-Queija and Prabhu 2008, Hajek and Zhu 2010, Zhu and Hajek 2011].

We present novel results on the stability of peer-to-peer swarming systems. To this aim, we consider parameters in the problem space that have not been previously analyzed, such as the impact of the neighbor selection algorithms on the system stability.

To illustrate the sort of results that we seek in this paper, let λ be the arrival rate of peers to the system and let U be the the service capacity of the stable publisher, measured in blocks/s. Hajek and Zhou [Hajek and Zhu 2010, Zhu and Hajek 2011], following up work by Mathieu and Reynier [Mathieu and Reynier 2006], have shown that if $\lambda > U$ and peers arrive according to a Poisson process, the number of peers increases unboundedly with time. It has also been shown that simple strategies can alleviate, and in some cases resolve, the instability problem. For instance, if peers reside in the system after completing their downloads, on average, the same time that they take to download a piece, then the system is always stable [Zhu and Hajek 2011]. Nevertheless, as peers have no incentive to stay in the system after completing their downloads, it is important to investigate whether other simple strategies that do not depend on providing incentives for peers to remain online after the download completion can improve system performance and stability.

The goal of this paper is to evaluate the impact of different system parameters and system's strategies on the stability of the system studying its throughput. We pose the following questions:

$a)$ how to increase the scalability of the system by letting peers strategically select their neighbors?

$b)$ how does the scalability depend on different system parameters?

We provide the following answers to the above questions. First, we derive an upper bound on the throughput when the publisher adopts the most deprived peer selection and

	peers	publisher
(a)	random peer/random useful piece (RP/RUB)	random peer/random useful piece (RP/RUB)
(b)	random peer/rarest first piece (RP/RFB)	random peer/rarest first piece (RP/RFB)
(c)	random peer/random useful piece (RP/RFB)	most deprived peer/rarest first piece (MDP/RFB)
(d)	random useful peer/random useful piece (RUP/RFB)	most deprived peer/rarest first piece (MDP/RFB)

Table 1. Neighbor and piece selection policies

rarest-first piece selection, while peers adopt random peer selection and random useful piece selection. The bound is significantly larger than the maximum attainable throughput in the scenarios studied in [Hajek and Zhu 2010], when both peers and publishers adopt random peer and random useful piece selection. Then, we use a simple Markov chain model to study how the throughput of the system scales with the number of peers.

The remainder of this paper is organized as follows. In the upcoming section we set the background and present related work in §3. In §4 we introduce our model. Then, we consider random peer selection, altruistic lingering and most deprived peer selection in §5. In §6 we show how the system throughput scales with the population size under different system settings and §7 concludes our work.

2. Background

In this section we describe the policies for peer and piece selection considered in this paper (see Table 1). Throughout this paper, we assume that at every transmission opportunity peers select a neighbor uniformly at random to exchange pieces (random peer selection). In §5 we briefly discuss the case where trackers dynamically inform, to each peer P, the members of the swarm in need of pieces owned by P. In this case, peers can select their neighbors uniformly at random among those that need the pieces they have. We refer to such neighbor selection policy as random *useful* peer selection.

After choosing a neighbor, each peer selects for transmission one of the pieces that it owns and that its neighbor does not have. If the piece is selected uniformly at random, the policy is referred to as random useful piece selection. If peers have access to a list of the number of replicas of each piece, they can build a rarest-piece set containing the indices of the pieces with least number of copies in the swarm [Legout et al. 2007]. This set can then be used by peers to select which piece to transmit, such policy being referred to as *rarest first piece selection*.

The publisher can select its neighbors (that is, peers to transmit pieces) and pieces in the same way as the rest of the population. In addition, publishers can also select their neighbors using the most deprived policy. Under this policy, the publisher prioritizes sending pieces to peers that own the least amount of pieces among those in the swarm. If the arrival rate of peers is large, these peers are likely to be content-less peers, also referred to as newcomers. The rationale behind the most deprived policy is to transmit rare pieces to peers that will linger longer in the system and as such will have more time to distribute the rarest pieces throughout the swarm.

3. Related Work

The literature on stability and throughput scaling laws in the realm of peer-to-peer swarming system is scarce, specially if contrasted with the vast literature on these topics in the

realm of queueing and wireless systems [Bertsekas and Gallagher 1992, Chapter 4].

The service capacity of peer-to-peer systems was first analyzed by Yang and de Veciana [Yang and De Veciana 2004]. Their analysis involves a closed system, wherein a peer departure immediately triggers a peer arrival, so that the population size remains constant. Using this system, they analyze the transient increase in throughput after a flash crowd. They also considered an idealized fluid model to study the system steady state. The fluid model was further explored by Zhang *et al.* [Zhang et al. 2010], among others. None of these works considered the instability problem that occurs due to the fact that one piece in the system might become rare compared to the others. This problem, referred to as the missing piece syndrome, was first pointed out by Mathieu and Reynier [Mathieu and Reynier 2006].

The most deprived policy was first proposed in the context of live streaming by Massoulie et al. [Massoulie and Vojnovic 2006]. In [Massoulie and Vojnovic 2006], the authors show that the most deprived policy yields a good balanced between high throughput and low delays. In this paper, in contrast, we are interested in the stability implications of the most deprived policy.

To the best of our knowledge, [Hajek and Zhu 2010, Zhu and Hajek 2011] and previous works considered only random peer selection [Nunez-Queija and Prabhu 2008], files with at most two pieces [Reittu 2009, Norros et al. 2009], or considered a different class of peer-to-peer networks as those considered here [Leskela et al. 2010]. Mathieu and Reynier [Mathieu and Reynier 2006] pointed out the potential advantages of most deprived peer selection, but did not pursue its in depth analysis since peers can cheat when announcing their ages. In this paper, in contrast, we analyze different peer selection strategies for peer-to-peer networks that resemble BitTorrent, but assuming peers that do not misbehave. The model considered in this paper is based on the one presented by Hajek and Zhu [Zhu and Hajek 2011], with two important modifications: 1) the publisher can adopt the most deprived peer policy and 2) a fraction of peers might reside in the system after completing their downloads ([Zhu and Hajek 2011] assumes that either all or no peers reside in the system as seeds after concluding their downloads).

Whereas previous work [Nunez-Queija and Prabhu 2008, Reittu 2009, Norros et al. 2009, Leskela et al. 2010] assume that peers have no information about the number of replicas of each piece in the system, in this paper, inspired by BitTorrent, we leverage the fact that most deprived peer/rarest-first piece selection are practical peer and piece selection mechanisms. As we will show, it suffices that only publishers adopt such mechanisms in order to improve the throughput of the whole population.

The implications of peer selection strategies on the stability of swarming systems was first studied in [Menasche et al. 2011]. This paper extends [Menasche et al. 2011] in multiple ways as we 1) allow a limited fraction of peers to reside in the system after completing their downloads; 2) consider publishers that serve only missing pieces and 3) compare the performance of the system under different peer transmission schemes.

4. Model

We consider a population of peers that arrive to a swarm with rate λ peers/s, interested in downloading a file. The file is divided into K pieces, that peers download from and upload to each other.

A file is divided into K pieces. Let $\{1, 2, \ldots, K\}$ be the set of pieces, and $\mathcal{C}$ be the set consisting of all subsets of $\{1, 2, \ldots, K\}$. A type C peer is a peer that has a collection C of pieces of the file, $C \in \mathcal{C}$. For instance, if $K = 2$ set $\mathcal{C}$ is $\{\{\}, \{1\}, \{2\}, \{1, 2\}\}$. A type $\{\}$ peer is a peer that has no pieces, also referred to as content-less or newcomer.

The publisher has service capacity U pieces/second. If the publisher adopts random peer selection, at the end of a service interval, which have mean duration $1/U$, it selects a new peer uniformly at random to transmit another piece. If the publisher adopts the most deprived peer selection, in contrast, and $U < \lambda$, a fraction U/λ of peers receive a piece from the publisher immediately after arriving to the system. This is because newcomers are content-less, so they will be serviced with priority by the publisher. The publisher cannot serve all newcomers due to its limited capacity with respect to the arrival rate of peers, and the peers that are served by the publisher are referred to as *gifted peers* (the motivation for the label *gifted* will become clear later). It should be clear then that the arrival rate of gifted peers and non-gifted peers are U and $\lambda - U$, resp. (see Fig. 1). Note that we are assuming that at every transmission opportunity, the publisher will find a newcomer with high probability.

Peers adopt the random peer, random useful piece selection. Transmission opportunities for each peer occur at their transmission rate, μ. At every opportunity, a peer selects a target peer uniformly at random to transmit a piece. The piece to be transmitted is selected uniformly at random among those that the target peer does not own. Alternatively, publishers or peers can adopt the rarest-first piece selection policy, according to which they select and transmit the rarest piece among those that the target peer does not own.

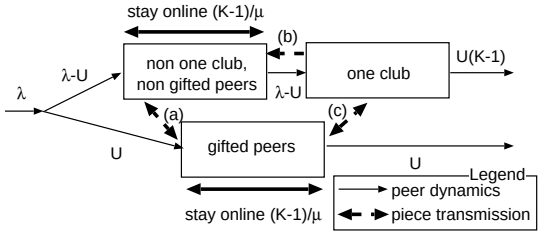

Figure 1. Publisher: most deprived peer-rarest first piece selection; peers: random peer-random piece selection.

A fraction F of peers that complete their downloads becomes seeds. The remaining fraction immediately leaves the system. Seeds remain in the system for an average of $1/\gamma$ and then leave (we assume $\gamma = \mu$, except stated otherwise).

4.1. Peer Dynamics and Content Flows

Next, we illustrate the main properties and insights of our model in the setting of Figure 1. Peers adopt the random peer/random piece policy, whereas the publisher adopts most deprived peer/rarest first piece policy. Peers arrive to the swarm at rate λ. Let $\lambda > U$. A fraction U/λ of the requesters immediately receives a piece from the publisher when they arrive to the system, and become *gifted peers*. The ones that do not receive a piece from the publisher are referred to as *non gifted peers*.

If gifted and non gifted peers could leave the system at rates $\lambda - U$ and U, respectively, flow balance yields a stable system. Nevertheless, this is not always the case.

To illustrate this point, consider for instance the scenario where we have a large number of peers in the system that have all pieces of the file except a tagged one. This piece c is referred to as the *missing piece*, and peers that don't have this piece are of type $C = \{1, 2, \ldots, K\} \setminus \{c\}$ and comprise the *one club*. Next, we analyze the implications of the publisher becoming a bottleneck due to the formation of the one club.

The only peers that can contribute to departures from the one club are the publisher and the gifted peers (recall that the gifted peers have the rarest piece c). Assuming that the publisher serves only newcomers, though, and noting that gifted peers stay online on average $(K-1)/\mu$ before departing, the departure rate of peers from the one club is limited by $(U(K-1)/\mu)\mu = U(K-1)$. That is because the mean number of gifted peers in the system is $U(K-1)/\mu$, and each of them contributes with capacity μ to the departure rate of peers from the one club. Since the arrival rate of peers to the one club is $\lambda - U$, the system will be stable only if the flow of peers into the one club is equal to the flow out. Flow balance occurs if the capacity to serve peers in the one club is larger than the rate at which peers join the one club, $i.e.$, $U(K-1) > \lambda - U$ or $\lambda < KU$. In the following sections, we will further formalize this claim, after introducing the details of our model. In §6 we will consider a special version of the model described above, wherein the mean time between transmissions are exponentially distributed, so as to obtain a Markov chain which can be solved numerically. Nonetheless, note that the arguments above are independent of such assumptions.

4.2. Model Details

Let n_C denote the number of peers of type C. The system described in the previous section has state space $\mathbf{n} = (n_C : C \in \mathcal{C})$. Let $\mathbf{e}_C$ be a vector of the same length as $\mathbf{n}$, with all its elements equal to zero, except the one corresponding to C, which equals one. If the system starts at state $\mathbf{n}$ and a peer of type C downloads piece i the system transitions to state $T_{C,i}(\mathbf{n})$.

If $\lambda < U$ and users adopt the random peer/random useful piece policy, the system is known to be stable [Hajek and Zhu 2010]. More generally, if $\lambda < U$ the system is expected to be stable for a broad range of policies, since there will be a drift towards states with small population size. Even in face of a large one club, the departure rate of peers, expected to be at least U, will be larger than the arrival rate, λ. In [Hajek and Zhu 2010] it is argued that the formation of a large one club constitutes a worst case scenario. Therefore, in what follows we assume $\lambda > U$.

If the publisher adopts the most deprived peer selection and rarest-first piece selection, whereas peers adopt random peer, random useful piece selection, the arrival rate of gifted peers is characterized by transitions from state $\mathbf{n}$ to $\mathbf{n} + \mathbf{e}_{\{r\}}$ with rate U, where r is the rarest piece in $\mathbf{n}$. The arrival rate of non-gifted peers is characterized by transitions from $\mathbf{n}$ to $\mathbf{n} + \mathbf{e}_{\emptyset}$, which occur with rate $\lambda - U$. Finally, we let $\mu_{\mathbf{n},C,i}$ denote the transition rate from $\mathbf{n}$ to $T_{C,i}(\mathbf{n})$. The dynamics of piece transmissions between peers is characterized through $\mu_{\mathbf{n},C,i}$, which is determined by the neighbor and piece policies. In what follows, we illustrate the definition of $\mu_{\mathbf{n},C,i}$ considering peers that adopt the random peer/random useful piece policy.

Under the random peer/random useful piece policy, a contact between two peers is said to be successful if it yields the transmission of a piece. For simplicity, assume that

the mean time between contacts of peers equals $1/\mu$, and is independent of whether the contact is successful or not. Adopting this simplifying assumption allows us to frame our results in the same reference setting as the one considered in [Hajek and Zhu 2010],

$$\mu_{\mathbf{n},C,i} = n_C \left(\mu \sum_{S \in \mathcal{C}: i \in S} n_S / (|S - C|) \right) / |\mathbf{n}| \quad (1)$$

According to (1), the rate at which a peer of type C receives piece i from a peer of type S when the population state is $\mathbf{n}$ equals the contact rate μ multiplied by rate at which users of type $S \in \mathcal{C}$ transmit piece i to peers of type C. The latter rate is proportional to n_S and to the probability of a user of type C being selected for transmission by a peer of type S, $n_C/|\mathbf{n}|$. Finally, the probability that piece i is selected among $S \setminus C$ is $1/|S \setminus C|$. This dynamics is a member of the class of dynamics for which the results presented in §5 holds. It will also be used in our numerical results in §6.

5. Stability Region: Implications of Altruistic Lingering and Most Deprived Peer Selection

Our goal in this section is to show the implications of altruistic lingering and most deprived peer selection on the stability region. First, we show that if a fraction F of peers reside in the system for mean time $1/\gamma$ and $F/\gamma > 1/\mu$ the system is stable. Then, we consider a stable publisher that adopts the most deprived peer/rarest first piece policy, and show that in this setting the system is stable if $\lambda < KU$. These results significantly broaden the stability region of the reference setting considered in the following proposition.

Proposition 5.1 *[Zhu and Hajek 2011] When peers and publisher adopt random peer and random piece selection, and peers leave the system immediately after concluding their downloads, the system is stable iff $\lambda < U$. Furthermore, if peers reside in the system for $1/\gamma$ after completing their downloads, the system is always stable.*

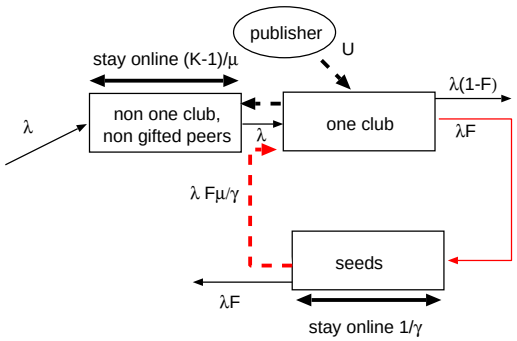

Figure 2. Random peer and random piece selection with altruistic lingering. The presence of a feedback loop (in red) is key to guarantee an always stable system.

Altruistic Lingering Let F be the fraction of peers that reside in the system after completing their downloads, and let $1/\gamma$ be the mean residence time after download completion.

Proposition 5.2 *When peers and publisher adopt random peer and random piece selection, and a fraction F of peers reside in the system for $1/\gamma$ after completing their downloads, the system is stable if $F/\gamma > 1/\mu$.*

Proof. [Sketch] Figure 2 indicates that each seed contributes with rate μ to the depletion of the one club. It follows from Little's result that there are on average $\lambda' F/\gamma$ seeds in the system, where λ' is the rate at which peers leave the one club ($\lambda' > 0$). Then, $\lambda' = \min(\lambda, U + \lambda' F\mu/\gamma)$. Equivalently, $\lambda' = \min(\lambda, U/(1-(\mu F)/\gamma))$. As far as $F\mu/\gamma > 1$, we have $\lambda' = \lambda$ and the system is stable.

For an alternative argument, let the *available bandwidth* of the seeds, α, be the difference of the service capacity of the seeds, $\mu\lambda' F/\gamma$ and the service required to make the one club stable, $\mu\lambda/\mu = \lambda$. If $\alpha \geq 0$ the system is stable. Therefore, if $\mu F/\gamma \geq \lambda/\lambda' \geq 1$ the system is stable. □

Publisher Policy: Most Deprived First In this section we study the system throughput when the publisher adopts the most deprived peer selection strategy and rarest-first piece selection, whereas peers adopt random peer, random useful piece selection.

Proposition 5.3 *If the publisher adopts most deprived peer selection and rarest-first piece selection and peers adopt random peer and random useful piece selection, the maximum achievable throughput is upper bounded by KU.*

Proof. [Sketch] In what follows, let $\lambda > KU$. First, we note that all states $\mathbf{n} = (n_C : C \in \mathcal{C})$ are achievable. Eventually, the system reaches a state in which a large number of peers have all pieces except a tagged one. These peers are also referred to as *one-club peers* (see Figure 1).

As a consequence of the random peer selection adopted by peers, if the one-club is large enough then gifted peers will transmit content only to one-club peers, with high probability. As shown next, if $\lambda > KU$ the one club grows unboundedly. Therefore, the effect of transmissions from gifted peers to members outside the one club reduces with time, and does not affect the maximum achievable throughput. For this reason, henceforth we neglect arrow (a) in Figure 1.

All uploads from the stable publisher are to newcomers, a fraction U/λ of which effectively receive pieces from the publisher. Each peer that receives a piece from the publisher has an additional expected lifetime of $(K-1)/\mu$. During this time, it will serve on average $K-1$ peers from the one-club, who will then leave the system. Therefore, the population of the one-club decreases at a rate of $U(K-1)$, and increases at a rate of $\lambda - U$. Hence, the total departure rate of peers is upper bounded by $U(K-1) + U = UK$. □

If the stable publisher uses random useful piece selection rather than rarest-first, the stability region degenerates to the case analyzed in [Hajek and Zhu 2010]. That is because in this case the argument presented in the above paragraph still holds, after replacing U by U/K, which yields a stable system if and only if $\lambda < U$.

Remark: Note that a simple modification in the publisher strategy yields significant gain in throughput. Consider, for instance, a typical piece size of 256KB and files varying being 600MB and 800MB. Adoption of the most deprived peer policy increases the

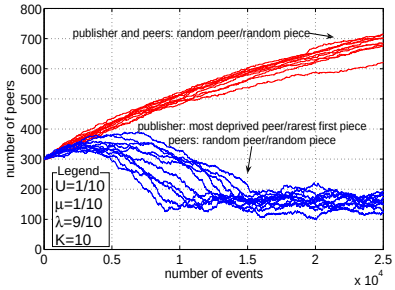

Figure 3. The effect of the publisher strategy.

throughput by a factor that varies roughly between 2,000 and 3,000 compared to the throughput when peers and publisher adopt random peer/random useful piece.

Next, we provide evidence that the bound of KU on the system throughput is achievable. To this aim, we consider simulation results shown in Figure 3. Let $\mu = U = 1/10$ pieces/s, $K = 10$, $\lambda = 9/10$, and all peers (except possibly the publisher) adopt random peer/random useful piece selection. Mean time between arrivals and contacts are exponentially distributed. Figure 3 shows, for 20 simulation runs, the population size as a function of simulation time (measured in number of events). When publishers adopt most deprived peer/rarest first piece selection (blue curves), the population size oscillates around its mean. This indicates that the system is likely to be stable and a throughput of 0.9 is achieved (see Proposition 5.3). In contrast, when publishers adopt random peer/random useful piece selection (red curves), the population grows unboundedly, which is in accordance to the results in [Hajek and Zhu 2010].

Peer Policy Implications Up to this point we considered peers that have minimal knowledge about the state of the system and adopt the random peer/random useful piece policy. Next, we consider the implications of the peer policy, assuming that the publisher adopts most deprived peer/rarest first piece policy. First, we analyze the random useful peer policy. Then, we analyze the case where peers also adopt the most deprived peer/rarest first piece policy.

Under the random peer policy, the rate at which a tagged peer A contacts a tagged peer B for transmissions is $\mu/|\mathbf{n}|$. Therefore, each peer is contacted by the rest of the population roughly at rate μ. The latter observation is key in the derivation of our results below. Nonetheless, note that if there is a large number of peers that have all pieces except a tagged one ($i.e.$, a large number of one-club members), most of the contact opportunities will occur among the one-club members, and will consist of unuseful contacts. In the sequel, we discuss how downlink constraints and reciprocity affect the system throughput when peers adopt random *useful* peer selection.

Recall that, according to the random *useful* peer policy, a peer A selects for transmission a peer that is in need of at least one of the pieces that A owns. Therefore, the mean time that gifted peers reside in the system after receiving a block from the publisher (also referred to as their mean lifetime), might be smaller that $(K-1)/\mu$. In particular, if the one club population grows unboundedly, the lifetime of gifted peers will be negligible. As discussed in the previous paragraph, the use of random peer selection by the peers, which yields gifted peers mean lifetime of $(K-1)/\mu$, is key. Alternatively, other

factors can also yield such a mean lifetime.

One factor that constraints the mean lifetime of gifted peers to $(K-1)/\mu$ is the limited download capacity of peers. A second constraining factor is the reciprocity that occurs among peers in most peer-to-peer swarming systems. Reciprocity, also known as tit-for-tat, enforces that peer A transmits a piece to peer B if peer B transmits a piece to peer A. To bootstrap peers that do not receive pieces from the publisher, peer A might also need to optimistically send pieces to resource-less peers. If non-gifted peers, when transmitting content to gifted peers, adopt a tit-for-tat strategy, the mean lifetime of gifted peers will be at least equal to $(K-1)/\mu$, independently of the peer selection strategy adopted. That is because gifted peers can only transmit packets at rate μ, hence receive packets at that rate from one-club members. Under any of these constraining factors, we conjecture that Propositions 5.1-5.3 still hold.

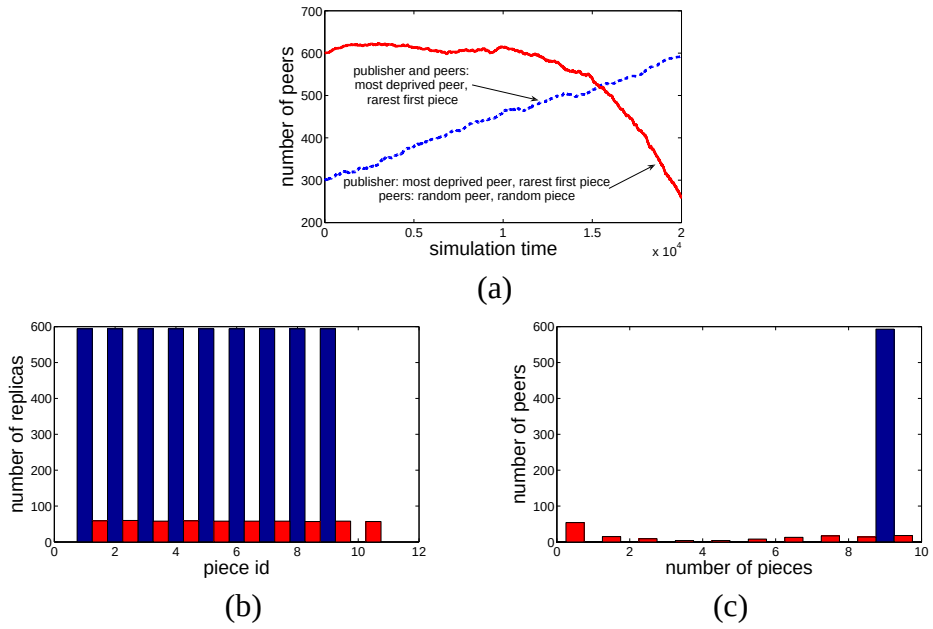

Figure 4. Publishers (but not peers) should adopt most deprived peer, rarest first piece selection: (a) Number of peers versus time; (b) number of replicas of each piece and (c) number of peers that own a given number of pieces. Blue (resp., red) lines and bars correspond to system in which all population (resp., only publisher) adopts most deprived peer and rarest-first piece selection.

In this section we have considered a publisher that adopts the most deprived peer policy. What happens if *all* peers adopt the most deprived policy? Figure 4 illustrates the behavior of the system when all users adopt the most deprived peer selection and rarest-first piece selection (blue line and bars), to be compared with the case where only the publisher adopts such a strategy and peers adopt random peer, random piece selection (red line and bars). The parameter values used to generate Figure 4 are $\lambda = 1$, $\mu = 0.1$, $K = 10$ and $U = 0.1$. Note that if all users adopt most deprived peer selection (blue line and bars), the population grows unboundedly (Figure 4(a)), the number of replicas of piece 10, the missing piece, is zero (Figure 4(b)) and almost all peers have 9 pieces (Figure 4(c)). If only the publisher adopts most deprived peer selection (red line and bars), in contrast, the population decreases with time, all pieces are roughly equally replicated

and peers have roughly the same number of pieces.

6. Throughput Scaling

The objective of this section is to study how the throughput of the system scales with the number of peers. To this goal, we consider a Markovian model of the peer-to-peer system such that every time a peer leaves a new one immediately arrives. This system with a fixed population size, N, and is referred to as a closed system. A detailed description of the Markov Chain and the corresponding Tangram II model are available at `http://www.land.ufrj.br/~sadoc/p2pthr/`.

Open and Closed Systems Although the closed and the open systems have different features, and one needs to beware of taking results that were discovered under the one and applying to the other [Schroeder et al. 2006], the closed system is helpful to give insight on the stability region of the open system. That is because the closed system characterizes the open system in its saturation point, when every departure triggers and arrival, hence the arrival rate equals the system throughput (in an open system, the throughput is smaller than or equal to the arrival rate). We describe some of the differences between the open and the closed system. Then, in the upcoming sections we will explore their similarities.

The states of the closed system are positive recurrent. Since the system is modeled with a Markov chain with finite state space, states are either positive recurrent or transient. This is different from the open system, in which the one club population increases unboundedly if the arrival rate is large enough.

The maximum throughput achieved in the closed system can be larger than in the open system. Consider, for instance, peers and publisher adopting random peer/random piece selection. In a closed system, due to the fact that all states are positive recurrent, there is a non-zero probability that the system will be at states in which a bounded fraction of peers are not in the one club. Therefore, the throughput can be larger than the maximum achievable throughput in the open system, wherein the one club increases unboundedly and the probability that the publisher will serve peers not in the one-club is negligible.

set of states	probability	throughput
non one club	0.000999	1.0227027
one club	0.999000	1.0000000
		1.0000217

Table 2. Non one club/one club

To illustrate this point, consider peers and publisher adopting random peer/random piece selection. Let $U = 1$. In the open system, the throughput is U. If $K = 2$, $\mu = 1000$ and $N = 5$, the throughput is 1.0000217. To see why, note that the system passes through periods at which all peers have all blocks except one, time at which the throughput is roughly 1.0 peers/s. However, when the system is not in the one club mode, the throughput is higher (see Table 2). Therefore, the throughput of the closed system can be higher than the maximum throughput achievable in the open system. In addition, note that the throughput of the closed system depends on μ, whereas the stability region of the open system is invariant to μ when $\gamma = \infty$ (see Proposition 5.1).

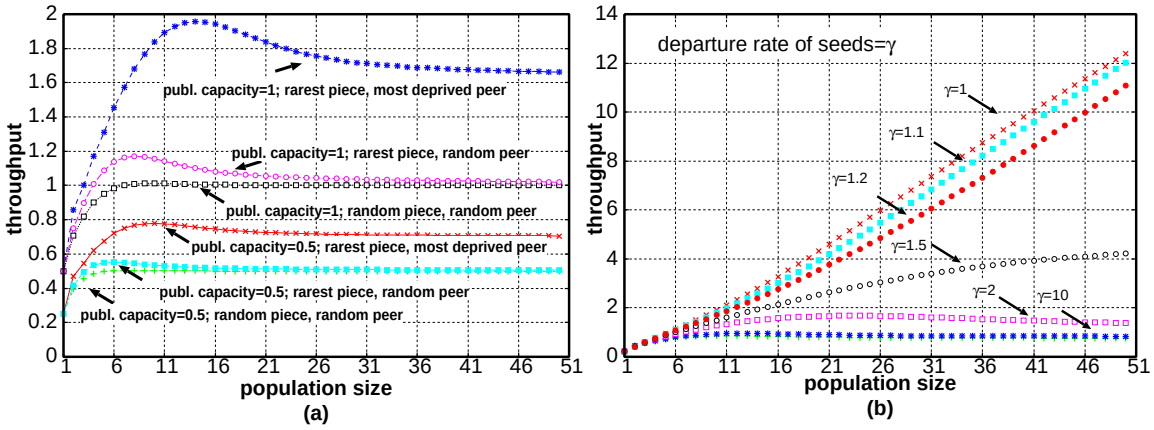

Figure 5. System throughput (μ=1) (a) immediate departures (γ=∞) (b) with lingering (γ<∞)

Experimental Setup In what follows, we make use of our Markov chain model to further study the implications of different policies on the throughput of peer-to-peer systems. To this aim, we consider a file divided into $K = 2$ pieces. Setting the number of pieces to 2 yields a manageable state space of size $(N + 2)!/(2!N!)$ when $\gamma = \infty$, and $(N + 3)!/(3!N!)$ when $\gamma < \infty$. Note that $K = 2$ is the simplest setting for which the peer-to-peer system does not degenerate into a client-server system. Nonetheless, we were able to reproduce and extend most of the insights gained from the open system in this simple setting.

Most Deprived Peer Selection Figure 5 plots the throughput as a function of the population size, for different publisher capacities U (varied between 0.5 and 1 blocks/s) and publisher strategies. Peers follow random peer, random useful piece selection. Figure 5 shows that the throughput obtained when publishers adopt rarest piece/most deprived peer selection is greater than that obtained with each of the other two strategies, which agrees with Proposition 5.3. The figure also shows that for large population sizes, the throughput of rarest first/random peer and random useful piece/random peer are roughly the same.

Altruistic Lingering Figure 5(b) shows results for the case where peers reside in the system as seeds after completing their downloads. The parameters are the same as those used to generate Figure 5. Recall that $1/\gamma$ is the mean time that peers reside in the system after completing their downloads. Note that if $\gamma = 1/\mu = 1$ the throughput increases with the population size and the system is scalable. As γ increases the throughput decreases, $\gamma = \infty$ corresponding to the scenario shown in Figure 5 (see Proposition 5.1).

Publisher Serves only Missing Pieces Figure 6 shows the throughput as a function of the population size when the publisher *serves only missing pieces*. As soon as one piece becomes unavailable among peers, the publisher serves that piece. If all pieces are available among peers, the publisher goes offline from that swarm to save energy and maintenance costs, or to serve other swarms that are in need of bandwidth.

This scenario yields results which, at first glance, are counter-intuitive in light of Proposition 5.3. In particular, consider peers that adopt random peer/random useful piece policy and $\gamma = \infty$. If the publisher only serves missing pieces, the adopting of most

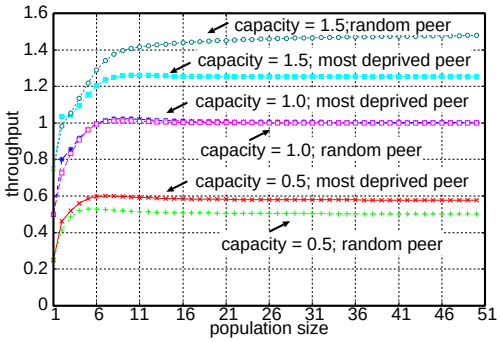

Figure 6. A publisher that only serves missing blocks

deprived peer policy by the publisher is not always beneficial. We have three cases to consider,

First, if the capacity of the publisher is equal to the capacity of the peers it is indifferent whether the publisher adopts most deprived peer or random peer policy. That is because the publisher will stop serving the missing piece in case the peer that receives the missing piece from the publisher remains in the system, and the publisher and the peers have the same capacity, so it is indifferent to whom the publisher transmits the piece. Second, if the publisher has more capacity than the peers, it is beneficial to the publisher to adopt random peer policy rather than most deprived peer policy. That is because if the former case, the publisher can send the missing piece to a peer that will reside in the system after completing the download of that piece. In this case, all pieces will be available among peers, and the publisher will go offline. If the publisher had served a peer that immediately leaves the system after concluding the download of the piece, the server would remain online which would benefit the system throughput, as $U > \mu$; Finally, if the publisher has less capacity than the peers, it is beneficial for the publisher to adopt most deprived peer selection, because as soon as a peer receives a piece from the publisher, such peer will serve other peers more efficiently than the publisher.

7. Conclusion

During the past decade, peer-to-peer systems have received considerable attention for their popularity and scalability. Nonetheless, it has been recently shown that such systems are not always stable [Mathieu and Reynier 2006, Hajek and Zhu 2010, Zhu and Hajek 2011, Menasche et al. 2011]. In this paper we considered different publisher and peer neighbor and piece selection policies. First, we presented a bound on the achievable system throughput when publishers that adopt the most deprived peer/rarest piece policy. The bound is proportional to the number of pieces in the file and simulations provide evidence that it is achievable in practice. Second, we presented numerical results obtained with a Markovian model of a closed system. These results also indicate considerable gains when publishers adopt the most deprived peer/rarest piece policy and when peers reside in the system as seeds after completing their download.

To deal with scalability problems in peer-to-peer systems (which is related to the missing piece syndrome), we consider three possible solutions, 1) *prevention*: to prevent the missing piece syndrome, one could enforce that peers stay in the system after completing their downloads, 2) *avoidance*: to avoid the missing piece syndrome, one could devise dynamic strategies to be implemented by the trackers publishers so as to enforce

that the *one-club* does not grow unboundedly. and 3) *detection*: once the missing piece syndrome is detected, one might increase the capacity of the publisher devoted to a swarm, or perform admission control.

References

Bertsekas, D. and Gallagher, R. (1992). *Data Networks*. Prentice Hall.

Hajek, B. and Zhu, J. (2010). The missing piece syndrome in peer-to-peer communication. In *IEEE ISIT*.

Legout, A., Liogkas, N., Kohler, E., and Zhang, L. (2007). Clustering and sharing incentives in bittorrent systems. In *ACM SIGMETRICS'07*.

Leskela, L., Robert, P., and Simatos, F. (2010). Stability properties of linear file sharing networks. In *Advances in Applied Probability*, volume 42(3), pages 834–854.

Massoulie, L. and Vojnovic, M. (2006). Coupon replication systems. In *SIGMETRICS'06*.

Mathieu, F. and Reynier, J. (2006). Missing piece issue and upload strategies in flashcrowds and p2p-assisted filesharing. In *AICT/ICIW*.

Menasche, D., Rocha, A., de Souza e Silva, E., Leao, R., Towsley, D., and Venkataramani, A. (2010). Estimating self-sustainability in peer-to-peer swarming systems. *Performance Evaluation*, 67:1243–1258.

Menasche, D. S., Aragao Rocha, A. A., de Souza e Silva, E., Towsley, D., and Meri Leao, R. M. (2011). Stability of p2p swarming systems: Implications of peer and piece selection strategies. *SIGMETRICS Perform. Eval. Rev. (to be published, see* `http://www.land.ufrj.br/~sadoc/p2pthr/`*)*.

Menasche, D. S., Rocha, A. A., de Souza e Silva, E. A., Leao, R. M., Towsley, D. F., and Venkataramani, A. (2009a). Modeling Chunk Availability in P2P Swarming Systems. *ACM SIGMETRICS Performance Evaluation Review*, 37:30–32.

Menasche, D. S., Rocha, A. A., Li, B., Towsley, D. F., and Venkataramani, A. (2009b). Content Availability and Bundling in Swarming Systems. In *CoNext*, pages 121–132.

Murai, F. (2011). Sobre dois fenômenos em redes p2p do tipo Bittorrent. *Master Thesis*.

Norros, I., Reittu, H., and Eirola, T. (2009). On the stability of two-chunk file-sharing systems. *Queueing Systems*.

Nunez-Queija, R. and Prabhu, B. (2008). Scaling laws for file dissemination in p2p networks with random contacts. In *IWQoS*, pages 75–79.

Reittu, H. (2009). A stable random-contact algorithm for peer-to-peer file sharing. In *IWS0S*.

Rocha, A. A., Menasche, D. S., Towsley, D. F., and Venkataramani, A. (2009). On P2P systems for enterprise content delivery. In *SBRC*, pages 379–392.

Schroeder, B., Wierman, A., and Harchol-Balter, M. (2006). Open versus closed: A cautionary tale. In *NSDI*.

Yang, X. and De Veciana, G. (2004). Service capacity of peer to peer networks. In *INFOCOM*.

Zhang, B., Borst, S. C., and Reiman, M. I. (2010). Optimal server scheduling in hybrid p2p networks. *Performance Evaluation*, 67(11).

Zhu, J. and Hajek, B. (2011). Stability of peer to peer systems. In *PODC*.