Stable and Efficient Piece-Selection in Multiple Swarm BitTorrent-like Peer-to-Peer Networks

Nouman Khan
ECE Division, EECS Department
University of Michigan
Ann Arbor, MI, 48105, USA
knouman@umich.edu

Mehrdad Moharrami
ECE Division, EECS Department
University of Michigan
Ann Arbor, MI, 48105, USA
moharami@umich.edu

Vijay Subramanian
ECE Division, EECS Department
University of Michigan
Ann Arbor, MI, 48105, USA
vgsubram@umich.edu

***Abstract*—Recent studies have suggested that the BitTorrent's rarest-first protocol, owing to its work-conserving nature, can become unstable in the presence of non-persistent users. Consequently, in any stable protocol, many peers are at some point endogenously forced to hold off their file-download activity. In this work, we propose a tunable piece-selection policy that minimizes this (undesirable) requisite by combining the (work-conserving) rarest-first protocol with only an appropriate share of the (non-work conserving) mode-suppression protocol. We refer to this policy as "Rarest-First with Probabilistic Mode-Suppression" or simply RFwPMS.**

We study RFwPMS under a stochastic model of the BitTorrent network that is general enough to capture multiple swarms of non-persistent users - each swarm having its own altruistic preferences that may or may not overlap with those of other swarms. Using a Lyapunov drift analysis, we show that RFwPMS is provably stable for all kinds of inter-swarm behaviors, and that the use of rarest-first instead of random-selection is indeed more justified. Our numerical results suggest that RFwPMS is scalable in the general multi-swarm setting and offers better performance than the existing stabilizing schemes like mode-suppression.

***Index Terms*—P2P Sharing, Mode-Suppression, Rarest-First.**

I. Introduction

Consider the task of distributing a large file to peers in a peer-to-peer (P2P) network. The file is initially available with a distinguished peer (usually termed as the seed) and each peer can initiate a transfer connection with any other peer [1].

One effective method to perform the above task is to chop the file into a large number of small and roughly equally-sized pieces, and to allow peers to share the pieces with each other. This strategy is the key idea behind the popular "BitTorrent protocol" [2]. Chopping the file allows peers to distribute parts of it before possessing it completely. Such an upload-while-download scheme not only reduces the average file-download time for the peers but also, more importantly, enables the network to scale its throughput with the number of peers. As a result, the BitTorrent protocol has gained a large popularity over the years. Even today, despite the growth of streaming services like Netflix and Youtube, BitTorrent sharing remains a significant source of internet traffic [3]. In the research literature also, the protocol has gained extensive interest. For instance, on the theoretical side, various mathematical models have been studied [1, 4–14], each model providing a high-level abstraction of the detailed workings of the actual protocol.

A common occurrence in BitTorrent-like networks is that a peer usually spends relatively more time downloading the last few pieces of the file. This phenomenon, known as the *delay-in-endgame-mode*, is because the last few pieces are often the rare ones in the network. Inspired by this, Hajek and Zhu [5] studied a stochastic abstraction and showed that an unstructured[1] BitTorrent-like network that employs a work-conserving[2] piece-selection policy (e.g., random-novel (RN), rarest-first (RF)) becomes unstable if the arrival rate of peers exceeds the fixed seed's upload rate and if each peer departs immediately upon completing their own file-download. The cause behind instability is a phenomenon called the *missing piece syndrome* [5], wherein the network converges to, and then cannot escape from, the *one-club scenario*, where a large group of users who possess all but exactly one piece of the file keeps growing.

A. Related Work

Zhu and Hajek [6], in a follow-up to [5], showed that if, after completing their file-download, each peer remains in the system long enough to upload one additional piece, then the network is stable under any positive seed uploading capacity and any peer arrival rate. This demands persistence of peers which might not hold, especially with wireless users who are sensitive to their energy and bandwidth usages.

Norros, Reittu and Eirola [1] proposed the *Enforced Friedman algorithm* in which a peer makes three contacts simultaneously (with replacement) and if there are 'minority pieces' (pieces possessed by exactly one of the three peers), then the peer downloads one of them uniformly at random. The stability was shown for a two-chunk set-up only. Oguz, Anantharam and Norros [9] proved the stability of the Enforced Friedman protocol for multi-chunk systems and also proposed a provably stable improvement, the *Common Chunk protocol*.

Bilgen and Wagner [10] proposed the *Group-Suppression Protocol* in which peers who share the same piece profile are

This work was funded in part by NSF via EPCN 1608361 (V. Subramanian) and EARS 1516075 (N. Khan and V. Subramanian), and the Rackham Predoctoral Fellowship from the University of Michigan (M. Moharrami).

[1] Each peer contacts an other peer uniformly at random.

[2] For a work-conserving piece-selection policy, a piece transfer always happens if the uploading peer has a piece that the downloading peer needs.

defined as a *group* and the group with the largest population is defined as the *largest club*. Assuming that a peer in the largest club uploads only to peers holding a greater number of pieces than it does, stability was shown in the two-chunk case only.

More recently, Reddyvari, Parag and Shakkottai [11] proposed the *Mode-Suppression* (MS) protocol. Here the transfer of pieces in the mode (present with the most number of peers) is prohibited, except when all pieces are in the mode, and a random-novel piece not in the mode (if any) is sent. Noting the explicit non-work-conserving nature of the protocol, developing a stable protocol with the smallest mean sojourn time[3] is an interesting design question, which we seek to answer.

The literature has also considered bundling different swarms together in the same network and then allowing content-sharing across them. This was proposed by Zhou, Ioannidis and Massoulié [7] with the claim that such "universal swarms" can increase the stability region of a BitTorrent network. Then Zhu, Ioannidis and Hegde [8] formally characterized the stability region of such networks when a work-conserving piece-selection policy is used. The stability region is indeed larger than in the single-swarm setting, however, yet again, it doesn't hold for all arrival rates. Developing a stable and efficient protocol for the multiple swarm (multi-swarm) setting is another interesting design question which we seek to answer.

B. Contribution

We make the first effort to abridge the performance gaps of the previous stable policies, and to respond to the findings and recommendations of [15]. We do that by proposing a simple and provably stable variant of the BitTorrent's original rarest-first protocol that includes only an appropriate amount of (probabilistic) mode-suppression; we call this protocol *Rarest-First with Probabilistic Mode-Suppression* (*RFwPMS*). To the best of our knowledge, we are the first to propose a RF-based piece-selection policy that is stable for all arrival rates. Then, numerically we show that RFwPMS outperforms MS [11].

Compared with [7], we extend RFwPMS to a multi-swarm model where peers arrive at any (positive) arrival rate and their caches are empty upon arrival. With regard to inter-swarm cooperation, we make the general assumption that each swarm has its own altruistic preferences. Leveraging the intuition that as long as each swarm is stable, the resulting network will be stable, we carefully establish the stability of RFwPMS using a more general version of the Lyapunov function from [11]. As the single-swarm model is a special case of the multi-swarm model, in the paper we only study the multi-swarm model.

The remainder of the paper is organized as follows. In Section II, we introduce the multi-swarm model which is built upon the model in [8]. Section III presents the main stability theorem along with the preliminary steps needed for the detailed proof in the appendix - Section VII. Section IV discusses the working of RFwPMS as well as a few types of inter-swarm behaviors that can be relevant in specific P2P environments. Various numerical results on stability, scalability and performance of RFwPMS are presented in Section V. Finally, our concluding remarks are in Section VI.

II. System Model

The model consists of a **master-file** which is divided into at least two equally-sized pieces. Let $\mathcal{F} = [K]^4$ denote the set of all the pieces ($K \geqslant 2$). There is a distinguished peer, the **seed**, which holds this master-file $\mathcal{F}$ and stays in the network indefinitely. We define **file-**W as any non-empty subset of the master-file, thus, $W \neq \phi$ and $W \subseteq \mathcal{F}$. The number of pieces in file-W is denoted by K_W, i.e., $K_W = |W|$. With this definition of file, **swarm**-W is defined as the set of peers who are primarily interested in downloading (pieces of) file-W. We note that peers entering the network can be interested in any file, i.e., the files need not be disjoint subsets of $\mathcal{F}$. Besides their primary interest in file-W, swarm-W peers may also have a secondary preference for some other pieces of the master-file. Thus, the set of pieces that a swarm-W peer can download during its stay in the network is given by $\mathcal{F}_W$ where $W \subseteq \mathcal{F}_W \subseteq \mathcal{F}$. It is assumed that a swarm-W peer enters the network according to a Poisson process of rate $\lambda_W > 0$, independent of other swarms. Let $\mathcal{W}$ denote the set of all swarms that join the network and let $\boldsymbol{\lambda} = (\lambda_W : W \in \mathcal{W})$ denote the vector of arrival rates. We now list three important assumptions of the model: *a) Empty Cache Upon Arrival*: Each peer maintains a **cache** to store the pieces it downloads. The cache is empty upon arrival and has a capacity of $|\mathcal{F}_W|$ pieces. The part of the cache that is devoted for the pieces of secondary interest is called the **excess-cache** (where pieces from the set $\mathcal{F}_W \backslash W$ are stored); *b) Ally-Swarms*: While peers interested in the same file exchange pieces with each other, they may or may not prefer to collaborate with peers who are interested in other files. Thus, each swarm has an associated set of ally-swarms. Formally, the **ally-set** of swarm-W, denoted by $\mathcal{W}_W$, is a non-empty subset of $\mathcal{W}$ that consists of swarm-W as well as any other swarms to which its peers can upload pieces; and *c) Non-persistent Peers*: Once a peer finishes downloading their file of (primary) interest, they leave the network immediately.

A. State

Peers can be classified into types according to *(a)* the swarm they belong to and *(b)* the set of pieces in their cache. Hence, a peer in swarm W holding $S \subseteq \mathcal{F}_W$ is denoted to be of type (W, S). We denote the number of (W, S)-type peers at any time $t \geqslant 0$ by $x_W^{(S)}(t) \in \mathbb{Z}_+ \triangleq \{0, 1, \ldots\}$. The state of the network is then given by

$$\mathbf{x}(t) = (x_W^{(S)}(t) : W \in \mathcal{W}, S \subseteq \mathcal{F}_W, \text{ and } W \backslash S \neq \varnothing). \quad (1)$$

Note that as a result of aforementioned assumptions, the cache-profile S of a swarm-W peer always satisfies the conditions $S \subseteq \mathcal{F}_W$ and $W \backslash S \neq \varnothing$. For the sake of brevity, from hereon,

[3] The time a peer remains in the system collecting all the pieces of the file.

[4] For $m, n \in \mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ and $m < n$, we use the standard notation $[m, n] := \{m + 1, \ldots, n\}$ and $[n]$ for $[0, n]$.

we will omit writing these two conditions. The **population** of swarm-W, i.e., the number of swarm-W peers at time $t \geqslant 0$ is then given by

$$|\mathbf{x}(t)|_W \triangleq \sum_S x_W^{(S)}(t). \tag{2}$$

Similarly, the total number of peers at time $t \geqslant 0$ is given by

$$|\mathbf{x}(t)| \triangleq \sum_{W \in \mathcal{W}} |\mathbf{x}(t)|_W. \tag{3}$$

From hereon, for brevity we will write $\mathbf{x}(t)$ as $\mathbf{x}$.

B. Peer-Contact and Piece-Selection Policies

1) Peer-Contact Policy: Consistent with the stochastic models of [1, 4–11], we assume that the network employs random peer-contacts. Each peer contacts some peer chosen uniformly at random from all other peers (excluding the seed), to upload (i.e. push) a piece to it; such contacts are called push-contacts. We assume that the seed makes independent random push-contacts according to a Poisson process of rate $U > 0$ whereas each normal peer does this independently according to a Poisson process of rate $\mu > 0$. Transfer of the piece is assumed to occur instantaneously with the push-contact.

2) Piece-Selection Policy (RFwPMS): If at time $t \geqslant 0$, a (V, T)-peer push-contacts a (W, S)-peer, then the piece-selection policy chooses the piece to upload from $(T \cap \mathcal{F}_W)\backslash S$. To describe RFwPMS, some definitions (*a'la* [11]) are needed.

Definition 1. *The **frequency** of a piece i in swarm-W's population is denoted by $\pi_W^{(i)}(\mathbf{x})$ and is defined as follows:*

$$\pi_W^{(i)}(\mathbf{x}) \triangleq \begin{cases} \frac{1}{|\mathbf{x}|_W} \sum\limits_{S: i \in S} x_W^{(S)}(t) & \textit{if } |\mathbf{x}|_W > 0, \\ 0 & \textit{if } |\mathbf{x}|_W = 0. \end{cases}$$

The maximum and minimum piece frequencies in swarm-W's population are denoted by $\overline{\pi}_W(\mathbf{x})$ and $\underline{\pi}_W(\mathbf{x})$ respectively,

$$\overline{\pi}_W(\mathbf{x}) \triangleq \max_{i \in W} \pi_W^{(i)}(\mathbf{x}) \quad \textit{and} \quad \underline{\pi}_W(\mathbf{x}) \triangleq \min_{i \in W} \pi_W^{(i)}(\mathbf{x}).$$

Importantly, both are computed over W instead of $\mathcal{F}_W$.

Definition 2. *The total number of copies of the pieces of file-W that are present with swarm-W peers is given by*

$$P_W(\mathbf{x}) \triangleq \sum_{i \in W} \pi_W^{(i)}(\mathbf{x}) |\mathbf{x}|_W.$$

Definition 3. *The **mismatch** in swarm-W at state $\mathbf{x}$ is defined as $(\overline{\pi}_W - \underline{\pi}_W)|\mathbf{x}|_W$.*

Definition 4. *The set of **rare pieces** in swarm-W's population, denoted by $R_W(\mathbf{x})$, is defined as follows:*

$$R_W(\mathbf{x}) \triangleq \begin{cases} \{i \in W : \pi_W^{(i)}(\mathbf{x}) < \overline{\pi}_W(\mathbf{x})\} & \textit{if } \overline{\pi}_W \neq \underline{\pi}_W, \\ W & \textit{if } \overline{\pi}_W = \underline{\pi}_W. \end{cases}$$

Definition 5. *The set of **non-rare pieces** in swarm-W is denoted by $\overline{R}_W(\mathbf{x})$ and is given by $W \backslash R_W$.*

Definition 6. *The set of **extra pieces** associated with swarm-W is given by $\mathcal{F}_W \backslash W$.*

We now list the rules of RFwPMS for a possible piece transfer from the (V, T)-peer to the (W, S)-peer:
a) *Ally Check*: If swarm-W is not considered an ally by swarm-V, i.e., if $W \notin \mathcal{W}_V$, then no piece is transferred to the (W, S)-peer. Otherwise, items (b), (c) and (d) are carried out in the mentioned order.
b) *Download of a Rare Piece*: If $(T \cap R_W(\mathbf{x}))\backslash S \neq \varnothing$, then the (V, T)-peer uploads the rarest piece it can offer, i.e., a piece chosen uniformly at random from the set

$$\underline{R}_{(T,W,S)}(\mathbf{x}) \triangleq \arg \min_{q \in (T \cap R_W(\mathbf{x}))\backslash S} \pi_W^{(q)}(\mathbf{x}).$$

c) *Download of a Non-rare Piece*: If $(T \cap R_W(\mathbf{x}))\backslash S = \varnothing$ and $(T \cap W)\backslash S \neq \varnothing$, then the (V, T)-peer chooses some non-rare piece $n \in (T \cap W)\backslash S$ uniformly at random, and decides to upload it with success probability given by

$$\zeta_W(\mathbf{x}) \triangleq \exp\left(-\frac{(\overline{\pi}_W - \underline{\pi}_W)|\mathbf{x}|_W}{\beta_W K_W} |\mathbf{x}|^{\alpha_W}\right) \mathbb{1}_{\{\beta_W > 0\}}, \tag{4}$$

where $\alpha_W > 0$ and $\beta_W \geqslant 0$ are constants. We refer to $\zeta_W(\mathbf{x})$ as the *non-rares' sharing factor*.
d) *Download of an Extra Piece*: If, from rules (b) and (c), no piece of file-W could be uploaded to the (W, S)-peer, then the (V, T)-peer uploads an extra piece chosen uniformly at random from the set $(T \cap \mathcal{F}_W)\backslash(S \cup W)$.

Note that when there is no rare piece at offer, a non-rare piece $n \in (T \cap W)\backslash S$ is transferred only *probabilistically*, so RFwPMS is not a work-conserving scheme. In order to extend our rules to the seed, we assume that the seed is a $(\dagger, \mathcal{F})$-type peer where $\dagger$ takes the swarm identity of whichever peer it push-contacts. Thus, the set of rare pieces when the seed push-contacts a normal peer is given by

$$\underline{R}_{(\dagger,W,S)}(\mathbf{x}) = \arg \min_{q \in R_W(\mathbf{x})\backslash S} \pi_W^{(q)}(\mathbf{x}).$$

C. Process Description

For the sake of notational simplicity, from hereon, we will write $f(\mathbf{x})$ as f when the dependence on state $\mathbf{x}$ is clear. At any given current state $\mathbf{x}$, the next state of the network depends solely on $\mathbf{x}$ as the piece-selection policy solely depends on $\mathbf{x}$. Hence, the evolution of the network described by the process $\{\mathbf{x}(t) : t \geqslant 0\}$ is a continuous-time, time-homogeneous and irreducible (easily shown) Markov chain with state space,

$$\mathcal{S} \triangleq \mathbb{Z}_+^{\sum_{W \in \mathcal{W}} |\{S : S \subseteq \mathcal{F}_W \text{ and } W \backslash S \neq \varnothing\}|}. \tag{5}$$

We now describe the generator matrix Q of the process $\{\mathbf{x}(t) : t \geqslant 0\}$ by listing its positive entries. Given a state $\mathbf{x}$, the different kinds of transitions possible are:
a) *Arrival of an Empty-peer*: Recall that a swarm-W peer with an empty cache enters the network according to a Poisson process of rate λ_W. This results in a unit increase in the number of swarm-W peers with no pieces, i.e., the new state is $\mathbf{x} + \mathbf{e}_{(W,\varnothing)}$ ($\mathbf{e}_{(\cdot,\cdot)}$ is the unit vector in the direction of $(\cdot,\cdot)$-axis) and the corresponding transition rate is given by

$$q(\mathbf{x}, \mathbf{x} + \mathbf{e}_{(W,\varnothing)}) = \lambda_W. \tag{6}$$

b) *Download of a Rare piece*: Let $r \in R_W$ denote some rare piece in swarm-W. The second type of transition is when a (W,S)-peer missing piece r, downloads it. The necessary condition for this is that the (W,S)-peer gets push-contacted by a peer holding piece r, which includes the seed and all (V,T)-peers such that $W \in \mathcal{W}_V$ and $r \in T$. By the properties of Poisson processes, the rate at which the seed push-contacts a (W,S)-peer is $U \times x_W^{(S)}/|\mathbf{x}|$, and the rate at which a (V,T)-peer push-contacts a (W,S)-peer is $\mu x_V^{(T)} \times x_W^{(S)}/(|\mathbf{x}|-1)$. Let $\xi(\mathbf{x}) = (|\mathbf{x}|-1)/|\mathbf{x}|$ and define

$$\Gamma_W^{(r)} \triangleq U + \mu \sum_{(V,T):W\in\mathcal{W}_V, r\in T} \frac{x_V^{(T)}}{\xi}.$$

Then, the aggregate rate at which a possible source of piece r, push-contacts the (W,S)-peer is given by $x_W^{(S)}/|\mathbf{x}| \times \Gamma_W^{(r)}$. After the (W,S)-peer has downloaded piece r, depending on S, it will either remain in the network or leave it immediately. Therefore, we have two cases:

i) if $W \backslash S \supsetneq \{r\}$, the peer stays in the system; the resulting state is $\mathbf{x} - \mathbf{e}_{(W,S)} + \mathbf{e}_{(W,S\cup\{r\})}$. For brevity, let us introduce the notation $q(\mathbf{x}, \mathbf{x} - \mathbf{e}_{(W,S)} + \mathbf{e}_{(W,S\cup\{i\})}) = q_{(W,S)}^{(i+)}$ for any S that satisfies $W \backslash S \supsetneq \{i\}$. Then, $q_{(W,S)}^{(r+)}$ is given by

$$\frac{x_W^{(S)}}{|\mathbf{x}|}\left(\frac{U \mathbb{1}_{\{r\in \underline{R}_{(\dagger,W,S)}\}}(\mathbf{x})}{|\underline{R}_{(\dagger,W,S)}|} + \mu \sum_{\substack{V:W\in\mathcal{W}_V,\\ T:r\in T}} \frac{x_V^{(T)} \mathbb{1}_{\{r\in \underline{R}_{(T,W,S)}\}}(\mathbf{x})}{\xi |\underline{R}_{(T,W,S)}|}\right), \tag{7}$$

where the two indicator terms ensure that piece r is transferred *only if it is the rarest of all the available rare pieces*.

When the piece distribution in swarm-W is uniform i.e., $\overline{\pi}_W = \underline{\pi}_W$, by definition, $R_W = W$ and the above transition rate takes a more tractable form. Therefore, for each swarm-W, we partition the state space into two regions, namely $\mathcal{S}_W^1$ and $\mathcal{S}_W^2$, where $\mathcal{S}_W^1 = \{\mathbf{x} : R_W(\mathbf{x}) \subsetneq W\}$, and $\mathcal{S}_W^2 = \{\mathbf{x} : R_W(\mathbf{x}) = W\}$. In the case when $\mathbf{x} \in \mathcal{S}_W^2$, both the indicator terms evaluate to 1. Thus, for a state $\mathbf{x} \in \mathcal{S}_W^2$, if we upper bound $|\underline{R}_{(\dagger,W,S)}|$ and $|\underline{R}_{(T,W,S)}|$ by K_W, then the transition rate $q_{(W,S)}^{(r+)}$ is lower bounded by

$$\frac{x_W^{(S)}}{|\mathbf{x}|} \frac{\Gamma_W^{(r)}}{K_W} = \frac{|\mathbf{x}|_W}{|\mathbf{x}|} \frac{x_W^{(S)}}{|\mathbf{x}|_W} \frac{\Gamma_W^{(r)}}{K_W}. \tag{8}$$

ii) If $W \backslash S = \{r\}$, the only piece of file-W that is missing with the (W,S)-peer is piece r. Consequently, for both $\mathbf{x} \in \mathcal{S}_W^1$ and $\mathbf{x} \in \mathcal{S}_W^2$, piece r is the rarest piece transferable to the (W,S)-peer, which leaves the network immediately upon downloading it. The resulting state is $\mathbf{x} - \mathbf{e}_{(W,S)}$. For brevity, let us introduce the notation $q(\mathbf{x}, \mathbf{x} - \mathbf{e}_{(W,S)}) = q_{(W,S)}^{(i-)}$ for any S that satisfies $W \backslash S = \{i\}$. Then $q_{(W,S)}^{(r-)}$ is given by

$$\frac{x_W^{(S)}}{|\mathbf{x}|} \Gamma_W^{(r)} = \frac{|\mathbf{x}|_W}{|\mathbf{x}|} \frac{x_W^{(S)}}{|\mathbf{x}|_W} \Gamma_W^{(r)}. \tag{9}$$

c) *Download of a Non-rare Piece*: The third type of transition is when a (W,S)-peer missing piece $n \in \overline{R}_W$, downloads it. We can upper bound both $q_{(W,S)}^{(n+)}$ and $q_{(W,S)}^{(n-)}$ by

$$\frac{x_W^{(S)}}{|\mathbf{x}|} \left(U + \mu \frac{|\mathbf{x}|}{\xi}\right) \zeta_W(\mathbf{x}). \tag{10}$$

Note that a non-rare piece is downloaded only when $\mathbf{x} \in \mathcal{S}_W^1$.

d) *Download of an Extra Piece*: Recall that extra pieces are preferred only when no pieces from the file of interest are transferable. We shall soon see that the stability of RFwPMS does not depend on the download of extra pieces. Consequently, we skip listing the associated rates.

III. Stability Analysis

In this section, we present the main result on the stability of RFwPMS. The proof is established using the Foster-Lyapunov theorem [16, 17] which is restated below.

Proposition 1. *Let $\{\mathbf{x}(t) : t \geqslant 0\}$ be a continuous-time, time-homogeneous and irreducible Markov chain with state space $\mathcal{S}$ and generator matrix Q. If there exist a non-negative function $V : \mathcal{S} \to \mathbb{R}_+$, an $\epsilon' > 0$, a finite set $\mathcal{A}$ and a finite constant B such that $\{\mathbf{x} : V(\mathbf{x}) \leqslant C\}$ is finite for all $C \in \mathbb{R}_+$, and the expected unit-transition drift $QV(\mathbf{x})$ is upper bounded as*

$$QV(\mathbf{x}) \leqslant -\epsilon' \mathbb{1}_{\{\mathbf{x}\in\overline{\mathcal{A}}\}}(\mathbf{x}) + B\mathbb{1}_{\{\mathbf{x}\in\mathcal{A}\}}(\mathbf{x}),$$

then the process $\{\mathbf{x}(t) : t \geqslant 0\}$ is positive recurrent; the expected unit-transition drift $QV(\mathbf{x})$ is given by

$$QV(\mathbf{x}) \triangleq \sum_{\mathbf{y}\in\mathcal{S},\mathbf{y}\neq\mathbf{x}} q(\mathbf{x},\mathbf{y}) \left(V(\mathbf{y}) - V(\mathbf{x})\right). \tag{11}$$

Our main stability result is the following theorem.

Theorem 1. *For the multi-swarm model with non-persistent peers in section II, RFwPMS is stable over the parameter region $\mu > 0$, $U > 0$, $\boldsymbol{\lambda} > 0$, and $K_W \geqslant 2$ for all $W \in \mathcal{W}$.*

Sketch of the proof. Our proof uses a more general version of the Lyapunov function used in [11]. The key ideas are
(i) show that the Lyapunov drift obtained from downloading rare pieces using rarest-first is upper bounded by the Lyapunov drift obtained from downloading rare pieces using random-selection (this is proved in Lemma 3 in Section VII); and
(ii) allow the (probabilistic) download of non-rare pieces only over a finite and bounded but sufficiently large population of the network (this holds by our choice of $\zeta_W(\mathbf{x})$).
Note that mode-suppression is covered as we allow $\beta_W = 0$.

The Lyapunov function we use is given by

$$V(\mathbf{x}) = \sum_{W\in\mathcal{W}} V_W(\mathbf{x}), \tag{12}$$

where $V_W(\mathbf{x}) \triangleq \sum_{i\in W} \left((\overline{\pi}_W - \pi_W^{(i)})|\mathbf{x}|_W\right)^2 +$

$$C_W^{(1)} \left(1 - \overline{\pi}_W\right) |\mathbf{x}|_W + C_W^{(2)} \left(M_W - P_W\right)^+, \tag{13}$$

with suitably large constants $C_W^{(1)}, C_W^{(2)} \in \mathbb{R}_+$ and $M_W \in \mathbb{Z}_+$. Note that $|\mathbf{x}| \to \infty$ only if $|\mathbf{x}|_W \to \infty$ for some $W \in \mathcal{W}$.

Then from Lemma 2, it follows that $V(\mathbf{x}) \to \infty$ as $|\mathbf{x}| \to \infty$. Consequently, the set $\{\mathbf{x} : V(\mathbf{x}) \leqslant C\}$ is finite for all $C \in \mathbb{R}_+$.

To evaluate the expected unit-transition drift for any state $\mathbf{x}$, we first evaluate the potential change for each possible transition. Note that, with our choice of V, any transition that occurs in swarm-W will only affect the term V_W. We have:

a) *Arrival of an Empty-peer*: The arrival of a $(W, \varnothing)$-peer results in a unit increase in $|\mathbf{x}|_W$ but does not affect the number of copies of each piece $i \in W$. Therefore,

$$V(\mathbf{x} + \mathbf{e}_{(W,\varnothing)}) - V(\mathbf{x}) = C_W^{(1)}. \tag{14}$$

b) *Download of a Rare piece*: Let $r \in R_W$ denote some rare piece. The potential change associated with the download of piece r is similar to that in [11], and depends on whether the current piece distribution is uniform or non-uniform. Therefore, we have two cases:

Case 1 - $\mathbf{x} \in \mathcal{S}_W^1$: Suppose piece $r \in R_W$ is downloaded by a (W, S)-peer. *i)* if $W \backslash S \supsetneq \{r\}$, the peer stays in the network upon downloading r. In this case, the associated potential change $V(\mathbf{x} - \mathbf{e}_{(W,S)} + \mathbf{e}_{(W,S\cup\{r\})}) - V(\mathbf{x})$ is given by

$$\Psi_W^{(r)}(\mathbf{x}) = \underbrace{1 - 2\left(\overline{\pi}_W - \pi_W^{(r)}\right)|\mathbf{x}|_W}_{\triangleq \psi_W^{(r)}(\mathbf{x})} - \underbrace{C_W^{(2)} \mathbb{1}_{\{M_W > P_W\}}(\mathbf{x})}_{\triangleq I_W^{(1)}(\mathbf{x}) \geqslant 0}. \tag{15}$$

Since $\mathbf{x} \in \mathcal{S}_W^1$, we have $(\overline{\pi}_W - \pi_W^{(r)})|\mathbf{x}|_W \geqslant 1$. Hence, $\psi_W^{(r)}(\mathbf{x}) \leqslant -1$ and the overall potential change $\Psi_W^{(r)}(\mathbf{x})$ is negative. *ii)* If $W \backslash S = \{r\}$, then the peer departs the network upon downloading r. In this case, the associated potential change $V(\mathbf{x} - \mathbf{e}_{(W,S)}) - V(\mathbf{x})$ can be upper bounded by

$$\psi_W^{(r)}(\mathbf{x}) + \underbrace{C_W^{(2)} (K_W - 1) \mathbb{1}_{\{M_W + K_W - 1 > P_W\}}(\mathbf{x})}_{\triangleq I_W^{(2)}(\mathbf{x}) \geqslant 0}, \tag{16}$$

Case 2 - $\mathbf{x} \in \mathcal{S}_W^2$: When $\mathbf{x} \in \mathcal{S}_W^2$, the download of piece $r \in R_W$ disturbs the uniform distribution with r attaining the highest frequency. *i)* If piece r is downloaded without an accompanying peer departure, the associated potential change $V(\mathbf{x} - \mathbf{e}_{(W,S)} + \mathbf{e}_{(W,S\cup\{r\})}) - V(\mathbf{x})$ is given by

$$K_W - 1 - C_W^{(1)} - I_W^{(1)}(\mathbf{x}), \tag{17}$$

which is made negative by choosing $C_W^{(1)} > K_W - 1$. *ii)* If the download of piece r is accompanied with the departure of the downloading peer, the associated potential change $V(\mathbf{x} - \mathbf{e}_{(W,S)}) - V(\mathbf{x})$ can be upper bounded by

$$K_W - 1 - C_W^{(1)} + I_W^{(2)}(\mathbf{x}). \tag{18}$$

c) *Download of a Non-Rare Piece*: As stated earlier, the download of a non-rare piece can occur only if $\mathbf{x} \in \mathcal{S}_W^1$. The potential changes $V(\mathbf{x} - \mathbf{e}_{(W,S)} + \mathbf{e}_{(W,S\cup\{n\})}) - V(\mathbf{x})$ and $V(\mathbf{x} - \mathbf{e}_{(W,S)}) - V(\mathbf{x})$ associated with the download of a non-rare piece $n \in \overline{R}_W$ can be upper bounded by

$$K_W^2 \left(1 + 2|\mathbf{x}|_W + C_W^{(2)}\right). \tag{19}$$

d) *Download of an Extra Piece*: It can be observed that for any swarm-W, no potential change is induced by the download of its extra pieces.

Having listed all the potential changes, we can now proceed to evaluate the expected unit-transition drift $QV(\mathbf{x})$. Note that for any state $\mathbf{x}$, we can write

$$QV(\mathbf{x}) = \sum_{W \in \mathcal{W}} QV_W^{(+)}(\mathbf{x}) + QV_W^{(R)}(\mathbf{x}) + QV_W^{(\overline{R})}(\mathbf{x}), \tag{20}$$

where $QV_W^{(+)}(\mathbf{x})$, $QV_W^{(R)}(\mathbf{x})$ and $QV_W^{(\overline{R})}(\mathbf{x})$ are contributions to the unit-transition drift from arrival of swarm-W peer, download of a rare piece in swarm-W and download of a non-rare piece in swarm-W, respectively. Next we evaluate each of these three terms below:

Arrival of an Empty-peer: Using (6) and (14), we have $QV_W^{(+)}(\mathbf{x}) = \lambda_W C_W^{(1)}$. We then upper bound $\sum_{W \in \mathcal{W}} QV_W^{(+)}(\mathbf{x})$ by

$$|\boldsymbol{\lambda}| C^{(1)}, \tag{21}$$

where $|\boldsymbol{\lambda}| \triangleq \sum_{W \in \mathcal{W}} \lambda_W$ and $C^{(1)} \triangleq \sum_{W \in \mathcal{W}} C_W^{(1)}$.

The terms $QV_W^{(R)}(\mathbf{x})$ and $QV_W^{(\overline{R})}(\mathbf{x})$ depend on whether $\mathbf{x}$ is in $\mathcal{S}_W^1$ or $\mathcal{S}_W^2$. Therefore, we have two cases.

Case 1 - $\mathbf{x} \in \mathcal{S}_W^1$: We first start with $QV_W^{(\overline{R})}(\mathbf{x})$. Let $\omega > 0$ be any real positive number, then by Lemma 4 in Section VII, we can upper bound $QV_W^{(\overline{R})}(\mathbf{x})$ by

$$\frac{|\mathbf{x}|_W}{|\mathbf{x}|} \mathcal{O}\left(|\mathbf{x}|_W^{-\omega}\right). \tag{23}$$

To compute $QV_W^{(R)}(\mathbf{x})$, we can write

$$QV_W^{(R)}(\mathbf{x}) = QV_W^{(R+)}(\mathbf{x}) + QV_W^{(R-)}(\mathbf{x}), \tag{24}$$

where $QV_W^{(R+)}(\mathbf{x})$ and $QV_W^{(R-)}(\mathbf{x})$ are contributions to the unit-transition drift from download of a rare piece without peer departure and with peer departure, respectively. The transitions that correspond to $QV_W^{(R+)}(\mathbf{x})$ include each piece $r \in R_W$ being downloaded by a (W, S)-peer such that $W \backslash S \supsetneq \{r\}$. Using (7), (15) and Lemma 3, we can upper bound $QV_W^{(R+)}(\mathbf{x})$ by

$$\frac{|\mathbf{x}|_W}{|\mathbf{x}|} \sum_{r \in R_W} \frac{\Gamma_W^{(r)}}{K_W} \left(1 - \pi_W^{(r)} - \gamma_W^{(r)}\right) \Psi_W^{(r)}, \tag{25}$$

where $\gamma_W^{(r)} \triangleq \sum_{S: W \backslash S = \{r\}} x_W^{(S)} / |\mathbf{x}|_W$ is the fraction of swarm-W peers who have all the pieces of file-W except r. The transitions that correspond to $QV_W^{(R-)}(\mathbf{x})$ include each piece $r \in R_W$ being downloaded by a (W, S)-peer such that $W \backslash S = \{r\}$. Using (9) and (16), we can upper bound $QV_W^{(R-)}(\mathbf{x})$ by

$$\frac{|\mathbf{x}|_W}{|\mathbf{x}|} \sum_{r \in R_W} \gamma_W^{(r)} \Gamma_W^{(r)} \left(\psi_W^{(r)}(\mathbf{x}) + I_W^{(2)}(\mathbf{x})\right). \tag{26}$$

Combining (25) and (26), we upper bound $QV_W^{(R)}(\mathbf{x})$ by

$$\frac{|\mathbf{x}|_W}{|\mathbf{x}|} \sum_{r \in R_W} \frac{\Gamma_W^{(r)}}{K_W} \bigg(\psi_W^{(r)}(\mathbf{x}) \left(1 - \pi_W^{(r)}\right)$$

$$-\left(1-\pi_W^{(r)}-\gamma_W^{(r)}\right)I_W^{(1)}(\mathbf{x})+\gamma_W^{(r)}K_W I_W^{(2)}(\mathbf{x})\Bigg). \quad (27)$$

Case 2 - $\mathbf{x} \in \mathcal{S}_W^2$: If $\mathbf{x} \in \mathcal{S}_W^2$, then $R_W = W$ and there are no non-rare pieces to download. Therefore,

$$QV_W^{(\overline{R})}(\mathbf{x}) = 0. \quad (28)$$

For $QV_W^{(R)}(\mathbf{x})$, we can use similar techniques as in Case 1 and show that it is upper bounded by

$$\frac{|\mathbf{x}|_W}{|\mathbf{x}|}\sum_{r\in W}\frac{\Gamma_W^{(r)}}{K_W}\Bigg(\left(K_W-1-C_W^{(1)}\right)(1-\overline{\pi}_W)-$$
$$\left(1-\overline{\pi}_W-\gamma_W^{(r)}\right)I_W^{(1)}(\mathbf{x})+\gamma_W^{(r)}K_W I_W^{(2)}(\mathbf{x})\Bigg). \quad (29)$$

Finally, using (21), (27), (23), (29), (28) and $\widetilde{QV}_W(\mathbf{x})$ from (29), we can upper bound the unit-transition drift $QV(\mathbf{x})$ by

$$\sum_{W\in\mathcal{W}}\frac{|\mathbf{x}|_W}{|\mathbf{x}|}\widetilde{QV}_W(\mathbf{x}).$$

The rest of the proof is in the appendix, i.e., Section VII. □

IV. Discussion

A. Sharing vs Suppression Trade-off for Non-Rare Pieces

As indicated earlier in Section I, the mode-suppression protocol [11] strictly forbids the download of non-rare pieces. While this maintains a uniform piece distribution, there is an accompanying undesirable effect: no piece is transferred in all those contacts in which only the non-rare pieces are offered to the contacted peer. As shown in Section V, this incurs a high penalty on file-delivery time during a flash-crowd[5]. Besides flash-crowds, under normal operating conditions also, completely suppressing transfers of non-rare pieces is unnecessary and, as indicated in [11], a trade-off exists between their suppression and sharing. RFwPMS allows for tuning this trade-off via β_W. Note that by choosing α_W sufficiently small, $\zeta_W(\mathbf{x})$ effectively becomes a function of the ratio of the mismatch to the number of pieces of file-W. The higher this ratio, the lesser the likelihood that the non-rare piece gets replicated; the choice of the ratio instead of just the mismatch matches the intuition that a file with large number of pieces should allow more sharing of non-rare pieces as opposed to a file with smaller number of pieces.

Taking $\beta_W \to \infty$, RFwPMS converges to rarest-first which is unstable, whereas taking $\beta_W \to 0$, RFwPMS converges to mode-suppression [11]. Thus, by choosing α_W sufficiently small, and β_W appropriately, we expect to find the right trade-off between sharing and suppression of the non-rare pieces. Via numerical simulations, we find that choosing β_W close to 1.5 appears to minimize the expected sojourn time.

By Lemma 4, RFwPMS is stable for any non-rares' sharing factor that is $\mathcal{O}(|\mathbf{x}|^{-2-\omega})$ for some $\omega > 0$. However, the specific choice of $\zeta_W(\mathbf{x})$ in (4) is presented as it allows the transfer of non-rare pieces even for large network population (by choosing $\alpha_W > 0$ to be small enough). We also remark that RFwPMS like mode-suppression is a centralized scheme as it requires the knowledge of piece distribution in the network; we assume that in practice, peers can keep such estimates either via gossiping or via a centralized tracker.

B. Rarest-First vs Random-Selection for Rare Pieces

Another notable observation about MS [11] is that it performs random-selection on the set of available rare pieces. We assert that knowing the distribution should allow for more advantage than just random-selection, and that a load-balancing scheme like rarest-first will reduce the duration of the transient phase. While this is intuitively clear, we provide theoretical proof for this in Lemma 3.

C. Inter-Swarm Collaboration

Our model in section II is general in terms of inter-swarm behavior of peers and their secondary download preferences. Here, we discuss three different behaviors that are covered:
a) *Altruistic Swarms*: In most wired P2P environments (e.g., the Internet), peers are generally insensitive to the consumption of their download and upload bandwidths, and they may download extra pieces in order to help other swarms. Such behavior can be captured in our model by setting $\mathcal{W}_W = \mathcal{W}$ and $\mathcal{F}_W = \mathcal{F}$ for every $W \in \mathcal{W}$. From [8], a network in which all swarms are altruistic is called a *universal swarm network*.
b) *Opportunistic Swarms*: A different type of altruism is when peers do not download any extra pieces but share their pieces with those of other swarms who need them. We obtain this by setting $\mathcal{W}_W = \mathcal{W}$ and $\mathcal{F}_W = W$ for every $W \in \mathcal{W}$.
c) *Selfish Swarms*: In wireless P2P networks, peers are generally sensitive to the consumption of their download and upload bandwidths. This holds by setting $\mathcal{W}_W = \{W\}$ and $\mathcal{F}_W = W$. Thus, the peers do not download any extra pieces nor do they upload any piece to other swarms.

D. Piece-Selection Policies for Excess-Cache

In our original version of RFwPMS, random-selection was assumed for the download of extra pieces, but the stability result in Theorem 1 extends to any piece-selection policy one

$$\widetilde{QV}_W(\mathbf{x}) = \begin{cases} |\boldsymbol{\lambda}|C^{(1)} + \mathcal{O}\left(|\mathbf{x}|_W^{-\omega}\right) + \sum\limits_{r\in R_W}\frac{\Gamma_W^{(r)}}{K_W}\left(\psi_W^{(r)}(\mathbf{x})\left(1-\pi_W^{(r)}\right) - \left(1-\pi_W^{(r)}-\gamma_W^{(r)}\right)I_W^{(1)}(\mathbf{x}) + \gamma_W^{(r)}K_W I_W^{(2)}(\mathbf{x})\right), & \text{if } \mathbf{x}\in\mathcal{S}_W^1 \\ |\boldsymbol{\lambda}|C^{(1)} + \sum\limits_{r\in W}\frac{\Gamma_W^{(r)}}{K_W}\left(\left(K_W-1-C_W^{(1)}\right)(1-\overline{\pi}_W) - \left(1-\overline{\pi}_W-\gamma_W^{(r)}\right)I_W^{(1)}(\mathbf{x}) + \gamma_W^{(r)}K_W I_W^{(2)}(\mathbf{x})\right), & \text{if } \mathbf{x}\in\mathcal{S}_W^2 \end{cases} \quad (29)$$

[5] A situation in which the network suddenly encounters a very large number of empty peers; this is commonly seen with torrents of popular files.

may use for the excess-cache because the Lyapunov function given by (12) and (13) is not affected by a piece download from the set of extra pieces. This is stated below.

Proposition 2. *For the multi-swarm model with non-persistent peers in section II, RFwPMS with any piece-selection policy for the excess-cache, is stable over the parameter region $\mu > 0$, $U > 0$, $\boldsymbol{\lambda} > 0$, and $K_W \geqslant 2$ for all $W \in \mathcal{W}$.*

E. Autonomous Swarms

One more case to consider is when all the swarms in the network operate in isolation from each other. Specifically, a peer belonging to swarm-W contacts and exchanges pieces with peers in the same swarm. The fixed seed, on the other hand, divides its uploading capacity across swarms, providing a static non-zero fraction of its total capacity to each swarm; optimal partition of the seed capacity is for future work. Such swarms are called *autonomous swarms* in [8]. The stability of RFwPMS holds for such swarms as well.

Corollary 1.1. *Consider a multi-swarm network where each swarm $W \in \mathcal{W}$ behaves autonomously and the seed has allocated a static non-zero fraction of its total capacity for each swarm-W (say $U_W > 0$). Then, the network is stable.*

Proof. Each swarm-W can be considered as an isolated single-swarm network with fixed seed of capacity $U_W > 0$. Stability then follows by applying Theorem 1 to each swarm. □

V. Simulations

Next, we investigate the stability, scalability and sojourn time performance of RFwPMS via numerical simulations. In all cases, we set the seed contact rate U to 1, peer contact rate μ to 1, α_W to 10^{-12} and β_W to 1.5.

A. Stability Check

We start by illustrating the stability of RFwPMS. Let us consider a two-swarm network consisting of a master-file $\mathcal{F}$ of 9 pieces, i.e., $\mathcal{F} = [9]$. The two swarms entering the network are denoted by W_1 and W_2, each having a peer-arrival rate of 20. Peers from swarm-W_1 wish to download file $W_1 = [6]$ whereas those from swarm-W_2 are interested in file $W_2 = [3, 9]$. We initiate the system in a state where both swarms are in the one-club scenario: both have 500 peers with all in swarm-W_1 missing piece 1 and all in swarm-W_2 missing piece 4. For the autonomous case, the seed's upload capacity is divided evenly among the two swarms, i.e., $(U_{W_1}, U_{W_2}) = (0.5, 0.5)$. Fig. 1 shows the evolution of the number of peers in the network for different swarm behaviors. Note that each swarm is able to escape the one-club in finite time, after which a stable regime persists with minimal fluctuations. An interesting observation is the sudden drop in the population of swarm-W_2 in altruistic and opportunistic swarms. This is because piece 4 (missing in swarm-W_2), is widely available in swarm-W_1. Thus, due to altruism on the part of swarm-W_1 peers, the one-club peers in swarm-W_2 quickly grab piece 4 and leave the network.

B. Scalability Check

Scalability is also necessary for a P2P network, i.e., its system-wide throughput should scale with the number of peers. For our multi-swarm model, this means that the expected sojourn time of peers should not increase by scaling up the arrival rate vector $\boldsymbol{\lambda}$. To check this we consider a two-swarm network in which the master-file comprises 14 pieces ($\mathcal{F} = [14]$) and the two swarms entering the network are interested in files $W_1 = [8]$ and $W_2 = [6, 14]$. For the autonomous case, the seed's upload capacity is again split equally. Table I lists the (steady-state) expected sojourn times for the arrival rate vectors $\boldsymbol{\lambda} = (8, 4)$ and $16\boldsymbol{\lambda}$. It can be seen that the expected sojourn time practically stays constant in all the four swarm behaviors. Based on this and many other similar empirical checks, we believe that RFwPMS is scalable; however, we leave the proof of this as a conjecture.

The expected sojourn time indeed improves when swarms are more altruistic. Another observation is that the expected sojourn time for autonomous swarms is lesser than that for altruistic swarms. This is not unexpected though; in autonomous swarms, every peer makes contacts within their own swarm, thus the likelihood that the next contact is useless is lower.

C. Sojourn Time Performance in Multiple Swarms

In Table II we compare the expected sojourn times of RFwPMS for two different files under different swarm behaviors. The values tabulated were computed on a two-swarm network in which the arrival rate vector was fixed at $\boldsymbol{\lambda} = (4, 1)$. Again the seed's upload capacity is split evenly in the autonomous case. Note that altruistic and autonomous cases are *Pareto* better than opportunistic and selfish cases.

D. RFwPMS vs. Mode-Suppression

Via simulations MS [11] has been shown to outperform other previously proposed piece-selection policies. Here, we compare the performance of RFwPMS and MS in a single-swarm network. Table III compares the expected sojourn times of RFwPMS and MS for different values of file pieces K with the arrival rate fixed at $\lambda = 4$. It can be seen that using RFwPMS (with $\beta_W = 1.5$), the expected sojourn time is indeed reduced. This reduction is not expected to be significant when K is large (roughly 100 or more) as the increased chunk diversity in steady-state reduces the number of contacts in which only the non-rare pieces are offered.

Fig. 2 compares the flash-crowd response of MS, RFwPMS and RNwPMS (swapping out RF for RN in our policy, see Lemma 3) for $K = 100$ ("large") and an initial population of 500 empty peers; RFwPMS empties the system in about half the time as MS whereas RNwPMS takes the most amount of time. The explanation behind this is indicated in second part of Fig. 2: MS [11] works towards minimizing the mismatch at all times and wastes many contacts; RNwPMS transfers many pieces that are not the rarest and gets stuck in a one-club like scenario that uses just the seed for uploads; and RFwPMS minimizes the mismatch rapidly after it has built up while waiting for the K^{th} piece to be shared by the seed.

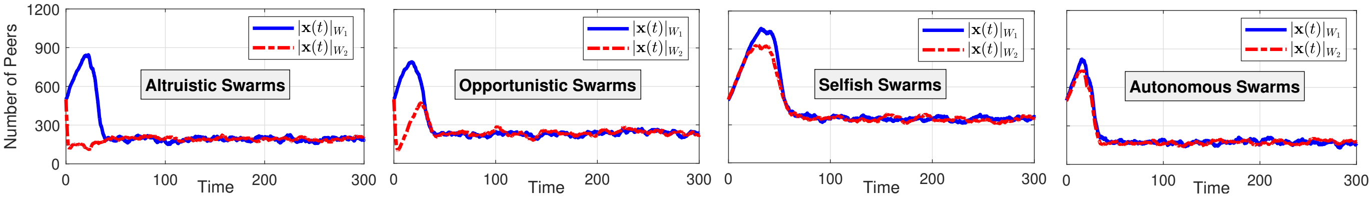

Fig. 1. Number of peers in the network for different swarm behaviours when $W_1 = [6]$, $W_2 = [3, 9]$ and $\boldsymbol{\lambda} = (20, 20)$.

TABLE I
EXPECTED SOJOURN TIMES FOR DIFFERENT ARRIVAL RATE VECTORS

Swarms' Behavior	Arrival Rates[a]	E[Sojourn Time]	
		$W_1 = [8]$	$W_2 = [6, 14]$
Altruistic	$\boldsymbol{\lambda}$	11.2	14.9
	$16\boldsymbol{\lambda}$	11.5	14.9
Opportunistic	$\boldsymbol{\lambda}$	15.1	20.9
	$16\boldsymbol{\lambda}$	15.4	21.9
Selfish	$\boldsymbol{\lambda}$	18.2	26.5
	$16\boldsymbol{\lambda}$	18.5	26.1
Autonomous	$\boldsymbol{\lambda}$	10.3	10.9
	$16\boldsymbol{\lambda}$	10.9	10.8

[a] $\boldsymbol{\lambda} = (8, 4)$. Simulation End-time: 1000 units

TABLE II
EXPECTED SOJOURN TIMES FOR DIFFERENT FILES

Swarms' Behavior	Files		E[Sojourn Time]	
	W_1	W_2	Swarm W_1	Swarm W_2
Altruistic	[2]	[1, 3]	5.5	5.1
	[20]	[10, 30]	23.6	34.5
Opportunistic	[2]	[1, 3]	6.9	14.2
	[20]	[10, 30]	23.9	40.0
Selfish	[2]	[1, 3]	7.8	16.5
	[20]	[10, 30]	35.9	72.8
Autonomous	[2]	[1, 3]	5.6	6.9
	[20]	[10, 30]	23.5	21.2

Simulation End-time: 1000 units

TABLE III
EXPECTED SOJOURN TIMES FOR RFWPMS AND MS

K	E[Sojourn Time]		Percent.
	MS	*RFwPMS*	Improv.
2	6.2	5.2	16.1
10	18.3	12.3	32.6
20	32.4	22.8	29.7
40	55.4	43.7	21.1
80	101.6	85.9	15.5
100	119.4	105.5	11.6
200	226.1	206.8	8.6
500	536.5	503.1	6.2

Simulation End-time: 5000 units

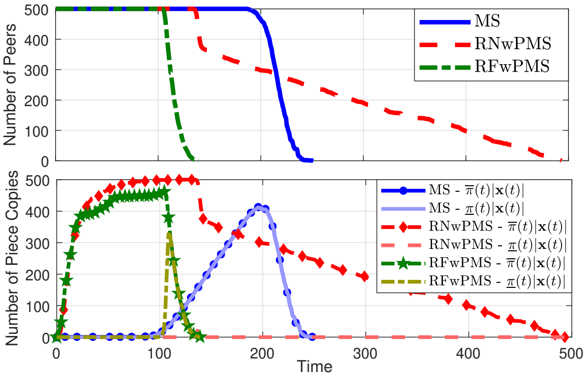

Fig. 2. Flash-crowd Response, 500 empty peers, $K = 100$ and $\lambda = 0$.

VI. Conclusion

In this work, we proposed and studied a piece-selection policy, RFwPMS, for a BitTorrent-like P2P network with multiple swarms. RFwPMS combines rarest-first (for rare pieces) with an adjustable sharing versus suppression choice for non-rare pieces. As a result, RFwPMS reduces the expected sojourn time of a peer and the file-delivery time during a flash-crowd. Using a Lyapunov drift analysis, we proved the stability of RFwPMS in a general multi-swarm setting and demonstrated a policy whose stability region is not affected by bundling swarms. Lastly, since RFwPMS uses rarest-first with a minor modification (use of the non-rares' sharing factor), it is amenable to be incorporated into the BitTorrent protocol.

Our work also opens several avenues for future work. One is on formally investigating the scalability of RFwPMS, and another is on designing excess-cache update policies that can further reduce the sojourn time in multi-swarms. An investigation of the performance of RFwPMS in real torrent deployments is also an interesting direction.

VII. Appendix

Lemma 2. *For any $W \in \mathcal{W}$, $V_W(\mathbf{x}) \to \infty$ as $|\mathbf{x}|_W \to \infty$.*

Proof. From Lemma 2 in [11], it follows that for any state $\mathbf{x}$, $\underline{\pi}_W \leqslant (K_W - 1)/K_W$. We have two cases. Case 1 - $\mathbf{x} \in \mathcal{S}_W^1$: Let $\rho_W \in (0, 1)$ be a constant such that $\rho_W - (K_W - 1)/(K_W) = \upsilon_W > 0$ and define $\mathcal{R}_W \triangleq \{\mathbf{x} : \overline{\pi}_W(\mathbf{x}) \geqslant \rho_W\}$. For all $\mathbf{x} \in \mathcal{S}_W^1 \cap \mathcal{R}_W$, V_W is lower bounded by $(\upsilon_W |\mathbf{x}|_W)^2$ which goes to infinity as $|\mathbf{x}|_W \to \infty$. For all $\mathbf{x} \in \mathcal{S}_W^1 \cap \overline{\mathcal{R}}_W$, $V_W(\mathbf{x})$ is lower bounded by $C_W^{(1)}(1 - \rho_W)|\mathbf{x}|_W$ which goes to infinity as $|\mathbf{x}|_W \to \infty$. Case 2 - $\mathbf{x} \in \mathcal{S}_W^2$: Let $\mathbf{x} \in \mathcal{S}_W^2$. Then using the fact that $\overline{\pi}_W = \underline{\pi}_W$, we can lower bound $V_W(\mathbf{x})$ by $C_W^{(1)}|\mathbf{x}|_W / K_W$ which goes to infinity as $|\mathbf{x}|_W \to \infty$. □

Lemma 3. *For any $\mathbf{x} \in \mathcal{S}_W^1$, $QV_W^{(R+)}(\mathbf{x})$ is upper bounded by*

$$\sum_{\substack{r \in R_W, \\ S : W \backslash S \supsetneq \{r\}}} \frac{x_W^{(S)}}{|\mathbf{x}|} \left(\frac{U}{|R_W \backslash S|} + \mu \sum_{\substack{V : W \in \mathcal{W}_V, \\ T : r \in T}} \frac{x_V^{(T)}}{\xi (T \cap R_W) \backslash S} \right) \Psi_W^{(r)},$$

which can be further upper bounded by

$$\frac{|\mathbf{x}|_W}{|\mathbf{x}|} \sum_{r \in R_W} \frac{\Gamma_W^{(r)}}{K_W} \left(1 - \pi_W^{(r)} - \gamma_W^{(r)}\right) \Psi_W^{(r)}.$$

Note that the first upper bound is the exact Lyapunov drift obtained when peers use random-selection on the available rare pieces (the RNwPMS policy).

Proof. Using (7) and (15), $QV_W^{(R+)}(\mathbf{x})$ is given by

$$\sum_{r \in R_W} \sum_{S : W \backslash S \supsetneq \{r\}} \frac{x_W^{(S)}}{|\mathbf{x}|} \left(\frac{U \mathbb{1}_{\{r \in \underline{R}_{(\dagger, W, S)}\}}(\mathbf{x})}{|\underline{R}_{(\dagger, W, S)}|} \Psi_W^{(r)} \right.$$
$$\left. + \mu \sum_{V : W \in \mathcal{W}_W} \sum_{T : r \in T} \frac{x_V^{(T)} \mathbb{1}_{\{r \in \underline{R}_{(T, W, S)}\}}(\mathbf{x})}{\xi |\underline{R}_{(T, W, S)}|} \Psi_W^{(r)} \right). \quad (30)$$

Let us define

$$\pi_W^{(S, \dagger)}(\mathbf{x}) \triangleq \min_{q \in R_W \backslash S} \pi_W^{(q)}(\mathbf{x}) \quad \text{and}$$

$$\Psi_W^{(S,\dagger)} \triangleq 1 - 2\left(\overline{\pi}_W - \pi_W^{(S,\dagger)}\right)|\mathbf{x}|_W - I_W^{(1)}(\mathbf{x}).$$

Then we can upper bound the first term in (30) as follows.

$$\begin{aligned}
\text{First Term} &= \sum_{\substack{r\in R_W,\\ S:W\backslash S\supsetneq\{r\}}} \frac{x_W^{(S)}}{|\mathbf{x}|} \frac{U\mathbb{1}_{\{r\in\underline{R}_{(\dagger,W,S)}\}}(\mathbf{x})}{|\underline{R}_{(\dagger,W,S)}|}\Psi_W^{(r)} \\
&= \sum_{\substack{S:|S\cap W|<K_W-1,\\ r:r\in\underline{R}_{(\dagger,W,S)}}} \frac{x_W^{(S)}}{|\mathbf{x}|} \frac{U}{|\underline{R}_{(\dagger,W,S)}|}\Psi_W^{(S,\dagger)} && \text{(A1)} \\
&= \sum_{S:|S\cap W|<K_W-1} \frac{x_W^{(S)}}{|\mathbf{x}|} \frac{|R_W\backslash S|U}{|R_W\backslash S|}\Psi_W^{(S,\dagger)} \\
&\leqslant \sum_{\substack{S:|S\cap W|<K_W-1,\\ r:r\in R_W\backslash S}} \frac{x_W^{(S)}}{|\mathbf{x}|} \frac{U}{|R_W\backslash S|}\Psi_W^{(r)} && \text{(A2)} \\
&= \sum_{\substack{r\in R_W,\\ S:W\backslash S\supsetneq\{r\}}} \frac{x_W^{(S)}}{|\mathbf{x}|} \frac{U}{|R_W\backslash S|}\Psi_W^{(r)}. && \text{(A3)}
\end{aligned}$$

For steps (A1) and (A3) we change the order of summation and inequality (A2) follows using $\Psi_W^{(S,\dagger)} \leqslant \Psi_W^{(r)}$ for every $r \in R_W\backslash S$. Now let us define

$$\pi_W^{(S,T)}(\mathbf{x}) \triangleq \min_{q\in(T\cap R_W)\backslash S} \pi_W^{(q)}(\mathbf{x}) \quad \text{and}$$

$$\Psi_W^{(S,T)}(\mathbf{x}) \triangleq 1 - 2\left(\overline{\pi}_W - \pi_W^{(S,T)}\right)|\mathbf{x}|_W - I_W^{(1)}(\mathbf{x}).$$

Then, using the same technique as in steps (A1) and (A3), and noting that $\Psi_W^{(S,T)} \leqslant \Psi_W^{(r)}$ for every $r \in (T\cap R_W)\backslash S$, we can upper bound the second term in (30) by

$$\sum_{\substack{r\in R_W,\\ S:W\backslash S\supsetneq\{r\}}} \frac{x_W^{(S)}}{|\mathbf{x}|}\mu \sum_{\substack{V:W\in\mathcal{W}_V,\\ T:r\in T}} \frac{x_V^{(T)}}{\xi(T\cap R_W)\backslash S}\Psi_W^{(r)}.$$

Combining the two results, we get the first desired upper bound. The second upper bound follows by noting that $\Psi_W^{(r)} \leqslant 0$ and upper bounding $|R_W\backslash S|$ and $|(T\cap R_W)\backslash S|$ by K_W. □

Lemma 4. *For any $\omega > 0$, $QV_W^{(\overline{R})}(\mathbf{x})$ is upper bounded by $|\mathbf{x}|_W/|\mathbf{x}| \times \mathcal{O}\left(|\mathbf{x}|_W^{-\omega}\right)$ over the region $\mathcal{S}_W^1$. Consequently, for any $\epsilon > 0$, there exists $N^* = N^*(\epsilon) \in \mathbb{R}_+$ such that $QV_W^{(\overline{R})}(\mathbf{x}) \leqslant \epsilon/2$ for all $|\mathbf{x}|_W \geqslant N^*$.*

Proof. Fix $\omega > 0$. With $\alpha_W > 0$, $\zeta_W(\mathbf{x}) \in \mathcal{O}(|\mathbf{x}|^{-2-\omega})$. From (10) and (19), we can upper bound $QV_W^{(\overline{R})}(\mathbf{x})$ by

$$\sum_{\substack{n\in N_W,\\ S:W\backslash S\supseteq\{n\}}} \frac{x_W^{(S)}}{|\mathbf{x}|} \underbrace{K_W^2\left(U + \mu\frac{|\mathbf{x}|}{\xi}\right)\left(1 + 2|\mathbf{x}|_W + C_W^{(2)}\right)}_{\mathcal{O}(|\mathbf{x}|^2)}\zeta_W(\mathbf{x})$$

Since $\zeta_W(\mathbf{x}) \in \mathcal{O}(|\mathbf{x}|^{-2-\omega})$, the Lemma follows. □

Lemma 5. *For any $\epsilon > 0$ and any $W \in \mathcal{W}$, there exists a countable set $\mathcal{A}_W$ and a finite constant $B_W \geqslant 0$ such that*

$$\widetilde{QV}_W(\mathbf{x}) \leqslant -\epsilon\mathbb{1}_{\{\mathbf{x}\in\overline{\mathcal{A}}_W\}}(\mathbf{x}) + B_W\mathbb{1}_{\{\mathbf{x}\in\mathcal{A}_W\}}(\mathbf{x}), \ \textit{and}$$

$$N_W = \max_{\mathbf{x}\in\mathcal{A}_W} |\mathbf{x}|_W < \infty.$$

Proof. The Lemma can be proved by using a similar approach as in the stability proof of [11]. The Big-$\mathcal{O}$ term present in $\widetilde{QV}_W(\mathbf{x})$ (when $\mathbf{x} \in \mathcal{S}_W^1$) is upper bounded by $\epsilon/2$ for all $|\mathbf{x}|_W \geqslant N^*$. The proof is omitted for brevity. □

Lemma 6. *For any $\epsilon' > 0$, there exists a finite set $\mathcal{A}$ and a finite constant $B \geqslant 0$ such that*

$$QV(\mathbf{x}) \leqslant -\epsilon'\mathbb{1}_{\{\mathbf{x}\in\overline{\mathcal{A}}\}}(\mathbf{x}) + B\mathbb{1}_{\{\mathbf{x}\in\mathcal{A}\}}(\mathbf{x}).$$

Proof. Fix $\epsilon' > 0$. For any $\epsilon > 0$ and any swarm $W \in \mathcal{W}$, it follows from Lemma 5 that $\widetilde{QV}_W(\mathbf{x}) \leqslant -\epsilon$, except possibly over $\mathcal{A}_W$ where its population and corresponding term $\widetilde{QV}_W(\mathbf{x})$ are bounded from above by N_W and $B_W \geqslant 0$ respectively. We can write the state-space as $\mathcal{S} = \bigcup_{\mathcal{H}:\mathcal{H}\subseteq\mathcal{W}} \mathcal{S}_\mathcal{H}$, where

$$\mathcal{S}_\mathcal{H} \triangleq \{|\mathbf{x}|_W \leqslant N_W \forall W \in \mathcal{H}, |\mathbf{x}|_U > N_U \forall U \in \mathcal{W}\backslash\mathcal{H}\}.$$

Case 1 - $\mathcal{H} = \varnothing$: Consider a state $\mathbf{x} \in \mathcal{S}_\varnothing$. Since $|\mathbf{x}|_W > N_W$ for all $W \in \mathcal{W}$, from Lemma 5, it follows that $\widetilde{QV}_W(\mathbf{x}) \leqslant -\epsilon$ for all $W \in \mathcal{W}$. This gives $QV(\mathbf{x}) \leqslant -\epsilon$. Choosing $\epsilon \geqslant \epsilon'$ ensures $QV(\mathbf{x}) \leqslant -\epsilon'$.

Case 2 - $\mathcal{H} = \mathcal{W}$: The set $\mathcal{S}_\mathcal{W}$ is finite. Thus, for any state $\mathbf{x} \in \mathcal{S}_\mathcal{W}$, we can write $QV(\mathbf{x}) \leqslant B_\mathcal{W} \triangleq \max_{\mathbf{x}\in\mathcal{S}_\mathcal{W}} (QV(\mathbf{x}))^+ < \infty$.

Case 3 - $\varnothing \neq \mathcal{H} \subsetneq \mathcal{W}$: Consider a state $\mathbf{x} \in \mathcal{S}_\mathcal{H}$. We can upper bound $QV(\mathbf{x})$ as follows.

$$\begin{aligned}
QV(\mathbf{x}) &= \sum_{W\in\mathcal{H}} \frac{|\mathbf{x}|_W}{|\mathbf{x}|}\widetilde{QV}_W(\mathbf{x}) + \sum_{U\in\mathcal{W}\backslash\mathcal{H}} \frac{|\mathbf{x}|_U}{|\mathbf{x}|}\widetilde{QV}_U(\mathbf{x}) \\
&< \frac{\sum_{W\in\mathcal{H}} N_W B_W}{\sum_{U\in\mathcal{W}\backslash\mathcal{H}} |\mathbf{x}|_U} + \frac{\sum_{U\in\mathcal{W}\backslash\mathcal{H}} |\mathbf{x}|_U}{\sum_{W\in\mathcal{H}} N_W + \sum_{U\in\mathcal{W}\backslash\mathcal{H}} |\mathbf{x}|_U}(-\epsilon).
\end{aligned}$$

Note that $\sum_{U\in\mathcal{W}\backslash\mathcal{H}} |\mathbf{x}|_U \to \infty$ over the set $\mathcal{S}_\mathcal{H}$, which implies

$$\frac{\sum_{W\in\mathcal{W}_1} N_W B_W}{\sum_{U\in\mathcal{W}_2} |\mathbf{x}|_U} \to 0 \text{ and } \frac{\sum_{U\in\mathcal{W}_2} |\mathbf{x}|_U}{\sum_{W\in\mathcal{W}_1} N_W + \sum_{U\in\mathcal{W}_2} |\mathbf{x}|_U} \to 1.$$

Thus, for any $\phi > 0$, there exists $N_\mathcal{H} = N_\mathcal{H}(\phi) \in \mathbb{R}_+$ such that $QV(\mathbf{x}) \leqslant \phi + (1-\phi)(-\epsilon)$ for any $\mathbf{x} \in \mathcal{S}_\mathcal{H}$ with $\sum_{U\in\mathcal{W}_2} |\mathbf{x}|_U \geqslant N_\mathcal{H}$. Choosing $\epsilon = 2\epsilon'$ and $0 < \phi \leqslant \frac{\epsilon/2}{1+\epsilon}$ ensures $QV(\mathbf{x}) \leqslant -\epsilon'$.

For all $\mathcal{H}$ such that $\varnothing \neq \mathcal{H} \subsetneq \mathcal{W}$, let us define

$$\mathcal{A}_\mathcal{H} \triangleq \mathcal{S}_\mathcal{H} \cap \left\{\sum_{U\in\mathcal{H}} |\mathbf{x}|_U < N_\mathcal{H}\right\} \qquad \text{(finite)},$$

$$\text{and then } \mathcal{A} \triangleq \left(\bigcup_{\phi\neq\mathcal{H}\subsetneq\mathcal{W}} \mathcal{A}_\mathcal{H}\right) \cup \mathcal{S}_\mathcal{W} \qquad \text{(finite)},$$

$$B \triangleq \max_{\mathbf{x}\in\mathcal{A}} (QV(\mathbf{x}))^+ < \infty.$$

Then, for any state $\mathbf{x}$, we can write

$$QV(\mathbf{x}) \leqslant -\epsilon'\mathbb{1}_{\{\mathbf{x}\in\overline{\mathcal{A}}\}}(\mathbf{x}) + B\mathbb{1}_{\{\mathbf{x}\in\mathcal{A}\}}(\mathbf{x}),$$

establishing the result. □

Combining Lemma 6 and Proposition 1, Theorem 1 holds.

References

[1] I. Norros, H. Reittu, and T. Eirola, "On the stability of two-chunk file-sharing systems," *Queueing Systems*, vol. 67, no. 3, pp. 183–206, March 2011.

[2] B. Cohen, "Incentives build robustness in bittorrent," *Workshop on Economics of Peer-to-Peer systems*, vol. 6, pp. 68–72, June 2003.

[3] Sandvine, "The global internet phenomena report," 2018. [Online]. Available: https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf

[4] L. Massoulie and M. Vojnovic, "Coupon replication systems," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 603–616, June 2008.

[5] B. Hajek and J. Zhu, "The missing piece syndrome in peer-to-peer communication," *Stochastic Systems*, vol. 1, no. 2, pp. 246–273, 2011.

[6] J. Zhu and B. Hajek, "Stability of a peer-to-peer communication system," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4693–4713, July 2012.

[7] X. Zhou, S. Ioannidis, and L. Massoulie, "On the stability and optimality of universal swarms," in *ACM SIGMETRICS*, ser. SIGMETRICS '11. ACM, June 2011, pp. 341–352.

[8] J. Zhu, I. Stratis, N. Hegde, and L. Massoulié, "Stable and scalable universal swarms," *Distributed Computing*, vol. 28, no. 6, pp. 391–406, December 2015.

[9] B. Oğuz, V. Anantharam, and I. Norros, "Stable distributed p2p protocols based on random peer sampling," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1444–1456, October 2015.

[10] O. Bilgen and A. B. Wagner, "A new stable peer-to-peer protocol with non-persistent peers," in *IEEE INFOCOM*, May 2017, pp. 1–9.

[11] V. Reddyvari, P. Parag, and S. Shakkottai, "Mode-suppression: A simple and provably stable chunk-sharing algorithm for p2p networks," in *IEEE INFOCOM*, April 2018, pp. 2573–2581.

[12] D. Qiu and R. Srikant, "Modeling and performance analysis of bittorrent-like peer-to-peer networks," *SIGCOMM Computer Communication Review*, vol. 34, no. 4, pp. 367–378, August 2004.

[13] D. S. Menasche, A. A. d. A. Rocha, B. Li, D. Towsley, and A. Venkataramani, "Content availability and bundling in swarming systems," *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 580–593, April 2013.

[14] E. de Souza e Silva, R. M. Leão, D. S. Menasché, and D. Towsley, "On the scalability of p2p swarming systems," *Computer Networks*, vol. 151, pp. 93–113, Mar 2019. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2019.01.006

[15] A. Legout, G. Urvoy-Keller, and P. Michiardi, "Rarest first and choke algorithms are enough," in *ACM SIGCOMM*, ser. IMC '06. ACM, 2006, pp. 203–216.

[16] B. Hajek, *Random Processes for Engineers*. Cambridge University Press, 2015.

[17] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2014.