On the Stability and Optimality of Universal Swarms

Xia Zhou
Dept. of Computer Science
UC Santa Barbara, CA, USA
xiazhou@cs.ucsb.edu

Stratis Ioannidis
Technicolor
Palo Alto, CA, USA
stratis.ioannidis@technicolor.com

Laurent Massoulié
Technicolor
Paris, France
laurent.massoulie@technicolor.com

ABSTRACT

Recent work on BitTorrent swarms has demonstrated that a bandwidth bottleneck at the seed can lead to the under-utilization of the aggregate swarm capacity. Bandwidth underutilization also occurs naturally in mobile peer-to-peer swarms, as a mobile peer may not always be within the range of peers storing the content it desires. We argue in this paper that, in both cases, idle bandwidth can be exploited to allow content sharing across multiple swarms, thereby forming a *universal swarm* system. We propose a model for universal swarms that applies to a variety of peer-to-peer environments, both mobile and online. Through a fluid limit analysis, we demonstrate that universal swarms have significantly improved stability properties compared to individually autonomous swarms. In addition, by studying a swarm's stationary behavior, we identify content replication ratios across different swarms that minimize the average sojourn time in the system. We then propose a content exchange scheme between peers that leads to these optimal replication ratios, and study its convergence numerically.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design–*distributed networks, store and forward networks*; C.4 [**Performance of Systems**]: Performance Attributes

General Terms

Theory, Algorithms

Keywords

Universal swarms, content distribution, peer-to-peer networks

1. INTRODUCTION

Peer-to-peer systems have been tremendously successful in enabling sharing of large files in a massive scale. This success has motivated several approaches of modeling BitTorrent swarms [6, 11, 12, 13]. Such models have illuminated important aspects of swarm behavior, including determining conditions for *swarm stability* and minimizing the system's *average sojourn time*. The study of swarm stability amounts to identifying conditions under which the swarm population remains finite as time progresses, while the sojourn time captures the time required until peers retrieve the file they request and leave the swarm.

Our aim in this work is to provide answers for similar questions in the context of *universal swarms* [15]. Rather than considering swarms as individual autonomous systems, we study scenarios in which peers from different swarms are permitted to exchange file pieces (chunks) with each other. Such inter-swarm exchanges make sense when bottlenecks in a single swarm lead to bandwidth under-utilization.

One application in which such bandwidth bottlenecks naturally arise is the peer-to-peer distribution of content over mobile opportunistic networks. Mobile peers wishing to retrieve a file can do so by downloading chunks from other peers they encounter opportunistically. These mobile content distribution systems have received considerable attention recently [1, 2, 4, 7, 8, 9, 14], as they alleviate the load on the wireless infrastructure by harnessing the bandwidth available during local interactions among mobile peers.

Bottlenecks in such peer-to-peer systems are a result of the opportunistic nature of the communication between peers: two peers meeting may not necessarily belong to the same swarm and may not be interested in the same content. Nevertheless, during encounters with peers from other swarms, a peer may use its idle bandwidth to obtain pieces of files in other swarms. Such exchanges can aid the propagation of under-replicated pieces that are otherwise hard to locate. If designed properly, such inter-swarm exchanges have the potential to improve the overall performance in terms of sojourn times and system stability.

Bottlenecks can also lead to bandwidth underutilization in online swarms. An example can be found in the recent work of Hajek and Zhu [6]. The authors considered the stability of a single BitTorrent swarm comprising a single seed and a steady stream of arriving peers (leechers). The peers share pieces they have retrieved while they are in the system but immediately depart once they download all pieces of a file. Hajek and Zhu observed that if the arrival rate of peers exceeds the upload capacity of the seed, the system becomes unstable in a very specific way: almost all peers arriving in the swarm very quickly obtain every missing piece *except for one*. The seed is unable to serve these peers with the

SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.

missing piece fast enough and, as a result, the size of this set of peers—termed the "one-club" [6]—grows to infinity.

When this so-called "missing piece syndrome" occurs, peers waiting for the missing piece are effectively idle, and their available upload bandwidth is essentially under-utilized. In this work we argue that, provided that peers have excess storage, this idle bandwidth capacity can be exploited in the presence of other swarms to store and to exchange pieces of other files. Such inter-swarm exchanges have the potential of improving the overall stability of the universal swarm system, as the peers in the "one-club" may be able to retrieve their missing piece from collaborating peers in other swarms. Most importantly, such transactions can be restricted to take place only when the intra-swarm bottleneck has rendered the peers idle, so inter-swarm exchanges do not hinder the delivery of the file in any way.

Our contributions can be summarized as follows:

- We propose a novel mathematical model for inter-swarm data exchange. Our model is simple but versatile enough to capture several different peer-to-peer file-sharing environments, both mobile and online.
- Using the above model, we analyze the stability of a universal swarm in which peers can retrieve items they miss from other swarms, but otherwise keep their caches static.
- Studying the stationary points of the data exchange process, we characterize the optimal replication ratios of pieces across swarms that minimize the system's average sojourn times.
- We propose BARON, a scheme for guiding data exchanges to yield optimal replication ratios, and study its convergence to these ratios numerically.

To the best of our knowledge, our work is the first systematic study of file sharing in a universal swarm system. Our results suggest that universal swarms can indeed achieve considerable performance improvements over independent autonomous swarm systems.

In particular, we establish the following surprising result: in a universal swarm where inter-swarm piece exchanges take place, *only one* swarm can become unstable. This is an interesting finding, especially when viewed in the context of the work of Hajek and Zhu [6]. An intuitive explanation of this phenomenon is this: a swarm growing to infinity attains an ever-growing capacity, which can be used to serve the missing pieces of other swarms. This service suppresses the growth of other swarms and, as a result, no two "one-clubs" can exist simultaneously.

Furthermore, our proposed scheme for guiding content exchanges can be used to enlarge the stability region for a universal swarm. Our design raises interesting open questions, such as the construction of schemes that work, *e.g.*, in fully distributed or non-cooperative environments. Though our model is simple, and our analysis is a first attempt at analyzing universal swarm behavior, we believe that these results are very promising. They indicate that universal swarms have very appealing stability properties, and certainly merit further investigation.

The remainder of this paper is structured as follows. We begin with an overview of related work in Section 2 and introduce our mathematical model for universal swarms in Section 3. We present our main results on convergence, stability, and optimality in Section 4, and provide their proofs in Section 5. We further propose BARON, a scheme to guide the system to the optimal stationary state, and evaluate it numerically in Section 6. We conclude by presenting future directions in Section 7.

2. RELATED WORK

Qiu and Srikant [13] were the first to introduce a fluid model for BitTorrent. Using an ordinary differential equation (ODE) to capture peer dynamics, they study sojourn times at the fixed points of this ODE, as well as the impact of incentive schemes. Our work is most similar to Massoulié and Vojnovic [12] who, contrary to [13], study directly the dynamics of the stochastic system determined by piece exchanges between peers. As in the present work, no seed exists: peers arrive already storing several pieces of a file, and exchange pieces by contacting uniformly at random other peers in the swarm. The authors identify conditions for system stability and determine the sojourn time at equilibrium. Massoulié and Twigg [11] study similar issues in the context of P2P streaming, which differs by requiring that pieces are retrieved in a certain order. Our work generalizes [12] by allowing piece exchanges across swarms and, as [11, 12], studies a fluid limit of the resulting system.

Recent work by Hajek and Zhu [6] identifies the "missing piece syndrome" described in the introduction. Their model differs from [12] in assuming that a single, non-transient seed exists while all other peers arrive with no pieces. The bandwidth bottleneck due to the missing piece syndrome partially motivates our study of universal swarms. We will further elaborate on the relationship of our work to [6] in our concluding remarks.

In the context of mobile peer-to-peer systems, BARON, our scheme for guiding content exchanges, is related to a series of recent papers on optimizing mobile content delivery. In general, the goal of these works is to ensure fast delivery of content to mobile users through opportunistic exchanges while using as few bandwidth and storage resources as possible. Schemes studied involve selecting which content to transmit during contacts [1, 2, 7], which information to cache in local memory [9, 14], or where to inject new content [8]. Our work differs both in considering an open system, where mobile users depart once obtaining the content they want, as well as in capturing several different (*e.g.*, contact or interference constrained) communication scenarios.

3. SYSTEM MODEL

3.1 Overview

The system that we model is a universal swarm, consisting of several peers wishing to retrieve different content items. Peers share content they store with other peers while they are in the system; once a peer retrieves the content item that it is interested in, it exits the system.

Our model describes both mobile and online peer-to-peer swarms. In both cases, we assume that downloads take place as in [12]: each peer is idle for an exponentially distributed time and then contacts a peer selected uniformly at random from the peers present in the system. During such contacts, peers may choose to exchange content items they store and all transfers are instantaneous.

In the wireless mobile case, the above contact process aims to model mobility. That is, two mobile peers come into contact whenever they are within each other's transmission range. In an online peer-to-peer network, the contact process captures *random sampling*. In particular, peers sample the system population uniformly at random to find the items they want. No "universal tracker" exists, and peers do not know which peers in other swarms may be storing the items they request, hence the need for random sampling.

We make the following assumptions. First, every peer entering the system is only interested in downloading *a single content item*; once retrieving this single item, the peer immediately exits the system. Second, whenever a peer contacts another peer that stores its requested item, it is able to retrieve the *entire item* immediately. Third, as in [12], peers arrive with non-empty caches, and begin to share immediately when they enter the system.

The above assumptions are obviously simplifications of real-life peer-to-peer system behavior. On one hand, if our items correspond to the granularity of files, a peer would not be able to download an entire file within one downloading session with another peer. If, on the other hand, items correspond to the granularity of chunks, peers would need to retrieve several items before exiting the system. Nevertheless, in spite of these simplifications, our analysis provides interesting insights in universal swarm behavior, especially in light of the "missing piece syndrome" observed by Hajek and Zhu [6]. We will revisit this issue in Section 7.

3.2 Peer Swarms and Classes

We consider a *universal swarm* in which content items belonging to a set $\mathbb{K}$, where $|\mathbb{K}| = K$, are shared among transient peers. Each peer arrives with a request $i \in \mathbb{K}$ and a cache of items $f \subset \mathbb{K}$, where $C = |f|$ is the capacity of the cache. We denote by $\mathbb{F} = \{f \subset \mathbb{K} : |f| = C\}$ the set of all possible contents of a peer's cache.

A peer *swarm* consists of all peers interested in retrieving the same item $i \in \mathbb{K}$. We partition the peers in the system into *classes* according to both (a) the item they request and (b) the content in their cache. That is, each pair $(i, f) \in \mathbb{C} = \mathbb{K} \times \mathbb{F}$ defines a distinct peer class.

We denote by $N_{i,f}(t)$ the number of peers requesting i and storing f at time t. We use the notation

$$\mathbf{N}(t) = [N_{i,f}(t)]_{(i,f)\in\mathbb{C}}$$

for the vector representing the system state, *i.e.*, the number of peers in each class. We also denote by

$$N(t) = \sum_{(i,f)\in\mathbb{C}} N_{i,f}(t) = \mathbf{1}^T \cdot \mathbf{N}(t)$$

the total number of peers in the system at time t.

3.3 Peer Arrival Process

Peers requesting item $i \in \mathbb{K}$ and storing $f \in \mathbb{F}$ arrive according to a Poisson process with rate $\lambda_{i,f}$, and that arrivals across different classes are independent. By definition, $\lambda_{i,f} = 0$ if $i \in f$. We denote by $\lambda = \sum_{(i,f)\in\mathbb{C}} \lambda_{i,f}$ the aggregate arrival rate of peers in the system. We also define

$$\lambda_{i,\cdot} = \sum_{f\in\mathbb{F}} \lambda_{i,f}, \quad \lambda_{\cdot,i} = \sum_{\substack{j\in\mathbb{K} \\ f: i\in f}} \lambda_{j,f}, \quad i \in \mathbb{K} \tag{1}$$

as the aggregate arrival rates of peers requesting and caching item i, respectively.

For some of our results, we require that λ tends to infinity; when doing so, we assume that the arrival rate corresponding to each class increases proportionally to λ, *i.e.*, the normalized arrival rate

$$\hat{\lambda}_{i,f} = \lambda_{i,f}/\lambda \tag{2}$$

is constant w.r.t. λ.

3.4 Contact Process

Opportunities to exchange items among peers occur when two peers come into contact. As mentioned in Section 3.1, contacts model different processes in a mobile network and an online peer-to-peer network. In the mobile case, a contact indicates that two mobile peers are within each other's transmission range. In the online case, contacts capture random peer sampling in the universal swarm.

Formally, if $N(t)$ is the total number of peers in the system at time t, then a given peer a present in the system contacts other peers according to a non-homogeneous Poisson process with rate

$$\mu \cdot (N(t))^{1-\beta}, \quad \beta \in [0, 2].$$

The peer with which peer a comes into contact is selected uniformly at random from the $N(t)$ peers currently present in the system. Moreover, the above contact processes are independent across peers.

The parameter β is used to capture different communication scenarios that may arise in a mobile or online network. We classify these below into *contact-constrained*, *constant-bandwidth*, and *interference-constrained* scenarios.

Contact-constrained communication. When $0 \le \beta < 1$, the contact rate of a peer is growing proportionally to the total peer population. This would be the case in a sparse, opportunistic or DTN-like wireless mobile network, where peers are within each other's transmission range very infrequently. In such cases, the bottleneck in data exchanges is determined by how often peers meet. Adding more peers in such an environment can increase the opportunities for contacts between peers. This is reflected in the increase of a peer's contact rate as the population size grows.

Constant-bandwidth communication. When $\beta = 1$ the contact rate of a peer does not depend on the population size. This reflects constant-bandwidth scenarios, where the system population has no effect on the bandwidth capabilities of a peer, and is thus a natural model of an online peer-to-peer network.

Interference-constrained communication. When $\beta \in (1, 2]$, the contact rate of a peer decreases as the total peer population grows. This captures a dense wireless network in which peers share a wireless medium to communicate. As the number of peers increases, the wireless interference can become severe, degrading the network throughput. This is reflected in our model by a decrease in successful contact events and, thus, in a peer's contact rate.

If $\beta > 2$, the aggregate contact rate over *all* peers in the system decreases as the total peer population grows. Assuming constant arrival rates, such a system will be trivially unstable; as such, we do not consider this case.

For simplicity of notation, we allow self-contacts. Contacts are not symmetric; when Alice contacts Bob, Bob does

not contact Alice, and vice versa. This, however, is not restrictive: symmetric contacts can be easily represented by appropriately defining symmetric interactions between two peers (*c.f.* the conversion probabilities appearing below).

Under the above assumptions, when the system state is $\mathbf{N}$, the aggregate rate with which users from class A contact users from class A' is

$$\mu_{A,A'}(\mathbf{N}) = \mu N_A N_{A'}/N^{\beta}, \quad A,A' \in \mathbb{C}. \tag{3}$$

We call $\mu_{A,A'}$ as the *inter-contact* rate between A and A'.

3.5 Content Exchanges During Contacts

When a peer in class $A \in \mathbb{C} = \mathbb{K} \times \mathbb{F}$ contacts another peer in class $B \in \mathbb{C}$, the two peers may exchange items stored in their respective caches. Such exchanges can lead to, *e.g.*, (a) the departure of a peer, because it obtains the item it requests, or (b) the change of its cache contents, as new items replace old items in the peer's cache.

In particular, given that the current state of the system is $\mathbf{N}(t)$, when a peer of class $A \in \mathbb{C}$ contacts another peer in $A' \in \mathbb{C}$, the peer in A is converted to a peer in $B \in \mathbb{C} \cup \{\varnothing\}$ and the peer in A' is converted to a peer in $B' \in \mathbb{C} \cup \{\varnothing\}$ with the following probability

$$\Delta_{A,A' \to B,B'}(\mathbf{N}(t)),$$

independently of any other event in the history of the process $\mathbf{N}(t)$ so far. In the above, we use the notation $\varnothing$ to indicate that a peer exits the system. We call the above Δ functions the *conversion probabilities* of the system. Conversion probabilities depend on the global state $\mathbf{N}(t)$ at the time of contact. We make the following technical assumption:

ASSUMPTION 1. *For every $s > 0$, and for every $A, A' \in \mathbb{C}$ and $B, B' \in \mathbb{C} \cup \{\varnothing\}$, $\Delta_{A,A' \to B,B'}(\mathbf{N}) = \Delta_{A,A' \to B,B'}(s\mathbf{N})$.*

In other words, the conversion probabilities are *invariant to rescaling*: if all peer classes are increased by the same factor, the conversion probabilities will remain unaltered. Let

$$\zeta^{A}_{A',A' \hookrightarrow B',B''} = \mathbb{1}_{B'=A} + \mathbb{1}_{B''=A} - \mathbb{1}_{A'=A} - \mathbb{1}_{A''=A}, \tag{4}$$

be an indicator function capturing how a conversion $A', A'' \to B', B''$ affects the population of class A. For example, (4) states that conversions can increase N_A by at most 2, when both classes A', A'' are converted to A.

We require that conversions follow what we call the "*grab-and-go*" principle: whenever two peers come into contact, if the first stores the second peer's requested item, the latter will retrieve it and exit the system. In other words, content exchanges that lead to departures are always enforced. Formally, the "grab-and-go" principle can be defined as:

$$\begin{aligned} &\textstyle\sum_{B' \in \mathbb{C} \cup \{\varnothing\}} \Delta_{(i,f),(i',f') \to \varnothing,B'}(\mathbf{N}) = 1 \text{ if } i \in f', \text{ and} \\ &\textstyle\sum_{B \in \mathbb{C} \cup \{\varnothing\}} \Delta_{(i,f),(i',f') \to B,\varnothing}(\mathbf{N}) = 1 \text{ if } i' \in f. \end{aligned} \tag{5}$$

The simplest interaction that satisfies the "grab-and-go" principle is the *static-cache* policy: peers never alter the contents of their caches for as long as they are in the system, other than as dictated by the "grab-and-go" principle. Formally, the static-cache policy can be stated as:

$$\Delta_{(i,f),(i',f') \to B,B'}(\mathbf{N}) = 1, \text{ where}$$
$$B = \begin{cases} \varnothing, & \text{if } i \in f' \\ (i,f) & \text{o.w.,} \end{cases} \quad \text{and } B' = \begin{cases} \varnothing, & \text{if } i' \in f \\ (i',f'), & \text{o.w.} \end{cases} \tag{6}$$

Table 1: Summary of Notation

$\mathbb{K}$	Set of items
C	Cache capacity
$\mathbb{F}$	Set of possible cache contents
(i,f)	Class of users requesting $i \in \mathbb{K}$ and storing $f \in \mathbb{F}$
$\mathbb{C}$	The set of classes $\mathbb{K} \times \mathbb{F}$
$N_{i,f}(t)$	The number of users in class (i,f)
$\mathbf{N}(t)$	The system state
$N(t)$	Number of peers in the system
$\lambda_{i,f}$	Arrival rate of peers in class (i,f)
λ	Aggregate arrival rate
$\hat{\lambda}_{i,f}$	Normalized arrival rate for class (i,f)
β	Decay exponent of the contact rate
μ	Contact rate constant
$\Delta_{A,A' \to B,B'}$	Conversion probabilities
$\zeta^{A}_{A',A' \hookrightarrow B',B''}$	Effect of conversion $A', A'' \to B', B''$ on class A
$\delta_{A',A' \hookrightarrow B',B''}$	Limit points of the conversion probabilities
$n_{i,f}(t)$	Fluid trajectory of class (i,f)
$\mathbf{n}(t)$	Fluid trajectories of the system state
$n(t)$	Sum of fluid trajectories
$n_{i,\cdot}$, $n_{\cdot,i}$	Demand and supply for $i \in \mathbb{K}$
$n^*_{i,\cdot}$, $n^*_{\cdot,i}$	Optimal demand and supply for $i \in \mathbb{K}$

Of course, there are many other conversion probabilities that satisfy the "grab-and-go" principle. In particular, (5) tells us nothing about how peers interact with each other when neither of them stores the other's requested item. Rather than leaving caches static, as in (6), such events can be exploited to change the number of replicas in the system, *e.g.*, to reach some global optimization objective, like increasing system stability or reducing the system sojourn times. We do precisely this in Section 6: we design interactions between peers (*i.e.*, determine the conversion probabilities) in a way that such a global optimization objective is met.

4. MAIN RESULTS

Having described our system model, we now present our main results. To begin with, we establish that, for arbitrary conversion probabilities the dynamics of our system can arbitrarily well approximated by a fluid limit (Section 4.1). We then describe the stability region of the static-cache policy (Section 4.2). Finally, we establish conditions under which interactions that follow the "grab-and-go" principle minimize sojourn times (Section 4.3).

4.1 Convergence to a Fluid Limit

Our first main result states that the evolution of the universal swarm through time can be approximated arbitrarily well by the solution of an ordinary differential equation (ODE). This result is very general: we prove convergence to such a fluid limit for *all* $\beta \in [0,2)$ and *all* conversion probabilities satisfying Assumption 1.

We begin by formally defining the notion of a fluid limit of the universal swarm. We say that the vector

$$\mathbf{n}(t) = [n_A(t)]_{A \in \mathbb{C}}$$

is a *fluid trajectory* of the system if, for every class $A \in \mathbb{C}$, the functions $n_A : \mathbb{R}_+ \to \mathbb{R}_+$ satisfy the following ODEs:

$$\dot{n}_A(t) = \hat{\lambda}_A + \sum_{\substack{A',A'' \in \mathbb{C} \\ B',B'' \in \mathbb{C} \cup \{\varnothing\}}} \zeta^{A}_{A',A' \hookrightarrow B',B''} \mu_{A',A''}(\mathbf{n}(t)) \delta_{A',A' \hookrightarrow B',B''}(\mathbf{n}(t)), \tag{7}$$

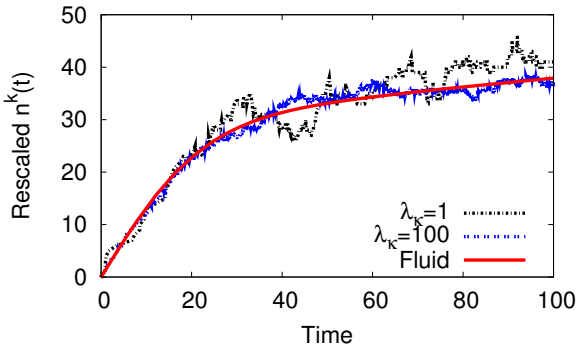


Figure 1: Comparing the rescaled trajectory of the original system to the fluid trajectory using the static-cache policy. We simulated a system where $\mathbb{K} = \{1, 2, 3\}, C = 1$, and $\beta = 0$, for $\lambda_k = 1$ and $\lambda_k = 100$ respectively. The rescaled trajectory clearly converges to the fluid trajectory as λ_k increases.

where $\hat{\lambda}_A$, $\mu_{A',A''}$, and $\zeta^A_{A',A''\hookrightarrow B',B''}$ are given by (2), (3), and (4) respectively, and $\delta_{A',A''\hookrightarrow B',B''} : \mathbb{R}_+^{|\mathbb{C}|} \to [0, 1]$ are any functions that satisfy the following property:

$$\delta_{\cdot\to\cdot}(\mathbf{n}) \in [\liminf_{\mathbf{n}'\to\mathbf{n}} \Delta_{\cdot\to\cdot}(\mathbf{n}'), \limsup_{\mathbf{n}'\to\mathbf{n}} \Delta_{\cdot\to\cdot}(\mathbf{n}')].$$

The δ functions are unique and coincide with the conversion probabilities if and only if the latter are continuous. In this case, the ODEs (7) also have a unique solution. For any $\mathbf{n}^* \in \mathbb{R}_+^{|\mathbb{C}|}$, let $S(\mathbf{n}^*)$ be the set of all fluid trajectories of the system with initial condition $\mathbf{n}(0) = \mathbf{n}^*$. The following theorem establishes two facts. First, $S(\mathbf{n}^*)$ is non-empty—*i.e.*, fluid trajectories exist for all initial conditions. Second, under appropriate rescaling, a trajectory of the universal swarm $\{\mathbf{N}(t), t \in \mathbb{R}_+\}$, can be arbitrarily well approximated by a fluid trajectory.

THEOREM 1. *Let $\alpha \equiv 1/(2-\beta)$. Consider a sequence of positive numbers $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_k = +\infty$, and a sequence of initial conditions $\mathbf{N}^k(0) = [N^k_A(0)]_{A\in\mathbb{C}}$ s.t. the limit $\lim_{k\to\infty} {\lambda_k}^{-\alpha}\mathbf{N}^k(0) = \mathbf{n}^*$ exists. Consider the rescaled process*

$$\mathbf{n}^k(t) = \lambda_k^{-\alpha}\mathbf{N}^k(\lambda_k^{\alpha-1}t), \quad t \in \mathbb{R}_+. \tag{8}$$

Then for all $T > 0$ and all $\epsilon > 0$,

$$\lim_{k\to\infty} \mathbf{P}\big(\inf_{\mathbf{n}\in S(\mathbf{n}^*)} \sup_{t\in[0,T]} |\mathbf{n}^k(t) - \mathbf{n}(t)| \geq \epsilon\big) = 0,$$

i.e., $\mathbf{n}^k$ *converges to a fluid trajectory in probability.*

The above convergence in probability is illustrated in Figure 1: the rescaled process $n^k(t) = \mathbf{1}^T\mathbf{n}^k(t)$ converges to the fluid trajectory as we increase the scaling factor λ_k from 1 to 100. The proof of this theorem can be found in our technical report [16] and follows closely the argument in [10], so we omit it for reasons of brevity.

Given a fluid trajectory $\{\mathbf{n}(t)\}_{t\in\mathbb{R}_+}$, we denote by

$$n_{i,\cdot}(t) = \textstyle\sum_f n_{i,f}(t), \;\; n_{\cdot,i}(t) = \sum_{j,f:i\in f} n_{j,f}(t), \quad i \in \mathbb{K} \tag{9}$$

the (rescaled) population of peers requesting and caching item i, respectively. We call $n_{i,\cdot}$, $n_{\cdot,i}$ the *demand* and the *supply* of item i, respectively.

4.2 Stability of the Static-Cache Policy

Armed with the above system characterization through its fluid trajectories, we turn our attention to the issue of system stability. Intuitively, we wish to understand what is the *stability region* of our system: what conditions should the arrival rates $\lambda_{i,f}$, $(i, f) \in \mathbb{C}$, satisfy, so that the total number of peers in the system remains bounded?

Surprisingly, a universal swarm evolving under the static-cache policy, arguably the simplest policy satisfying the "grab-and-go" principle, has a very wide stability region. We demonstrate this below by studying (a) the stability of the fluid trajectories of the static-cache policy and (b) the ergodicity of the original stochastic system.

We begin by stating our main result regarding fluid trajectories. We say that the system of ODEs (7) is stable if the fluid trajectories $\{\mathbf{n}(t)\}_{t\geq 0}$ remain bounded for all $t \geq 0$ irrespectively of the initial conditions $\mathbf{n}(0)$. In other words, irrespectively of how many peers are originally in the system, the population never blows up to infinity. Denote by

$$\lambda_{i,j} = \textstyle\sum_{f:j\in f} \lambda_{i,f}, \qquad i, j \in \mathbb{K}, \tag{10}$$

the aggregate arrival rate of peers requesting item i and caching item j.

A sufficient condition for stability of the static-cache policy is stated in the following theorem, whose proof can be found in Section 5.1.

THEOREM 2. *Assume that all rates $\lambda_{i,j}$ are positive and*

$$\textstyle\sum_{j:j\neq i} \lambda_{j,i}/\lambda_{i,j} > 1, \quad \forall i \in \mathbb{K}. \tag{11}$$

Then, for all $\beta \in [0, 2)$, the system of ODEs (7) under the static-cache policy is stable.

There are several important conclusions to be drawn from Theorem 2. To begin with, the stability region remains the same for all values of $\beta \in [0, 2)$: this is quite surprising, as it implies that (7) applies to all the different contact regimes we reviewed in Section 3.4 (contact-constrained, constant-bandwidth, and interference-constrained communications).

In addition, recall that $\lambda_{i,\cdot}$ and $\lambda_{\cdot,i}$, given by (1), are the aggregate arrival rates of peers requesting and caching item i, respectively. Inequality (11) implies that the system can be stable even if $\lambda_{i,\cdot} > \lambda_{\cdot,i}$ for some i, *i.e.*, *peers requesting i arrive at a higher rate than peers storing i*. In particular as long as for every i there exists an item j such that $\lambda_{j,i} > \lambda_{i,j}$, then (11) are satisfied, and the system is stable. Intuitively, if such a j exists, the size of its swarm will grow large enough to provide the upload capacity necessary to serve peers requesting item i.

The above theorem has a direct equivalent w.r.t. the stochastic process $\{\mathbf{N}(t)\}_{t\in\mathbb{R}_+}$:

THEOREM 3. *Assume that all rates $\lambda_{i,j}$ are positive and that (11) holds. Then, for all $\beta \in [0, 2)$, the stochastic process $\{\mathbf{N}(t)\}_{t\in\mathbb{R}_+}$ under the static-cache policy is ergodic.*

We provide a proof of this theorem in Section 5.2.

A very interesting aspect of static-cache stability is in the manner in which the universal swarm becomes unstable when (11) is violated. In particular, recall by (9) that $n_{i,\cdot}(t)$ is the demand for item i, *i.e.*, the size of the swarm of peers requesting item i (in the fluid limit). The following theorem then holds:

THEOREM 4. *Assume that all rates $\lambda_{i,j}$ are positive, and $\beta \in [0,2)$. Then there exists at most one item $i \in \mathbb{K}$ for which*

$$\sum_{j:j\neq i} \lambda_{j,i}/\lambda_{i,j} < 1. \quad (12)$$

Moreover, if such an item i exists, there exist initial conditions $\mathbf{n}(0)$ such that

$$\lim_{t\to\infty} n_{i,\cdot}(t) = \infty, \text{ and } \limsup_{t\to\infty} n_{j,\cdot}(t) n_{i,\cdot}^{1-\beta}(t) < \infty, \ \forall j \neq i.$$

In other words, for $\beta \in [0,1]$, *only one swarm can become unstable*. There can be only one item that satisfies (12), and although the swarm of peers requesting this item grows to infinity, the product $n_{j,\cdot} n_{i,\cdot}^{1-\beta}$ remains bounded. As a result, for $\beta \in [0,1]$, no other swarm than the one satisfying (12) can become unstable. This property is very appealing, as it suggests that even if the arrival rates are outside the stability region, the stability of all but one swarm remains unaffected. Note that, when $\beta > 1$, *i.e.*, in the interference-constrained case, the product $n_{j,\cdot} n_{i,\cdot}^{1-\beta}$ is also bounded; however, this does not imply that other swarms do not grow. Nevertheless, these swarms grow at a slower rate than $n_{i,\cdot}$.

This stability property arises precisely because the universal swarm utilizes available bandwidth for inter-swarm communication. Intuitively, a swarm that becomes unstable has unbounded uploading capacity. As a result, as long as all arrival rates are positive, a fraction of this unbounded capacity can be used serve to other swarms at a very high rate; when $\beta \in [0,1]$, this rate is in fact high enough to suppress the growth of any other swarm.

4.3 Optimality Under the "Grab-and-Go" Principle

Despite the interesting stability properties of the static-cache policy, it is still tempting to see whether we can design more sophisticated policies that achieve a wider stability region. Preferably, given that the system is stable we would like a design that minimizes average sojourn time. In this section, we characterize the minimum sojourn time achievable by any system satisfying the "grab-and-go" principle. We will use this to propose a content exchange policy that minimizes the average sojourn time in Section 6.

By Little's Theorem, minimizing the average sojourn time is equivalent to minimizing $N(t)$, the total number of peers in the system. We approach this problem by studying the stationary points of the fluid trajectories. This is a heuristic: by studying the stationary points of (7), we implicitly assume that the Markov process $\{\mathbf{N}(t)\}_{t\in\mathbb{R}_+}$ exhibits some form of concentration around these stationary points. Nevertheless, we believe that there is important intuition to be gained through our approach; we demonstrate that this is indeed the case through our numerical study of a sojourn minimizing system in Section 6.2.

Recall by (9) that $n_{i,\cdot}$ and $n_{\cdot,i}$ are the demand and supply of item i, respectively. The "grab-and-go" principle (5) implies that the fluid trajectories given by (7) satisfy the following set of equations:

$$\dot{n}_{i,\cdot}(t) = \hat{\lambda}_{i,\cdot} - 2\mu \cdot (n(t))^{-\beta} n_{i,\cdot}(t) n_{\cdot,i}(t). \quad i \in \mathbb{K} \quad (13)$$

The above equations state that the swarm of peers requesting i grows with new peer arrivals and decreases at encounters between peers in the swarm and peers that cache i. However, they do not specify what type of conversions take place during other types of encounters between peers. Nevertheless, a stationary point $\mathbf{n} \in \mathbb{R}^{|\mathbb{C}|}$ of (13) must satisfy:

$$n_{i,\cdot} n_{\cdot,i} - \hat{\lambda}_{i,\cdot} (2\mu)^{-1} (n)^{\beta} = 0, \forall i \in \mathbb{K}.$$

Since peers cache at most C items, the number of cached items must be no more than the total cache capacity, *i.e.*,

$$\sum_{i\in\mathbb{K}} n_{\cdot,i} \leq Cn = C \sum_{i\in\mathbb{K}} n_{i,\cdot}$$

We now pose the following problem: among all stationary points of content exchange policies that satisfy the "grab-and-go" principle, which stationary point has the minimum aggregate peer population? More formally, we wish to solve:

$$\begin{aligned}
\text{Minimize} \quad & \textstyle\sum_{i\in\mathbb{K}} n_{i,\cdot} && (14a)\\
\text{subj. to:} \quad & n_{i,\cdot} n_{\cdot,i} - \tfrac{\hat{\lambda}_{i,\cdot}}{2\mu} \big(\textstyle\sum_{i\in\mathbb{K}} n_{i,\cdot}\big)^{\beta} = 0, \quad \forall i \in \mathbb{K} && (14b)\\
& \textstyle\sum_{i\in\mathbb{K}} n_{\cdot,i} \leq C \sum_{i\in\mathbb{K}} n_{i,\cdot} && (14c)\\
& n_{i,\cdot} \geq 0, \ \forall i \in \mathbb{K}. && (14d)
\end{aligned}$$

When $\beta \in [0,1]$, the above problem is convex [3] and its solution is given by the following theorem:

THEOREM 5. *For $\beta \in [0,1]$ and $\rho_i = \hat{\lambda}_{i,\cdot}(2\mu C)^{-1}$ the unique optimal solution to (14) is*

$$n^*_{i,\cdot} = \sqrt{\rho_i} \big(\textstyle\sum_{j:j\in\mathbb{K}} \sqrt{\rho_j}\big)^{\beta/(2-\beta)} \quad (15a)$$

$$n^*_{\cdot,i} = C\sqrt{\rho_i} \big(\textstyle\sum_{j:j\in\mathbb{K}} \sqrt{\rho_j}\big)^{\beta/(2-\beta)}, \quad \forall i \in \mathbb{K}, \quad (15b)$$

The proof of Theorem 5 can be found in Section 5.4. Note that the theorem does not hold for $\beta \in (1,2]$, as (14) is not convex for these values of β. Moreover, (15) describes the optimal steady state demand and supply but not the size of each individual class. By (15), the optimal supply is proportional to the square-root of the aggregate arrival rates of peers requesting this item. This was also observed in the closed caching system described in [5]; our result can thus be seen as an extension of [5] for an open system with peer arrivals and departures. Finally, by (15)

$$n^*_{i,\cdot} = C n^*_{\cdot,i} \quad (16)$$

i.e., the demand is C times the supply. In Section 6, we use this to propose a sojourn-minimizing item-exchange policy.

5. ANALYSIS

5.1 Proof of Theorem 2

Using (6), the ODE (7) for the fluid trajectories under the static cache policy assumes the following simple form.

$$\dot{n}_{i,f}(t) = \hat{\lambda}_{i,f} - 2\mu n(t)^{-\beta} n_{i,f}(t) n_{\cdot,i}(t), \quad i \in \mathbb{K}, f \in \mathbb{F}, \quad (17)$$

where $n_{\cdot,i}(t) = \sum_{j,f:i\in f} n_{j,f}(t)$. The above differential equation has an explicit solution in terms of $g_i(t) := 2\mu n_{\cdot,i}(t) n(t)^{-\beta}$, given by $n_{if}(t) = \big\{n_{if}(0) + \int_0^t \hat{\lambda}_{if} e^{\int_0^s g_i(u)du} ds\big\} e^{-\int_0^t g_i(u)du}$. Consider now the ratio $n_{if}(t)/n_{if'}(t)$ for two distinct indices f, f'. In the view of the previous formula, it reads

$$\frac{n_{if}(t)}{n_{if'}(t)} = \frac{n_{if}(0) + \int_0^t \hat{\lambda}_{if} \exp\big(\int_0^s g_i(u)du\big)\, ds}{n_{if'}(0) + \int_0^t \hat{\lambda}_{if'} \exp\big(\int_0^s g_i(u)du\big)\, ds}.$$

Since the function g_i is non-negative, the argument in the integrals is lower-bounded by a positive constant. As a result, it follows by L'Hospital's rule that $\frac{n_{if}(t)}{n_{if'}(t)} = \frac{\hat{\lambda}_{if}}{\hat{\lambda}_{if'}} + O(1/t)$.

This implies that for large t, the individual variables $n_{ij}(t)$ are related by proportionality constraints, and as a result we can focus on tracking a smaller set of variables. Namely, we introduce the variables $u_i(t) := \frac{n_{i\cdot}(t)}{\hat{\lambda}_{i\cdot}}$. Each individual variable $n_{if}(t)$ verifies $n_{if}(t) = \hat{\lambda}_{if}u_i(t) + O(1/t)$, then

$$\begin{aligned}\dot{u}_i(t) &= 1 - u_i(t)n_{\cdot i}(t)n(t)^{-\beta} \\ &= 1 - u_i(t)\big(\tfrac{\sum_{j\neq i}\hat{\lambda}_{ji}u_j(t)}{[\sum_j \hat{\lambda}_{j\cdot}u_j(t)]^\beta} + O(1/t)\big),\end{aligned}$$

where $\hat{\lambda}_{ij}$ as in (10) and $\hat{\lambda}_{i\cdot}$ as in (1). Hence, for large enough T the evolution of u_i within a finite interval $[T, T+t]$ can be arbitrarily well approximated by the ODE:

$$\dot{u}_i = 1 - u_i\textstyle\sum_{j\neq i}\hat{\lambda}_{ji}u_j\big[\sum_j \hat{\lambda}_{j\cdot}u_j\big]^{-\beta}. \tag{18}$$

We therefore focus on (18)—keeping in mind that our analysis below holds for large enough T. We will show that if (11) is satisfied for every $i \in \mathbb{K}$, then $\sup_i u_i(t)$ is bounded for all t. In particular, the following lemma holds:

LEMMA 1. *For $M>0$ large enough, there exist $\delta>0$ and $\epsilon>0$ s.t. if $\sup_i u_i(0)=M$, then $\sup u_i(M\delta)\leq M(1-\epsilon)$.*

PROOF. To show this, for a given M, fix a $\delta > 0$. If $\sup_i u_i(\delta M) < M(1-\delta)$, then the lemma obviously holds for $\epsilon = \delta$. Suppose thus that there exists an i such that $u_i(\delta M) \geq M(1-\delta)$. By (18), for $t \in [0, \delta M]$ we have $u_i(t) \leq u_i(0) + t \leq M + \delta M$ and $u_i(t) \geq u_i(\delta M) + t - \delta M \geq M(1-\delta) - \delta M$. Hence $u_i(t) \in [M(1-2\delta), M(1+\delta)]$. This in turn implies that, for $t \in [0, \delta M]$, $n(t) = \Theta(M)(1+O(\delta))$, where the constants involved depend on $\hat{\lambda}_i$ but not on t. As a result, for $j \neq i$, and $t \in [0, \delta M]$, we have

$$\dot{u}_j(t) = 1 - u_j\Theta(M^{1-\beta})(1+\epsilon_1(\delta)),$$

where $\epsilon_1(\delta) = 1 - \frac{1+O(\delta)}{(1+O(\delta))^\beta} = O(\delta)$. Thus for $t \in [0, \delta' M]$, where $\delta' < \delta$, we have

$$\begin{aligned}u_j(t) &= [u_j(0) + \int_0^t e^{s\Theta(M^{1-\beta})(1+\epsilon_1(\delta))ds}]e^{-t\Theta(M^{1-\beta})(1+\epsilon_1(\delta))} \\ &= u_j(0)e^{-t\Theta(M^{1-\beta})(1+\epsilon_1(\delta))} + \frac{1-e^{-t\Theta(M^{1-\beta})(1+\epsilon_1(\delta))}}{\Theta(M^{1-\beta})(1+\epsilon_1(\delta))}.\end{aligned}$$

Fix a $0 < \delta' < \delta$, then

$$\begin{aligned}u_j(\delta' M) &= O(Me^{-\Theta(\delta' M^{2-\beta})(1+\epsilon_1(\delta))}) + \frac{1-e^{-\Theta(\delta' M^{2-\beta})(1+\epsilon_1(\delta))}}{\Theta(M^{1-\beta})(1+\epsilon_1(\delta))} \\ &= \Theta(M^{-(1-\beta)})(1+\epsilon_2(M,\delta,\delta')),\end{aligned}$$

where $\epsilon_2 = O(\epsilon_1(\delta) + M^{2-\beta}e^{-\Theta(\delta' M^{2-\beta})(1+\epsilon_1(\delta))})$. From this and (18) we get that for $t \in [\delta' M, \delta M]$

$$\dot{u}_j(t) = 1 - u_j\hat{\lambda}_{i,j}\hat{\lambda}_{i,\cdot}^{-\beta}M^{1-\beta}(1+\epsilon_3(M,\delta,\delta')),$$

where $\epsilon_3 = O(\delta + O(M^{\beta-2}) + O(M^{\beta-2}\epsilon_2) = O(\delta) + O(M^{\beta-2}) + O(e^{-\Theta(\delta' M^{2-\beta})(1+\epsilon_1(\delta))})$. From this refined bound on the ODE, we can repeat the steps above to get that for $t \in [\delta'' M, \delta M]$, where $\delta' \leq \delta'' < \delta$, we have

$$u_j(t) = \hat{\lambda}_{i,\cdot}^{\beta}\hat{\lambda}_{i,j}^{-1}M^{\beta-1}(1+\epsilon_4(M,\delta,\delta',\delta'')),$$

where $\epsilon_4 = O(\epsilon_3) + O(M^{2-\beta}e^{-\Theta(\delta'' M^{2-\beta})})$. As a result, for $t \in [\delta'' M, \delta M]$,

$$\begin{aligned}\dot{u}_i(t) &= 1 - u_i\frac{\sum_{k\neq i}\hat{\lambda}_{ki}u_k}{n^\beta} \\ &= 1 - u_i(t)\frac{\sum_{k\neq i}\hat{\lambda}_{k,i}\frac{\hat{\lambda}_{i\cdot}^\beta}{\hat{\lambda}_{ik}}[M^{\beta-1}(1+\epsilon_4)]}{[\hat{\lambda}_{i\cdot}M(1+O(M^{2-\beta})(1+\epsilon_4)]^\beta} \\ &= 1 - u_i\sum_{k\neq i}\frac{\hat{\lambda}_{k,i}}{\hat{\lambda}_{i,k}}[M(1+\epsilon_5(M,\delta,\delta',\delta''))]^{-1},\end{aligned}$$

for $\epsilon_5 = O(\epsilon_4) + O(M^{2-\beta})$. Let $\gamma_i = \sum_{k\neq i}\frac{\hat{\lambda}_{ki}}{\hat{\lambda}_{ik}} > 1$, by (11). Then

$$u_i(\delta M) = u_i(\delta'' M)e^{-\gamma_i(1+\epsilon_5)(\delta-\delta'')} + M\frac{1-e^{-\gamma_i(1+\epsilon_5)(\delta-\delta'')}}{\gamma_i(1+\epsilon_5)}.$$

By a Taylor expansion, $u_i(\delta M)$ becomes

$$\begin{aligned}&u_i(\delta'' M)[1-\gamma_i(1+\epsilon_5)(\delta-\delta'')+O(\delta^2)] + M[(\delta-\delta'')+O(\delta^2) \\ \leq& M[1+\delta''+(\delta-\delta'')[1-\gamma_i(1+\epsilon_5)]+O(\delta^2)]\end{aligned}$$

as $u_i(\delta'' M) \leq M(1+\delta'')$ by (18). Assume now that M is large, and set $\delta = \Theta(M^{(\beta-2)/2})$ and δ', δ'' to be proportional to δ, such that $\delta' < \delta'' < \delta$ and $\delta'' + (\delta - \delta'')[1-\gamma_i] < 0$. It then follows that $\epsilon_5 = O(\delta)$. Hence for large enough M (and small enough δ) $u_i(\delta M) = M[1+\delta''+(\delta-\delta'')[1-\gamma_i]+O(\delta^2)] < 0$ and the lemma follows. □

Hence, outside a bounded set, $\sup_i u_i$ has to decrease (*i.e.*, is a Lyapunov function), and the theorem follows. □

5.2 Proof of Theorem 3

We now establish that under condition (11), the original Markov process $\mathbf{N}(t)$ is ergodic. To this end, we shall rely on the *fluid limit* approach. That is to say, we shall identify a Lyapunov function F, and establish that, for initial condition $\mathbf{N}(0)$ such that $F(\mathbf{N}(0)) = M$, then , for large enough M, it holds that

$$\mathbf{E}F(\mathbf{N}(\delta M)) \leq (1-\epsilon)M, \tag{19}$$

for suitable positive constants $\delta, \epsilon > 0$. Unsurprisingly, the line of argument parallels that of Theorem 2's proof, with some additional elements introduced to take care of the random fluctuations in the process.

The Lyapunov function to be considered is

$$F(N) := \sup_{i\neq j} N_{ij}/\lambda_{ij}.$$

Define the event $\Omega_{ij} = \{N_{ij}(M\delta) \geq \lambda_{ij}M(1-\delta)\}$. We first establish the following intermediate result.

LEMMA 2. *On the event Ω_{ij}, for some positive constants $\gamma, c > 0$, for all $k \neq i$, with probability $1 - e^{-\Theta(M)}$ one has*

$$N_{ij}(t) \in [\lambda_{ij}M(1-c\delta), \lambda_{ij}M(1+c\delta)],\ t \in [0, M\delta], \tag{20}$$

$$N_{ik}(M\delta) \in [\gamma M, \lambda_{ik}M(1+c\delta)], k \neq j. \tag{21}$$

PROOF. Let E_{ik} denote the unit rate Poisson processes used to generate the arrival times of type (ik)-users in the system. Consider the event $\Omega_1 = \{|E_{ik}(\lambda_{ik}M\delta) - \lambda_{ik}M\delta| \leq M\lambda_{ik}\delta/2, k \neq i\}$. Then using Chernoff bounds, it is readily seen that its probability is at least $1 - e^{-\Theta(M)}$.

To establish (20), it suffices to note that

$$N_{ij}(M\delta) - E_{ij}(\lambda_{ij}M\delta) \leq N_{ij}(t) \leq N_{ij}(0) + E_{ij}(M\delta),$$

and on the event $\Omega_1 \cap \Omega_{ij}$, the left-hand side is at least $\lambda_{ij}M(1-(5/2)\delta)$ and the right-hand side is at most $\lambda_{ij}M(1+(3/2)\delta)$. (20) thus holds with $c = 5/2$.

Consider next $k \neq j$. We introduce now the notation $D_i(t)$ to represent the number of departures of users requesting object i in time interval $[0,t]$. On the event Ω_1, necessarily $D_i(M\delta) \leq rM$ for some suitable constant r. Indeed, $D_i(M\delta)$ cannot exceed $N_{i\cdot}(0) + \sum_{k\neq i} E_{ij}(M\delta\lambda_{ij})$, which in turn is no larger than $\sum_{k\neq i} M\lambda_{ik}(1+(3/2)\delta)$ on Ω_1, given the initial condition $F(N(0)) = M$.

Introduce now $D_{ik}(t)$ to represent the number of departures of type (ik)-users during time interval $[0,t]$. This process is generated as follows: at each jump time T of the counting process $D_i(\cdot)$, conditional on the past of the process before time T, a type (ik)-user is chosen to leave the system with probability $N_{ik}(T^-)/N_{i\cdot}(T^-)$. An explicit construction of this selection mechanism can be made by attaching a uniform random variable U_n to each jump point T_n of the process D_i in $[0, M\delta]$, and by letting

$$D_{ik}(t) = \sum_{n:T_n \geq t} \mathbb{1}_{U_n < N_{ik}(T_n^-)/N_{i\cdot}(T_n^-)}.$$

As previously established, on the event $\Omega_1 \cap \Omega_{ij}$, one has $N_{i\cdot}(t) \geq M\gamma$ for all $t \in [0, M\delta]$, and $D_i(M\delta) \leq rM$. This entails that, on this event, $D_{ik}(M\delta) \leq \sum_{n=1}^{rM} Z_n$, where $Z_n := \mathbb{1}_{U_n < \left(X - \sum_{\ell=1}^{n-1} Z_\ell\right)/M\gamma}$, and $X := N_{ik}(0) + E_{ik}(\lambda_{ik} M\delta)$. Indeed, type (ik)-departures are more likely if arrivals occur at the beginning of the interval $[0, M\delta]$.

This yields a first lower bound:

$$N_{ik}(M\delta) \geq Y := X - \sum_{n=1}^{rM} Z_n. \tag{22}$$

To simplify this further, one can note that the resulting random variable Y is stochastically reduced if one replaces X in both this expression and the definition of the random variables Z_n by a lower bound. On the event Ω_1, such a lower bound consists in $M\rho$ with $\rho = \lambda_{ik}\delta/2$.

We now control the probability that the lower bound Y in (22) is below a threshold τM for some constant $\tau > 0$, taking $X = M\rho$. We have the following representation: $\mathbf{P}(Y < \tau M) = \mathbf{P}(\sum_{n=0}^{(\rho-\tau)M} V_n \leq rM)$, where the random variables V_n are independent, geometrically distributed with parameter $(\rho M - n)/(\gamma M)$. We omit details, but Chernoff's bounding technique can be used, by evaluating the Laplace transform of the random variable $\sum_{n=0}^{(\rho-\tau)M} V_n$, to show that, for small enough constant $\tau > 0$, the probability $\mathbf{P}(Y < \tau M)$ is at most $\exp(-\Theta(M))$. This concludes the proof of the Lemma. □

We next need the following Corollary.

Corollary: On the event Ω_{ij}, for any $\delta' < \delta$, with probability $1 - \exp(-\Theta(M))$, the following holds for all $k \neq i$:

$$N_{k\cdot}(M\delta') \leq \mathrm{Bin}(O(M), e^{-\Theta(M^{2-\beta})}) + \mathrm{Poi}(\Theta(M^{\beta-1})),$$

where Bin denotes a Binomial random variable, Poi a Poisson variable, that are mutually independent.

Proof. On Ω_{ij}, with probability $1 - e^{-\Theta(M)}$ it holds that $N_{ij}(M\delta') \geq M(1-(5/2)\delta)$. The previous Lemma, suitably modified, therefore applies, and thus, there must exist a constant $\delta'' < \delta'$ such that with probability $1 - e^{-\Theta(M)}$, the following holds: $N_{ik}(t) = \Omega(M),\ t \in [M\delta'', M\delta']$. Consider now the dynamics of $(N_{k\cdot})$. Arrivals occur at a rate $\lambda_{k\cdot}$, and departures occur at a time-varying rate $N_{k\cdot}(t)N_{\cdot k}(t)N(t)^{-\beta}$. The product $N_{\cdot k}(t)N(t)^{-\beta}$ is at least $\Omega(M^{1-\beta})$ on the interval $[\delta'' M, \delta' M]$ by the previous argument. Thus its state at time $M\delta'$ can be upper-bounded by that of a M/M/∞/∞ queue, with initial state $N_{k\cdot}(M\delta'')$ at time $M\delta''$, arrival rate $\lambda_{k\cdot}$ and death rate $\Omega(M^{1-\beta})$. Now, with probability $1 - e^{-\Theta(M)}$, it holds that $N_{k\cdot}(M\delta'') = O(M)$, and the result follows. □

We are now ready to conclude the proof of the Theorem. To this end, we place ourselves on the event Ω_{ij}, and derive bounds on the trajectories $N_{ij}(t)$ for t in the interval $[M\delta', M\delta]$, relying on the previous results.

As we have just seen, with probability $1 - e^{-\Theta(M)}$, the components $N_{ik}(M\delta')$ are of order $\Theta(M)$. Furthermore, following the same lines as in the proof of (21), we can deduce from the fact that $N_{ij}(t) = N_{ij}(0)(1+O(\delta)),\ t \in [0, M\delta]$ that

$$N_{ik}(t) = N_{ik}(M\delta')(1+O(\delta)),\ t \in [M\delta', M\delta]. \tag{23}$$

Let us now introduce dedicated unit rate Poisson processes $\Delta_{k\ell}$ for each user type $(k\ell)$, and consider the representation

$$\begin{aligned} N_{k\ell}(t) = \quad & N_{k\ell}(M\delta') + E_{k\ell}(\lambda_{k\ell}(t - M\delta')) \\ & -\Delta_{k\ell}(\mu \textstyle\int_{M\delta'}^t N_{k\ell}(s) N_{\cdot k}(s) N(s)^{-\beta} ds). \end{aligned}$$

Replacing in the above $N_{\cdot k}(s)$ by an upper bound of order $N_{ik}(M\delta')(1+O(\delta))$, and $N(s)$ by a lower bound of order $N_{i\cdot}(M\delta')(1+O(\delta))$, we obtain a process $N^+_{k\ell}(t)$ that is an upper bound to $N_{k\ell}(t)$, and that is an $M/M/\infty/\infty$ process with arrival rate $\lambda_{k\ell}$ and death rate $N_{ik}(M\delta')N_{i\cdot}(M\delta')^{-\beta}(1+O(\delta))$.

Subsequently, we can also derive lower-bounding processes $N^-_{k\ell}(t)$ by upper-bounding $N_{\cdot k}(s)$ by

$$N_{\cdot k}(s) \geq N_{ik}(M\delta')(1+O(\delta)) + \textstyle\sum_{m\neq i,k} N^+_{mk}(s),$$

and lower-bounding $N(s)$ by $N(M\delta')(1+O(\delta))$ in the argument of $\Delta_{k\ell}$. Note now that the processes $N^+_{k\ell}$ have a stationary distribution that is Poisson with parameter $O(M^{\beta-1})$. Thus with high probability, their supremum over $[M\delta', M\delta]$ is small compared to $N_{i\ell}$, itself of order M. Eventually, we obtain that with high probability, $N_{k\ell}(t)$ admits lower bounds $N^-_{k\ell}(t)$ that are $M/M/\infty/\infty$ processes with arrival rate $\lambda_{k\ell}$ and death rates again equal to

$$N_{ik}(M\delta')N_{i\cdot}(M\delta')^{-\beta}(1+O(\delta)).$$

These lower bounds in turn will provide upper bounds on N_{ik}, by writing

$$\begin{aligned} N_{ik}(t) \leq \quad & N_{ik}(M\delta') + E_{ik}(\lambda_{ik}(t - M\delta')) \\ & -\Delta_{ik}(\textstyle\int_{M\delta'}^t N_{ik}(s)[\sum_{\ell\neq i} N^-_{\ell i}(s)]N(s)^{-\beta} ds. \end{aligned} \tag{24}$$

The argument of Δ_{ik} is lower-bounded by

$$N_{ik}(M\delta')N_{i\cdot}(M\delta')^{-\beta}(1+O(\delta)) \int_{M\delta'}^t \sum_{\ell\neq i} N^-_{\ell i}(s) ds.$$

By the ergodic theorem, applied to the $M/M/\infty/\infty$ processes $N^-_{\ell i}$, this integral reads with high probability with respect to M:

$$(t - M\delta') \sum_{\ell\neq i} \lambda_{\ell i} \frac{N_{i\cdot}(M\delta')^\beta}{N_{i\ell}(M\delta')}(1+O(\delta)).$$

Upon simplification, we have with high probability, replacing in (24) the Poisson processes E_{ik} and Δ_{ik} by their expectation, up to some error vanishing as M increases,

$$\begin{aligned} N_{ik}(M\delta) - N_{ik}(M\delta') \leq \quad & M(\delta - \delta')\,[\lambda_{ik} - -(1+O(\delta)) \times \dots \\ & \dots N_{ik}(M\delta') \textstyle\sum_{\ell\neq i} \frac{\lambda_{\ell i}}{N_{i\ell}(M\delta')}\Big]. \end{aligned}$$

Let $a_{ik}(t) = \lambda_{ik}^{-1} N_{ik}(t)$. The previous equation reads

$$a_{ik}(M\delta) - a_{ik}(M\delta') \le M(\delta - \delta') \times \cdots$$
$$\cdots \times \Big[1 - a_{ik}(M\delta') \textstyle\sum_{\ell \neq i} \frac{\lambda_{\ell i}}{\lambda_{i\ell}} \frac{1}{a_{i\ell}(M\delta')}(1 + O(\delta))\Big].$$

Now, for the index k for which $a_{ik}(M\delta')$ is largest,the right-hand side of the above is no larger than

$$M(\delta - \delta')\big[1 - (1 + O(\delta)) \sum_{\ell \neq i} \frac{\lambda_{\ell i}}{\lambda_{i\ell}}\big],$$

itself strictly smaller than $-M\epsilon$ for some positive ϵ, if we chose δ small enough, when condition (11) is in force. This enables to conclude that, with high probability,

$$\sup_{i,j} N_{ij}(M\delta)/\lambda_{ij} = F(N(M\delta)) \le (1 - \epsilon) F(N(0)).$$

The same bound applies to the expectation of the left-hand side, using a uniform integrability argument. This establishes the desired contraction property of the Lyapunov function F and, hence, the ergodicity of the original Markov process. □

5.3 Proof of Theorem 4

To show that there can be at most one i for which (12) holds, we observe that if it holds for some i, then for any other $j \neq i$, we have $\hat{\lambda}_{j,i}/\hat{\lambda}_{i,j} < 1$. This implies that any other j satisfies (11), so no $j \neq i$ can also satisfy (12).

To prove the remainder of the theorem, we use the notation $z = x \pm y$ to indicate that $z \in [x - y, x + y]$. Suppose that $\gamma_i = \sum_{k \neq i} \hat{\lambda}_{ki}/\hat{\lambda}_{ik} < 1$ for some i and assume that $u_i(0) = M > 0$, for some large M. Assume further that for $j \neq i$, $u_j(0) = u_i^{1-\beta} \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1 \pm \epsilon)$, for some small $\epsilon > 0$. The following lemma then holds:

LEMMA 3. *For all $\epsilon > 0$, there exists $M_0 > 0$ s.t. for all $M > M_0$ there exists $\delta > 0$ such that if $u_i(0) = M$ and $u_i^{1-\beta}(0)u_j(0) = (1 \pm \epsilon)\hat{\lambda}_{i\cdot}^{\beta}/\hat{\lambda}_{i,j}$, then $u_j(t)u_i^{1-\beta}(t) = (1 \pm \epsilon)\hat{\lambda}_{i\cdot}^{\beta}/\hat{\lambda}_{i,j}$ for all $t \in [0, \delta]$.*

PROOF. Fix some $\epsilon > 0$. The lemma follows by the continuity of the fluid trajectories if $u_i^{1-\beta}u_j$ is in the interior of $\frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1 \pm \epsilon)$. Suppose thus that it is at the boundary. We consider the upper boundary case, *i.e.*, $u_j u_i^{1-\beta} = \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1+\epsilon)$ and show that $\frac{d}{dt}u_j u_i^{1-\beta}$ is negative, so that $u_j u_i^{1-\beta}$ is forced in the interior; the same argument can be used to show that the derivative is positive when at the lower boundary, so we omit this case. Indeed

$$\frac{d}{dt}u_i^{1-\beta}u_j = (1-\beta)\dot{u}_i u_i^{-\beta}u_j + u_i^{1-\beta}\dot{u}_j$$
$$= (1-\beta)u_i^{-\beta}u_j(1 - u_i\frac{\sum_{k\neq i}\hat{\lambda}_{ki}u_k}{n^\beta}) + u_i^{1-\beta}(1 - u_j\frac{\sum_{k\neq j}\hat{\lambda}_{kj}u_k}{n^\beta}).$$

Note that $u_j(0) = \Theta(u_i^{\beta-1}(0))$, where the asymptotic notation is as $M \to \infty$. Observe that, since $\hat{\lambda}_{i',j'} > 0$ for all $i', j' \in \mathbb{K}$, we have that at time $t = 0$, $0 < \frac{\sum_{k\neq i}\hat{\lambda}_{ki}u_k}{n^\beta} = \Theta(u_j u_i^{-\beta}) = \Theta(u_i^{-1})$, and $0 < \frac{\sum_{k\neq j}\hat{\lambda}_{kj}u_k}{n^\beta} = u_i^{1-\beta}\frac{\hat{\lambda}_{i,j}}{\hat{\lambda}_{i,\cdot}^{\beta}}(1 + o(1))$. We thus have that, at $t = 0$,

$$\frac{d}{dt}u_i^{1-\beta}u_j = \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1+\epsilon)\Big(\frac{1}{u_j} - u_i^{1-\beta}\frac{\hat{\lambda}_{i,j}}{\hat{\lambda}_{i,\cdot}^{\beta}}(1 + o(1)) + O(u_i^{-1})\Big)$$
$$= \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1+\epsilon)\Big(\frac{1}{u_j} - u_i^{1-\beta}\frac{\hat{\lambda}_{i,j}}{\hat{\lambda}_{i,\cdot}^{\beta}}(1 + o(1))\Big)$$

as $u_i^{-1} = o(u_i^{1-\beta})$ for $\beta < 2$. On the other hand as $u_j u_i^{1-\beta} = \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1+\epsilon)$ implies that $u_j^{-1} = u_i^{1-\beta}\frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}\frac{1}{1+\epsilon} < u_i^{1-\beta}\frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}$, so for M large enough the above quantity is negative. □

Consider now a fluid trajectory in which $u_i(0) = M$ and $u_i^{1-\beta}(0)u_j(0) = \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1 \pm \epsilon)$. Then we have that

$$\dot{u}_i = 1 - u_i\frac{\sum_{k\neq i}\hat{\lambda}_{ki}u_k}{n^\beta} = 1 - u_i\frac{\sum_{k\neq i}\hat{\lambda}_{ki}\frac{\hat{\lambda}_{i,\cdot}^{\beta}}{\hat{\lambda}_{i,k}}u_i^{\beta-1}1 \pm \epsilon}{(\sum_k \hat{\lambda}_{k,\cdot}u_k)^\beta}$$
$$= 1 - u_i\frac{\sum_{k\neq i}\frac{\hat{\lambda}_{k,i}}{\hat{\lambda}_{i,k}}\hat{\lambda}_{i\cdot}^{\beta}u_i^{\beta-1}(1 \pm \epsilon)}{\hat{\lambda}_{i\cdot}^{\beta}u_i^{\beta}(1 + o(1))},$$

which for large enough M and a small enough ϵ becomes $1 - \sum_{k\neq i}\frac{\hat{\lambda}_{ki}}{\hat{\lambda}_{ik}}(1 + o(1)) > 0$. This, along with Lemma 3 implies we can select an $\epsilon > 0$ such that, for large enough M, if $u_i(0) = M$ and $u_i^{1-\beta}(0)u_j(0) = \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1 \pm \epsilon)$, then there exists a $\delta > 0$ s.t. $u_i'(0)$ is positive and bounded away from zero uniformly in M and $u_i^{1-\beta}(t)u_j(t) = \frac{\hat{\lambda}_{i\cdot}^{\beta}}{\hat{\lambda}_{i,j}}(1 \pm \epsilon)$, for all $t \in [0, \delta]$. This in turn implies that the above is true for all $t \ge 0$, and, in particular, that u_i diverges to infinity. □

5.4 Proof of Theorem 5

Let us define $x_i = n_{i,\cdot}$ and $\rho_i = \hat{\lambda}_{i,\cdot}(2\mu C)^{-1}$, $i \in \mathbb{K}$. By (14b), we have

$$n_{\cdot,i} = C\rho_i(\textstyle\sum_{j:j\in\mathbb{K}} x_j)^\beta / x_i. \tag{25}$$

Using (25), we can rewrite (14) as the following equivalent convex optimization problem involving only x_i:

$$\begin{aligned} \text{Minimize} \quad & \textstyle\sum_{i\in\mathbb{K}} x_i \\ \text{subj. to:} \quad & \textstyle\sum_{i\in\mathbb{K}}(\rho_i x_i^{-1}) \le (\sum_{i\in\mathbb{K}} x_i)^{1-\beta}, \quad i \in \mathbb{K} \quad (26a) \\ & x_i \ge 0, \quad i \in \mathbb{K}. \quad (26b) \end{aligned}$$

We can write its Lagrangian function as

$$\Lambda(\mathbf{x}, \varphi, \mathbf{w}) = \textstyle\sum_{i\in\mathbb{K}} x_i + \varphi h(\mathbf{x}) + \sum_{i\in\mathbb{K}} w_i g(x_i),$$

where φ and $\mathbf{w} = [w_i]_{i\in\mathbb{K}}$ are Lagrangian multipliers, $\mathbf{x} = [x_i]_{i\in\mathbb{K}}$, $h(\mathbf{x}) = \sum_{i\in\mathbb{K}}\rho_i x_i^{-1} - (\sum_{i\in\mathbb{K}} x_i)^{1-\beta}$, and $g(x_i) = -x_i$. Hence, any $\tilde{\mathbf{x}} = \{\tilde{x}_1, ..., \tilde{x}_K\}$ is optimal if and only if it satisfies the following KKT conditions [3]:

$$h(\tilde{\mathbf{x}}) \le 0, \quad g(\tilde{x}_i) \le 0, \quad i \in \mathbb{K}, \tag{27a}$$
$$\varphi \ge 0, \quad w_i \ge 0, \quad i \in \mathbb{K}, \tag{27b}$$
$$\varphi h(\tilde{\mathbf{x}}) = 0, \quad w_i g(\tilde{x}_i) = 0, \quad i \in \mathbb{K}, \tag{27c}$$
$$\frac{d\Lambda}{dx_i}(\tilde{\mathbf{x}}, \varphi, \mathbf{w}) = 0, \quad i \in \mathbb{K}. \tag{27d}$$

We know that $x_i > 0, \forall i \in \mathbb{K}$ from (26a) and (26b). Thus condition (27c) requires $w_i = 0, \forall i \in \mathbb{K}$ and (27d) needs $\varphi \neq 0$. Then by $\varphi h(\tilde{x}) = 0$ in (27c) and condition (27d), any optimal solution $\tilde{\mathbf{x}}$ must satisfy the following two equations:

$$h(\tilde{\mathbf{x}}) = 0, \quad \frac{d\Lambda}{dx_i}(\tilde{\mathbf{x}}, \varphi, [0]) = 0$$

Solving these two equations leads to the unique solution $x_i = \sqrt{\rho_i}(\sum_{j:j\in\mathbb{K}}\sqrt{\rho_j})^{\frac{\beta}{2-\beta}}, i \in \mathbb{K}$, as shown in (15a), and the Lagrangian multiplier $\varphi = (\sum_{i\in\mathbb{K}}\sqrt{\rho_i})^{\frac{2\beta}{2-\beta}}(2-\beta)^{-1}$.

By plugging x_i into (25), we can derive the value of $n^*_{\cdot,i}$ as (15b), and (15) is the unique optimal solution of (14). □

6. BARON: GUIDING CACHE REPLACEMENT VIA VALUATIONS

Our analysis in Section 5.4 has identified the optimal stationary points that minimize the average sojourn time. However, we have not described a method for leading the system to such points. In this section, we present BARON to bridge this gap. BARON dictates how peers should exchange content items so that the system converges to the optimal points defined in Theorem 5. We also demonstrate BARON's performance using numerical simulations.

BARON is a centralized scheme. In particular, it requires estimating the demand and supply of each item $i \in \mathbb{K}$, captured by the population of peers requesting and storing i, respectively. In practice, individual peers may maintain estimates of these quantities, *e.g.*, either by gossiping or sampling. However, studying decentralized schemes for estimating the demand and supply is beyond the scope of this paper. As a result, we focus on scenarios in which these quantities are readily monitored through at a centralized tracker.

6.1 Designing BARON

To lead the system to the optimal point, one intuitive way is to first identify which items are over-replicated and which are under-replicated. Whenever two peers come into contact, if one has an over-replicated item i and the other has an under-replicated item j, then the first peer replaces its item i with item j. This replacement increases the current supply $n_{\cdot,i}$ of the under-replicated item.

Valuations in BARON. BARON keeps track of whether an item is currently over-replicated or under-replicated in following way. In particular, for each content item i, BARON maintains a real-valued variable v_i. We will call this variable the *valuation* of item i.

Our choice of valuation is inspired by (16), which states that at an optimal point the supply of an item is C times the demand. Motivated by this, the valuations are given by

$$v_i(t) = Cn_{i,\cdot}(t) - n_{\cdot,i}(t), \quad i \in \mathbb{K}. \tag{28}$$

A positive valuation $v_i > 0$ indicates that item i is currently under-replicated. Similarly, a negative valuation $v_i < 0$ indicates that item i is currently over-replicated.

One appealing property of (28) is that it requires prior knowledge *only* of the cache capacity C; in particular, it does not require knowledge of the arrival rates $\lambda_{i,f}$ of each peer class. Nevertheless, this valuation requires to track the supply and demand for each item.

Content exchange guided by valuations. BARON is a centralized design that relies on a central controller to maintain the valuations (28). In addition, this central controller lists the valuations on a public board, and makes them available to all peers.

The content exchanges between peers are guided by these valuations following a *negative-positive rule*. More specifically, during a contact event between a peer A with cache f and a peer B with f', each peer checks if it has any over-replicated items. If so, it further checks whether the other peer has any under-replicated items that it has not already stored in its cache. If such a pair of items exists, a replacement takes place. In particular, the first peer A replaces the item with the minimal negative valuation in its cache, *i.e.*, peer A removes item i such that

$$i = \text{argmin}\{v_x | x \in f, v_x < 0\}.$$

Then, among the under-replicated items in the peer B's cache f' yet not in peer A's cache f, peer A replicates the item with the maximal positive valuation, *i.e.* peer A selects item j such that

$$j = \text{argmax}\{v_y | y \in f' \setminus f, v_y > 0\}.$$

After retrieving item j from peer B, peer A replaces i with j. Hence its cache f changes to $(f \setminus \{i\}) \bigcup \{j\}$. A similar procedure follows for peer B.

Clearly, there are other ways to design valuations and the rules for guiding content exchanges via valuations. In Section 6.3, we will examine other options for these two design components of BARON.

Based on the above definitions, the conversion probabilities of BARON satisfy Assumption 1 because of the positive-negative rule. As a result, by Theorem 1, we can study the dynamics of BARON through its fluid trajectory.

6.2 Evaluating BARON

We evaluate BARON's fluid trajectories using numerical simulations in MATLAB. Our main observation is that, by guiding the content exchanges through valuations, BARON converges to the optimal stationary points defined in (15), which minimize the average sojourn time.

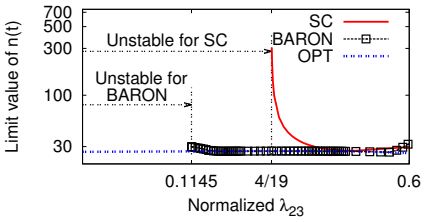
6.2.1 BARON vs. Static-Cache Policy

We compare BARON to the static-cache policy by the examining system stability and optimality when using each design. Then we further use the static-cache policy as an example to demonstrate that only one swarm becomes unstable when instability occurs.

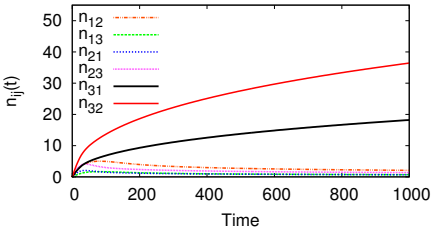
We simulate the following scenario. Assume there are three items $\{1, 2, 3\}$ in the system, and peer's cache size is one. Hence we have six peer classes, where each class of peers requesting item i and caching item j $(\neq i)$ has a normalized arrival rate of $\hat{\lambda}_{i,j}$. Peers requesting one item form one swarm, leading to three swarms in total. We set the contact process parameters as $\beta = 0$ and $\mu = 0.002$. We assume initially no peer is in the system.

Stability. We begin with comparing the system stability under BARON and the static-cache policy. In particular, we aim to understand under which conditions of arrival rates, the system stabilizes when using each design. So we leave $\hat{\lambda}_{23}$ as a free variable, and fix the relative ratios of the other five classes as $\frac{1}{5}, \frac{1}{15}, \frac{2}{15}, \frac{1}{5}, \frac{2}{5}$, respectively of $(1 - \hat{\lambda}_{23})$. To identify the system stability for a given $\hat{\lambda}_{23}$ value, we examine the system's fluid trajectory over a significantly long time ($t \approx 10^5$).

Figure 2(a) shows the rescaled peer population when the system can stabilize as we vary $\hat{\lambda}_{23}$. We see that when using the static-cache policy, the system stabilizes only when $\hat{\lambda}_{23}$ is above $\frac{4}{19}$. This verifies the conclusion in Theorem 2 since $\frac{4}{19}$ is the arrival rate $\hat{\lambda}_{23}$ that violates condition (11). In con-

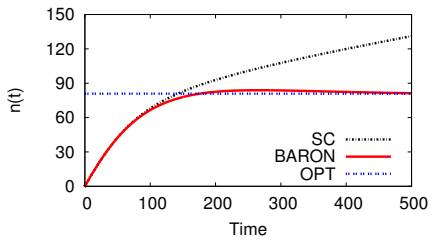


(a) Limit points of the fluid trajectory for variable $\hat{\lambda}_{23}$.

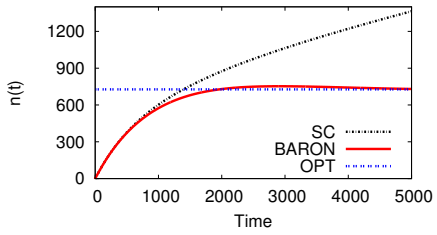


(b) Static-cache policy w/ $\hat{\lambda}_{23} = \frac{4}{19}$.

Figure 2: Comparing BARON and static-cache (SC) policy by varying arrival rate configurations.

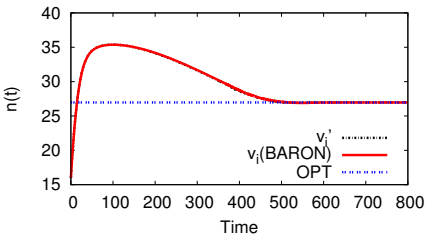


(a) Contact-constrained communication w/ $\beta = 0.5$.

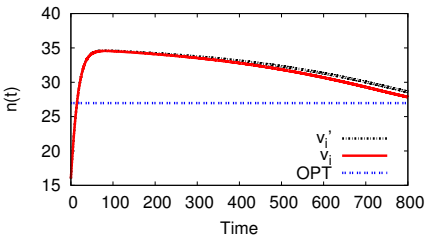


(b) Constant-bandwidth communication w/ $\beta = 1$.

Figure 3: BARON under various β.



(a) Varying valuation designs under NPR.



(b) Varying valuation designs under LHR.

Figure 4: Examining the peer population under various design options in BARON.

trast, when using BARON, the system has a much larger stability region. More specifically, the system is able to stabilize when $\hat{\lambda}_{23}$ is larger than 0.1145. This demonstrates BARON's effectiveness of guiding content exchanges.

Optimality. We further examine the stationary state that the system converges to when using BARON and the static-cache policy. As shown in Figure 2(a), the system under the static-cache policy converges to a non-optimal state. Moreover, the closer $\hat{\lambda}_{23}$ is to the stability boundary $\frac{4}{19}$, the more peers in the stationary state. In contrast, BARON is able to guide the system to the optimal stationary state if the system stabilizes. This demonstrates that BARON achieves optimality by the use of valuations.

Single swarm instability. Now we examine how peer classes evolve in time when instability occurs under the static-cache policy. Figure 2(b) shows the population of each peer class along the time when $\hat{\lambda}_{23} = \frac{4}{19}$, demonstrating the conclusion in Theorem 4: only *one* swarm can become unstable.

Recall that peers requesting the same item form one swarm. Our main observation is that only the swarm requesting item 3 blows up. This is because item 3 is the one that does not satisfy (11). As this swarm grows, peers in other swarms can obtain their requested items quickly and depart. Hence the supply for item 3 further decreases.

6.2.2 *Dependence on β*

To comprehensively understand BARON's performance, we extend to cases with other β values. In particular, we examine two cases with $\beta = 0.5$ and $\beta = 1$ respectively. As shown in Section 3.4, a larger β indicates a smaller contact rate. The case when $\beta = 1$ is the constant-bandwidth communication scenario where a peer's contact rate is constant regardless of the peer population. We do not simulate the case where $\beta > 1$, because the optimality result identified in Section 4.3 does not hold for such β. We configure the other parameters as in Figure 2 with $\hat{\lambda}_{23} = \frac{1}{6}$.

Figures 3(a) and (b) show the evolution of the total number of peers in time. The main observation is that, while the system does not stabilize when using the static-cache policy, the system under BARON converges to the optimal in both cases. This demonstrates the effectiveness of the valuations under various communication settings. Even though the aggregate contact rate decreases as β increases, BARON is still able to adapt the item supply according to the demand, guiding the system towards the optimal.

Furthermore, as the contact rate becomes smaller when β increases, the system with BARON takes longer time to stabilize to the optimal. This is because the item replacement and replication only occur during contact events. A smaller contact rate slows down the adjustment of the item distribution, leading to a slower convergence.

6.3 Comparing to Other Designs

BARON has two design components – the valuations in (28) and the negative-positive rule. Now we experiment with other designs for these two components, and examine their performance in comparison to BARON.

An alternative valuation v_i' is

$$v_i'(t) = n^*_{\cdot,i} - n_{\cdot,i}(t), \quad i \in \mathbb{K}, \tag{29}$$

i.e., item i's valuation is defined as the distance of its current supply $n_{\cdot,i}$ to the optimal $n^*_{\cdot,i}$ as given by (15b). Note that in (29), computing the optimal supply $n^*_{\cdot,i}$ requires the knowledge of several system parameters, including the arrival rates $\lambda_{i,f}$, $(i,f) \in \mathbb{C}$, and the contact process parameters μ and β. Obtaining the values of these parameters could be difficult in practice.

Moreover, instead of the *negative-positive rule* (NPR) in BARON, another rule of guiding content exchanges via val-

uations is replacing one item with another as long as the other item has a higher valuation and the item is not already stored. We refer to it as the *lower-higher rule* (LHR).

We examine all four combinations of these design options, where BARON is the combination of NPR with valuations v_i defined in (28). We use the same configuration as Section 6.2.2, and assume initially 16 peers request item 3.

Figures 4(a) and (b) plot the trajectories of peer population under various design combinations. We observe that none of the other combinations performs better than BARON.

In addition, in terms of the comparison of design options for each component, we make the following observations. First, the two valuations perform similarly, and v_i performs better than v_i' under LHR. Second, the system with NPR converges to the optimal faster than LHR. This is interesting because NPR is stricter than LHR, and one would expect that LHR leads to a faster convergence by enabling more frequent replications and replacements. Indeed, from Figures 4(a) and (b), we observe that the peer population stays around its peak ($\approx$35) for a longer time when using NPR. Nonetheless, NPR is able to catch up later. While this demonstrates the efficiency of restricting the replacement to over-replicated item only, the analytical reason beneath is worthwhile to further explore.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we made the first attempt towards a systematical understanding of universal swarms, where peers share content across peer-to-peer swarms. We have rigorously proved that such content exchange across swarms significantly improves stability compared to a single autonomous swarm. We also have proved convergence to a fluid limit for a general class of content exchanges; our theorem thus paves the way for the analysis of more complicated exchange schemes than the one described in the present work.

An important future research direction lies in further investigating the parallels between our work and "missing piece syndrome" in single swarms [6]. In particular, once a "one-club" forms in a swarm, each "one-club" peer has idle bandwidth capacity. It can thus contact uniformly at random peers and seeders at other swarms to obtain C items and place them in its cache, and subsequently continue to sample other swarms to see if it can retrieve and/or offer a missing piece. From this point on, our model applies: peers wish to retrieve one item (their "missing piece"), and leave the system immediately once they retrieve it (corresponding to the "grab-and-go" principle).

This is of course a simplification of the above system, as it ignores the "growing" phase when peers acquire all chunks of a file but the last one, as well as the cache-filling phase. However, in light of the stability properties we observed in this work, understanding if, *e.g.*, the stability region increases through such exchanges, is an interesting open question.

Our analysis leaves several additional open questions, including formally characterizing the stability conditions of BARON, and analytically studying BARON's convergence to optimal stationary points. Our model can also be extended in various ways, including multi-item request, heterogeneous cache sizes and contact rates. The case where arrival rates are not strictly positive, and peers arrive with a partially-filled cache are also worth considering.

Finally, while our model assumes peers are cooperative, it would be interesting to investigate the strategic behavior of peers in universal swarm systems.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions and insights. This work is supported in part by the FP7 EU project "SCAMPI" and the ANR French project "PROSE".

8. REFERENCES

[1] Altman, E., Nain, P., and Bermond, J.-C. Distributed storage management of evolving files in delay tolerant ad hoc networks. In *INFOCOM* (2009), pp. 1431 –1439.

[2] Altman, E., Neglia, G., De Pellegrini, F., and Miorandi, D. Decentralized stochastic control of delay tolerant networks. In *INFOCOM* (2009), pp. 1134 –1142.

[3] Boyd, S., and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

[4] Chaintreau, A., Le Boudec, J.-Y., and Ristanovic, N. The age of gossip: spatial mean field regime. In *SIGMETRICS* (2009), pp. 109–120.

[5] Cohen, E., and Shenker, S. Replication strategies in unstructured peer-to-peer networks. *SIGCOMM Comput. Commun. Rev. 32* (2002), 177–190.

[6] Hajek, B., and Zhu, J. The missing piece syndrome in peer-to-peer communication. *Information Theory Proceedings, 2010 IEEE International Symposium on* (June 2010), 1748 –1752.

[7] Hu, L., Le Boudec, J.-Y., and Vojnoviae, M. Optimal channel choice for collaborative ad-hoc dissemination. In *INFOCOM* (2010).

[8] Ioannidis, S., Chaintreau, A., and Massoulié, L. Optimal and scalable distribution of content updates over a mobile social network. In *INFOCOM* (2009).

[9] Ioannidis, S., Massoulié, L., and Chaintreau, A. Distributed caching over heterogeneous mobile networks. In *SIGMETRICS* (2010), pp. 311–322.

[10] Massoulié, L. Structural properties of proportional fairness: Stability and insensitivity. *The Annals of Applied Probability 17*, 3 (2007), 809–839.

[11] Massoulié, L., and Twigg, A. Rate-optimal schemes for peer-to-peer live streaming. *Perform. Eval. 65* (November 2008), 804–822.

[12] Massoulié, L., and Vojnovic, M. Coupon replication systems. *Networking, IEEE/ACM Transactions on 16*, 3 (June 2008), 603 –616.

[13] Qiu, D., and Srikant, R. Modeling and performance analysis of BitTorrent-like peer-to-peer networks. *SIGCOMM Comput. Commun. Rev. 34* (2004), 367–378.

[14] Reich, J., and Chaintreau, A. The age of impatience: optimal replication schemes for opportunistic networks. In *CoNEXT* (2009), pp. 85–96.

[15] Towsley, D. The internet is flat: a brief history of networking in the next ten years. In *PODC* (2008), pp. 11–12.

[16] Zhou, X., Ioannidis, S., and Massoulié, L. On the stability and optimality of universal swarms. Tech. rep., Technicolor, 2011.