Efficient Content Delivery via Interest Queueing

Tianchong Gao*, Feng Li*
*Indiana University-Purdue University Indianapolis, Indianapolis, IN, U.S.A.
tgao@iupui.edu, fengli@iupui.edu

***Abstract*—Content sharing is an approach to relieve the congestion of cellular networks with alternative communication technologies such as the Wi-Fi and bluetooth. Through a Content Delivery Network (CDN), only a small portion of users need to download the data directly. Other users obtain packets from these users through short-range communications. However, the uncertainty of movement of mobile users challenges the effectiveness of CDNs.**

Unlike previous CDN solutions, in this paper, we present a novel scheme that studies the probabilistic meeting of users. When the accessibility to the cellular network is limited, we apply the queueing theory to guide the downloading or waiting strategies of users. In this system, the users who hold the content become seeds in the CDN and benefit their neighbors. Therefore we also consider the seed growing performance in the strategy. The purpose of our scheme is to let every user efficiently obtain their target content with restricted cellular data. The evaluation results show that our scheme gains significant satisfaction throughput improvements compared to the performance of basic downloading strategies.

***Index terms*—Content delivery network; queueing theory; probabilistic algorithm; satisfaction.**

I. Introduction

Because of the explosion in use of mobile devices, the demand for content sharing is growing rapidly. The wireless data traffic of AT&T has grown 20 000% from 2007 to 2011 [8]. In 2019, about 72% of the all mobile data traffic will be occupied by mobile video data according to the Cisco report [12]. One piece of mobile content may be small, however, the aggregate downloading requests can easily cause network congestion on the server's side [22]. Existing wireless networks face a significant challenge for satisfying user quality of service as well as delivering content efficiently. Furthermore, many mobile users are suffering from the limits of mobile data.

Because mobile users often share the same interests, like the news and popular videos, network providers try to use short-range low-cost communications like bluetooth and Wi-Fi to replace the cellular network as much as possible [10, 11]. Additionally, the fifth generation (5G) cellular network introduces device-centric architecture, millimeter wave short-range communication, and native support for machine-to-machine communication, which opens opportunities [5]. When two users are close to each other, they can perform as content servers and exchange information with each other. Building up a mobile CDN means that popular contents, requested by a majority of users, only needs to be downloaded a few times. Most users then obtain the packets through exchanging, which not only alleviates the cellular network congestion, but also saves money for users.

Previous works demonstrated several design strategies for mobile data offloading and building up mobile Content Delivery Networks (CDNs) [14, 24]. However, the uncertainty of movement of users becomes a significant issue in mobile CDNs. Some researchers used the public transportation system as the server of the CDN [20], but the delay of the shuttles threatened to break the schedule. Especially in the case of users who have relative motion with the bus line and can only build communications in a short time period, communication opportunities are lost. Without considering the irregular moving may cause massive rescheduling works in CDN and damage the system.

Hence, this paper mainly focus on the two problems: *How to model the uncertain movement of users? How to use this model to guide the CDN behavior?* In the proposed scheme, we apply probabilistic theory and queueing theory, respectively, to solve the two problems. We model the meeting time of user as the exponential distribution, we model the content requests as the Poisson distribution, and finally we manage the downloading queue of each user with the help of the queueing theory.

In this paper, the CDN design problem is transformed into an optimization problem, that we want all users to efficiently obtain their target contents when their accessibilities to the cellular network is limited. In order to solve this problem, the proposed scheme needs to smartly choose the right users as seeds to directly download the content. Specifically, our scheme considers the impact of being seeds from both the whole network aspect and the individual aspect, so this scheme has the coordinator calculation phase and the individual calculation phase.

In the coordinator calculation phase, the scheme globally compares the benefit of downloading with waiting. When simulating the benefits, queueing theory is applied because users' interests are randomly generated. Moreover, the expected waiting time for meeting other people is also probabilistic, so it can be easily absorbed in the queueing model. After the simulation, the coordinator announces the number of directly downloading users to make the system achieve highest satisfaction throughput, i.e., most efficient content delivery. In the individual calculation phase, the seeds, who directly download the content, are chosen based on their local effects to the network. In particular, receiving the number of seeds from the coordinator, a user decides whether to be a seed or not based on two factors: fairness and urgency. Fairness links

This is the author's manuscript of the article published in final edited form as:
Gao, T., & Li, F. (2019). Efficient Content Delivery via Interest Queueing. ICC 2019 - 2019 IEEE International Conference on Communications (ICC), 1–6. https://doi.org/10.1109/ICC.2019.8761339

to the length of downloading queue, which is estimated using queueing theory. This scheme aims to reduce the probability that one user holds too many requests. Urgency links to the requests of the user and the user's neighbors for the same piece of content. If a user has many neighbors who urgently demand one piece of content, he or she has high probability to meet other users who request that content in the future. These two factors are combined with the fairness-scaling parameter β to get the optimal satisfaction gain.

The main contributions of our work are as follows: First, we apply the queueing theory to CDNs to model users interests generating, users meeting, and users downloading strategies. Second, we introduce a two-phase algorithm to choose seeds, which aims to achieve the highest satisfaction throughput by balancing fairness and urgency. Third, we experimentally analyze the scheme with two real-world datasets and show the proposed scheme obtaining higher satisfaction than the basic strategy.

II. Related Work

CDNs distribute high-performance service to end-users according to their spatial position [6]. Some end-users, who directly download the data through a cellular network, can behave as the data servers. CDN then has the ability of mobile data offloading, making the trade-off between the low-cost short-range communications and the high-quality but expensive cellular network. It is first proved that Wi-Fi could be used to build the CDN [1]. The feasibility of communication with bluetooth is discussed [10]. The edge caching technique and the new 5G technique contain the possibility of mobile data offloading in device-to-device (D2D) communications [2, 13]. Although some of their model also use the Poisson process to model the download requests, these techniques lack a design to guide the behavior of the end-users from the point of view of the global network. In [9], helper caches, i.e., seeds, are totally randomly chosen. Recently, researchers analyzed the topology of the network and proposed specific seeding algorithms to build CDNs [14]. Our work absorbs the idea of seeding and also introduces queueing theory to simulate the seeding results.

Some current researches are about the caching problems in CDNs. Berger et al. studied the algorithm to choose the hot object to download [4]. Their work is orthogonal to the proposed scheme. When their work is about choosing the right contents to download, our work is about choosing the right users as the servers. Retal et al. designed a platform to provide Content Delivery Network as a Service (CDNaaS), which is another good addition to our work [15].

Some other researchers employed content delivery cloudlets to improve the network performance [18]. However, the users should wear a GPS sensor and the system is assumed to be perfectly aware of the moving path, which is unrealistic in real mobile environments. Wang et al. proposed a probabilistic model about the mobility of users [21]. This model analyzed spatial properties and temporal properties. In [24], the authors employed the probabilistic model to embed the social relationships in CDN design, but their scheme is restricted by the particular social network. Nevertheless, previous studies give us insights to design caching schemes based on the probabilistic mobility model.

Meaning	Symbol
user	u, V-Z in the graph
interest arriving rate	λ
content	c
expected meeting or downloading rate	μ
probability	P
interest amount	I
satisfaction	Sat
queue length	L
content size	S
fairness scaling parameter	β

TABLE I: Symbol table

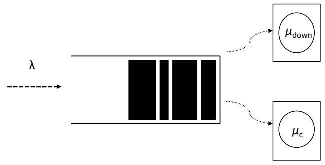


Fig. 1: Serving queue

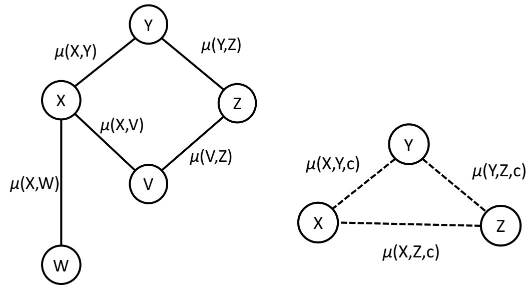
III. System Model

For reference, some important symbols used in this paper are given in Table I.

In this paper, each mobile user acquires a piece of content through the cellular network or the CDN. The cellular network is persistent, but the cost is higher, while the CDN connection is probabilistic but nearly free. Our general purpose is to design a scheduling scheme to maximize the satisfaction of users when their downloading rate through the cellular network is limited. Specifically, for each user u in the user set U, we use the serving queue shown in Figure 1 to model the behavior of u. In the queue, the interest arriving process is an aggregated Poisson process with rate λ. The interest is satisfied either through the downloading service with rate μ_{down} or through the CDN service with rate μ_c.

On the CDN side, the content requests are served if and only if u meets other users holding that content. Fig. 2(a) gives an example of the meeting network. Researches studying random movement suggest that the meeting time is exponentially distributed under the random waypoint mobility model and the Brownian motion model [7, 17]. When talking about the specified content c, users not interested in this content are removed in Fig. 2(b).

In this example, only three users have interest in that content. User V and W are omitted. The interest graph is based on the original meeting graph. But the weight μ is different. Suppose Z wants to get c from X: although the two users are not linked in the meeting network, Z still has the probability to receive the content with the help of Y or V. And the expected rate $\mu(X, Z, c)$ is given by $\frac{1}{\mu^{-1}(X,Y)+\mu^{-1}(Y,Z)} + \frac{1}{\mu^{-1}(X,U)+\mu^{-1}(U,Y)}$. In conclusion, the



(a) Meeting network (b) Interest network

Fig. 2: Network example. A solid edge means two users have opportunities to meet. $\mu(X, Y)$ shows the expected rate of their meeting. A dotted edge means two users can exchange the content. $\mu(X, Y, c)$ shows the expected rate that X and Y exchange the content c.

expected rate of interest exchanging for two arbitrary users u_1 and u_m is given by

$$\mu(u_1, u_m, c) = \sum_{all\ paths} \frac{1}{\sum_{1 \leqslant i < j \leqslant m} \mu^{-1}(u_i, u_j)} \tag{1}$$

where u_i and u_j are intermediate nodes in the path between u_1 and u_m.

Considering all the nodes in the CDN who can help u to get the content c, the expected rate u to get that content is given by μ_c.

$$\mu_c = \sum_{u_m} \mu(u, u_m), \quad u_m \in U_c \tag{2}$$

where U_c means the set of users holding the content c.

On the cellular network side, users' ability to access the Internet is often limited. In other words, if a user have unlimited access to the Internet, he/she will directly download the content instead of using CDN for efficiency consideration. Although there are several aspects to limit users' accessibility, e.g., cellular speed and cellular server congestion, today users are often limited by their cellular data plans other than those network limitation. In our network model, we assume the user u utilizes a cellular data plan of size DP_u in the system each month. Suppose $D_u(t)$ is the total size of content downloading from the cellular network until time t. In order to strictly guarantee the downloading data does not exceed the data limit, u should make

$$D_u(t) \leqslant DP_u \cdot \frac{t}{t_f} \qquad \text{for all t} \tag{3}$$

when t_f shows the total time, e.g., a month.

At first glance, Equation 3 seems to be a soft limit: it is possible that users download more data than the limit in some time periods but still get the total downloading size below the quota at the end of the month. However, the possibility is based on the expectation of future interests and sometimes dangerous. In this paper, we use Equation 3 as a strict limit to guide the downloading behavior. If the used data reaches the limit when $D_u(t) = \frac{DP_u}{t_f} \cdot t$, the downloading queue is saturated, and u will download data at the rate of $\mu_{down} = \frac{DP_u}{t_f}$. When $D_u(t) < \frac{DP_u}{t_f} \cdot t$, the downloading queue is unsaturated, and u will download data at the speed of the cellular network, which is significantly larger than μ_{down}.

With the network model, we need to define our optimization goal of this efficient content delivery problem. In this paper, we assume a user may randomly has interest to a content at any time after this content is published. The interest of the user u to content c is defined as $I_{\{u,c\}}$. After final obtaining that content, the user has a satisfaction gain $Sat(u, c)$, and this gain is a monotonous decreasing function of t, where t is the time between raising the interest and obtaining the content. In particular, the satisfaction gain is modeled as

$$Sat(u, c) = I_{\{u,c\}} \cdot e^{-\alpha \cdot t} \tag{4}$$

where α is a scaling parameter to show the decreasing speed of satisfaction.

Thus our goal is to design the downloading queue to maximize the total satisfaction of all users to the system.

$$\begin{gathered} \max \sum_{u \in U} \sum_{c} Sat(u, c) \\ \text{subject to } D_u(t) \leqslant DP_u \cdot \frac{t}{t_f} \end{gathered} \tag{5}$$

IV. Scheme

Once a piece of content is released, our scheme has the following steps: First, there is a coordinator, i.e., a centralized server, that knows the content before most of the users generating interests. The coordinator estimates the number of users the content will attract, and publishes a suggestion about the number of users to download the content directly. Second, receiving the information and suggestion, each individual needs to make the decision whether to add the content to the downloading queue directly or to wait the CDN. The decision considers both fairness and urgency.

A. Coordinator Calculation

Given a new piece of content c with size S_c, the coordinator needs to estimate the percentage of users interested in this content and the percentage of users to download the content through the cellular network. While the first estimation P_{inte} is based on the content itself, the second estimation P_{seed} is related to the satisfaction benefits of downloading.

In order to get the best P_{seed}, we focus on two groups of people. The first group is the users who have no interest in the content, but download it. Because the coordinator calculation happens as soon as the content is released, many devices make the downloading decision before they know the interest of users. Then, for users in this group, their downloading service is occupied and their satisfaction is decreased. If the total number of users having the corresponding interest in that content is $|U_c|$, the size of the first group will be $|U_c| \cdot (1 - P_{inte}) \cdot P_{seed}$. We use $\bar{S}$ to denote the mean size of the content, $\bar{I}$ to denote the mean interest. Then the loss of satisfaction from downloading content the user is not interested in is like the satisfaction gain of downloading a piece of arbitrary content with size S_c.

$$Sat_{loss} = |U_c| \cdot (1 - P_{inte}) \cdot P_{seed} \cdot \bar{I} \frac{\bar{S}}{S_c} \cdot e^{-\alpha \cdot t_m} \tag{6}$$

t_m shows the expected time of waiting on the downloading queue, which will be given in the following section.

In contrast, the second group of users get benefits from the CDN. They are interested in the content but they do not use the downloading service. The size of this group is $|U_c| \cdot P_{inte} \cdot (1 - P_{seed})$. The gain in satisfaction is due to the CDN and independent from other pieces of content on the downloading queue. If we use $\bar{\mu}$ to show the average expected rate of meeting a person in the meeting network, the person who has the desired content is given by probability P_{seed}. Hence, the expected rate of meeting the right person is $\bar{\mu}P_{seed}$ and the satisfaction gain is given as follows:

$$\begin{aligned} Sat_{gain} &= |U_c| \cdot P_{inte} \cdot (1 - P_{seed}) \cdot I_c \cdot e^{-\alpha/(\bar{\mu}P_{seed})} \\ \bar{\mu} &= \frac{\sum_{u_1}\sum_{u_2}\mu(u_1,u_2)}{|U_c|}, \quad u_1 \in U_c,\ u_2 \in U \end{aligned} \tag{7}$$

I_c is an estimated average of interest given by the coordinator.

Choosing the ratio of people to download the content and become the seeds affects the strategy of each individuals and eventually results in the change of satisfaction. Hence, in order to get the maximum satisfaction, the coordinator can choose P_{seed} of people to be the seeds.

$$P_{seed} = \arg\max(Sat_{gain} - Sat_{loss}) \tag{8}$$

B. Individual Calculation

Taking P_{seed} as the suggestion, each user should have his/her own strategy to download or to wait. Analyzing the downloading queue of each individual can give some useful properties of downloading.

According to the assumption in our model, the user u generates interest in a Poisson process with rate λ, and downloads an interest with probability P_{down}, which is independent of the generating process. Hence, the downloading queue of an arbitrary user u is an M/D/1 queue model. According to the coordinator calculation, total number of people downloading the content equals the number of seeds.

$$\sum_{U} P_{down} = |U_c| \cdot P_{seed} \tag{9}$$

When an arriving process is in the unit of interests, the downloading service takes S/μ_{down} time to serve that interest with size S. Hence, the service process has the rate μ_{down}/S. With the help of the Poisson Arrivals See Time Averages (PASTA) property [23] and residual time analysis [19], we can get the utilization ρ and the expected length of downloading queue L.

$$\begin{aligned} \rho &= \frac{P_{down} \cdot \lambda \cdot S}{\mu_{down}} \\ L &= \rho + \frac{1}{2}\left(\frac{\rho^2}{1-\rho}\right) \end{aligned} \tag{10}$$

Here L is only a rough estimation. Besides downloading the content, the downloading queue can also be served with the CDN, and that serving process does not follow the property of queueing, which requires first come first serve and serving one by one. Hence, the overall exchanging service could not be modeled by queues. However, L can still help the heuristic algorithms to choose seeds.

To maximize the satisfaction, the proposed scheme considers two aspects: fairness and urgency. At one extreme, each downloading queue keeps the same length L. It prevents one seed holds an extremely long downloading queue and other users are all waiting this seed. However, the overall system is not an ideal queue. Some nodes may have low downloading rate, low data plan size, or low probability to meet other users, then they cannot take responsibilities to behave like a seed. At the other extreme, the seeds are chosen according to the requests of the user and his/her neighbors' interests. When that area of users are urgent to get the content, there is a high probability that the central user will meet other users who urgent to get the content as well. If a user meet with enough people urgent to that content, we will let it directly download the content. However, some nodes could gain importance across different interest graphs because of the network structure, and then these seeds get congested.

The scheme begins with keeping the maximum fairness. If we choose $P_{down} \propto \frac{\mu_{down}}{\lambda}$ for an arbitrary interest, each user holds the same length of downloading queue. According to Equation 9, we have

$$P_{down} = \frac{\mu_{down}}{\lambda} \cdot \frac{P_{seed}}{\sum_U \frac{\mu_{down}}{\lambda}} \tag{11}$$

Then we introduce a fairness scaling parameter β, $\beta \in [0, 1]$, to balance fairness and urgency. Algorithm 1 is a real-time adaptive algorithm and it chooses two groups of seeds. β of users become seeds to contribute to the CDN as soon as the content is published, considering fairness. The rest of the users become seeds after getting enough requests from their neighbors in meeting network and themselves, considering urgency.

In Algorithm 1, the Sat_c calculated in Lines 11 and 12 is the same as the satisfaction that the user u and u's neighbors could get when u immediately downloads the content. Because the calculation does not consider the downloading time, Sat_c is ofter greater than the real value, and eventually many users may have Sat_c over the threshold I_c. However, the total seeds number is controlled by the coordinator with a counter B. The expected number of seeds is still $P_{seed} \cdot |U_c|$. Hence, the user who gains more requests from his/her neighbors at the beginning has more possibility of becoming a seed in the second group. These urgent seeds can make contribution to the CDN efficiently.

C. Overhead

Algorithm 1 shows that the total communication overhead in the cellular network is relatively small. After the content first becomes available, the coordinator broadcasts some related information like P_{seed}, I_c. Some network data, like μ_c, $\sum_U \frac{\mu_{down}}{\lambda}$, can be stored in each device. The first group of seeds report their decisions together and the coordinator only needs one broadcast. In contrast, the second group of seeds

Algorithm 1: Seeding strategy for u with content c

```
Input: Seed probability $P_{seed}$, estimated interest $I_c$ published by the coordinator
Output: Becomes the seed (download) or not
$u$ becomes the seed with probability $P'_{down} = \beta \cdot P_{down}$
Seeds tell the coordinator, and the coordinator broadcasts which members are to be seeds
Initialize $B = (1-\beta) \cdot P_{seed} \cdot |U_c|$
If $u$ is not a seed, then
while true do
    Meet neighbor $n$ at time $t_n$.
    If $n$ has interests and not a seed, get the interest $I_n$, time $t_n$
    Calculate the future satisfaction $Sat_n = I_n \cdot e^{-\alpha*\mu_c(u,n)}$
    Meet neighbor $n'$, who has been met before
    If $n'$ already has the content, or becomes a seed, delete $Sat_{n'}$.
    At time $t$, calculate the total satisfaction $Sat_c = e^{-\alpha*S/\mu_{down}} \cdot \sum_{\{n\}} Sat_n \cdot e^{-\alpha*(t-t_n)}$
    If $u$ has interest $I_u$ at time $t_u$, $Sat_c = Sat_c + I_u \cdot e^{-\alpha*(t-t_n)}$
    If $Sat_c > I_c$, $u$ becomes the seed, $B = B - 1$.
    If $B \leqslant 0$, then Break
end while
```

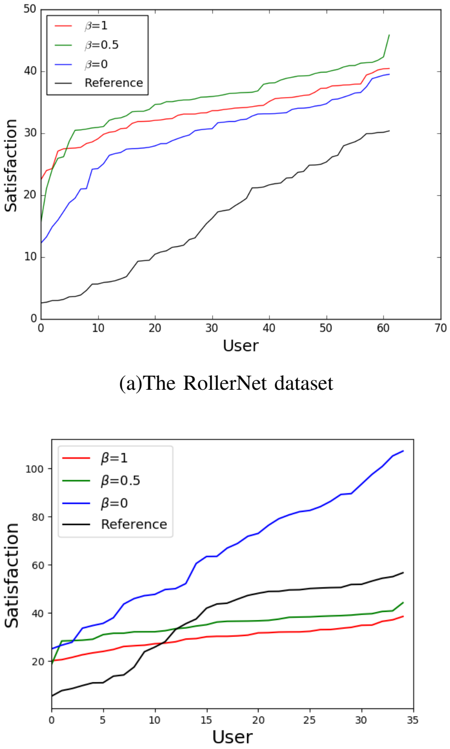
need to upload their decisions to the coordinator and then the coordinator broadcasts these seeds separately. The total times of cellular transmission is less than $(1-\beta) \cdot P_{seed} \cdot |U_c| + 3$ times of broadcast, and each broadcast has several bits of data. Because the overall seeding strategy avoids broadcasting interests and calculation is split among the coordinator and each user, our scheme of building the CDN does not occupy too much cellular data.

V. Evaluation

The evaluations are driven by two publicly available real-world contact traces, the RollerNet dataset [3], and the Haggle dataset [16]. The RollerNet trace took records of communications between rollerbladers. There were 15 000 people in the three-hour roller tour. The Haggle trace was conducted for four days during the INFOCOM meeting. In the experiment, attendees carried 78 mobile devices with bluetooth for three days. The experiment simulates a closed-world environment, in which users have repetitive communications in a short time.

For simplicity, we use the records of communications between the most frequent users to build the CDN. Sixty-two users from the RollerNet dataset have 11.86 contacts per second on average. However, 35 users from the Haggle dataset only have 0.02 contacts per second on average. Hence, these evaluations test the performance of both heavy and weak communications. Having the communication records within a time period, the proposed algorithm calculates the expected CDN serving rate μ through meeting frequency. In the following experiments, each user only knows the average waiting time, but not the precise meeting time.

In order to build the downloading scenario, the average interests arriving interval is assumed to be 2 min; then each user independently decides whether he/she cares about the content or not. After that, the algorithms with various fairness concerns (β) are applied. For comparison purpose, the performance of the basic strategy, which directly downloads the content from the Internet, is also tested and marked as ‘Reference’. If a user decides to download the content directly, the downloading rate μ_{down} is 1.8Mbyte/min to 18Mbyte/min, which simulates the limit of the data plan. Each piece of content has a uniform distributed size from 1Mbyte to 20Mbyte.



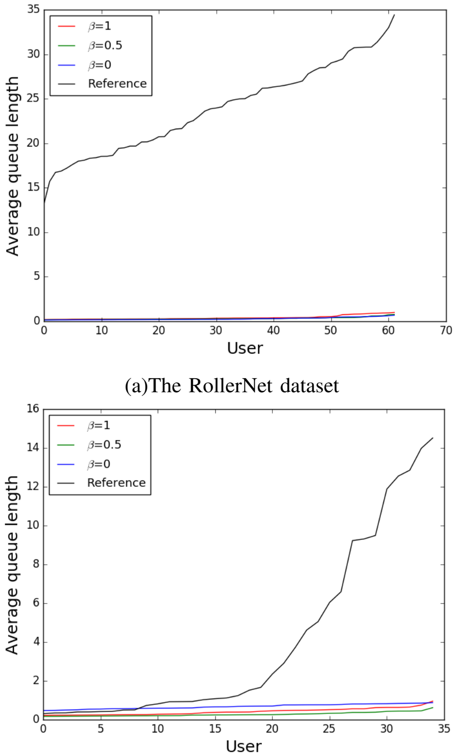
(a)The RollerNet dataset

(b)The Haggle dataset

Fig. 3: Total satisfaction gain

Fig. 3(a) shows the satisfaction gain of each user in the whole evaluation period in the RollerNet dataset. The average satisfaction gain is 33.33, 35.42, 29.73 and 16.09 corresponding to the case of β=1, β=0.5, β=0 and the reference. In this experiment, all the interest queueing methods outperform the reference method, especially the one combines both fairness and urgency. That method reduces some downloading time of the busy users and asks them to wait on the CDN. Because the seeding algorithm can avoid the congestion in downloading and make full use of the CDN, it is more efficient than direct downloading.

Fig. 3(b) shows the satisfaction gain in the Haggle dataset. The average satisfaction gain is 34.99, 29.62, 65.16, and 36.43, respectively. In this experiment, the method only considering urgency outperforms the reference method, while other methods considering fairness have poor performance. The Haggle dataset is a weak communication scenario that the average contacts per second is only 0.02. It makes the seed users have low probability to meet other users. While the urgency method focuses on existing requests, the fairness method is an estimation of future requests before the user and



(a)The RollerNet dataset

(b)The Haggle dataset

Fig. 4: Average queue length over time

his/her neighbors raise interests. In the weak communication case, if the node himself/herself does not raise interest, the downloaded content has low probability to benefit other users. Finally, it is a waste to the downloading resource, and the fairness methods have bad performance.

Fig. 4(a) shows the average queue length over the whole evaluation period in the RollerNet dataset. The average queue length over all users is 0.41, 0.31, 0.29, and 23.90. In Fig. 4(b) of the Haggle dataset, it is 0.42, 0.28, 0.68, and 4.02. This result shows that in both environments, the proposed algorithm can help to significantly reduce the queue length. The reference algorithm needs to add the request on the downloading queue as soon as the user needs it. However, the seeding algorithm enforces that only the seeds can add the request, which avoids the congestion on the downloading queue and then improves the satisfaction.

VI. Conclusion

This paper presents a downloading algorithm by applying interest-based queueing. The proposed algorithm uses seeds to reduce the waste of direct downloading and it has three phases. The coordinator calculation phase studies the probability of communication and uses this information to determine the number of seeds. The distribution phase builds the downloading queue for analyzing the influence of seeds and then makes decision on seeds. The evaluation results show that by reducing the average queue length, the proposed algorithm can have more satisfaction than the basic strategy.

References

[1] Aruna Balasubramanian, Ratul Mahajan, and Arun Venkataramani. Augmenting mobile 3g using wifi. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 209–222. ACM, 2010.

[2] Ejder Bastug, Mehdi Bennis, and Mérouane Debbah. Living on the edge: The role of proactive caching in 5g wireless networks. *IEEE Communications Magazine*, 52(8):82–89, 2014.

[3] Farid Benbadis and Jeremie Leguay. CRAWDAD dataset upmc/rollernet (v. 2009-02-02), February 2009.

[4] Daniel S Berger, Ramesh K Sitaraman, and Mor Harchol-Balter. Adaptsize: Orchestrating the hot object memory cache in a content delivery network. In *NSDI*, pages 483–498, 2017.

[5] Federico Boccardi, Robert W Heath, Angel Lozano, Thomas L Marzetta, and Petar Popovski. Five disruptive technology directions for 5g. *IEEE Communications Magazine*, 52(2):74–80, 2014.

[6] Rajkumar Buyya, Mukaddim Pathan, and Athena Vakali. *Content delivery networks*, volume 9. Springer Science & Business Media, 2008.

[7] Tracy Camp, Jeff Boleng, and Vanessa Davies. A survey of mobility models for ad hoc network research. *Wireless communications and mobile computing*, 2(5):483–502, 2002.

[8] J. Donovan. Wireless data volume on at&t's network continues to double annually. http://goo.gl/JtNKT, 2012.

[9] Salah-Eddine Elayoubi, Antonia Maria Masucci, J Roberts, and Berna Sayrac. Optimal d2d content delivery for cellular network offloading. *Mobile Networks and Applications*, 22(6):1033–1044, 2017.

[10] Bo Han, Pan Hui, and Aravind Srinivasan. Mobile data offloading in metropolitan area networks. *ACM SIGMOBILE Mobile Computing and Communications Review*, 14(4):28–30, 2011.

[11] Bo Han, Pan Hui, VS Anil Kumar, Madhav V Marathe, Jianhua Shao, and Aravind Srinivasan. Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*, 11(5):821–834, 2012.

[12] Cisco Visual Networking Index. Global mobile data traffic forecast update 2014–2019. white paper c11-520862.

[13] Shahid Mumtaz and Jonathan Rodriguez. *Smart device to smart device communication*. Springer.

[14] Wei Peng, Feng Li, Xukai Zou, and Jie Wu. The virtue of patience: Offloading topical cellular content through opportunistic links. In *2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems*, pages 402–410. IEEE, 2013.

[15] Sara Retal, Miloud Bagaa, Tarik Taleb, and Hannu Flinck. Content delivery network slicing: Qoe and cost awareness. In *Communications (ICC), 2017 IEEE International Conference on*, pages 1–6. IEEE, 2017.

[16] James Scott, Richard Gass, Jon Crowcroft, Pan Hui, Christophe Diot, and Augustin Chaintreau. CRAWDAD dataset cambridge/haggle (v. 2009-05-29), May 2009.

[17] Gaurav Sharma, Ravi Mazumdar, and Ness B Shroff. Delay and capacity trade-offs in mobile ad hoc networks: A global perspective. *IEEE/ACM Transactions on Networking (ToN)*, 15(5):981–992, 2007.

[18] Hassan Sinky and Bechir Hamdaoui. Cloudlet-aware mobile content delivery in wireless urban communication networks. In *Global Communications Conference (GLOBECOM)*, pages 1–7. IEEE, 2016.

[19] William J Stewart. *Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling*. Princeton University Press, 2009.

[20] Qiankun Su, Katia Jaffres-Runser, Gentian Jakllari, and Charly Poulliat. An efficient content delivery infrastructure leveraging the public transportation network. In *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 338–347. ACM, 2016.

[21] Rui Wang, Xi Peng, Jun Zhang, and Khaled B Letaief. Mobility-aware caching for content-centric wireless networks: modeling and methodology. *IEEE Communications Magazine*, 54(8):77–83, 2016.

[22] Mike P Wittie, Veljko Pejovic, Lara Deek, Kevin C Almeroth, and Ben Y Zhao. Exploiting locality of interest in online social networks. In *Proceedings of the 6th International COnference*, page 25. ACM, 2010.

[23] Ronald W Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[24] Chen Xu, Caixia Gao, Zhenyu Zhou, Zheng Chang, and Yunjian Jia. Social network-based content delivery in device-to-device underlay cellular networks using matching theory. *IEEE Access*, 5:924–937, 2017.