A Locality-based Mobile Caching Policy for D2D-based Content Sharing Network

Yali Wang†, Yujin Li†, Wenye Wang† and Mei Song*
†Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606
*School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China 100876
Email: {ywang116, yli27, wwang}@ncsu.edu, songm@bupt.edu.cn

***Abstract*—As the explosion of Internet traffic is quickly leading to overloaded cellular network, device-to-device (D2D)-based content sharing is proposed as a method to offload mobile data traffic. The performance of D2D-based content sharing is dramatically affected by the success rate of content fetching from nearby devices and quality of content transmission, which is determined by the geographic distribution of mobile devices, the number of devices having contents in their caches, and the condition of D2D links. Hence, a key problem is how to cache various contents in the limited storage of mobile devices for improving the success rate of content fetching. In this paper, we aim to design a caching policy by considering the joint impact of locality of real-world mobile data traffic and device contact pattern to improve the success rate of content fetching. To do this, we first study the characteristics of network traffic and device contact pattern by analyzing traces from realistic networks. Then, we design a locality-based caching policy and derive the content caching probability and hit ratio through mathematical analysis. Through numerical evaluation and trace-driven simulations, we not only quantify how content popularity, content active lifetime, content size, content bit rate, device storage, transmission rate, and closeness centrality affect the content hit ratio, but also provide comparison on hit ratio and storage cost in different caching policy, which is a strong evidence that the joint impacts from characteristics of content and device are the necessary consideration when to design a caching policy.**

I. Introduction

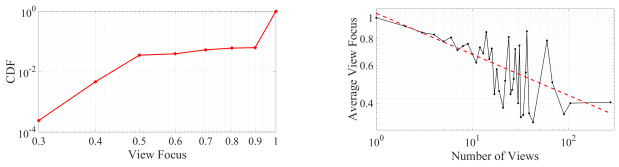
The rapidly increasing number of personal mobile devices, e.g., smartphones and tablets, and various mobile applications over the last few years have resulted in an exponential growth of data traffic in cellular networks. Cisco reported that cellular networks may need to support a 1000 fold increase in capacity by 2020 [1]. In particular, video stream continues to be the major application generator for mobile data traffic growth and will account for 75% of global mobile data traffic by 2021 [2]. Particularly, online video-sharing services (such as YouTube, Yahoo! Video), which have gained an audience of billions of users including educators and scholars. Such tremendous increase in mobile data traffic is predicted to overload cellular and WiFi networks.

In order to mitigate the load on cellular networks, D2D-based content sharing has been proposed as a method to offload traffic. The performance of D2D-based content sharing is dramatically affected by the success rate of content fetching from nearby clients and reliably QoE-oriented content transmission, which is determined by geographic distribution of mobile devices, the number of devices having contents in storage, the category of content cached in devices, and devices sharing index. In practice, the advantage of D2D-based content sharing may be limited when a significant number of mobile devices in a large network area must be served with a finite caching storage on the move. Hence, design a efficient caching policy is important for content fetching. Generally, on one hand, the geographical distribution of devices, the number and the category of cached content determine that *what kind of content can be fetched*. On the other hand, varying device contact pattern and social properties determine that *which device can successfully share content to its close-by devices*. Therefore, the realistic network traffic, users requests pattern, and device contact pattern jointly determine the implement and performance of D2D-based content sharing, which motivates us to think a key problem in D2D-based content sharing that is *how to make decision on caching various kinds of contents in different mobile devices for achieving optimal success rate of contents fetching in D2D-based content sharing network*.

Existing works on D2D communication-based content sharing are mainly concerned with interference avoidance and energy efficiency [3] and sharing strategies [4]. Only a minority of the existing studies investigate and design caching policy in D2D communication-enabled mobile devices for improving content sharing performance. For example, Jingjie et al. [5] studied the problem of maximizing cellular traffic offloading via D2D communication, by selectively caching popular content locally, and by exploring maximal matching for sender-receiver pairs. Bastug et al. [6] and Kang et al. [7] studied on caching scenarios that exploiting the file popularity, correlations among users-files patterns, and social structure of the network, to minimize the average caching failure rate. These works, however, do not consider the joint impacts of properties of realistic network traffic, user requests pattern, device contact pattern, and wireless transmission on making caching decision in different mobile device and QoS of D2D communication. There still lacks a comprehensive understanding of content caching policy in D2D-based content sharing network.

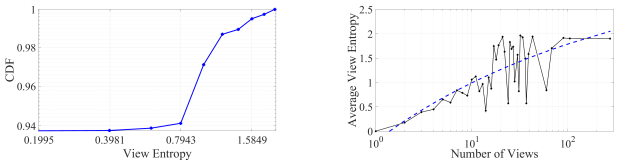
In this paper, we specifically focus on the D2D-based video-sharing scenario. In this scenario, we focus on designing a

This work is supported by NSF CNS-1423151. Yali Wang is a visiting student from Beijing University of Posts and Telecommunications.



(a) CDF of view focus (b) Average view focus

Fig. 1. View focus as a function of the video views.



(a) CDF of view entropy (b) Average view entropy

Fig. 2. View entropy as a function of the video views.

locality-based caching decision policy by considering the joint impact of spatial and temporal locality of content popularity, content active lifetime, content size, content bit rate, device transmission rate and sharing index on the content hit ratio. We firstly characterize network traffic and mobile devices opportunistic contact pattern through analyzing realistic traces of YouTube Video [8, 9], and Bluetooth encounters [10]. Then we define network model, traffic pattern and design the locality-based caching policy. Finally, we derive content caching probability and hit ratio through mathematical analysis. Numerical results not only quantify the performance of content fetching, but also provide critical points to be used to making optimal decisions on contents caching.

The rest of this paper is organized as follows. In Section II and III, we examine the characteristics of network traffic and devices encounter, respectively. Section IV defines system model and caching policy. In Section V, mathematical analysis about content caching probability and hit ratio are performed. Finally, section VI concludes this paper.

II. Network Traffic Trace Analysis

In this section, we analyze the spatial and temporal locality of network traffic and characteristics of user request patterns via two groups of YouTube traffic traces [8, 9] to achieve a comprehensive study.

A. Spatial and Temporal Locality of Content Popularity

One of the design criteria for content caching policy is the content popularity, which is represented by the number of client views. Obtaining and analyzing the popularity of a video enables caching policy designers to decide which contents to cache, which affects the number of copies of a content in a network, thus influencing the hit ratio. Motivated by this, we analyze spatial locality of content popularity by employing one trace from Zink's dataset: youtube.parsed.dat [8].

Let $1 \leq r \leq R$ be a region in the network. Denote $v_{o_r}^t$ as the number of content o's view in region r at time t, and $V_o^t = \sum_{i=1}^{R} v_{o_i}^t$ as the total number of content o's view at time t. To examine the spatial locality of content popularity, we define the view focus as $F_o^t := \max_r(\frac{v_{o_r}^t}{V_o^t})$, and the view entropy as $H_o^t := -\sum_{i=1}^{R} F_o^t \log_2 F_o^t$, where the sum function is running only over regions for which $v_{o_r}^t \neq 0$.

Fig. 1 (a) shows the cumulative distribution function (CDF) of the view focus of videos in our dataset. We observe that about 90% of YouTube videos that enjoy at least 80% of their views in a single region. This is a strong evidence that videos tend to be popular in a locally confined area, rather than in a globally wide region. Furthermore, as videos get more views, they tend to be watched in a more disparate set of regions, as shown in Fig. 1 (b), where the average view focus decreases as the number of views grows. Moreover, Fig. 2 (a) shows the CDF of the view entropy of videos in our dataset. As shown in the distribution, there are about 94% of YouTube videos with view entropy lower than 1 bit and about 2% of videos with view entropy larger than 1.4 bits. On the contrary, view entropy grows as the number of views grows, as shown in Fig. 2 (b). We observe that a large percents of videos' views are geographically concentrated with high view focus value. Since the view entropy is high, videos with larger number of views need to be accessible from all over the network.

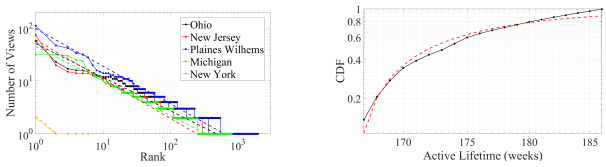
In addition, we extract the rank of YouTube videos according to the number of views in different regions, as shown in Fig. 3(a). Surprisingly, we find that although a large fraction of videos have high view focus and low view entropy, the distribution of popularity in a certain region (e.g, Ohio or Michigan) is a reasonably good fit for a Zipf distribution, as observed in other studies [11]. Thus it is reasonable to assume Zipf content popularity in different region r with various Zipf's law exponents γ_r in this paper.

After analyzing the spatial locality of video popularity, we begin to study the temporal evolution of video views. In this part, we are interested in studying whether videos exhibit a steady and uniform level of popularity across their lifetime or, instead, their popularity changes over time by analyzing one trace from Cheng's dataset [9]. We define that if the ratio of the increasing number of views directed to a video o at time t in region r is less than ϵ from the previous time $(t-1)$, the video's active lifetime is given as $l_o := \inf\{t \in \mathbb{R} : \frac{v_{or}^t}{v_{or}^{t-1}} - 1 \leq \epsilon\}$. Let $\epsilon = 1\%$. The active lifetime gives us a way to estimate the temporal locality.

Fig. 3(b) shows the CDF of active lifetime for approximately 161 thousand videos in log-log scale, which has a reasonably good fit for exponential distribution with parameters $\lambda_L = 0.1135$. We observe that there are approximately 80% contents which has active lifetime equal to or less than 180 weeks after they are uploaded, which implies that most videos are requested and viewed frequently during their early period, and then fewer and fewer clients will request them after the video's active lifetime. This characteristic can be applied in caching design, which can help mobile device to make decisions on caching.

B. User Request Pattern

As caching storage of mobile devices is limited and each mobile device can only cache contents for a limited period of time, content sharing through D2D communications become possible only if nearby devices have these contents which are



(a) YouTube videos are ranked according to the number of views in different regions.

(b) CDF of content active lifetime.

Fig. 3. Spatial and temporal locality of content popularity.

requested by users around the same time. Hence, contents caching duration and user request pattern significantly affect the hit ratio, which can be applied in caching design to achieve optimal hit ratio. By employing one trace from Zink's dataset, we extract the inter-arrival time of users' consecutive requests and the time intervals between two consecutive requests directed to the same content. From Fig. 4, we observe that there are approximately 90% consecutive requests with inter-arrival time smaller than 700 seconds (approximately 12 minutes) and approximately 90% consecutive requests directed to the same content with time intervals smaller than 3 hours. It indicates that users tend to request videos multiple time within a short period, which shows a great motivation of designing a efficient caching policy for content sharing.

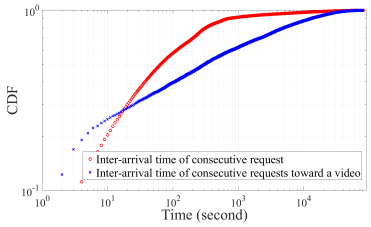


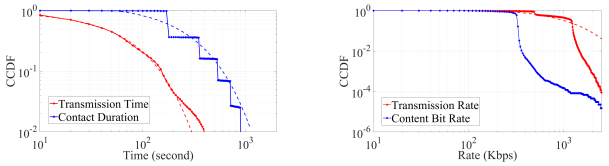
Fig. 4. The time intervals between a client's two consecutive request and the time intervals of two consecutive requests directed to the same content

III. Devices Contact Pattern Analysis

In this section, we analyze the characteristics of the devices contact pattern in terms of Bluetooth encounters between two devices by using one trace from A. Caputo's dataset [10]. In the following, we study the contact duration, transmission time and rate, and network centrality.

A. Contact Duration Vs. Transmission Time

We define that a user u can successfully transmit the content o to its neighbour v only if their opportunistic communication is established and their contact duration is long enough for the video transmission. Fig. 5 (a) shows the complementary CDF of contact duration t_C comparing with the transmission time. From Fig. 5 (a), we observe that a large percentage of video's transmissions take a short time. There are approximately 50% of videos taking less than 40 seconds and 90% of transmissions taking less than 100 seconds. The complementary CDF of contact duration has a heavier tail than that of video transmission time. Specifically, there are 99.6% of contacts between mobile devices lasting longer than 40 seconds and approximately 50% of contacts lasting longer than 1000 seconds. Hence, we find that there is a relatively high probability that a large parts of



(a) Contact duration Vs. Transmission time

(b) Transmission rate Vs. Bit rate

Fig. 5. Complementary CDF of contact duration and transmission rate.

videos can be successfully transmited through D2D communications. In this paper, we assume that the distribution of contact time t_C and transmission time τ both fit well for exponential distribution with parameters λ_C and λ_T respectively, which has been used by other existing studies, such as [12] and shown to be a good approximation.

B. Transmission Rate Vs. Content Bit Rate

In addition, we define that QoE-satisfied content transmission can be achieved when the transmission rate is large enough for the content bit rate. Fig. 5 (b) shows approximately 90% of videos' bit rate are smaller than 200 $Kbps$ and 10% of them are larger than 320 $Kbps$, which indicates that a large percentage of videos' bit rate are within range $[200\ Kbps, 300\ Kbps]$. Moreover, approximately 90% of the transmission rate between two devices r_T are within range $[500\ Kbps, 1000\ Kbps]$. We can find that there is relatively high probability that content delivery between two D2D communication-enabled devices can satisfy the QoE.

C. Closeness Centrality

We apply centrality, which is a social network analysis metrics to measure the social connectivity of devices in the network. That has potential influence in the probability of content sharing between devices. In graph theory and network analysis, centrality is a quantification of the relative importance of a vertex within a graph. For node u, we define $C_u = \frac{1}{F}\sum_{v=1}^{F} D(u,v)$ as the closeness centrality, which measures the average contact duration between node u and its Facebook friend v, which means the node has larger probability of opportunistically contact with other devices and the node has more higher capability for content sharing. F is the total number of contacts with its total friends and $D(u,v)$ is the contact duration between node u and its friend v.

Fig. 6 shows the mobile devices ranked according to the closeness centrality. From this figure, we can see that there is a reasonably good fit for a Zipf distribution with parameter σ, then we define $p_u = \frac{u^{-\sigma}}{\sum_{j=1}^{n} j^{-\sigma}} (1 \le u \le n)$ as *sharing index* for device u. n is the total number of mobile devices in the network.

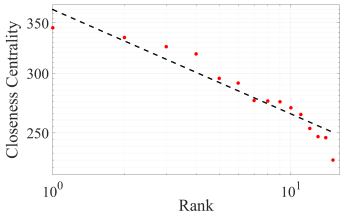


Fig. 6. Mobile devices are ranked according to decrease order of closeness centrality.

IV. Caching Policy in D2D-based content sharing Network

In this section, motivated by the trace analysis in previous sections, we present the D2D-based content sharing network as well as the traffic and content transmission model. Then, we define a locality-based caching policy an derive the hit ratio through mathematical analysis. Finally, numerical simulation is performed to quantify the performance of content fetching.

A. Network Model

We suppose numbers of n mobile user equipments (UEs) are distributed in the network modeled as independent homogeneous poisson point process Φ_U with density λ_U [13]. Each mobile UE can communicate with its neighbouring UE through Bluetooth directly. In this network, a mobile UE can either download content from network servers or fetch content from its neighbouring UEs through D2D communications. The probability that number of m UEs exist within a transmission range d from a reference device is given as $f_n(m,d) = \frac{(\pi d^2 \lambda_U)^m}{m!} e^{-\pi d^2 \lambda_U}$.

B. Traffic and Content Transmission Model

We denote the set of contents by $\mathcal{O}$, where $O := |\mathcal{O}|$ is the total number of contents. Let $\mathcal{R}$ be the set of regions in the network, where $R := |\mathcal{R}|$ is the number of regions in the network. Motivated by the results in Fig. 3 (a), we assume that the content popularity in a certain region $r \in \mathcal{R}$ to be distributed according to the Zipf distribution, given as

$$p_o^r = \frac{o^{-\gamma_r}}{\sum_{c=1}^{O} c^{-\gamma_r}}, \quad 1 \le o \le O, \tag{1}$$

where γ_r is the Zipf's law exponent in region r. Thus, content o is requested across whole network with probability $P_o = \sum_{r=1}^{R} \mathbb{1}_o^r p_o^r$, where let $\mathbb{1}_o^r$ be a $(0/1)$ variable that indicate whether a user requests content o in region r. To proceed, we denote the inter-arrival time of two consecutive requests from a user by random variable T. We assume that the content request process is modeled as a renewal process, in where $N(t)$ denotes the number of requests from a user up to time t.

In this paper, we focus on the content fetching from users' neighbours through D2D communications over a single hop. Motivated by the results in Fig. 5 (b), we assume the transmission rate r_T to be distributed according to exponential distribution with parameter λ_T. Let z_o be the size of content o, and b_o is the bit rate. Then we define that the transmission time of content o as $\tau_o := z_o / r_T$, which is determined by content size and transmission rate r_T. We suppose that when a user u has a content o in its cache, content transmission from node u to node v is successful only if the contact time between u and v is larger than τ_o and transmission rate is higher than b_o. Hence, the success rate of a content transmission q is defined as $q_o := Pr(t_C \ge \tau_o, r_T \ge b_o)$, which takes account into real condition of wireless transmission and QoE.

C. Locality-based Caching Policy

Considering the joint effect of spatial and temporal locality of network traffic and devices contact pattern, i.e., spatial locality of popularity, content active lifetime, content bit rate, inter-arrival time of requests, device contact duration and transmission rate, we define the locality-based caching policy for improving performance of content sharing and aiming to contribute to guideline for making optimal decision on caching.

Definition 1 (Locality-based Caching Policy)**.** Let $\mathbb{1}_{or}^t$ be the $(0/1)$-indicator function that determine whether devices cache content o at time t in region r for a random period of time CT. For all $1 \le o \le O$ and $1 \le r \le R$, let $\mathbb{1}_{or}^t := f(\mathbb{1}_o^r, l_o, \tau_o, b_o)$, given as

$$\mathbb{1}_{or}^t := \begin{cases} 1 & \text{if } \mathbb{1}_o^r = 1,\ l_o > t,\ t_C \ge \tau_o,\ r_T \ge b_o \\ 0 & \text{if } \mathbb{1}_o^r = 0 \end{cases}. \tag{2}$$

When content o is requested, devices will decide whether to copy this content by checking content active lifetime l_o, and evaluating the success rate of a content transmission q_o, which is determined by whether device's contact duration t_C is longer than content transmission time τ_o, and its transmission rate r_T is larger than content bit rate b_o. Hence, if the content o is with higher popularity, longer active lifetime and device is with higher success rate of content transmission q_o, the content will be cached with larger probability. Conversely, the content has little chance of being cached. When the cache is full, the user will discards the least recently viewed content first to make a space for the new content.

V. Content Fetching Performance

In this section, we analyze how likely users can find content o in neighbouring mobile devices' caches when they request a content o at time t, which is determined by the probability $p_f(t)$ that a user can successfully fetch content o from its neighbours (defined as hit ratio). The hit ratio is affected by the probability $p_c(t)$ that nodes having content o in cache within communication coverage d, and users' sharing index p_u. Firstly, we study the content caching probability $p_c(t)$.

A. Content Caching Probability

Suppose that the content request process of a user is a renewal process, we have the following lemma.

Lemma 1. *When inter-arrival time T of a user's content requests follows general distribution with expectation μ, the expected number of times that a content o is requested by a user within time $[t-\Delta t, t]$ is asymptotically equal to $\frac{P_o \Delta t}{\mu}$ for a fixed Δt as t goes to infinity, where P_o is the request probability of content o.*

Proof: Denote by $N_o(t)$ the number of times that content o is requested by a user by time t. Let $N_o(t) = \sum_{i=1}^{N(t)} \mathbb{1}_o^i$, where let $\mathbb{1}_o^i$ be a $(0/1)$ variable that indicate whether the i-th content request is directed to content o, and $Pr(\mathbb{1}_o^i = 1) = P_o$. According to Blackwell's theorem on renewal process, as $t \to \infty$, for any fixed Δt, $E[N_o(t)] - E[N_o(t-\Delta t)] \to \frac{P_o \Delta t}{\mu}$. ■

In the following, we analyze the content caching probability under the defined caching policy. Based on the above lemma, we derive an upper bound on the probability p_c that a user has a content o in storage.

Theorem 1. *Suppose that users' content requests are modeled as renewal process and the locality-based caching policy is applied in each device. Within time interval $[t-CT,t]$, the probability $p_c(t)$ that a user has a content o in storage at time t is asymptotically upper bounded by $\frac{P_o E[CT]}{\lambda_L\lambda_C\lambda_T\mu\tau_o b_o t}(1-p_d(t))$, where CT is a random content caching time, $p_d(t)$ is the probability that user will delete content o from its cache.*

Proof: A node has content o in its cache at time t only if it requested content o within time interval $[t-CT,t]$, i.e., the number of requests directed to o within time interval CT is equal to or lager than 1. Moreover, as defined in locality-based caching policy, the probability of caching content o is also proportional to the active lifetime and success rate of a content transmission. Hence, the probability p_c that a user store a content o at time t is given as $Pr(N_o(t)-N_o(t-CT)\geq 1, l_o > t, t_C > \tau_o, r_T > b_o)$. According to the Markov's inequality, $Pr(N_o(t)-N_o(t-CT)\geq 1)$ is asymptotically upper bounded by $E[N_o(t)]-E[N_o(t-CT)]$ as $t\to\infty$. Based on Lemma 1, $E[N_o(t)]-E[N_o(t-\Delta t)]$ is asymptotically equal to $\frac{P_o\Delta t}{\mu}$. Considering random variable CT, $E[E[N_o(t)]-E[N_o(t-CT)]|CT]$ is asymptotically equal to $P_oE[CT]/\mu$ as $t\to\infty$. Motivated by analysis results in Fig. 3 (b) and 5, we suppose the active lifetime has exponential distribution with parameter λ_L, the contact duration and transmission rate both have exponential distribution with parameters λ_C and λ_T respectively. According to the Markov's inequality, $Pr(l_o > t)$ is asymptotically upper bounded by $\frac{1}{t\lambda_L}$, $Pr(t_C > \tau_o)$ is asymptotically upper bounded by $\frac{1}{\tau_o\lambda_C}$, and $Pr(r_T > b_o)$ is asymptotically upper bounded by $\frac{1}{b_o\lambda_T}$. Moreover, we assume that nodes have limited storage capacity K measured in unit of number of contents and content request pattern follows Poisson process with rate $1/\mu$. Thus, the probability that a device requests a content with in time interval CT is given by $1-e^{-\frac{E[CT]}{\mu}}$. When devices' cache is full, it will discard the least recently requested items first to make a space the new content. At time t, when a user's cache has K or less than K contents, there is no contents be discarded. On the other hand, a user will delete content o from its cache within time interval $[t-CT,t]$ only if up to time t user has at least $(K+1)$ requests, within in time interval CT there is at least one request, and its K recent requests are not directed to content o while the $(K+1)$-th recent request is directed to content o. Subsequently, the probability $p_d(t)$ that user will delete content o from its cache within time interval $[t-CT,t]$ is given as

$$\begin{aligned} p_d(t) &= \sum_{i=K+1}^{\infty}\sum_{j=1}^{i}[P_o(1-P_o)^K Pr(N(t)=i) \\ &\quad Pr(N(t)-N(t-CT)=j)] \\ &\approx \frac{P_o(1-P_o)^K E[CT]}{\mu}(1-e^{-\frac{t}{\mu}}\sum_{i=0}^{K}\frac{(\frac{t}{\mu})^i}{i!}). \end{aligned} \tag{3}$$

Subsequently, $p_c(t)$ is asymptotically upper bounded by $\frac{P_oE[CT]}{\lambda_L\lambda_C\lambda_T\mu\tau_o b_o t}(1-p_d(t))$. This completes our proof. ∎

B. Hit Ratio

Suppose that a user u requests a content o at time t, the user can successfully fetch the content from its neighbours only if at least one neighbour has the content in its cache and stay in connection for content transmission. Hence, we have the following theorem.

Theorem 2. *The probability $p_f(t)$ that a user successfully fetch a content o from its neighbours at time t is asymptotically equal to $\frac{\pi d^2\lambda_U p_u P_o E[CT](1-p_d(t))}{\mu e^{(\lambda_L t+\lambda_C\tau_o+\lambda_T b_o)}}$, where p_u is the sharing index.*

Proof: Results in Theorem 1 shows that a user has a content o in its cache with probability p_c at time t. Accordingly, the number of neighbours X_c that have a content o is given as $X_c=\sum_{u=1}^{n}\mathbb{1}_o^u$, where let $\mathbb{1}_o^u$ be a $(0/1)$ variable that indicates whether a node u has content o in its cache, and $Pr(\mathbb{1}_o^u=1)=p_c$. Thus, $X_c=np_c$. Hence, nodes having content o are also distributed as a independent homogeneous poisson point process Φ_C with density $\lambda_C=p_c\lambda_U$. Besides, let user sharing index p_u be the probability that devices stay in communication. Subsequently, the probability that a user can successfully fetch a content o from its neighbours is given as

$$\begin{aligned} p_f(t) &= p_u(1-e^{-\pi d^2\lambda_U p_c(t)}) \approx \pi d^2\lambda_U p_u p_c(t) \\ &= \frac{\pi d^2\lambda_U p_u P_o E[CT](1-p_d(t))}{\mu e^{(\lambda_L t+\lambda_C\tau_o+\lambda_T b_o)}}. \end{aligned} \tag{4}$$

∎

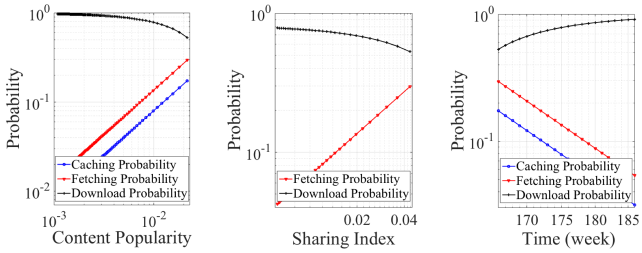
Remark 1. From the analysis above, we observe that the probability $p_f(t)$ of fetching content o various with time passes and is indeed impacted by content popularity, content active lifetime, contact duration, transmission rate, expected caching time, user request rate, user sharing index, and the density of users. This further indicates that to achieve optimal content sharing performance, each device should take account of joint impacts of these parameters in making right decision on caching contents.

C. Numerical and Trace-driven Evaluation

In addition to mathematical analysis, we use numerical simulation to evaluate the content fetching performance under the locality-based caching policy applied in D2D-based content sharing network. Let $1-p_c-p_f$ as the probability that a user download a content o from servers. Motivated by the results in the trace analysis, we assume that the content popularity in a specific region r is according to Zipf distribution with Zipf's law exponent $\gamma_r=1$, and the user sharing index is according to Zipf distribution with Zipf's law exponent $\sigma=1$. Then, we have $p_o^r\approx 1/o\ln O$, and $p_u\approx 1/u\ln n$. We set parameters: $n=100$, u various from 5 to 35, $O=10000$, o various from 5 to 100, $R=5$, $CT=24$ (hours), $\mu=3$ (hours), $d=10\ m$, $l_m=166$ (week), $\beta=15$, $\lambda_U=0.5$, $\lambda_C=0.0042$, $\lambda_T=0.0012$, $\lambda_L=0.1135$.

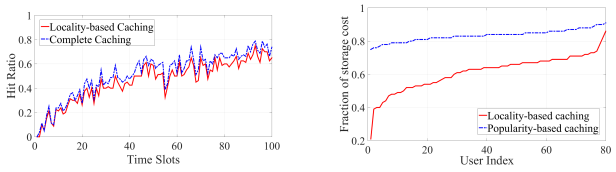
Fig. 7 gives a first glance of content fetching performance, which shows that content fetching from neighbouring mobile devices through D2D communication can indeed offload traffic from network servers significantly, especially for popular contents in strong sharing devices: 50% servers' traffic can be

offloaded through D2D communication and fetched from own cache. From Fig. 7 (a) and (b), we observe that when contents are more popular and device has a larger sharing index, we can obtain higher fetching probability and offload traffic from servers dramatically. Besides, because of majority of contents' growth trend factor is less than 1 and popularity decreases with time passes, caching and fetching probability decrease, which leads to the increase of traffic load in network servers, as shown in Fig. 7 (c). In addition, we evaluate the proposed caching policy by using real-world traces, i.e., YouTube Video [8, 9] and Bluetooth contact pattern [10]. In this trace-driven simulation, we define the successful fetching from neighbours only if the transmit rate is higher than the requested content bit rate and the contact duration is longer enough to complete transmission. In Fig. 8, we compare the hit ratio and storage cost of locality-based caching policy with the popularity-based caching policy. We observe that the hit ratio of these two policy are almost same, although the hit ratio in locality-based caching policy is a little lower than the popularity-based caching policy. However, the storage cost in the locality-based caching policy is significantly smaller than the other one.



(a) Content popularity (b) User sharing index (c) Time passes

Fig. 7. Client caching probability, neighbouring fetching probability and servers downloading probability.



(a) Hit Ratio (b) Storage Cost

Fig. 8. Trace-driven simulation results.

Remark 2. From the above numerical and trace-driven simulation results, we can find that content can be successfully fetched from close-by devices through D2D communication, which can indeed reduce network and server load. Besides, the trace-driven simulation results is a strong evidence that considering the joint impacts of content popularity, content active lifetime, caching duration, user request rate, contact duration, transmission rate, and the density of users in the network is necessary and effective when making caching decision. It is also a guideline for designing caching policy.

VI. Conclusion

In this paper, we study how mobile devices make right decision on caching various contents for achieving optimal content fetching performance in D2D-based content sharing network. To validate the impact of real-world traffic and devices opportunistic communication on content fetching performance, we first extract the spatial and temporal characteristics of network traffic, and devices contact pattern through analyzing real-world traces. Motivated by these results in traces analysis, we design the locality-based caching policy. Through mathematical analysis, we derive the upper bound of content caching probability and hit ratio. Numerical and trace-driven simulation results not only evaluate the content caching probability, content fetching and traffic offloading performance, but also offer a strong evidence and guideline for each mobile device making optimal decision on content caching.

References

[1] C. V. Forecast, "Cisco visual networking index: Global mobile data traffic forecast update 2015-2020," *White Paper, February*, 2016.

[2] Ericsson, "Ericsson mobility report: On the pulse of the networked society," *Ericsson, Jun*, 2016.

[3] Y. Zhao, Y. Li, H. Mao, and N. Ge, "Social community aware long-range link establishment for multi-hop d2d communication networks," in *Communications (ICC), 2015 IEEE International Conference on*, June 2015, pp. 2961–2966.

[4] T. Wang, Y. Sun, L. Song, and Z. Han, "Social data offloading in d2d-enhanced cellular networks by network formation games," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 7004–7015, Dec 2015.

[5] J. Jiang, S. Zhang, B. Li, and B. Li, "Maximized cellular traffic offloading via device-to-device content sharing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 82–91, Jan 2016.

[6] E. Baştuğ, M. Bennis, and M. Debbah, "Social and spatial proactive caching for mobile data offloading," in *Communications Workshops (ICC), 2014 IEEE International Conference on*, June 2014, pp. 581–586.

[7] H. J. Kang, K. Y. Park, K. Cho, and C. G. Kang, "Mobile caching policies for device-to-device (d2d) content delivery networking," in *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*, April 2014, pp. 299–304.

[8] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Computer networks*, vol. 53, no. 4, pp. 501–514, 2009.

[9] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, June 2008, pp. 229–238.

[10] A. Caputo, A. Socievole, and F. D. Rango, "CRAWDAD dataset unical/socialblueconn (v. 2015-02-08)," Downloaded from http://crawdad.org/unical/socialblueconn/20150208, Feb 2015.

[11] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '11. New York, NY, USA: ACM, 2011, pp. 305–316. [Online]. Available: http://doi.acm.org/10.1145/1993744.1993776

[12] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, June 2007.

[13] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, November 2011.