Caching on the Move: Towards D2D-based Information Centric Networking for Mobile Content Distribution

G. Chandrasekaran, N. Wang and R. Tafazolli
Institute of Communication Systems
University of Surrey
United Kingdom
{g.chandrasekaran, n.wang, r.tafazolli}@surrey.ac.uk

***Abstract*—With the advent of device-to-device (D2D) communications, user equipment (UE) such as smart phones will become an integral part of the (mobile) network for content distribution operations. In this context, we introduce a novel Information centric networking (ICN) framework for mobile content caching and distribution based on direct D2D communications in cellular network environments. Specifically, small pieces of mobile content can be cached at incentivised mobile UEs known as helpers, and the ICN-aware cellular network edge (e.g. base stations) are able to resolve content requests from local mobile clients to the helpers in their D2D proximity. With simple POI (point of interests) based selection of helpers as well as content caching/eviction control at the base station side, a significant proportion of mobile content requests can be locally resolved to helpers in proximity of clients, thus achieving very effective content traffic offloading away from the cellular network infrastructure.**

I. INTRODUCTION

Mobile content access keeps increasing day by day due to rapidly growing popularity of smart devices (e.g. smart phones and tablets) as well as emerging social media applications. According to Cisco's Visual Network Index (VNI) published in 2015, mobile data traffic is expected to grow at a rate of 57% from 2014 to 2019, reaching 24.3 Exabyte per month and there will be over 11.5 billion mobile-connected devices by 2019 [1]. Giving that cellular network infrastructures are considered as one of the most common communication platforms for mobile content access and delivery, how to support the explosive growth of mobile content traffic volume will become a key research issue for the design of future 5G-based mobile cellular systems.

During the past few years, research efforts have been mainly spent on new network architectures which are able to natively handle networked content at large scale, with content/information centric networking (CCN/ICN) paradigms being a typical example [2]. The design rationale of ICN/CCN is to make the network infrastructure to be content-aware, in which case the network is able to intelligently perform content-based operations such as content resolution/searching, delivery, as well as storage/caching [2]. On the other hand, it is worth mentioning that, up to now the mainstream of CCN/ICN research have been focusing on fixed ISP networks, in the sense that network routers are made to be content aware, including the support of in-network caching of popular content objects. Most recently, proposals have also been made towards the support of content caching in LTE based mobile cellular network environments [3, 4]. Specifically, with popular content objects being cached at the mobile network edge (e.g. at base stations (BSs)), it is envisaged that the end-to-end access and delivery delay of mobile content cached at the network edge can be substantially reduced.

Recent technology development in Device-to-device (D2D) communications has enabled direct communication between mobile devices, but without the data traffic being delivered through the cellular network, including base stations [5-8]. Meanwhile, emerging mobile devices such as smart phones are featured with high computational power, larger storage capacity, long battery life, accurate GPS location functions, as well as heterogeneous wireless interfaces such as cellular, WiFi and Bluetooth. Such advanced features allow mobile devices to play a more active role in content distributions rather than simply being a dummy device for consuming mobile data. A typical example in this case is D2D based traffic offloading [5], where (relatively small) mobile content objects can be carried by mobile devices and disseminated to nearby interested consumers directly without involving the cellular infrastructure in the data plane. In this case, mobile devices can be deemed as an integral part of the mobile network with own contribution of resources for serving other clients. The obvious benefit is to effectively save expensive cellular spectrum resources through offloading to D2D based communications in support of content delivery.

In this paper, we introduce an ICN-based network architecture which expands content intelligence further to the mobile device level, and by leveraging on less expensive communication channels such as D2D WiFi, mobile content can be efficiently cached at selected mobile devices and opportunistically distributed to other mobile content consumers in proximity. Such a scheme offers an ideal solution for efficiently distributing small pieces of social media content which normally attracts large group of audience. Recent studies in [9] report that user interests in social media content have significant impact on its *locality* and *homophily* characteristics. This means that people geographically close to each other may have common or similar interests of content objects (locality) and also the users are both clustered by regions and interests (homopily), which is also called "birds-of-a-feather" effect

[10]. An efficient device level content sharing scheme could take advantage of these characteristics to offload these social media traffic via D2D.

Firstly, we introduce in the paper a full suite of content manipulation protocols based on ICN design principles, including content resolution, delivery and caching/eviction management. Specifically, content resolution functions and caching decision intelligences are embedded at the mobile network edge, typically eNodeBs or base stations (BSs) close to content clients. Each BS has the awareness of the mobile content popularity under its coverage, and based on such knowledge, it pushes the content to a set of strategically selected mobile devices (called incentivised helpers, or simply *helpers*) for caching. Upon receiving incoming content requests, each BS (as a content resolver) resolves them directly to one of the helpers who have already cached the requested content, and also is in proximity of geographical locations of the requester. Then based on direct D2D based communications, the helper is able to directly serve the content to the requester in D2D proximity. It is worth mentioning that the proposed scheme only allows content to be served by a helper in proximity which has already cached it. This is contrast to most of other DTN (delay tolerant network) schemes that allow multi-hop content delivery. The rationale behind is to assure user experience for mobile content access within reasonable waiting time.

It can be easily inferred that, the selection of helpers will have a significant impact on the efficiency of D2D based content distribution. Helpers are set of users in the network who are incentivised to disseminate mobile content to interested consumers, based on their own cache space, via direct D2D communication. Incentives can be in terms of better QoS or additional data allowance offered from the mobile operator [11]. A key technical issue in helper selection is how to achieve efficient D2D-based content distribution based on opportunistic encounter between helpers and potential content clients. This requires necessary knowledge on (predicted) mobility patterns on the helper side. The good news is that, it has been observed that the daily mobility patterns of the majority of mobile users are regular and easy to predict [12, 13]. Based on this, we propose in this paper a simple strategy for selecting a small proportion of helpers that are responsible for distributing mobile content to encountered consumers. Through extensive mobility trace-driven simulations over two different mobility models, namely the shortest path map based movement model (SPMBM) and the working day movement model (WDMM) [13], we show that the upper bound of mobile content offloading ratio to D2D is 50.1% based on the proposed ICN scheme with a Point of Interest (POI) based helper selection strategy. Even though we have not yet specifically considered lower layers parameters in this paper (i.e. simply assuming congestion-free WiFi channel for D2D communication), we believe that this paper sheds lights on a potentially promising solution for future mobile content distribution services based on the natural combination of ICN and D2D technologies.

The remainder of the paper is organised as follows. In Section II we present the overall system design, including the basic illustration of content handling functions at the BS side. We also provide the specification of content resolution and delivery mechanisms for direct content forwarding from designated helpers towards mobile clients. In section III we discuss the proposed content caching and eviction techniques based on the proposed architecture. We present our performance evaluations in section IV based on two representative mobility models. Related works are summarised in section V, followed by our conclusions in Section VI.

II. Overall Framework Design

A. Overview

While the mobile content distribution takes place directly between devices (user equipment – UE) in the data plane, the control plane functions for handing content are mainly located at the cellular network edge, typically at the eNodeB or base station (BS) level. Specific ICN-based mobile content control functions at the BS side include:

- *Content resolution*: Whenever a content client requests a certain mobile content, it always issues the content requests towards the BS it is currently associated with. Such a content request can be either resolved to remote content sources, or alternatively towards a helper in proximity which has already cached the content. Upon an incoming mobile content request, the BS takes the ICN content resolver role by looking up its locally maintained *content resolution table* (CRT). If the content is currently being carried by a helper which is within the D2D communication range of the requester (both commonly under the radio coverage of the BS as well), the BS will simply forward the request towards the targeted helper in proximity of that client. Towards this end, the BS (resolver) should have the knowledge about the actual locations of both content clients and the helpers, typically through location-based services like GPS. In addition, the BS should also be aware what popular mobile content is currently being carried by each helper. The latter function is associated with the content caching/eviction control which will be introduced at a later stage.
- *Content delivery setup*: Once the content request has been resolved to the helper in the D2D proximity of the requesting client, the helper is then expected to transmit the locally cached content to the client through direct D2D link. In this case, the link establishment for such content transmission between the helper and the client is handled by the BS. For the sake of minimising complexity, the proposed scheme only allows *single-hop* content delivery, which can be relatively easily handled by the BS. This is in contrast to the traditional multi-hop delay tolerant networking (DTN) in which content objects are forwarded across a chain of devices autonomously [6, 8, 14].
- *Content caching control at the helper device level*: In addition to the handling of client requests as an ICN resolver, a BS also takes the responsibility of content caching controller for those helpers currently located in its own radio coverage. Specifically, since all the content requests are first sent to the BS as the content resolver, the BS has the knowledge on the popularity of the content objects being requested in the local region. Based on such knowledge, each BS autonomously makes decisions on what need to be cached and carried by the

local helpers. For this purpose, the BS also maintains a *helper table* (HT), which records the cache utilisation/status of each helper. There are two options for enforcing the caching of selected content at the helpers. The first option is for the BS to directly push the content to the helper, even if the helper may not be interested in the content. Alternatively, the BS can instruct the helper to cache its previously requested (popular) content and serve other clients in proximity which are also interested in it. In both cases, specific service level agreements (SLAs) should be established by the mobile operator and the helpers on their caching operations.

As can be seen from the overview, there is a clear separation of the content control plane and the data plane in the proposed scheme. While the BS is not directly involved in the data plane for transmission of mobile content, it plays an essential role in the control plane for orchestrating the content distribution operations between devices, including resolving content requests to helpers, coordination of D2D communication link establishment, as well as decision-making for content caching. Now we proceed to the detailed introduction of the proposed scheme based on a simple illustrative example.

B. Content resolution and delivery operations

Each BS needs to maintain a CRT which records the mapping of the mobile content identifier and its cached location(s) (i.e. the helper(s)). Additional meta-data can be also included, such as content size, popularity information which is useful for efficient caching decision making by each BS.As shown in Fig. 1, the client UE makes a request carrying a unique content identifier. Its currently associated BS resolves the content request and finds that the helper UE has that particular content. If a helper UE is in proximity range of the client according to the location knowledge obtained by the BS, then the BS will direct the content request to that helper, which carries the address of the requester as the destination of the content. In this case, since it will be the helper that actually initiates the D2D communication for content forwarding, the BS should at the same time coordinate with the helper regarding the establishment of the D2D radio link. Such a procedure can be simply adapted from existing D2D communication link establishment protocols [15, 16].

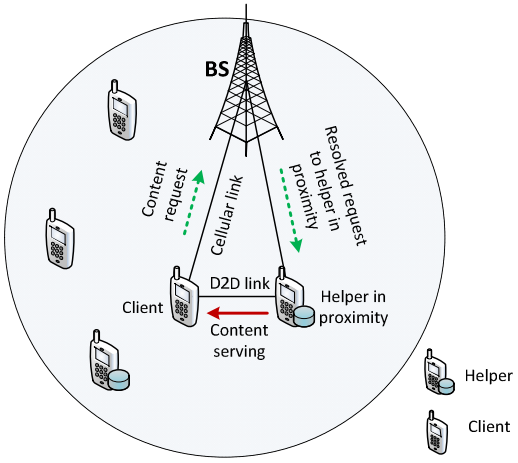


Figure 1. Basic scheme illustration

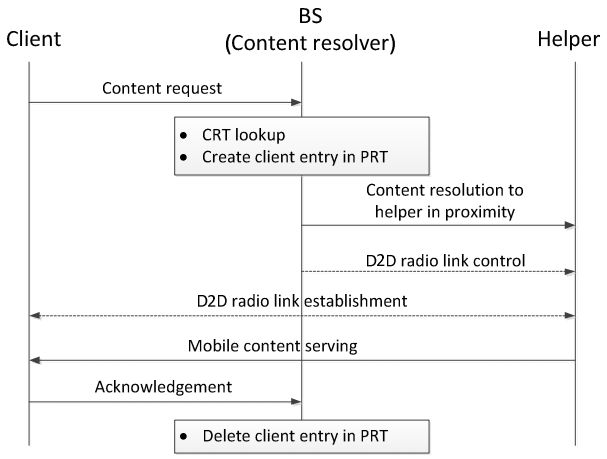


Figure 2. Signalling chart for D2D based content resolution and delivery

In general, D2D communication between a helper and a client can be supported by one of the following three ways: direct WiFi communication or in-band spectrum communication or out-band spectrum communication. In in-band spectrum communication, communication between two devices utilises the same frequency as the current BS cell operates. Whereas in the out-band option, communication between two devices takes place in the frequency assigned explicitly for D2D communications.

Details on the usage of the spectrum are outside the scope of this paper but will certainly remain as part of our future work. Once receiving the directed content request from the BS, the helper is ready to forward the content to the requester. Upon the success of the content transmission, the client UE then sends an acknowledgement message back to the BS, indicating it has already obtained the requested content from the designated helper. This message can also be used for crediting helpers that have fulfilled the serving task [11]. However, in case the BS fails to receive any acknowledgement from the requester following its forwarding of the content request to a local helper within a given threshold period, it will resolve again the pending request to the original content source rather than relying on opportunistic D2D based content distribution. Such a situation can happen especially when the helper is moving outside the proximity range.

In order to successfully resolve every incoming content request, the BS need to maintain a pending request table (PRT) which records the states for those content requests that have been directed towards local helpers. Upon the receipt of a successful acknowledgement, the corresponding request is removed from the PRT. Fundamentally, the function of the PRT is similar to the pending interest table (PIT) in the CCN scheme [2]. The overall signalling chart for a successful content resolution and delivery is presented in Fig. 2. The solid lines represent the signalling for the proposed content resolution and delivery operations, while the dash lines indicate the necessary D2D radio establishment arrangement according to 3GPP [17].

C. Device-level content caching control

Now we discuss how a BS, taken the role of a content caching controller, actually manages the cached content at the local helpers under its coverage. First of all, there are different options for decision-making on what mobile content to cache. This can be either based on recommendation of the mobile content source (e.g. breaking news which by nature is expected to have a large crowed of audience), or alternative based on the actual popularity measurement according to received content requests at the BS side. In addition to popularity information, other factors can also be taken into account such as content size and other attributes. In this paper we mainly focus on the second scenario where the BS needs to have the intelligence to determine what to cache based on its own observations of content popularity according to received requests. While dealing with normal incoming content requests, the BS can build the knowledge on the most popular set of content objects which are desired to be cached at helpers in order to offload mobile content traffic from the cellular infrastructure to D2D. Towards this end, a popularity threshold can be preconfigured, which can be in terms of number of received requests within a unit of time period. If the popularity of a specific content exceeds the threshold, then the BS can identify one or multiple helpers to cache the content.

An interesting technical issue is mobile content replacement/eviction control, as newly published mobile content with higher popularity will need to replace old ones already cached on helpers when there is no free space for caching new content on the device. According to our scheme, the BS, with the overview of content popularity (according to the captured statistics on the received content requests), is responsible for content eviction decision making. With the information maintained at the helper table (HT), the BS knows what has been already cached at each helper, and also the *least popular* content cached there (including content size). So when there is a new content that needs to be cached at the helper, the BS knows exactly which old content needs to be evicted from the cache. Detailed description on this will be presented in Section III. In case a helper is moving out of the radio coverage of the BS, the BS should be able to realise this situation based on the legacy LTE mobility management function of the cellular system. In this case both the HT and the CRT should be updated in order to exclude the helper from the list.

D. Practicality Considerations

In this paper we do not cover the advanced scenario how a helper device can remain an active helper when traversing a number of cells/BSs. It will be our future study to design an advanced scheme that allows intelligent helper control coordination between neighbouring BSs for a mobile helper to serve a larger geographical area.

On the device energy side, there is no doubt that additional energy needs to be consumed for opportunistically serving content on the helper side, as compared to other conventional UEs. Incentives such as higher priority in energy harvesting services for device battery sustainability on the helper side could potentially address this problem (i.e. can serve as win-win case for both helpers and mobile operators). The energy consumption pattern depends on a wide range of the factors such as, signal strength, transmission range, size of data transferred and interference level. Energy consumption dependency between the sizes of transferred data using different network interfaces are obtained and studied from [18]. The general observation is that WiFi radio is the desired communication medium for our purposes, whereas LTE and 3G is less energy efficient for transferring smaller amount of data. On transferring small-size data by LTE, energy per bit decreases as bulk data size increases[18]. WiFi-enabled D2D based offloading schemes can be used to offload these small pieces of content to subscribers with minimum energy consumption (performance analysis in Section IV). When a helper's remaining energy drops below a given threshold, a signalling message is triggered to the BS, which will therefore stop using that particular user as helper, in terms of both resolving requests and pushing new content for caching (same as the out-of-range case aforementioned).

The proposed scheme with content-awareness is fully in-line with the Mobile Edge Computing (MEC) paradigm proposed in 3GPP [17]. MEC provides IT, storage and computing capabilities within the Radio Access Network in close proximity to mobile subscribers. Some MEC functionalities can directly support the required intelligence in our framework: E.g., On-Premises location (deployed either at the LTE macro base station (eNB) site, or at the 3G Radio Network Controller site) meaning that it is located close to users and can run isolated from the rest of the network, while having access to local resources. Location awareness, network measurement based tracking of active (GPS independent and network determined) users, and access to real-time network context data (e.g. dynamic link quality, radio and network information) can be used to develop a more intelligent scheme with network condition awareness. Security is another concern in direct D2D communication. In case of *general* D2D content distribution, a malicious UE could modify the content before passing it to other clients. In fact, according to the existing D2D-based traffic offloading schemes [6, 8, 14] all the UEs can participate in D2D based content transmission, and also across multiple hops. In this paper the BS (using embed secured computing features of MEC) only re-directs content requests from clients to the pre-selected, *authorized* and *authenticated* helpers for serving clients in proximity. Such a feature ensures higher degree of security as compared to previous schemes. While D2D communication is foreseen to be an integral part of cellular network [17], implementing such framework will involve no additional deployment cost for network operators (apart from installing content control functionality at the network edge).

E. Helper Selection Strategies

We first discuss the strategies of helper candidate selections in the ecosystem, as an *offline* procedure. In the literature, helper selection schemes have largely been based on previous

user mobility patterns (e.g. how active the user is?), connectivity history e.g. (how many users has it encountered?) or based on some simple algorithms like greedy, heuristic and random selection [8, 14]. In our scenario, we select incentivised users as helpers based on its willingness and time spent in proximity to given POIs. In a real time scenario, a set of POIs can be identified such as tourist attractions, shopping centres and even large business/office areas (in particular on working days). Based on the identified POIs, we select a top proportion of helper candidates that are predicted to spend *the longest duration* with them. Similar to the literature work, this is based on the derived knowledge on their long term mobility patterns. It is worth mentioning, it is *not* our intention in this paper to design the most appropriate helper selection algorithm, and indeed we do not believe there is a universally best solution for supporting different D2D content distribution scenarios. Instead, we would like to indicate that, the actual content distribution efficiency indeed can be substantially influenced by the helper selection, and even a simple but POIs based strategy can substantially improve the performance, as will be presented in Section IV.

III. Content Control Algorithm

As mentioned earlier, all helpers have cache memory for caching mobile content. When cache memory at a helper becomes full, and when a new content object has to be accommodated, an algorithm has to make a decision on which existing content should be evicted from the device memory. When a BS decides to push a new content to be cached at the helper side, it selects a helper from the cell which has the currently *least popular content* and also with *larger content size* than the new content to be cached. For this purpose, the BS should also have the knowledge on the size of the content, and this can be obtained by the previous delivery of same content through the BS before it is considered for caching. If the existing content object is less popular, then the content replacement will take place. The process of pushing a more popular one to helper's cache memory is called an *implant*. This concept of content eviction is described in Algorithm figure below.

Decision making on content caching/eviction at a Helper

INPUT:

- A new content object x potentially to be cached
- P_x – Content popularity for x in the current window size (T_{sw}) and C_x – Content size of x
- K : A list of already cached content objects at the helpers, including their popularity in the current window (P_k) and content size (C_k), $1 \leq k \leq |K|$

DO:

Sort K in ascending order according to P_k;
For $k = 1$ to $|K|$ in the sorted cached list
  if ($P_k < P_x$) **and** ($C_k \geq C_x$) **then**
    Identify the helper H_k that is currently caching k
    Replace content k with x at H_k;
  else skip;
End for

We are employing this simple eviction scheme to be executed on the BS side, where content with the least number of request counts are removed and replaced with content with more number of request counts, according to a sliding window size (T_{sw}). The actual algorithm is an adapted version of the generic Least Frequently Used (LFU) policy. In LFU only the number of requests for each content object is noted whereas, in our scheme we also note the approximate time on which the content was requested. LFU in general can be a very effective algorithm but it becomes less efficient in caching content with highly dynamic popularity patterns such as social media (compared to professional VoD content which cannot be supported by D2D caching). This is because the trend of data popularity might change quickly over time, which for LFU to adopt it and retain popular content will be time consuming and less effective. So in our algorithm we follow the below steps:

- Check if a request for i □ R' arises, then BS increments value of P_i. For scalability reasons, we keep P_i's count for only a fixed period of time T_{sw} (using a sliding window). R' is the set of requests for content, which are not cached at the helper side but are eligible to be in helpers due to its popularity.

- To add a new content x, find content k = arg min P_i ∀ ($1 \leq k \leq |K|$) where P_k's lifetime expires after T_{sw}. Compare P_x against P_k, if $P_k < P_x$ and $C_k \geq C_x$ then replace content k with x at H_k.

By this algorithm, popularity of content can be learned and adapted quickly, and hence data cached can be used more effectively. It is worth mentioning that, alternative content caching/eviction schemes can be directly applied in the proposed architecture in a plug and play fashion, and it will be our future work to investigate more intelligent decision making algorithms.

IV. Performance Evaluations

A. Simulation experiment setup

We use the city road map obtained from [19] for our performance evaluation. A new application layer was coded over Opportunistic Network Environment (ONE) to enable the eviction algorithm functions and Hybrid D2D caching algorithm [20]. The percentage of offloaded requests (OR) and percentage of implanted requests (IR) are the main performance metrics considered in our scheme. OR is the ratio of content requests served via device helpers against the total number of requests made by all clients, i.e. directly proportional to the number of D2D based offloaded requests. IR is the ratio of implants made on helper's cache for future dissemination against the total number of requests made by all clients, i.e. directly proportional to the number of implanted requests. We also evaluate the signalling overhead ratio,

$$\text{Signalling overhead ratio} = (N_r - N_d) / N_d$$

Where, N_r is the number of messages relayed and N_d is the number of messages delivered, i.e. is the average number of forwarded copies per message. Low value of overhead means less processing required for delivering the messages. It is worth

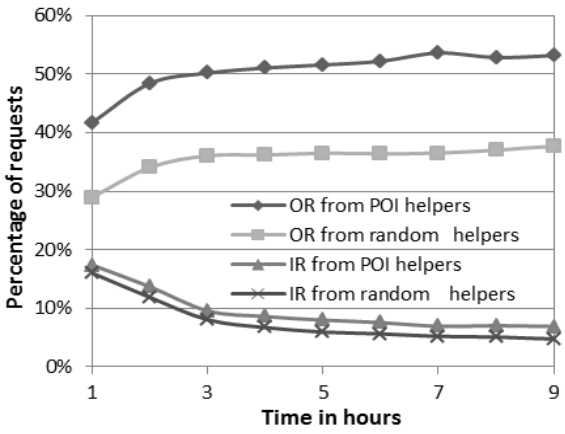
mentioning that, total relayed messages in our scheme will be the combination of implants made on helpers and a hop from helper to subscriber. We simulated the network with 500 mobile nodes and set the maximum number of data items to be 1000 (randomly distributed from 100 KB to 10 MB). The area of the map was 2300 x 2400 m^2, i.e. not densely populated for 500 mobile nodes. Cache size of each helper node was set to be 500MB (i.e. roughly 10% of total content in network). We are using WiFi for D2D communication, hence D2D transmission range of each node was 30 meters, rate of transmission was set to be 2Mbps, and our T_{sw} was set to be 2 hours. Experimental results have showed that two phones can exchange up to 1.48 MB of data during their short inter-contacts using WiFi [5].

Zipf distribution is well known for the use in content popularity modelling. All client nodes generating content requests follow Zipf distribution with skew value of 0.6. By default we used POI based helper selection algorithm to select helpers. We also set the default delay tolerable or access time (D_t) to be 5 seconds, i.e. difference between requested time and start of content reception time should be less than 5 seconds. That means a pending request can remain in the PRT for up to 5 seconds, and the BS will resolve a request in the PRT to the original source after its 5 second lifetime without receiving the acknowledgement on the D2D based distribution from a helper. Unlike other proposed D2D offloading schemes, parameter D_t makes sure that content is offloaded to subscribers with realistic access delay.

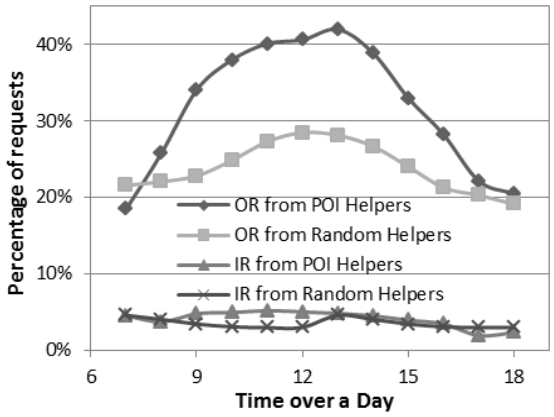
B. Performance Evaluations

We exam the proposed scheme based on two different mobility models, SPMBM [20] and WDMM [13]. User in SPMBM selects a destination from list of POIs on the map and takes shortest route to that point. These POIs may be any popular real-world destinations such as restaurants or tourist attractions or popular shops [13]. SPMBM emulates mobility of users who are tourists or user's weekend mobility characteristics. WDMM emulates human mobility behaviours on working day basis, e.g. between home and offices considering daily working hours. WDMM is proved to have characteristics such as, inter-contact time and contact time distributions similar to that of realistic real-time mobility traces like, Reality Project of MIT, Cambridge trace gathered by the Haggle project and Dartmouth traces [13].

In Fig. 3 (a) we plot the percentage of OR and IR under SPMBM model over time. Both the curves undergo two stages i.e. variable phase (VP) and saturation phase (SP). VP is a phase where the caching schemes tend to find the right popular contents in helpers cache memories (before the caching space is fully used). As observed, the OR ratio keeps increasing in both random and POI based helper selections, with increasing numbers of content objects being cached at helpers. Over time this phase transforms into SP, in which all the cache memories of helpers are occupied with maximum popular contents. Any popularity shift on data in the network will affect the SP, and hence a sliding window approach (with T_{sw}) is followed to make this phase as steady as possible.

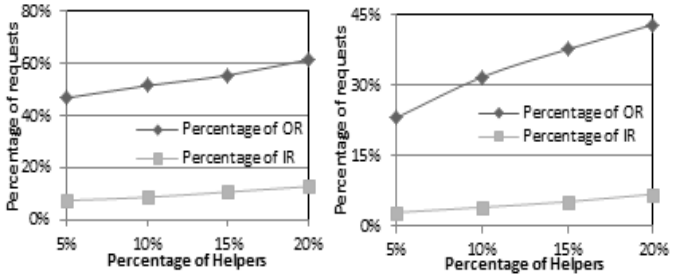


(a)

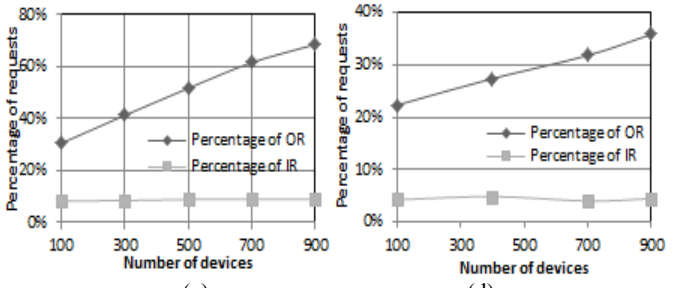


(b)

Figure 3. OR/IR performance based on (a) SPMBM and (b) WDMM



(a) (b)



(c) (d)

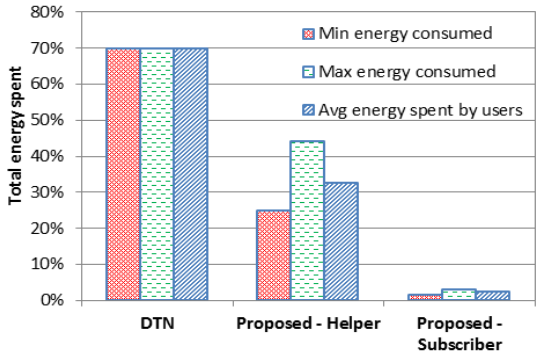
Figure 4 (a) OR/IR performance with varied helper proportions on SPMBM and (b) WDMM and (c) OR/IR performance with various population density on SPMBM and (d) WDMM

From Fig. 3 (a) we see that, with random helper selection algorithm 35.9% of total requests were offloaded on an average with just 7.2% of implants. So it's clear that with our framework a simple POI based helper selection can outsmart the random selection scheme. We now evaluate the performance based on the WDMM. It can be easily inferred that during idle hours (e.g. mid-night) where human mobility is extremely low, and hence the proposed D2D based content dissemination technique cannot be applied in practice. As such, in our simulation study we only focus on the peak-hour period of a single day. As depicted in Fig. 3 (b), with random helper selection, up to 28.5% of mobile content requests was offloaded to helpers during the given period. In contrast, with the POI based helper selection algorithm, up to 41.9% of requests was offloaded. In both cases, the success rate of offloading to D2D helpers reaches its peak during middle of the day when the overall number of inter-contact rate between users becomes maximum. Understandably, with lower volume of contact rate early in the morning or in the evening, the OR performances are not as promising as those during the peak. IR in both the cases was fairly the same. The rate of IR depends entirely on the caching algorithms used.

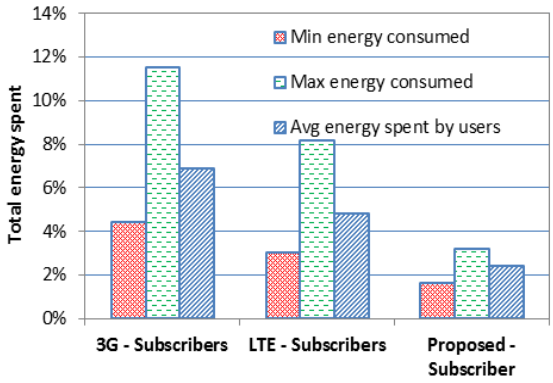
We used POI based helper selection scheme to evaluate upcoming scenarios. We now evaluate how the population of helpers can influence rate of OR on SPMBM. In Fig. 4 (a) we can see that percentage of OR is linearly proportion to the percentage of the helpers. For instance, even as low as 5% of helpers used can lead to up to 46% of content requests being offloaded to D2D helpers. However, it is worth noting that this value is only a promising upper bound, as in our study we do not include constraint imposed from lower layer networking mechanisms and potential network traffic conditions. Same observation is observed for WDMM as in Fig. 4 (b). However, compared to Fig. 4 (a) and (b), the actual OR rate is lower than the SPMBM scenario. Meanwhile, Fig. 4 (c) indicates the offloading performances with different user density on SPMBM. As observed, higher the density of subscribers, higher the rate of OR, thanks to increased probability of encounters between the helpers and the mobile clients. Whereas, the rate of IR is almost the same because percentage of helpers in the network is unchanged. It can be seen from Fig. 4 (d) that characteristics of both percentage of OR and IR in WDMM is similar to that of SPMBM. However, again this should be deemed as an upper bound of the actual performance, as we do not take into account lower layer constraints with increased number of devices.

TABLE I. PERCENTAGE OF DATA OFFLOADED AND OVERHEAD RATIO

Percentage of data offloaded over various D_t			
Tolerable access time	Proposed framework	Greedy selection	Spray and wait routing
50s	54.1%	0.5%	2.2%
30m	56.6%	28.7%	36.7%
Infinite	57.1%	55.5%	66.9%
Overhead ratio over various user density			
No. of nodes	Proposed framework	Greedy selection	Spray and wait routing
100	0.2	4.63	2.99
500	0.12	138.97	5.03
900	0.09	370.38	5.13



(a)



(b)

Figure 5 (a) Total energy spent by users on various frameworks and (b) Total energy spent by subscribers in proposed framework, LTE and 3G network.

In addition to the performance analysis on our proposed scheme, we also evaluate its benefit and cost against some existing (multi-hop) D2D offloading frameworks in literature. In Table I below, we compare the percentage of data offloading under various tolerance degree of content access time (D_t). Due to the single hop-content delivery policy in our proposed scheme, it can be inferred that it is able to offload more mobile content traffic in limited time duration compared with other multi-hop based schemes [7, 8]. Indeed, in general non-real mobile content delivery applications do require content access time at the time scale of seconds for the sake of user experiences. On the other hand, even without any bounded content access time constraints, our proposed scheme is comparable with existing schemes, we also see that overhead ratio of existing schemes are substantially higher. Overhead ratio of our scheme decreases over increase in population density, whereas it is in contrast to the existing frameworks [7, 8]. This is because fewer helpers can satisfy more users in a densely populated network (thanks to ICN enabled BS), whereas with existing schemes data has to route via many hops to reach destination.

In Fig. 5 (a) we show the energy consumption of helper and subscriber with same set of input parameters as before. We show that energy consumed by helpers in our scheme is much less compared to the existing multi-hop frameworks [8]. At the same time in Fig. 5 (b) we also show that energy consumed by

subscribers is lower compared to subscribers obtaining data directly from LTE or 3G network. With the help of resolver at network edge (e.g. BS) our approach can serve as an ideal solution for offloading smaller social media traffic away from cellular network with minimal energy consumption. To note that median content size in Youtube is 8.4MB [21].

V. Related work

ICN is deemed as one of the promising network architectures in recent years which tends to moves current Internet architecture away from host-centric paradigm to a network architecture in which the focal point is "named information" [2]. Caching and replacement of addressable content in every cache-equipped network devices (e.g. routers or switches) can be done at line-speed [2].

Recent study by Intel Corporation showed that, a content based distributed caching technology in cellular network can provide improved backhaul, transport savings and improved QoE. Content caching has the potential to reduce backhaul capacity requirements by up to 35%. Local DNS caching can reduce web page download time by 20% [22]. Opportunistic offloading by direct D2D communication was exploited widely in latest works [6, 8, 14, 23] and proved to offload more than 70% of delay tolerant traffic from cellular infrastructure. In case of D2D based offloading schemes, [6, 8] all the devices are actively involved in data forwarding. Devices start to share content with other devices on receiving it, i.e. involving multi-hop opportunistic forwarding. Interference and processing are much higher in these cases. Practically not all users in the network will be willing to disseminate content to peers. Whereas, in our scheme only helpers are allowed to disseminate content, and we select users as helpers based on their willingness to disseminate, and also we also try to offload contents at a realistic time interval. In our scheme we follow one-hop based direct delivery routing, with central coordination of BS as the content resolver and caching controller.

Li et al. [14] studied the optimal subset selection as a problem of utility function maximization under multiple constraints, but it is used only to offload delay tolerant traffic and also is not network/BS assisted. Andreev et al. proposed a network assisted D2D offloading system model [23], in which all the users in the network (with content) can participate in offloading traffic. Complexity of the scheme [23] was not pushed towards the edge, where a dedicated application server is responsible for resolving every single request that comes to D2D server, which will invoke high signalling overhead. Our proposed scheme is content/information centric and content resolution happens at the BS, therefore complexity is pushed towards the network edge.

At this stage, it is worth pointing out that 3GPP has addressed D2D ProSe communication technology as a viable offloading solution [17]. Also IRTF have recently produced a set of baseline scenario of using ICN in Opportunistic Content Sharing [24]. Certainly our work is first in the literature to naturally combine ICN and D2D technologies for content offloading. Technical challenges of using radio and unlicensed spectrum for D2D communication has been widely analysed in many works [15, 16]. G. Khoshkholgh et al. derived the maximum allowable density of spectrum based D2D users for a given collision probability [25]. Parameters such as QoS/delay, power constraint and interference were taken into account and their results prove that with proper assistance from the cellular network D2D communication can be made a viable approach.

VI. Conclusion

Information centric networking (ICN) has been widely investigated in the research community for efficient content distribution at large scale. In this paper we aim to push the ICN application boundary from network core towards the device level. Our contribution includes a proposed cellular- based network framework with mobile content awareness and intelligence for content resolution to designated helpers in proximity and coordinated content caching/eviction control on the helper side. It is well-known that vanilla in-network caching results in edge caches holding popular content while core caches are cold. The proposed scheme is thus a way of spreading popular content from the edge caches into the wireless cell. We also illustrated that a POI based helper selection scheme is able to potentially lead to a significant proportion of offloaded content requests away from cellular network to helper devices. Our proposed framework for offloading small pieces of social media content was proved to have less overhead ratio, access delay and energy consumption compared to some existing frameworks.

Acknowledgements

This work was supported in part by the EPSRC KCN Project (EP/L026120/1). We also would like to acknowledge the support of the University of Surrey 5GIC (http://www.surrey.ac.uk/5gic) members for this work.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019," February 3, 2015.

[2] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher and B. Ohlman. "A survey of information-centric networking," in Communications Magazine, IEEE 50(7), pp. 26-36. 2012.

[3] B. D. Higgins, J. Flinn, T. J. Giuli, B. Noble, C. Peplin and D. Watson. "Informed mobile prefetching," in Proceedings of Mobile Systems, Applications and Services, 2012.

[4] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch and G. Caire. "FemtoCaching: Wireless video content delivery through distributed caching helpers," in INFOCOM, 2014 Proceedings IEEE, 2012.

[5] B. Han, P. Hui, V. Kumar, M. V. Marathe, J. Shao and A. Srinivasan. "Mobile data offloading through opportunistic communications and social participation," in Mobile Computing, IEEE Transactions On 11(5), pp. 821-834. 2012.

[6] F. Rebecchi, M. Dias de Amorim and V. Conan. "DROid: Adapting to individual mobility pays off in mobile data offloading," in Networking Conference, 2014 IFIP. 2014.

[7] T. Spyropoulos, K. Psounis and C. S. Raghavendra. "Spray and wait: An efficient routing scheme for intermittently connected mobile networks," in Proceedings of the 2005 ACM SIGCOMM Workshop on DTN. 2005.

[8] B. Han, P. Hui, V. Kumar, M. V. Marathe, G. Pei and A. Srinivasan. "Cellular traffic offloading through opportunistic communications: A case study," in Proceedings of the 5th ACM Workshop on Challenged Networks. 2010.

[9] X. Wang, M. Chen, Z. Han, D. Oliver Wu , T. Kwon, "TOSS: Traffic Offloading by Social Network Service-Based Opportunistic Sharing in Mobile Social Networks," in INFOCOM, Proceedings IEEE, 2014.

[10] M.D. Choudhury, H. Sundaram, A. John, D.D. Seligmann, and A. Kelliher, "Birds of a Feather: Does User Homophily Impact Information Diffusion in Social Media?," in CoRR, 2010.

[11] X. Zhuo, W. Gao, G. Cao and Y. Dai. "Win-coupon: An incentive framework for 3G traffic offloading," in Network Protocols (ICNP), 19th IEEE International Conference, 2014.

[12] M. C. Gonzalez, C. A. Hidalgo and A. Barabasi. "Understanding individual human mobility patterns," in Nature 453(7196), pp. 779-782. 2008.

[13] F. Ekman, A. Keranen, J. Karvo and J. Ott. "Working day movement model," in MobilityModels '08: Proceeding of the 1st ACM SIGMOBILE Workshop on Mobility Models. 2008.

[14] Y. Li, G. Su, P. Hui, D. Jin, L. Su and L. Zeng. "Multiple mobile data offloading through delay tolerant networks," in Proceedings of the 6th ACM Workshop on Challenged Networks. 2011.

[15] P. Phunchongharn, E. Hossain and D. I. Kim. "Resource allocation for device-to-device communications underlaying LTE-advanced networks," in Wireless Communications, IEEE 20(4), pp. 91-100. 2013.

[16] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson and Y. Koucheryavy. "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," in Selected Areas in Communications, IEEE Journal On 33(1), pp. 67-80. 2015.

[17] 3GPP, "Feasibility study for Proximity Services (ProSe) and Potential collaboration on Mobile-Edge Computing", 3rd Generation Partnership Project (3GPP), in TR 22.803, 2013 and TDoc ID: S3-151009 respectively.

[18] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen and O. Spatscheck. "A close examination of performance and power characteristics of 4G LTE networks," in Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services. 2012.

[19] C. P. Mayer, "Osm2wkt - OpenStreetMap to WKT Conversion," mayer2010osm, from OpenStreetMaps-ONE, 2010.

[20] A. Keränen, J. Ott and T. Kärkkäinen. "The ONE simulator for DTN protocol evaluation," in SIMUTools, 2009.

[21] A. Abhari and M. Soraya. "Workload generation for youtube," in Multimedia Tools and Applications, 46(1):91-118, 2010.

[22] Intel Corp, "Smart cells revolutionalize service delivery," White Paper, 2013.

[23] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina and Y. Koucheryavy. "Cellular traffic offloading onto network-assisted device-to-device connections," in Communications Magazine, IEEE 52(4), pp. 20-31. 2014.

[24] K. Pentikousis,B. Ohlman, D. Corujo, G. Boggia and et al., "Information-Centric Networking: Baseline Scenarios," RFC 7476 standardisation document, March 2015.

[25] M. G. Khoshkholgh, Yan Zhang, Kwang-Cheng Chen, K. G. Shin and S. Gjessing. "Connectivity of cognitive device-to-device communications underlying cellular networks," in Selected Areas in Communications, IEEE Journal On 33(1), pp. 81-99. 2015.