Coping With Emerging Mobile Social Media Applications Through Dynamic Service Function Chaining

Tarik Taleb, *Senior Member, IEEE*, Adlen Ksentini, *Senior Member, IEEE*, Min Chen, *Senior Member, IEEE*, and Riku Jantti, *Senior Member, IEEE*

***Abstract*—User generated content (UGC)-based applications are gaining lots of popularity among the community of mobile internet users. They are populating video platforms and are shared through different online social services, giving rise to the so-called mobile social media applications. These applications are characterized by communication sessions that frequently and dynamically update content, shared with a potential number of mobile users, sharing the same location or being dispersed over a wide geographical area. Since most of UGC content of mobile social media applications are exchanged through mobile devices, it is expected that along with online social applications, these content will cause severe congestion to mobile networks, impacting both their core and radio access networks. In this paper, we address the challenges introduced by these applications devising a complete framework that 1) identifies such applications/sessions and 2) initiates multicast-based delivery (or offload through WiFi) of the relevant content. The proposed framework leverages the network function virtualization (NFV) paradigm to dynamically integrate its functionalities to the operators' service function chaining (SFC) process, allowing fast deployment and lowering both capital and operational expenditures (CAPEX and OPEX) of the mobile operators. The performance of the proposed framework is evaluated through mathematical analysis and computer simulations, taking Twitter-like social applications as an example.**

***Index Terms*—Social media networking, network function virtualization, service chaining, evolved packet system, EPS, and mobile network.**

I. Introduction

THE EMERGENCE of nearly-ubiquitous mobile data connectivity is revolutionizing the way people live, work, interact, and socialize. Social network applications are in the heart of this social revolution and have been attracting ever-increasing interest from users. Social network platforms (e.g., Twitter and Facebook) or news tickers (e.g. CNN and sport events) are known to be based on a one-to-many communication paradigm, i.e., one entity posts a message of the same content which is then received by many users that have "subscribed" to this "news feed". There are also many other mobile web applications that involve the delivery of the same content to multiple users being in the same location. These applications offer location-based "check in" services. Notable examples are Foursquare (1 million users), Facebook places, Gowalla, Brightkite, Yelp, and Google's Latitude. These applications allow users, particularly mobile users, to check in at locations they visit as a way to find other friends, coordinate gatherings and exchange content of common interest among a "social network" of users. The problem today is that every user establishes a point-to-point communication to the Web server to request the HTML/XML data. While this solution works fine for low-interest information (i.e., where only few users are interested), for high-interest feeds (i.e., information that are "followed" by many users in real-time) this solution introduces a significantly high, and above all, unnecessarily duplicate load on the mobile network, wasting mobile core network resources and resulting in undesirable delays and poor quality of experience (QoE) for users.

In this paper, we propose a complete framework to handle the emerging social media networks and mobile applications behaving in the above-described communication pattern. This framework exploits the Service Function Chaining (SFC) [1] of the mobile network domain, particularly when it is "virtualized" as per the Network Function Virtualization (NFV) paradigm [2], [3]. Hereby, virtualized SFC refers to having one or multiple service functions, "implemented in one or more software instances running on physical or virtual hosts", being applied to traffic flows using routing in a virtual network [4]. The proposed framework consists of two modules. The first one identifies mobile web applications and services that are characterized by a dynamic and frequent transmission/reception of the same content by a group of users in the same neighborhood. The application and service identification can be done by calling the Data Packet Inspection (DPI) mechanism available at SFC and deploying it as a Virtualized Network Function (VNF). The second module uses the data identification to apply adequate policies to the relevant data flows. The applied policies can, for example, enforce a dynamic offload of the relevant

Manuscript received August 9, 2015; revised November 28, 2015; accepted December 2, 2015. Date of publication December 24, 2015; date of current version April 7, 2016. This research work is partially supported by the TAKE 5 project funded by the Finnish Funding Agency for Technology and Innovation (TEKES), a part of the Finnish Ministry of Employment and the Economy. The associate editor coordinating the review of this paper and approving it for publication was M. Li.

T. Taleb is with Sejong University and Aalto University, Espoo FI-00076, Finland (e-mail: talebtarik@ieee.org).

A. Ksentini is with the Department of Mobile Communications, EURECOM, Sophia-Antipolis, France (e-mail: aksentini@ieee.org).

M. Chen is with Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: minchen2012@hust.edu.cn).

R. Jantti is with Aalto University, Espoo 00076, Finland (e-mail: riku.jantti@aalto.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2015.2512274

data to a well-designated access network [5]. VNFs of the offload points/gateways can be dynamically instantiated for this purpose. Other policies could reroute data traffic of target mobile web applications/services through a proxy server that may dynamically establish a multicast group (i.e., based on any available multicasting technology such as Multimedia Broadcast Multicast Service – MBMS [6]) for content that is pushed to many User Equipment (UEs) and may trigger the concerned UEs to join the relevant multicast group using one or more suitable multicast technology [7]. For this aim, we propose adding a new SFC entity, preferably as VNF, which defines relevant policies to apply and handles multicast procedure establishment in the mobile domain.

In this research work, we particularly target online social applications, characterized by the involvement of multiple sessions with frequently and dynamically updated content, shared in a push manner with a potential number of mobile users, sharing the same location as in Foursquare or being dispersed over a wide geographical area as in Twitter. It has been demonstrated by different research work that many mobile social network traffic has such characteristics [8]. Furthermore, in our analysis, we use a model that is based on real Twitter traffic data. Indeed, as shown in Fig. 8, we clearly confirm the findings in [8], which show that social traffic arrivals follow a lognormal distribution. This means that the traffic has a heavy tail but with a peak of connections in the beginning, which is very realistic as users click on the link to an information that got just posted. However, the click number decreases not instantly but taking a long period before reaching zero, which is well captured with the heavy tail of the lognormal distribution.

The remainder of this paper is organized as follows. Section II describes some related work on social-aware network optimizations and portrays a general SFC architecture. Section III details the proposed framework functionalities and architecture. An analytical model based on Markov chains is presented in Section IV. Simulation results are presented in Section V. Section VI concludes this paper with a summary recapping the main advantages of the proposed framework.

II. Related Work

A. Social-Aware Network Optimizations

Several studies have been conducted to identify the traffic generated by social network applications and its relation with users' behaviors. In [9], the authors concentrate on Twitter platform in order to classify users and to identify their behaviors and their geographic growth patterns. In [10], an empirical model is used to study users' behaviors and traffic patterns in social networking services. A special focus was on validating the Zipf law assumption regarding the content popularity in Youtube and Twitter. Another important aspect of online social networks is the content spread among users according to their social activities. In [11], the authors have studied the impact of users' re-tweets on information diffusion in Twitter. An epidemic model was used in [12] to investigate information propagation in social connections. Based on the analysis done on traffic patterns and content propagation in social networks, several research works have been conducted to optimize network resources, particularly for content delivery focusing on social-based data replication and caching. Research works cited herein mainly focus on analyzing the behavior of social application traffic and how to model it. The proposed framework uses the outputs of these research works as a basis for analyzing the performance of the proposed solutions.

Indeed, as in [13], replicating videos to different geographic regions is an interesting solution to increase user's QoE in social video services. In [14], the authors proposed a social partition and replication middleware in order that data of users' friends are collocated in the same server. In [15], a social-media partition was proposed to balance social load among servers. In [16], the authors proposed a social-aware content replication strategy using a hybrid edge-cloud and a peer-assisted architecture. The proposed replication strategy is based on three replication indices, namely geographic influence index, content propagation index – to indicate the way to cache video content on the edge cloud, and social influence index – to indicate for peers which videos to cache for their friends. Based on the fact that social video services are coming from microblog recommendations, another work in [17] proposed a proactive service deployment of a video sharing system. Based on microblog advertisement, the authors predict the upcoming video demand and proactively react by pushing content of interest nearby corresponding users. Most of the mentioned social-oriented network optimization solutions are dedicated to content delivery. However, mobile networks are highly affected by social network traffic as an important portion of UGC is uploaded from mobile devices. Besides considering caching as a solution, the proposed framework considers also the use of multicast and data offload techniques to further optimize the use of mobile network resources.

To cope with mobile user generated content, some researches have considered the concept of Delay Tolerant Networks (DTN) for uploading user's content to mobile networks, designing different delay tolerant forwarding and data transferring algorithms [18]. For example, in [19], it was observed that mobile user generated content delivery is a user-behavioral problem, as most content uploads occur at small number of locations (e.g., users' home or work locations) with significant lag between the content generation time and the content upload time. Based on these observations, it was proposed that mobile user generated content uploads shall happen at selective locations, called drop zones, that are intelligently placed across the cellular network taking into account deployment cost and daily movement patterns of a large number of mobile users. In [20], usage patterns of mobile data users in large 3G cellular networks were characterized. It was found that most of the mobile users access mobile data services occasionally, whereas only a few of heavy users contribute to a majority of data usage in cellular networks, that is due to usage of a small number of data-intensive mobile applications, video browsing and streaming, and popular social media sites. Similar to these research works (i.e. [18], [19] and [20]), the objective of the proposed framework is to locally handle the social application traffic. However, in addition to data offload, we propose the use of multicast communications to locally distribute the shared content among users.

In [21], flow-level dynamics of cellular traffic are studied, proposing a ZIPF-like model and a Markov model to capture

the volume distributions of application traffic and the volume dynamics of aggregate Internet traffic, respectively. In [22], the authors proposed a method for uplink distribution of live video content by considering the popularity of the content, the video characteristics and the available resources. In the proposed solution, users are connected to a video portal, which gathers the videos generated by users. The video portal is responsible of ranking these videos according to their popularity and shares this list with a central entity (evolved NodeB – eNB). The latter schedules and allocates radio resources among the users. Focus on the impact of the over-the top video on cellular networks is provided in [23]. Unlike the above mentioned research works (i.e. [21],[22]), we propose a framework of optimization solutions at the application layer, which rely on the link layer optimization techniques such as multicast. Indeed, the multicast solution is established at higher layers but strongly depends on the multicast capability of eNBs.

In [24], the characteristics of cellular HTTP-based traffic are analyzed with respect to a group of applications, namely those related to social, news, and video (e.g., Flickr, Google Videos, WorldPress, YouTube, and Blogsport). Many observations were made about the size of wireless sessions, the number of flows per wire-less sessions, the packet size used in wireless sessions, and the temporal distribution of demands for mobile services, in comparison to wireline networks. An important observation pertains to the fact that inter-packet gaps differ significantly among different service types, suggesting advanced optimizations such as application-oriented handling of bearer and terminal states, which is in line with the objective of this paper.

B. Service Function Chaining

SFC is not a novel concept. It has been deployed by mobile operators as well as fixed network operators for many years. It simply consists of a set of network services, such as Deep Packet Inspection (DPI), firewall, Intrusion Detection System (IDS), and Network Address Translation (NAT), which are interconnected through networks. In case of mobile networks, SFC is located between the Packet Data Network Gateway (P-GW) and the Packet Data Network (PDN) (e.g., Internet) in the so-called SGi Local Area Network (SGi LAN). Fig. 1 shows the current Long Term Evolution (LTE) architecture, namely the Evolved Packet System (EPS), including the SGi LAN. SFC is used to control and manage traffic coming from and going into mobile networks. It is used to enforce operators' policies to optimize mobile traffic. For instance, an email service chain would include virus, spam and phishing detection and would be routed through connections ensuring no excessive delay. Web traffic would be routed through a chain that includes virus scanning and a Transmission Control Protocol (TCP) optimizer. The chain created for video and voice traffic would include traffic shaping so that traffic would be routed over links with the level of delay and jitter guarantees ensured for each customer. For video traffic, a chain would include video transcoding system to adapt the video stream to the user context (e.g., screen, size, and CPU).

The main weakness of the above mentioned architecture is the difficulty to build a scalable and flexible SFC. Indeed,

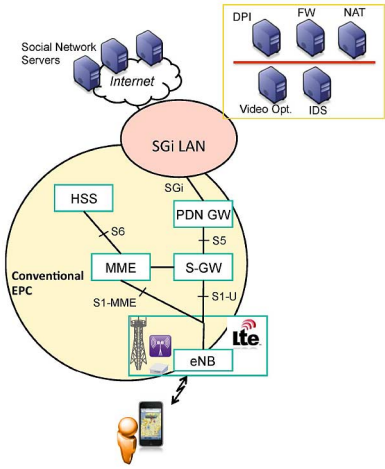


Fig. 1. Conventional SFC architecture.

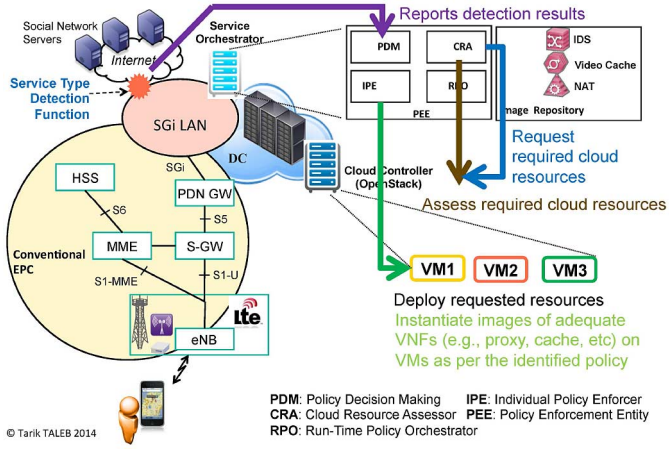


Fig. 2. Envisioned NFV-based SFC architecture.

deploying a service chain to support a new application requires time and effort. Each service requires a special hardware device, and each device has to be individually configured with its own command syntax. Recent trends in Software Defined Networking (SDN) and NFV open up new ways to deploy SFC in a more efficient and scalable fashion; above all, on-demand [1]. In [30], IETF has already started investigating new solutions for virtualizing SFC based on the concept of NFV. With this regard, it is worth stating that the SGi LAN architecture (depicted in Fig. 2) we envision in this paper is based on the SFC architecture, built on NFV, as defined in [30].

III. Proposed Framework

As stated earlier, the proposed framework has two main design goals; namely identifying traffic from mobile applications with frequently updated content sent to many users and enforcing an adequate policy to cope with the congestion that may be caused by the identified applications.

A. Social Traffic Identification Function

In the envisioned SFC-based architecture (Fig. 2), data identification takes place at the DPI NFV instance. This could be

achieved by adding a new function dedicated to detect and identify social network traffic, namely Social Traffic Detection Function (STDF). Indeed, upon a trigger, UEs issue a HTTP GET request to get a content common to all of them. These HTTP GET requests are intercepted and analyzed by the DPI function. Note that data identification and traffic detection could be carried out at any point on the path between UEs and the application server. For instance, it could be carried at a function collocated with a data anchor gateway (e.g., P-GW in case of EPS). If these HTTP GET requests are issued at a frequency higher than a predetermined threshold (i.e., inter-arrival time between two consecutive HTTP GET requests from the same UE is shorter than a certain threshold, and/or the number of HTTP GET requests from the same UE issued during a particular time interval exceeds a certain threshold), STDF qualifies/identifies the application relevant to the HTTP GET requests as "an application with frequently and dynamically requested content". Alternatively, STDF can also monitor the traffic sent to UEs and identify sessions that are delivered to many UEs and send frequent content updates, based on configurable thresholds.

Since STDF identifies requests/sessions that lead to the delivery of frequent content updates to many UEs (i.e., to more than a predetermined number of UEs), STDF informs and forwards the traffic to the Policy Enforcement Entity (PEE), residing in the SGi-LAN. A high-level diagram architecture of PEE is shown in Fig. 2. Shown are also the interactions of PEE with STDF and the Cloud Controller that is in charge of instantiating required resources/Virtual Machines (VMs) on the virtual infrastructure platform of the SGi LAN, using a suitable cloud management tool (e.g., OpenStack). PEE principally consists of four units, namely Policy Decision Making (PDM), Cloud Resource Assessor (CRA), Individual Policy Enforcer (IPE), and Run-Time Policy Orchestrator (RPO). Upon detection of an application with frequently and dynamically requested content, SDTF reports this event to the PDM unit. Depending on the characteristics of the identified application, PDM decides an adequate policy with regard to the relevant data traffic, such as offloading the relevant data traffic, rerouting the relevant data traffic through a proxy server, which dynamically establishes a multicast group if that content is sent to many UEs, or requesting the concerned UEs to join a relevant multicast group. The identified policy is then communicated to CRA that assesses the required cloud resources to enforce it on the traffic of the relevant application. For instance, in case PDM decides to multicast the content of the application to its respective users, a proxy or multiple proxies to fully or partially cache the content of the application, a MBMS gateway, and/or a Broadcast Multicast Service Centre (BM-SC) may become required. Resources for instantiating images of the virtualized network functions of these elements become therefore required. Once the needed resources are identified, they are communicated to the cloud controller that deploys them, e.g., using OpenStack. IPE then instantiates images of adequate VNFs (e.g., proxy, cache, etc) on deployed VMs as per the identified policy. RPO then orchestrates the underlying policy during its run-time and per changes in the application detection and based on internal as well as external triggers (e.g., cloud resource monitoring, service level agreement controller, etc) [2]. Exploiting cloud computing technologies, a policy orchestration could indicate the turning on of a new proxy to scale up, turning off another to scale down, replacing VNF running on a VM with another one as per changes in the social traffic load and the behavior of its users. With its four units, PEE follows a common lifecycle policy management model whereby policy design is conducted by PDM and CRA, policy implementation and deployment are carried out by IPE, and policy provisioning, runtime and operation, and disposal are conducted by RPO. With the flexibility that cloud computing and the PEE architecture offer, a network operator may instantiate VMs and run on them suitable VNFs to specifically handle traffic of a particular social network application.

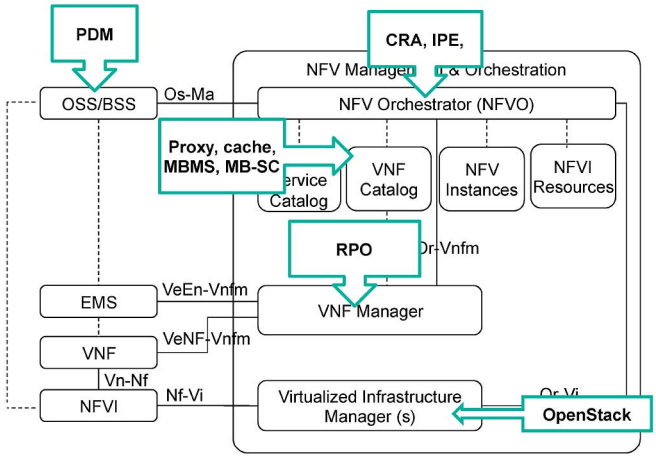


Fig. 3. Possible integration of the proposed framework within the reference ETSI NFV architecture.

Fig. 3 shows how the proposed framework can be integrated within the ETSI NFV reference architecture [31]. In the envisioned architecture, the VNF manager maps unto RPO. Indeed, RPO is responsible of the lifecycle management of VNF instances and the interaction with Element Management System (EMS) provided by the Cloud Controller, which allows to turn on/off VNF instances to scale up or down; e.g., adding new instances of proxy or MBMS or deleting existing instances. The PDM functional block can be part of the Operations Support System (OSS) and Business Support System (BSS) functions. Indeed, all policies that react to the detection of an application with frequently and dynamically requested content are implemented at this functional block. The CRA and IPE functions are part of the NFV Orchestrator. According to the ETSI definition, the NFVO is in charge of the orchestration of NFVI resources across multiple VIM and lifecycle management of Network services, which corresponds to the functions carried out by the CRA and IPE. Finally, the cloud controller functions map unto a VIM.

It shall be noted that the above mentioned thresholds can be dynamically updated; depending on the time of the day and the location. The applied policy could also depend on these thresholds. For instance, if the frequency of the HTTP GET requests exceeds a certain value (Val_1), the operator may offload the relevant data traffic to WiFi. If the frequency of the HTTP GET

requests is within the range of $[Val_2; Val_1]$, UEs may become requested to join an adequate multicast group. If the frequency of the HTTP GET requests is lower than (Val_2), the operator and UEs are requested to do nothing.

B. Multicast Delivery

Whilst live content and IPTV services are the best "traditional" use cases for multicast, in this section, we will show how multicast can be used to mitigate the issues raised by the services targeted in this research work: services whereby one entity posts a message and then that message is delivered (either in a push or pull mode) to many users. As stated earlier, in addition to news tickets, many emerging social networks exhibit this characteristic, and that include location-based check-in services (Foursquare, Facebook Places, Gowalla, etc) and Twitter. For these Over The Top (OTT) services that involve simultaneous (or even near-simultaneous) delivery of the same content to potential number of users, it is very trivial that multicast could be of potential use to reduce the redundancy of content over the network and therefore ensure efficient usage of network resources.

Hereunder, we show how once a mobile service is identified as an application with frequently and dynamically updated content (i.e., by STDF), its subsequent relevant data traffic is sent via multicast. To this end, the core idea is to extend the proxies (i.e., VNFs of web proxies) hosted in the SFC pool and clients (e.g., browsers) with a functionality (e.g., through a plug-in) that enables efficient delivery of the same Web content, requested by many users, by instantiated VNFs of Web proxies using 3GPP multicasting technologies (i.e., MBMS [6]) and seamless integration/embedding of the multicast content into normal Web pages/services by the clients (i.e., Web browsers). A VNF of web proxy enhanced with this functionality would be either statically configured or would dynamically decide (e.g., upon receiving many requests for the same content feed) to allocate a multicast address to this content feed and then start multicasting the content (i.e., text and images) using HTML/XML encoded via UDP – User Data Protocol (or an alternative multicast transport protocol). The VNF of the web proxy would respond to any HTTP request for that content with a well-defined Content Type and the Multicast Address, which would, upon arrival at the Web browser, activate the Multicast Browser functionality (or launch the respective browser plug-in). The multicast plug-in would then join the respective Multicast Address/Group and start listening for content message. The content messages received via the multicast channel would then be rendered according to the XML/HTML format in the browser window. To interact with the MBMS system, the VNF of the web proxy may also incorporate some functions of the Broadcast Multicast Service Centre (BM-SC). Fig. 4 depicts the overall architecture of the envisioned solution (omitting PEE and Cloud Controller of Fig. 2). For the Web client/browser (i.e., at UEs) to support multicasting, it is enhanced by an internal cache where it stores any content that is received via the multicasting channel. Prior to requesting any missing content objects from the server – as it would do normally – the client/browser checks if the content object

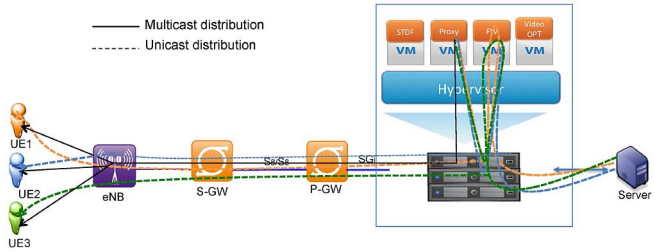


Fig. 4. Envisioned solution, illustrated for 3GPP's Evolved Packet System [6]; content in common with all relevant UEs is multicast from a VNF of a web proxy on top of UDP and received by a suitable plugin at UEs.

has already been received via the multicasting channel. If so, and the cached content is still topical, it will omit the HTTP request to the server and uses the cached content. This allows the Web server/proxy's VNF to push embedded content objects (e.g., images or other content types) in addition to normally configured XML/HTML content. If the content is not in the local cache, the Web client/browser could request the content as usual. To avoid that the client/browser requests any missing content based on a regular HTTP request, the server/proxy's VNF needs to ensure that all embedded contents are delivered prior to the XML/HTML of the page. Alternatively, to relax this requirement, the client/browser could also wait for a configurable timeout for content objects that are related to previously "multicasted" content, and only if the content does not arrive via the multicast channel within the specified timeout, it would place the normal HTTP request. The Content-Feed can be either declared, using adequate XML/HTML tags, "replacing" or "additive". In the former case, a multicast Content-Feed update will replace the previously received content, while in the latter case, the update will be "added" at the end of the previously delivered content. It shall be stated that since in the envisioned framework, the multicast communication occurs within the mobile network domain, other 3GPP schemes can be used to deal more rigorously with packet losses. In the context of MBMS, [34] introduces a number of mechanisms for packet loss recovery. In 3GPP mobile networks, packet losses can be also mitigated through the use of Forward Error Correction (FEC) mechanisms at the physical layer. At the link layer, multicast transmissions can be handled in a different way by the radio access network. Indeed, feedbacks on Channel Quality Indicators (CQI) from each UE belonging to a multicast group can be used to determine the corresponding CQI level. Accordingly, data is sent to all or a subset of UEs. The node with the lowest CQI level limits the data rate for all other nodes, as all available resources (RBs) are allocated to the single activated CQI level; this procedure may highly mitigate the impact of packet losses on multicast communication.

The services that UEs are receiving may have content that is common to the UEs and other part of the content that is specific to each UE. The content that is not common is delivered in unicast over TCP – as usual (Fig. 4). The proposed solution is only relevant to the common content, which is delivered via multicast over UDP. At the UE side, there is an application layer logic that integrates the two portions of the content; the one received in unicast and the one received in multicast.

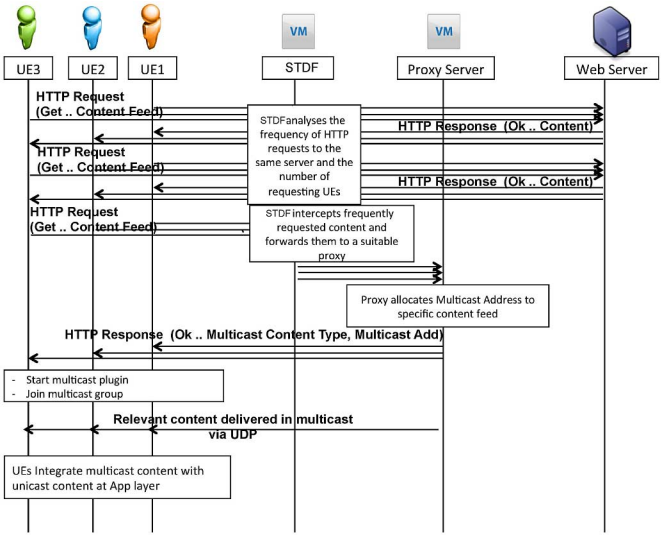


Fig. 5. Detailed message flow of Web-based multicast content delivery in the proposed framework.

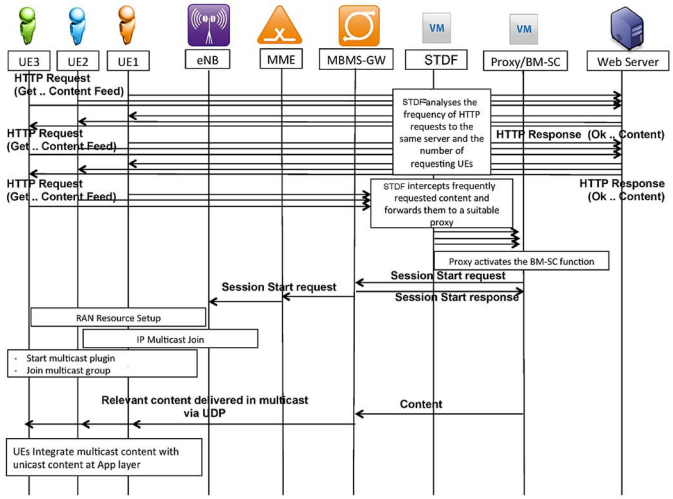


Fig. 6. Detailed message flow of MBMS-based multicast content delivery in the proposed framework.

Fig. 5 shows an example implementation scenario of the proposed solution whereby the VNF of the proxy is located in the SGi pool of the mobile operator and does not run the mobile network's multicast plugin presented here. In the envisioned implementation scenario, a number of UEs send HTTP request messages asking for a particular content from the same server. STDF initially analyses the frequency of these HTTP requests and the number of requesting UEs. If they satisfy particular conditions as explained above, STDF intercepts new HTTP requests to the same server and forwards them to the VNF of the proxy hosted in the SGi pool (e.g., in case the Web server is not owned by the mobile network or the application server does not run the mobile network's plugin for multicasting common content). The VNF of the proxy then allocates a multicast address to the specific content feed, and sends back a HTTP response to the relevant UEs indicating the Multicast Content Type and the Multicast Group Address. Upon receiving these HTTP responses from the VNF of the proxy, UEs launch their multicast plugin and join the multicast group. In the envisioned scenario, the service that the UEs are receiving may have content that is common to the UEs and other part of the content that is specific to each UE. The uncommon content is delivered in unicast using TCP from the server or the proxy server. The content in common is then delivered in multicast using UDP to the UEs. The content could be either received by a VNF of the proxy server caching the information locally, or a VNF of the proxy server could request it from the web server and immediately relay it to the UEs in multicast. The UEs integrate the content delivered in multicast and the content delivered in unicast at the application layer.

To further optimize the transmission over the mobile network, the mobile operator GW (e.g. P-GW / GGSN - Gateway GPRS (General Packet Radio Service) Service Node) and/or VNF of a proxy could also trigger the establishment of a MBMS session over which the common content can be transmitted efficiently. Fig. 6 shows this case, where two new 3GPP entities are involved to handle multicast in the mobile network, namely BM-SC and the MBMS gateway. Similar in spirit to the precedent scenario, STDF detects an application with common content and informs the VNF of the proxy. In this scenario, we assume that the proxy VNF also implements BM-SC functions; therefore the proxy VNF sends the session start request to the MBMS-GW in order to create the multicast group and notifies the Mobility Management Entity (MME), eNB and UE about the IP address of the multicast group and mobile network-related parameters, such as Flow Identifier, Quality of Service (QoS), MBMS service area, and C-TEID (Common Tunnel End Identifier) for bearer establishment. Since the eNB is aware of the created group, it reserves Radio Access Network (RAN) resources to be shared by the group of UEs participating to the multicast communication, and notifies the implicated UEs about these parameters. Then, the UE joins the multicast group, and the multicast delivery process begins. Note that the content is first sent from the proxy VNF/BM-SC to the MBMS GW, which then forwards it, in a multicast manner, to the mobile network domain.It is worth mentioning that the multicast communications used by the proposed solution is done at the mobile network operator domain, which means that all communication are one hop-based (similar to the Internet Group Management Protocol - IGMP). Therefore, to be part of a multicast group, a user equipment needs only to accept packets destined to the multicast group address [6], [32]. So unlike traditional multicast routing algorithms such as Multicast Open Short Path First (MOSPF) or Protocol Independent Multicast (PIM), there is no need to create a multicast tree. For more details on the multicast delivery architecture in 3GPP, interested readers may refer to [6], [32].

IV. Analytical Model

In this section, we present an analytical model for the proposed framework. The aim of this model is to investigate the impact of the thresholds used by STDF to detect a social network traffic and to subsequently establish an associated multicast channel. Let $X(t)$ denote the number of active users

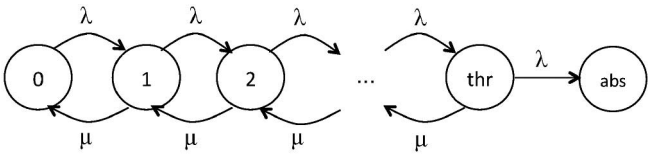


Fig. 7. Semi-Markov Chain representing the proposed framework.

having clicked on a link shared by a member of their social group at time instant t. These users stay connected until they complete the download of the shared object. Here, we assume that after the detection of a social group by STDF, a certain number of users remain connected. We assume that this number exceeds a specific threshold, denoted as thr. We also assume that when the communication switches in multicast mode, the system becomes absorbed in one state, representing multicast communication. Assuming that the time duration for downloading the shared object follows an exponential distribution μ, the stochastic process $X(t)$ becomes then a semi-Markov chain as the inter-arrival times, denoted by λ, is not exponentially distributed, and the system evolution depends only on the current state. Fig. 7 shows the Semi-Markov chain representing the proposed framework for one social group.

Indeed, it is generally agreed that the inter-arrival times distribution of social applications exhibits a long tail, which is well captured by a lognormal distribution. For instance, research work conducted in [26] and [8] indicate that the inter-arrivals in Orkut as well as Social Mobile Instant Messaging follow a lognormal distribution. To further confirm this observation, we use the tool in [27] to simulate a twitter traffic with the following features:

- Twitter followers: 85000
- Fraction of followers who are watching their twitter feeds: 10%
- Initial fraction of watchers who click: 45%
- Background clicks: 15%
- Change in number of Tweet per minute: 25

The obtained results are shown in Fig. 8, which also plots a lognormal fit distribution. From the figure, it becomes apparent that the inter-arrivals of Twitter users also follow a lognormal distribution with parameters $\sigma = 2.5$ and $\mu = 1.6$, confirming the findings of [26] and [8] also for Twitter. Hereunder, we will consider this Twitter model as a basis for the inter-arrival times of users in Fig. 7.

To transform the above-mentioned semi-Markov model to a Markov model, we propose replacing the lognormal distribution, representing the user inter-arrivals, by a Phase Type distribution characterized by the same mean and variance. By doing so, we can resolve the Markov chain and derive the performance of the proposed framework for different values of the thresholds. The Phase Type distribution is widely used to approximate an arbitrary continuous distribution (with $x > 0$) with a sequence of "Phase-type" distributions, which results in a generalized Erlang distribution. There are different approaches to approximate an arbitrary distribution with a Phase Type distribution. Notable examples are the method of Moment, the method of Maximum Likehood, and the method of Maximum Entropy. In this work, we use the Maximum Likehood method, which is widely used in the literature. The

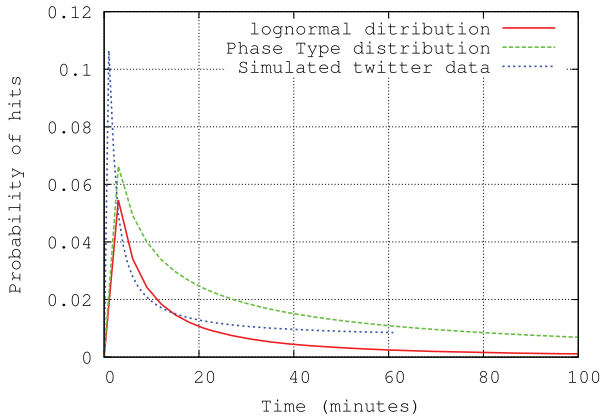


Fig. 8. Twitter real data in comparison to Lognormal Fit and Phase Type distributions.

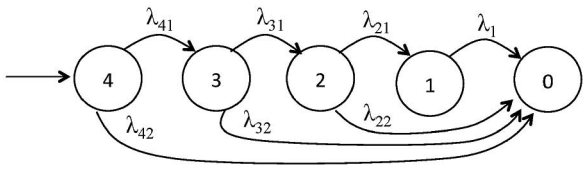


Fig. 9. Phase Type distribution.

obtained Phase Type distribution representation is illustrated in Fig. 9. The PDF of the Phase Type distribution is plotted in Fig. 8 and is compared to the real data and the lognormal distribution. It becomes apparent that the Phase Type distribution can approximate the real data as well as the lognormal distribution. Note that the Phase Type distribution, illustrated in Fig. 9, is defined by its transition rate matrix Q_T with state space {0, 1, 2, 3, 4} where 0 is the absorbing state. Q_T is given as follows:

$$Q_T = \begin{bmatrix} 0 & 0 \\ t & T \end{bmatrix}$$

where T is a defective transition matrix of a continuous time Markov chain with finite space $\{1, 2, 3, 4\}$. That is, T has non-negative off-diagonal entries and negative diagonal elements such that $t = -T1 \geq 0$ but $t \neq 0$ (1 is the vector of ones and 0 is the null vector). By denoting the initial distribution of this Markov chain by α, the distribution phase type is denoted by PH(α,T), where $\alpha = (0, 0, 0, 1)$. When the Markov chain reaches State 0, it starts over again with the same initial distribution α. Therefore, a renewal process, which counts the absorbing times, is defined and is adopted as the arrival process (user arrival) in the proposed model. The stochastic process $(X(t), J(t))$ represents the new Markov chain, where $X(t)$ is the number of users downloading the shared object and $J(t)$ is the current phase of the renewal process at time instant t. Fig. 10 shows the new Continuous Time Markov Chain. The system evolves since the shared object is sent using unicast. Once STDF detects the social group (i.e. exceeding the predefined threshold), the system becomes absorbed in the state $X(t) = abs$. In this model, we are interested in the number of objects sent in a unicast manner (i.e. before the system gets absorbed in State abs). This metric allows us to see the impact of the thresholds on the proposed framework, as the goal

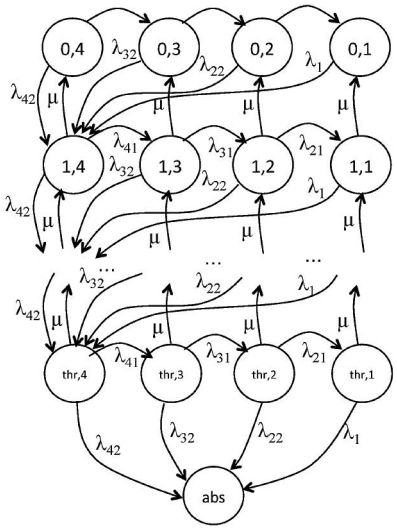


Fig. 10. Continuous Time Markov Chain.

is to minimize the number of downloaded objects in a unicast fashion. To do so, we need to compute the proportion of time the Markov chain spent in each state before absorption. Let the state space $S = A \cup N$ be partitioned into the set $A = \{abs\}$ of absorbing states and the set N for non-absorbing states. The time spent before absorption is obtained by considering the $lim_{t\to\infty} L_N(t)$ restricted to the set N. To compute $L(\infty)$, the infinitesimal generator matrix and the initial probability vector are restricted to the states of set N, and are denoted by Q_N and $\pi_N(0)$. It is worth noting that Q_N is not an infinitesimal generator. By counting each state (in Fig. 10) from right to left, the Q_N format can be expressed as follows:

$$Q_T = \begin{bmatrix} B_0 & A & 0 & 0 & \dots & 0 \\ B_1 & B_2 & A & 0 & \dots & 0 \\ 0 & B_1 & B_2 & A & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ \vdots & \vdots & \vdots & \vdots & B_1 & B_2 \end{bmatrix}$$

where:

$$B_0 = \begin{pmatrix} -\lambda_1 & 0 & 0 & 0 \\ \lambda_{21} & -(\lambda_{21}+\lambda_{22}) & 0 & 0 \\ 0 & \lambda_{31} & -(\lambda_{31}+\lambda_{32}) & 0 \\ 0 & 0 & \lambda_{41} & -(\lambda_{41}+\lambda_{42}) \end{pmatrix}$$

$$B_1 = \mu(4,4)$$

$$B_2 = \begin{pmatrix} -\lambda_1-\mu & 0 & 0 & 0 \\ \lambda_{21} & -(\lambda_{21}+\lambda_{22}+\mu) & 0 & 0 \\ 0 & \lambda_{31} & -(\lambda_{31}+\lambda_{32}+\mu) & 0 \\ 0 & 0 & \lambda_{41} & -(\lambda_{41}+\lambda_{42}+\mu) \end{pmatrix}$$

$$A = \begin{bmatrix} 0 & 0 & 0 & \lambda_1 \\ 0 & 0 & 0 & \lambda_{21} \\ 0 & 0 & 0 & \lambda_{31} \\ 0 & 0 & 0 & \lambda_{41} \end{bmatrix}$$

Since Q_N is a square matrix ($4thr$,$4thr$) and has a regular structure, $L_N(\infty)$ can be obtained by the linear equation:

$$L_N(\infty) Q_N = -\pi_N(0)$$

Having the L_i values, the Mean Time To Absorption (MTTA) is obtained by:

$$MTTA = \sum_{i \in N} L_i(\infty)$$

where $\pi_N(0) = (0,0,0,1,0,\dots,0)$. Finally, the expected number of unicast messages (objects) downloaded in a unicast manner is obtained as follows:

$$E[msg] = \sum_{i}^{t} \left(\sum_{j=i*4}^{(i+1)*4} \frac{L_i(\infty) * i}{MTTA} \right)$$

V. Performance Evaluation

Having described in details the proposed framework, we focus, in this section, on its evaluation. The evaluation results were obtained using both the analytical model presented in Section IV and computer simulations. As stated earlier, the analytical model is used to evaluate the responsiveness of STDF to detect social network groups for different traffic load configurations. The analytical model aims for: (i) evaluating the impact of the threshold in terms of the number of users and (ii) showing the impact of the traffic intensity. The simulation model complements the analytical model taking in consideration real social network traffic (i.e., twitter) traces. It aims for (i) evaluating the impact of the threshold in terms of time duration to detect the social traffic and (ii) for evaluating the impact of user popularity on the proposed framework. Accordingly, the analytical model results are not compared against the simulation results; rather the two results are complementing each other. It is worth noting that the simulation model implements only the PDM module. Therefore, we can simulate only the impact of the PDM policies, i.e. either offload through WiFi or the use of multicast communications, on the overall performance of the proposed framework.

A. Numerical Results

In order to investigate the performance of the proposed framework under different traffic loads, we consider $\rho = \frac{\lambda}{\mu}$ as an indicator of the load. Three cases are considered: (i) low traffic load ($\rho = 0.5$); (ii) medium traffic load ($\rho = 1$); and (iii) high traffic load ($\rho = 2$). Note that λ is the mean inter-arrival time obtained from the Phase Type distribution as follows:

$$\lambda = -\alpha Q_T^{-1}$$

In the model presented in Fig. 8, $\lambda = 48$ and the standard deviation is 111.

Fig. 11 shows the expected number of unicast messages for different threshold values and different traffic loads. In addition

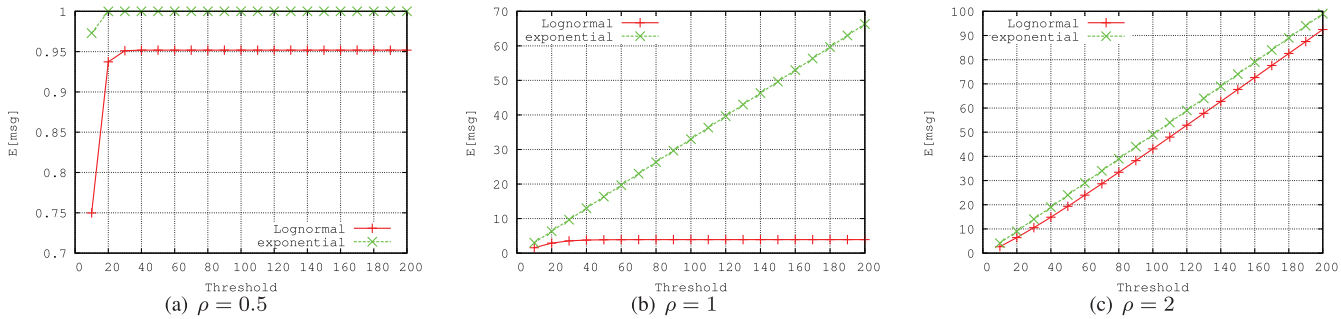


Fig. 11. Expected unicast messages.

to the results obtained using a lognormal distribution for the inter-arrival times, we also include results obtained using an exponential distribution. As a comparison term, for low traffic load (for instance the case of social network applications with low interest), the threshold has no impact on the performance of the proposed framework as it (and hence the absorbing state) is rarely reached. The gain in this case is minimal. We also notice that no major difference is seen between using exponential and lognormal distributions, which is logical as the traffic is low. In case of medium traffic, we notice an important difference between using exponential and lognormal distributions. Exponential distribution is overestimating the incoming traffic, which heavily impacts the performance results. We remark that in case of lognormal traffic, the threshold is rarely reached in comparison to the exponential case, and again the threshold value does not have impact on the proposed framework. The most interesting case is when the traffic load is high (highly popular traffic). In this case, both exponential and lognormal distributions exhibit the same performance. In this case, the threshold is reached each time, and the impact of the threshold becomes remarkable. Indeed, we notice that the number of messages is convex to the threshold. We argue this by the fact that when the traffic load is high, multicast usage allows considerable reduction in the number of exchanged messages. As it will be confirmed by the simulation results, the proposed framework is more efficient when the traffic load is high. When the traffic load is low, the gain beneath using multicast communication for few messages is negligible.

B. Simulation Results

To evaluate the performance of the proposed framework, we developed a C-based events simulator. This simulator mimics the behavior of a simple LTE network (with multicast capability) and simulates social traffic following the same principle of Twitter. That is, a group of users, noted as followers, hit an embedded link in a tweet according to a certain probability (p). To be realistic, we assume that the number of followers for a tweet is dynamic and is randomly selected from within the range of [100, 200]. We consider the same concept as detailed in [25], wherein the number of hits decreases with time according to the probability (p), meaning that a tweet will not interest users after the elapse of some time (e.g., in the order of minutes). Two scenarios were considered. The first one assumes that a group of users are located in the same region and are

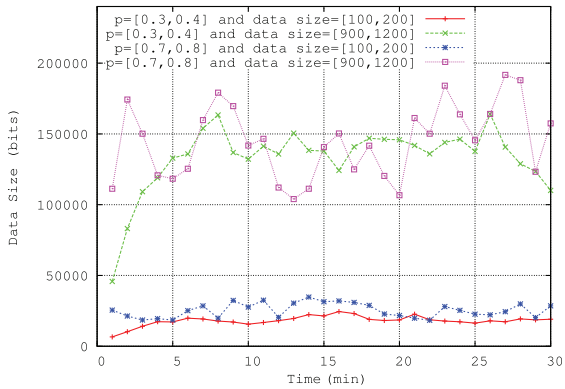
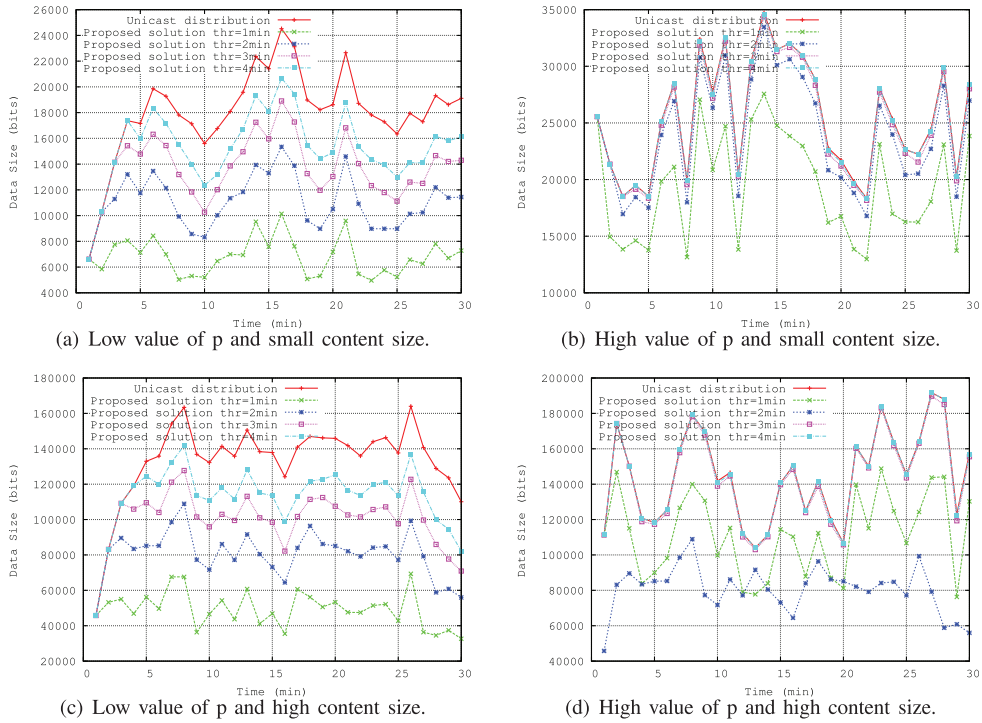


Fig. 12. The amount of traffic exchanged between the social group.

connected to the mobile network through the same eNB. Each minute, a user tweets a message, and according to the probability (p), the followers hit the embedded link in the tweet and download the content. The simulation duration is 30 min. The average runs of each simulation is 100 times. The envisioned scenario represents the case that a group of users actively communicates and shares the same content among them. In the second scenario, we assume that the probability of hits depends on the popularity of the user who tweets a message. For this aim, we assume that the probability that a user hits the link to the message depends on the user popularity and follows a Zipf law with s = 0.56 [8]. In this scenario, we increase the size of the social network to 10000 subscribers. This reflects the case of a social network of subscribers residing in the same city and connecting to the same mobile network through different eNBs.

For each scenario, we changed the size of the exchanged content between users: (i) to simulate a tweet with a video link with a content size randomly drawn from within [900, 1200] kbits; (ii) to simulate a tweet with a low quality video link with a content size selected randomly from within [200, 300] kbits. For both scenarios, a static method based on varying thresholds is used for detecting social media applications. Once an application is detected as a frequently requested social media application involving the delivery of the same content among many users, its traffic data start being multicast (or offloaded through WiFi) rather than being delivered in unicast.

1) Simulation Model and Scenarios: Fig. 12 plots the amount of traffic exchanged between the social group in the first scenario. This figure shows the traffic for four cases: (i) the probability that a follower hits the link is high (randomly



(a) Low value of p and small content size.

(b) High value of p and small content size.

(c) Low value of p and high content size.

(d) High value of p and high content size.

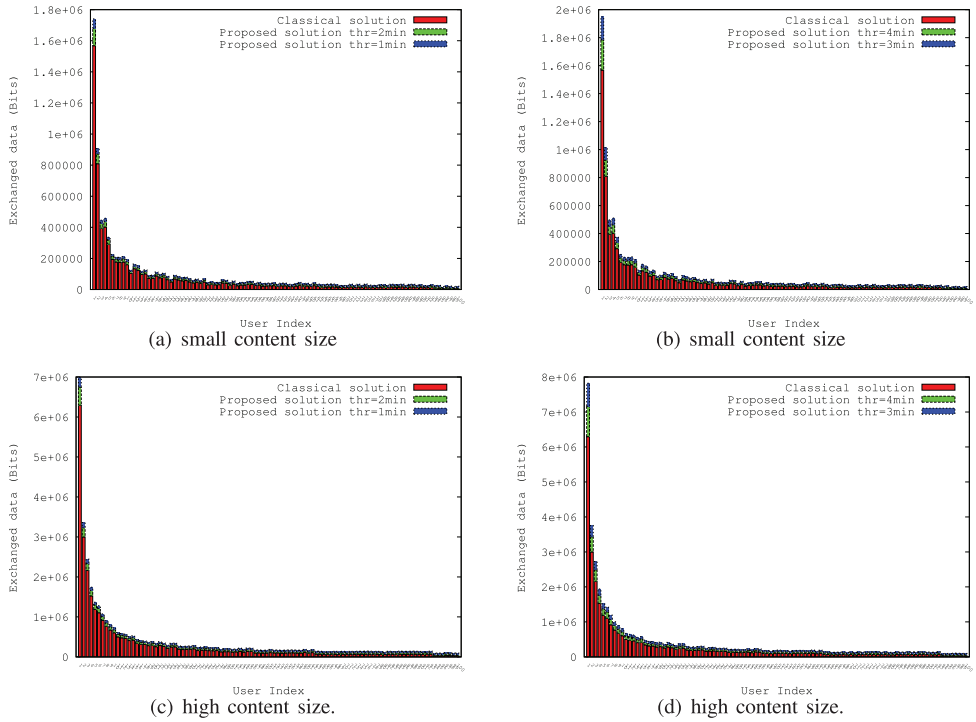
Fig. 13. Total amount of data exchanged: proposed solution vs unicast distribution.

TABLE I
OFFLOAD CASE

		thr=1min	thr=2min	thr=3min	thr=4min
Low value of p	small content size	535.5 (Kbits)	525.2 (Kbits)	511 (Kbits)	493.7 (Kbits)
	high content size	3.948 (Mbits)	3.864 (Mbits)	3.755 (Mbits)	3.636 (Mbits)
High value of p	small content size	737.2 (Kbits)	715.8 (Kbits)	697.3 (Kbits)	677.8 (Kbits)
	high content size	4.227 (Mbits)	4.052 (Mbits)	3.902 (Mbits)	3.781 (Mbits)

selected from within [0.7,0.8]) and the size of the exchanged content is also high (randomly selected from within [900-1200] kbits); (ii) the probability that a follower hits the link is low ([0.3,0.4]) and the size of the exchanged content is large ([900-1200] kbits); (iii) the probability that a follower hits the link is high ([0.7,0.8]) and the size of the exchanged content is small ([200-300] kbits); and finally the case whereby (iv) the probability that a follower hits the link is low ([0.3,0.4]) and the size of the exchanged content is also small ([200-300] kbits). It is obvious that the higher the size of the exchanged content and the higher the probability to hit a link, the higher the amount of traffic exchanged among the users.

2) Results: Based on the traffic model presented in Fig. 12, Fig. 13 compares the proposed solution, when implemented with different thresholds for detecting social media applications, against the classical unicast distribution, and that is for the four simulated cases. Unlike the analytical model, wherein the threshold was considered as the actual number of users downloading the shared content, in the simulation model, it is assumed to be a period of time. This is more realistic as most of the DPI process does not depend on the number of users but on a specific filter of data. From the figure, it becomes apparent that the proposed solution significantly reduces the exchanged data over the network for each traffic case. This is trivial due to the fact that using multicast communications to deliver content common among users is more efficient than the otherwise repeated unicast delivery. Furthermore, the threshold for detecting social media applications has a direct impact on the performance of the proposed solution: the higher the threshold, the lower the gain of the proposed solution in aggregate traffic reduction. For some scenarios, we observe that there is no gain if the threshold is equal to four minutes. This is attributable to the fact that after some minutes, a tweet loses its interest among other users, so the number of exchanged content becomes low which limits the efficiency of using multicast communication. Whilst acknowledging the impact of the threshold on the performance of the proposed solution, it shall be mentioned that as soon as STDF detects a social network application with frequently requested common content (i.e., after the threshold is met), it reports this to PDM and subsequently the service orchestrator instantiates a proxy to intercept the upcoming requests and deliver their relevant traffic over multicast. So until the proxy has indeed established the multicast communication, all the coming requests continue to be served by unicast as in the conventional way. However, it shall be noted that the time required to instantiate and orchestrate the needed VNFs to establish the multicast connection shall be in the order of few msec, e.g., using ClickOS technology as demonstrated in [33]. Table I illustrates the performance of the proposed mechanism, when the frequently requested content



(a) small content size (b) small content size

(c) high content size. (d) high content size.

Fig. 14. Amount of data exchanged according to users' popularity: proposed solution vs unicast distribution.

traffic is offloaded through WiFi rather than using multicast communication. The table shows the amount of offloaded traffic through WiFi for varying thresholds and in case of the fourth scenarios defined earlier. We observe a trend similar to that of Fig. 13, whereby the smaller the threshold value is, the higher the percentage of offloaded social traffic is (i.e. saving LTE bandwidth). Moreover, we remark that the content size has more impact on the performance of the proposed mechanism than the probability p, which is trivial as the higher the size of data is, the higher the LTE saved bandwidth is. However, the high gain in the offload case comes at a high cost, as mobile operators have to invest in deploying another access network (i.e., WiFi network) in addition to the LTE infrastructure.

From this figure, we notice that the gain of the proposed solution depends particularly on the size of the exchanged content and the number of followers that hit the link embedded in the tweet. The higher the size of the exchanged content and the higher the number of hits, the higher the gain. To investigate this behavior with more clarity, we use users' popularity as a parameter. Fig. 14 plots the amount of data exchanged for each user that tweets a message. The x-axis shows the users' indices ordered according to their popularity. As stated before, we used a Zipf law with parameter $s = 0.56$. Fig. 14 considers two cases: high content size and low content size. For both cases, we notice that the gain achieved by the proposed solution is proportional to the popularity of the content. This confirms the behavior observed earlier, i.e., the higher the number of hits (frequent transfer of the content in the mobile network) the higher the gain achieved by the multicast communication. This indeed shows the need for dynamically refining the proxy functionalities in order to achieve the highest gain. Indeed, establishing a multicast channel could be more costly in comparison to the gain achieved by the proposed solution. Therefore, it may be worthwhile having a process (i.e., as part of the PDM unit) that decides whether to establish multicast communication for a particular content and that is based on its popularity. Such process may further decide the characteristics of VMs to instantiate for running the VNFs (e.g., MBMS GW and MB-SC) that will be handling the traffic of the popular content. This decision can be part of the CRA unit in the envisioned framework.

VI. Conclusion

In this paper, we proposed a complete framework to identify and efficiently handle social-based mobile applications that may waste the scarce resources of mobile networks. The proposed framework consists of two modules, namely a traffic identification module and a policy enforcement entity. Both entities are integrated into the Service Function Chain at the SGi LAN of mobile networks and identify an application/session as "an application/session with frequently and dynamically updated content" based on the frequency at which its content or part of its content is delivered to a UE or a set of UEs. Indeed, an application is qualified of such if the inter-arrival time between two consecutive HTTP GET requests from a UE or a set of UEs using the application is shorter than a certain threshold, and/or the number of HTTP GET requests from a UE or a set of UEs, using the application, issued during a time interval exceeds a certain value, or if the traffic delivered over a session exceeds a certain data volume threshold and is delivered to many users.

Upon detection of such application, the policy enforcement entity, also hosted in the SFC, enforces a suitable policy with regard to the relevant data traffic; offloading the relevant data traffic to a particular access network, rerouting the relevant data traffic to/from a different server or via a proxy server, and/or requesting the concerned UEs to join a relevant multicast group. The policy enforcement entity also ensures the deployment and lifecyle management of cloud resources needed by each policy, and orchestrates the service underlined by the policy. The proposed solution makes efficient usage of the available MBMS technology and alleviates congestion at both the mobile core network and RAN by reducing the load of duplicate content. It shall be noted that all the proposed modules are based on the NFV concept and may be hosted on VMs in SFC; incurring limited CAPEX and OPEX to mobile network operators and benefiting from the numerous advantages cloud computing offer (e.g., on-demand, self-service, elasticity, cost-efficient scalability, agility, and pay-as-you-go). Both analytical analysis and simulations were conducted considering the case of Twitter-like services and significant gain in terms of core network load reduction was achieved under different simulated scenarios, particularly when the shared content is highly popular.

REFERENCES

[1] W. Liu *et al.*, "Service function chaining uses cases," IETF draft, expires Mar. 29, 2014.

[2] T. Taleb *et al.*, "EASE: EPC as a service to ease mobile core network deployment over cloud," *IEEE Netw. Mag.*, vol. 29, no. 2, pp. 78–88, Mar. 2015.

[3] T. Taleb, "Towards carrier cloud: Potential, challenges, & solutions," in *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, pp. 80–91, Jun. 2014.

[4] S. Mackie *et al.*, "Service function chains using virtual networking," Internet draft, Oct. 2014.

[5] K. Samdanis, T. Taleb, and S. Schmid, "Traffic offload enhancements for eUTRAN," *IEEE Commun. Surv. Tuts. J.*, vol. 14, no. 3, pp. 884–896, Third Quart. 2012.

[6] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description," 3GPP TS 23.246 v6, 2005, 8. European Telecommunications Standards Institute.

[7] N. Bouten *et al.*, "A multicast-enabled delivery framework for QoE assurance of over-the-top services in multimedia access networks," *J. Netw. Syst. Manage.*, vol. 21, no. 4, pp. 677–706, 2013.

[8] X. Zhou *et al.*, "Understanding the nature of social mobile instant messaging in cellular networks," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 389–392, Mar. 2014.

[9] M.-B. Krishnaurtly and P. Gill, "A few chirps about twitter," in *Proc. ACM Workshop Online Social Netw. (WOSN)*, 2008, pp. 19–24.

[10] M. Franck-Ytter and H. Overby, "An empirical study of valuation and user behavior in social networking services," *Proc. World Telecommun. Congr. (WTC)*, 2012, pp. 1–6.

[11] H. Kwark, C. Lee, H. Park, and S. Moon, "What is Twitter, A social network or News Media," in *Proc. ACM WWW*, 2010, pp. 591–600.

[12] P. Doods and D. Watts, "A genaralize model of social and biological contagion,"*J. Theor. Biol.*, vol. 232, pp. 587–605, 2005.

[13] Z. Wang, L. Sun, S. Yang, and W. Zhu, "Prefetching Strategy in peer-assisted Social video streaming," in *Proc. ACM Multimedia*, 2011, pp. 1233–1236.

[14] JM. Pujol *et al.*, "The little engine(s) that could: Seatting online social networks," *SIGCOMM Comput. Commun. Rev.*, vol. 41, pp. 375–386, Aug. 2010.

[15] X. Cheng and J. Liu, "Load balancing migration of social media to content clouds," in *Proc. ACM NOSSDAV*, 2011, pp. 51–56.

[16] Z. Wang *et al.*, "Propagation-based Social-aware replication for social video contents," in *Proc. ACM Multimedia*, 2012, pp. 29–38.

[17] Z. Wang, L. Sun, C. Wu, and S. yang, "Guiding internet-scale video service deployment using microblog-based prediction," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2012, pp. 2901–2905.

[18] S. Medjiah, T. Taleb, and A. Toufik, "Sailing over data mules in delay tolerant networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 5–13, Jan. 2014.

[19] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Taming user-generated content in mobile networks via drop zones," *IEEE/ACM Trans. Netw. (TON)*, vol. 20, no. 4, pp. 1010–1023, Aug. 2012.

[20] Y. Jin *et al.*, "Characterizing data usage patterns in a large cellular network," in *Proc. SIGCOMM Workshop Cellular Netw. Oper. Challenges Future Des. (CellNet)*, Helsinki, Finland, Aug. 2012, pp. 7–12.

[21] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proc. ACM SIGMETRICS*, Jun. 2011, pp. 305–316.

[22] A. Essail *et al.*, "QoE-driven resource optimization for user generated video content in new generation mobile networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2011, pp. 913–916.

[23] J. Erman, A. Gerber, K. K. Ramadrishnan, S. Sen, and O. Spatscheck, "Over the top video: The gorilla in cellular networks," in *Proc. ACM SIGCOMM Internet Meas. Conf.*, Berlin, Germany, Nov. 2011, pp. 127–136.

[24] Y. Zhang and A. Arvidsson, "Understanding the characteristics of cellular data traffic," in *Proc. SIGCOMM Workshop Cellular Netw. Oper. Challenges Future Des. (CellNet)*, Helsinki, Finland, Aug. 2012, pp. 461–466.

[25] T. Taleb and A. Ksentini, "Impact of emerging social media on mobile networks," in *Proc. IEEE Int. Conf. Commun. (ICC'13)*, Budapest, Hungary, Jun. 2013, pp. 5934–5938.

[26] F. Benevenuto *et al.*, "Characterizing user behavior in online social networks," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf. (IMC'09)*, 2009, pp. 49–62.

[27] (2015) M. Bean [Online]. Available: http://forio.com/simulate/mbean/simulated-bit-ly-traffic/overview/

[28] T. Taleb and A. Ksentini, "VECOS: A vehicular connection steering protocol," *IEEE Trans. Veh. Technol.*, vol. 64, no. 3, pp. 1171–1187, Mar. 2015.

[29] T. Taleb, K. Samdanis, and S. Schmid, "DNS-based solution for operator control of selected IP traffic offload," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011, pp. 1–5.

[30] [GS-NFV] ETSI NFV Group Specification, Network Functions Virtualizsation (NFV), Use Cases. ETSI GS NFV 001v1.1.1, 2013.

[31] ETSI GS NFV-MAN 001, "Network function virtualization (NFV) Management and Orchestration (MANO)," V. 1.1.1, 12-2014, Dec. 2014, ETSI Industry Specification Group (ISG), [Online]. Available: http://www.etsi.org/

[32] 3GPP, "Introduction of the multimedia broadcast/multicast service (MBMS) in the radio access network (RAN), stage 2," 3GPP TS 25.346, Dec. 2007, European Telecommunications Standards Institute.

[33] A. Ksentini, T. Taleb, and F. Messaoudi, "A LISP-based implementation of follow me cloud," *IEEE Access*, vol. 2, pp. 1340–1347, Nov. 2014.

[34] 3GPP, "Multimedia broadcast/multicast service (MBMS); Protocols and codecs," 3GPP TS 26.346, Sep. 20, European Telecommunications Standards Institute.

Tarik Taleb (S'05–M'05–SM'10) received the B.E degree (with distinction) in information engineering, and the M.Sc. and Ph.D. degrees in information sciences from the Graduate School of Information Sciences, Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a Professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. Prior to his current academic position, he was working as a Senior Researcher and 3GPP Standards Expert with NEC Europe Ltd., Heidelberg, Germany. He was then leading the NEC Europe Labs Team working on R&D projects on carrier cloud platforms, an important vision of 5G systems. He was also serving as a Technical Leader of the main work package, Mobile Core Network Cloud, in EU FP7 Mobile Cloud Networking project, co-ordinating among 9 partners including NEC, France Telecom, British Telecom, Telecom Italia, and Portugal Telecom. Before joining NEC and until March 2009, he worked as an Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a laboratory fully funded by KDDI, the second largest network operator in Japan. From October 2005 to March 2006, he worked as a Research Fellow with the Intelligent Cosmos Research Institute, Sendai, Japan. His research interests include architectural enhancements to mobile core networks (particularly 3GPP's), mobile cloud networking, network function virtualization, software-defined networking, mobile

multimedia streaming, inter-vehicular communications, and social media networking. He has been also directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture working group. He is a Distinguished Lecturer of the IEEE Communications Society (ComSoc) and a member of the IEEE Communications Society Standardization Program Development Board. As an attempt to bridge the gap between academia and industry, he founded the IEEE Workshop on Telecommunications Standards: from Research to Standards, a successful event that was awarded the Best Workshop Award by IEEE ComSoc. Based on the success of this workshop, he has also founded and has been the Steering Committee Chair of the IEEE Conference on Standards for Communications and Networking. He is the General Chair of the 2019 edition of the IEEE Wireless Communications and Networking Conference (WCNC'19) to be held in Marrakech, Morocco. He is/was on the Editorial Board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the *IEEE Wireless Communications Magazine*, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, and a number of Wiley journals. He is serving as the Chair of the Wireless Communications Technical Committee, the largest in the IEEE ComSoC. He also served as the Vice Chair of the Satellite and Space Communications Technical Committee of IEEE ComSoc (2006–2010). He has been on the technical program committee of different IEEE conferences, including Globecom, ICC, and WCNC, and chaired some of their symposia. He was the recipient of the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher Award (June 2009), the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation (March 2008), the 2007 Funai Foundation Science Promotion Award (April 2007), the 2006 IEEE Computer Society Japan Chapter Young Author Award (December 2006), the Niwa Yasujirou Memorial Award (February 2005), and the Young Researcher's Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (October 2003).

Adlen Ksentini (SM'14) received the M.Sc. degree in telecommunication and multimedia networking from the University of Versailles Saint-Quentin-en-Yvelines, Versailles, France, and the Ph.D. degree in computer science from the University of Cergy-Pontoise, Cergy-Pontoise, France, in 2005. He is currently an Associate Professor with the University of Rennes 1, Rennes, France. He is a member of the Dionysos Team with INRIA, Rennes, France. He is involved in several national and European projects on QoS and QoE support in future wireless and mobile networks. He has coauthored over 80 technical journal and international conference papers. His research interests include future Internet networks, mobile networks, QoS, QoE, performance evaluation, and multimedia transmission. He is the TPC Chair of the Wireless and Mobile (WMN) Symposium of the IEEE ICC 2016. He was a Guest Editor of the *IEEE Wireless Communication Magazine*, the *IEEE Communication Magazine*, and two ComSoc MMTC letters. He has been on the technical program committee of major IEEE ComSoc, ICC/Globecom, ICME, WCNC, and PIMRC conferences. He was the recipient of the Best Paper Award from the IEEE ICC 2012 and ACM MSWiM 2005.

Min Chen (M'08–SM'09) is a Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China. He was an Assistant Professor with the School of Computer Science and Engineering, Seoul National University (SNU), Seoul, South Korea, from September 2009 to February 2012. He was the R&D Director with Confederal Network Inc. for half a year. He worked as a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of British Columbia (UBC), Vancouver, BC, for three years. Before joining UBC, he was a Postdoctoral Fellow with SNU for one and half years. He has more than 260 paper publications, including 120+ SCI papers. He has authored a book on IoT *OPNET IoT Simulation* (HUST Press, 2015), and a book on big data *Big Data Related Technologies* (series in computer science) (Springer, 2014). His Google Scholars Citations reached 5,400 with an h-index of 35. His top paper was cited 568 times, while his top book was cited 420 times as of June 2015. His research interests include Internet of Things, mobile cloud, body area networks, emotion-aware computing, healthcare big data, cyber physical systems, and robotics. He serves as an Editor or an Associate Editor for *Information Sciences, Wireless Communications, and Mobile Computing*, *IET Communications*, *IET Networks*, international journal of *Security and Communication Networks* (Wiley), *Journal of Internet Technology*, *KSII Transactions on Internet and Information Systems*, and *International Journal of Sensor Networks*. He is the Managing Editor for IJAACS and IJART. He is a Guest Editor for IEEE NETWORK, and the *IEEE Wireless Communications Magazine*. He is the Chair of IEEE Computer Society (CS) Special Technical Communities (STC) on Big Data. He is a Co-Chair of IEEE ICC 2012-Communications Theory Symposium, and the IEEE ICC 2013-Wireless Networks Symposium. He is the General Co-Chair for the 12th IEEE International Conference on Computer and Information Technology (IEEE CIT-2012) and Mobimedia 2015. He is the General Vice Chair for Tridentcom 2014. He is a Keynote Speaker for CyberC 2012, Mobiquitous 2012, and Cloudcomp 2015. He was the recipient of the best paper award from the IEEE ICC 2012 and best paper runner-up award from QShine 2008.

Riku Jäntti (M'02–SM'07) received the M.Sc (with distinction) in electrical engineering and the D.Sc. degree (with distinction) in automation and systems technology from Helsinki University of Technology (TKK), Espoo, Finland, in 1997 and 2001, respectively. He is an Associate Professor (tenured) of communications engineering and the Head of the Department of Communications and Networking with Aalto University School of Electrical Engineering, Espoo, Finland. Prior to joining Aalto (formerly known as TKK) in August 2006, he was a Professor (*pro tem*) of computer science, University of Vaasa, Vaasa, Finland. His research interests include radio resource control and optimization for machine type communications, cloud-based radio access networks, spectrum and coexistence management, and RF inference. He is an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.