Interdependent Multi-layer Spatial Temporal-based Caching in Heterogeneous Mobile Edge and Fog Networks

Vu San Ha Huynh[a] and Milena Radenkovic[b]
School of Computer Science, University of Nottingham, Nottingham, U.K.

Keywords: Content Caching, Mobile Edge and Fog Networks, Network Multilayer Interplay, Spatial-temporal Locality.

Abstract: Applications and services hosted in the mobile edge/fog networks today (e.g., augmented reality, self-driving, and various cognitive applications) may suffer from limited network coverage and localized congestion due to dynamic mobility of users and surge of traffic demand. Mobile opportunistic caching at the edges is expected to be an effective solution for bringing content closer and improve the quality of service for mobile users. To fully exploit the edge/fog resources, the most popular contents should be identified and cached. Emerging research has shown significant importance of predicting content traffic patterns related to users' mobility over time and locations which is a complex question and still not well-understood. This paper tackles this challenge by proposing K-order Markov chain-based fully-distributed multi-layer complex analytics and heuristics to predict the future trends of content traffic. More specifically, we propose the multilayer real-time predictive analytics based on historical temporal information (frequency, recency, betweenness) and spatial information (dynamic clustering, similarity, tie-strength) of the contents and the mobility patterns of contents' subscribers. This enables better responsiveness to the rising of newly high popular contents and fading out of older contents over time and locations. We extensively evaluate our proposal against benchmark (TLRU) and competitive protocols (SocialCache, OCPCP, LocationCache) across a range of metrics over two vastly different complex temporal network topologies: random networks and scale-free networks (i.e. real connectivity Infocom traces) and use Foursquare dataset as a realistic content request patterns. We show that our caching framework consistently outperforms the state-of-the-art algorithms in the face of dynamically changing topologies and content workloads as well as dynamic resource availability.

1 INTRODUCTION

User mobile devices today are increasingly intelligent which leads to the explosive development of new applications involving distributed real-time mobile processing and increasing traffic demands (e.g. HD video streaming, remote health care, critical applications for public safety communications, augmented/virtual reality apps and automatic driving/traffic control). Varying mobility patterns, network topology changes, potential disconnections and resource restrictions in mobile environments pose many challenges for the design and implementation of future mobile network algorithms, particularly content caching with an aim to bring contents proactively as close as possible to the users and improve the reliability and efficiency of mobile edge/fog networks and users' services. Typical edge/fog networks consist of heterogeneous nodes which can include end users and edge devices with different computing resources and communication capabilities (Liu et al., 2018). Our architecture design assumes that the communication between edge/fog nodes is handled in mobile opportunistic multi-hop manner.

Existing opportunistic caching policies in mobile edge/fog networks such as (Wang et al., 2014; Fricker et al., 2012) typically cannot capture, predict and adapt to the spatial-temporal locality of content requests needed for more accurate content popularity-based caching decisions because they rely on assumptions that content interests occur independently of users' mobility and resources. More specifically, when content caching predictions are not sufficiently fine-grained, they result in increased cache miss (i.e. the content request needs to traverse

[a] https://orcid.org/0000-0002-5472-5328
[b] https://orcid.org/0000-0003-4000-6143

Huynh, V. and Radenkovic, M.
Interdependent Multi-layer Spatial Temporal-based Caching in Heterogeneous Mobile Edge and Fog Networks.
DOI: 10.5220/0008167900340045
In *Proceedings of the 9th International Conference on Pervasive and Embedded Computing and Communication Systems (PECCS 2019)*, pages 34-45
ISBN: 978-989-758-385-8

from subscribe to publisher), delay and resource consumption. In order to tackle this complex challenge, we propose a next-generation content caching approach that is able to capture more accurately the dynamic spatial-temporal correlation of mobility and traffic patterns as well as the mobility-content traffic interplay needed to support more reliable, adaptive caching algorithms. Previous research (Wang et al., 2017; Radenkovic et al., 2018) has shown that the centralised solution is not scalable in mobile complex heterogeneous network topologies due to its high complexity and single point of failure problem. Moreover, distributed solution technique (Wang et al., 2017) may still cause high connectivity overheads although it provides a cheaper computable lower bound compared to the centralised solution. In addition, previous research has also shown that collaborative caching usually outperforms both locally and centrally optimized algorithms (Saha et al., 2013). Therefore, we propose a fully-distributed predictive analytic and heuristic-based caching approach which comprises of multi-layer complementary real-time distributed predictive heuristics to maintain the best possible trade-off between caching performance and resource consumption. Note that our focus is not to build a protocol that forces nodes to cooperate to achieve mutual benefit, but rather to design an underlying algorithm that ensures no node attains lower utility by collaborating with others, similarly to (Radenkovic and Huynh, 2017; Wang et al., 2017; Radenkovic et al., 2018). Existing research utilizes different approaches to deal with trend prediction such as reinforcement learning, Bayesian learning, Markov chain (Ruan et al., 2019), or Exponential Moving Average (Radenkovic and Huynh, 2017; Radenkovic et al., 2018; Huynh and Radenkovic, 2018). Due to the continuous nature of users' mobility and content requests, in this paper, we propose to apply the concept of high-order Markov chain to our complex real-time analytics to more accurately predict the content traffic trends based on the historical information of content traffics related users' mobility.

This paper extends the multilayer multidimensional predictive heuristics integrated in CafRepCache (Radenkovic and Huynh, 2017; Radenkovic et al., 2018) which is an predictive adaptive collaborative cognitive forwarding and caching framework in heterogeneous opportunistic mobile networks. We propose a novel prediction model that allows improved congruency with both the underlying network and users demands, more accurately capture the temporal and spatial locality of mobility and content requests as well as mobility-content traffic patterns interplay. We build on two integral complementary multidimensional predictive analytics and heuristics: i) temporal predictive analytics and heuristics that captures the temporal locality of content requests upon request access time, enables more responsiveness to the rising trend of newly high popular contents and fading out of older contents over time as well as avoid one-timer contents and mitigate flash crowd effect; ii) spatial predictive analytics and heuristics that capture the spatial locality of user mobility and content requests as well as balance the trade-off between serving different regions of contents' subscribers. We exploit mobility-content traffic patterns interplay and do not focus on resource analytics (Radenkovic and Huynh, 2017; Radenkovic et al., 2018) in this paper.

The paper begins by providing an overview of the related work in section 2, section 3 describes and discusses our models and multilayer novel predictive heuristics, section 4 evaluates the effect of each complement heuristic on caching performance in mobile DTN across a range of metrics over two different network topologies and a real-world location-based service dataset for content workloads. Section 5 gives a conclusion.

2 RELATED WORK

Authors in (Le et al., 2015) propose a forwarding and cache replacement policy for SocialCache based on content popularity driven by frequency and freshness of content requests. As part of its replacement policy, SocialCache may remove a cached content from the network, thus reduce the cache hit ratio and increase delays. This problem is exacerbated when the resource is limited and the replacement rate is high as (Le et al., 2015) is not resource and congestion aware. Authors in (Zhang et al., 2014) propose Optimal Cache Placement Based on Content Popularity (OCPCP) that takes a caching decision based on the frequency of content requests at a caching node. That is, the more frequent requests for content, the higher chance that content will be requested again. Time Aware Least Recent Used (TLRU) (Bilal and Kang, 2014) is an extension of the simple LRU in which the time stamp of an arriving content is recorded locally by a single node. The arriving content is cached if the average request time is smaller than the time stamps of the stored contents. Authors in (Mardham et al., 2018) propose Location-based Caching that uses decay function measured by the function of distances between subscribers and caching points and some varying attributes, such as time or number of requests to classify contents and replace them when cache memory is full. (Mardham et al., 2018) considers the fairness problem that some contents belonging to some specific location may be more important and is cached often across the network, even if it is not very popular. Authors in (Flores et al., 2017) proposed a social-aware hybrid offloading strategy for load balancing and computation sharing based on node's stability which is measured by contact frequency and duration in order to improve the availability of offloading support for mobile users. However, (Flores et al., 2017) does not support high topology dynamics with intermittent disconnections and dynamically changing publishers and subscribers workload patterns.

Existing opportunistic caching policies such as (Wang et al., 2014; Fricker et al., 2012) rely on an assumption that the content distribution in the networks approximately follows Zipf's law (Yoneki et al., 2008; Breslau et al., 1999) or the characterisation of content requests have been based on Independent Reference Model (IRM) which assumes that content requests occur independently. However, authors in (Dabirmoghaddam et al., 2014; Dán and Carlsson, 2010) have questioned the validity of the IRM model and Zipf's law model. According to the proposal in (Dabirmoghaddam et al., 2014), content requests often exhibit both temporal locality and spatial locality. Researches on today's social networks (Dabirmoghaddam et al., 2014; D'Silva et al., 2018) suggest that if content is requested at a certain period in time, more likely it will be requested again in the near future. In fact, the content is not requested scattered randomly and independently over time; but rather particular contents are interested at a certain time interval, while its popularity gradually fades out. Spatial locality of content requests is based on the fact that content requests of the same content are more likely to be issued by geographically close areas. More precisely, the requests coming from a specific region in space are more likely to be similar than those collected over regions far apart (D'Silva et al., 2018).

3 FULLY-DISTRIBUTED PREDICTIVE MOBILE EDGE AND FOG CACHING

In this section, we briefly describe CafRepCache (Radenkovic and Huynh, 2017; Radenkovic et al., 2018) framework, then we propose to extend its integral multilayer fully-distributed predictive heuristics.

CafRepCache (Radenkovic and Huynh, 2017; Radenkovic et al., 2018) is a multi-path content and interest forwarding and replication with adaptive collaborative caching framework in heterogeneous opportunistic mobile networks. It utilises fully-localised and ego networks multi-layer predictive heuristics about dynamically changing topology, resources and content popularity to manage dynamic trade-offs between minimizing the end-to-end latency and maximising content delivery while enabling resource efficiency and congestion avoidance. CafRepCache relies on three theory foundation: network science, social science and information science as shown in Fig. 1.

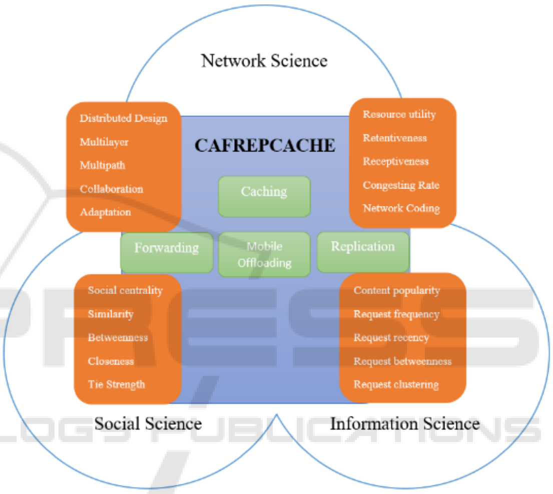

Figure 1: CafRepCache theory foundation.

CafRepCache system is modelled as a network *G* that consists of a set *N* of nodes n_i ($n_i \in N$) and a set *E* of edges, *G = (N, E)*. As the connectivity of the network and the state of the nodes change over time, each of these sets is modelled as time series, thus $N = \{N^t : \mathrm{t} \in \mathrm{T}\}$ and $E = \{E^t : \mathrm{t} \in \mathrm{T}\}$. We denote with O a set of content files that can be requested by the network. Each content $o_k^t \in O$ (or o_k for simplicity) is published at time *t*. Node $n_i \in N$ may act as either a subscriber of content o_k, denoted by $S_{i,k}^t$ or a publisher $P_{i,k}^t$ or a caching point $C_{i,k}^t$ at any time *t* from any location while being mobile. Thus for any content o_k, a set of subscribers who are interested in o_k is denoted as $S_k^t = \{S_{i,k}^t | n_i \in N\}$ and so on. Each node in the network is able to perform predictive analytics of multivariate mixed data (e.g. content and mobility) as well as collaborate and exchange its local observations with other neighbour nodes when two nodes are in contact in order to capture and detect various events (e.g. user connectivity patterns, request patterns) in more accurate and responsive manner without the need of global knowledge. We define "ego network" of each node $n_i : EN_i$ as a network consisting of n_i together with the nodes they are connected most recently (i.e. k-hops neighbours) or nodes it meets at regular interval (most frequently) over the time duration ΔT and all the links among those nodes. In this way, ego network allows each node to give its own regional or temporal perspective of the network (or both are included).

3.1 Dynamic Complex Temporal-Graph Heterogeneous Network Topologies

In order to provide insights of users' content traffic-related mobility, connectivity patterns, this section analyses CafRepCache in different underlying network topologies with different degree of mobility, connectivity patterns which helps us to better design novel modelling analytics to capture and predict the spatial-temporal correlations of content request patterns with underlying network topologies. Empirical evidence shows fixed and wired networks are mostly scale-free, whereas mobile and opportunistic networks can be either random or scale-free. Thus, we analyse theoretically CafRepCache in extremely heterogeneous random topologies as well as scale-free topologies to understand fundamental performance limitations of CafRepCache in different realistic networks.

3.1.1 Complex Temporal Random Network Topologies

In random networks, the majority of previous studies in mobile networks assume Poisson contact processes and model user encounters as independent Poisson processes with rate λ due to the time between two consecutive contacts of a pair of nodes follows exponential distribution (Bornholdt and Schuster, 2006). The probability of some node $n_i \in N$ connecting with some other node n_j (Bornholdt and Schuster, 2006) is: $P(n_i) = \frac{1}{N}\int P(v_{ij})dn_i$.

The probability that node $n_i \in N$ connects with exactly n other nodes (Bornholdt and Schuster, 2006) within time T: P(n encounters) ≈ $\frac{e^{-\lambda \mathrm{T}}(\lambda \mathrm{T})^n}{n!}$. Let p'_k be the probability that content o_k is stored in the cache of an arbitrary nodes, then the probability of the cache miss in random network is:

Prob(content o_k is not in the cache) * Prob(interest request never reaches correct caching points within time T)
= $(1 - p'_k)$ * $\sum_n^N \mathrm{Prob}(\mathrm{n\ encounters}) *$ $\mathrm{P}(\mathrm{none\ of\ the\ n\ encounters\ has\ requested\ contents})$
= $(1 - p'_k)$*$\sum_n^N \frac{e^{-\lambda \mathrm{T}}(\lambda \mathrm{T})^n}{n!}(1 - {}'p_k)^n$.

3.1.2 Complex Temporal Scale-free Network Topologies

Complex temporal scale-free networks are characterized by a highly heterogeneous degree distribution, which follow a power-law distribution (Yoneki et al., 2008). Although the network may change significantly over time, the degrees of its nodes obey the power-law model at any time (Yoneki et al., 2008).

The probability P(n encounters) of a node in the network goes for large values of n as: $f(n) \sim n^{-\gamma}$ where γ is the shape parameter of the power-law distribution and represents the degree of the power-tail.

Then the probability of a cache miss of content o_k in scale-free networks is $(1 - p'_k) * \sum_n^N n^{-\gamma}(1 - p'_k)^n$.

3.2 Spatial-temporal Analytics and Heuristics for Content Caching

In this section, we describe two non-trivial predictive analytics and heuristics (i.e. K-order Markov chained-based temporal heuristics and spatial clustering heuristics) that cover different dimensions of the spatial-temporal dynamics and mobility-content traffic interdependence problem. When combined together, they allow forming dynamic transient interest and data dissemination topologies based on predictive analysis and commonalities between their interests, caches and retrieval histories as well as connectivity histories.

3.2.1 Dynamic K-order Markov Chained-based Temporal Heuristics

Each node in the network resolves the request frequency, recency and betweenness in fully-localised distributed manner. When two nodes are in contact, they exchange their local observations and continuously resolve the value of dynamically changing predictive heuristic based on both its local observation and the collaborative observations it gets from others. We apply the concept of K-order Markov chain on our complex analytics to predict the content traffic trends based on the historical information of content traffics related users' mobility. When K = 1, a lot of historical state information is ignored and only the information of the current moment is used. Such limitations make the practical application of 1-order Markov chain prediction method lack of prediction accuracy. In line with (Ruan et al., 2019), compared with other existing time-order-based prediction methods, the K-order Markov chain performs much better when the order number K = 2. We apply K-order Markov chain to leverage efficiently historical information of content requests and subscribers' mobility to predict the future trends of content traffics. The K-order is defined as:

$$P(C_t = i_t \,|C_{t-1} = i_{t-1}, \ldots, C_0 = i_0) =$$
$$P(C_t = i_t \,|C_{t-1} = i_{t-1}, \ldots, C_{t-k} = i_{t-k})$$

We then introduce our method to measure the temporal heuristics based on request frequency, recency and betweenness as below.

Request frequency counters are additively increased upon the arrival of a content request and decreased through time. This is in order to ensure that if the number of requests for a content has been reduced, its popularity counter will be reduced accordingly and the content will be subject to eviction or offload to other nodes. Given a caching point observes average *f* interests of content *k* during the interval Δt, the request frequency is measured as:

$RequestFrequency = f^{-\lambda\Delta t}$ where λ is a control parameter. Request frequency implies that the more content requests have been observed by a caching point during a short interval time Δt, the more likely that content will be requested in the same interval, thus capture the temporal locality of content requests

Request recency enables our caching design to capture the content popularity trend in responsive manner based on the recorded time stamp of recent requests in different locations for each caching point, thus allow to predict adaptively the emerging contents that may become highly popular in near future and contents that are currently considered as high popular but will be less popular soon. Given a caching point *n* observes the most recent interest of content *k* at $t_{recentReq}$, then we denote $t_{recentReq-1}$, $t_{recentReq-2}$, $t_{recentReq-3}$, etc. as the time that previous interests have been recorded by the caching point. The request recency is measured as how likely a recent content request will trigger a subsequent request at the current time.

$$RequestRecency =$$
$$\frac{2}{1 + e^{-2\frac{\left(t_{currentTime} - t_{recentReq}\right) + \cdots + \left(t_{currentTime} - t_{recentReq-f}\right)}{f}}}$$

Request recency analytic shows that the smaller gap between current time and recent requests observed in the past of content *k*, the higher chance that content *k* will be requested soon in the future.

Request betweenness provides the trade-off between current observed content popularity versus long terms interest in it in order to balance between potentially one-timer contents or fake news and long-term useful content. The time gap between continuous requests in the period of time Δt is measured as: $t_{currentTime} - t_{recentReq}, t_{recentReq} - t_{recentReq-1}, t_{recentReq-1} - t_{recentReq-2}, \cdots$

Then the request betweenness heuristic is measured as:

$$RequestBetweeness =$$
$$\sqrt{\frac{1}{f}\sum_{i=1}^{f}\left(t_{recentReq-i} - t_{recentReq-i-1} - T_{averageTimeGap}\right)^2}$$

The novel temporal heuristics are the combination of *request frequency*, *recency* and *betweenness* which allow CafRepCache to choose the best suitable contents to cache and when to cache by predicting in real time the locality trend of content request patterns over time in different locations and avoid losing valuable contents by reducing the caches for one-timers contents and fake news.

$$TemporalHeuristics = Freq(o_k) + Rec(o_k) + Bet(k_k).$$

3.2.2 Dynamic Spatial Clustering Heuristics

As the requests are more likely to be similar in a specific region, we propose the fully-localised distributed spatial heuristics that aim to classify, recognize and predict content interests coming from dynamic changing clusters of subscribers. The spatial heuristic shows that the higher request rate coming from the same localised region or dynamic cluster of different subscribers, the higher likely that content will be requested again by other subscribers within that location. Given a caching point observes interest requests of content *k* from a set of subscribers $s \in S_k$ within a time interval Δt, the spatial heuristic is measured by the clustering coefficient (Nicosia et al., 2013), similarity, closeness and tie strength (Radenkovic and Huynh, 2017; Radenkovic and Grundy, 2011; Daly and Haahr, 2007) between different subscribers of the same content as below:

$$SpatialHeuristic = CCoef(S_k) + Sim(s_1, s_2) + Close(s_1, S_k) + StrTieStr(s_1, s_2)$$
$$\forall s_1, s_2 \in S_k$$

in which the similarity value (Daly and Haahr, 2007; Nicosia et al., 2013) between $s_1, s_2 \in S_k$ within a time interval Δt is: $\frac{\sum_t |N_{s1} \cap N_{s2}|}{\Delta t}$ where $|N_{s1} \cap N_{s2}|$ is the similarity in contacts between two subscriber s_1, s_2 of the content *k*. The closeness centrality (Daly and Haahr, 2007; Nicosia et al., 2013) of s_1 is a measure of how close s_1 is to any other node in S_k. It is measured as the inverse of the average distance from s_1 to any other node in S_k: $\frac{S-1}{\sum_j d_{1j}}$ where d_{1j} is the distance between s_1 and s_j in the set of subscribers $s \in S_k$.

The node tie strength value (Radenkovic and Huynh, 2017; Radenkovic and Grundy, 2011; Daly and Haahr, 2007) between of subscribers $s_1, s_2 \in S_k$ within a time interval Δt is:

$$\sum\nolimits_{\mathrm{t}} \frac{f(s_1)}{F(s_2)- f(s_1)} + \frac{rec(s_1)}{T(s_2)- rec(s_1)} + \frac{d(s_1)}{D(s_2)- d(s_1)}$$

where $\frac{f(s_1)}{F(s_2)- f(s_1)}$ measures the frequency of contacts between s_1, s_2 ; $\frac{rec(s_1)}{T(s_2)- rec(s_1)}$ measures the recency of contacts between s_1, s_2 and $\frac{d(s_1)}{D(s_2)- d(s_1)}$ indicate the relative distance by hops between s_1, s_2. The complex temporal graph metrics of contact frequency, recency and topology distance allow congruency with the underlying dynamic changing network topology and connectivity. In order to balance the trade-off between serving contents requested from a specific local region of highly connected subscribers and from multiple less-connected subscribers, we favours the weak tie strength which helps to give a wider and broader long-term predictive content popularity instead of only favouring and serving the contents requested from a local region of highly connected subscribers.

Spatial heuristics allow CafRepCache to choose the most suitable caching points to cache popular contents based on its relative location with the subscribers and between the subscribers themselves.

3.3 CafRepCache's Combined Heuristics

In section 3.2, we described our novel approach for opportunistic caching protocol that, we argue, is essential to be able to capture the spatial-temporal locality of mobility and content requests as well as the mobility-content traffic interplay. The above two non-trivial heuristics $h(n_i, o_k)$ cover different dimensions of the mobility-content traffic pattern interdependences problem and when combined they allow managing a number of trade-offs we identified. Each caching point n_i in the network resolves and combines the two heuristics to measure $P(n_i, o_k, \Delta t)$ denoted as the popularity of o_k during the interval time Δt. $P(n_i, o_k, \Delta t)$ implies the probability of how likely the content o_k will be requested in a period of time:

$$P(n_i, o_k, \Delta t) = \sum_{h\, \in H} \alpha_h Util_h(n_i, \mathrm{C}(n_i), o_k)$$

where α_h is the weighting factor of each heuristic, $Util_h(n_i, \mathrm{C}(n_i), o_k)$ is the respective utilities of each heuristic as measurements of their relative gain, loss or equality, calculated as pair-wise comparison between the node's own heuristics and that of the encountered contacts:

$$Util_h(n_i, \mathrm{C}(n_i), o_k) = \frac{h(n_i, o_k)}{h(n_i, o_k) + h(\mathrm{C}(n_i), o_k)}$$

$$h \in H = \{RFreq, RRec, RBet, DCluster\}$$

The predictive analytics are resolved by caching nodes' local observations and collaborative observations from neighbours within its ego network in order to allow each caching point have a more regional converged perspective of the network without the need of global network knowledge.

The total content popularity heuristic is measured as:

$$P(EN_i, o_k, \Delta t) = \frac{1}{EN} \sum\nolimits_{j\, \#\, i \in EN_i}^{EN_i} \alpha_j P\big(n_j, o_k, \Delta t\big).$$

4 EXPERIMENT SETUP AND EVALUATION

This section presents an evaluation of our caching algorithm, first introducing a set of state-of-the-art caching policies as competitive caching algorithms and metrics for comparing the experimental results. For the underlying network topology and mobility patterns, we use a simulation-driven data trace with two very different network topologies: random network and real-world mobility traces Infocom (Scott et al., 2006) in ONE simulator (Keränen et al., 2009) as scale-free network in order to give a deeper and more accurate performance overview of CafRepCache. We use Foursquare New York Dataset (Yang et al., 2014) as a real trace for content requests. This dataset is collected through location-based service Foursquare API (https://developer.foursquare.com/) describing the spatial-temporal locality of content requests in terms of user interests at public venues, it contains 227,428 subscriptions of 18,201 users in different locations of New York city during the period of 10 months. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning (represented by fine-grained venue-categories).

We compare and evaluate CafRepCache on the overall caching performance measured by different criteria (e.g. cache hit ratio, latency, eviction rate, etc.) against multiple state-of-the-art and benchmark proposals: SocialCache (Le et al., 2015), Optimal Cache Placement Based Content Popularity (OCPCP) (Zhang et al., 2014), Time Aware Least Recent Used

Table 1: Values of the simulation parameters.

Parameter	Value
Complex temporal network topologies	Random network, Scale-free network (Infocom)
Content request pattern	Foursquare New York
Number of nodes	50-100
Simulation duration	1 - 3 hours
Request rate	1-25 request/min
Number of contents	10^3 - 10^5
File size	1 MB - 8.4 MB
Interest packet size	8 kB - 128 kB
$\frac{\text{Cache size}}{\text{Total content population}}$	0.1 - 0.6%

(TLRU) (Bilal and Kang, 2014) and LocationCache (Mardham et al., 2018). We have run six increments of the number of subscribers and publisher ranging from 10% to 60% of the total number of nodes. Due to limited space, we report on experiments with increasing number of subscribers, but note that the results for increasing number of publishers are similar to the ones presented here. Without losing generality, we assume that 25% of node population are publishers and varying the number of subscribers. All experiments are repeated ten times and averaged with different random subscribers and publishers. For each experiment, either the cache hit ratio or average latency (in hops) or eviction ratio will be shown. The general simulation parameters details are shown in Table 1.

Fig. 2 shows the spatial and temporal correlation of content traffic (i.e. temporal requests pattern of mobile subscribers) in Foursquare dataset for a content in different locations over time.

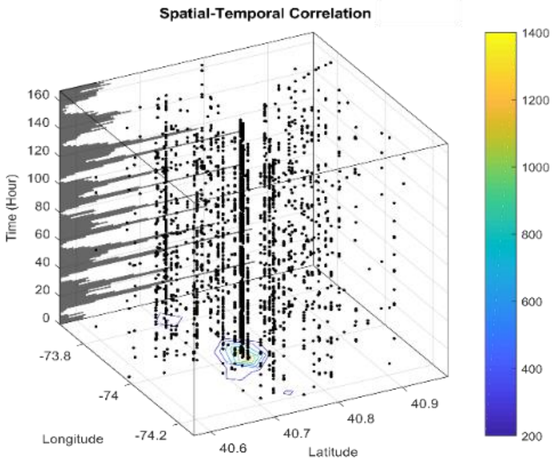

Figure 2: Spatial-temporal correlation of content requests.

It shows the temporal patterns (similarity) of content traffic during weekdays and weekend: if a content is requested at a certain point in time, more likely it will be requested again in near future. In fact, nor are the content references scattered randomly and independently over time; rather, a content might be of particular interest at a certain time interval, while its popularity gradually fades out. The locations of mobile subscribers feature different degrees of similarity in content request such that the location 1 and 2 which are relatively close to each other have similar request patterns compared to that of location 3 which are far away from others. This captures the impact that the geographical diversity of the users has on the observed trace of requested contents by them. More precisely, the requests coming from a specific region in space are more likely to be similar than those collected over regions far apart.

In order to understand the scalability of caching points with regarding to the increasing number of subscribers, we vary the number of subscribers to evaluate the number of caching points in Fig.3, average latency (measured by the number of hops) in Fig. 4 and cache hit ratio (which refers to how many interest packets are matched with the contents in caching points without being forwarded to publishers) in Fig.5 that indicate the efficiency of caching decisions and locations in random topology and scale-free topology.

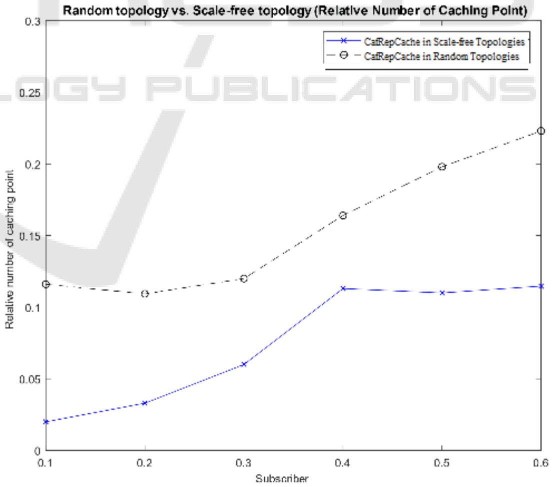

Figure 3: Number of caching points vs. number of subscribers.

Fig. 3 shows that the relative number of CafRepCache caching points increases from 12% to 23% in random networks and from 2 to 11% in scale-free networks regarding the growth of subscribers. We argue that random networks with short average paths and low clustering require more number of caching points to serve the dynamic mobile subscribers while scale-free networks with high social character need less number of caching points which converge to 11% regarding the increasing number of subscribers from 40% to 60%. Fig. 4 and 5 shows that CafRepCache achieves 79.4% and 92.4% cache hit ratio while the hop count average from the caching points to subscribers are 3.01 and 2.7 in random network and scale-free network respectively.

We argue that in random network, CafRepCache benefits from its cache redundancy mechanism that select highly suitable locations for caching and replication when needed as adaptive replication and caching are both necessary to address multi-user data communications in dynamic fragmented and sparse topologies. In scale-free network, CafRepCache utilities its multidimensional predictive analytics and complex temporal graph metrics to make caching decisions in a predictive manner and congruent with the underlying network mobility, connectivity and content interest.

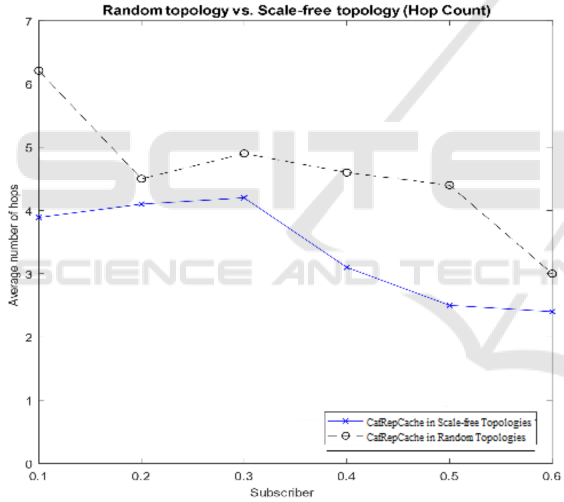

Figure 4: Average hops count between caching points and subscribers.

We show graphs of random network topologies as a worst case scenario in Fig.3-5 and we will focus on scale-free network topologies for the rest of our experiments as it allows our caching decision makings to leverage the spatial-temporal correlations of mobility and traffic patterns as well as mobility-traffic interdependence.

To evaluate the effectiveness of the content request frequency heuristic, we vary the content request patterns which follows different content popularity skewness α (0.6-1.1) utilized in Hawkess process (Dabirmoghaddam et al., 2014) and measure the cache hit ratio. In line with (Dabirmoghaddam et al., 2014), when α becomes larger, there is small number of contents that account for the majority of total requests, or being requested more frequently compared to the others.

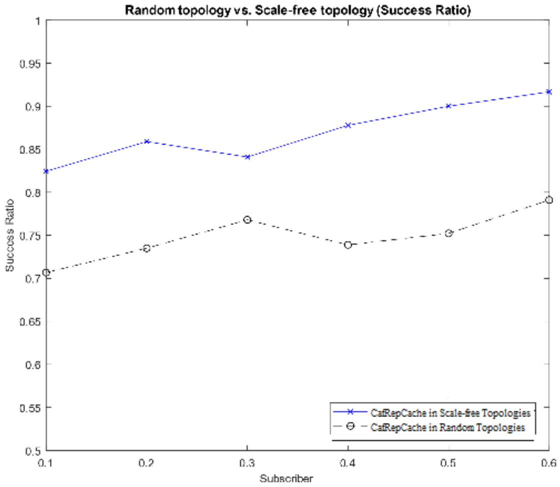

Figure 5: Success ratio vs. number of subscribers.

Fig. 6 shows that our cognitive caching achieves the highest cache hit ratio (ranging from 31% to 86.4%) compared to other competitive algorithms (OCPCP, TLRU, SocialCache, LocationCache) regarding the increase of popularity skewness α. We observe that higher α leads to bigger gap between CafRepCache and others. This is because the request frequency heuristic takes advantage of highly skewed content popularity and content request temporal locality to predict efficiently the incoming content requests and adapts strongly with the content requests while others neglect or could not adapt with the temporal locality of content. CafRepCache is followed by LocationCache and SocialCache that manage up to 76.6% cache hit ratio regarding the increase of content popularity skewness. We observe that LTRU and OCPCP achieve the lowest cache hit ratio (ranging from 24,3% to 60%) as it relies only on simple request frequency or recency metric, thus could not be able to predict adaptively the temporal locality of content request patterns.

As the content popularity changes in real-time regarding the variation in user interests, we further study the impact of varying content popularity fluctuation on the caching performance in order to evaluate the effectiveness of the content request recency heuristic. The Fig.7 shows the popularity variation range from 20 to 120 popularity rank for each round of the experiment. As caching performance of slight variation on content popularity has advantage over sharp variation, it is

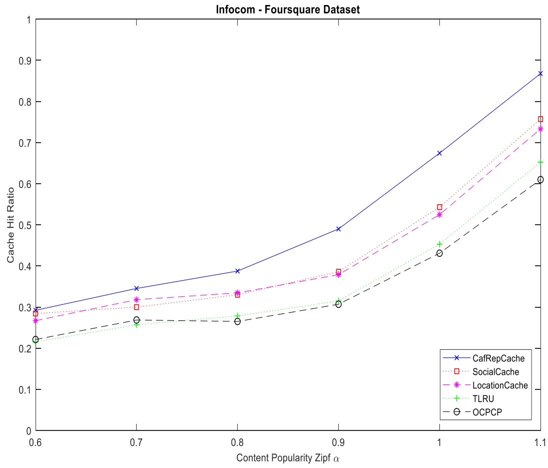

Figure 6: Cache hit ratio vs Popularity zipf alpha.

challenging for the caching algorithms to adapt quickly to the severe and quick popularity alteration.

As shown in Fig.7, CafRepCache's request recency metric keeps good stability regarding the popularity change (typically above 83.1% cache hit ratio), predict adaptively the emerging contents that may become highly popular in near future and contents that are currently considered as high popular but will be less popular soon. The sensitiveness of our proposed algorithm provides better adaptation to rapidly changing content popularity and different network environment while all other competitive caching algorithms could not be able to adapt with sharp variation in popularity.

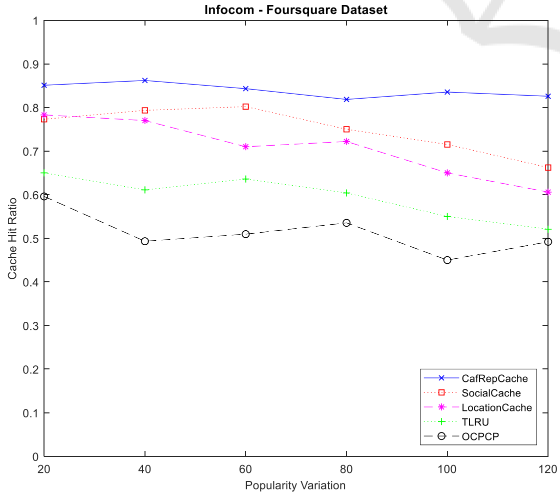

Figure 7: Cache hit ratio vs Popularity variation.

We vary the cache size and measure the cache eviction ratio to evaluate the effectiveness of the content *request betweeness* heuristic. Eviction is one of the popular metrics for the performance evaluation of content caching. When the cache space is full and caching point makes a decision to cache a new arrived content then one of the cached contents is evicted or offloaded to free up the buffer. When the resource is limited and the eviction rate is potentially high due to one-timer contents, the overall network throughput is affected in terms of cache hit ratio and content retrieval latency. In other words, if a stored content is evicted incorrectly and a new request arrives for it, then there is a cache miss and the content has to be retrieved from other nodes or publishers directly. As a result, this increases the content retrieval latency and decreases the cache hit ratio of the caching services. Smaller cache buffer size offers more selective cached contents, thus requires more accurate content popularity prediction.

Fig. 8 shows that CafRepCache achieves the lowest eviction ratio regarding the dynamic changing in size of the caching points (decreasing from 0.62 to 0.34 eviction ratio when the cache size is increased) compared to the state-of-the-arts caching algorithms. CafRepCache is followed by SocialCache and LocationCache (ranging from 0.75 to 0.46 eviction ratio) while TLRU and OCPCP have the worst performance, especially when the cache space is relatively small. This is due to the request betweeness heuristic allows CafRepCache balance the trade-off between current observed content popularity versus long terms interest in order to avoid caching quickly potentially one-timer contents or fake news and losing the long-term useful contents.

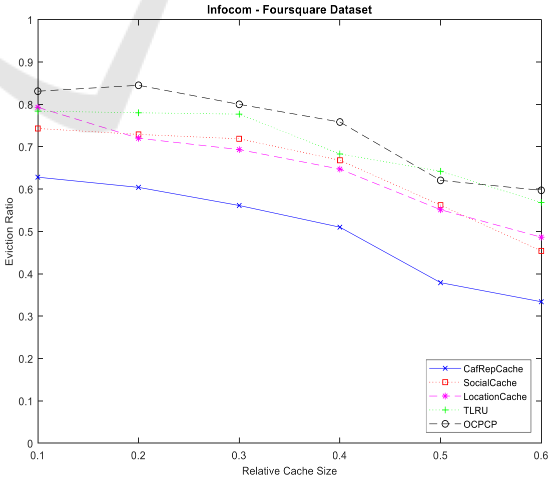

Figure 8: Eviction rate vs Relative cache size.

In order to understand the request spatial-based heuristics integrated in our caching algorithm, we vary the content request patterns which follows different spatial localisation factor β (0.1-0.85) utilized in Hawkess process (Dabirmoghaddam et al., 2014) and measure the cache hit ratio. In line with (Dabirmoghaddam et al., 2014), with low localization factor β, content requests are generated independently. The generated trace, therefore, conforms to the IRM assumption and hence, one single time content is considered. With a localization factor of 0.85, other near nodes in a region more likely request the same content after a node requests it.

Fig. 9 shows that the request spatial clustering heuristic enables CafRepCache to adapt with the spatial locality of content requests. CafRepCache achieves the best performance (around 87% cache hit ratio) followed by LocationCache (80%) and SocialCache (72%). TLRU and OCPCP have no (or little) improvement to the accuracy of predicting content popularity, ranging from 52% to 63% cache hit ratio regarding the dynamically changing of spatial localization factor.

We observe that higher β even leads to bigger gap between CafRepCache and other state-of-the-art content solutions. It is due to the spatial locality heuristic helps to classify and recognise content interests coming from localised group of subscribers, then adaptively predict that the contents will be requested again by other subscribers within that location.

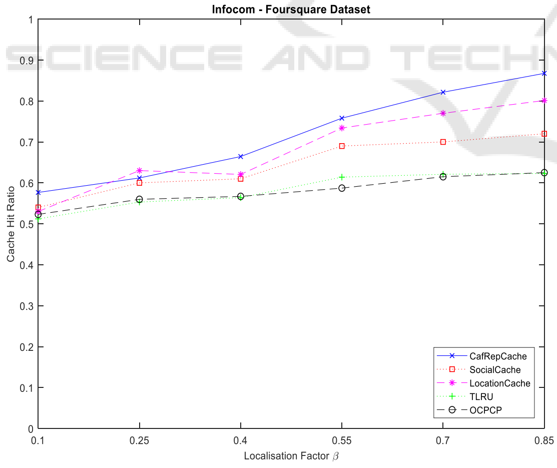

Figure 9: Cache hit ratio vs Spatial localisation factor.

We evaluate the effect of subscriber-caching point connectivity on the performance of different state-of-the-art caching protocols in order to understand the nature of caching points regarding the dynamic connectivity of mobile subscribers.

Fig. 10 shows that CafRepCache outperforms other competitive caching protocols in terms of average delay measured by the number of hops.

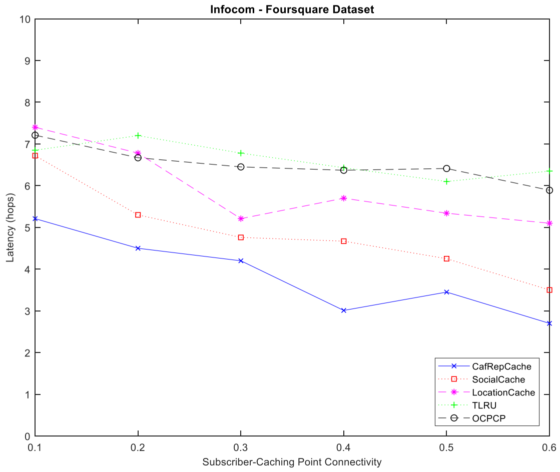

Figure 10: Average latency (hops) vs. Subscriber-Caching point connectivity.

CafRepCache is able to bring the cached contents to only a few hops away from subscribers (2.7 hops) regarding the dynamic centrality of caching points. This is due to CafRepCache benefits from multidimensional multilayer analytics that allow it to place the most suitable set of contents in the most suitable set of caching points which are not only highly central but also have similarity in contacts and interest requests with the subscribers.

5 CONCLUSIONS

We proposed multilayer adaptive predictive distributed collaborative analytics and heuristics for enabling spatial-temporal locality awareness of mobility and traffic patterns as well as mobility-content traffic interplay for opportunistic caching in mobile edge/fog networks. Our combined heuristics allow the caching protocol to be more flexible and responsive to the dynamic mobility and complex content request patterns in the unreliable scenarios imposed by varying publishers and subscribers as well as dynamic resource availability. We performed extensive real trace-driven experiments in ONE simulation (Keränen et al., 2009) and showed that the proposed predictive heuristics help CafRepCache caching framework to perform better than the state-of-the-art caching solutions in terms of success ratio, cache hit ratio, average delay and eviction rate.

We aim to explore our ego-network analytics and heuristics in greater detail and propose adaptive context-aware weighting of the complementary analytics and utilities. We plan to deploy our approach in different application scenarios which have complex temporal network topologies such as smart agriculture (Wietrzyk and Radenkovic, 2010; Brun-Laguna et al., 2016), urban emergency (Huynh and Radenkovic, 2017), intelligent transport system (Loscri et al., 2019), and smart manufacturing (Radenkovic et al., 2015) using software-defined networking (SDN) or network function virtualization (NFV) as in (Radenkovic, 2016; Radenkovic and Huynh, 2016).

REFERENCES

Bilal, M. and Kang, S.-G. (2014) Time aware least recent used (TLRU) cache management policy in ICN, *16th International Conference on Advanced Communication Technology*. IEEE.

Bornholdt, S. and Schuster, H. G. (2006) *Handbook of graphs and networks: from the genome to the internet*. John Wiley & Sons.

Breslau, L., Cao, P., Fan, L., Phillips, G. and Shenker, S. (1999) Web caching and Zipf-like distributions: Evidence and implications, *Ieee Infocom*. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE).

Brun-Laguna K., Diedrichs A. L., Dujovne D., Leone R., Vilajosana X. and Watteyne T. (2016). (Not so) intuitive results from a smart agriculture low-power wireless mesh deployment. *Proceedings of the Eleventh ACM Workshop on Challenged Networks*. ACM.

Dabirmoghaddam, A., Barijough, M. M. and Garcia-Luna-Aceves, J. (2014) Understanding optimal caching and opportunistic caching at the edge of information-centric networks, *Proceedings of the 1st ACM conference on information-centric networking*. ACM.

Daly, E. M. and Haahr, M. (2007) Social network analysis for routing in disconnected delay-tolerant manets, *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. ACM.

Dán, G. and Carlsson, N. (2010) Power-law revisited: large scale measurement study of P2P content popularity, *IPTPS*.

D’Silva, K., Noulas, A., Musolesi, M., Mascolo, C. and Sklar, M. (2018) Predicting the temporal activity patterns of new venues. *EPJ Data Science*, 7(1), 13.

Flores, H., Sharma, R., Ferreira, D., Kostakos, V., Manner, J., Tarkoma, S., Hui, P. and Li, Y. (2017) Social-aware hybrid mobile offloading. *Pervasive and Mobile Computing*, 36, 25-43.

Fricker, C., Robert, P. and Roberts, J. (2012) A versatile and accurate approximation for LRU cache performance, *2012 24th International Teletraffic Congress (ITC 24)*. IEEE.

Huynh, V. S. H. and Radenkovic M. (2017). A novel cross-layer framework for large scale emergency communications. *2017 13th International Wireless Communications and Mobile Computing Conference* (IWCMC), IEEE.

Huynh, V. S. H. and Radenkovic, M. (2018). Understanding information centric layer of adaptive collaborative caching framework in mobile disconnection-prone networks. *2018 14th International Wireless Communications & Mobile Computing Conference* (IWCMC), IEEE

Keränen, A., Ott, J. and Kärkkäinen, T. (2009) The ONE simulator for DTN protocol evaluation, *Proceedings of the 2nd international conference on simulation tools and techniques*. ICST (Institute for Computer Sciences, Social-Informatics and ….

Le, T., Lu, Y. and Gerla, M. (2015) Social caching and content retrieval in disruption tolerant networks (DTNs), *2015 International Conference on Computing, Networking and Communications (ICNC)*. IEEE.

Liu, Z., Yang, X., Yang, Y., Wang, K. and Mao, G. (2018) DATS: Dispersive Stable Task Scheduling in Heterogeneous Fog Networks. *IEEE Internet of Things Journal*.

Loscri, V., Manzoni P., Nitti M., Ruggeri G. and Vegni A. (2019). A social internet of vehicles sharing SIoT relationships.

Mardham, D., Madria, S., Milligan, J. and Linderman, M. (2018) Opportunistic distributed caching for mission-oriented delay-tolerant networks, *2018 14th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*. IEEE.

Nicosia, V., Tang, J., Mascolo, C., Musolesi, M., Russo, G. and Latora, V. (2013) Graph metrics for temporal networks, *Temporal networks*Springer, 15-40.

Radenkovic, M. (2016). Cognitive privacy for personal clouds. *Mobile Information Systems 2016.*

Radenkovic, M. and Grundy, A. (2011) Congestion aware forwarding in delay tolerant and social opportunistic networks, *2011 Eighth International Conference on Wireless On-Demand Network Systems and Services*. IEEE.

Radenkovic, M. and Huynh V. S. H. (2016). Low-cost mobile personal clouds. *2016 International Wireless Communications and Mobile Computing Conference* (IWCMC), IEEE.

Radenkovic, M. and Huynh, V. S. H. (2017) Collaborative cognitive content dissemination and query in heterogeneous mobile opportunistic networks, *Proceedings of the 3rd Workshop on Experiences with the Design and Implementation of Smart Objects*. ACM.

Radenkovic, M., Huynh, V. S. H. and Manzoni, P. (2018) Adaptive real-time predictive collaborative content discovery and retrieval in mobile disconnection prone networks. *IEEE Access*, 6, 32188-32206.

Radenkovic, M., Kostadinov I. and Wietrzyk B. (2015). Increasing communication reliability in manufacturing environments. *2015 International Wireless Communications and Mobile Computing Conference* (IWCMC), IEEE.

Ruan, M., Chen, X. and Zhou, H. (2019) Centrality prediction based on K-order Markov chain in Mobile Social Networks. *Peer-to-Peer Networking and Applications*, 1-11.

Saha, S., Lukyanenko, A. and Ylä-Jääski, A. (2013) Cooperative caching through routing control in information-centric networks, *2013 Proceedings IEEE INFOCOM*. IEEE.

Scott, J., Gass, R., Crowcroft, J., Hui, P., Diot, C. and Chaintreau, A. (2006) CRAWDAD dataset cambridge/haggle (v. 2006-09-15). *CRAWDAD wireless network data archive*.

Wang, L., Tyson, G., Kangasharju, J., Crowcroft, J., Wang, L., Tyson, G., Kangasharju, J. and Crowcroft, J. (2017) Milking the cache cow with fairness in mind. *IEEE/ACM Transactions on Networking (TON)*, 25(5), 2686-2700.

Wang, T., Hui, P., Kulkarni, S. and Cuff, P. (2014) Cooperative caching based on file popularity ranking in delay tolerant networks. *arXiv preprint arXiv:1409.7047*.

Yang, D., Zhang, D., Zheng, V. W. and Yu, Z. (2014) Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1), 129-142.

Yoneki, E., Hui, P. and Crowcroft, J. (2008) Distinct types of hubs in human dynamic networks, *Proceedings of the 1st Workshop on social Network Systems*. ACM.

Zhang, G., Tang, B., Wang, X. and Wu, Y. (2014) An optimal cache placement strategy based on content popularity in content centric network. *JOURNAL OF INFORMATION & COMPUTATIONAL SCIENCE*, 11(8), 2759-2769.

Wietrzyk, B. and M. Radenkovic (2010). Realistic large scale ad hoc animal monitoring. *International Journal On Advances in Life Sciences 2(1 and 2)*.