Proactive Caching in 5G Small Cell Networks

Ejder Baştuğ[◇], Mehdi Bennis[⋆] and Mérouane Debbah[◇],

[◇]CentraleSupélec, Gif-sur-Yvette, France

[⋆]Centre for Wireless Communications, University of Oulu, Finland

{ejder.bastug, merouane.debbah}@centralesupelec.fr, bennis@ee.oulu.fi

This research has been supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering), the SHARING project under the Finland grant 128010 and the project BESTCOM.

Proactive Caching in 5G Small Cell Networks

Abstract

Massive deployment of small-cell base stations (SBSs) is going to play a key role for capacity and coverage enhancements in 5G networks. However, the backhaul for these networks remains one of the important issue to solve. Ideally, the capacity of backhaul has to be in the same order of wireless links in order to avoid bottlenecks in the delivery and sustain huge traffic generated by mobile users, especially due to video streaming and content sharing in social networks. In reality, the deployment of such high-speed backhauls is not straightforward due to its costly nature. Thus, one promising way of tackling this backhaul bottleneck and satisfying users' demand is to cache the strategic contents at the edge of the network, namely at the SBSs and user terminals (UTs). So far, most of the existing solutions are based on the *reactive* networking paradigm in which users' content requests are served immediately upon their arrival or causing outages otherwise. In this chapter, we first provide an overview for recent research in small cell networks (SCNs), and then we explore the novel paradigm of *proactive* caching in SCNs that leverages the latest developments in storage, context-awareness, and social networking. With this approach, we show that important gains can be obtained, with backhaul offloadings and higher ratios of satisfied users reaching up to 22% and 26%, respectively.

I. Small Cell Networks: Past, Present and Future Trends

Smartphones have exponentially increased the traffic load in current cellular networks showing no signs of slowing down [1], [2]. It is now well understood that a very effective way to increase network capacity is making cells smaller by reducing the distance to the users [3]. Indeed, cell densification has gone from the order of hundreds of square kilometers (back in the eighties) to a fraction of a square meter or less with the advent of hotspots. There has been recently a great interest to deploy relays, distributed antennas and small cellular access points (such as micro/pico/femto cells) in residential homes, subways, enterprises, and hot-spot areas. These network architectures, which are either operator-deployed or user-deployed are referred to heterogeneous networks (HetNets) or small cell networks (SCNs) [3], [4]. By deploying additional network nodes within local-area range and making the network closer to end-users, small cells can significantly improve spatial reuse and coverage, boost capacity, and offload traffic more efficiently [4].

There exists a comprehensive literature on the topic of HetNets and SCNs tackling various aspects from interference management, cell association, stochastic network modeling, inter-cell interference coordination (ICIC), energy-efficiency, self-organizing networkss (SONs), mobility management, LTE/Wi-Fi interworking, among others (see [4] for a comprehensive survey). One of the key take-away drawn from these studies is that tight interference coordination among macro and femto/picocell tiers is necessary for achieving cell splitting gains. This hinges on the availability of low-latency and high-capacity backhauls [5]. Network modeling approaches based on stochastic geometric tools have shown reasonably-close performance gains (i.e., lower bound) in terms of system-wide and per-user capacities. Their attractive feature is attributed to the fact that unlike time-consuming system-level simulations, fundamental insights can be gleaned from these tools, some of which have been corroborated by industry field trials and observations from detailed simulations [6]. In parallel to that, mobility management has received significant attention from the wireless industry, research community, and standardization bodies [7]. Unlike conventional homogeneous networks where user terminals (UTs) typically use the same set of handover parameters (i.e., hysteresis margin, time-to-trigger (TTT), etc.), using the same set of handover parameters in HetNets for all cells and/or for all UTs may degrade mobility performance. This is because high-mobility macro UTs may run deep inside coverage areas of small cells before the TTT optimized for macrocells expires, thus incurring handover failure (due to degraded signal-to-interference-plus-noise ratio (SINR)) [8]. Decentralized interference management/mitigation strategies in co-channel interference scenarios have also been studied in details, whereby small cells are able to self-organize based on local information and optimize their transmission strategies (i.e., power/frequency) based on minimum information exchange [9]. This leads to a number of tradeoffs in terms of faster/slower convergence at the cost of partial/full information. Carrier aggregation (CA) and its single/multiflow enhancements have also been investigated as a means of further boosting network capacity and per-user throughput, in which users may be served on several bands simultaneously [10]. Furthermore, with the increasing traffic asymmetry in the uplink (UL) as compared to the downlink (DL), novel cell association mechanisms and architectures are needed to cope with new types of inter-node interferences (DL-to-UL), thereby opening new avenues for research such as flexible DL/UL communication, massive multiple-input multiple-output (MIMO), device-to-device (D2D), full-duplexing, etc. [3] [11]. Finally, the topic of LTE and Wi-Fi coexistence has received tremendous attention due to the

multi-mode capability of small base stations (SBSs)[1] and the possibility of using both licensed and unlicensed bands. Therein, dynamic load balancing and traffic steering mechanisms have been proposed leveraging the availability of Wi-Fi for best-effort services, traffic load, delay tolerance, etc [12].

While small cell densification is clearly the way to go, a number of technical challenges remain unsolved. Indeed, while small cell densification was shown to boost capacity, simply adding small cells may turn out to be energy-inefficient [13]. In addition, backhaul optimization and the optimal location of small cells represent one of the main limiting factors before a full rollout of small cells takes place. The importance of the backhaul is further underscored with the unabated proliferation of smartphones with the vast array of new wireless services (i.e., multimedia streaming, web-browsing applications, etc.). As a result, novel approaches to backhaul-aware small cell networking have been recently proposed in the literature [14] such as how to optimally decouple control and data planes to make cells more adaptive to traffic dynamics and network state while having a global view of the network, backhaul offloading via smart edge caching [15]–[17], cloud radio access network (C-RAN) [18], software defined networking (SDN) [19], resource/network virtualization, ultra-dense networks, massive MIMO, etc. Among these approaches, in this chapter, we focus on proactive edge caching as a way of dealing with backhaul offloading in SCNs, which is especially crucial in dense deployments.

Rest of this chapter is organized as follows. We give an introduction of cache-enabled proactive SCNs in Section II. Our system model and corresponding problem formulation is presented in Section III. The details of proactive caching at the SBSs and UTs are given in Sections IV and V respectively and discussions of numerical results are carried out in the same sections. The current directions of caching in wireless networks and relevant works are discussed in Section VI. Finally, Section VII draws some conclusions and future work.

II. Cache-Enabled Proactive Small Cell Networks

Most of the existing studies in SCNs are so far based on the classical networking paradigm, called as *reactive*, in which users' content requests are served immediately or yielding outages otherwise. In such a situation, sustaining peak traffic demands in these networks requires expensive high-speed backhaul, resulting in tremendous operational expenditures (OPEX). Given

[1]The term "SBS" will be used interchangeably with "small cell" in this work.

the fact that such a cost may not be affordable, a *novel* networking paradigm is clearly needed for densely deployed SCNs. This can be done by exploiting recent advances in storage, context-awareness, social networking and D2D [17].

This novel network paradigm is *proactive* in the sense that the nodes at the edge of the network (i.e., SBSs and UTs) predicts users' context information and pre-store intelligently strategic contents, in order offload the backaul and satisfy users' quality-of-service (QoS). This goes beyond the scope of traditional cellular networks which they have been designed assuming *dumb* UTs with limited storage and processing features. Nowadays, UTs are much more sophisticated than before, giving the opportunity to exploit their capabilities in conjunction with SCNs by storing the predicted content at the network edge. This in turn yields significant gains in terms of network resources, minimizing operational and capital expenditures [3].

The fact that the huge amount of users' information is often available and the human behaviour has a certain predictability [20], users' future events can be inferred. Therefore, in this chapter, we explore such a proactive caching framework by leveraging context-awareness and storage capabilities at the edge of the network in order to sustain peak data demands and offload the backhaul. More precisely, estimating users' future demands and content popularity can be used to proactively store the content before the actual requests take place. In addition, whenever a D2D communication is available, the proactive caching approach exploits users' social relationships (and their influence within the social community), as well as users' storage for content dissemination and physical proximity.

As stated before, recent results have shown that the human behaviour is correlated and predictable to a large extent [20]. Therefore, SBSs are assumed to be equipped with storage units and the low-speed backhaul is used for their broadband connections. Then, as will be shown, proactively caching users' content at SBSs alleviates the backhaul load and incurs higher users' satisfaction. The proactive caching procedure is based on the idea of storing the popular content at the SBSs. To achieve this, the popularity of the content has to be estimated. Using tools from machine learning and analysing the infrastructure logs (such as in [21]), a trove of hidden information about users' behaviour can be revealed. Analysing these traces falls into the *big data* phenomenon where collaborative filtering (CF) methods can be successfully applied for inference.

Yet another approach for bringing contents at the edge of the network is via caching at the

users' devices, and leveraging D2D communications for content dissemination. Online social networks (Facebook, Twitter, Digg) have become instrumental in disseminating various contents across social communities [2]. Typically, users tend to value highly recommended contents by their friends or people with similar interests. Thus, exploiting users' social relationships, and proactively storing the content in users' devices can alleviate peak traffic demands. Notably, the strategic contents in the caches of popular/influential users can ease backhaul congestion and yield considerable network savings. In order to show such network savings, we first detail our system model in the following section.

III. System Model

Let us consider a scenario that consists of M SBSs $\mathcal{M} = \{1, ..., M\}$ and N UTs $\mathcal{N} = \{1, ..., N\}$. The broadband connection of every SBS $m \in \mathcal{M}$ is provided by a central scheduler (CS) via a limited backhaul link with capacity c_m.[2] We suppose that the capacity of the wireless small cell link between SBS m and UT n is given by $c_{m,n}$. Depending on the content availability and users' proximity, the SBSs can establish D2D communications between users n and n', whereas the corresponding D2D link capacity is denoted by $\ddot{c}_{n,n'}$. This scenario is illustrated in Fig. 1. Suppose that user n requests a content from a library of F contents, represented by $\mathcal{F} = \{1, ..., F\}$, according to probabilities $\mathcal{P}_n = \{p_{n,1}, ..., p_{n,F}\}$. In this library, the length of contents are $\mathcal{L} = \{l_1, ..., l_F\}$ and the bitrate are given by set of $\mathcal{B} = \{b_1, ..., b_F\}$. Now, suppose that R number of content requests are drawn by users randomly during T time slots. Then, we say that a request $r \in \mathcal{R} = \{1, ..., R\}$ is *satisfied* if the rate of delivery is equal or greater than the bitrate of the requested content as follows:

$$\frac{l_r}{t'_r - t_r} \geq b_r, \tag{1}$$

where $l_r \in \mathcal{L}$ represent the length of the requested content, t_r (t'_r) is the start (end) time of the delivery, and $b_r \in \mathcal{B}$ is the bitrate of the content $f_r \in \mathcal{F}$. Given this definition, the *satisfaction ratio* can be expressed as:

$$\eta(\mathcal{R}) = \frac{1}{R} \sum_{r \in \mathcal{R}} \mathbb{1} \left\{ \frac{l_r}{t'_r - t_r} \geq b_r \right\}, \tag{2}$$

[2]This controller is typically a network entity located at the evolved packet core (EPC) or at the network edge (small cell gateway)

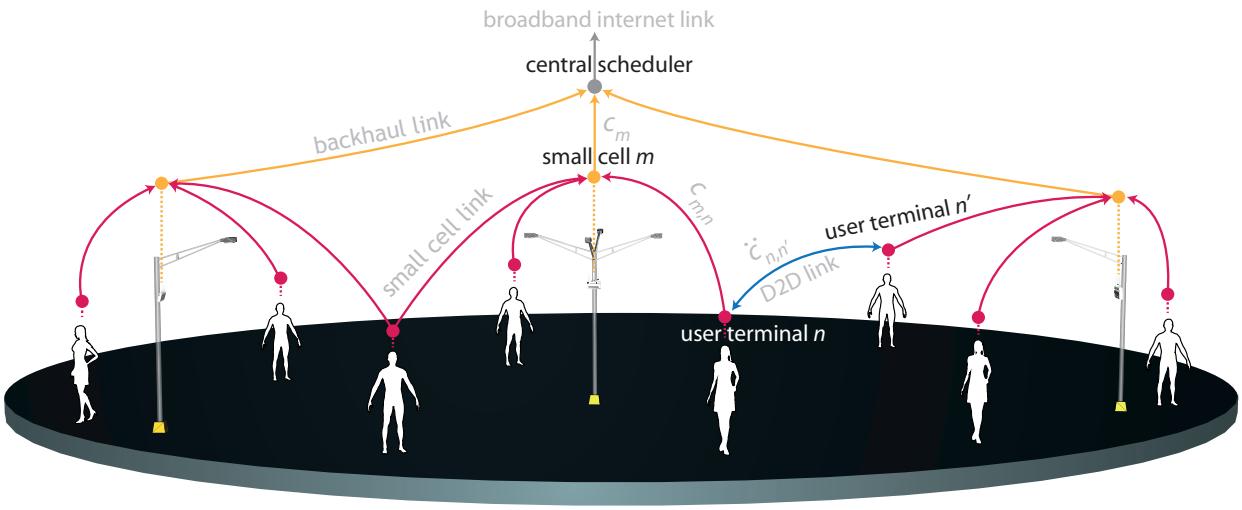


Figure 1: A sketch of the scenario given in the system model. A central scheduler is in charge of providing broadband connection to M SBSs via backhaul links. Depending on the users' content availability in the caches of SBSs and UTs, the SBSs serve their user either via wireless small cell links or D2D communications.

where $\mathbb{1}\{...\}$ is the indicator function which yields 1 when the condition holds and 0 otherwise.

Our target as the network operator is to keep the satisfaction ratio above a threshold, while minimizing the usage of the backhaul. As stated before, this can be done via proactive caching in SBSs and UTs, in which we detail these two case studies separately in the following sections.

IV. Proactive Caching at Base Stations

Results have shown that the backhaul constitutes one of the most important challenges for SCN deployments and this is going to increase dramatically due to the densely deployed SBSs. From this observation, suppose that the total capacity of the backhaul is lower than the available wireless link capacity between SBSs and UTs, such as $\sum_{m\in\mathcal{M}} c_m \ll \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} c_{m,n}$. Since in this case we suppose that the backhaul is the bottleneck, one reasonable option is to avoid its usage by storing the users' content proactively at the SBSs, during peak-off hours. In other words, if the users' content can be stored at SBSs before the users' actual contents arrives, the backhaul will not be used for a certain level, depending on on how smartly the content is placed.

Let us consider that the rate of the backhaul link during the content delivery for request r at time t is $\lambda_r(t)$. Then, the *backhaul load* under given these definitions can be expressed as

follows:

$$\rho(\mathcal{R}) = \frac{1}{R}\sum_{r\in\mathcal{R}}\frac{1}{l_r}\sum_{t=t_r}^{t=t'_r}\lambda_r(t). \tag{3}$$

Additionally, suppose that the storage capacity of SBS m is given by s_m and the amount of its consumption at time t is denoted by $\kappa_m(t)$. Hence, the backhaul minimization problem subject to the link capacities, storage and QoS constraints can be formulated as follows:

$$\begin{aligned} \underset{t'_r, r\in\mathcal{R}}{\text{minimize}} \quad & \rho(\mathcal{R}) && \\ \text{subject to} \quad & \lambda_r(t) \le c_m, && \forall m \in \mathcal{M}, \\ & \kappa_m(t) \le s_m, && \forall m \in \mathcal{M}, \\ & \eta(\mathcal{R}) \ge \eta_{min}, && \forall r \in \mathcal{R}, \end{aligned} \tag{4}$$

where η_{min} is the target satisfaction ratio. Since dealing with (4) is computationally intractable, a heuristic approach similar to the one in [22] can be performed by storing users' popular content in the cache of SBSs. Before such a caching procedure is applied, we suppose that each SBS m has to track, learn and build its user' content profile to infer their future demands. Assume that $\mathbf{P}_m$ is the discrete content probabilities of users in SBS m in which we refer as *popularity matrix*, each row representing the users and columns are content popularities/ratings. Indeed, a perfectly known $\mathbf{P}_m$ could easily allow us to store the content according to this caching procedure. Unfortunately, this situation in practice is not the case, in which the matrix is not perfectly known, large and indeed sparse. Given these observations and inspired from the *Netflix paradigm* [23], supervised machine learning tools can be used to exploit users-content correlations. Inferring the probability that user n requests content f (namely estimating the popularity matrix), and storing the predicted content accordingly can clearly offload the backhaul.

The proposed proactive caching procedure is composed of training and placement steps. The first step is the training step in which each SBS m builds a model for the popularity matrix $\mathbf{P}_m$ based on the available information. The estimation of $\mathbf{P}_m$ boils down to solving a least square minimization problem as follows:

$$\min_{\{b_n, b_f\}} \sum_{n,f}\Big(r_{nf} - \hat{r}_{nf}\Big)^2 + \lambda\Big(\sum_n b_n^2 + \sum_f b_f^2\Big), \tag{5}$$

where the sum is over the (n, f) user/content pairs in the training set, containing how user n rated content f (i.e., r_{nf}). The total number of users in the training set is N and F is the total number

of contents, thus, the minimization is done over all the $N + F$ parameters. In this formulation, $\hat{r}_{nf} = \bar{r} + b_n + b_f$ is the baseline estimator where b_f is the relative quality of each content f compared to the average $\bar{r}$. The bias of each user n relative to b_n is given by $\bar{r}$. Additionally, the parameter λ is used for balancing the regularization and fitting the training data.

In the numerical setup, we use the regularized singular value decomposition (SVD) due to its numerical accuracy (see [24] for comprehensive study of CF methods). Roughly speaking, since the entries of $\mathbf{P}_m$ are not fully known, the model construction is done via gradient descent by using the least-squares property of the SVD. Thus, $\hat{\mathbf{P}}_m$ is constructed as the low rank version of $\mathbf{P}_m$.

So far, we have described the first step. In the last step (namely, the placement step of the caching procedure), the content is cached proactively by storing the most popular content based on the estimation of $\hat{\mathbf{P}}_m$, until the storage capacity is fulfilled. In the following, we show the gains of proactive caching in a numerical setup and discuss the impact of various parameters of interest. A sketch of the proactive caching procedure at the base stations is summarized in Fig. 2.

A. Numerical Results and Discussions

The list of parameters used in the numerical study is provided in Table I. In order to see the impact of the parameters of interest, the length and bitrate of the content, wireless small cell links and storage capacities are set to the identical values. We consider three regimes of interest: (i) low load, (ii) medium load, and (iii) high load.

In the numerical study, R number of requests are drawn over a time duration T, given the fact that the arrival times of these requests are sampled uniformly at random. The users' content requests are drawn from the ZipF(α) distribution. Given that knowledge, at $t = 0$, the perfect popularity matrix $\mathbf{P}_m$ is constructed for each SBS m. Removing 20% of the entries of this matrix uniformly at random, the remaining entries are used for the model construction in CF. The prediction of missing entries are then carried out by the regularized SVD [25]. Once the popularity matrix is estimated, the proactive caching is applied by greedily storing the most popular content subject to the storage size of the SBS. In the numerical setup, afer completing the training and placement steps of the proactive procedure at $t = 0$, the users' are served depending on their request arrival time until all content delivery processes finish. We use random caching

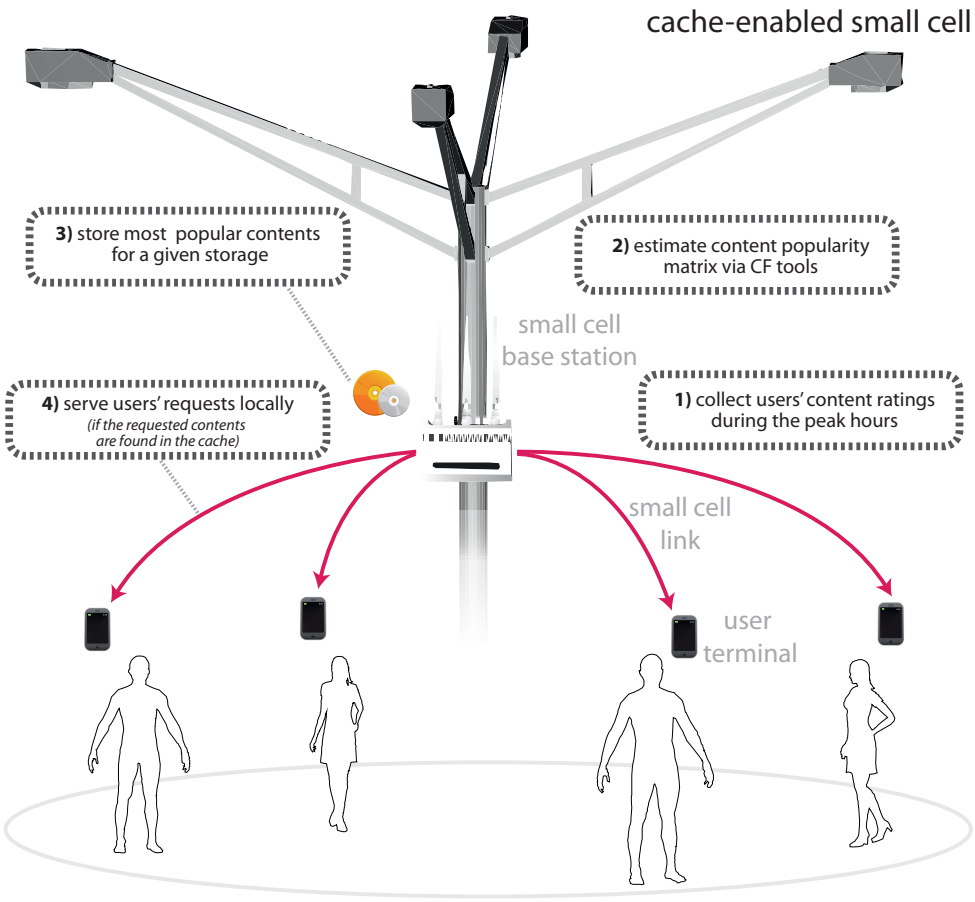


Figure 2: A practical procedure for proactive caching at the base stations.

as a baseline referred to as *reactive*.

In order to compare the benefits of caching both for proactive and reactive cases, three parameters of interest are detailed: (i) number of requests R, (ii) total cache size S, and (iii) ZipF distribution parameter α. The gains in the plots are normalized for ease of understanding. The evolution of the satisfaction ratios and the backhaul loads with respect to the variation of these parameters are given in Fig. 3.

In the figures, we see that the satisfaction ratio decreases as the number of users' content requests increases. The reason is somewhat obvious as the capacity constraints starts to be limiting factor for the delivery of high amount of requests. Concerning the backhaul load in very small number of requests, the reactive approach is generating less load compared to the proactive case which can be explained by *cold start* phenomenon of the CF used in the proactive

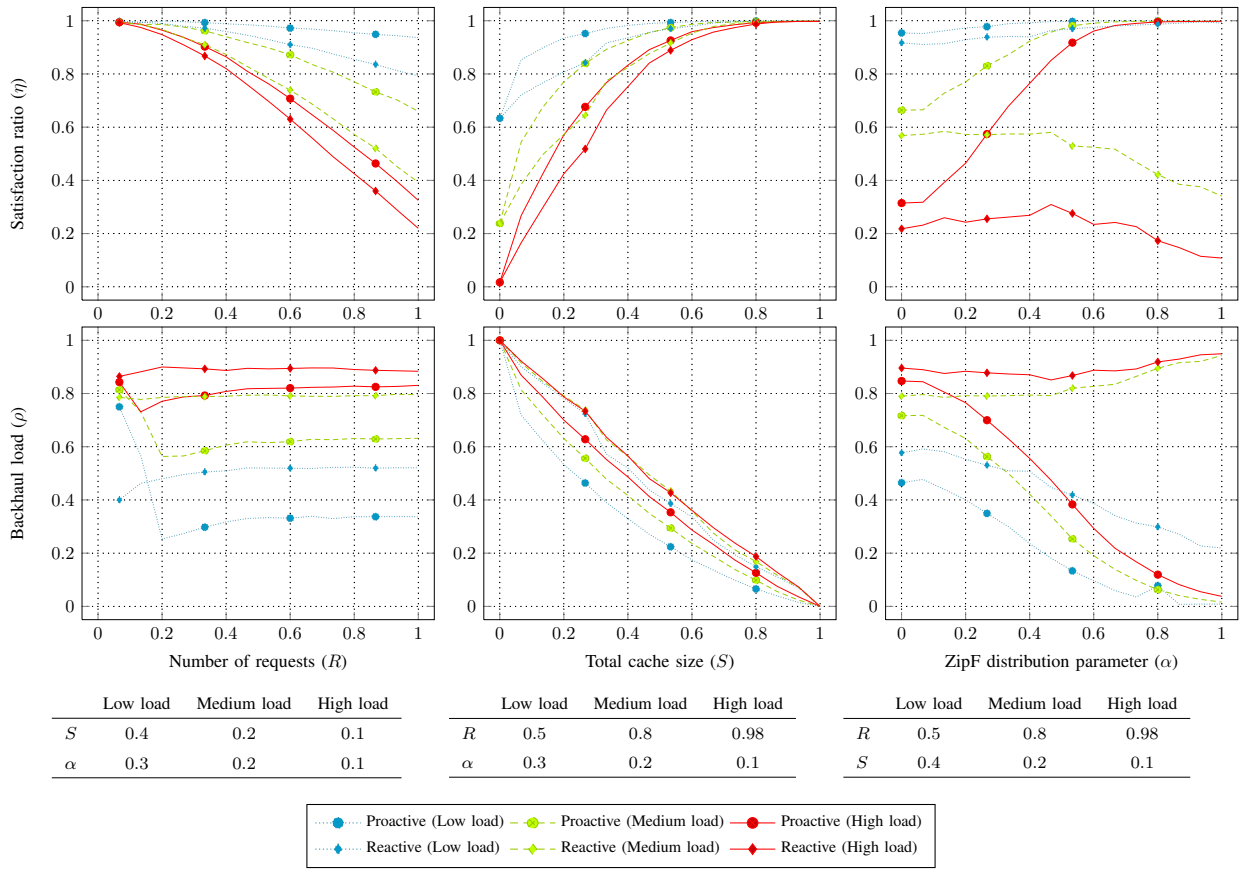
Parameter	Description	Value
T	Time slots	1024 seconds
M	Number of SBSs	4
N	Number of UTs	32
F	Number of contents	128
l_f	Length of content f	1 Mbit
b_f	Bitrate of content f	1 Mbit/s
$\sum_m c_m$	Total backhaul link capacity	2 Mbit/s
$\sum_m \sum_n c_{m,n}$	Total wireless small cell link capacity	64 Mbit/s
R	Number of requests	$0 \sim 2048$
S	Total cache size	$0 \sim l_f \times F$
α	ZipF parameter	$0 \sim 2$

Table I: The numerical setup parameters for proactive caching at the SBSs.

case. However, as the number of request increases, the amount of information given to the CF for training step increases. Therefore, in the end, the proactive approach with sufficient amount of information outperform the reactive approach with an almost constant gain.

One important parameter of interest in our scenario is the total storage size of SBSs. As we increase the storage, the SBSs gains more capability to store the content from the catalog, yielding the satisfaction ratio up to 1 and backhal load up to 0 in the extreme values of the storage size. Looking at more practical situations in which the storage size is somewhere between 0 and 1, we see that the proactive approach outperforms the reactive case in terms of the satisfaction ratio as well as the backhaul load.

The content popularity parameter α indeed has an impact on the performance metrics. In the low values of α where the distribution follows a uniform behaviour, the proactive approach outperforms the reactive case with a relatively low difference. However, as the α increases, a few amount of content become highly popular than the rest of the content in the catalog. Thus, the difference between the gain of proactive and reactive approaches become quite visible in terms of the satisfaction ratio and the backhaul load.



	Low load	Medium load	High load
S	0.4	0.2	0.1
α	0.3	0.2	0.1

	Low load	Medium load	High load
R	0.5	0.8	0.98
α	0.3	0.2	0.1

	Low load	Medium load	High load
R	0.5	0.8	0.98
S	0.4	0.2	0.1

Figure 3: Backhaul Offloading via Proactive Caching: Dynamics of the satisfied requests and backhaul load with respect to the number of requests, total cache size and ZipF parameter.

V. Proactive Caching at User Terminals

Yet another mean of offloading the traffic at SBSs (thus, offloading the backhaul as a consequence) can be achieved by caching users' content at the UTs and exploiting D2D communications for content dissemination. For this purpose, the interplay between users' social ties and physical proximity can be taken into account for proactive caching decision. In particular, when a content request arrives to the network, the SBS can take benefit of the influential users who have the content, requesting them to join the content delivery via D2D opportunities. If such a opportunity does not exist and the requested content is not available, as a last resort, the content can be delivered by the SBS but with the cost of using the backhaul.

Let us consider that the storage capacity of UT n is $\ddot{s}_n$ and its usage at time t is given by $\ddot{\kappa}(t)$. Also suppose that $\dot{\lambda}_r(t)$ is the total rate of the SBSs during the content delivery of request

r at time t and the D2D link rate is $\ddot{\lambda}_r(t)$. Then, *small cell load* can be expressed as follows:

$$\ddot{\rho}(\mathcal{R}) = \frac{1}{R} \sum_{r \in \mathcal{R}} \sum_{t=t_r}^{t=t'_r} \frac{\dot{\lambda}_r(t)}{\dot{\lambda}_r(t) + \ddot{\lambda}_r(t)}. \tag{6}$$

Given that definition and using a formulation similar to (4), the D2D caching optimization problem can be written as:

$$\begin{aligned}
\underset{t'_r, r \in \mathcal{R}}{\text{minimize}} \quad & \ddot{\rho}(\mathcal{R}) && && (7)\\
\text{subject to} \quad & \dot{\lambda}_r(t) \leq c_{m,n}, && \forall m \in \mathcal{M}, \forall n \in \mathcal{N},\\
& \ddot{\lambda}_r(t) \leq \ddot{c}_{n,n'}, && \forall (n, n') \in \mathcal{N},\\
& \ddot{\kappa}_n(t) \leq \ddot{s}_n, && \forall n \in \mathcal{N},\\
& \eta(\mathcal{R}) \geq \eta_{min} && \forall r \in \mathcal{R}.
\end{aligned}$$

According to our scenario, the first step for solving (7) is to infer the set of influential users. This, as mentioned before, is done via the notion of *centrality* metric [26]. In general, the centrality measures is used to quantify the social influence of a node in the network and also related to how the node is well connected. A node with higher value of this measure in turns means that such a node is more central (thus influential) than the nodes who have lower values of this measure. Several definition of centrality metrics exist on literature [26], whereas we only focus on the eigenvector centrality for exposition. Let $G = (\mathcal{N}, \mathcal{E})$ be the social graph which consists of N nodes/users, where $\mathcal{N}$ represents the set of nodes and $\mathcal{E}$ is set of the links between them. We now that, the graph G can be represented by its adjacency (or D2D connectivity) matrix $\mathbf{A}_{N \times N}$, where the entry $a_{n,n'}$, $n, n' = 1, ..., N$ is 1 if link (or edge) $\ddot{c}_{n,n'}$ exists, or 0 otherwise. For this matrix, let the eigenvalues to be represented by $\lambda_1 \geq ... \geq \lambda_N$ in decreasing order, and the corresponding eigenvectors of these eigenvalues be given by $\mathbf{v}_1, ...\mathbf{v}_N$. The eigenvector-centrality in this case is basically the eigenvector $\mathbf{v}_1$ that has the largest eigenvalue λ_1. Knowing K-most influential users of the social network via notion of centrality, a clustering method (i.e., K-means [27]) can be then formed around the users for community formation.

Once the set of influential users is identified and their communities are formed, the next step is to analyze the content dissemination within each social community. By doing so, the critical content of each community can be stored in the cache of influential users. To show this, suppose that there is a set number of available contents, denoted by $\mathcal{F} = \mathcal{F}_0 + \mathcal{F}_h$, where $\mathcal{F}_h$ is

the set of contents with viewing history and $\mathcal{F}_0$ represents the set of contents without history. We further assume that each user is interested to only one type of available contents $\mathcal{F}$. Let π_f be the probability that content f is chosen by a given user, and as a prior [28], assume that the distribution follows a Beta distribution [28]. Then, the selection of user n given as the conjugate probability of the Beta distribution has a Bernoulli distribution. This in turn shows that the resulting user-content partition is analogous to that of the Chinese restaurant process (CRP) [28]. The CRP is a metaphor in which the objects are customers in a restaurant, and the classes are represented by the tables which the customers sit. More precisely, in CRP, there exist a restaurant with a large number of tables, each with infinite number of sets, and customers arrive sequentially each of them choosing a table at random.

In the CRP with concentration parameter β, each customer decides to occupy a table with a probability proportional to the number of occupiers of that table, and chooses the next available table with proportional to the parameter β. Being more specific, the first customer selects the first table with probability $\frac{\beta}{\beta} = 1$. The second customer selects the first table with probability $\frac{1}{1+\beta}$, and the second table with probability $\frac{\beta}{1+\beta}$. Once the second customer selects the table, in the next, the third customer selects the first table with probability $\frac{1}{2+\beta}$, the second table with probability $\frac{1}{2+\beta}$ and the third table with probability $\frac{\beta}{2+\beta}$. This selection process continues until all customers have seats, yielding a distribution over allocation of customers to tables. In this process, the decision of subsequent customers are affected by the feedback of previous customers, where customers learn the previous selections to update their beliefs and probabilities in which they select the tables.

From this point, the behaviour of the content dissemination in the social network is similar to the table selection in an CRP. Looking to the social network as a Chinese restaurant, the contents as the large number of tables and the users as the customers, we can model the content dissemination process by an CRP. This means that, within each social community, users intend to request the sought-after content sequentially, and once a content is downloaded, a hit is recorded (i.e., history). This, in turn, changes the probability that this content will be requested by others within the same social community, where popular contents will be requested more frequently and new contents less frequently. Suppose a random binary matrix $\mathbf{Z}_{N\times F}$, indicating the selection of contents by users, where $z_{nf} = 1$ if user n chooses content f and 0 otherwise. Then, we can

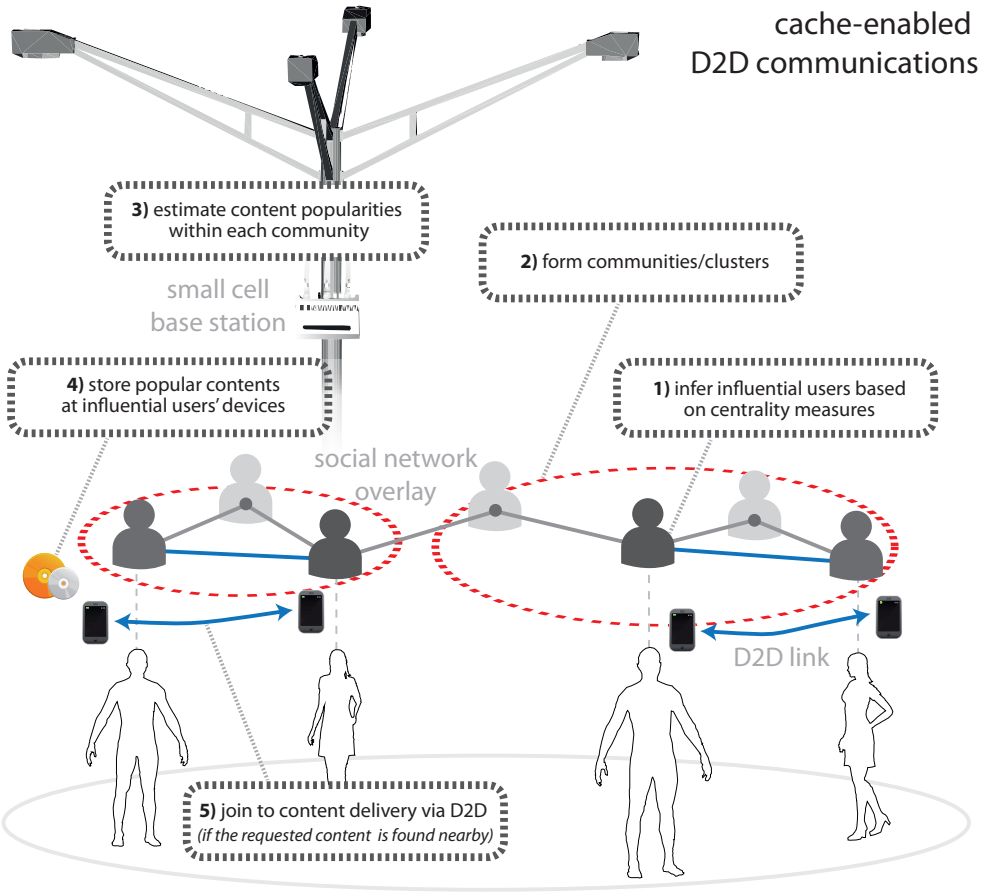


Figure 4: A practical procedure for proactive caching at the user terminals.

show that [28]:

$$P(\mathbf{Z}) = \frac{\beta^{F'}\Gamma(\beta)}{\Gamma(\beta+N)}\prod_{f=1}^{F'}(m_f-1)! \tag{8}$$

where $\Gamma(.)$ is the Gamma function [29], m_f is the number of users already assigned to content f (i.e., viewing history) and F' is the number of partitions with $m_f > 0$. Therefore, for a given $P(\mathbf{Z})$, the popular contents of each community can be stored inside the cache of influential users. A sketch of the proactive caching procedure at the user terminals is summarized in Fig. 4.

A. Numerical Results and Discussions

In the numerical setup, for similar purposes as in the previous section, the wireless link capacities are assumed to be equal among the users. The total D2D link capacity of each user

is shared among the number of social links. The list of parameters are given in Table II.

Parameter	Description	Value
T	Time slots	1024 seconds
M	Number of SBSs	4
K	Number of communities	3
N	Number of UTs	32
F	Number of contents	128
l_f	Length of content f	1 Mbit
b_f	Bitrate of content f	1 Mbit/s
$\sum_m \sum_n c_{m,n}$	Total SBSs link capacity	32 Mbit/s
$\sum_n \sum_{n',n' \neq n} \ddot{c}_{n,n'}$	Total D2D link capacity	64 Mbit/s
R	Number of requests	$0 \sim 9464$
S	Total D2D cache size	$0 \sim l_f \times F$
β	CRP concentration parameter	$0 \sim 100$

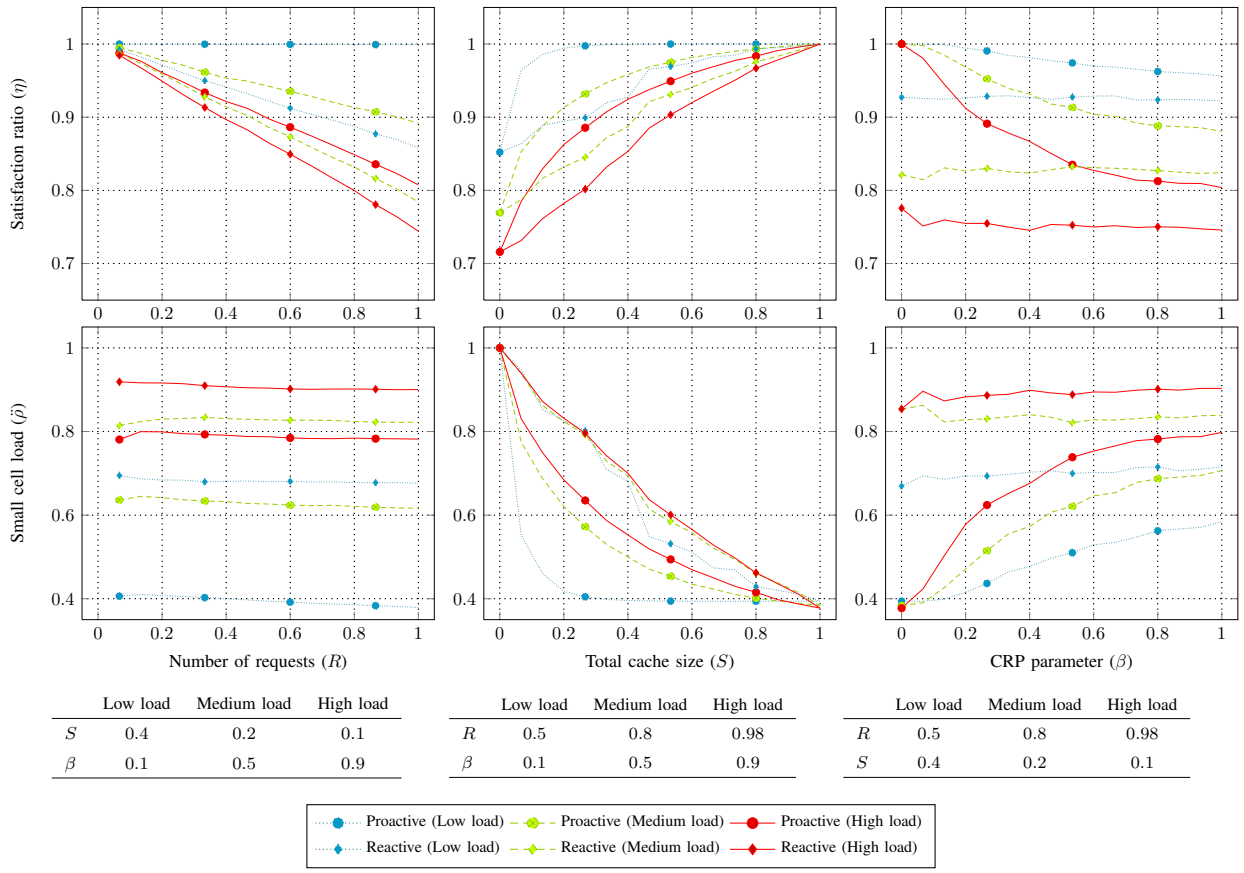
Table II: The numerical setup parameters for proactive caching at the UTs

Starting from $t = 0$, request arrival times are drawn uniformly at random until the time T. The social network is constructed by using the preferential attachment model [30]. As states before, the eigenvector centrality is used to quantize the influential users in the social network, then, K-most influential are formed into K communities via K-means clustering [27]. In each community, the content popularity distribution is sampled from the CRP(β). Given the content popularity, the proactive caching is done by storing the popular files greedily inside the influential users until no storage space remains. Similar to the case study in previous section, random caching is used as a baseline.

Parameters of interests in this case are: (i) number of requests R, (ii) total D2D cache size S, and (iii) CRP concentration parameter β. The results are normalized for ease of understanding. The impact of parameter of interests on the satisfaction ratio and small cell load are given in Fig. 5.

In the figure, increasing the number of requests, we see that the satisfaction ratio decreases rapidly and the small cell load decreases at a low pace. The gains of proactive caching approach are higher than the reactive approach in all regimes.

When an increment of D2D size is the case, we observe an increment in the satisfaction ratio



	Low load	Medium load	High load
S	0.4	0.2	0.1
β	0.1	0.5	0.9

	Low load	Medium load	High load
R	0.5	0.8	0.98
β	0.1	0.5	0.9

	Low load	Medium load	High load
R	0.5	0.8	0.98
S	0.4	0.2	0.1

Figure 5: Social-Aware Caching via D2D: Dynamics of the satisfied requests and small cell load with respect to the number of requests, total cache size and CRP concentration parameter β.

and decrement in the small cell load. Even though both proactive and reactive cases have the gains, the proactive approach has more desirable performance compared to the reactive approach.

The concentration parameter β has also an impact on the performance. When β increases (i.e., the number of distinct contents grows), the satisfaction ratio and the small cell loads tends to be almost constant in the reactive approach. On the other hand, as β increases, the satisfaction ratio in the proactive approach decreases and the small cell increases. The performance gap between the proactive and reactive approaches gets closer and closer as β increases. This is due to the facts that the contents catalog size is growing while UTs having a limited cache size.

VI. Related Work and Research Directions

In this work, we have highlighted the proactive caching framework given in [17]. Indeed, the idea of caching goes back to the sixties in the context of algorithm design in operating systems [31]. According to [31], the optimal content removing strategy in the case of new content arrival is to evict the content from the memory which is not going to be requested in the near future. Beside this line of work, there has been also extensive studies on web caching schemes in the past decades, aiming to improve the scalability of world wide web and offloading the network, by caching content in the proxy servers and/or intermediate nodes of the network (see [32] for a brief literature). Numerous caching algorithms for content delivery network (CDN) have emerged in the recent years [33], allowing content providers to reduce access delays to the requested contents. Conceptually, there exist also information-centric networks (ICNs) which aims to change the way of accessing the content on the internet, by uniquely naming the contents and smartly distribute these across the network, rather than traditionally having one source for the content access [34] (see also [35] for a recent survey). Beside these line of works, the caching problem as a way of offloading the wireless communications infrastructure is recent. Similar to what we have presented here, the growing literature is mostly based on caching at the edge of network. In the following, we summarize some of these works based on their similarities and directions.

A. Proactive Caching and Content Popularity Estimation

Proactive caching in SCNs with perfect knowledge of the content popularity is given in [22]. In [17], exploiting context-awareness, social networks, D2D communications, the proactive caching approaches for SCNs are studied both at the SBSs and UTs, showing that several gains are possible under the given numerical setup. Therein, instead of perfect knowledge of the content popularity, an estimation is done via machine learning tools (the CF in particular), by exploiting correlations of human behaviour on their preferences. Thus, having such an estimation, the caching decision is applied more efficiently, yielding better performance in terms of the users' satisfaction and offloading of the network. On the other hand, a well-known problem in the CF literature is the cold-start problem which can occur in the case of estimation with very few amount of information. Therefore, to boost the content popularity estimation, one approach harnessing the machine learning literature is *transfer learning*, based on the idea of *smartly* transferring

information from a target domain to a source domain (see [36] for a survey). Inspired from this, a preliminary study on transfer learning for caching in SCNs is conducted in [37]. Even though it has naturally its own challenges (i.e., negative transfer), it is shown in [38] that the content popularity estimation via CF can be improved by this approach. Further investigations are needed to combine this approach with the proactive caching in SCNs. Additionally, in the context of proactive caching, the centrality measures for the content placement are exploited in [39]. Therein, a simple content dissemination process is introduced and the preliminary performance results of this centrality-based content placement methods are given via numerical simulations. Alternative to these proactive approaches, a game theoretical formulation of the proactive caching problem as a many-to-many matching game is introduced in [40]. A matching algorithm that reaches a pairwise stable outcome is provided for the caching problem, showing that the number of satisfied requests can be reach up to three times the satisfaction of a random caching policy.

B. Approximation Algorithms

The idea of *femtocaching* is given in [16], in which the SBSs (helpers) with low-rate backhaul but high storage units are in charge of delivering the content to the users via short-range transmissions. The analysis is carried out both for coded and uncoded cases, showing that the optimum content assignment is NP-hard, whereas the coded case is formulated as a convex problem that further can be reduced to a linear program. A greedy algorithm for coded case and numerical results are provided, showing that video throughput can be improved by a factor $3-5$ in realistic settings. Extensions to this work, including D2D case, is given in [41], [42]. Alternatively, a multicast aware caching problem is formulated in [43] and a heuristic algorithm is provided for that purpose, showing that servicing cost can be reduced down to 52% compared to the multicast-agnostic case case.

Optimal content placement in a SBS with limited backhaul capacity is also studied in [44], showing that the problem can be reduced to a knapsack problem when the content popularity distribution is known. Assuming that the content popularity distribution is not known in advance, the problem is formulated as a multi-armed band (MAB) problem so that the content popularity distribution can be learned online and content placement can be done. Three different caching algorithm is provided to show the exploration vs. exploitation trade-offs of this problem. As an extension, a derivation of regret bounds and more extensive analysis of the algorithms through

numerical simulations are presented in [45]. Additionally, a distributed caching model with multiple SBS is given in [46] in the framework of MAB problem, showing that coded caching can outperform the uncoded case. Beside MAB approaches, an approximation framework based on the facility location problem is given in [47]. Also, for a given traffic demand, a distributed caching algorithm based on Alternating Direction Method of Multipliers (ADMM) is presented in [48].

C. Coded Caching Gains

Information-theoretic formulation of the caching problem is studied by [49]. Therein, local and global caching gains, which depend on the available memory of each user and cumulative memory of all users respectively, are derived based on a coded caching scheme. The proposed scheme consists of placement and delivery phases (i) is given for a *centralized* setup where the content placement is handled by a central server, (ii) is essentially *offline* as there is no content placement during the delivery phase, (iii) is shown to outperform conventional uncoded schemes under *uniform content popularities*, and (iv) works in a single shared link instead of *more general networks*. These results are then extended to non-uniform content popularities in [50], [51], non-uniform cache access in [52], heterogeneous cache sizes in [53], online caching systems in [54], hierarchical caching networks in [55] and multi-server case in [56]. Moreover, the improved bounds are given in [57], [58], delay-sensitive content case is studied in [59] and the information-theoretic security aspects are shown in [60]. With similar line to these works, a decentralized approach for D2D networks with random coded caching is studied in [61], [62] in terms of scaling laws where a protocol channel model similar to [63] is taken into account. In the same vein, the performance of decentralized random caching placement with a coded delivery scheme is given in [64], [65], where the expected rate is characterized for random demands with Zipf popularity distribution.

In the context of distributed storage systems and coding, the performance of simple caching, replication and regenerating codes is studied in a D2D scenario in [66], in which a simple decision rule for choosing simple caching and replication is derived for minimizing the expected total cost in terms of energy consumption. On the other hand, the study of the physical layer functionality of wireless distributed storage systems is given in [67] from point of space-time storage codes. Based on that work, a wireless storage system that communicates over a fading channel is studied in

[68] and a novel protocol for the transmission is proposed based on algebraic space-time codes, in order to improve the system reliability while keeping the decoding at a feasible level. It is shown that the proposed protocol performs better than the simple time-division multiple access (TDMA) protocol and falls behind the optimal diversity-multiplexing gain tradeoff (DMT). Alternatively, a triangular network coding approach for cache content placement is presented in [69], in which the uncoded content placement and the triangular network coding strategies are compared in a numerical setup. Additionally, a coded caching scheme over wireless fading channel is presented in [70], whereas [71] casts the caching problem into a multi-terminal source coding problem with side information.

D. Joint Designs

In terms of joint designs, a two time-scale joint optimization of power and cache control is given in [72] for cache-enabled opportunistic cooperative MIMO. First, for the short time scales, the closed-form expressions for the power control is derived from an approximated Bellman equation. Then, for the long time scales, the caching problem is translated into a convex stochastic optimization problem and a stochastic subgradient algorithm is provided for its solution. The proposed solution is shown to be asymptotically optimal for high signal-to-noise ratio (SNR) whereas its comparison with baseline approaches are done via simulations. Another mixed time-scale solution for cooperative MIMO is given in [73]. Therein, in order to minimize the transmit power under the QoS constraint, the MIMO precoding is optimized in the short time scale and cache control is done in the long time scale. Additional to these approaches, the joint optimization of cache control and playback buffer management for video streaming is given in [74]. The joint caching and beamforming for backhaul limited caching networks is studied in [75], and finally the joint caching and interference alignment (IA) in MIMO interference channel under limited backhaul capacity is presented in [76].

E. Mobility

Mobility aspects of coded content delivery is analyzed in [77] based on a discrete-time Markov chain model. In order to minimize the probability of using the main base station in this model, a distributed approximation algorithm based on large deviation inequalities is introduced and numerical experiments on a real world dataset are conducted for the proposed algorithm. Another

caching scheme that exploits users' mobility is given in [78], in which the influence of the system parameters on the delay gains are investigated via the system level simulations. The works in [79] and [80] also consider the impact of mobility in cache-enabled networks.

F. Energy Consumption

Energy consumption aspects of caching both in terms of area power consumption and energy efficiency are investigated in [81]. Therein, the cache-enabled base stations are distributed according to a homogeneous Poisson point process (PPP) and the optimization is done using a detailed power model. On the other hand, energy harvesting aspects of proactive caching is highlighted in [82], and an effective push mechanism for energy harvesting powered small-cell base stations is proposed in [83]. Also, a joint caching and base station activation for green cellular networks is proposed in [84].

G. Deployment Aspects

Concerning the deployment aspects of cache-enabled SBSs with limited backhaul, a study is given in [85]. In that study, the cache-enabled SBSs are stochastically distributed for the analysis rather than the traditional grid models. The expressions for the outage probability and average content delivery rate are derived as a function of the SINR, SBSs intensity, target content bitrate, cache size and shape of content popularity distribution. Following the work in [85], the results in [86] shows that storing the most popular contents is beneficial only in some particular deployment scenarios. On the other hand, for cache-enabled D2D communications, another stochastic framework is shown in [87], by relying on two performance metrics that quantify the local and global fraction of served content requests. Yet another study for the stochastically distributed cache-enabled nodes is given in [88]. Given the fact that the cost is defined as a function of distance, the expected cost of obtaining the complete content under coded as well as uncoded content allocation strategies is investigated. As an extension to [88], the expected deployment cost of caches vs. the expected content retrieval from the caches is analyzed in [89].

VII. Conclusions

In this chapter, we discussed the current advances in SCNs and proposed a novel proactive network paradigm based on caching at the edge of the network. Using tools from machine

learning, we exploited users' predictable behaviour and their social relationships for caching at the edge of the network. Our approach showed that peak mobile traffic demands can be significantly minimized, yielding backhaul offloadings and resource savings. According to our findings and the growing literature, caching is seen as a disruptive solution for 5G SCNs. An interesting direction of the work presented here would be the *estimation of content popularity* when the time and spatial dynamics of mobile users are involved. This clearly requires the development of novel algorithms and machine learning tools which can infer the content popularity patterns from available data. Additionally, the benefits of caching in *complex network structures* (i.e., hierarchical networks, multi-hop networks, heterogeneous networks, combination of them, etc.) could be investigated while considering network constraints and physical-layer aspects. On the other hand, *adaptive proactive caching schemes* which can predict users' behaviour online and cache the contents accordingly are still in infancy, and in this regard, establishing trade-offs between the feedback overhead and possible performance gains would be interesting. Also, *joint designs* (i.e., caching and scheduling, network/index coding aided caching, etc.) is yet another direction to reveal. On top of these, *experimental test-beds* would allow network operators to see the practical gains for cache-enabled 5G SCNs.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," *White Paper, [Online] http://goo.gl/l77HAJ*, 2014.

[2] Ericsson, "5G radio access - research and vision," *White Paper, [Online] http://goo.gl/Huf0b6*, 2012.

[3] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, 2013.

[4] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[5] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.

[6] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[7] 3GPP TR 36.839 v11.0.0, "Mobility enhancements in heterogeneous networks (release 11)," *[Online] http://www.3gpp.org/DynaReport/36913.htm*, sept 2014.

[8] M. Simsek, M. Bennis, and I. Guvenc, "Mobility management in hetnets: a learning-based perspective," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 26, 2015.

[9] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, 2013.

[10] M. Simsek, M. Bennis, and I. Guvenc, "Enhanced intercell interference coordination in hetnets: Single vs. multiflow approach," in *IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2013, pp. 725–729.

[11] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic uplink-downlink optimization in tdd-based small cell networks," *arXiv preprint arXiv:1402.7292*, 2014.

[12] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 44–50, June 2013.

[13] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," *arXiv preprint arXiv:1305.5585*, 2013.

[14] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *arXiv preprint arXiv:1312.0229*, 2013.

[15] S. Göbbels, "Disruption tolerant networking by smart caching," *IEEE International Journal of Communication Systems*, pp. 569–595, Apr 2010.

[16] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *IEEE INFOCOM*, 2012, pp. 1107–1115.

[17] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82 – 89, 2014.

[18] C. M. R. Inst., "C-RAN: The Road Towards Green RAN," *White Paper, [Online] http://goo.gl/jZt7TR*, 2011.

[19] S. Sezer, S. Scott-Hayward, P.-K. Chouhan, B. Fraser, D. Lake, J. Finnegan, N. Viljoen, M. Miller, and N. Rao, "Are we ready for SDN? implementation challenges for software-defined networks," *Communications Magazine, IEEE*, vol. 51, no. 7, 2013.

[20] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[21] V. Etter, M. Kafsi, and E. Kazemi, "Been There, Done That: What Your Mobility Traces Reveal about Your Behavior," in *Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing*, 2012.

[22] E. Baştuğ, J.-L. Guénégo, and M. Debbah, "Proactive small cell networks," in *20th International Conference on Telecommunications (ICT)*, Casablanca, Morocco, 05/2013 2013.

[23] Netflix, "Netflix prize," *[Online] http://www.netflixprize.com*, 2009.

[24] J. Lee, M. Sun, and G. Lebanon, "A comparative study of collaborative filtering algorithms," *[Online] arXiv: 1205.3193*, 2012.

[25] P. Arkadiusz, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD cup and workshop Vol. 2007.*, 2007.

[26] M. Newman, *Networks: an introduction*. Oxford University Press, 2009.

[27] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651 – 666, 2010.

[28] T. L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review," *J. Mach. Learn. Res.*, vol. 12, pp. 1185–1224, Jul. 2011.

[29] M. A. I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing, 1972.

[30] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[31] L. A. Belady, "A study of replacement algorithms for a virtual-storage computer," *IBM Syst. J.*, vol. 5, no. 2, p. 78–101, 1966.

[32] J. Wang, "A survey of web caching schemes for the internet," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 5, pp. 36–46, 1999.

[33] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[34] A. Araldo, M. Mangili, F. Martignon, and D. Rossi, "Cost-aware caching: optimizing cache provisioning and object placement in ICN," *arXiv preprint arXiv:1406.5935*, 2014.

[35] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26–36, 2012.

[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, October 2010.

[37] E. Baştuğ, M. Bennis, and M. Debbah, "Anticipatory caching in small cell networks: A transfer learning approach," in *1st KuVS Workshop on Anticipatory Networks*, Stuttgart, Germany, 09/2014 2014.

[38] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," *arXiv preprint arXiv:1503.05448*, 2015.

[39] E. Baştuğ, K. Hamidouche, W. Saad, and M. Debbah, "Centrality-based caching for mobile wireless networks," in *1st KuVS Workshop on Anticipatory Networks*, Stuttgart, Germany, 09/2014 2014.

[40] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *WNC3 workshop, WiOpt*, Hammamet, Tunisia, 2014.

[41] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[42] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, "Caching eliminates the wireless bottleneck in video-aware wireless networks," *arXiv preprint arXiv:1405.5864*, 2014.

[43] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," *arXiv preprint arXiv:1402.7314*, 2014.

[44] P. Blasco and D. Gunduz, "Learning-based optimization of cache content in a small cell base station," *arXiv preprint arXiv:1402.3247*, 2014.

[45] P. Blasco and D. Gündüz, "Content-level selective offloading in heterogeneous networks: Multi-armed bandit optimization and regret bounds," *arXiv preprint arXiv:1407.6154*, 2014.

[46] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *International Symposium on Wireless Communication Systems (ISWCS)*, Barcelona, Spain, 08/2014 2014.

[47] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3665–3677, October 2014.

[48] A. Abboud, E. Baştuğ, K. Hamidouche, and M. Debbah, "Distributed caching in 5G networks: An alternating direction method of multipliers approach," *[Online] http://goo.gl/vBdhV7*, 2015.

[49] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[50] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *arXiv preprint arXiv:1308.0178*, 2013.

[51] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," *arXiv preprint arXiv:1404.6563*, 2014.

[52] ——, "Coded caching for heterogeneous wireless networks with multi-level access," *arXiv preprint arXiv:1404.6560*, 2014.

[53] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," *arXiv preprint arXiv:1504.01123*, 2015.

[54] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *arXiv preprint arXiv:1311.3646*, 2013.

[55] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. Diggavi, "Hierarchical coded caching," in *IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 2142–2146.

[56] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *arXiv preprint arXiv:1503.00265*, 2015.

[57] Z. Chen, "Fundamental limits of caching: Improved bounds for small buffer users," *arXiv preprint arXiv:1407.1935*, 2014.

[58] A. Krishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," *arXiv preprint arXiv:1501.02549*, 2013.

[59] U. Niesen and M. A. Maddah-Ali, "Coded caching for delay-sensitive content," *arXiv preprint arXiv:1407.4489*, 2014.

[60] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *arXiv preprint arXiv:1312.3961*, 2013.

[61] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv preprint arXiv:1305.5216*, 2014.

[62] ——, "Fundamental limits of caching in wireless d2d networks," *arXiv preprint arXiv:1405.5336*, 2014.

[63] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 388–404, 2000.

[64] A. F. M. Mingyue Ji, Giuseppe Caire, "On the average performance of caching and coded multicasting with random demands," in *11th International Symposium on Wireless Communication Systems (ISWCS'14)*, Barcelona, Spain, Aug. 2014.

[65] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv preprint arXiv:1502.03124*, 2015.

[66] J. Paakkonen, C. Hollanti, and O. Tirkkonen, "Device-to-device data storage for mobile cellular systems," in *IEEE Globecom Workshops (GC Workshops)*, 2013, pp. 671–676.

[67] C. Hollanti, D. Karpuk, A. Barreal, and H.-f. F. Lu, "Space-time storage codes for wireless distributed storage systems," *arXiv preprint arXiv:1404.6645*, 2014.

[68] A. Barreal, C. Hollanti, D. Karpuk, and H.-f. Lu, "Algebraic codes and a new physical layer transmission protocol for wireless distributed storage systems," *arXiv preprint arXiv:1405.4375*, 2014.

[69] P. Ostovari, A. Khreishah, and J. Wu, "Cache content placement using triangular network coding," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 1375–1380.

[70] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv preprint arXiv:1504.01452*, 2015.

[71] C.-Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *arXiv preprint arXiv:1504.00553*, 2015.

[72] A. Liu and V. Lau, "Cache-enabled opportunistic cooperative mimo for video streaming in wireless systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 390–402, Jan 2014.

[73] ——, "Cache-induced opportunistic mimo cooperation: A new paradigm for future wireless content access networks," in *IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 46–50.

[74] ——, "Exploiting base station caching in mimo cellular networks: Opportunistic cooperation for video streaming," *IEEE Transactions on Signal Processing*, vol. 63, no. 1, pp. 57–69, January 2015.

[75] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'14)*. Washington, DC, USA: IEEE, September 2014.

[76] M. Deghel, E. Baştuğ, M. Assaad, and M. Debbah, "On the benefits of edge caching for MIMO interference alignment," *[Online] http://goo.gl/uqp5Uj*, 2015.

[77] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, July 2013, pp. 1017–1021.

[78] V. A. Siris, X. Vasilakos, and G. C. Polyzos, "Efficient proactive caching for supporting seamless mobility," *arXiv preprint arXiv:1404.4754*, 2014.

[79] P. Sermpezis, L. Vigneri, and T. Spyropoulos, "Offloading on the edge: Analysis and optimization of local data storage and offloading in HetNets," *arXiv preprint arXiv:1503.00648*, 2015.

[80] G. Alfano, M. Garetto, and E. Leonardi, "Content-centric wireless networks with limited buffers: when mobility hurts," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 1815–1823.

[81] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching on the edge: a green perspective for 5G networks," in *IEEE International Conference on Communications (ICC'15)*, London, UK, June 2015.

[82] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: Proactive content caching and push with energy harvesting based small cells," *arXiv preprint arXiv:1503.04254*, 2015.

[83] J. Gong, S. Zhou, Z. Zhou, and Z. Niu, "Proactive push with energy harvesting based small cells in heterogeneous networks," *arXiv preprint arXiv:1501.06239*, 2015.

[84] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Joint caching and base station activation for green heterogeneous cellular networks," in *IEEE International Conference on Communications (ICC'15)*, London, UK, June 2015.

[85] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, no. 1, p. 41, February 2015.

[86] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," *arXiv preprint: 1409.7626*, 2014.

[87] A. Altieri, P. Piantanida, L. R. Vega, and C. Galarza, "On fundamental trade-offs of device-to-device communications in large wireless networks," *arXiv preprint arXiv:1405.2295*, 2014.

[88] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," *arXiv preprint arXiv:1309.0604*, 2013.

[89] M. Mitici, J. Goseling, M. de Graaf, and R. J. Boucherie, "Deployment vs. data retrieval costs for caches in the plane," Enschede, the Netherlands, December 2013. [Online]. Available: http://doc.utwente.nl/88064/