TRANSACTIONS ON EMERGING TELECOMMUNICATIONS TECHNOLOGIES
Trans. Emerging Tel. Tech. 2014; **25**:468–478
Published online 7 November 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/ett.2583

RESEARCH ARTICLE

Energy savings scheme in radio access networks via compressive sensing-based traffic load prediction

Rongpeng Li, Zhifeng Zhao*, Xuan Zhou and Honggang Zhang

York-Zhejiang Lab for Cognitive Radio and Green Communications, Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

ABSTRACT

In radio access networks, the base stations' (BSs) power consumption does not merely depend on the traffic loads within its coverage. The auxiliary devices, especially the cooling system in BSs, contribute to significant energy exhaustion. As the traffic loads fluctuate spatially and temporally, the BSs consequently suffer from heavy energy wastage when the traffic loads of their coverage are low. In this paper, an energy saving scheme over predicted traffic loads is proposed to tackle this energy inefficiency problem in incumbent radio access networks induced by the fluctuations of traffic loads. The proposed scheme firstly takes advantage of the spatial–temporal pattern of traffic loads and employs the compressive sensing method to predict the future traffic loads. Then, a grid-based energy saving scheme is developed to improve the energy efficiency through turning some BSs into sleeping mode while ensuring the quality of experience. Results of the simulation with real traffic load statistics finally validate the accuracy of the traffic load prediction and large improvement of energy efficiency. Copyright © 2012 John Wiley & Sons, Ltd.

***Correspondence**
Z. Zhao, Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China.
E-mail: zhaozf@zju.edu.cn

Received 30 March 2012; Revised 1 August 2012; Accepted 23 August 2012

1. INTRODUCTION

With the unprecedented popularity of smartphones and other mobile terminals, there emerges an explosive demand for radio network access and incurs large power consumption simultaneously [1]. Furthermore, it is coming to a consensus that the information and communication technology (ICT) industry has become one of the major contributors to the world's power consumption and greenhouse gas emission. Recent studies show that the ICT industry accounts for 2 to 10 per cent of the world's overall power consumption [2, 3] and it is envisioned that the energy exhaustion of mobile networks would reach 124.48 kWh in the year of 2011 [4], and the power bill will doubly enlarge in 5 years for China Mobile [5]. Beyond that, increasing awareness of the exhaustion of non-renewable energy resources also spurs the need to improve the energy efficiency of telecommunication systems.

In addition to environmental concerns, there are also economical benefits for cellular network operators to reduce the power consumption of their networks. The energy expenditure accounts for a significant proportion of the overall cost, and cellular network operators would save a lot in capital and operating expenditure through improving energy efficiency [6].

Currently, the subscribers cost around 1 per cent of the overall energy in ICT industry. Comparatively, over 80 per cent of the power consumption takes place in the radio access networks (RAN), especially the base stations (BSs) [7]. The reasons lie in that the networks at present stage are designed for maximum throughput, and the deployment of BSs is usually optimised for peak user demand operation. However, there are significant variations in traffic loads, both temporally and spatially [8] because the typical day–night division of user behaviour causes the temporal periodic pattern (tidal effect of traffic loads), and the spatial difference occurs when people move between their apartments and downtown districts back and forth, taking along their mobile terminals. Unfortunately, the present communication infrastructure's activities in reality weakly depend on the amount of the traffic loads, which means certain heavily underutilised BSs still have to stay active and consume relatively large amount of energy even when there is little user accessing requirement, resulting in severe energy wastage [9]. Hence, there is a strong incentive to make the working status of RAN adaptive to

the traffic loads, in order that the energy efficiency of BSs can be improved.

Paper Scope and Contributions: In a nutshell, this paper considers how to dynamically optimise the number of active BSs, depending on the predicted traffic loads. Further speaking, the paper proposes how and when to turn some BSs into sleeping mode, so as to solve the underutilisation problem of BSs and its induced energy inefficiency in RAN. Thus far, there has been a substantial body of work towards it. In [2] and [8], the authors propose to dynamically adjust the BSs based on the traffic loads, but the authors assume that the traffic loads follow some kind of mathematical distributions. Therefore, it can not catch the real traffic characteristics of the network precisely. In [10], certain BSs' on/off operations at specific moments can only occur once within 24 h. Accordingly, the network adaptation operation can not timely follow the actual spatial–temporal traffic dynamics, leading to user experience quality degradation. The authors in [11] propose a snapshot-based global cellular network greening scheme. To catch the variance of traffic loads, the BSs need to operate between switching on and switching off frequently, which results in high computation cost and wastage of bandwidth for transferring data. In [12], Niu, Wu, *et al.* propose a traffic load transferring scheme. However, they do not consider the underlying coverage adjustment problem and the increase of energy consumption caused by zooming coverage, thus merely obtaining a suboptimal energy consumption result.

A preliminary versions of our results appear in [13] and [14]. This submission is distinct because of the comprehensive analysis of the cause of energy inefficiency in BSs and fuller and deeper introduction of power consumption model in RAN. The contributions of this paper lie in that it employs the spatial–temporal characteristics of traffic loads and adopts a compressive sensing-based method to predict the future traffic loads. After attaining the predicted traffic loads, instead of merely minimising the number of active BSs, a grid-based scheme is proposed to minimise the power consumption of active BSs (GM-PAB) by balancing the number of active BSs and their coverage (and also their served traffic loads). Beyond that, to give an objective assessment for our research, the simulation is conducted with real traffic load statistics in Hangzhou, China and validates large energy efficiency improvement.

Paper Organisation: Section 2 covers some background on RAN and power consumption model of BSs. Sections 3 and 4 introduce the method for traffic load prediction and the scheme for energy saving, respectively. Section 5 evaluates our proposed scheme and presents our simulation results. Section 6 concludes with a summary of this paper and a discussion of future work.

2. BACKGROUND

The RAN of cellular mobile networks is responsible for handling traffic and signalling. Figure 1 illustrates a typical architecture of RAN in Global System for Mobile Communications (GSM) network.* As Figure 1 depicts, RAN connects the subscribers and the core network and mainly consists of several base station controllers (BSCs) and their controlled BSs. Usually, a single BSC can take control of tens or even hundreds of BSs.

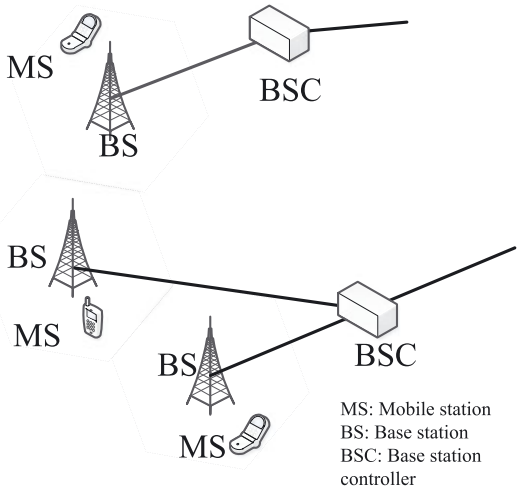

Figure 1. An illustration of the radio access networks.

As is the concentration of the paper, one BS consists of a communication subsystem and a supporting subsystem. The communication subsystem includes antenna array, remote radio unit (RRU), base band unit (BBU) along with the fibre-optic cable connecting BBU and RRU as illustrated in Figure 2. Each BS may instal several RRUs near the antennas to provide larger coverage and capacity. BBU is the main unit and takes charge of all the communication functionalities: scrambling, modulation, scheduling, adaptive coding, link quality measurements, soft handovers, and so on. The supporting subsystem includes the cooling system and other auxiliary devices and aims to maintain an appropriate temperature and sufficiently monitor the environment. In the viewpoint of power consumption, the energy exhaustion model of each BS in [8, 15] is adopted, and it is composed of two categories. One is from the traffic loads transmission, whose power consumption P_t can be linearly approximated as $P_t = P_\alpha \cdot V + P_\beta$ with respect to the traffic loads V [16]. The variance of P_t is the result of RRU and BBU. For example, RRU has to support more active links if the traffic loads are heavy, whereas BBU has to do baseband processing unless the BS turns into sleeping mode. Moreover, all the other operations such as signalling control will incur energy overhead even when the traffic loads are low. As to the coefficient P_α, it depends on the transmission distance because one

*It is worthwhile to note here that although the remaining part of the paper mainly focuses on GSM network, the energy saving scheme in GSM network can also be applied in third generation and its long-term evolution networks, given the similarities of RANs of these networks.

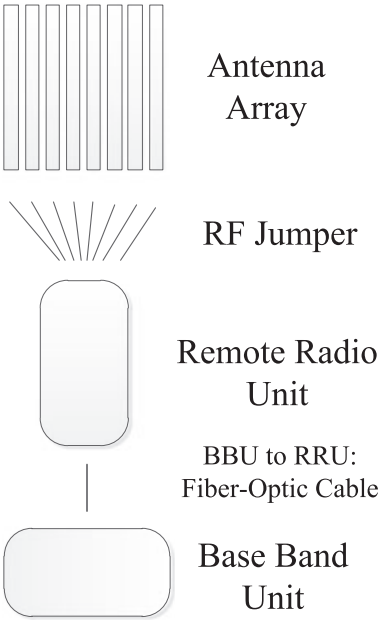

Figure 2. An illustration of communication subsystem structure of typical base station in Global System for Mobile Communication network.

BS will exhaust more energy to serve the traffic loads from longer distance. Meanwhile, there exists certain energy consumption, owing to the supporting subsystem and some communication modules, especially the cooling system. This category of power consumption P_s mainly depends on the working environment, and thereby, it is almost invariant to the traffic loads. Hence, the supporting subsystem is assumed to stay constant on a daily basis in the paper. In this paper, let us denote the constant power consumption as $P_{\text{steady}} = P_\beta + P_s$, which is irrespective of traffic loads and contributes to the overall power consumption as high as 50 per cent [7]. Therefore, the RAN can save a large amount of energy if some of the BSs are turned into sleeping mode when few traffic loads exist.

3. COMPRESSIVE SENSING-BASED TRAFFIC LOAD PREDICTION

3.1. Traffic load prediction model and literature review

In this paper, it is assumed that there exists a traffic load vector to timely record the volumes of BSs' traffic loads.[†] Suppose there are n BSs under one particular BSC's control. Thereby, there will be an n-length vector as traffic vector, each of whose elements archives the volume of one BS's traffic at one specific moment. For monitoring purposes, the volumes of traffic loads would be traced periodically to better know the working status of BSs. Hence, a traffic load matrix is referred to as the set of traffic load vectors at different moments. For example, $X_{i,t}$ in a traffic load matrix X denotes the volume of traffic of BS i at the moment t. In other words, every row vector of traffic matrix denotes the volumes of traffic at one specific BS with respect to the time, whereas every column vector denotes volumes of traffic of several adjacent BSs at one specific moment. Therefore, the traffic load prediction problem can be interpreted as determining the traffic load vector in the near future moment τ.

Thus far, there are two main streams concerned with the traffic load matrix prediction. One employs the characteristics of the traffic loads, such as spatial and temporal relevancies [17, 18] or self-similarity [19], and tries to find a fitting model (i.e. ON-OFF model [20], FARIMA model [18], mobility model [21, 22], network traffic model [22] and alpha-stable model [23, 24]) and then adopts an appropriate prediction method to obtain the future traffic loads. The other attacks the traffic matrix estimation problem by using some modern signal processing techniques, such as principal components analysis method to study the intrinsic dimensionality [25, 26] or Kalman filtering method to capture the evolution of traffic loads [25, 27].

Because some traffic volumes cannot be recorded because of the possible overloading of recording server, the traffic load matrix might be incomplete or inaccurate in reality. Therefore, the aforementioned prediction methods suffer and might become infeasible under this scenario. Consequently, this paper adopts a spatial–temporal compressive sensing-based traffic prediction method, which not only involves the signal processing technique (i.e. compressive sensing [28]) but also exploits the characteristics of spatial and temporal relevancies. Hence, by taking advantage of compressive sensing and exploring more characteristics of traffic loads, our proposed method should outperform the existing ones.

3.2. Spatial–temporal compressive sensing and traffic prediction method

3.2.1. Brief introduction of compressive sensing.

In real world, a great many of signals or datasets exhibit characteristics such as structure or redundancy. To be more specific, the representation of signals or datasets contains only a small number of non-zero elements in some transformation bases. In other words, the signals are sparse in certain transformation bases. Compressive sensing, which aims to handle these sparse signals problems, has developed and attracted large attention recently [28]. According to the compressive sensing theory, if the signals match the sparsity condition, much smaller dimensional measurements are required for both sampling and reconstruction of signals or datasets. Meanwhile, compressive sensing or its usually embedded greedy algorithm is an

[†]According to our survey, there exists such statistical records on the volumes of traffic concerning all the BSs because the records will help make the maintenance more easy. Above all, the assumption is feasible.

Trans. Emerging Tel. Tech. **25**:468–478 (2014)
DOI: 10.1002/ett

alternative way to solve underconstrained linear inverse problems as well. It leads to efficient compression, estimation and modelling [29].

In the context of matrix compressive sensing, if one matrix has very low rank, the spectrum of singular values of this matrix will be sparse, thus making the matrix have sparsity. Because the traffic load matrix has certain periodic pattern [8], spatial and temporal relevancies [17, 18] or certain structure and redundancy [30], the traffic load matrix can thereby be approximated as a low-rank matrix. So, compressive sensing can be applied in this traffic load prediction problem.

3.2.2. Spatial–temporal compressive sensing method.

Given the potential breakdown of monitoring equipments or overloading of BSs, there might exist some measurement errors or inaccuracies in the recorded volume of traffic loads. This section focuses on how to find or recover the absent data from the 'flaw' traffic matrix. Later on, it will be introduced how to extend the proposed idea to the prediction problem in Section 3.2.3.

Let us denote a low-rank matrix $\hat{X}$, whose entries are almost equal to those in X, except the missing or inaccurate values in the latter matrix X. Considering that the traffic load matrix has certain structure or redundancy, the approximation matrix $\hat{X}$ should have the least rank and meanwhile satisfy the following equation

$$\begin{aligned} &\min \ \operatorname{rank}\left(\hat{X} = LR^T\right) \\ &\text{s.t.} \quad M .* (LR^T) = M .* X \end{aligned} \tag{1}$$

where M is an $m \times n$ matrix, the same size with X, and $M .* X$ is the element-wise multiplication. Moreover, the approximation matrix $\hat{X}$ is replaced with two factorization matrices L and R, the sizes of which are $m \times r$ and $n \times r$, respectively. Here, r is a factor involving the factorization precision and takes the value far less than $\min(m, n)$. Because the factorization results L and R contain only $r \times (m + n)$ entries, far less than $m \times n$ ones in X, the factorization method has fewer unknown entries and contributes to the low-rank approximation. The entries of M[‡] are given by

$$M(i, j) = \begin{cases} 0, & X(i, j) \text{ is missing or inaccurate,} \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

When $\hat{X}$ is determined by the factorization result LR, the missing values in X can be approximated using their counterparts in $\hat{X}$. Therefore, the only remaining problem is to solve Equation (1). Unfortunately, the objective of rank minimization is not convex and is perhaps not solvable. But under certain particular constraints (i.e. the restricted isometry property in compressive sensing [26, 31, 32]), the rank minimization problem can be solved by the following equivalent form:

$$\begin{aligned} &\min \ \|L\|_F^2 + \|R\|_F^2 \\ &\text{s.t.} \quad M .* (LR^T) = M .* X \end{aligned} \tag{3}$$

where the form $\|\cdot\|_F$ denotes the Frobenius norm, that is $\|Y\|_F = \sqrt{\sum_{i,j} Y(i, j)^2}$ is the Frobenius norm for matrix Y.

Besides, because of the relative low-rank nature of the original data matrix X and inaccuracy of the measurements, the matrix factorization that perfectly fits the constraints may not exactly exist. As a result, the precision of approximation is also included as a part of the optimization process. Thus, Equation (3) transforms into

$$\min \ \|M .* \left[\left(LR^T\right) - X\right]\|_F^2 + \alpha \left(\|L\|_F^2 + \|R\|_F^2\right) \tag{4}$$

where α is a weight factor. In this regard, this approach finds a low rank regularised factorization, which fits the intrinsic characteristic of the data matrix X. In addition, it does not strictly ensure the constraints but still keeps the accuracy by minimising the difference.

When deriving the factor matrices, the alternating least squares method is adopted by fixing one of the two factor matrices and optimising the other. The optimization process is completed alternatively until the convergence of the two factor matrices.

Actually, the spatial and temporal relevancies in the original datasets is exploited to make the approximation more precise [26]. In temporal dimension, the traffic values in adjacent moments are often close. Also, in spatial dimension, the traffic values between neighbouring places also have some kind of smoothness because of the space relevancy in some sense. Therefore, after utilising this prior knowledge to reconstruct the data matrix, Equation (4) can be formulated as

$$\begin{aligned} \min \ \|M .* \left[\left(LR^T\right) - X\right]\|_F^2 + \alpha \left(\|L\|_F^2 + \|R\|_F^2\right) \\ + \beta \left\|SLR^T\right\|_F^2 + \gamma \left\|LR^T T^T\right\|_F^2 \end{aligned} \tag{5}$$

where T and S denote the temporal relevancy matrix and spatial relevancy matrix, respectively. β and γ are weight factors like α as well. These two matrices contain our prior knowledge about the spatial and temporal structure of the traffic datasets. The temporal relevancy matrix T describes the temporal smoothness of the traffic datasets, which means that the traffic values at adjacent moments are usually flat. The spatial relevancy matrix S is used to express the spatial correlation of the traffic data at different places. By minimising $\|LR^T T^T\|_F^2$ and $\|SLR^T\|_F^2$, Equation (5) intends to make the low-rank approximation inherit the intrinsic temporal and spatial properties of X.

[‡]Because the monitoring equipments record the operating status of BSs simultaneously, it can be known where and when the traffic load records are missing and inaccurate.

3.2.3. Traffic load prediction solution.

Previous sections state the way to determine the missing values in the incomplete traffic load matrix by the compressive sensing-based method. Indeed, the compressive sensing-based method can be extended to the prediction problem, where the values prediction problem is similar to determining the missing values beforehand.

According to different sources of original traffic datasets and applications, different time and space relevancy matrices S and T can be set to make low-rank approximation more real and precise. But for simplicity purpose, the spatial relevancy matrix is not used when predicting the network traffic in this paper.

If the known incomplete traffic load matrix is augmented with the traffic load vector at the future moment τ, the future traffic load vector can be achieved following the compressive sensing-based algorithm mentioned earlier. As to the temporal relevancy matrix, after taking into account the design insight 2 of Peng *et al*. [8]: *at any time, the traffic load difference in two consecutive days is less than* 20% *for* 70% *BSs*, a temporal relevancy matrix can be constructed as

$$T = \begin{bmatrix} \overbrace{1 & 0 & \cdots & -1}^{24\text{ h}} & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & -1 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots & -1 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \end{bmatrix} \tag{6}$$

This temporal relevancy matrix means that the traffic datasets of a network at the same moment of two consecutive days have similar patterns. Then, by employing the spatial–temporal compressive sensing method with the temporal relevancy matrix,[§] the traffic loads of future time can be predicted based on the historical traffic loads and underlying trends.

4. TRAFFIC PREDICTION-BASED BASE STATION ENERGY SAVING SCHEME

Currently, in typical cellular networks, many BSs are under utilisation at most time of the day, which results in much energy wastage and heavy energy inefficiency. For instance, to realise reliable transmission of high quality at the rare peak time, there might be a great many of BSs with highly overlapping coverage [33]. For example, as Figure 3 shows, the dark mobile terminal can be simultaneously covered and served by the surrounding three BSs. Commonly, one of the three BSs is enough to provide communication with acceptable quality of experience (QoE), thus making the other two BSs under utilisation. Hence, if the traffic loads can be assembled to only an optimised number of BSs (active BSs) according to the predicted traffic loads, the energy inefficiency problem can be solved. To be specific, after switching off the rest BSs (sleeping BSs), the active BSs can take advantage of modern techniques such as beamforming to cover the blank space left by the inactive BSs [12]. When these inactive BSs are needed because of traffic rebound in the network, they are switched on again. With this scheme, a great energy efficiency improvement is achieved. Moreover, the BS switching operations can be maintained in an off-line manner based on the predicted data beforehand.

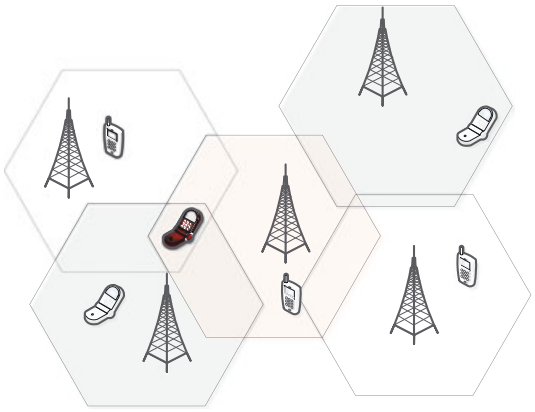

Figure 3. An illustration of the overlapping coverage among adjacent base stations.

4.1. Introduction to basic approach for base station energy saving

An intuitive solution [11] to solve the energy inefficiency of BSs is to minimise the number of active BSs (M-NAB) when some of the BSs has very low traffic loads. Meanwhile, the M-NAB solution needs to guarantee that the active BSs can serve at least the volume of predicted traffic loads as well. Thus, the state-of-the-art behind the M-NAB solution is to transfer all the traffic loads of certain BSs, which will be turned into sleeping mode, to their adjacent ones. At the same time, the solution should also concern and satisfy the capacity requirements. So the M-NAB solution can be formulated as

$$\begin{aligned} \min\ & \sum_{j=1}^{n} \operatorname{sgn}\left(\sum_{i=1}^{n} V_{i,j}\right) \\ \text{s.t. } & \textstyle\sum_{j=1}^{n} V_{i,j} \geqslant \hat{X}_{i,\tau}, && \forall i \in 1 \cdots n \\ & \textstyle\sum_{i=1}^{n} V_{i,j} \leqslant C_j, && \forall j \in 1 \cdots n \\ & V_{i,j} = 0, && \forall (i,j) \notin \varepsilon \cap i \neq j \\ & V_{i,j} \geqslant 0, && \forall (i,j) \in \varepsilon \cup i = j \end{aligned} \tag{7}$$

where n denotes the number of BSs, consistent with the length of traffic load vector. $\operatorname{sgn}(\cdot)$ is the sign function, that is $\operatorname{sgn}(x) = x/|x|$ if $x \neq 0$ and $\operatorname{sgn}(x) = 0$ if $x = 0$.

[§]The time relevancy matrix can be chosen as other format, depending on the computation and precision requirement.

Trans. Emerging Tel. Tech. **25**:468–478 (2014) © 2012 John Wiley & Sons, Ltd.
DOI: 10.1002/ett

Moreover, $\hat{X}_{i,\tau}$ denotes the predicted traffic loads in BS i at time τ, whereas C_i represents the capacity threshold of BS i. $V_{i,j}$ denotes the volume of traffic loads reallocated from BS i to BS j. ε is the set of neighbouring relation pairs, and every element $(i, j) \in \varepsilon$ denotes BS i and j are adjacent and ensures the traffic loads originally served by BS i can be served by BS j. The constraints make sure that the volume of traffic loads transferred from some BSs is no less than the corresponding predicted traffic loads, thus guaranteeing the QoE.

4.2. Grid-based base station energy saving scheme

The M-NAB solution gives the approach for energy saving in RAN, but there are several drawbacks in it. First of all, it does not concern how to adjust the BSs' coverage to transfer or reallocate the traffic loads. Secondly, the M-NAB solution does not consider the increased power consumption for one BS to zoom out its coverage or serve larger volumes of traffic loads. Therefore, a grid-based BS energy saving scheme is proposed to compensate the drawbacks of M-NAB.

In the first place, the coverage of every BS is divided as grids with equal size as Figure 4(a) illustrates. The size of every grid depends on the precision requirement of traffic transferring. Because of the coverage of one BS is usually small, it is assumed that the predicted traffic loads of one BS will be equally distributed in every grid in its coverage. Then, the BSs and grids are mapped as the vertices of an undirected graph $G = (V, E)$, which are represented as triangles and squares, respectively, in Figure 4(b) for the convenience of representation. The grid (square) k and its one surrounding BS (triangle) i is connected to form an edge $e_{i,k} \in E$ so long as this grid k is located in the BS i's maximum transmission range. Thus, the N-MAB solution can be regarded as selectively choosing the edges in the graph while keeping in mind that the square needs to connect with one triangle (it ensures that the grid must be served by one BS). Further, the triangle should be connected with limited squares (it ensures that every BS is able to provide so many traffic loads). Figure 4(b) illustrates the grid-based energy saving scheme. For example, the subscribers in every grid will finally be served by the BS connected by the bold line, whereas the two 'black' BSs will be turned into sleeping mode to save energy.

According to the aforementioned description, the M-NAB solution in Equation (7) can be rewritten as follows (grid M-NAB or GM-NAB):

$$\begin{aligned} \min \ & \sum_{i=1}^{n} \operatorname{sgn}\left(\sum\nolimits_{e_{i,k}\in E} I_{e_{i,k}}\right) \\ \text{s.t.} \ & \textstyle\sum_{e_{i,k}\in E} V_{k,\tau} \times I_{e_{i,k}} \leqslant C_i, && \forall i \in 1 \cdots n \\ & \textstyle\sum_{e_{i,k}\in E} I_{e_{i,k}} \geqslant 1, && \forall k \\ & I_{e_{i,k}} \in \{0, 1\}, && \forall e_{i,k} \in E \end{aligned} \tag{8}$$

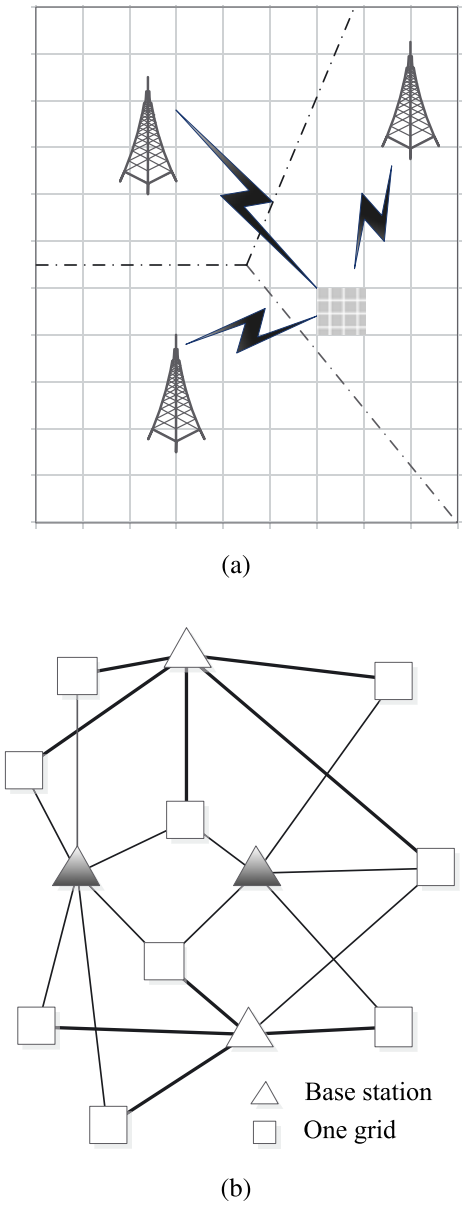

Figure 4. (a) Grids of base stations' coverage area and (b) the logic model.

where $I_{e_{i,k}} = 1$ indicates $e_{i,k}$ is included in the optimal solution, whereas $I_{e_{i,k}} = 0$ indicates not. C_i is consistent with that in Equation (7), representing the capacity threshold of BS i. $V_{k,\tau}$ denotes the predicted volume of traffic loads in grid k at time τ. Assuming that grid k originally belongs to the coverage of BS i, $V_{k,\tau}$ will equal the predicted volume of traffic loads $\hat{X}_{i,\tau}$ divided by the number of grids within the coverage of BS i. Hereafter, the subscript τ in $V_{i,\tau}$ is dropped for convenience of representation. The first constraint guarantees that the traffic loads distributed to any BS do not exceed its capacity.

And the second constraint ensures that every grid area can be covered by one BS.

The GM-NAB solution helps explain how to adjust the coverage of active BSs, yet it still consider the power consumption of one active BS to be steady. But the practical situation is that some proportion of the power consumption is relevant to the traffic loads and transmission distance even though the dominant power consumption is steady as explained in Section 2. Therefore, edge weight $P_{e_{i,k}}$ in graph $G = (V, E)$ is introduced to denote the power consumption for BS i to serve the traffic loads in its representing grid k, that is $P_{e_{i,k}} = P_{\alpha,i} \cdot V_k$, where V_k denotes the volume of predicted traffic loads in grid k. Moreover, $P_{\alpha,i}$ is the amount of energy consumption per volume of traffic loads, the value of which depends on distance between BS i and grid k, as discussed in Section 2. Hence, the GM-NAB solution in Equation (8) can be extended to the grid-based minimization of power consumption for Active BSs (GM-PAB):

$$\begin{aligned} \min \quad & \sum_{e_{i,k} \in E} P_{e_{i,k}} \times I_{e_{i,k}} \\ & + \sum_{i=1}^{n} P_{\text{steady},i} \times \operatorname{sgn}\left(\sum\nolimits_{e_{i,k} \in E} I_{e_{i,k}}\right) \\ \text{s.t.} \quad & \textstyle\sum_{e_{i,k} \in E} V_k \times I_{e_{i,k}} \leqslant C_i, \quad \forall i \in 1 \cdots n \\ & \textstyle\sum_{e_{i,k} \in E} I_{e_{i,k}} \geqslant 1, \quad \forall k \\ & I_{e_{i,k}} \in \{0, 1\}, \quad \forall e_{i,k} \in E \end{aligned} \tag{9}$$

where $P_{\text{steady},i}$ denotes the (almost) constant part of power consumption when BS i is active, which is irrelevant to the traffic loads.

As there is a sign function in Equation (9), whose discontinuation makes the equation difficult to solve, Equation (9) is transformed and solved by

$$\begin{aligned} \min \quad & \sum_{e_{i,k} \in E} P_{e_{i,k}} \times I_{e_{i,k}} + \sum_{i=1}^{n} P_{\text{steady},i} \times I_{s_i} \\ \text{s.t.} \quad & \textstyle\sum_{e_{i,k} \in E} V_k \times I_{e_{i,k}} \leqslant C_i, \quad \forall i \in 1 \cdots n \\ & \textstyle\sum_{e_{i,k} \in E} I_{e_{i,k}} \geqslant 1, \quad \forall k \\ & \textstyle\sum_{e_{i,k} \in E} I_{e_{i,k}} - I_{s_i} \times N_{s_i} \leqslant 0, \quad \forall i \in 1 \cdots n \\ & I_{e_{i,k}} \in \{0, 1\}, \quad \forall e_{i,k} \in E \\ & I_{s_i} \in \{0, 1\}, \quad \forall i \in 1 \cdots n \end{aligned} \tag{10}$$

where $I_{s_i} = 1$ indicates BS i will remain active in the final energy saving scheme, whereas $I_{s_i} = 0$ indicates not. And N_{s_i} is the number of edges connected to BS i's corresponding triangle in graph $G = (V, E)$. The newly added third constraint ensures that when BS i is selected to enter into sleeping mode ($I_{s_i} = 0$), there is no need for its service to any grid (all the corresponding $I_{e_{i,k}} = 0$).

Then, GM-PAB in Equation (10) is a typical binary integer programming problem, which is usually an *NP-hard* problem. Fortunately, there is a vast of mathematical algorithms by providing good approximation solutions, such as primal and dual algorithm [34, 35] and branch-and-bound algorithm [36]. In our simulation, our GM-PAB solution is solved by the integration of branch-and-bound algorithm and primal and dual algorithm embedded in Mosek Optimization Tools [37], which is a world-class large-scale linear optimization platform.

5. EVALUATION AND ANALYSIS

5.1. Analysis of traffic load characteristics

To evaluate and testify the algorithm's performance, one two-week traffic load records of 64 BSs in Hangzhou is adopted. Moreover, the interval between two consecutive time moments in the database is 1 h. Therefore, the traffic load matrix is one 64 × 336 matrix. (Given the space limitation of the paper, the traffic load matrix is constructed by merely augmenting traffic load vectors at different moments, whereas the relationship between spatial locations and traffic loads of neighbouring BSs is ignored.) Figure 5 depicts the traffic loads of two typical BSs in 1 week. In Figure 5, it is easily found that the traffic load matrix of BSs exhibit several characteristics, which have been described before and are worthwhile to list here. Firstly, the traffic loads for one specific BS usually follow daily periodic fluctuations. Secondly, the records might be incomplete because of the breakdown of the recording system or other reasons just like the traffic load record at the 84th and 146th hour in Figure 5. Thirdly, the traffic loads at different BSs vary heavily. Certain BSs are under high utilisation, whereas others have very few traffic loads, depending on the locations or regions they belong to. These characteristics are not only the evidence for the assumptions in this paper but also the effectiveness guarantee of the algorithms.

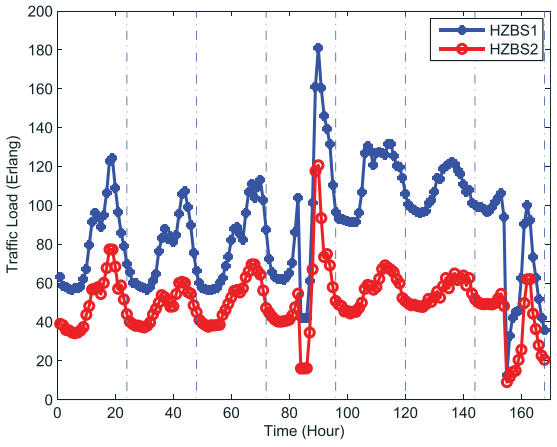

Figure 5. An example of two selected base stations' traffic loads.

Trans. Emerging Tel. Tech. **25**:468–478 (2014) © 2012 John Wiley & Sons, Ltd.
DOI: 10.1002/ett

5.2. Performance of traffic load prediction algorithm

To evaluate the performance of compressive sensing-based traffic load prediction algorithm, the following methodology is adopted: some existing data are intentionally hidden, and then, our prediction algorithm is applied on the pseudo-unknown data. Furthermore, the prediction accuracy is measured by the normalized root mean square error (NRMSE) in the predicted values, namely

$$\mathrm{NRMSE}(\tau) = \frac{\sqrt{\sum_{i=0}^{n}\left(X(i,\tau)-\hat{X}(i,\tau)\right)^2}}{\sqrt{\sum_{i=0}^{n}X(i,\tau)^2}} \qquad (11)$$

where $\hat{X}(:,\tau)$ denotes the predicted traffic load vector at moment τ and $X(:,\tau)$ is the original traffic load vector. n denotes the number of BSs.

To better assess the performance of compressive sensing-based prediction algorithm, the simulations are independently run twice. All of the two simulations utilise six-day training database to predict the next-day traffic loads. As Figure 6 shows, the NRMSE for these two independent simulations with compressive sensing-based prediction algorithm is less than 0.3 at most of the time. Along with the performance of compressive sensing prediction algorithm, the performance using the prediction method based on alpha-stable model [23, 24] is also presented in Figure 6. However, Figure 6 shows that the NRMSE for alpha-stable processes prediction method varies heavily with inferior performance. Especially, as the time to be predicted become longer, the performance may degrade a lot. The proposed algorithm's superior performance lies in that instead of merely utilising the temporal relevancy to predict the future loads, it also exploits the low-rank characteristic of traffic matrix and takes the advantages of compressive sensing techniques.

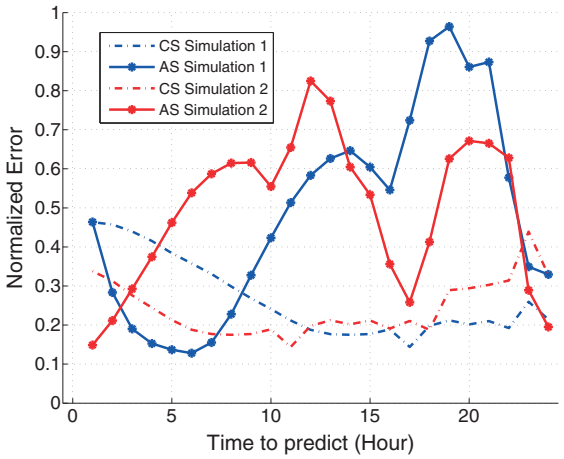

Figure 6. The normalized root mean square error of two independent traffic load prediction simulations using spatial–temporal compressive sensing method (CS) and alpha-stable prediction method (AS).

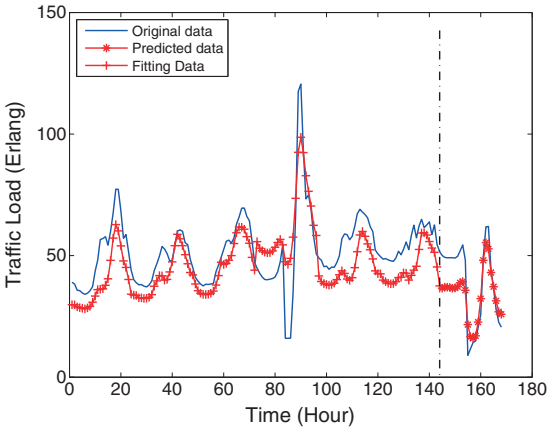

Figure 7. Performance of traffic load prediction using spatial–temporal compressive sensing method: prediction result for one selected base station in one simulation.

Figure 7 shows the approximation precision between the predicted traffic loads and original data. In the whole, the performance of the introduced prediction algorithm is good and stable under different traffic load situations. To be more specific, the prediction algorithm will provide a good premise for the introduced GM-PAB solution to make the communication infrastructure more energy efficient.

5.3. Performance of grid-based base station energy saving solution

To verify the performance of our grid-based solution, it is assumed that there are 64 BSs with initially equal size of coverage (1 km in diameter). And the main parameters in our simulation are heterogeneous and assumed as follows [8]: (i) the capacity of each BS is 110 per cent of the maximum traffic loads for a given BS; (ii) the maximum transmission range also varies in BSs, and it can be 1.6 and 2 km consistent with many available products; (iii) the constant part of the power consumption P_β[¶] can be 2100 and 2800 W; (iv) P_α differs in 4, 6 and 8 as the transmission distance varies in 1, 1.6 and 2 km; and (v) depending on the BS capacity, one third of BSs' constant power consumption is 2800 W with 2 km maximal transmission range. Other BSs' constant power consumption is 2100 W with 1.6 km maximal transmission range.

Here, two metrics, namely power saving ratio and number of active BSs ratio, are defined to measure the performance of GM-PAB solution and GM-NAB solution. To be specific, power saving ratio is the power consumption after utilising the GM-PAB or GM-NAB solution over the original power consumption without any energy

[¶]The meaning of P_α, P_β, and so on have been introduced in Section 2.

saving scheme while number of active BSs ratio is similarly defined. Figures 8 and 9 present the performance of our energy saving schemes at each hour of the day in the case where each BS's coverage is divided into 64 grids of equal size and shows the performance is consistent with the overall traffic load variations. Figure 9 demonstrates that half of the BSs can be turned into sleeping mode in the earliest 10 h of the day and save almost 40 per cent of the power consumption during these first 10 h after performing energy saving schemes as indicated in Figure 8. Beyond that, if the energy saving schemes are utilised, the power consumption of the whole day can be reduced as almost 68 per cent of the consumption before. Especially, the proposed GM-PAB solution can save 34.6 per cent of the power consumption, roughly saving 4.3 per cent more energy over the GM-NAB solution as noted in Figure 8. The reason for energy saving improvement lies in that if the number of active BSs is kept as few as possible, certain BSs have to zoom out their coverage, which in turn leads to larger power consumption and a suboptimal energy saving result. Furthermore, whereas the GM-NAB solution drastically change the number of active BSs in the whole day, the GM-PAB solution can work and adjust more smoothly.

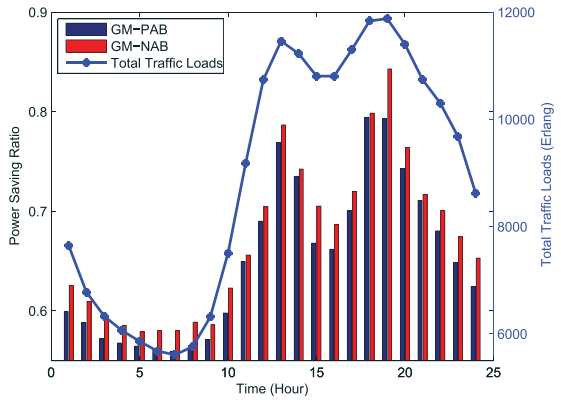

Figure 8. Power saving ratio of GM-PAB and GM-NAB solution in one whole day and the corresponding traffic loads.

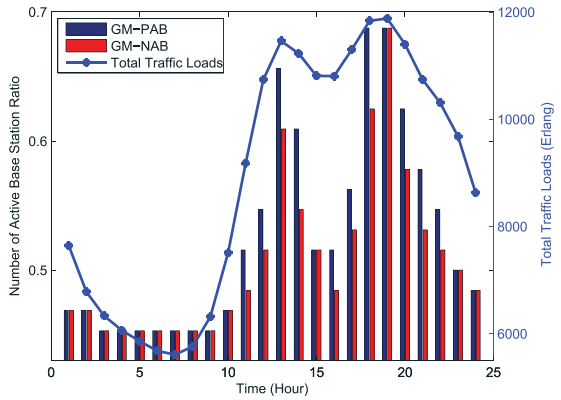

Figure 9. Number of active base station ratio of GM-PAB and GM-NAB solution in one whole day and the corresponding traffic loads.

6. CONCLUSION AND FUTURE WORK

This paper has studied how the power consumption can be saved by turning some BSs into sleeping mode in RAN, on the basis of the predicted traffic loads. The traffic loads of RAN follow some spatial and temporal patterns, thus making the traffic load matrix satisfy the low-rank property. In that regard, the compressive sensing-based prediction method can be employed to forecast the future traffic loads. According to the prediction result of traffic loads, an energy saving scheme has been proposed by turning some BSs into sleeping mode while guaranteeing the QoE, and the solution is finally formulated as a binary integer programming problem. The simulation results verified the precision of our prediction method and proved the effectiveness of our energy saving schemes at last.

In this paper, dynamic BS switching operation schemes have been developed in RAN on the basis of predicted traffic loads and have shown the potential of improving energy efficiency. Nevertheless, the operators might be reluctant to turn off their BSs because of concerns about possible QoE degradation. In the future, the work can be extended to design a component-level energy saving technique in BS operation, which is more conservative than turning BSs into sleeping mode. However, the technique needs to be more sensitive to the traffic load fluctuations in order to timely turn some components into sleeping mode. In other words, it requires to improve the compressive sensing-based traffic load prediction method such that a more precise result can be obtained in a smaller time scale. This will be carefully addressed in future work.

ACKNOWLEDGEMENTS

This paper is partially supported by the National Basic Research Program of China (973 Program 2012CB316000) and the National Natural Science Foundation of China (NSFC) under grant number 61071130.

REFERENCES

1. Zhang H, Gladisch A, Pickavet M, Tao Z, Mohr W. Energy efficiency in communications. *IEEE Communications Magazine* 2010; **48**(11): 48–49.
2. Marsan M, Chiaraviglio L, Ciullo D, Meo M. Optimal energy savings in cellular networks, In *Proceedings of IEEE ICC 2009*, Dresden, Germany, 2009; 1–5.
3. Global Action Plan Rep. Global action plan, an inefficient truth, 2007. Available at: http://globalactionplan.org.uk

4. ABI Research. Mobile networks go green-minimizing power consumption and leveraging renewable energy, 2009. Available at: http://www.abiresearch.com/
5. China Mobile Research Institute. C-RAN: Road towards green radio access network. *Technical Report*, 2010.
6. Reviriego P, Maestro J, Larrabeiti D. Study of the potential energy savings in ethernet by combining energy efficient ethernet and adaptive link rate. *Transactions on Emerging Telecommunications Technologies* 2012; **23**: 227–233.
7. Fettweis GP, Zimmermann E. ICT energy consumption-trends and challenges, In *Proceedings of WPMC 2008*, Vol. 4, Lapland, Finland, 2008; 6.
8. Peng C, Lee SB, Lu S, Luo H, Li H. Traffic-driven power savings in operational 3G cellular networks, In *Proceedings of ACM Mobicom 2011*, Las Vegas, Nevada, USA, 2011; 121–132.
9. Herault L, Strinati EC, Zeller D, Blume O, Imran MA, Tafazolli R, Lundsjö J, Jading Y, Meyer M. Green Communications: a global environmental challenge, In *Proceedings of WPMC 2009*, Sendai, Japan, 2009.
10. Oh E, Krishnamachari B. Energy savings through dynamic base station switching in cellular wireless access networks, In *Proceedings of IEEE Globecom 2010*, Miami, Florida, USA, 2010; 1–5.
11. Zhou S, Gong J, Yang Z, Niu Z, Yang P. Green mobile access network with dynamic base station energy saving, In *Proceedings of ACM Mobicom 2009*, Beijing, China, 2009.
12. Niu Z, Wu Y, Gong J, Yang Z. Cell zooming for cost-efficient green cellular networks. *IEEE Communication Magazine* 2010; **48**(11): 74–79.
13. Wei Y, Zhao Z, Zhang H. Dynamic energy savings in heterogeneous cellular networks based on traffic prediction using compressive sensing, In *Proceedings of IEEE ISCIT 2011*, Hangzhou, China, 2011; 460–465.
14. Li R, Zhao Z, Wei Y, Zhou X, Zhang H. GM-PAB: a grid-based energy saving scheme with predicted traffic load guidance for cellular networks, In *Proceedings of IEEE ICC 2012*, Ottawa, Canada, 2012; 1160–1164.
15. Son K, Kim H, Yi Y, Krishnamachari B. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. *IEEE Journal On Selected Areas in Communications* 2011; **29**(8): 1525–1536.
16. Richter F, Fehske AJ, Fettweis GP. Energy efficiency aspects of base station deployment strategies in cellular networks, In *Proceedings of IEEE VTC 2009 Fall*, Anchorage, Alaska, USA, 2009; 1–5.
17. Cappe O, Moulines E, Pesquet JC, Petropulu A, Xueshi Y. Long-range dependence and heavy-tail modeling for teletraffic data. *IEEE Signal Processing Magazine* 2002; **19**(3): 14–27.
18. Shu Y, Jin Z, Zhang L, Wang L, Yang O. Traffic prediction using FARIMA models, In *Proceedings of IEEE ICC 1999*, British Columbia, Canada, 1999; 891–895.
19. Leland WE, Taqqu MS, Willinger W, Wilson DV. On the self-similar nature of ethernet traffic. *IEEE/ACM Transactions on Networking* 1994; **2**(1): 1–15.
20. IEEE 80216 Boradband Wireless Access Working Group. IEEE 802.16m evaluation methodology document (EMD), July 2008. Available at: http://ieee802.org/16
21. Ashtiani F, Salehi J, Aref M. Mobility modeling and analytical solution for spatial traffic distribution in wireless multimedia networks. *IEEE Journal on Selected Areas in Communications* 2003; **21**(10): 1699–1709.
22. Tutschku K, Tran-Gia P. Spatial traffic estimation and characterization for mobile communication network design. *IEEE Journal on Selected Areas in Communications* 1998; **16**(5): 804–811.
23. Ge X, Yu S, Yoon WS, Kim YD. A new prediction method of α-stable processes for self-similar traffic, In *Proceedings of IEEE Globecom 2004*, Dallas, Texas, USA, 2004; 675–679.
24. Xiang L, Ge X, Liu C, Shu L, Wang C. A new hybrid network traffic prediction method, In *Proceedings of IEEE Globecom 2010*, Miami, Florida, USA, 2010; 1–5.
25. Soule A, Lakhina A, Taft N, Papagiannaki K, Salamatian K, Nucci A, Crovella M, Diot C. Traffic matrices: balancing measurements, inference and modeling, In *Proceedings of ACM SIGMETRICS 2005*, Banff, Alberta, Canada, 2005; 362–373.
26. Zhang Y, Roughan M, Willinger W, Qiu L. Spatio-temporal compressive sensing and internet traffic matrices, In *Proceedings of ACM SIGCOM 2009*, Barcelona, Spain, 2008; 267.
27. Falvo MC, Gastaldi M, Nardecchia A, Prudenzi A. Kalman filter for short-term load forecasting: an hourly predictor of municipal load, In *Proceedings of IASTED International Conference on ASM 2007*, Palma de Mallorca, Spain, 2007; 364–369.
28. Donoho D. Compressed sensing. *IEEE Transaction on Information Theory* 2006; **52**(4): 4036–4048.
29. Xu W, Lin J, Niu K, He Z. A joint recovery algorithm for distributed compressed sensing. *Transactions on Emerging Telecommunications Technologies* 2012; **23**(6): 550–559.
30. Lakhina A, Papagiannaki K, Crovella M, Diot C, Kolaczyk ED, Taft N. Structural analysis of network traffic flows, In *Proceedings of ACM SIGMETRICS 2004*, New York, USA, 2004; 61–72.
31. Li Z, Zhu Y, Zhu H, Li M. Compressive sensing approach to urban traffic sensing, In *Proceedings of IEEE ICDCS 2011*, Minneapolis, Minnesota, USA, 2011; 889–898.

32. Recht B, Fazel M, Parrilo PA. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization, 2007. ArXiv preprint arXiv:0706.4138.
33. Gong J, Zhou S, Niu Z, Yang P. Traffic-aware base station sleep-ing in dense cellular networks, In *Proceedings of IEEE IWQoS 2010*, Beijing, China, 2010; 1–2.
34. Bradley SP, Hax AC, Magnanti TL. Chapter 9: Integer programming. In *Applied Mathematical Programming*, Bradley SP, Hax AC, Magnanti TL (eds). Addison-Wesley Press: Boston, Massachusetts, USA, 1997.
35. Goemans M, Williamson D. The primal-dual method for approximation algorithms and its application to network design problems. In *Approximation Algorithms for NP-Hard Problems*, Hochbaum D (ed.). PWS Publishing Co.: Boston, Massachusetts, USA, 1997; 144–191.
36. Chinneck JW. Chapter 12: Integer/discrete programming via branch and bound. In *Practical Optimization: A Gentle Introduction*, Chinneck JW (ed.). Systems and Computer Engineering, Carleton University: Ottawa, Ontario, Canada, 2006.
37. The mosek optimization tools manual, 2011. Available at: http://docs.mosek.com/6.0/tools/index.html?id=2

Trans. Emerging Tel. Tech. **25**:468–478 (2014) © 2012 John Wiley & Sons, Ltd.
DOI: 10.1002/ett