EUROPEAN COOPERATION
IN THE FIELD OF SCIENTIFIC
AND TECHNICAL RESEARCH

EURO-COST

COST 2100 TD(09)724
Braunschweig, Germany
February 16-18, 2009

SOURCE: Dpt. Ingeniería de Comunicaciones
University of Málaga, Málaga, Spain

Performance Analysis of Location Area Re-Planning in a Live GERAN System

M. Toril
Dpt. Ingeniería de Comunicaciones
ETSI Telecomunicación, Universidad de Málaga
Campus Universitario de Teatinos, s/n
E-29071 Málaga
Spain
Phone: +34 952137120
Fax: +34 952132027
Email: mtoril@ic.uma.es

Performance Analysis of Location Area Re-Planning in a Live GERAN System

M. Toril, V. Wille, S. Luna-Ramírez, K. Jarvinen

***Abstract*— In mobile networks, Location Area (LA) planning has a strong influence on network performance. In this paper, the problem of optimising an existing LA plan based on statistical measurements in a live GSM-EDGE Radio Access Network (GERAN) is addressed. Unlike prior work, the main focus in this paper is on estimating the impact of changes in the LA plan on the air-interface signalling load. For this purpose, a performance model is proposed to estimate changes in the traffic load in dedicated signalling channels. With such a model, it is possible to estimate the change that a new LA plan will cause in terms of the required number of signalling channels. Model assessment is carried out based on measurements taken before and after changing the LA plan of a limited geographical area. Then, a comparison of several LA re-planning algorithms is performed over measurements in a larger area by using the proposed performance model. Results show that the signalling traffic (and, hence, the required signalling resources) can be significantly reduced by improving the existing LA plan in a live GERAN system.**

***Index Terms*— mobile network, location area, network management, optimisation, signalling, location update, paging, graph partitioning**

I. INTRODUCTION

In recent years, the size and complexity of mobile communications networks have increased exponentially, which makes it very difficult for operators to manage their networks. One of the basic, yet most time-consuming task in operators' activities is the mere structuring of the network. Network structure is key to providing an adequate quality of service with minimum infrastructure. This motivates operators to invest much time and effort in optimizing network structure frequently. For scalability, mobile networks are given a hierarchical structure. Thus, network structuring aims to find the best clustering of elements in lower layers (often, cells or base stations) to be assigned to controllers (or control areas). Examples of these clustering problems are the assignment of base stations to base station controllers, [1], packet control units, [2], mobile switching centres, [3], and location and routing areas, [4].

M. Toril and S. Luna-Ramírez are with the University of Málaga, Communications Engineering Dept., Málaga, Spain (email: {mtoril, sluna}@ic.uma.es).

V. Wille and K. Jarvinen are with Nokia Siemens Networks, Performance Services, Huntingdon, Cambs., United Kingdom (email: volker.wille@nsn.com).

In GSM-EDGE Radio Access Network (GERAN), Location Area (LA) planning has a strong impact on the air-interface signalling load. In current networks, the coverage area is divided into several LAs, each comprising a large number of cells. Any terminal in idle mode crossing a LA boundary performs a Location Update (LU) request (referred to as *mobility LU*). Likewise, idle terminals update their location periodically (referred to as *periodic LU*). In parallel, the network notifies the user about an incoming call via a paging message in all cells belonging to the same LA, [5]. Thus, operators have to define the optimal LA size and shape. An ideal LA plan should minimise the number of users moving between LAs to reduce the number of mobility LUs. At the same time, the LA plan should keep the total number of mobile terminated calls per LA below certain limits in order not to exceed the paging capacity of cells in the LA.

As many other clustering problems, the LA planning problem can be formulated as a graph partitioning problem, [6]. Based on this approach, many methods have been proposed to determine the optimal grouping of cells into LAs based on network measurements (e.g., [4][7]). These methods rely on Hand-Over (HO) statistics for users in connected mode, as current networks do not provide mobility statistics for idle users or users in packet-transfer mode. In all methods, the aim is to minimise the number of HOs between cells in different LAs in the hope that the signalling load is thus also reduced. In [8], it was shown that the number of incoming inter-LA HOs and the number of LUs in a cell are highly correlated. However, experience shows that the number of HOs is not directly linked to the load in signalling channels. For instance, users in idle mode, which generate most LU requests, might not follow exactly the same movement pattern as users in connected mode, for which HO statistics are collected. In [9], the authors have recently shown that, in hierarchical networks, HOs fail to explain mobility LUs in cells of lower layers (e.g., microcells and picocells) due to the influence of traffic management algorithms on HO statistics. Even if such a correlation exists, the signalling load on the air interface is not exactly proportional to the number of mobility LUs. Note that mobility LU traffic is only a fraction of signalling traffic and its relative weight depends on the magnitude of the other traffic components.

This paper investigates the relationship between changes in inter-LA HOs and signalling load due to a new LA plan in GERAN. As a result, a network performance model is proposed to estimate the change that a new LA plan will cause in terms of the required number of signalling channels based on statistics in the network management system. Such a model will enable GERAN operators to predict the benefit that a new LA plan will provide in terms of "saved" time slots. Model assessment is carried out by analysing changes in the LA plan of a limited geographical area in a live GERAN system. Based on the proposed model, a performance comparison of several LA re-planning algorithms is then performed over measurements in a larger network area.

To the author's knowledge, no previous work has investigated the impact of changing the LA plan in a live network. Thus, the main contributions of this paper are: a) a new proof of the correlation between HOs and mobility LUs, which, unlike in other studies, is based on differential (i.e., before and after) rather than absolute measurements; b) a model to estimate signalling traffic changes caused by a new LA plan from measurements in the network management system; and c) a comparison of several classical graph partitioning algorithms over graphs built from real network statistics.

The rest of the paper is organised a follows. Section II introduces the LA planning problem. Section III describes several LA planning methods. Section IV presents the model to quantify the impact of LA changes on the required signalling resources. Section V presents the performance analysis based on measurements taken from a live network. Finally, Section VI presents the main conclusions of the study.

II. Location Area Planning in GERAN

In this section, the LA planning problem is described from the operator's perspective. First, the signalling capacity of current GERAN systems is estimated. Then, the planning problem is formulated analytically and the current practice is described.

A. Signalling Capacity in GERAN

In GERAN, signaling capacity largely depends on the capacity of two radio interface channels: the Common Control CHannel (CCCH) and the Stand-alone Dedicated Control CHannel (SDCCH), [5]. The CCCH is made up of four logical channels, amongst which are the Paging CHannel (PCH) and the Access Grant CHannel (AGCH). The PCH is used to broadcast paging messages to notify idle users about an incoming call, while the AGCH is used to assign resources to a user requesting access to the network. The SDCCH is involved in call set-up, mobile station registration and location update, and data services, i.e., Short Message Service (SMS), Multimedia Messaging Service (MMS), Wireless Application Protocol (WAP) and Suplementary Services (SS).

The capacity of the PCH and the AGCH depends on the configuration of the CCCH, [10]. In this work, a non-combined SDCCH configuration for the CCCH is assumed network wide. In this configuration, CCCH and SDCCH are allocated in different Time Slots (TSLs), but PCH and AGCH share CCCH resources of TLS 0 in the beacon transceiver, consisting of 9 blocks in a TDMA multiframe. Normally, priority on the CCCH is given to the PCH, but several CCCH blocks can be reserved for the AGCH, only. Thus, the capacity of the PCH depends on the number of blocks reserved for AGCH, which can be up to 7. Table I presents the maximum hourly capacity of the PCH for different configurations of the AGCH in the non-combined SDCCH case. In the table, it is taken into account that each paging request message can page 2, 3 or 4 Mobile Stations (MSs), depending on the type of paging request message used by the Base Station (BS). Note that the latter is selected dynamically and, hence, values in the tables are upper bounds. A preliminary analysis has shown that operators often reserve 1 block for AGCH and most paging request messages are of Type 2 and 3 (i.e., 3 and 4 paged mobiles per message). The resulting capacity values are highlighted in grey. Hereafter, it is assumed that the maximum PCH capacity is 400000 individual paging requests per hour.

The capacity of the SDCCH depends on the number of TSLs devoted to it on a per-cell basis. In a non-combined SDCCH configuration, each TSL comprises 8 SDCCH sub-channels. Therefore, the number of SDCCH sub-channels in a cell is a multiple of 8. In some networks, one of these sub-channels is used for the Cell Broadcast CHannel (CBCH), in which case the capacity takes values in the set $8i - 1$, $i \in \mathbb{N}_+$.

B. Problem Formulation

Because of the sheer size of mobile population and the limited capacity of the PCH, operators have to sub-divide their network into several Location Areas (LAs), each comprising a large number of cells. Within an LA, all BSs will receive the same paging messages. The number of periodic LUs is related to the number of MSs attached to a BS. In contrast, the number of mobility LUs depends on the location of a BS within a LA, as BSs at the border of LAs will also carry mobility LUs from MSs entering the LA.

To optimise system performance, operators have to choose the LA of each BS in the network carefully. If the network is divided into many small LAs, the paging traffic per LA will be small, but the mobility LU traffic will be high. This means that more SDCCH resources have to be deployed on BSs at the LA border to carry the

TABLE I
MAXIMUM HOURLY CAPACITY OF THE PCH FOR DIFFERENT CONFIGURATIONS OF THE AGCH.

Reserved AGCH blocks	PCH blocks	Nbr. of MSs with Pag. Req. Type 1	Nbr. of MSs with Pag. Req. Type 2	Nbr. of MSs with Pag. Req. Type 3
0	137628	275257	412886	550514
1	122336	244673	367009	489346
2	107044	214089	321133	428178
3	91752	183504	275257	367009
4	76460	152920	229381	305841
5	61168	122336	183504	244673
6	45876	91752	137628	183504
7	30584	61168	91752	122336

mobility LU signalling traffic. If the network is divided into fewer but larger LAs, then the likelihood of MSs crossing LA borders is comparatively smaller. Therefore, the mobility LU traffic will be lower, thus requiring less SDCCH capacity. However, the paging traffic per LA will be high compared to the case where the network is divided into many LAs.

In the ideal LA plan, the capacity of the paging channel determines the size of the LA, i.e., how many MSs (or equivalently BSs) should be served by a LA. At the same time, the LA plan should also ensure that as few BSs as possible are located at a LA border so that the number of mobility LUs is minimised. Thus, the total SDCCH traffic is reduced and SDCCH congestion periods are avoided. Likewise, the required SDCCH capacity in terms of TSLs devoted to signalling is reduced. Minimising the SDCCH TSL requirements is desirable as each TSL devoted to carry mobility LU traffic cannot be used to carry revenue generating user traffic. Hence, schemes to optimise LA plans are of direct benefit to operators. In [11], it was shown that SDCCH resources are often overdimensioned. In these conditions, it is the change in the SDCCH TSLs requirements that ultimately relates to the benefit of a new LA plan as it indicates how many SDCCH TSLs can be converted from carrying signalling traffic to carrying actual payload.

The definition of LAs can be formulated as a graph partitioning problem, [6]. In this approach, the optimised network area is modeled by a non-directed weighted graph. The vertices of the graph represent the cells in the network, while the undirected edges represent the adjacencies defined by the operator for HO purposes. Since edges are non-directed, it is assumed that adjacencies are bi-directional in nature (i.e., the adjacency between two cells is unique, regardless of the direction of user movement). The weight of each vertex, ω_i, denotes the number of paging requests due to mobile terminated calls in the cell, while the weight of each edge, γ_{ij}, denotes the number of idle users moving between cells on its ends. The latter can be estimated from HO statistics. The former must be estimated from counters of paging requests and mobile terminated calls, as

$$\omega_i = PA_i \frac{MTC_i}{\sum\limits_{j \in LA_i} MTC_j} \tag{1}$$

where PA_i is the number of paging attempts in the LA of cell i, LA_i, and MTC_i is the number of mobile terminated calls in cell i.

The partitioning of the graph, performed by grouping vertices into disjoint subsets, referred to as *subdomains*, reflects the assignment of cells to LAs. The resulting partition defines a set of edges that join vertices in different subdomains, δ, which is referred to as a cut. Formally, the LA planning problem can be modeled as a *bounded, min-cut problem*, [12], described as follows. Let G be an undirected weighted graph G=(V,E), consisting of a set of vertices V and edges E, vertex weights ω_i and edge weights γ_{ij}. Let B_{aw} be a real number, such that $0 < B_{aw} \leq \sum\limits_{i=1:|V|} \omega_i$, where $|V|$ is the number of vertices in the graph. The problem stands for the partition of V into an arbitrary number of subdomains k, S_1, S_2, $\cdots$, S_k, such that

$$\|S_n\| = \sum_{i \in S_n} \omega_i \leq B_{aw} \qquad \forall\, n \in \{1, 2, ..., k\} \tag{2}$$

(i.e., the weight of each subdomain is bounded), and

$$\sum_{(i,j) \in \delta(S_1, \ldots, S_k)} \gamma_{ij} \tag{3}$$

(i.e., the sum of the weights of edges in the cut, referred to as *edge-cut*) is minimised.

It should be pointed out that, in the past, LAs could not cross the boundaries of Mobile Switching Centres (MSCs). Thus, LA planning had to be done on a per-MSC basis and, thus, a whole operator's network comprised several instances of the problem (i.e., one per MSC). Current equipment allows multiple physical

MSC Server Systems to be logically grouped into one MSC Server Systems. Therefore, it is possible that some LAs cross MSC borders. LA planning can be performed network wide instead of on a per-MSC basis. The latter approach is followed in this work.

C. Current Solution Technique

For simplicity, operators have traditionally defined LAs by assigning whole Base Station Controllers (BSCs), and not individual cells, to LAs. In such an approach, the original graph with cell resolution is coarsened to obtain a simplified graph with BSC resolution, where each vertex is a BSC. Thus, the LA plan has BSC resolution, i.e., all cells in the same BSC are assigned to same LA. Also for simplicity, a LA seldom consists of more than one BSC. Despite its inherent simplicity, such an approach does not take full advantage of LA planning. On the contrary, with the limited size of current BSCs, many LAs could be merged without exceeding the maximum paging capacity of cells in the LA.

To cope with the increased LU traffic from non-optimal LA plans, operators have to re-plan the number of SDCCH TSLs on a cell-by-cell basis relatively frequently. For this purpose, a very conservative approach is followed, based on SDCCH congestion statistics. In those cells where SDCCH congestion occurs, new SDCCH TSLs are added. As a result of this reactive approach, SDCCH resources are often over-dimensioned in live networks, [11]. Thus, it is often possible that SDCCH resources can be reduced without affecting network performance. This reduction can be even larger if the peak SDCCH traffic in cells is reduced by improving the current LA plan.

In this work, the benefit of an optimal LA configuration is investigated. For this purpose, the next section describes several methods to design a LA plan taken from the graph partitioning literature.

III. Solution Algorithms

The *bounded, min-cut problem* is known to be NP-complete, [13]. Therefore, in general, it is not possible to compute optimal partitionings for graphs of interesting size in a reasonable amount of time. Thus, several efficient heuristics have been proposed in the literature (for a survey, see [6]). However, unlike in other applications, the size of graphs is small when LA planning is performed with BSC resolution. In this case, exact approaches are also feasible. The following paragraphs outline four graph partitioning algorithms that can be used to re-plan LAs. These methods are the common benchmark against which more sophisticated methods are compared in the graph partitioning literature.

A. Exact Method

As shown in [3], the graph partitioning problem can be formulated as the Integer Linear Programming (ILP) model

$$\text{Min} \quad \sum_{(i,j)\in E} \gamma_{ij}(1-\sum_{n\in N} Z_{ijn}) \tag{4}$$

$$\text{s.t.} \quad \sum_{n\in N} X_{in} = 1, \qquad \forall\, i \in V, \tag{5}$$

$$\sum_{i\in V} \omega_i X_{in} \leq B_{aw}, \qquad \forall\, n \in N, \tag{6}$$

$$Z_{ijn} \leq X_{in}, \qquad \forall\, (i,j) \in E,\; n \in N, \tag{7}$$

$$Z_{ijn} \leq X_{jn}, \qquad \forall\, (i,j) \in E,\; n \in N, \tag{8}$$

$$Z_{ijn} \geq X_{in} + X_{jn} - 1, \;\; \forall\, (i,j) \in E,\; n \in N, \tag{9}$$

$$X_{in} \in \{0,1\}, \qquad \forall\, i \in V,\; n \in N, \tag{10}$$

$$Z_{ijn} \in \{0,1\}, \qquad \forall\, (i,j) \in E,\; n \in N, \tag{11}$$

where X_{in} and Z_{ijn} are binary variables that reflect the assignment of cell i and adjacency (i,j) to LA n, respectively, and $N = \{1, 2, \cdots, k\}$. (4) reflects the goal of minimizing the number of HOs between cells in different LAs. (5) ensures that a cell belongs to only one LA. (6) reflects the paging capacity limit of cells in a LA, B_{aw}. (7)-(9) show the dependence between decision variables by linear constraints and (10)-(11) are binary constraints.

The model (4)-(11) can be solved exactly by the *Branch-and-Cut* (BC) algorithm, [14]. The BC algorithm is a refined enumeration method that discards groups of non-promising solutions without explicitly testing them. For space reasons, the reader is referred to [14] for more details on the BC algorithm.

It should be pointed out that the value of k (i.e., the number of LAs) in the optimal solution is not known a priori. A lower bound, k_{min} can be obtained as

$$k_{min} = \left\lceil \frac{\sum_{i\in V} \omega_i}{B_{aw}} \right\rceil . \tag{12}$$

However, there might not be a valid solution with that number of subdomains. Even if this is the case, the optimal solution might not be a valid solution with the lowest number of subdomains. Therefore, in theory, all values in the interval $[k_{min}, |V|]$ should be tested. Nonetheless, it is clear that the best solutions should have a small number of subdomains. Hence, if only a single trial is to be performed for computational efficiency, the value of k should be selected close to but larger than k_{min} as not to discard the optimal solution. In this work, k has been fixed to the value in the best heuristic solution, which is computed by one of the method described next.

B. Heuristic Methods

1) Local Refinement: The simplest approach to partition a graph is by improving an existing solution. For this purpose, the *Kernighan-Lin* (KL) refinement algorithm, [15], can be used. Given a partition of a graph, the KL algorithm swaps those subsets of vertices in different subdomains that yield the greatest possible edge-cut reduction. The main strength of the algorithm is its ability to escape from local minima, because it explores movements that temporarily increase the edge-cut. In most applications, the variant of the KL algorithm proposed by *Fiduccia and Mattheyses* (FM), [16], is implemented. The FM algorithm differs from the KL algorithm in that it moves only a single vertex at a time instead of swapping pairs of vertices. Thus, the time complexity of each step is reduced.

Unlike classical refinement algorithms, the FM algorithm used in this work can reduce the number of subdomains in the initial solution for increased edge-cut performance, i.e., subdomains can become empty after moving a vertex. Hence, it is not possible to know a priori the number of subdomains in the final solution. This behaviour is also observed in the algorithms described next, since the latter include local refinement algorithms to improve their partitions.

2) Random Greedy Graph Growing Partitioning: When the initial solution displays a poor quality, it is advantageous to build the partition from scratch. In the *Greedy Graph Growing Partitioning* (GGGP) algorithm, [12], k initial seed vertices are chosen. Then, a partition of the graph is built by growing subdomains incrementally around a arbitrary seed vertices. At each step, the growing region with the smallest weight is selected. Then, the algorithms adds the unassigned vertex with the heaviest edge weight to vertices already in the growing region.

Two different strategies can be used to choose the seed vertices in the GGGP algorithm. The first one aims to maximize the average distance among seed vertices in the graph, [17]. More robust partitions are thus obtained at the expense of an increased runtime due to the calculation of the distance between every pair of vertices. In the second strategy, the seed vertices are chosen arbitrarily. To improve the robustness of the method, a limited number of trials can be conducted with randomly selected seed vertices and the best solution in terms of edge-cut is selected, [18]. Such an approach is referred to as *Random GGGP* (RGGGP). Unfortunately, the efficiency of this naive multi-start approach is limited, as the chance to improve a solution quickly diminishes from one iteration to the next, [19]. Nonetheless, the diversity provided by random seed selection proves to be extremely valuable in instances of the graph partitioning problem with few vertices per subdomain.

Again, it should be noted that the optimal value of k is not known a priori. Thus, several trials of R-GGGP with different values of k must be performed.

3) Multi-Level Refinement: In the Multi-Level (ML) approach, [20], the original graph is coarsened by collapsing vertices (often pairs) to reduce the size of the graph. Then, an initial solution is efficiently computed on the coarsened graph. Finally, the solution is progressively uncoarsened to obtain the partition of the original graph. After each uncoarsening step, a local refinement algorithm (often, the FM algorithm) is applied on small portions of the graph that are close to the partition boundary. This technique avoids the problem of local minima in local refinement algorithms, improving both solution quality and runtime. The reader is referred to [20] and [21] for more details about the ML algorithm.

The ML algorithm adopted in this work has several differences with that commonly used in the supercomputing literature, [20]. First, coarsening is performed by selecting the heaviest edges on the graph in each coarsening step, unlike classical approaches, where the heaviest edge of a randomly selected vertex is chosen for computational efficiency. Second, the graph is coarsened until the number of vertices in the coarsened graph coincides with the number of subdomains. Finally, the optimal value of k is not known a priori. Therefore, different values of k must be tested.

IV. SDCCH Performance Model

Before modifying an existing LA plan, operators need to know the exact impact of changes on SDCCH performance. Likewise, operators might want to quantify the number of SDCCH TSLs that could be converted into TSLs for payload traffic if the new LA plan was implemented. In this section, a model is proposed to estimate the change that a new LA plan will cause in terms of the required number of SDCCH TSLs. For this purpose, the peak traffic during the busy hour with the old and new plan is estimated on a per-cell basis from statistics in the network management system. This is aligned to the conservative rule used by operators to plan SDCCH resources.

The input data to the model consists of HO statistics, number of paging requests, number of SDCCH attempts classified per cause, mean SDCCH holding time as well as average and peak SDCCH traffic, gathered with the old plan. Ideally, all statistics (except HO) should be collected on a cell basis. However, current BSCs only reflect the sum of paging requests directed to users in a location area, since the paging algorithm is MSC functionality. Thus, all cells in a LA have the same paging traffic. Using (1), it is possible to estimate the number of paging requests originated per cell. Likewise, all measurements should be collected on the SDCCH Busy Hour (BH). Unfortunately, some operators only gather HO statistics on a daily basis for storage reasons. In this case, some transformation between daily and BH statistics will be needed. Finally, measurements can be collected for several days and some averaging be performed to increase the robustness of estimates.

The model starts by estimating the difference in the total daily sum of HOs coming from cells in other LAs, referred to as *incoming inter-LA HOs* (IIHOs) on a cell basis. For this purpose, a directed graph is built from HO statistics. This graph differs from the graph used to build the new LA plan, as incoming and outgoing adjacencies were aggregated into a single bi-directional adjacency for computational efficiency. Then, the sum of edge weight incident to each vertex from edges in the cut is computed for each plan. By subtracting the amount obtained with the new plan from that of the old plan, the difference in the total daily sum of IIHOs in cell i, $\Delta IIHO_{day}(i)$, is obtained. For convenience, $\Delta IIHO_{day}(i) = IIHO_{day}(i) - IIHO_{day}^{(0)}(i)$, where $IIHO_{day}$ and $IIHO_{day}^{(0)}$ are the values with the new and old plans, respectively.

Second, the difference in the total daily sum of LUs, ΔLU_{day}, is calculated per cell as

$$\Delta LU_{day}(i) = c_{LU/IIHO}\ \Delta IIHO_{day}(i)\ , \tag{13}$$

where $c_{LU/IIHO}$ is a constant reflecting the ratio of idle to connected moving users network wide. In (13), it has been assumed that: a) mobility LUs are proportional to IIHOs, and b) the number of periodic LUs does not change with LA planning. Both assumptions will be shown to be valid in the next section. Note that, unlike prior studies, the model works on differential rather than absolute measurements. Thus, it is expected that the influence of periodic LUs is avoided, without the need for estimating this component of LU traffic explicitly, as in [9]. Also important, any relationship between IIHOs and LUs through a third variable (e.g., cell population) is reduced.

Third, daily LU attempt figures are translated into BH LU figures. Here, it is taken into account that LUs are not distributed the same in all cells during a day. In some cells, all LUs take place during the BH, while in other cells LUs are evenly distributed along the day. Thus, the difference in the number of LUs in the BH is computed as

$$\Delta LU_{BH}(i) = c_{BH/day}(i)\ \Delta LU_{day}(i)\ , \tag{14}$$

where $c_{BH/day}$ is a proportionality constant between $\frac{1}{24}$ and 1 showing the BH-to-day ratio of LU traffic on a cell basis. From this value, the difference in the average SDCCH carried traffic (in Erlangs) is calculated as

$$\Delta A_{c\,BH}(i) = \frac{\Delta LU_{BH}(i)\ MHT_{LU}}{3600}\ , \tag{15}$$

where MHT_{LU} is the mean holding time of LU attempts. Preliminary analysis has shown that MHT_{LU} can greatly vary from network to network. Therefore, instead of using values reported in the literature, a multiple regression analysis is performed on counters of SDCCH attempts per cause and mean SDCCH holding time gathered on a cell basis, as

$$MHT(i) = \sum_c \beta_c \frac{N_c(i)}{3600}, \tag{16}$$

where MHT is the mean SDCCH holding time, N_c is the number of SDCCH BH attempts for cause c, and β_c is the regression coefficient for cause c. For instance, β_{LU} is the estimated value of MHT_{LU} for the entire network. Thus, the new average SDCCH traffic, $A_{c\,BH}$, can be easily computed from the old value, $A_{c\,BH}^{(0)}$, as

$$A_{c\,BH}(i) = A_{c\,BH}^{(0)}(i) + \Delta A_{c\,BH}(i)\ . \tag{17}$$

To convert average traffic values into peak traffic values, a regression curve is built from network statistics. To explain the reason, Fig. 1 depicts a scatter plot of average versus peak SDCCH traffic values in cells of a live network. Each point corresponds to measurements in a cell and hour. It is observed that peak traffic values are highly variable, i.e., for the same value of average traffic, many different peak values are possible. A regression curve, f_{pk2av}, is derived to reflect the main trend. However, the value of the coefficient of determination is small (i.e., R^2 =0.16). Nonetheless, note that the aim of the study is to quantify the overall impact of a new LA plan on a long-term basis, and not the impact on every single cell on an hourly basis. Thus, it is expected that this lack of fit should not affect estimates when aggregated across the network for several days. By means of this curve, the difference in the peak values can be calculated as

$$\begin{aligned} \Delta A_{pk\,BH}(i) &= A_{pk\,BH}(i) - A_{pk\,BH}^{(0)}(i) \\ &= f_{pk2av}(A_{c\,BH}(i)) - f_{pk2av}(A_{c\,BH}^{(0)}(i))\ . \end{aligned} \tag{18}$$

Finally, the difference in the number of required SDCCH TSLs is estimated. A rough estimate could be made by dividing $\Delta A_{pk\,BH}$ by 8, since 1 SDCCH TSL comprise 8 subchannels. A more accurate estimate is obtained by considering that: a) 1 SDCCH TSL

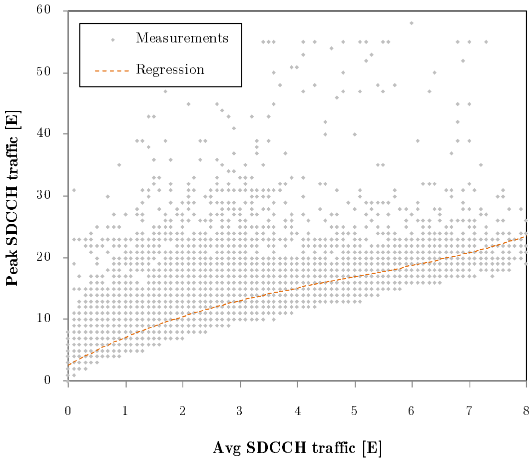


Fig. 1. Relationship between average and peak SDCCH traffic in a live network.

is saved per each multiple of 8 that falls within the interval $[A_{pk_{BH}}, A_{pk_{BH}}^{(0)}]$, even if $\Delta A_{pk_{BH}} < 8$, and b) a minimum of 1 SDCCH TSL is required in the non-combined configuration. Thus, the number of saved TSLs, ΔN_{TSL}, is estimated as

$$\begin{aligned} \Delta N_{TSL}(i) &= \min\left\{1, \left\lceil \frac{A_{pk_{BH}}(i)}{8} \right\rceil \right\} \\ &- \min\left\{1, \left\lceil \frac{A_{pk_{BH}}^{(0)}(i)}{8} \right\rceil \right\} . \end{aligned} \tag{19}$$

It is worth noting that classical analytical queueing models fail to give accurate estimations of SDCCH blocking. A preliminary analysis has shown that, in many cells, SDCCH requests are concentrated in short periods of time, and, hence, the arrival process is not Poisson. Thus, Erlang-B formula tends to under-estimate SDCCH blocking in a live network. More importantly, burstiness greatly varies from cell to cell, which make analytical approaches difficult to apply. This is the main reason not to use a classical birth-death model to estimate peak traffic from average traffic values.

V. Performance Analysis

The proposed performance model is validated from measurements taken before and after changing the LA plan of a live GERAN. Then, the above-described LA planning algorithms are applied to a graph constructed with real data. These algorithms are compared using the proposed SDCCH performance model. The aim of the analysis is to: a) check the validity of the SDCCH performance model, b) identify the best LA planning algorithm among those in the graph partitioning literature, and c) quantify the reduction in the number of TSLs from an optimal LA plan in a live network. For clarity, model assessment is presented first and method assessment is then discussed. In both cases, the preliminary conditions are described first and subsequently the performance results are discussed.

A. Model Assessment

Due to the way operators currently assign LAs, re-allocating a BS to a different BSC causes a change in the LA plan. The analysis presented here is based on a BSC splitting event.

1) Analysis Set-up: The network area consists of 519 BSs, distributed over 177 sites, covering a geographical area of 4000 km^2. Initially, these BSs were controlled by 4 BSCs, each representing a different LA. As a result of BSC splitting, a new BSC was added and 77 BSs of the old BSCs were re-assigned to new BSC. Thus, 96 BSs experienced changes in the incoming inter-LA adjacencies. In these BSs, measurements were gathered one week before and after implementing the changes. During the analysis, differential measurements are computed by comparing the same day of the week in the before and after periods (note that the BH might not be the same). More specifically, the difference is defined as the value with the new plan minus the value with the old plan.

Fig. 2 (a)-(b) show the old and new LA plans of the area, respectively. In the figure, a symbol represents the physical location of a BS. Note that all BSs in a site are represented in the same location. Each of the 4-5 different symbols represent a different LA (i.e., BSC). Lines represent adjacencies between BSs. From the figure, it can be understood that the analysis focuses on BSs adjacent to those re-allocated, represented by a 'x' symbol.

The model is validated by comparing estimates of the peak SDCCH traffic in the BH with the new LA plan from statistics with the old plan. To cope with the high variability of peak traffic measurements, a weekly average of estimates is performed, resulting in a single value per cell. Thus, more robust estimates are obtained.

2) Analysis Results: The following paragraphs show the strengths and weaknesses of the proposed model. First, the different steps in the model are justified and subsequently the overall performance is evaluated.

Fig. 3 (a)-(c) show the relationship between changes of different network performance indicators, as they are used in the model (i.e., (13), (14) and (15), respectively). Each point in the figures represent a day of the week in a cell. A regression line is superimposed, together with the value of the coefficient of determination.

Fig. 3 (a) shows the relationship between changes in IIHOs and LUs. First, it is observed that most samples have positive values. This is aligned to the fact that the total number of IIHOs in the area increased by 18% by changing from 4 to 5 LAs. From the large value of

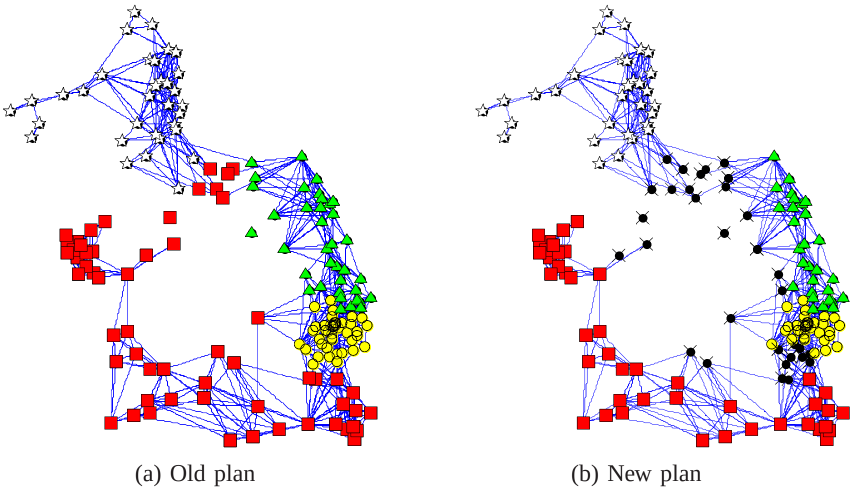


Fig. 2. Location area plans before and after BSC splitting.

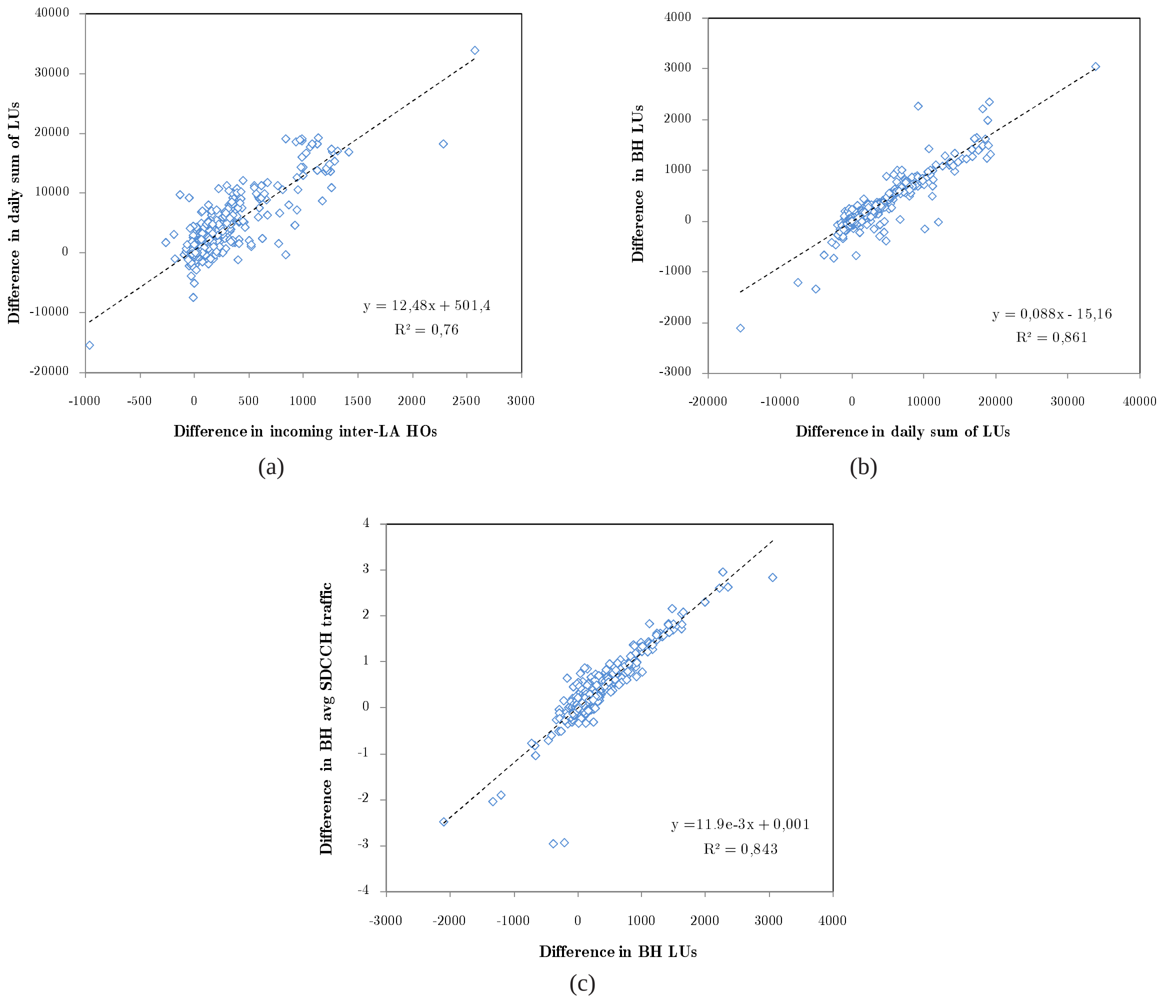


Fig. 3. A scatter plot of changes in several performance indicators.

R^2, it can be deduced that both indicators are strongly correlated. This is proof of the correlation between idle and connected mobility statistics. By using differential measurements, it is expected that only mobility LUs (and not all types of LUs) are included in the comparison. From the line equation, it can be inferred that there are 12.5 idle users per connected user moving. This value is slightly less than that obtained by absolute measurements, [9].

Fig. 3 (b) and (c) show the relationship between changes in the daily sum of LUs and average SDCCH traffic, as well as between number of LUs in the BH and average SDCCH traffic. In the latter indicator, the variation of SDCCH traffic due to attempts other than LU has been subtracted. This term is calculated from attempt counters in the before and after periods and the mean holding time on a per-service basis, obtained from (16). From the figures, it can be concluded that the three indicators are highly correlated. Note that the slope of the regression line in Fig. 3 (c) gives a rough estimation of the value of $MHT_{LU}/3600$. The value of MHT_{LU} thus obtained (i.e., 4.3 seconds) coincides with β_{LU} in (16). It is worth noting that such a tight link does not exist between changes in the average and peak traffic, since the absolute values of these variables are related by a non-linear function (i.e., f_{pk2av}). Therefore, similar changes in the average traffic can lead to totally different changes in peak traffic.

Hitherto, the analysis has justified the steps in the estimation process. The last figure shows the overall accuracy of the model. Fig. 4 (a)-(b) shows a scatter plot of estimates versus measurements for weekly averages of the difference in average and peak SDCCH BH traffic. From the figures, it can be concluded that the model can estimate changes in the average SDCCH traffic caused by the new LA plan quite accurately. However, it only gives a rough approximation of changes in the peak traffic.

B. Method Assessment

In spite of its limitations, the above-described model is used to compare the performance of several graph partitioning methods over real graphs. For this purpose, a larger geographical area is considered.

1) Analysis Set-up: The extended network area covers 100000 km^2, comprising 9807 BSs, 3353 sites, 76 BSCs, 20 MSCs and 71 LAs. Statistical performance measurements were gathered for the whole area during 2 weeks. Table II shows some statistics of the area. From the table, it can be deduced that the number of TSLs devoted to SDCCH is 7% of total network capacity. An upper bound of the number of saved TSLs achieved by optimising the LA plan is the number of SDCCH TSLs in BSs with more than 1 TSL minus the minimum number of SDCCH TSLs in these BSs (i.e., 6902-3237=3665).

TABLE II

STATISTICS IN THE SCENARIO WITH EXISTING CONFIGURATION.

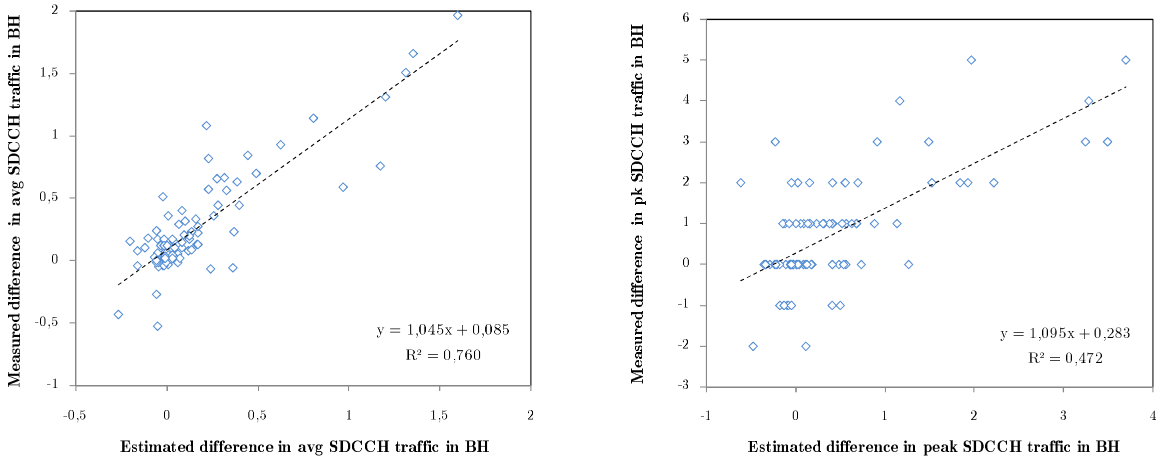
Total nbr. of TSLs	202240
Total nbr. of SDCCH TSLs	13472
Nbr. of BSs with $N_{TSL} > 1$	3237
Total nbr. of SDCCH TSLs in BSs with $N_{TSL} > 1$	6902
Maximum BH paging load per LA [req./h]	134959
Total sum of BH paging load [req./h]	5208060
Total sum of BH Avg. SDCCH traf. [E]	11294.2
Total sum of Pk BH SDCCH traf. [E]	71300

Likewise, it is observed that the maximum paging load per LA (i.e., 134959) is far from the maximum capacity value (i.e., 400000). This evidence clearly indicates that the size of LAs can be made at least three times larger without compromising paging performance. From the sum of paging load, it can be deduced that the minimum number of LAs is 14.

Although the area consisted of several MSCs, a single instance of the LA planning problems is considered, as it is assumed that the core network includes the feature that allows LAs to cross MSC borders. The network graph is built from statistics collected during 2 weeks. Edge weights are computed by scaling the number of HOs in 2 weeks by a factor 1/14 to reflect the daily average. Vertex weights are the maximum number of paging requests originated in cells during the SDCCH BH in 2 weeks. As an optimization constraint, the sum of paging requests from cells in a LA must be below 400000 (i.e., B_{aw}=400000).

During the analysis, several methods are compared. To quantify the improvement of a new LA plan, the initial operator solution (denoted as OI for Operator Initial) is evaluated first. Then, several graph partitioning algorithms are tested: a) the BC algorithm over the ILP model of the problem (denoted as BC), b) the FM refinement of the operator solution (denoted as FM), c) the traditional ML refinement algorithm (denoted as ML), and d) the RGGGP algorithm, based on the repeated use of GGGP with random seed selection and FM refinement. The exact method uses the BC algorithm in SCIP, [22], whereas the heuristic methods were implemented from scratch in Matlab, [23]. In all heuristic methods, the number of passes in the FM refinement algorithm is 4. All methods are tested with different target values of k, ranging from 14 to 20, and the best solution is selected. Note that FM and ML are deterministic and, therefore, a single run is performed per value of k. In contrast, RGGGP is randomised and hence produces a different solution for each different random seed. Reported values for RGGGP correspond to the best solution found in 100 attempts per value of k.

During the comparison, heuristic methods are applied to the graphs of site and BSC resolution. The exact



(a) Average SDCCH traffic

(b) Peak SDCCH traffic

Fig. 4. A scatter plot of measurements versus estimates of changes in SDCCH traffic.

method is only tested on the graph of BSC resolution for computational efficiency. Performance assessment is based on solution quality in terms of edge-cut. By using the proposed performance model, edge-cut figures are translated into changes in SDCCH traffic and saved TSLs. Time efficiency is not considered in the analysis as it is not a critical issue, since LA re-planning is performed, at most, once a week.

2) Analysis Results: Table III presents the main results obtained by the methods on the graph of the extended area. For clarity, the methods have been grouped by the granularity of the graph. For each method, the edge-cut, the maximum subdomain weight and the number of subdomains are shown (i.e., total daily sum of IIHOs, maximum BH paging load and number of LAs, respectively). From the table, it is clear that any of the algorithms can reduce the edge-cut of the initial operator's solution by fully exploiting the paging capacity limit, regardless of graph resolution. On the BSC-level graph, RGGGP and BC are the best method in terms of edge-cut. Not shown is the fact that RGGGP finds the optimal solution in minutes while BC takes hours. Specifically, RGGGP can reduce the edge-cut of the initial solution by 73%. Its superiority over other heuristics is due to the small number of vertices per subdomain, which makes balancing subdomain weights more complicated. In this situation, the diversity provided by random seed selection is extremely valuable. In contrast, on the site-level graph, ML outperforms FM and RGGGP. This is due to the fact that, as the graph gets larger, the solution space becomes larger and the optimisation surface has many local minima. Nonetheless, the surface has a globally convex structure, where the best local minima are clustered in a central

TABLE III

PERFORMANCE OF PARTITIONS BUILT BY METHODS.

Method	Resolution	Edge-cut [$\cdot 10^6$]	Max. subd. wgt.	Nbr. of subd.
OI	BSC	29.66	134959	71
FM	BSC	8.35	394896	16
ML	BSC	8.69	395508	16
RGGGP	BSC	7.94	394896	15
BC	BSC	7.94	394896	15
FM	Site	20.21	126224	70
ML	Site	4.13	397076	16
RGGGP	Site	7.87	384020	15

position of the solution space. Thus, FM and RGGGP get trapped in the local minima. Multiple trials in RGGGP do not alleviate this problem, as random local minima in the graph partitioning problem tend to all have average quality and little variance, [19]. This problem is solved in ML by coarsening the graph first. Thus, the optimisation surface is smoothed by eliminating many local minima, which makes local refinement algorithms more effective, [21].

The rest of the analysis is focused on the best heuristic methods, i.e., RGGGP with BSC resolution and ML with site resolution. Table IV presents the expected impact of the new plans on SDCCH performance. The performance of the the old LA plan is also included in the table. For a fair comparison, the total number of SDCCH TSLs with the existing plan is estimated (as in the other methods) by the sum of peak traffic in the area during the BH, instead of using the value currently configured by the operator, reported in Table II (i.e., 13597 instead of 13472). Results show that re-planning LAs can reduce the average SDCCH BH traffic by up to 28%. Most of the benefit (i.e., 85%) is already obtained with the solution

TABLE IV
ESTIMATED SDCCH PERFORMANCE OF BEST HEURISTIC METHODS.

Method	Resolution	Avg. SDCCH BH traf. [E]	Diff. Avg. SDCCH BH traf. [E]	Peak SDCCH BH traf. [E]	Diff. peak SDCCH BH traf. [E]	Nbr. of SDCCH TSLs	Diff nbr. of SDCCH TSLs
OI	BSC	11294.2	-	71300.0	-	13597	-
RGGGP	BSC	8667.9	-2626.3	61814.1	-9485.9	12638	-959
ML	Site	8167.2	-3127.0	58945.8	-12354.2	12368	-1229

with BSC resolution. The reduction in the peak SDCCH traffic is not so large (i.e., 17%). This is due to the fact most changes occur in cells with a large average traffic value, where sensitivity to changes is lower, as shown in Fig. 1. Nonetheless, the reduction in peak SDCCH traffic is noticeable. Finally, it is observed that the number of "saved" TSLs can be as low as 10%. This is partly due to the fact that only 33% of the BSs have more than 1 TSL. Specifically, the best of LA plans considered (i.e., ML with site resolution) can obtain 33% of the maximum achievable reduction reported in Table II (i.e., 3665). This figure is remarkable, as LU traffic is only 50% of the SDCCH traffic in live networks.

VI. CONCLUSIONS

In future cellular networks, higher traffic demand and smaller cell size will make network structuring a complex issue. Therefore, operators will need automatic network structuring methods. For optimal performance, these methods should use accurate mobility models and effective clustering algorithms. In this paper, the problem of optimising an existing LA plan in a live GERAN based on statistical measurements has been addressed. Unlike prior work, the main focus has been on estimating the impact of changes in the LA plan on the signalling load. For this purpose, a network performance model has been proposed to estimate changes in the peak traffic on the SDCCH. Model assessment has been carried out based on measurements before and after changing the LA plan of a limited area during a BSC splitting event. Based on the proposed model, a comparison of several LA re-planning algorithms has been carried out over graphs built from measurements in a larger area.

Results show that the proposed model can estimate the changes in the average SDCCH traffic caused by a new LA plan accurately. Based on this model, it has been shown that the average SDCCH traffic can be reduced by up to 28% by proper planning of LAs. However, such a reduction only implies a reduction in the required number of SDCCH TSLs by 10%. Hence, it can be concluded that the benefit from re-planning LAs is reducing SDCCH blocking, rather than assigning additional resources for payload traffic.

Although the models and methods presented here have been conceived for GERAN, they can be easily extended to other radio access technologies. Future work is to develop more sophisticated SDCCH traffic models that reflect correlated arrivals and can be tuned on a cell basis.

REFERENCES

[1] M. Toril and V. Wille, "Optimization of the assignment of base stations to base stations controllers in GERAN," *IEEE Communications Letters*, vol. 12, no. 6, pp. 477–479, June 2008.

[2] M. Toril, V. Wille, and R. Barco, "Optimization of the assignment of cells to packet control units in GERAN," *IEEE Communications Letters*, vol. 10, no. 3, pp. 219 – 221, Mar 2006.

[3] A. Merchant and B. Sengupta, "Assignment of cells to switches in PCS networks," *IEEE/ACM Transactions on Networking*, vol. 3, no. 5, pp. 521–526, Oct 1995.

[4] J. Plehn, "The design of location areas in a GSM-network," in *Proc. 45th IEEE Vehicular Technology Conference*, Jun 1995, pp. 871–875.

[5] M. Mouly and M.-B. Pautet, *The GSM system for mobile communications*. Cell & Sys, 1992.

[6] K. Schloegel, G. Karypis, and V. Kumar, "Graph partitioning for high performance scientific simulations," in *CRPC Parallel Computing Handbook*, J. Dongarra, I. Foster, G. Fox, K. Kennedy, and A. White, Eds. Morgan Kaufmann, 2000.

[7] I. Demirkol, C. Ersoy, M. U. Caglayan, and H. Delic, "Location area planning and cell-to-switch assignment in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 3, no. 3, pp. 880–890, May 2004.

[8] T. Kurner and A. Hecker, "Performance of traffic and mobility models for location area code planning," in *Proc. 61st IEEE Vehicular Technology Conference*, vol. 4, Jun 2005, pp. 2111–2115.

[9] M. Toril, S. Luna, V. Wille, and R. Skehill, "Analysis of user mobility statistics for cellular network re-structuring," accepted in 69th IEEE Vehicular Technology Conference, Apr 2009.

[10] T. Halonen, J. Melero, and J. Romero, *GSM, GPRS & EDGE Performance: Evolution Towards 3G/UMTS*. J. Wiley & Sons, 2002.

[11] S. Pedraza, V. Wille, M. Toril, R. Ferrer, and J. Escobar, "Dimensioning of signaling capacity on a cell basis in GSM/GPRS," in *Proc. 54th IEEE Vehicular Technology Conference*, vol. 1, April 2003, pp. 155–159.

[12] R. Krishnan, R. Ramanathan, and M. Steentrup, "Optimization algorithms for large self-structuring networks," in *Proc. INFOCOM '99*, vol. 1, Mar 1999, pp. 71–78.

[13] M. Garey and D. Johnson, *Computers and Intractability: A Guide to NP-Completeness*. California: W.H. Freeman & Co., 1979.

[14] G. Nemhauser and L. Wolsey, *Integer and Combinatorial Optimization*. John Wiley & Sons, 1999.

[15] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell System Technical Journal*, vol. 49, pp. 291–307, 1970.

[16] C. Fiduccia and R. Mattheyses, "A linear time heuristic for improving network partitions," in *Proc. 19th ACM/IEEE Design Automation Conference*, 1982, pp. 175–181.

[17] R. Ramanathan and M. Steenstrup, "Hierarchically-organized, multi-hop mobile wireless networks for quality-of-service support," *Mobile Networks & Applications*, vol. 3, no. 2, pp. 101–119, Aug 1998.

[18] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 359–392, 1998.

[19] K. D. Boese, A. Khang, and S. Muddu, "A new adaptive multi-start technique for combinatorial global optimizations," *Operation Research Letters*, vol. 16, pp. 101–113, 1994.

[20] G. Karypis and V. Kumar, "Multilevel k-way partitioning scheme for irregular graphs," *Journal of Parallel and Distributed Computing*, vol. 48, no. 1, pp. 96–129, 1998.

[21] C. Walshaw, "Multilevel refinement for combinatorial optimisation problems," *Annals of Operations Research*, vol. 131, pp. 325–372, 2004.

[22] T. Achterberg, "SCIP - a framework to integrate constraint and mixed integer programming," Zuse Institute Berlin, Tech. Rep. TR 04-19, 2004, http://www.zib.de/Publications/abstracts/ZR-04-19/.

[23] The MathWorks, Inc., "Getting started with Matlab (version 6)," July 2002, available at http://www.mathworks.com.