This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2011 proceedings.

Mobility-Aware Streaming Rate Recommendation System

Tarik Taleb[1], Abdelhakim Hafid[2], and Apollinaire Nadembega[2]
[1] NEC Laboratories Europe, taleb@neclab.eu
[2] University of Montreal, {ahafid,nadembea}@iro.umontreal.ca

***Abstract*—In mobile multimedia streaming services, important requirements consist of the support of service continuity, the guarantee of acceptable Quality of Service (QoS) and insurance of steady Quality of Experience (QoE). How to get a uniform data exchange rate during the entire (or partial) course of a streaming service while a user is on the move is an important challenge. Generally speaking, the streaming rate of a multimedia service may heavily fluctuate due to the unavailability or deficiency of resources along the movement path of a user. To cope with this challenge, this paper proposes a framework that integrates user mobility prediction models with resource availability prediction models to keep a constant or less fluctuating streaming rate and to ultimately ensure steady QoE. Simulations are conducted to evaluate the performance of the proposed framework in achieving its design objectives and encouraging results are obtained.**

I. Introduction

For mobile multimedia streaming services, important requirements consist of the support of service continuity, the guarantee of acceptable Quality of Service (QoS), and insurance of steady Quality of Experience (QoE). Whilst there has been a large body of research work on the two first requirements, there is a little on QoE. Indeed, regarding QoE, due to users' mobility, mobile users freely, and sometimes frequently, change their points of attachment to the network; along the trajectory of mobile users, the amount of bandwidth available at the different points of attachment may vary. This bandwidth disparity can be due to differences in traffic load in the traversed wireless cells. To cope with these bandwidth disparities, mobile terminals are forced to adjust their data exchange rates whenever the available bandwidth varies; indeed, a mobile terminal generates less traffic to avoid packet drops in case less bandwidth becomes available and generates more traffic to ensure efficient network resource utilization in case more bandwidth becomes available [1]. Adjustment in the data exchange rate may sometimes lead to significant variation in the perceived QoS, which ultimately impacts the overall QoE, and the credibility of the whole service. Ideally, a mobile user should be able to get a uniform exchange rate during the entire (or partial) course of the service and his/her movement; this rate should not exceed the minimum available bandwidth along his/her trajectory. This should be beneficial for both users and service providers. From the customer's perspective, it is beneficial as users will not experience frequent changes in the streaming rates and thus will not perceive a major change in QoE. It is also beneficial for users in case the adopted pricing model charges users for only the bandwidth they have indeed consumed (e.g., packet count based). Indeed, in such schemes, it will be better for customers to receive the streaming service at a constant optimal bandwidth from the start of the service, rather than having the service initiated at a high rate and ending up later at lower rates. At the service provider side, the system scalability can be improved as savings in the network resources become possible and more users can be then served, not to mention that when users are satisfied with the service, more subscribers join the service, resulting in higher revenues.

In this paper, we propose to exploit existing tools and methods for the prediction of user's mobility and bandwidth availability in order to build a framework that reflects mobility awareness in the initial bandwidth recommendation to mobile users; the objective is to avoid frequent changes in the streaming rates and to ultimately ensure acceptable QoE.

The remainder of this paper is structured as follows. Section II provides a survey on existing tools for the prediction of mobility patterns and the assessment of bandwidth availability. The proposed solution along with the envisioned mobile network architecture is described in Section III. Section IV evaluates the performance of the proposed solution and showcases its potential in achieving its design objectives. The paper concludes in Section V.

II. Related Work

According to [4], mobility models should emulate real life mobility in a reasonable way; therefore, they should be associated with a specific place. Mobility models can be classified into two groups: (1) Random-based mobility models; and (2) non-random based models. Random based models are not realistic and thus are not able to emulate the real-life mobility of users. The non-random based models are more suitable to model the mobility of users; in general, they take into account three facts: temporal dependency (constrains of physical laws; e.g., speed), spatial dependency (constraints of neighbouring nodes), and geographic restriction (constraints of the environment; e.g., highways). In [3], three basic types of mobility model are presented: (1) the city area model (used for location management); (2) the area zone model (used for radio resources management scheme); and (3) the street unit model (used for users' mobility behaviour). In this paper, we make use of street unit model since our objective is to provide mobile users with the most optimal rate that provides acceptable QoE.

Most existing mobility models are based on historical data of motion or mobility trace files of the mobile users; more

978-1-4244-9268-8/11/$26.00 ©2011 IEEE

specifically, they make use of the historical data to predict, using different schemes, the movements of users. The authors in [2], [5], abd [6] use Hidden Markov Model (HMM) while the authors in [8] use Bayesian network theory to compute the probability of next destination (e.g., next wireless cell). In [7], Kalman filter is used to extract parameters, such as speed and pause time, from real user traces; it is reported that these parameters follow a log-normal distribution and depend on roads and walkways. These techniques are based on an assumption that the user's movements follow a specific pattern and exhibit some regularity. In this case, a training phase is first required during which regular movement patterns are detected and stored. User's movement behaviour may be highly uncertain and assumptions about user's movement patterns should be made with utmost care. Therefore, whenever the user is located in new locations or when there is a slight change in the user's mobility patterns, the accuracy of the prediction considerably suffers.

Another set of mobility models is based on users' knowledge or their habits. Samaan et al. [9] apply the Dempster-shafer' theory to the knowledge of user's preferences and goals to predict his/her mobility; they did not make any assumption about the availability of users' movement history. The authors in [10] and [11] apply the social theory to the structure of the relationships among individual users to predict their movements while the authors in [12] and [13] define mobility models based on daily planned activities; they assume that users move from home to work, from work to restaurant, from restaurant to work, from work to leisure, and return home in the evening.

Bandwidth prediction management is crucial in providing QoS. However, most existing contributions focus only on handover from one cell to a neighbouring cell and not on the whole trajectory of the user from the start to the end of the requested service. Zhou et al. [14] study the performance of mobile real time service in 802.11 multi-hop network; they report that uncertain handoff latencies and lack of QoS guarantee are the main performance bottlenecks for real-time applications; thus, predicting the next handoff will help resolve these bottlenecks. In [15], a scheme is proposed to improve resource reservation performance and call admission control for cellular networks; bandwidth is allocated to neighbouring cells based on mobility prediction. To support on-going calls during handoff, when needed, bandwidth is borrowed following the prediction of non-conforming calls and existing adaptive calls without impacting the minimum QoS guarantees of these calls. Li et al. [16] propose the integration of RSVP and a flow reservation scheme in wireless LANs in order to provide end-to-end solution for QoS guarantee in wired-cum-wireless networks. They developed an efficient handoff scheme that considers both the requested flow rate demand and network resource availability for continuous QoS support. The scheme is based on renegotiating QoS when moving from one cell to a neighbouring one. However, it does not guarantee the same QoS across all the cells along the trajectory of the flow movement.

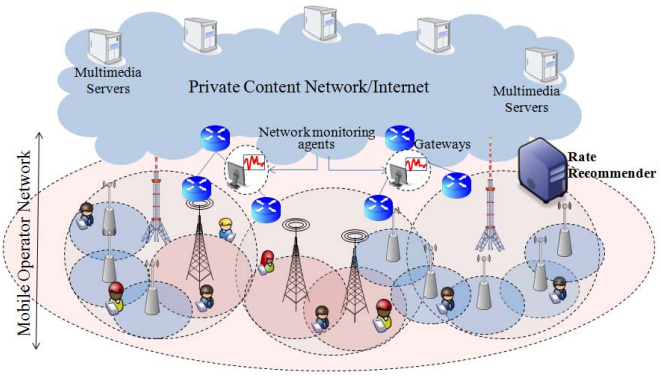

Fig. 1. The envisioned network topology.

In this paper, we propose to use any existing mobility prediction model combined with bandwidth availability measurement methods (for profiling the usage of network resources) in order to compute the most optimal rate that a mobile user can have for the duration of the requested service.

III. PROPOSED SCHEME

The network architecture and its components are conceptually depicted in Fig. 1. The figure portrays a network consisting of two parts, fixed and mobile operator networks, which are inter-connected via gateways. The wired part can be the Internet, a large scale peer-to-peer network, or any private content network (e.g., Akamai) comprising a number of media servers with a wide library of multimedia contents. The mobile operator part of the network consists of a number of wireless domains. A wireless domain comprises a number of access points, using the same or different wireless access technologies (i.e., 4G networks), and a population of mobile users. The mobile operator administrates a new entity, called Rate Recommender (RR), that recommends to mobile users, upon request, the streaming rate at which they should be receiving data from multimedia servers. A number of network monitoring agents are deployed over the entire mobile operator network to assess bandwidth availability and to report it on a regular basis to the Rate Recommender. RR uses this information to assess the current network resource availability and also to form a statistical profile of the network bandwidth availability over time. The main role of RR consists of using these statistical profiles, together with the mobility features of mobile users, in order to recommend the most optimal rate these users should use in order to minimize bandwidth fluctuations during the course of the streaming service.

As depicted in Fig. 2, the design of RR consists of three main modules, namely Rate Recommendation, Network Resource Profiling (NRP) and User Profile Repository (UPR). As the name infers, UPR forms a repository for users' profiles. UPR consists of four units, Context Repository Service (CxRS), Context Gathering Service (CxGS), Context Aggregation Service (CxAS) and Context Distribution Service (CxDS). At regular times, CxGS gathers context information from users

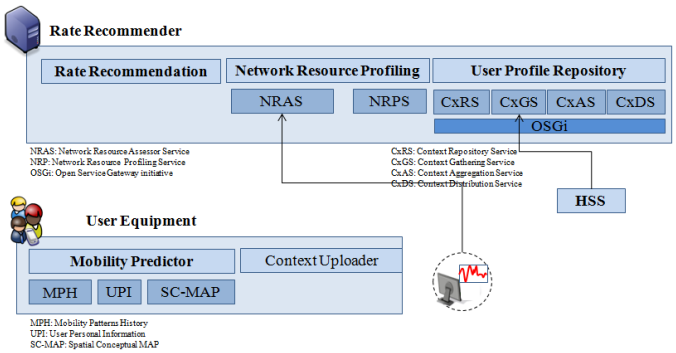

Fig. 2. Key components of RR and mobile terminals.

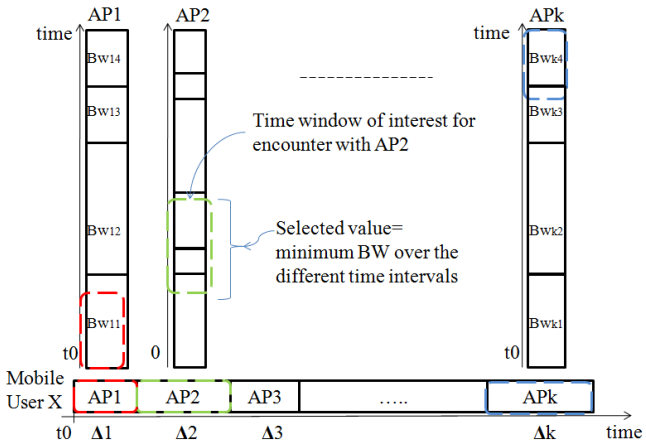

Fig. 3. Envisioned matrix for assessing optimal value of the streaming rate by RR.

(i.e., Mobility Predictor and Context Uploader). Contextual information may also include users' personal information and preferences provided by the user when he/she first subscribes to the service and users' mobility patterns predicted by a Mobility Predictor (MP) entity implemented at terminals. Users can also upload further contextual information, if required, using their Context Uploader (CU). CxGS forwards the obtained values to both CxRS and CxAS. CxRS persistently maintains context information that is required to build-up and update users' profile. On the other hand, CxAS aggregates context information, obtained from different context sources (e.g., CxRS) or on demand (e.g., from Home Subscriber Server (HSS) in case of 3GPP networks), that forms the basis for streaming rate recommendation. Finally, CxDS is in charge of sharing and distributing the aggregated context data to the interested entities (e.g., RR of another mobile operator in case of roaming).

The Network Resource Profiling module of RR consists of two units, namely Network Resource Assessor Service (NRAS) and the actual Network Resource Profiling Service (NRPS). NRAS is regularly updated with network conditions from a number of network monitoring agents intelligently deployed over the network. NRPS uses instantaneous information from NRAS to form a statistical profile over time of network resource availability and that is for each access point or a set of access points (i.e., forming a routing area in case of UTRAN or a tracking area in case of e-UTRAN) based on existing time-series model.

The Rate Recommendation module of RR gets information about the mobility of a user from UPR, accordingly sorts out the access points (or areas) to be visited by the mobile user, and from the Network Resource Profile computes the optimal streaming rate which the mobile user should be receiving the multimedia data at. Fig. 3 shows the envisioned matrix for assessing the optimal value of the streaming rate by RR for a mobile user X. In the figure, upon request for rate recommendation, the user is connecting to AP1. Based on feedback from the mobile terminal, RR predicts that the mobile user will be connecting to AP1, AP2 APk, for durations $\Delta_1, \Delta_2, ..., \Delta_k$, respectively, during a time window of interest (e.g., a predetermined period of time, duration of requested content play-back, etc). For the statistical profile of the bandwidth availability at each AP, RR recommends a rate smaller or equal to the minimum available bandwidth at the different access points that the mobile user is predicted to be connecting to. It should be noted that the above described rate recommendation operation can be performed over only a time interval out of the entire service duration or only a portion of the user's path. For instance, if a user is predicted to connect to 10 access points, the rate recommendation could be done for only 3 access points with instructions that the mobile user should consult RR later at the fourth access point. This shall cope with possible errors in the prediction of the user's mobility.

In the envisioned network topology, a mobile terminal comprises two tools, namely Mobility Predictor and Context Uploader (Fig. 2). The Context Uploader gives mobile users the flexibility to provide RR with further context information according to their emerging needs. The Mobility Predictor makes estimates of the users' mobility features and notifies them to the Context Gathering Service (CxGS) unit of RR. For mobility prediction, the application layer of a mobile terminal can refer to a set of tools forming the terminal's Mobility Predictor (MP) (Fig. 2) to sort out the access points to which the mobile node is most likely going to be connected to during the streaming service. Indeed the application layer may use history on the user's mobility pattern (Mobility Patterns History - MPH in Fig. 2) to predict the access points. Referring to a Spatial Conceptual Map (SC-MAP), along with the User's Personal Information (UPI) (e.g., diary), his/her current position, and his/her moving direction, the application layer can predict the most probable future access points. Prior knowledge on the topology of the mobile network can further increase the accuracy of the prediction. After this operation, UPR at RR is informed of the list of access points that the mobile node is most likely going to be connected to during the streaming service time. The mobility prediction is performed only at the beginning of the service or when a mobile client

enters a new mobile domain: it shall incur no significant overhead at the end-terminal. As discussed in the related work section, different mobility prediction models can be used and implemented on platforms such as Google Android or J2ME (Java 2 Platform, Micro Edition). When a mobile user intends to request the stream of a particular video title, it first informs RR of the video characteristics (e.g., video duration, video encoding rate, etc) in addition to information regarding the mobility of the mobile user. Using the developed statistical profile of the network resource availability, RR recommends the optimal bandwidth at which the video should be streamed and that is while taking into account the access points to which the mobile user will be connecting to during the movement, as described earlier.

IV. PERFORMANCE EVALUATION

Having described details on the main functions of our proposed RR entity, we now direct our focus towards the evaluation of its performance in terms of minimizing packet losses and the fluctuations of the streaming rate during the course of the streaming service. The performance evaluation is based on the network simulator NS2. We envision a wired-cum-wireless network topology, consisting of a number of servers connecting to 100 access points via a number of gateway routers as shown in Fig. 1. Wired network nodes are connected via bidirectional links with a propagation delay of 2ms and a capacity of 5Mbps. The 914 MHz Lucent WaveLAN DSSS radio is considered for access points' MAC. The propagation model is TwoRayGround, providing a coverage radius of 250 meters. Access points are deployed over an area of 5x5 km^2 in a way that the longest distance across the overlapping area between two adjacent cells is equal to 10 meters. A population of 100 mobile nodes is simulated with a mobility model as in [3], receiving UDP packets at a constant bit rate from the servers. The size of packets is set to 1K bytes. The simulations are run for ($\theta = 700s$) and the presented results are an average of 7 simulation runs. To consider different mobility prediction schemes, we consider different scenarios by varying the length of a user's mobility path that the system can predict. For example, we consider a scheme that can predict the whole trajectory of a mobile user, or just a part of the trajectory. This shall cope with the possible inaccuracy in the mobility path prediction.

Fig. 4 plots the aggregate packet losses, averaged over three seconds, experienced by mobile nodes when RR is employed and when it is not. When RR is employed, six scenarios are envisioned; each with a particular portion of the trajectory that could be predicted and that is from different time instances (T) during the simulation run time ($\theta = 700s$). Scenario 1 refers to when the whole trajectory could be predicted. When RR is not employed, all mobile nodes receive their streams at rates adjusted to the available bandwidth at the current cell. It should be noted that during handoff, till the streaming rate is adjusted to the available bandwidth, mobile nodes keep receiving data at previous streaming rates. In case less bandwidth becomes available at the target cell (in comparison to that at the source cell), packet drops become inevitable. This explains why in Fig. 4, high packet drops are experienced when RR is not employed. In contrast, when RR is used, less packet drops are experienced and the longer part of a mobile node trajectory is predicted, the further reduction in the packet drops is. The main reason beneath this performance consists in the fact that RR recommends to users rates that are optimal during the whole duration or an interval of the service time while they are on the move. This assists in avoiding streaming rate fluctuations as indicated in Fig. 5. Indeed, Fig. 5 plots the ratio of the number of changes in the streaming rate to the total number of individual handoffs performed by a subset of 16 mobile nodes, selected based on their different mobility patterns. The figure indicates that in case RR is not employed, the vast majority of the mobile nodes experience a change in their streaming rate that is in 20% to 40% of the performed handoffs. However, with the help of RR, the frequency of fluctuations in the streaming rate is suppressed to 5% of the

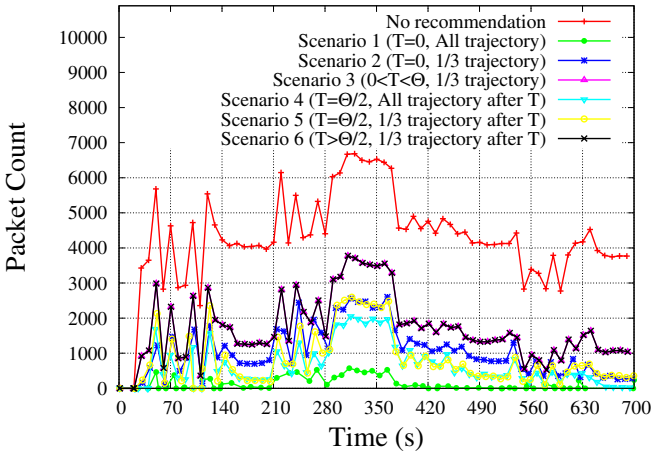

Fig. 4. Average packet drops experienced by the total number of mobile nodes in case of the different simulated scenarios.

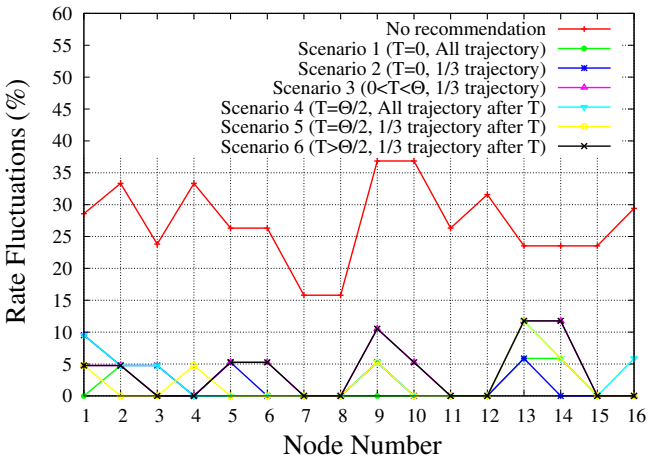

Fig. 5. Frequency of the streaming rate fluctuations for a subset of 16 mobile nodes.

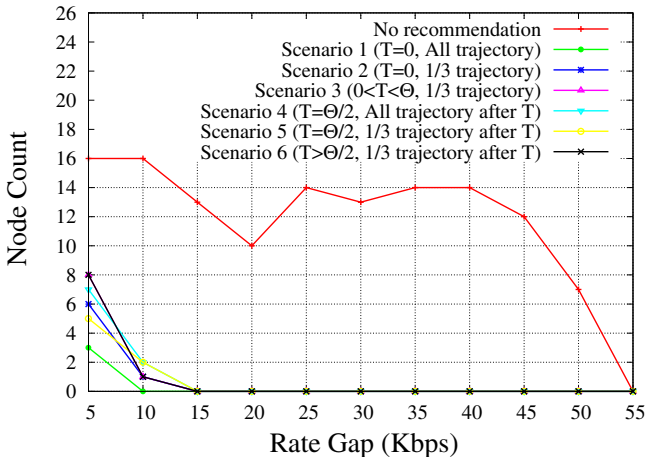

Fig. 6. Distribution of the streaming rate fluctuations.

total performed handoffs, for most mobile nodes.

The superior performance of RR is further evident from Fig. 6 that exhibits how many mobile nodes experienced a particular change in the streaming rate and the relevant magnitude of the streaming rate fluctuation. From the figure, it becomes apparent that few mobile nodes experienced fluctuations in the streaming rate but the relevant magnitude remains relatively small, within the range of 5kbps and 10kbps in case of all simulated scenarios. In contrast, in the absence of streaming rate recommendation, quite a high number of mobile nodes experience important fluctuations in the streaming rate, sometimes in the order of 40kbps to 50kbps.

V. CONCLUDING REMARKS

In this paper, we exploit existing tools and methods for the prediction of user's mobility and bandwidth availability in order to build a framework that recommends to users adequate streaming rates to avoid fluctuations of the multimedia streaming rate and to ultimately ensure acceptable perceived QoS. The proposed framework involves a new entity, called Rate Recommender (RR), administrated by the mobile network operator. RR recommends to mobile users, on demand, the streaming rate at which they should be receiving data from multimedia servers. Based on feedback from network monitoring agents, deployed over the network, RR assesses the current network resource availability and forms a statistical profile of the network bandwidth availability over time. With the help of inputs from mobile users regarding their mobility features, RR is then capable to recommend them at what rate they should be receiving a multimedia service to have a steady QoE. The performance of the proposed framework is evaluated through computer simulations based on NS2. The simulation results confirm the good performance of the framework in reducing packet losses and minimizing both the frequency and the magnitude of fluctuations in the multimedia streaming rate.

REFERENCES

[1] T. Taleb, K. Kashibuchi, A. Leonardi, S. Palazzo, K. Hashimoto, N. Kato, and Y. Nemoto, "A Cross-Layer Approach for an Efficient Delivery of TCP/RTP-based Multimedia Applications in Heterogeneous Wireless Networks", in IEEE TVT, Vol. 57, No. 6, Nov. 2008. pp: 3801-3814.

[2] P. S. Prasad and P. Agrawal, "Movement prediction in wireless networks using mobility traces," in Proc. 7th IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, Jan. 2010.

[3] J. Markoulidakis, G. Lyberopoulos, D. Tsirkas, and E. Sykas, "Mobility modeling in third-generation mobile telecommunications systems," in IEEE Personal Communications, Vol. 4, No. 4, Aug. 1997. pp. 41-56.

[4] F. Bai and A. Helmy, "A survey of mobility models," Book chapter in Wireless Ad Hoc and Sensor Networks, Kluwer academic Publishers, Jun. 2004.

[5] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," in Wireless Communications and Mobile Computing, Vol. 2, No. 5, Aug. 2002. pp. 483-502.

[6] S. Hongbo, W. Yue, Y. Jian, and S. Xiuming, "Mobility prediction in cellular network using hidden Markov model," in Proc. 7th IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, Jan. 2010.

[7] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in Proc. 26th Annual IEEE Conference on Computer Communications (INFOCOM'06), Barcelona, Spain, Apr. 2006.

[8] S. Akoush and A. Sameh, "Mobile user movement prediction using bayesian learning for neural networks," in Proc. 2007 IWCMC, Honolulu, Hawaii, USA, Aug. 2007.

[9] N. Samaan and A. Karmouch, "A mobility prediction architecture based on contextual knowledge and spatial conceptual maps," in IEEE Transactions on Mobile Computing, Vol. 4, No 6, Nov.-Dec. 2005. pp. 537-551.

[10] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research," in Proc. 2nd ACM/SIGMOBILE International Workshop on Multi-hop Ad Hoc Networks: from theory to reality (REALMAN'06), Florence, Italy, May. 2006.

[11] M. Musolesi and C. Mascolo, "Designing mobility models based on social network theory," in ACM/SIGMOBILE Mobile Computing and Communications Review, Vol. 11, No. 3, Jul. 2007. pp. 59-70.

[12] F. Ekman, A. Kernen, J. Karvo, and J. Ott, "Working day movement model," in Proc. 1st ACM/SIGMOBILE workshop on Mobility models for Networking Research (MobilityModels'08), Hong Kong, China, May. 2008.

[13] A. Aymen, K. Houda, A. Mohamed, Z. Ez, and T. Sami, "Probabilistic model for mobility in cellular network subscriber," in Proc. 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE 2009), Aalborg, Denmark, May. 2009.

[14] T. Zhou, H. Sharif, M. Hempel, P. Mahasukhon, W. Wang, and S. Ci, "A Quantitative study of mobility impact for real-time services on a Wi-Fi multi-hop network," in Proc. IEEE VTC 2008 Spring, Singapour, Republic of Singapore, May. 2008.

[15] S. Intarasothonchun, S. Thipchaksurat, and R. Varakulsiripunth, "Effect of mobility on predictive mobility support dynamic resource reservation in cellular networks," in Proc. 8th International Conference on ITS Telecommunications (ITST 2008), Phuket, Thailand, Oct. 2008.

[16] M. Li, H. Zhu, I. Chlamtac, and B. Prabhakaran, "End-to-end QoS framework for heterogeneous wired-cum-wireless networks," in ACM/Baltzer WINET, Vol. 12, No. 4, Aug. 2006. pp. 439-450