Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

MODELING CELLULAR NETWORK TRAFFIC WITH MOBILE CALL GRAPH CONSTRAINTS

Junwhan Kim
V. S. Anil Kumar
Achla Marathe
Guanhong Pei
Sudip Saha
Balaaji S.P. Subbiah

Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute
Virginia Tech, Blacksburg, VA 24061, USA

ABSTRACT

The design, analysis and evaluation of protocols in cellular and hybrid networks requires realistic traffic modeling, since the underlying mobility and traffic model has a significant impact on the performance. We present a unified framework involving constrained temporal graphs that incorporate a variety of spatial, homophily and call-graph constraints into the network traffic model. The specific classes of constraints include bounds on the number of calls in given spatial regions, specific homophily relations between callers and callees, and the indegree and outdegree distributions of the call graph, for the whole time duration and intervals. Our framework allows us to capture a variety of complex behavioral adaptations and study their impacts on the network traffic. We illustrate this by a case study showing the impact of different homophily relations on the spatial and temporal characteristics of network traffic as well as the structure of the call graphs.

1 INTRODUCTION

Modeling traffic in cellular and hybrid networks is a first step in a number of applications. For instance, the design, analysis and evaluation of protocols requires realistic traffic modeling, since the underlying mobility and traffic model has a significant impact on the performance. Dynamic spectrum access and spectrum trading require realistic modeling primary/licensed user behavior (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010b),(Willkomm, Machiraju, Bolot, and Wolisz 2008), in order to enable efficient opportunistic usage by secondary users. It is usually very difficult to obtain data on a large spatio-temporal scale, because of the proprietary nature of network infrastructure. The results of Willkomm et al. (Willkomm, Machiraju, Bolot, and Wolisz 2008), present one of the few studies for cellular network traffic characteristics at a large scale. Most of the other studies with real data are for very small networks, e.g., local area networks within small regions, such as a campus or a conference site, e.g., (Balachandran, Voelker, Bahl, and Rangan 2002, Kotz and Essien 2005, Tang and Baker 2000). All these results show significant spatio-temporal variation in the network traffic – these cannot be adequately modeled by simple stochastic processes that can match aggregate characteristics (Balachandran, Voelker, Bahl, and Rangan 2002, Kotz and Essien 2005, Tang and Baker 2000). Similar issues come up in the analysis of Internet structure and traffic, and researchers have suggested the use of "first principles" approaches for realistic network and traffic modeling. Beckman et al. (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a) and Kroc et al. (Kroc, Eidenbenz, and Smith 2009) develop agent based

first-principles approaches for synthetic network traffic modeling, which integrate a number of different data sets, and match aggregate properties, such as the known call arrival rate and call duration distributions.

In most applications, modeling network traffic to match known aggregate distributions is not adequate. User behavior changes in response to new wireless technologies, changes in pricing and new incentives by providers and emergency and disaster situations. For instance, the increased data traffic by iPhone users caused severe strain on AT&T's infrastructure. Barrett et al. (Barrett, Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010) model the changes in calling behavior during an evacuation, and show that this leads to a significant strain on the communication network. An important aspect of user behavior in such settings is "homophily", which implies specific correlations between callers and callees. For instance, (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a),(Kroc, Eidenbenz, and Smith 2009) use location based homophily to specify call patterns, in which people who are co-located during their activities in a day have a higher likelihood of calling each other. There can be other kinds of homophilies, e.g., age-based, in which people are more likely to call others in the same age group (Jackson 2008).

Data for such homophily relations is likely to be even more challenging to obtain, because of the obvious privacy issues. Aggregate statistical properties related to this have been studied in the form of the "call graph", which has one or more edges between people who participate in a call. Seshadri et al. (Seshadri, Machiraju, Sridharan, Bolot, Faloutsos, and Leskove 2008) use data on a long time scale from the Sprint network and compute properties such as the degree distribution of the call graph. Nanavati et al. (Nanavati, Singh, Chakraborty, Dasgupta, Mukherjee, Gurumurthy, and Joshi 2008) study other properties besides the degree, such as the diameter, cores and cliques. The degree distribution of call graphs and indegree-outdegree correlations show that they are not Poisson, and have several properties similar to the web-graph. Realistic modeling of network traffic requires preserving both aggregate traffic characteristics (e.g., arrival rates), as well as the call graph properties. We are not aware of any simulation framework that captures all these aspects, and is the motivation for our work.

In this paper, we develop a unified framework involving constrained temporal graphs that incorporates a variety of spatial, homophily and call-graph constraints into the network traffic model. Our main contributions are summarized below. (1). *Unified dynamic graph framework for spatial and homophily relations*: We develop a unified framework that satisfies traffic properties as well as spatial, homophily based and call-graph based constraints. The specific classes of constraints we incorporate include: (i) bounds on the number of calls in given spatial regions – this captures, e.g., base station capacity constraints, which limit how many calls can be made in a base station cell, (ii) specific homophily relations between callers and callees, e.g., between age groups or other demographic classes, (iii) constraints on the indegree and outdegree distributions of the call graph, for the whole time duration, as well as for specific time intervals. We show that these constraints can be viewed in terms of degree constraints for different subsets in the dynamic call graph, which gives a general approach to formulate them. We extend the session generation module in (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a) to incorporate these features – incorporating these constraints requires extending this framework to use and keep track of location and demographic information for the population. (2). *Illustrative case study*: our extended framework allows us to capture a variety of complex behavioral adaptations and study their impacts on the network traffic. We illustrate this by a case study showing the impact of different homophily relations on the spatial and temporal characteristics of network traffic as well as the structure of the call graphs.

The remainder of the paper is organized in the following manner. In Section 2, we give a brief overview of traffic characteristics and SSRSM. In Section 3, we discuss traffic and call constraints. We describes our algorithm and implementation in Section 4. Our results are discussed in Section 5.

2 PRELIMINARIES

2.1 Notation for Mobility and Traffic Characteristics

We discuss some of the notation needed to describe the mobility and traffic model and the temporal graphs used here. We build on the notation from (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a). Details about the other components of SSRSM, such as the mobility and device ownership models are described later in Section 2.3, and build on models from (Barrett, Beckman, Khan, Kumar, Marathe, Stretz, Dutta, and Lewis 2009). We assume the geographic region is split into cells, which correspond to base station coverage areas. For our framework, the exact shapes and sizes of the cells do not matter, and they can be uniform. For concreteness, we use the results of (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a), which assumes the cells to be rectangular. Let $\mathscr{C}$ denote the set of cells. If an individual u moves from cell C to C', we assume that this movement happens along the shortest path through the city's road network connecting the end points. The sequence of cells would correspond to the cells intersected by this shortest path. For a person u, let $f(u,t)$ denote his spatial position at time t; for most of the paper, we will use grid cells to identify the node's position. We denote a call from a person u to a person v starting at time t_1 and ending at time t_2 by a tuple $(u,v,[t_1,t_2])$. We consider the following measures. For a cell C and a time interval $[t',t'']$, we define $A(C,[t',t''])$ as the set of all calls $(u,v,[t_1,t_2])$ such that the caller u is in cell C and $[t',t''] \cap [t_1,t_2] \neq \phi$. We have a similar definition on the callee's side.

The inter-arrival time is defined as the time between two consecutive call arrivals, and we consider different variants of this measure. For all calls originating at cell C, we determine the inter-arrival times and construct its PDF. Further, we are interested in observing the temporal changes in this parameter and consider the inter-arrival times in cell C within $[t_1,t_2]$. For example, we compute the inter-arrival times for all calls in the set $A(C,[t_1,t_2])$ and observe its variation during the day. The call duration (τ) or call holding time at a cell C is the length of the call. We consider both the average value and the distribution of τ within each cell C. For calls that straddle the cell boundaries (originate in one cell C_1 and terminate in another C_2), we compute τ for cell C_1 with a duration the caller is in C_1 and use the remaining for the cell C_2. We also determine τ for different intervals $[t_1,t_2]$ during the day. The spatial footprint of a session $(u,v,[t_1,t_2])$, is defined as the distance traveled by i during the time frame $[t_1,t_2]$. Average Load, $X(C,[t_1,t_2])$, which is defined as $\sum_{t' \in [t_1,t_2]} h(C,t')/|A(C,[t_1,t_2])|$, where $h(C,t')$ is the number of calls in $A(C,[t_1,t_2])$ that are simultaneously active at time t'—this captures the average number of simultaneous calls in cell C during this time interval; similarly, we also consider the peak load in the cell.

2.2 Notation for Call Graphs

The call graph is denoted by a directed graph $G = (V,E)$ where V denotes the population and an edge $e = (u,v)$ denotes a call from u to v. The edges are labeled, and let $\phi(e)$ denote the label corresponding to edge e; $\phi(e)$ could consist of a sequence of the time intervals during which a $u \leftarrow v$ call was made. It could also capture other attributes, e.g., the activities and locations of the end points, when the call was being made. We use $V(G)$ and $E(G)$ as the set of nodes and edges in G, respectively.

We consider various induced subgraphs and restrictions of G: (i) let G_T denote the subgraph of G restricted to time interval T, and (ii) let $G_T[V']$ denote the subgraph of G induced by the subset $V' \subseteq V$ of nodes. The outdegree of a node v in graph $G_T[V']$, denoted by $deg_{G_T[V']}(v)$, is the number of out-neighbors of v; we drop the subscript, when the graph is clear from the context. We will consider induced subgraphs $G_T[V']$, where V' could denote subpopulations with specific demographic characteristics, e.g., in an age group, or the set of people within a cell C. In section 3, we will specify constraints for our framework in terms of constraints on such induced subgraphs.

2.3 Synthetic Population and Traffic Modeling

As discussed earlier, we use a synthetic demand generation model, SSRSM (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a) to generate the demand for wireless spectrum at detailed spatio-temporal scale for the city of Portland, Oregon. It combines a number of different data sets and models, including: an urban mobility model, synthetic social network of Portland (Barrett, Beckman, Khan, Kumar, Marathe, Stretz, Dutta, and Lewis 2009), road network data, data sets for device ownership from NHIS (National Health Interview Survey), and aggregate characteristics for cellular communication traffic, such as call arrival rate and call duration (see (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010a, Barrett, Beckman, Khan, Kumar, Marathe, Stretz, Dutta, and Lewis 2009) for more details). The various components of SSRSM are shown in Figure 1. The demand model consists of the following steps:

1. Creation of synthetic urban populations: this step integrates a variety of commercial and public data sources to produce a synthetic population that is statistically indistinguishable from census data (Barrett, Beckman, Khan, Kumar, Marathe, Stretz, Dutta, and Lewis 2009). The process preserves the confidentiality of the original individuals and produces synthetic individuals with realistic attributes and demographics. The synthetic population is a set of synthetic people and households, located geographically, each associated with a set of demographic variables drawn from the census. It is a collection of synthetic objects, each associated with a set of attributes. Each synthetic individual is placed in a household with other synthetic people and each household is placed geographically in such a way that a census of the synthetic population is statistically indistinguishable from the original census, if aggregated to the block group level. Synthetic populations are thus statistically indistinguishable from the census data. See (Barrett, Beckman, Khan, Kumar, Marathe, Stretz, Dutta, and Lewis 2009) for additional details.
2. Activity and location choice: in this step, a set of activity templates for individuals in the households are determined, based on US census and survey data on activity and time-use surveys. These activity templates describe the sort of activities each household member performs and the time of day they are performed. Each synthetic household is then matched with one of the survey households using a decision tree based on demographics such as the number of workers in the household, number of children, their ages, etc. The synthetic household is assigned the activity template of its matching survey household. For each household and each activity performed by this household, a preliminary assignment of a location for the activity is made based on observed land-use patterns, tax data, etc.
3. Route selection: in this step, route plans are assigned to individuals who need to move between locations in order to perform their daily activities. The movements respect road network constraints such as speed limits, number of lanes etc. The final mobility is obtained by interpolating the movement of individuals on the road network.
4. Device assignment: this step chooses the number and type of wireless devices to be used by each individual. This assignment is based on the age and gender of the person, if he/she is a worker, and the income of the household as suggested by the survey data set from NHIS (Center for Disease Control). This model assumes that every individual has a wireline connection at their locations. However, during transit between locations, only individuals who have been assigned wireless devices can use them.
5. Session Generation: Finally, the demand model generates wireless calls (or call sessions) for each individual from the synthetic population for the duration of a day using statistics in (Willkomm, Machiraju, Bolot, and Wolisz 2008). This module uses a discrete event simulation approach to model calling patterns, using distributions for call duration and arrival rates from those reported in (Willkomm, Machiraju, Bolot, and Wolisz 2008). It uses the number of arrivals and session duration distribution as input, however these can be changed as needed, depending on any other available information.

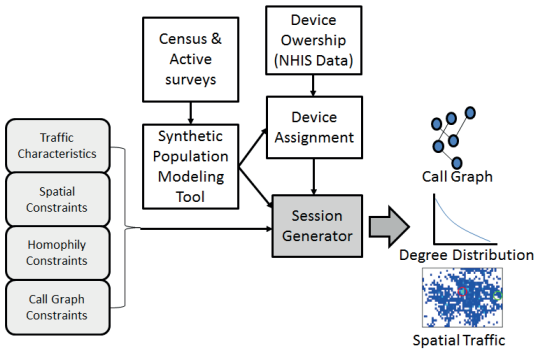
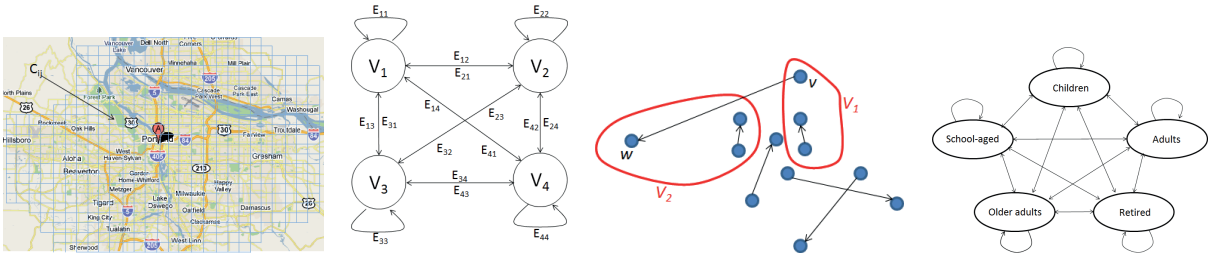


Figure 1: The Components of the synthetic mobility and demand generation framework. The rounded boxes represent modules of SSRSM and the rectangles represent input datasets.

3 TRAFFIC AND CALL GRAPH CONSTRAINTS

We now describe the different classes of constraints for the network traffic and call graph that our framework is able to handle. As discussed earlier, a unifying feature is the formulation of these as temporal graph constraints.



(a) Cell-based Portland (b) Homophily (c) Call Graph (d) Age-based Grouping

Figure 2: Examples of Call Graph Constraints: (a) spatial constraints on the number of calls, (b) homophily constraints that specify the number of calls between different subsets, (c) example of a call graph, and (d) example of homophily constraints between different age groups.

1. *Aggregate traffic characteristics*: we consider constraints on traffic characteristics, such as call arrival rate and call duration distributions. They can be specified as a single distribution for the whole region and time interval, or different distributions for different parts and times. We will discuss it in Section 4.2.2.
2. *Spatial constraints*: these are of the following form – "the number of calls within an interval $[t_1, t_2]$ in a spatial region is bounded". Figure 2(a) shows that the specific spatial region would typically be a cell C_{ij}. This is formulated as a constraint on the graph $G_{[t_1,t_2]}(V)$ – a simplest form of such a constraint is that the sum of all node degrees in $G_{[t_1,t_2]}(V)$ is bounded by a given parameter.
3. *Call graph degree constraints*: these constraints can be of the following kinds: (i) for a given subset $V' \subseteq V$ and parameter B, this is formulated as $\sum_{v \in V'} outdeg_G(v) \leq B$, (ii) for a given partition $W_1, \ldots, W_k$ of the set V, and a sequence $\beta = (\beta_1, \ldots, \beta_k)$, with $\sum_i \beta_i = 1$, we have $\sum_{v \in W_i} outdeg_G(v) = \beta_i |E(G)|$, (iii) The outdegree and indegree sequences $(outdeg_G(v) : v \in V)$ and $(indeg_G(v) : v \in V)$ satisfy specific power-law constraints. In general, there could be other constraints, e.g., on the clustering coefficient.

4. *Homophily constraints*: these are of the form – "the number of edges between subsets V_i and V_j is bounded". Given a partition of V into sets $V_1, \ldots, V_k$, let $\mathbf{H} = (H_{ij})$ denote a homophily distribution, with $\sum_j H_{ij} = 1$ for each i. Let $E_{ij} = \{e = (u,v) : u \in V_i, v \in V_j\}$. This constraint implies $|E_{ij}| = H_{ij} \sum_{j'} |E_{ij'}|$.

4 ALGORITHM DESCRIPTION

4.1 Pseudo-Code Level Description

We describe our algorithm at a high level here. It is implemented as a discrete event simulation, which uses arrival rates and call durations as the triggering events to determine connections. Each such event involves starting a call (which requires choosing the caller and callee), or ending a call. In order to ensure degree constraints are satisfied, we use biased sampling for choosing the caller and callee for each call. This is described in Algorithm 1. Our implementation of SG is fairly general and flexible, and constraints can be specified through simple configuration files, as described later. This make large studies, involving different parameter choices, easy to implement.

Algorithm 1: Pseudo-code description of SG

```
Input: Population, mobility model (i.e., function f(u,t) for each person u and time t, device assignment for
       each person (wireless and wireline), demographics
Output: Model of traffic (who calls whom, and when)
Initialize event queue Q
Initialize list of free callers
while event queue Q is not empty do
    if event is start a session then
        Choose type of call (wireless or wireline)
        Choose subset V1 from which to pick a caller – this depends on the current spatial or homophily
        constraint being enforced.
        Choose a free caller v ∈ V1.
        Choose subset V2 from which to pick a callee, using spatial and demographic properties of V1
        Choose a callee w ∈ V2
        if spatial constraint at V1 or V2 not satisfied then
            Report as call drop
        if callee is not free then
            Report as call drop
        if current calls ≥ capacity of the cell of caller v, or current calls ≥ capacity of the cell of callee w
        then
            Report as call drop
        if call is successful then
            Choose end-time by sampling from call duration distribution, and add trigger to event queue
        else
            Retries to make a call after a random time
    else
        End an ongoing session – free up caller and callee
```

We now discuss details of some of the steps in Algorithm 1, and how they help in enforcing the various traffic and call graph constraints.

1. Choosing the type of call: we assume we are given the fraction of wireless calls as a parameter, which is used for deciding whether the call should be a wireless-wireless, wireless-wireline or wireline-wireline call. This affects the choices of the sets V_1, V_2 in the subsequent steps.
2. Choosing set V_1 for possible callers: this is based on the following criteria: (i) Is the call a wireless-wireless or wireless-wireline call? (ii) Spatial or activity constraints: the caller is in a specific location or region of the city, or is doing a specific activity – in this case, V_1 is the subset of people who satisfy such constraints, and are currently free. (iii) Demographic constraints: this is specified by the distribution M, where M_i specifies the fraction of calls originating from the subset W_i of V.
3. Choosing the caller $v \in V_1$: sample from the set V_1 using distribution D_{out}, conditioned on the set V_1.
4. Choosing set V_2 for possible callers: this involves the following steps
   (a) Suppose the chosen caller v is in group W_i. We use the Homophily matrix H to determine the class V_2 – it is chosen to be the group W_j with probability H_{ij}.
   (b) If there are additional spatial constraints, e.g., the callee is required to be within a cell C, we choose V_2 to be $W_j \cap \{w : f(w,t) \in C\}$.
5. Choosing the callee $w \in V_2$: we choose the callee based on the indegree distribution D_{in}, conditioned on the set V_2.

4.2 Implementation

Our implementation involves configuration files that specify different kinds of parameters, making the framework flexible and easy to use. We discuss the main components below.

4.2.1 Main Configuration

The system has different kinds of configuration files that specify attributes of the people, devices, activities and cells, as well as a main configuration file that specifies the global parameters (i.e., path and names of the other input files, simulation duration, output file, and seed number).

The *person file* represents the list of individuals in the population. The file has some redundant information as the same format as the Device_Assignment module in the SG can be used. The file consists of *hid*, *pid*, *age*, *gender*, and *critical worker*. The hid is a household identifier, the pid is an identification for the person, and the critical worker identifies if the person forms a critical individual for the work place.

The *device assignment file* is generated by the *Device_Assignment* module (see Figure 1). The file is composed of *pid*, *device ID*, *device type*, *service provider*, *interface*. The device ID is an identification number for a device. The device type indicates whether it can support voice, voice and data, smart phone, etc. The SG currently supports only devices capable of voice alone and devices capable of making voice with data calls. The service provider can be imposed that some specific market share of the device is based on the latest information regarding market share of various service providers. We consider 6 service providers (e.g., Verizon, AT&T, T-Mobile, US Cellular, Sprint, and Alltel). The interface is used to differentiate between landline (wired) and cell phones (wireless) devices. We can add other name into this for differentiating the technology of the phone into 2G, 3G, 4G and WCDMA.

The *activity file* contains *pid*, *hid*, *purpose*, *start time*, *duration*, and *location* for activities in the populations. The activity performed by a person has a significant impact on it behavior. While generating sessions, the activity file should be taken to create sessions for realistic loads on the network. For example, if a person is at home, the duration of the calls tend to be longer and usually are with certain contacts. In case of work place, the call durations tend to be shorter, maybe the frequency of the calls may be higher and contacts are official contacts. All these aspects are important to consider while generating sessions. Some people may not have a cell phone and can only originate or receive calls from the landline accessible at locations. When a caller calls such people, it is important to also consider the availability of a phone to this person before assuming that the call can actually go through. Sometimes when a certain person

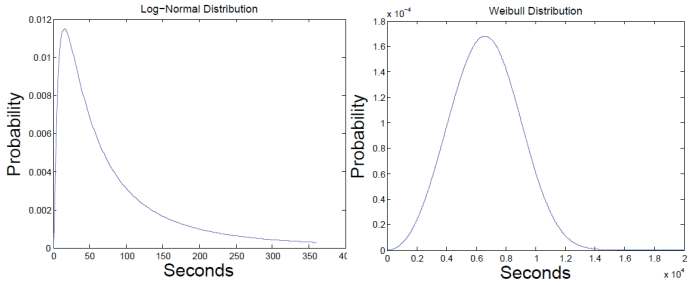
cannot be reached, people tend to retry the person at a later time or contact the same person on another device almost immediately.

The *capacity file* defines a capacity for each base station. We assume that each cell has a base station having capacities to simultaneously serve 50 calls. Whenever a caller tries to create a session, the session is dropped if the current number of successful sessions is more than the predefined capacity in the caller's cell. Even if the caller creates a session, the session is dropped unless the capacity in the callee's cell is sufficient. We can adjust some capacities depending on the amount of load. For example, people may try to make a lot of calls in a shopping mall or downtown, so more capacity may be assigned.

The *social network file* is for the region considered and this file is generated by EpiSimdemics (Barrett, Bisset, Eubank, Feng, and Marathe 2008) and represents the contact between individuals and stored as a graph file. The format for the graph file is according to *Graph Library package* (GaLib) of NDSSL.

4.2.2 Mobile and Landline Configurations

The mobile file configures the parameters for the mobile call generation program. The landline file configures the parameters for the wireline call and their patterns. The configuration for landline and mobile calls defines the arrival rate and duration of sessions. Arrival and duration depend on the activity that the person and the time of the day. For example, a person is less likely to make frequent personal calls during the office hours and more likely to make them after office hours. Also, the device used for making work related calls is different from the device used for personal calls. The SG supports both log-normal and Weibull distributions for arrival and duration for each session. The mobile and landline files define these two distributions for arrival and duration.



(a) Log-Normal Distribution (b) Weibull Distribution

Figure 3: Traffic constraints: aggregate call arrival rate and duration distributions.

Figure 3(a) shows the probability density function (PDF) for a log normal distribution. For a mobile call configuration, the mean and deviation of a session duration are set to 4.07 and 1.15, respectively. Figure 3(b) indicates the PDF of a Weibull distribution with shape and scale parameters–(7400.0, 3.2) to define an arrival rate. As long as the probability of the arrival rate increases, the number of sessions increases.

4.2.3 Outputs

For conducting the simulation of sessions in a realistic scenario, each person begins the day at a different time and the sessions should ideally start after the start of the day. So, each person's first session should independently start off other people. Some people may make their first call after a few hours, some might call in an hour, some will receive calls from others, while some others may not receive any calls until a certain time. The social network used for determining the callee also will be different.

The output of the SG contains the call information relating to the wireline and wireless calls. The format of the output consists of *pid caller*, *device id*, *device type*, *service provider*, *pid callee*, *device id*,

device type, *service provider*, *start time*, *end time*, *session type*. A caller indicates initiator for the session. A callee is another person belonging to the social contact list of the initiator (,or caller). The device id shows both a caller and a callee of the session. The SG currently generates only voice calls, but has the provision of generating 'data call' if required. Each session has to be assigned a certain type. The types of sessions currently are voice and data. The voice sessions can originate from phones capable of only voice communications.

5 RESULTS AND DISCUSSION

We illustrate our framework by a case study to construct a synthetic mobile network traffic and call graph model for Portland, OR. We consider age based partitioning of the population and arrival rates for each group based on callers' age. We compare two kinds of homophilies— age based and co-location based, which are captured in terms of the social network in the synthetic population generated by (Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei 2010b). In this setting, the callers are picked at random but the callees are picked either from the same-age groups or from the callers' social network. The arrival rates stay the same under both age based and social network based scenarios.

Table 5 indicates the age range for each of the groups. We let x denote a parameter, and the arrival rates for different groups are multiples of x, as shown in Table 5. For instance, SG generates x as a number of sessions for group 1 and 10 times x for group 2 etc. We assume that the school-aged children and adults make significantly more calls than children and retired. The parameters used in this table can easily be modified to reflect more realistic numbers as they become available.

Table 1: Age based grouping of the population. These groupings are used to determine the arrival rates and for age-based callee selection.

	Age Range	Category	Arrival Rate
Group 1	0-5	Children	x
Group 2	6-18	School-aged	10*x
Group 3	19-35	Adults	20*x
Group 4	36-65	Older adults	30*x
Group 5	66+	Retired	x

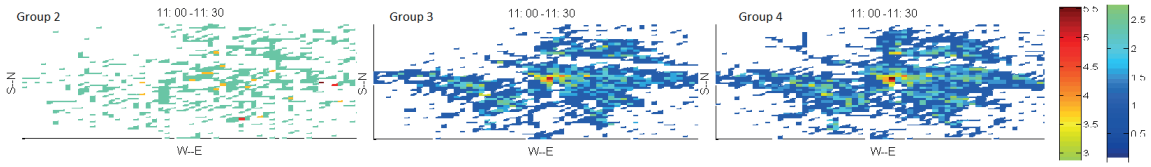


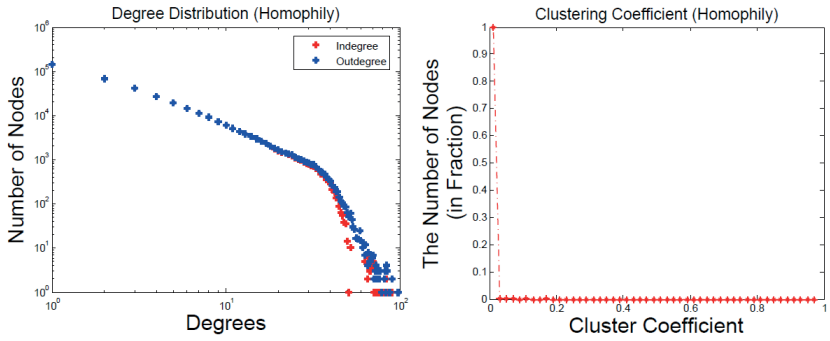
Figure 4: Spatio-temporal variation in traffic load for groups 2, 3 and 4.

Figure 4 shows the the spatio-temporal variation in the traffic load for Groups 2, 3 and 4 from 12:00 to 12:30 for the region of Portland, OR. The differences in the panels reflect the different call volumes for each of the three groups shown here. Observe that Group 4 has substantially more calls than group 2.

5.1 Degree Distributions and Clustering Coefficient

Figure 5(a) shows the indegree and outdegree distributions of the call graph for callees selected based on age-based homophily. The gap between indegree and outdegree curves indicates the difference between the number of callers and callees with the same edges.

The clustering coefficients (CCs) is the likelihood of a random pair of neighbors of a node being connected, and is defined as: $CC_v = \frac{NE(v)}{\binom{d(v)}{2}}$, where $NE(v)$ is the number of edges among neighbors of



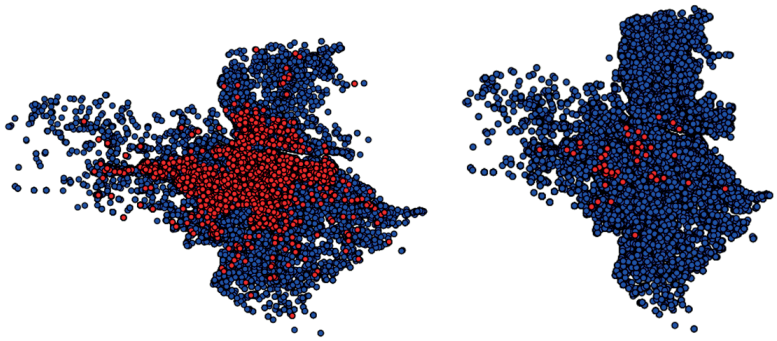
(a) Degree Distribution (b) Clustering Coefficient

Figure 5: Degree distribution and clustering coefficient distributions of the call graph.

nodes v and $d(v)$ is the degree of v. Figure 5(b) shows cluster coefficient distribution in the call graph with age-based homophily. We observe that the clustering coefficient is small for majority of the nodes.

5.2 Spatial Diversity in Call Graphs under Homophily Constraints

We compare two models to select callees. A callee is selected either from the social network of the caller or the same age group as the caller. Figures 6 depicts the spatial separation in call graphs for the entire population under social network and age based homophily constraints for a half an hour interval of 12:00-12:30pm. Each blue circle represents the fact that the caller and callee are located in different cells. Each red circle indicates that both caller and callee are located in the same cell. Note that the social network is based on co-location and that is why a large number of social contacts are confined to the same cell. As a result, when callee selection is based on the social network, we observe many more red circles.

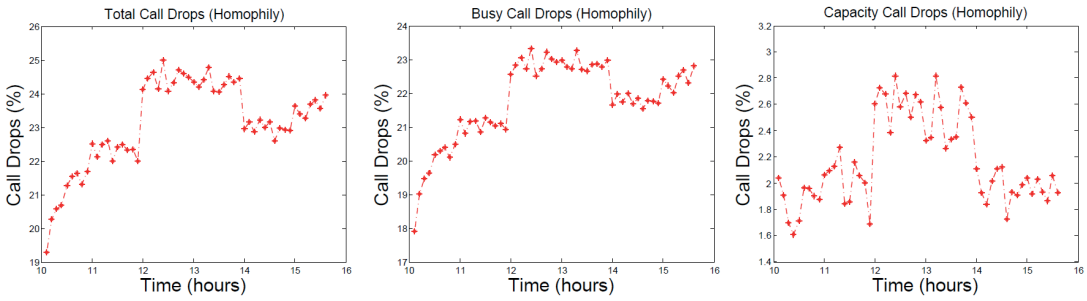


(a) Call Graph Based on Social Networks (b) Call Graph Based on Age homophily

Figure 6: Call graphs based on social network and age based homophily for the city of Portland between 12 and 12:30pm.

5.3 Call Drops under Age Based Homophily Constraints

Next we consider the number of call drops resulting from constraints on cell capacities. Figure 7(a) shows the total number of call drops. Figures 7(b) and 7(c) indicate the number of call drops due to callees' busy signal and due to the cell capacity constraint, respectively. The results show that majority of calls drop due to callee being busy and not from the cell tower reaching its capacity, for the specific parameters we chose.



(a) Total Call Drops (b) Call Drops due to a busy signal (c) Call Drops due to over capacity

Figure 7: Call Drops Under Age Homophily Constraint.

6 CONCLUSIONS

This paper examines modeling cellular network traffic with mobile call graph constraints. We develop a unified framework involving constrained temporal graphs that incorporate a variety of spatial, social, and homophily constraints into the network traffic model. We also evaluate the framework through a case study showing the impact of different homophily relations on the spatial and temporal characteristics of network traffic as well as the structure of the call graphs.

ACKNOWLEDGMENTS

We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. This work has been partially supported by NSF Nets Grant CNS-0626964, NSF HSD Grant SES-0729441, NSF PetaApps Grant OCI-0904844, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C-0113, NSF NETS CNS-0831633, DOE DE-SC0003957, NSF CNS-0845700, NSF Netse CNS-1011769 and NSF SDCI OCI-1032677.

REFERENCES

Balachandran, A., G. M. Voelker, P. Bahl, and P. V. Rangan. 2002. "Characterizing user behavior and network performance in a public wireless LAN". In *Proceedings of the 2002 ACM SIGMETRICS*, 195–205. New York: ACM.

Barrett, C., D. Beckman, M. Khan, V. S. A. Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. 2009, December. "Generation and analysis of large synthetic social contact networks". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1003 –1014. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barrett, C., R. Beckman, K. Channakeshava, F. Huang, V. S. A. Kumar, A. Marathe, M. V. Marathe, and G. Pei. 2010. "Cascading failures in multiple infrastructures: From transportation to communication network". In *CRIS'10: 2010 International Conference on Critical Infrastructure*, 1–8.

Barrett, C. L., K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe. 2008. "EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks". In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, SC '08, 37:1–37:12. Piscataway, NJ, USA: IEEE Press.

Beckman, R., K. Channakeshava, F. Huang, V. S. A. Kumar, A. Marathe, M. V. Marathe, and G. Pei. 2010a. "Implications of Dynamic Spectrum Access on the Efficiency of Primary Wireless Market". In *IEEE DySPAN*, 1 –12.

Beckman, R., K. Channakeshava, F. Huang, V. S. A. Kumar, A. Marathe, M. V. Marathe, and G. Pei. 2010b. "Synthesis and Analysis of Spatio-Temporal Spectrum Demand Patterns: A First Principles Approach". In *IEEE DySPAN*, 1 –12.

Center for Disease Control. "National Health Interview Survey (NHIS)". http://www.cdc.gov/nchs/about/major/nhis/nhis_2007_data_release.htm.

Jackson, M. O. 2008. "Average distance, diameter, and clustering in social networks with homophily". In *Proceedings of the 4th International Workshop on Internet and Network Economics*, 4–11: Springer-Verlag.

Kotz, D., and K. Essien. 2005, January. "Analysis of a campus-wide wireless network". *Wirel. Netw.* 11:115–133.

Kroc, L., S. Eidenbenz, and J. Smith. 2009, December. "SessionSim: Activity-Based Session Generation for Network Simulation". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 3169 –3180. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Nanavati, A. A., R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjee, G. Gurumurthy, and A. Joshi. 2008. "Analyzing the Structure and Evolution of Massive Telecom Graphs". *EEE Transactions on Knowledge and Data Engineering* 20 (5): 703 –718.

Seshadri, M., S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. 2008. "Mobile Call Graphs: Beyond Power-law and Lognormal Distributions". In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 596–604: ACM.

Tang, D., and M. Baker. 2000. "Analysis of a Local-Area Wireless Network". In *ACM MobiCom*, 1–10.

Willkomm, D., S. Machiraju, J. Bolot, and A. Wolisz. 2008, Oct.. "Primary Users in Cellular Networks: A Large-Scale Measurement Study". In *IEEE DySPAN*, 1–11.

AUTHOR BIOGRAPHIES

JUNWHAN KIM is a Ph.D. student in the Dept. of ECE and the Virginia Bioinformatics Institute at Virginia Tech. He received his B.S. degree in Computer Science from Dankook University in 1999, and his M.S. degree in Computer Science from Texas A&M in 2001. Before joining Virginia Tech in 2007, he worked with Samsung Electronics and ETRI, Korea, as a Research Staff Member. junwhan@vbi.vt.edu.

ANIL KUMAR is is currently at the Dept. of Computer Science and the Virginia Bioinformatics Institute at Virginia Tech. His interests are in the broad areas of algorithms, combinatorial optimization, probabilistic techniques and distributed computing, and their applications to wireless networks, epidemiology, and the modeling, simulation and analysis of social and infrastructure networks. akumar@vbi.vt.edu.

ACHLA MARATHE is an associate professor at the Virginia Bioinformatics Institute and, at the Dept. of Agricultural and Applied Economics at Virginia Tech. She is also the lead economist and social scientist at the Network Dynamics and Simulation Science Laboratory at the Virginia Bioinformatics Institute. Before joining Virginia Tech, she worked at the Los Alamos National Laboratory for ten years where she worked on a number of projects involving fraud detection, data mining, and simulation and modeling of the socio-technical systems. She has been doing research in critical infrastructures, behavioral economics, and the economics of epidemiology. amarathe@vbi.vt.edu.

GUANHONG PEI is a Ph.D. student in the Dept. of ECE and the Virginia Bioinformatics Institute at Virginia Tech. Before that, he received his B.S.'s in both ECE and Business Administration from Shanghai Jiao Tong University, China, in 2006. He worked as research interns at Alcatel-Lucent Bell Labs, NJ and at DOCOMO USA Labs, CA. somehi@vbi.vt.edu.

SUDIP SAHA a Ph.D. student in Computer Science at Virginia Tech. He received his BS in computer science and engineering from Bangladesh University of Engineering and Technology in 2006 and MS in computer science from The University of Memphis, TN in 2010. ssaha@vbi.vt.edu.

BALAAJI SUNAPANASUBBIAH is a master student in the Dept. of ECE and the Virginia Bioinformatics Institute at Virginia Tech. basp@vbi.vt.edu.