Week 9: Continuous-time Markov chains
Cellular network design

As the name suggest, cellular communication systems—such as the one providing service to your smartphone right now—divides the area of service into small units called *cells*. Each of these cells is serviced by a base station (for which we use the unsightly abbreviation BS). There are two reasons for this arrangement: (i) electromagnetic power decay with distance and (ii) the electromagnetic spectrum is scarce. Power decay is the reason why there are places in which your phone does not work (has no coverage) and you get a no-service message. Spectrum scarcity is the reason why there sometimes you cannot place a call and get a busy signal or your Internet connection drops even though you have bars. Power decay is the main concern in sparsely populated areas where each BS has to cover a large area. Spectrum scarcity takes precedence in densely populated areas where a large number of customers may simultaneously require service in a small area.

The effect of having a limited amount of bandwidth is that there is a limit in the number of calls that a BS can handle. Exactly what this limit is and how the system behaves close to this limit depends on the type of technology used. For our purposes, suffices it to say that there is a maximum number K of calls that the system can handle. A common problem faced by systems engineers is determining whether an existing cell needs to be subdivided using statistical traffic information collected by the BSs.

To model this problem, assume that customers behave independently. It then follows that the time between call requests T_c is exponentially distributed with mean λ, i.e.,

$$T_c \sim \exp(\lambda). \tag{1}$$

If the BS has no more channels available, something that happens if there are already K calls established, service is denied and the customer request to establish a call fails. Otherwise, the call is established and a channel is assigned to the customer for the duration of that call. The duration of calls, T_d, is also modeled by an exponentially distributed random variable with parameter μ, i.e.,

$$T_d \sim \exp(\mu). \tag{2}$$

This exercise is roughly divided in two sections. In the first section, comprised of parts A–I, you are asked to build and analyze a stochastic model for the placement of calls in the service area of a BS. In the second section, made up of parts J and K, you are asked to decide when a new BS needs to be added.

A Departure process. We say a departure occurs whenever a call is completed. Fix a given time t and let $1 \leq k \leq K$ be the number of calls in service at that time. Consider $t + T_{di}$, with $1 \leq i \leq k$, to be the random time at which the i-th customer hangs up. Recall from (2) that T_{di} is exponentially distributed with parameter μ. If T_k is the random time until the next departure (until the next costumer hangs up), write T_k as a function of the T_{di} and show that its probability distribution is exponential with parameter $k\mu$.

B Four simple questions on the departure process. Given that there are k calls established, what is the probability that Customer 1 will be first to complete his call? Customer i has been talking for $s_i = 2$ minutes, while Customer j has been on the phone for $s_j = 10$ minutes. What is the probability of Customer i hanging up before Customer j? If $1/\mu = 3$ minutes, what is the probability of $T_{di} > 3$ minutes? What is the probability of $T_{dj} > 3$ minutes?

C Continuous time Markov chain (CTMC) model. The number $X(t)$ of calls established at time t can be modeled as a CTMC with states $0 \leq k \leq K$. Explain why and specify the transition rates q_{ij} from state i to state j. Notice that most of these transition rates are null. Draw a transition diagram.

D Alternative CTMC representation. Give expressions for the transition rates ν_k out of state k and the transition probabilities P_{ij}. Recall that the P_{ij} denote the probability of going from state i to state j given that the CTMC is transitioning out of state i.

E Embedded Markov chain (MC) and ergodicity of the CMTC. Specify the embedded discrete time MC associated with the CTMC $X(t)$. Explain why the CTMC $X(t)$ is ergodic.

F System simulation. Write a function to simulate the placing of calls in the service area of a BS. Inputs to the function are the call rate λ, average call duration $1/\mu$, maximum number of channels K, and the duration $t_{\max}$ of the simulation. The outputs are a vector of transition times $\boldsymbol{t}$ and a vector $\boldsymbol{X}$ with the corresponding values of the number of calls in service. Run your simulation for call rate $\lambda = 25$ calls/minute, average call duration $1/\mu = 56$ seconds, number of available channels $K = 32$, and duration $t_{\max} = 30$ minutes.

G Limit distribution. Define the limit distribution of the CTMC as the set of probabilities P_k that the CTMC is in state k for t sufficiently large, i.e.,

$$P_k := \lim_{t\to\infty} P_{ik}(t). \tag{3}$$

Find the probabilities P_k for all $0 \leq k \leq K$. A good approach is to express all probabilities in terms of P_0 and then use the fact that $\sum_{k=0}^{K} P_k = 1$.

H Ergodic limits. As for discrete time MCs, it is possible to relate the limit probabilities in (3) with the proportion of time spent visiting state k, i.e., with the ergodic limits

$$\bar{p}_k = \lim_{t\to\infty} \bar{p}_k(t) = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{I}\left[X(s) = k\right] ds. \tag{4}$$

We have claimed that the average times in (5) are more important than the probabilities in (3) because they give information about a single realization of the process (ergodic average), whereas the probabilities (3) are a measure across all realizations (ensemble average).

In many cases, both of this quantities coincide and we can write

$$\bar{p}_k = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{I}\left[X(s) = k\right] ds = P_k \quad \text{almost surely.} \tag{5}$$

Explain under which conditions (5) is true and discuss if they are valid for the problem considered here. Write down the values of $\bar{p}_k$ for $0 \leq k \leq K$. Does the expression in (5) means that $\bar{p}_k = P_k$ holds for *all* realizations of the process $X(t)$? You are advised to think twice before answering this question.

I Approximating P_k using a simulation. For the same parameters of Part F (except for the value of $t_{\max}$ that you are asked to choose), compute an approximate value of the probabilities in (3) using your simulation code. You can choose to answer this question using a single run of the simulation (if this is at all possible) or using multiple runs. What is the best option? Whatever you choose to do, you must select $t_{\max}$ and/or the number of runs to guarantee that your approximation has an accuracy of at least 5×10^{-3}.

J Blocked call probability. As a first step to determine whether a new BS is necessary, compute the probability that a customer is denied service. This is referred to as the *blocked call probability* P_b. Express P_b in terms of λ, μ, and K.

K Do we need another BS? Whenever a user makes a call request, the BS logs the attempt in a database. Typically, the BS reports an aggregate metric like "number of call requests" at periodic intervals, say every half hour. The information you are given as a system engineer is a large database containing the total number of call requests in 30 minutes intervals during the last year. I have done part of the work for you, which is the selection of the largest 10 values:

Date	**Time**	**Call attempts**	**Date**	**Time**	**Call attempts**
12/24	22:00 - 22:30	1,498	12/24	22:30 - 23:00	1,390
11/23	17:00 - 17:30	1,134	12/24	23:00 - 23:30	1,127
11/23	17:30 - 18:00	1,109	10/13	16:00 - 16:30	913
9/15	17:30 - 18:00	892	8/18	17:00 - 17:30	872
6/13	16:00 - 16:30	865	3/9	17:30 - 18:00	851

A typical design criteria is to discard the two busiest days of the year. Of the remaining values, also discard the two largest. The next largest value is your design target. The company plans a 5% increase in traffic for the upcoming year, which you should take into account in your design target. Then, using this target number of calls in a half hour interval, estimate the target arrival rate λ. Let the average call duration be $1/\mu = 56$ seconds, the number of available channels be $K = 32$, and maximum blocked call probability requirement be 0.02. Does the company need another BS? (pun intended).

Addendum 1

Although we used a cellular system as an example, the problem you just solved appears in many different areas. Keeping close to communications, this type of service dimensioning is also needed to determine the number of customer representatives in a call center or servers in a datacenter. A minor variation would tell you about the shelving of products in a supermarket or stocking of parts in a factory.

Addendum 2

On a different note, the fact that we discarded the busiest days when designing the system might give you a hint as to why it is pretty much impossible to place a call on Christmas or at the end of a football game. Still, the system is dimensioned for a very busy half hour of a very busy day. In fact, the BS is grossly over-dimensioned for most of the days and times. This problem is common to all utilities, most notably to the production and distribution of electric energy. The power capacity installed has to be able to support the most demanding time of the most demanding day of the year—most likely a hot summer day. This is one of the biggest limitations of renewable energies: because their availability cannot be guaranteed—the sun might not be shining or the wind might not be blowing—, renewable energy requires conventional energy as backup, consequently duplicating investment. This is another CTMC that you might want to study. A conclusion you will find is that you want to pool renewable energy from different sources and different geographical areas. This is why you hear about the need to develop a national "power superhighway."

What do I need to solve this exercise?

You can solve parts A–F given the material covered last week. The remaining parts can be solved after Monday's lecture. The topics on Wednesday and Friday are similar to what you are asked to do in parts G, H, J, and K. However, you do not need to wait until then and I strongly advise you not to do so.