Data Mining in Big Dynamic Networks

Tutorial 1: simulating dynamic networks

Ernst C. Wit

May 7, 2024

1 Preliminary information

1.1 Theoretical Overview

Counting Process. A **counting process** is a stochastic process $N(t), t \geq 0$ that represents the total number of events that have occurred up to time t. The function $N(t)$ satisfies the following properties:

- $N(0) = 0$ (initial condition);
- $N(t)$ is integer-valued for all $t \geq 0$;
- $N(t)$ is non-decreasing as t increases; that is, if $s < t$, then $N(s) \leq N(t)$;
- The function $N(t)$ is right-continuous, meaning that for each t, $\lim_{s \to t+} N(s) = N(t)$

Poisson Process. A counting process is said to be a **Poisson process** with rate $\lambda(t) > 0$ if:

- $N(0) = 0$;
- The number of events in disjoint time intervals are independent;
- The probability of more than one event occurring in a small time interval of length Δt is negligible.
- The rate is proportional to the probability of an event occurring in a small interval,

$$P(N(t+\Delta t) - N(t) = 1) \approx \Delta t \times \lambda(t).$$

For a Poisson process with constant rate λ, the number of events in any interval of fixed length Δt follows a Poisson distribution with parameter $\lambda \Delta t$.

Exponential Waiting Times and Multinomial selection. The probability distribution of the waiting time for the next event in a Poisson process with constant rate λ is exponential with parameter λ.

Let's consider k counting processes simultaneously, $N_1, \ldots, N_k$. Let $T_1, \ldots, T_k$ be independent exponentially distributed random variables with rate parameters $\lambda_1, \ldots, \lambda_k$ Then the random variable:

$$\min\{T_1, \ldots, T_k\}$$

is also exponentially distributed, with parameter equal to:

$$\lambda = \lambda_1 + \cdots + \lambda_k$$

Conditional on an event happening at time t_1, the probability that the event type Y is j is multinomial, i.e.,

$$P(Y = j) = \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i.}$$

Conditional that the first event was j, the distribution of the additional waiting times for the other events is — surprisingly — still exponentially distributed with the same hazard. This is known as the memoryless property of the exponential distribution.

A brief overview on Relational Event Models A relational event is a behavioural action from a sender $s \in \mathscr{V}_1$ towards a receiver $r \in \mathscr{V}_2$ at a certain time $t \in [0, \tau]$ (Butts, 2008).

$$e = (s, r, t)$$

Relational events may be seen as time-stamped edges of a dynamic graph.

$$\begin{aligned}\mathscr{G} &= (\mathscr{V}, \mathscr{E}) \\ \mathscr{V} &= (\mathscr{V}_1, \mathscr{V}_2) \\ \mathscr{E} &= \{e_1, \ldots, e_n\}\end{aligned}$$

A counting process $N_{sr}(t)$ is aimed to count the number of events occurred for the couple (s, r) up to time t. We will focus on the underlying intensity rate of this process, that contributes to the computation of the probability of event (s, r) to occur at time t. According to the formulation above the described counting process is a Non Homogeneous Poisson Process (NHPP) with time-varying rate $\lambda_{sr}(t)$.

$\lambda_{sr}(t)$ may be considered as function of the baseline hazard function and a number of covariates that are allowed to vary in time.

$$\lambda_{sr}(t) = \lambda_0(t) \cdot \exp\left[\beta^T \boldsymbol{x}_{sr}(t)\right]$$

1.2 Packages and Functions

You will need the library ggplot in R.

2 Homogeneous and Non Homogeneous Poisson Processes

Email exchange in a company. We consider a company with n employees and we aim to simulate email exchange within the company. We start the simulation on January 1, 2023, which is a Sunday.

1. **Constant Rate:** We will commence by adopting an exponential inter-arrival time model across the entire period. In this approach, the mean number of emails $\lambda_{sr} = \lambda$ sent per day between any pair of colleagues is constant.

   Perform the following sampling:

   - Sample interarrival times $\Delta t \sim \text{Exp}(n(n-1)\lambda)$
   - Randomly sample pairs of colleagues (Sender, Receiver).

   Until $\sum \Delta t > n_d$ where the entire period is composed of n_d days.

   Create the function email1 that produces a simulated dataset (Sender, Receiver and Timestamp) of email exchanges according to the above simulation method. Inputs of the function are the number of colleagues, the average rate of emails between any pair of colleagues sent per day, and the duration of the simulation in days.

2. **Changing baseline hazard.** Up to this point, emails have been simulated using a constant mean number of emails per day. Now, let's consider a scenario where the rate is λ_{WD} emails per day from Monday to Friday and λ_{WE} emails per day on Saturday and Sunday.

   Perform the following sampling:

   - Sample interarrival times $\Delta t \sim \text{Exp}(n(n-1)\lambda)$, where

   $$\lambda = \begin{cases} \lambda_{WD} & \text{if week day} \\ \lambda_{WE} & \text{if weekend} \end{cases}$$

   - Randomly sample pairs of colleagues (Sender, Receiver).

   Until $\sum \Delta t > n_d$ where the entire period is composed of n_d days.

   Implement the function email2 that produces a simulated dataset (Sender, Receiver and Timestamp) of email exchanges according to the above simulation method. Inputs of the function are the number of colleagues, average number of weekday and weekend emails per day between any pair of colleagues.

3. **Relational event process.** Up to this point, the simulation of senders and receivers for different emails has been random, with no assumption that the sender and receiver of one email have any influence on the subsequent emails. However, it's quite natural to consider that when individual A sends an email to individual B, this may increase the likelihood of B reciprocating and sending an email back to A, representing a form of reciprocity. Furthermore, one can reasonably assume that if A has written to B once, the probability of this event recurring is higher, reflecting a sense of repetition. Express $\lambda_{sr}(t)$

   $$\lambda_{sr}(t)dt = \lambda_0(t)\exp\{\beta_1 \times \text{reciprocity} + \beta_2 \times \text{repetition}\}$$

   where:

   $$\lambda_0(t) = \begin{cases} \lambda_{0WD} & \text{if week day} \\ \lambda_{0WE} & \text{if weekend} \end{cases}$$

   and:

   - Reciprocity:

   $$\text{Reciprocity}_{sr}(t) = \begin{cases} 1 & \text{if } t_{rs} > t - 2 \text{ days} \\ 0 & \text{otherwise} \end{cases}$$

   - Repetition:

   $$\text{Repetition}_{sr}(t) = \begin{cases} 1 & \text{if } t_{sr} > t - 2 \text{ days} \\ 0 & \text{otherwise} \end{cases}$$

   (a) *Time Simulation*:
   Waiting times (time intervals between events) are generated according to an Exponential distribution with a rate equal to the sum of the rates of all the possible pairs of colleagues.

   $$\Delta T \sim \text{Exponential}\left(\sum_{(s,r)\in\text{Colleague Pairs}} \lambda_{sr}(t)\right) \tag{1}$$

   Note: if you select a time that jumps from a week day to the weekend (or vice versa), you can make use of the memoryless property of the exponential distribution: discard the sampled time and simply sample another time starting from midnight with the new rate.

(b) *Pair Selection*:

Given that at a certain point t a relational event occurs, you calculate the probability that it an email from employee s the colleague r as:

$$\frac{\lambda_{sr}(t)}{\sum_{(s',r')\in\text{Colleague Pairs}} \lambda_{s'r'}(t)} \tag{2}$$

Create the function `email3` that produces a simulated dataset (Sender, Receiver and Timestamp) of email exchanges according to the above simulation method. Inputs of the function are the number of colleagues n, the coefficients for reciprocity β_{rec} and repetition β_{rep}, the duration of the simulation in days n_d, and the vector `lambda0`.

Explore which possible values may be adequate for the elements of `beta` and `lambda0`. To do so, we suggest using positive values for the parameters in `beta`.