

# Patent mining

Antonin Bergeaud, Yoann Potiron and Juste Raimbault

April 11, 2016

We define  $N$  the number of patents in the database. We let  $t_1, \dots, t_N$  their corresponding submission date by increasing-time order. For each patent  $i = 1, \dots, N$ , we consider  $C_i$  the number of citations made by the  $i$ th patent as well as  $C_i^{(tec)}$  the number of citations made to patents of its own technological class and  $C_i^{(sem)}$  the number of citations made to patents of its own semantic class. Finally, we define respectively  $N_i^{(tec)}$  and  $N_i^{(sem)}$  the number of patents which belonged to the technological (semantic) class of the  $i$ th patent when it was submitted.

Let  $z \in \{tec, sem\}$ . If we assume that the type of citations made  $C_i^{(z)}$  are independent of the technological or semantic classes, we can expect  $C_i^{(z)}$  to follow a  $Bin(C_i, \frac{N_i^{(z)}}{N_i-1})$  conditioned on  $C_i$  and  $N_i^{(z)}$ . This assumption is not satisfied in practice. Consequently, we define the parameter  $0 < \theta^{(z)} < 1$  which indicates the propensity for any patent to cite patents of its own technological or semantic class. We assume that  $C_i^{(z)}$  follows a  $Bin(C_i, \min(1, \frac{N_i^{(z)}}{N_i-1} + \theta))$ .

We assume that the number of citations are a sequence of IID random variables, and also independent of any other quantity. This will facilitate the likelihood analysis in the following. We have

$$\log \mathcal{L}(C_N, C_N^{(z)}, \dots, C_1, C_1^{(z)}) = \sum_{i=2}^N \log \mathcal{L}(C_i, C_i^{(z)} | C_{i-1}, C_{i-1}^{(z)} \dots, C_1, C_1^{(z)}) + \log \mathcal{L}(C_1, C_1^{(z)})$$

and bla bla