MAP554 – Networks: PC 6

November 16, 2016

Threshold for giant component in the stochastic block model

For fixed $K > 0$ let $a_1, \ldots, a_K$ be positive numbers such that $\sum_{k=1}^K a_k = 1$ and let $B = (b_{ij})_{1 \le i,j \le K}$ be a symmetric matrix with strictly positive entries. Consider the stochastic block model, that is a random graph G on node set $[n]$ (which we will let go to infinity), partitioned into K blocks, where block $k \in [K]$ consists of $a_k n$ nodes. For two nodes i, j from respective blocks k, ℓ, edge (i, j) is present with probability $b_{k\ell}/n$, independently of the presence of other edges.

Let $C^{(1)}$ denote the size of the largest connected component of G, and ρ the spectral radius of matrix $(a_i b_{ij})_{1 \le i,j \le K}$. Our aim in this problem is to establish the following:

$$\rho < 1 \Rightarrow \exists A > 0 \text{ such that } \lim_{n \to \infty} \mathbb{P}(C^{(1)} > A \log(n)) = 0. \tag{1}$$

We consider the following process for enumerating the connected component of G containing an arbitrary node from some block $k_0 \in [K]$. At each step one deactivates all nodes active at the previous step, and activates all neighbors of these previously active nodes that have not yet been activated.

1.1 Let $Z_\ell(t)$ denote the number of type ℓ nodes that are active at step t. Show that conditionally on $\mathcal{F}(t-1) := \{Z_k(s)\}_{k \in [K], s < t}$, one has

$$Z_\ell(t) \le \sum_{k=1}^K \sum_{s=1}^{Z_k(t-1)} X_{k,\ell}(s, t), \tag{2}$$

for some random variables $X_{k,\ell}(s, t)$ that are independent from $\mathcal{F}(t-1)$, mutually independent, and where $X_{k,\ell}(s, t)$ follows a Binomial distribution with parameters $(a_\ell n, b_{k\ell}/n)$.

1.2 Deduce for all $\theta_\ell > 0$, $\ell = 1, \ldots, K$ the inequality

$$\mathbb{E}\left[e^{\sum_{\ell=1}^K \theta_\ell Z_\ell(t)} | \mathcal{F}(t-1)\right] \le e^{\sum_{k,\ell=1}^K (e^{\theta_\ell} - 1) a_\ell b_{k\ell} Z_k(t-1)}. \tag{3}$$

1.3 Show that under the hypothesis $\rho < 1$, there exists $\theta > 0$ and $B < +\infty$ such that

$$\mathbb{E}\left[e^{\theta \sum_{k=1}^K \sum_{t \ge 0} Z_k(t)}\right] \le B < \infty. \tag{4}$$

Hint: one can consider a vector $\{x_\ell\}$ with strictly positive coordinates such that for all $k \in [K]$, $\sum_\ell x_\ell a_\ell b_{k\ell} = \rho x_k$, and then use (3) with $\theta_\ell = \epsilon x_\ell$ for $\epsilon > 0$ small enough.

1.4 Deduce the announced result (1).

Randomized content search and adaptive content replication

Consider a system consisting of n servers, collectively in charge of storing m content items. For each item $i \in [m]$, n_i servers hold a copy of it. We assume that, when a request for content item i is put on the system, the following randomized search takes place. Servers are checked in an i.i.d. uniformly distributed fashion until a copy of content i is found.

2.1 Determine the distribution of the number of servers contacted before a copy of i is found and its average.

2.2 Assume that the total number of content copies that can be stored in the system is Cn for some constant C, so that necessarily $\sum_{i\in[m]} n_i \leq Cn$. Let π_i be the probability that a new content request is for item $i \in [m]$.

Relaxing the constraint that $n_i \in \mathbb{N}$ to $n_i \in \mathbb{R}_+$, determine the optimal replication numbers n_i^* which minimize the average search time per requested item as a function of C and $\{\pi_j\}_{j\in[m]}$.

2.3 Assume that there is on average one request per time unit, and after each request, if it is for content i and it took T_i steps to find a server holding a copy of it, then T_i new copies of content i are distributed across servers. Assume further that each content copy is suppressed within time interval $[t, t + dt]$ with probability λdt, independently of everything else.

Heuristically establish a differential equation for the number of copies $n_i(t)$ of content i at time t.

2.4 Determine the stationary points of the previous differential equation and relate them to the allocations $n*_i$ previously identified. Prove convergence of the ODE's solution to these stationary points.

The above content replication scheme has been proposed by Cohen and Schenker in the context of peer-to-peer systems. There, successive servers are visited by performing a random walk along a graph connecting them; the new T_i replicas are then placed at the T_i servers visited before finding a replica for content i. This approach is therefore referred to as path replication.