Newer bounds in the Switching Algorithm: fixed point and time convergence.

Andrea Gobbi[1,2], Giuseppe Jurman[2]
[1] Department of Mathematics, Univerisity of Trento, Italy
[2] Fondazione Bruno Kessler, Trento, Italy

The Switching Algorithm SA is the most common algorithm used for the generation of a random version of a given graph preserving the degree distribution ([1, 2, 3]). Let $\mathcal{G} = (E, V)$ a undirected graph without loops with e edges and with edge density d. The SA is composed by a sequence of N Switching Steps SS in which the following actions are performed: two edges (a, b) and (c, b) are randomly selected; if $(a, d), (c, b) \notin E$ the edges $(a, b), (c, b)$ are removed and $(a, d), (c, b)$ are added and the SS ends; otherwise, if $(a, c), (b, d) \notin E$ the edges $(a, b), (c, b)$ are removed and $(a, c), (b, d)$ are added and the SS ends; if both the previous situations are satisfied then one pair is chosen randomly (Fig.1). The SA is widely used for the generation of null models preserving the degree distribution. Usually in literature the number N of SSs is chosen empirically (e.g. $N = 100e$ [3]). In this work we first formalize the concept of "random version" of a graph and then we compute a theoretical lower bound N in order to obtain such random version.

Let $x^{(k)}$ be the number of common edges between $\mathcal{G}$ and its rewired version $\mathcal{G}^{(k)}$ after k SSs and let $0 < p_r \leq 1$ be the probability to perform a SS. We measure the "randomness" between $\mathcal{G}^{(k)}$ and $\mathcal{G}$ through the Jaccard Index JI $s^{(k)}$ that in this case reads as $s^{(k)} = \frac{x^{(k)}}{2e - x^{(k)}}$. Since the JI is an injective function we need only to consider the behaviour of $x^{(k)}$. We prove that:

$$x^{(k)} \to \overline{x} \qquad \text{for} \qquad k \to +\infty \qquad \text{and that}$$

$$|x^{(k)} - \overline{x}| < 1 \qquad \text{for} \qquad k > N = \frac{e(1-d)}{2p_r} \ln e(1-d).$$

In the case of bipartite graphs:

$$N = \frac{e}{2(1-d)} \ln e(1-d)$$

as proved in an our recent work [4], otherwise

$$N = \frac{e}{2d^3 - 6d^2 + 2d + 2} \ln e(1-d).$$

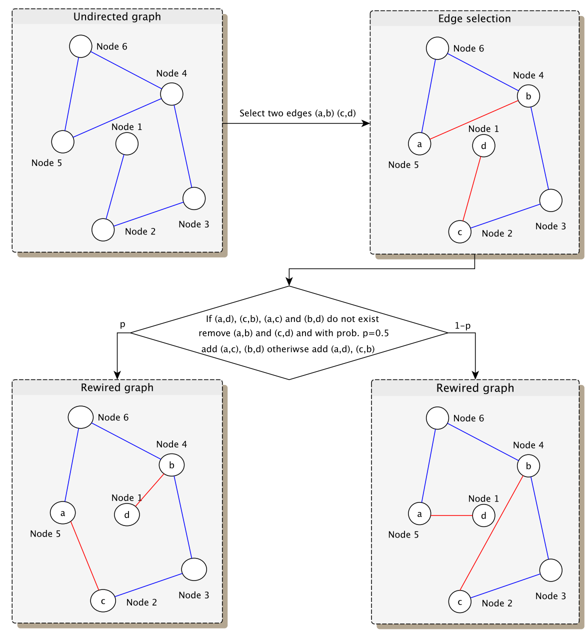

Figure 1: Scheme of a SS in the SA.

In order to derive such bound, we estimate the mean value of $x^{(k+1)}$ using the value $x^{(k)}$; indeed, after a SS, the number of common edges between $\mathcal{G}^{(k+1)}$ and $\mathcal{G}$ can increase (decrease) by 1, increase (decrease) by 2 or remain unchanged w.r.t. the number of common edges between $\mathcal{G}^{(k)}$ and $\mathcal{G}$. For each of these five possibilities $f_i(x^{(k)})$ we compute the relative probabilities $p_i(x^{(k)})$. We estimate the mean number of common edges at the $(k + 1)$-th step as

$$x^{(k+1)} = \sum_{i=1}^{5} p_i(x^{(k)}) f_i(x^{(k)})$$

and the result is a second order linear sequence for which a closed form reads as

$$x^{(k+1)} = m^{k+1} + q \qquad \text{with} \qquad m = \frac{et - 2p_r(t - z) - e^2 - ez}{(t - e - z)e} \qquad \text{and} \qquad q = \frac{2ep_r}{t - e - z}.$$

We prove that $0 < m < 1$ and so the unique fixed point of the recursive equation is q. Finally we show that $|x^{(k)} - \overline{x}| < 1$ of $k > N(z) = \log_m \frac{t-z}{(t-e-z)e}$. The value z is an unknown parameter related to the number of possible random graphs that could be obtained from $\mathcal{G}$ preserving the degree distribution. We also prove that $N(z)$ has a maximum for $z = 0$ and so the value $N = N(0)$ can be chosen as a bound. Finally, by using the same strategy described above, we prove that for each step k of the SA, the number of common edges between $\mathcal{G}^{(k)}$ and $\mathcal{G}$ is lager than the number of common edges between $\mathcal{G}^{(k)}$ and $\mathcal{H}^{(k)}$ where $\mathcal{H}^{(k)}$ is a different instance of the SA starting from $\mathcal{G}$.

References

[1] Ciriello, G. and Cerami, E. and Sander, C. and Schultz, N.(2012) Mutual exclusivity analysis identifies oncogenic network modules, *Genome Research*, **22**, 398-406.

[2] Miklòs I, Podani J. Randomization of presence-absence matrices: comments and new algorithms. Ecology. Eco Soc America; 2004;85(1):86–92.

[3] Milo, R. and Kashtan, N. and Itzkovitz, S. and Newman, M. E. J. and Alon, U. (2004), On the uniform generation of random graphs with prescribed degree sequences *arXiv*.

[4] Gobbi, A. and Iorio, and Wedge, D. and Dawson, K. and Ludmil, A. F. and Jurman, G. and Saez-Rodriguez, J. (in preparation), *Preserving genomic-event distributions in randomized next generation sequencing data*.