SHORT COMMUNICATION

A GENERATING FUNCTION APPROACH TO RANDOM SUBGRAPHS OF THE n-CYCLE

XAVIER GOURDON AND HELMUT PRODINGER

Abstract. Given a cycle with n nodes a random subgraph is created by 'accepting' edges with probability p and 'rejecting' them with probability $q = 1-p$. The parameter of interest is the order of the largest component. There are some partial answers to this question in the literature. Using an appropriate encoding by formal languages, we present here a complete solution. Singularity analysis of generating functions gives good approximations of the probabilities, and the asymptotic evaluation of expectation and variance is performed by the Mellin (integral) transform. For instance, the expected order is like a logarithm of n plus an oscillating function.

1. Introduction

We denote by C_n the directed cycle of size n and by $M_n(H)$ the order of the largest connected component in a random subgraph H of C_n, where H has the same vertex set as C_n, and its edge set is defined by independently selecting, with the same constant probability p, each of the edges of C_n. The random variable M_n appears in various papers: the probability $\Pr(M_n = k)$ has been studied in [5] for the case $k = 1$ and $\lfloor n/2 \rfloor \le k \le n$; in [3], the case $k = 2$ is treated and an asymptotic value for fixed k as n tends to infinity is given.

Here, we give an asymptotic value of the value $\Pr(M_n \le k)$ for all k (varying from 1 to n) when p is fixed, as n tends to infinity. As a consequence, we show that M_n has an expected value asymptotic to $\log_{p^{-1}}(n)$, and that the variance of M_n is asymptotically constant. The precise asymptotic expansion (given below) can in principle be extended to any order of accuracy.

2. The generating function model

The generating function approach is classical to study such problems, and as we shall see on this particular case, it is a natural tool to work with.

First we describe an equivalent model which translates easily in terms of generating functions. Suppose we construct from C_n what we call a random n-ab-cycle C, that is an n-cycle having two types of edges a and b, such that every edge has a constant probability p ($0 < p < 1$) to be of type a, $q = 1-p$ to be of type b. By selecting edges of type a on C, you can construct a random subgraph $H(C)$ of C_n and conversely, each subgraph H of C_n can be seen as the selection of edges of type a in an n-ab cycle C. This one-to-one construction has also the property that a random n-ab-cycle C has the same probability of appearance than the random subgraph $H(C)$, this random subgraph being constructed as described in the introduction. If we denote by m_n the maximal number of consecutive a's in a random n-ab-cycle, we have

$$m_n(C) = M_n(H(C)) - 1 \quad \text{if} \quad m_n(C) < n, \qquad m_n(C) = M_n(H(C)) \quad \text{if} \quad m_n(C) = n. \tag{1}$$

The work of the second author was done while he visited INRIA. He is thankful for the warm hospitality encountered there.

The ab-cycles are in bijection with the words over the alphabet $\{a,b\}$, but do not have the same topological properties. From this observation and from (1), it is easy to see that the ab-cycles C such that $M_n(H(C)) \le k$ can be written as

$$\sum_{i+j<k} a^i (ba^{<k})^* ba^j + a^{\le k}$$

where

$$a^{<k} = \varepsilon + a + a^2 + \cdots + a^{k-1}.$$

If we mark edges of type a by the variable u and edges of type b by v, the corresponding generating function is

$$F_k(u,v) = \sum_{i+j<k} u^i \frac{1}{1 - v(1+u+\cdots+u^{k-1})} vu^j + (1+u+\cdots+u^k). \tag{2}$$

Because edges of type a can appear with probability p and edges of type b with probability q, it follows by a standard argument using generating functions that[1]

$$\Pr(M_n \le k) = [z^n]G_k(z), \qquad G_k(z) = F_k(pz, qz).$$

¿From (2), a short calculation leads to

$$G_k(z) = \frac{qz(1-pz)}{1 - z + \frac{q}{p}(pz)^{k+1}} A_k(pz) + B_k(pz),$$

where

$$A_k(z) = \sum_{i=0}^{k-1} (i+1) z^i = \frac{1-(k+1)z^k + kz^{k+1}}{(1-z)^2} \quad \text{and} \quad B_k(z) = \sum_{i=0}^{k} z^i = \frac{1-z^{k+1}}{1-z}.$$

3. Analysis

This section is devoted to the asymptotic analysis of $[z^n]G_k(z)$, which is quite analogous to that of [4]. For that purpose, we look for the dominant singularity of the function $G_k(z)$. This function being rational, its singularities are poles given by the roots of the polynomial

$$P_k(z) = 1 - z + \frac{q}{p}(pz)^{k+1}.$$

When k is large, the value $P_k(1)$ is small, thus it is natural to expect that the polynomial $P_k(z)$ has one root close to 1, namely $1+\varepsilon_k$. In fact, we have the following result.

Lemma 1. *The polynomial $P_k(z)$ has only one root in $|z| < 1/p$. Calling this root $1+\varepsilon_k$, we have*

$$\varepsilon_k = qp^k(1 + O(kp^k)) \quad as \quad k \to \infty.$$

Proof. Because $|\frac{q}{p}(pz)^{k+1}|$ is strictly less than $\frac{q}{p}$ for $|z| = \rho < 1/p$, hence less than $|1-z|$ on this circle, we can apply Rouché's theorem (see [1, p. 247] or [6, p. 116]) implying that $P_k(z)$ has the same number of roots as the polynomial $1 - z = P_k(z) - \frac{q}{p}(pz)^{k+1}$ in $|z| \le \rho$. This being true for all $\rho < 1/p$, we have proved the first part of the lemma.

Now, noticing that $P_k(1+1/k)$ is negative for large k, we find that $\varepsilon_k = O(1/k)$. Since

$$\varepsilon_k = qp^k(1+\varepsilon_k)^{k+1} \tag{3}$$

we have $\varepsilon_k = O(p^k)$, thus by equality (3) again ("bootstrapping"), we find

$$\varepsilon_k = qp^k(1 + O(kp^k))$$

[1] The notation $[z^n]G(z)$ denotes the coefficient of z^n in the series $G(z)$.

and the lemma is proved. □

We return to the study of the asymptotic behavior of $\Pr(M_n \le k) = [z^n]G_k(z)$. We fix ρ such that $1 < \rho < 1/p$. By the residue theorem, we have, when k is large enough,

$$\Pr(M_n \le k) + \frac{c_k}{(1+\varepsilon_k)^{n+1}} = \frac{1}{2i\pi} \int_{|z|=\rho} G_k(z)\, dz,$$

where c_k denotes the residue at $z = z_k = 1 + \varepsilon_k$ of the function $G_k(z)$. We have

$$c_k = \frac{q z_k (1 - p z_k)}{-1 + (k+1)\frac{q}{p}(p z_k)^k} A_k(p z_k),$$

and since we have $G_k(z) = O(1)$ uniformly for $|z| = \rho$, the Cauchy integral shows that

$$\left| \Pr(M_n \le k) + \frac{c_k}{(1+\varepsilon_k)^{n+1}} \right| = O(1/\rho^n) \tag{4}$$

uniformly in k. The explicit value we found for $A_k(z)$ implies that

$$A_k(p z_k) = \frac{1}{(1 - p z_k)^2} + O((p z_k)^k) = \frac{1}{(1-p)^2} + O(p^k),$$

because $z_k = (1 + \varepsilon_k)$ with $\varepsilon_k = O(p^k)$. ¿From this, we derive

$$c_k = \frac{q(1-p) + O(p^k)}{-1 + O(kp^k)} \cdot \left(\frac{1}{(1-p)^2} + O(p^k) \right) = -\frac{q(1-p)}{(1-p)^2} + O(kp^k) = -1 + O(kp^k).$$

We do not present the following analysis in full detail, because it is quite technical and resembles the one in [4]. When k is "not too large" and "not too small" with respect to n (that is something like $n^{-3} \le p^k \le \log n/n$) we have

$$\begin{aligned}(1+\varepsilon_k)^{-(n+1)} &= \exp\Big[-(n+1)(qp^k + O(kp^{2k})) \Big] = \exp(-nqp^k)\Big(1 + O(p^k) + O(nkp^{2k})\Big) \\ &= \exp(-nqp^k)\left(1 + O\left(\frac{\log^3 n}{n}\right)\right),\end{aligned}$$

and thus for such values of k, we find by means of (4)

$$\Pr(M_n \le k) = \exp(-nqp^k) + O\left(\frac{\log^3 n}{n}\right). \tag{5}$$

When k is not in this range, it is possible to show that this estimation holds with an error term of $O(n^{-2})$ instead of $O(\log^3 n/n)$.

4. Mean and variance

Thanks to a classical Abel transformation (summation by parts), the mean value of the random variable M_n can be expressed as

$$\mathrm{E}(M_n) = \sum_{k=0}^{n} \Pr(M_n > k) = \sum_{k=0}^{n} \Big(1 - \Pr(M_n \le k)\Big).$$

¿From (5) — which holds for $O(\log n)$ values of k — and from our previous discussion, one can show

$$\mathrm{E}(M_n) = \sum_{k=0}^{n} \left(1 - e^{-nqp^k}\right) + O\left(\frac{\log^4 n}{n}\right)$$

and since $\sum_{k>n}(1-e^{-nqp^k})$ is exponentially small, we have

$$\mathrm{E}(M_n) = E_n + O\left(\frac{\log^4 n}{n}\right), \qquad E_n = \sum_{k=0}^{\infty}\left(1-e^{-nqp^k}\right).$$

The asymptotic expansion for E_n is a classical application of Mellin transform techniques (see [7, p. 278, 292]). The result is

$$E_n = \log_{p^{-1}}(qn) - \frac{\gamma}{\log p} + \frac{1}{2} + \delta(\log_{p^{-1}}(qn)) + O(n^{-1}),$$

where $\delta(x)$ is a 1-periodic function of x,

$$\delta(x) = \frac{1}{\log p}\sum_{k\neq 0}\Gamma\left(\frac{2ik\pi}{\log p}\right)e^{2ik\pi x}.$$

The function δ has mean value 0 and a small amplitude. Such a calculation can be achieved for the variance of M_n, by means of to the relation

$$\mathrm{E}(M_n^2) = \sum_{k=0}^{n}(2k+1)\Big(1-\Pr(M_n\le k)\Big).$$

Finally we obtain the following result.

Theorem 1. *The average order of the largest component in a random cycle C_n is*

$$\mathrm{E}(M_n) = \log_{p^{-1}}(qn) - \frac{\gamma}{\log p} + \frac{1}{2} + \delta(\log_{p^{-1}}(qn)) + O\left(\frac{\log^4 n}{n}\right);$$

its variance is given by

$$\mathrm{Var}(M_n) = \frac{\pi^2}{6\log^2 p} + \frac{1}{12} + \delta_1(\log_{p^{-1}}(qn)) + O\left(\frac{\log^5 n}{n}\right).$$

Here, $\delta(x)$ is the function from above and $\delta_1(x)$ is a 1-periodic function of x, with a small amplitude (namely something like $< 10^{-10}$).

This time, the mean of δ_1 is *not* exactly zero since it contains the square of δ. Roughly speaking, the theorem says that the variance is approximatively a constant.

5. Conclusion

We have seen that a problem on the random circle can be easily analyzed by an appropriate encoding as a formal language and by analysis of generating functions. For paths, this is more obvious and known in the literature ("forbidden subwords", [2]), but with slight modifications it also works in this instance.

This approach is powerful enough to deal with some other parameters of a similar type. For a systematic account we again refer to [7]. Note that equation (5) characterizes the shape of the limit distribution of the random variable M_n.

References

[1] Dieudonné, J. *Calcul Infinitésimal.* Hermann, Paris, 1968.

[2] Guibas, L. J., and Odlyzko, A. M. Periods in strings. *Journal of Combinatorial Theory,* Series A *30* (1981), 19–42.

[3] Katona, G. O. H., and Quintas, L. V. The largest component in a random subgraph of the n-cycle. *Discrete Mathematics 121* (1993).

[4] Knuth, D. E. The average time for carry propagation. *Indagationes Mathematicae 40* (1978), 238–242.

[5] Palka, Z., and Quintas, L. V. Random subgraphs of the n-cycle. *Discrete Mathematics 52* (1984), 243–251.

[6] Titchmarsh, E. C. *The Theory of Functions*, second ed. Oxford University Press, 1939.

[7] Vitter, J. S., and Flajolet, P. Analysis of algorithms and data structures. In *Handbook of Theoretical Computer Science*, J. van Leeuwen, Ed., vol. A: Algorithms and Complexity. North Holland, 1990, ch. 9, pp. 431–524.

Xavier Gourdon, Algorithms Project, INRIA Rocquencourt, F-78150 Le Chesnay, France

E-mail address: `Xavier.Gourdon@inria.fr`

Helmut Prodinger, Institut für Algebra und Diskrete Mathematik, Technische Universität Wien, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria

E-mail address: proding@email.tuwien.ac.at