LEVEL-WISE NODE SIZE DISTRIBUTION OF RANDOMLY GENERATED REGULAR TREES

ELITZA MANEVA, TAKEYA SHIGEZUMI, YUSHI UNO, OSAMU WATANABE

1. Introduction

For analyzing a random graph model for explaining a hierarchical clique structure of large scale Web networks, some statistical properties of random regular trees have been used in [SUW10]. In this note, we give a detailed analysis of these properties.

We consider random k-regular trees for any integer $k \geq 2$ that will be fixed throughout this paper. We consider a branching process that has been known as *Galton-Watson process*. For a given parameter μ_0, $0 < \mu_0 < 1$, the process, starting from an initial node, generates a tree in the following way:

(1) For each *open node* v,
   (a) with probability $p_k = \mu_0/k$, create k new open nodes, add them to v as its child nodes, and change the status of v to *closed*,
   (b) otherwise (i.e., with probability $1 - p_k$), change the status of v to *closed* without adding any child nodes.

(2) Repeat the above until all nodes are closed.

Let T denote a tree generated by this process. The initial node is called a *root* node and a node with no children is called a *leaf* node. For each node v of T, we define its height $h(v)$ and level $l(v)$ inductively as follows.

$$h(v) = \begin{cases} 0, & \text{if } v \text{ is a root node, and} \\ h(v') + 1, & \text{otherwise} \\ & \text{(where } v' \text{ is the parent node of } v\text{);} \end{cases}$$

$$l(v) = \begin{cases} 0, & \text{if } v \text{ is a leaf node, and} \\ \max\{l(v_1), \ldots, l(v_k)\} + 1, & \text{otherwise} \\ & \text{(where } v_1, \ldots, v_k \text{ are child nodes of } v\text{).} \end{cases}$$

The height of a tree is the maximum height of nodes in T. Note that the height of a tree equals the level of the root node of the tree.

The height of T as well as the number of nodes with a given height h have been studied in depth in the literature (see, e.g., [Agr74]). On the other hand, less is

A part of this work has been done while the first and the fourth authors were visiting Centre de Recerca de Matemàtica, Barcelona, Spain.

known about the number of nodes of given level l. The purpose of this note is to show reasonable upper and lower bounds for the expected number of nodes of given level l.

2. Analysis

Let T denote a random k-regular tree generated by the above process. In the following, we assume that T is finite; thus, precisely speaking, probabilities and expectations discussed below are all conditional on the fact that T is finite. Recall that we assume that $kp_k = k(\mu_0/k) = \mu_0 < 1$; on the other hand, it has been known (see, e.g., [Fel68]) that T is finite with probability 1 in this case.

Fix any $l \geq 0$. Let $M(l)$ denote the expected number of nodes with level l in T. Our goal is to give good upper and lower bounds for $M(l)$. For this, we use $P(l)$, the probability that the root has level l, i.e. the depth of T is l.

We analyze $M(l)$ by estimating all possible contributions to it. First, consider the case that the root has level l. If the root has level l, other nodes cannot have level l, so there is only one level l node in T. The root has level l with probability $P(l)$; hence, this contributes $P(l) \cdot 1$ to $M(l)$. Then consider the other case. Since $M(l)$ would be 0 for $l \geq 1$ if the root were not expanded, consider the situation that the root is expanded (which occurs with probability p_k). Let $v_1, \ldots, v_k$ denote the child nodes of the root and let $T_1, \ldots, T_k$ denote the trees rooted by these nodes. Then the contributions from $T_1, \ldots, T_k$ are p_k times the expected number of level l nodes of those trees. Each T_i follows the same probability distribution as T; thus, we may use $M(l)$ for the expected number of level l nodes of T_i. Since these are all contributions, we have

$$M(l) = P(l) + p_k \cdot k \cdot M(l),$$

and, since we assumed that the number of nodes on the tree T is finite, this implies that

$$M(l) = \frac{P(l)}{1 - p_k k} = \frac{P(l)}{1 - \mu_0}. \tag{1}$$

Now our task is to estimate $P(l)$, and we will discuss it in the rest of this note. Let $g(z)$ denote the probability generating function (p.g.f.) of the number of children of a node in our process; that is, $g(z) = 1 - p_k + p_k z^k$. Note that $g'(1) = \mu_0$ is the expected number of children of one node and that we assumed $\mu_0 < 1$. The p.g.f. of the number of nodes with height i, denoted by Z_i, is $g_i(z)$ where $g_1(z) = g(z)$ and $g_j(z) = g(g_{j-1}(z))$ for $j > 1$ [Fel68]. However, it is hard to obtain the closed-form of $g_i(z)$.

Let $q(l)$ denote the probability that the root has level at least l. We here note some basic equations of $P(l)$ and $q(l)$.

$$P(l) = q(l) - q(l+1) \quad (\text{for } l \geq 0) \tag{2}$$

$$q(l) = p_k \left\{ 1 - (1 - q(l-1))^k \right\} \quad (\text{for } l \geq 1) \tag{3}$$

$$q(l) < \mu_0 q(l-1). \quad (\text{for } l \geq 1) \tag{4}$$

Bound (4) is derived from (3) as follows:

$$q(l) = p_k \left\{ 1 - (1 - q(l-1))^k \right\} < p_k \left\{ 1 - (1 - kq(l-1)) \right\} = \mu_0 q(l-1).$$

For an upper bound of $P(l)$, we have the following Lemma.

Lemma 1. *We have $P(0) = 1 - p_k$ and $P(1) = p_k(1-p_k)^k = \frac{\mu_0}{k}(1 - \frac{\mu_0}{k})^k$. For any $l > 1$, we have*

$$P(l) < \frac{\mu_0^l}{k} \left(1 - \frac{\mu_0}{k} \right)^k.$$

Proof. By definition, $P(0)$ and $P(1)$ are the probability that the root node has level 0 and 1 respectively, so we immediately have $P(0) = 1 - p_k$ and $P(1) = p_k(1-p_k)^k$. For any $0 < x < y < 1$, it is easy to show that

$$(1-x)^k - (1-kx) < (1-y)^k - (1-ky).$$

Using this with (2) and (3), we have

$$\begin{aligned} P(l) &= q(l) - q(l+1) = p_k \left[\left\{ 1 - (1 - q(l-1))^k \right\} - \left\{ 1 - (1 - q(l))^k \right\} \right] \\ &< p_k \left[\left\{ 1 - (1 - kq(l-1)) \right\} - \left\{ 1 - (1 - kq(l)) \right\} \right] \\ &= p_k k \left(q(l-1) - q(l) \right) = \mu_0 P(l-1). \end{aligned}$$

Hence we obtained $P(l) < \mu_0^{l-1} P(1) = \frac{\mu_0^l}{k}(1 - \frac{\mu_0}{k})^k$. □

For analyzing a lower bound of $P(l)$, we need both upper and lower bounds of $q(l)$. An upper bound is derived inductively from (4). Noting that $q(1) = p_k = \frac{\mu_0}{k}$, we have

$$q(l) < \frac{\mu_0^l}{k}. \tag{5}$$

For showing a lower bound of $q(l)$, we make use of facts that have been shown in the literature. Note first that $q(l)$ satisfies the following relationships with the p.g.f. $g_l(z)$:

$$\begin{aligned} 1 - q(l) &= \Pr[\, \text{the level of the root node} < l \,] \\ &= \Pr[\, \text{the number of nodes with height } l \text{ is } 0 \,] \\ &= \Pr[\, Z_l = 0 \,] = g_l(0). \end{aligned}$$

However, as mentioned before, the closed-form of $g_l(z)$ is hard to obtain. In [Agr74], Agresti used a fractional linear generating function (f.l.g.f.) to obtain good upper/lower bounds of $g_l(z)$. We follow their analysis and obtain the following lower bound.

Lemma 2. *For any $l \geq 1$, we have*

$$q(l) > \frac{\mu_0^l}{k}(1-\mu_0).$$

Proof. For any p.g.f. $g(z)$, let $U(z)$ be any p.g.f. satisfying $g(z) \leq U(z)$ for $0 \leq z \leq 1$. We first recall the following fact shown by Seneta (Lemma A of [Sen67]).

Fact 1. *For any $l \geq 1$, and for any $0 \leq z \leq 1$, we have*

$$g_l(z) \leq U_l(z),$$

where U_l is defined inductively by $U_l(z) = U(U_{l-1}(z))$ and $U_1(z) = U(z)$.

Proof. Since $U_l(z)$ is a p.g.f., it is an increasing function; also since $g(z)$ is a p.g.f., it satisfies $0 \leq g(z) \leq 1$ for any $0 \leq z \leq 1$. Thus we have

$$\begin{aligned} g_l(z) &= g_{l-1}(g(z)) \\ &\leq U_{l-1}(g(z)) \quad \text{(by induction)} \\ &\leq U_{l-1}(U(z)) \quad (U_{l-1}(z) \text{ is increasing}) \\ &= U_l(z). \end{aligned}$$

□(Fact 1)

Thus, by using some appropriate $U(z)$, we can give the following lower bound of $q(l)$:

$$\begin{aligned} 1-q(l) &= \Pr[\,\text{the level of the root} < l\,] \\ &= \Pr[\,\text{the number of nodes with height } l \text{ is } 0\,] \\ &= g_l(0) \leq U_l(0). \end{aligned} \tag{6}$$

For $U(z)$, we use the following fractional linear generating function (f.l.g.f.) introduced by Agresti ([Agr74], Lemma 3 (i)).

Fact 2. *Define $U(z)$ by*

$$U(z) = 1 - p_k + \frac{p_k z}{k-(k-1)z}.$$

Then, $U(z)$ satisfies $g(z) \leq U(z)$ for any $0 \leq z \leq 1$.

Proof. By definition of $g(z)$ and $U(z)$, it suffices to show

$$g(z) = 1 - p_k + p_k z^k \leq 1 - p_k + \frac{p_k z}{k-(k-1)z}$$

for all z, $0 \leq z \leq 1$. This holds if and only if

$$t(z) = 1 - kz^{k-1} + (k-1)z^k = 1 - z^k - k(1-z)z^{k-1} \geq 0$$

for all z, $0 \le z \le 1$. Note that $t(1) = 0$ and that

$$\begin{aligned} t'(z) &= -kz^{k-1} - k(k-1)z^{k-2}(1-z) + kz^{k-1} \\ &= -k(k-1)z^{k-2}(1-z) \le 0 \end{aligned}$$

for all z, $0 \le z \le 1$. Hence, $t(z) \ge 0$ for $0 \le z \le 1$. □ (Fact 2)

Since $U(z)$ is a f.l.g.f., we can obtain the closed form of $U_l(z)$, the lth iterate of $U(z)$, which is stated as follows (see Appendix for its derivation):

$$U_l(z) = 1 + \frac{\mu_0^l(1-\mu_0)(z-1)}{(k-1)\left(\mu_0^l - 1\right)z + \left(k - \mu_0 - (k-1)\mu_0^l\right)}.$$

Thus from (6) it follows

$$\begin{aligned} 1 - q(l) &\le U_l(0) = 1 - \frac{\mu_0^l(1-\mu_0)}{k - \mu_0 - (k-1)\mu_0^l} \\ &< 1 - \frac{\mu_0^l(1-\mu_0)}{k}, \end{aligned}$$

and hence

$$q(l) > \frac{\mu_0^l}{k}(1-\mu_0).$$ □(Lemma 2)

By (5) and Lemma 2, $q(l)$ can be represented as

$$q(l) = \frac{\mu_0^l}{k}(1-\mu_0) + \epsilon_l,$$

where $0 < \epsilon_l < \frac{\mu_0^{l+1}}{k}$. Now by (4), we have

$$\frac{\mu_0^{l+1}}{k}(1-\mu_0) + \epsilon_{l+1} = q(l+1) < \mu_0 q(l) = \mu_0\left(\frac{\mu_0^l}{k}(1-\mu_0) + \epsilon_l\right).$$

Hence we have $\epsilon_{l+1} < \mu_0\epsilon_l < \epsilon_l$, from which it follows $\epsilon_l - \epsilon_{l+1} > 0$. Thus, we have

$$\epsilon_l - \epsilon_{l+1} = q(l) - q(l+1) - \left\{\frac{\mu_0^l}{k}(1-\mu_0) - \frac{\mu_0^{l+1}}{k}(1-\mu_0)\right\} > 0.$$

From this bound, we obtain the following lower bound of $P(l)$:

$$P(l) = q(l) - q(l+1) > \left\{\frac{\mu_0^l}{k}(1-\mu_0) - \frac{\mu_0^{l+1}}{k}(1-\mu_0)\right\} = \frac{\mu_0^l}{k}(1-\mu_0)^2.$$

Then from this bound and Lemma 1, we obtain the following upper and lower bound of $P(l)$:

$$\frac{\mu_0^l}{k}(1-\mu_0)^2 < P(l) < \frac{\mu_0^l}{k}\left(1 - \frac{\mu_0}{k}\right)^k.$$

We now obtained both upper and lower bounds of $P(l)$, using them we have

$$\frac{\mu_0^l}{k}(1-\mu_0) < M(l) < \frac{\mu_0^l}{k}\left(1-\frac{\mu_0}{k}\right)^k \frac{1}{1-\mu_0} < \frac{\mu_0^l}{k}\frac{1}{1-\mu_0}.$$

From this, the following Theorem is derived.

Theorem 3. *Let* $C_1 = 1-\mu_0$ *and* $C_2 = \frac{1}{1-\mu_0}$. *Then for any* $l \geq 0$, *we have*

$$C_1\mu_0^l\frac{1}{k} < M(l) < C_2\mu_0^l\frac{1}{k}.$$

3. Concluding remarks

In this note, we discuss a branching process and give detail analysis for the expected number of nodes with level l. We focus on the special p.g.f. $g(z) = 1 - p_k + p_k z^k$. Many detailed analysis of $P(l)$ and $q(l)$ of other p.g.f. were given in the literature, e.g., [Fel68, AN72, Har63], so we can apply these analysis to Equation 1, and obtain the expected number of nodes with level l.

References

[Agr74] Alan Agresti, Bounds on the extinction time distribution of a branching process, *Advances in Applied Probability*, 6(2):322–335, 1974.

[AN72] Krishna B. Athreya and Peter E. Ney, *Branching Processes*, Springer-Verlag, Berlin, Heidelberg, 1972.

[Fel68] William Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, 3 edition, January 1968.

[Har63] Theodore E. Harris, *The Theory of Branching Processes*, Springer-Verlag, Berlin, Heidelberg, 1963.

[Sen67] E Seneta, On the transient behavior of a Poisson branching process, *Journal of the Australian Mathematical Society*, 7:465–480, 1967.

[SUW10] Takeya Shigezumi, Yushi Uno, and Osamu Watanabe, A new model for a scale-free hierarchical structure of isolated cliques, in *Proc. of Workshop on Algorithms and Computation* (WALCOM'10), LNCS, 2010, to appear.

Appendix

We here derive the lth iteration of $U(z)$. Let us recall our definition of $U(z)$, that is,

$$U(z) = 1 - p_k + \frac{p_k z}{k-(k-1)z} = \frac{(k-1-\mu_0)z-(k-\mu_0)}{(k-1)z-k}.$$

Also recall that its lth iteration $U_l(z)$ is defined inductivey by $U_l(z) = U(U_{l-1}(z))$ for $l > 1$ and $U_1(z) = U(z)$.

To derive $U_l(z)$, we use a linear function $L(z)=az+b$ and $f(z)=L^{-1}(U(L(z)))$. Due to the following lemma, for evaluating $U_l(z)$, it suffices to get good a and b such that $f_l(z)$ is easily calculated.

Lemma 4.

$$U_l(z) = L(f_l(L^{-1}(z))).$$

Proof. By $f(z) = L^{-1}(U(L(z)))$, we have $U(z) = L(f(L^{-1}(z)))$. Then we prove the lemma by induction. We already have it for $l = 1$. Let us assume that $U_l(z) = L(f_l(L^{-1}(z)))$. Then we have

$$\begin{aligned} U_{l+1}(z) &= U(U_l(z)) = L(f(L^{-1}(U_l(z)))) \\ &= L(f(L^{-1}(L(f_l(L^{-1}(z)))))) = L(f(f_l(L^{-1}(z)))) \\ &= L(f_{l+1}(L^{-1}(z))). \end{aligned}$$

□

Let $a = \frac{1-\mu_0}{k-1}$ and $b = 1$; then we have

$$\begin{aligned} f(z) &= L^{-1}(U(L(z))) = \frac{1}{a}(U(az+1)-1) \\ &= \frac{a(k-1-\mu_0)z + (k-1-\mu_0) - (k-\mu_0) - \{a(k-1)z + (k-1) - k\}}{a\{a(k-1)z + (k-1) - k\}} \\ &= \frac{a(k-1-\mu_0)z - 1 - a(k-1)z + 1}{a\{a(k-1)z - 1\}} \\ &= \frac{-\mu_0 z}{a(k-1)z - 1} = \frac{-\mu_0 z}{(1-\mu_0)z - 1} \qquad (\text{by } a = \tfrac{1-\mu_0}{k-1}) \\ &= \frac{z}{\left(1 - \frac{1}{\mu_0}\right)z + \frac{1}{\mu_0}}. \end{aligned}$$

Lemma 5. *Let $K = \frac{1}{\mu_0}$. Then we have*

$$f_l(z) = \frac{z}{K^l + (1-K^l)z}.$$

Proof. For $l = 1$, we have

$$f_1(z) = \frac{z}{\frac{1}{\mu_0} + \left(1 - \frac{1}{\mu_0}\right)z} = \frac{z}{K + (1-K)z},$$

and the lemma holds. For $l \geq 1$, we prove by induction as follows:

$$\begin{aligned} f_{l+1}(z) &= \frac{f_l(z)}{K + (1-K)f_l(z)} = \frac{\frac{z}{K^l + (1-K^l)z}}{K + (1-K)\frac{z}{K^l + (1-K^l)z}} \\ &= \frac{z}{K^{l+1} + (1-K^l)Kz + (1-K)z} = \frac{z}{K^{l+1} + (1-K^{l+1})z}. \end{aligned}$$

□

We now have the closed form of $f_l(z)$. That is,

$$f_l(z) = \frac{z}{K^l + (1-K^l)z} = \frac{z}{\left(\frac{1}{\mu_0}\right)^l + \left(1 - \left(\frac{1}{\mu_0}\right)^l\right)z} = \frac{\mu_0^l}{\left(\frac{1-z}{z}\right) + \mu_0^l}.$$

By using Lemma 4, we obtain the closed form of $U_l(z)$ as follows:

$$\begin{aligned}
U_l(z) &= L(f_l(L^{-1}(z))) = a\left(f_l\left(\frac{z-1}{a}\right)\right)+1 \\
&= a\frac{\mu_0^l}{\left(\frac{a+1-z}{z-1}\right)+\mu_0^l}+1 \\
&= 1+\frac{a\mu_0^l z - a\mu_0^l}{\left(\mu_0^l-1\right)z+\left(a+1-\mu_0^l\right)} \\
&= 1+\frac{\mu_0^l(1-\mu_0)(z-1)}{(k-1)\left(\mu_0^l-1\right)z+\left(k-\mu_0-(k-1)\mu_0^l\right)}.
\end{aligned}$$

Elitza Maneva
Dept. of Applied Mathematics and Analisis
Universitat de Barcelona
E-mail address: elitza.maneva@gmail.com

Takeya Shigezumi
Dept. of Industrial and Systems Eng.
Aoyama Gakuin Univ.

Yushi Uno
Graduate School of Science
Osaka Prefecture University
E-mail address: uno@mi.s.osakafu-u.ac.jp

Osamu Watanabe
Dept. of Mathematical and Computing Sci.
Tokyo Institute of Technology
E-mail address: watanabe@titech.ac.jp