Appendices to the Paper "Random Generation of Git Graphs"

Julien Courtiel Martin Pépin

A The unlabelled model

This appendix proves the main theoretical results on the unlabeled model for Git graphs. We start by proving combinatorial bounds in Section A.1 on the numbers $g_{n,k}$, from which our results follow (Sections **??** and A.2).

A.1 Estimates on the number of Git graphs

The following Lemma gives us an approximation of the coefficients $g_{n,k}$ from which we can show that $g_{n,k}$ is negligible compared to $g_{n,\lfloor n/2 \rfloor}$ when k is away from $n/2$.

Lemma 1. *For all $1 \leq k \leq n$, we have*

$$\frac{(k-1)!}{(2k-n-1)!} \leq g_{n,k} \leq \binom{n-2}{k-2} \frac{(k-1)!}{(2k-n-1)!} \qquad \text{when } n-k \leq k-1$$

$$\binom{n-k-1}{k-2}(k-1)! \leq g_{n,k} \leq \binom{n-2}{k-2}(k-1)! \qquad \text{when } n-k > k-1$$

Proof. We start by proving the lower bounds. To this end, we construct an easy-to-enumerate subclass of the Git graphs of size (n,k). Start with a size k main branch and attach a size-1 branch to as many of its k commits as possible, starting with the rightmost/most recent commits.

- If $(n-k) \leq (k-1)$, we reach a total of n commits at some point, and the number of possibilities for connecting these branches to one of the main commits on the left is

$$(k-1) \cdot (k-2) \cdots (k-(n-k)) = \frac{(k-1)!}{(2k-n-1)!}.$$

- If $(n-k) > (k-1)$, then we can attach a branch to every commit on the main branch, and there remains $n-(2k-1)$ unused commits at the end of this process. There is then $\binom{n-k-1}{k-2}$ ways of adding these remaining commits to the $(k-1)$ branches. In that case, the number of possibilities for connecting the branches is $(k-1)!$.

This gives the lower bound.

For the upper bounds, first consider all the possible ways to assign the $(n-k)$ commits that are not on the main branch to one of the $(k-1)$ commits on the main branch to which they will be attached. There are $\binom{n-2}{k-2}$ ways of doing this (this is the number of compositions of $(n-k)$ into $(k-1)$ terms). Then, for each of theses configurations, and for each feature branch, we have to choose where it forks on the main branch. In the worst case, all the non-free commits are after the free commits. Furthermore, there is at most $(n-k)$ branches. This yields the same enumeration as for the lower bound, thus

$$g_{n,k} \leq \binom{n-2}{k-2} \frac{(k-1)!}{\max(0,2k-n-1)!}$$ □

Let $g_n(u)$ be $\sum_{k=0}^{n} u^k g_{n,k}$. Using these bounds, we proceed to show that for any choice of $u > 0$, the mass is concentrated around $k = \lfloor (n+1)/2 \rfloor$ in $g_n(u)$.

An upper bound for k below the mean Let $0 < \alpha < \frac{1}{2}$. Whenever $k \leq \alpha n$, we have

$$g_{n,k} \leq \binom{n-2}{k-2}(k-1)! = \frac{(n-2)!}{(n-k)!}(k-1) \leq \frac{(n-1)!}{((1-\alpha)n)!}\alpha n.$$

And by Stirling's formula,

$$\frac{(n-1)!}{((1-\alpha)n)!}\alpha n \underset{n\to\infty}{\sim} \frac{\alpha}{\sqrt{1-\alpha}}\left(\frac{e^{-\alpha}}{(1-\alpha)^{1-\alpha}}\right)^n n^{\alpha n}.$$

Similarly, when $k = \lfloor (n+1)/2 \rfloor$, we have that

$$g_{n,k} \geq \left\lfloor \frac{n-1}{2} \right\rfloor ! \geq \left(\frac{n}{2}-1\right)! \underset{n\to\infty}{\sim} \sqrt{4\pi}(2e)^{-\frac{n}{2}} n^{n/2-1/2}.$$

These two bounds allow to control the probability that the length of the main branch of a uniform Git graph is less than αn. Indeed, for any choice of $0 < \alpha < \frac{1}{2}$ and $u > 0$, we have

$$\frac{1}{g_n(u)} \sum_{k=0}^{\lfloor \alpha n \rfloor} u^k g_{n,k} \leq \frac{\max(1, u^{\alpha n})}{u^{\lfloor (n+1)/2 \rfloor}} \frac{n g_{n,\lfloor \alpha n \rfloor}}{g_{n,\lfloor (n+1)/2 \rfloor}} = O\left(n^{3/2} \cdot \left(\frac{\max(1,u^\alpha)\sqrt{2e}}{\sqrt{u}\, e^\alpha (1-\alpha)^{1-\alpha}} \right)^n \cdot n^{-n(\frac{1}{2}-\alpha)} \right) \quad (1)$$

which tends rapidly to zero as $n \to \infty$.

An upper bound for k above the mean By Lemma 1, for $k \geq \frac{n+1}{2}$ we have that

$$g_{n,k} \leq \frac{(n-2)!(k-1)}{(n-k)!(2k-n-1)!} =: h_{n,k}$$

where the quantity $h_{n,k}$ on the right is unimodal as soon as $n \geq 3$. Indeed, we observe that

$$\frac{h_{n,k+1}}{h_{n,k}} = \frac{(n-k)k}{(2k-n)(2k-n+1)(k-1)} = \frac{\left(\frac{n}{2}-t\right)\left(\frac{n}{2}+t\right)}{2t(2t+1)\left(\frac{n}{2}+t-1\right)} \qquad \text{where } t = k - \frac{n}{2} \in \left[\frac{1}{2}; \frac{n}{2}\right],$$

and we can show that this ratio in decreasing in t. Furthermore, this ratio evaluates to $\frac{n+1}{4} \geq 1$ at $1/2$ and to 0 at $n/2$. There is thus a unique value t_n of t such that this ratio is 1, and this value satisfies $t_n \sim \sqrt{\frac{n}{8}}$. As a consequence of this observation, we have that for any given $\alpha > 1/2$ there exists an $n_0(\alpha)$ such that for all $n \geq n_0(\alpha)$ and for all $k \geq \alpha n$, we have

$$\begin{aligned} g_{n,k} &\leq \frac{(n-2)!(k-1)}{(n-k)!(2k-n-1)!} \\ &\leq \frac{(n-2)!(\alpha n-1)}{((1-\alpha)n)!((2\alpha-1)n-1)!} \underset{n\to\infty}{\sim} \alpha\sqrt{\frac{2\alpha-1}{2\pi(1-\alpha)}} \cdot \left(\frac{e^{\alpha-1}}{(1-\alpha)^{1-\alpha}(2\alpha-1)^{2\alpha-1}}\right)^n \cdot n^{(1-\alpha)n-1/2} \end{aligned}$$

where the equivalent is obtained using Stirling's formula. Similarly as before, we conclude using this bound that the probability that a uniform Git graph has more than αn commits on main (with $\alpha > 1/2$ and $u > 0$) tends rapidly to zero as $n \to \infty$:

$$\frac{1}{g_n(u)} \sum_{\alpha n \leq k \leq n} u^k g_{n,k} = O\left(n \cdot \left(\frac{\max(u, u^\alpha)\, e^{\alpha-1}\sqrt{2e}}{\sqrt{u}\,(1-\alpha)^{1-\alpha}(2\alpha-1)^{2\alpha-1}} \right)^n \cdot n^{-n(\alpha-\frac{1}{2})} \right). \quad (2)$$

By Equations (1) and (2), we have

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\left| \frac{k(\gamma)}{n} - \frac{1}{2} \right| \geq \varepsilon \right) \to 0,$$

which means that $\frac{k(\gamma)}{n}$ converges in probability to $\frac{1}{2}$.

A.2 The special case $k = O(\sqrt{n})$

Lemma 1 also gives an estimate of $g_{n,k}$ up to a constant factor in the particular case when $k = O(\sqrt{t})$. Let $h_{n,k}$ denote the sequence $\binom{n-2}{k-2}(k-1)!$ counting the generalization $\mathcal{H}$ of Git graphs, discussed in Section 2.3 of the paper, where branches are allowed to have zero commits. By Lemma 1, when $k < \frac{n+1}{2}$, we have that

$$1 \geq \frac{g_{n,k}}{h_{n,k}} \geq \binom{n-k-1}{k-2}\binom{n-2}{k-2}^{-1} = \prod_{j=0}^{k-3}\left(1 - \frac{k-1}{n-2-j}\right).$$

Furthermore, when $k \leq t\sqrt{n}$ for some constant $t > 0$ independent of n, we can split this product as follows

$$\prod_{j=0}^{k-3}\left(1 - \frac{k-1}{n-2-j}\right) = \exp\left(-\sum_{j=0}^{k-3}\left(\frac{k-1}{n-2-j} + O\left(\frac{(k-1)^2}{(n-2-j)^2}\right)\right)\right)$$

where the big O under the summation is uniform in j. Consequently,

$$\begin{aligned}\prod_{j=0}^{k-3}\left(1 - \frac{k-1}{n-2-j}\right) &= \exp\left(-(k-1)(H_{n-2} - H_{n-k}) + O\left(\frac{k^3}{n^2}\right)\right)\\ &= \exp\left(-\frac{k(k-1)}{n} + O\left(n^{-1/2}\right)\right)\\ &= e^{-\frac{k(k-1)}{n}} \cdot \left(1 + O\left(n^{-1/2}\right)\right)\end{aligned}$$

where $(H_n)_{n\geq 0}$ denotes the harmonic series. The fact that $e^{-\frac{k(k-1)}{n}} \geq e^{-t^2}$ suffices to conclude that $g_{n,k}$ and $h_{n,k}$ are of the same order of magnitude. This result can be rephrased as follows: the probability that a uniform $\mathcal{H}$ structure is a Git graph is lower-bounded by a uniform constant as $n \to \infty$ in the domain $k \leq t\sqrt{n}$. This implies directly that the rejection-based algorithm from Section 2.3 of the paper terminates after a constant number of rejections.

B Labeled model

B.1 Asymptotic analysis

Recall that the generating function $\tilde{G}(z,u)$ of Git graphs (exponential in u, ordinary in z) is given by

$$\tilde{G}(z,u) \coloneqq \sum_{0\leq k\leq n} g_{n,k}\frac{u^k}{k!}z^n = \exp\left(\frac{1-z}{z}\ln\frac{1}{1-\frac{uz^2}{1-z}}\right) \tag{3}$$

We first study its domain of analyticity in order to show that it is amenable to singularity analysis. We then apply the classical techniques of analytic combinatorics to estimate the mean and variance of the number of commits on the main branch of Git graph under the Boltzmann model.

B.1.1 Domain of analyticity

For any fixed value of $u \in \mathbb{R}_+$, the expression given in equation (3) is well-defined for all complex numbers in $\mathbb{C} \setminus F$ where $F = \left\{z \in \mathbb{C} \mid 1 - \frac{z^2 u}{1-z} \in \mathbb{R}_-\right\}$. Within the set F of forbidden values for z, maybe a countable number of points are actually valid values of z for expression (3) because $\frac{1-z}{z} \in \mathbb{Z}$. Since there is only a countable number of such points, the domain of analyticity of $\tilde{G}$ (for a fixed value of u) is $\mathbb{C} \setminus F$. (Also note that the formula extends naturally at $z = 0$ to $\tilde{G}(0,u) = 1$).

For any $x \in \mathbb{R}_+$, we have that $1 - \frac{z^2 u}{1-z} = -x$ if and only if $z = \rho(u,x)$ or $z = \bar{\rho}(u,x)$ where

$$\rho(u,x) = \frac{1+x}{2u}\left(\sqrt{1+\frac{4u}{1+x}} - 1\right), \text{and}$$
$$\bar{\rho}(u,x) = -\frac{1+x}{2u}\left(\sqrt{1+\frac{4u}{1+x}} + 1\right).$$

Moreover, ρ and $\bar{\rho}$ are respectively increasing and decreasing in x so that $\rho(u,\mathbb{R}_+) = [\rho(u);1[$ and $\bar{\rho}(u,\mathbb{R}_+) =]-\infty;\bar{\rho}(u)]$ where

$$\rho(u) = \rho(u,0) = \frac{\sqrt{1+4u}-1}{2u}, \text{and} \tag{4}$$
$$\bar{\rho}(u) = \bar{\rho}(u,0) = -\frac{\sqrt{1+4u}+1}{2u} = -\frac{1}{u\rho(u)}. \tag{5}$$

Finally, we remark that $\frac{-\bar{\rho}}{\rho} = 1 - \bar{\rho} > 1$ so that ρ is the dominant singularity (as expected). The domain of analyticity of $\tilde{G}$ is pictured in Figure 1.

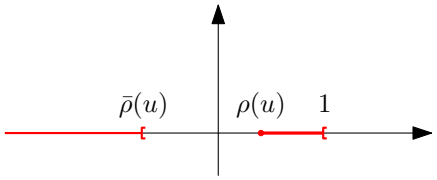

Figure 1: Domain of analyticity of $\tilde{G}(z,u)$ when u is a fixed non-negative real number

B.1.2 Singularity analysis

Per the previous section, $\tilde{G}$ is analytic on a delta domain and has only one dominant singularity. Moreover, near $\rho(u)$, we have the following estimate for $\tilde{G}$:

$$\tilde{G}(z,u) = \left(\frac{C}{1-\frac{z}{\rho}}\right)^{\frac{1-\rho}{\rho}} \cdot \left[1 + \frac{1}{\rho}\cdot\left(1-\frac{z}{\rho}\right)\left(\ln\frac{1}{1-\frac{z}{\rho}} + \mu\right) + O\left(\left(1-\frac{z}{\rho}\right)^2 \ln\left(1-\frac{z}{\rho}\right)^2\right)\right]$$

where

$$C = \frac{1-\rho}{2-\rho} \quad \text{and} \quad \mu = \ln C + \frac{1}{2-\rho}$$

As a consequence, the transfer theorem from **[FS2009]** applies and we have that

$$[z^n]\tilde{G}(z,u) \underset{n\to\infty}{=} \frac{C^{\frac{1-\rho}{\rho}}}{\Gamma\left(\frac{1-\rho}{\rho}\right)} n^{\frac{1-2\rho}{\rho}} \rho^{-n} \cdot \left[1 + \frac{1-2\rho}{\rho^2 n}\left(\ln n + \frac{1-\rho}{2} + \mu - \psi\left(\frac{1-2\rho}{\rho}\right)\right) + O\left(\frac{(\ln n)^2}{n^2}\right)\right] \tag{6}$$

where ψ denotes the digamma function, that is the logarithmic derivative of Euler's gamma function. We can obtain a similar estimate for the coefficients of $\partial_2\tilde{G}$ by observing that:

$$u\partial_2\tilde{G}(z,u) = \frac{uz}{1-\frac{uz^2}{1-z}}\tilde{G}(z,u) = \frac{z(1-z)}{(\rho-z)(z-\bar{\rho})}\tilde{G}(z,u).$$

It follows that

$$\frac{u\partial_2\tilde{G}(z,u)}{\tilde{G}(z,u)} \underset{z\to\rho}{=} \frac{(1-\rho)^2}{\rho(2-\rho)}\left(1-\frac{z}{\rho}\right)^{-1}\left[1 - \frac{1-3\rho+\rho^2}{(1-\rho)(2-\rho)}\left(1-\frac{z}{\rho}\right) + O\left(\left(1-\frac{z}{\rho}\right)^2\right)\right]$$

and thus

$$u\partial_2\tilde{G}(z,u) \underset{z\to\rho}{=} \frac{1-\rho}{\rho}\cdot\left(\frac{C}{1-\frac{z}{\rho}}\right)^{\frac{1}{\rho}} \cdot\left[1+\frac{1-\frac{z}{\rho}}{\rho}\left(\ln\frac{1}{1-\frac{z}{\rho}}+\mu-\frac{\rho(1-3\rho+\rho^2)}{(1-\rho)(2-\rho)}\right)+O\left(\left(1-\frac{z}{\rho}\right)^2\ln\left(1-\frac{z}{\rho}\right)^2\right)\right].$$

Using the transfer theorem, it follows that

$$[z^n]u\partial_2\tilde{G}(z,u) \underset{n\to\infty}{=} \frac{1-\rho}{\rho}\frac{C^{1/\rho}}{\Gamma(1/\rho)}n^{\frac{1}{\rho}-1}\rho^{-n} \cdot\left[1+\frac{1-\rho}{\rho^2 n}\left(\ln n-\psi\left(\frac{1}{\rho}-1\right)+\mu+\frac{1}{2}-\frac{\rho(1-3\rho+\rho^2)}{(1-\rho)(2-\rho)}\right)+O\left(\frac{(\ln n)^2}{n^2}\right)\right]. \tag{7}$$

For the sake of computing the variance of the number of commits on the main branch of a Git graph under the labeled-main distribution, we also analyze the second derivative. For any function $f = f(z,u)$, let $f^\bullet$ denote the function $u\partial_2 f(z,u)$. Then we have that

$$\tilde{G}^{\bullet\bullet}(z,u) \underset{z\to\rho}{=} \frac{1-\rho}{\rho^2}\cdot\left(\frac{C}{1-\frac{z}{\rho}}\right)^{\frac{1}{\rho}+1} \cdot\left[1+\frac{1-\frac{z}{\rho}}{\rho}\left(\ln\frac{1}{1-\frac{z}{\rho}}+\mu+\frac{\rho(\rho^3-5\rho^2+8\rho-2)}{(1-\rho)(2-\rho)}\right)+O\left(\left(1-\frac{z}{\rho}\right)^2\ln\left(1-\frac{z}{\rho}\right)^2\right)\right].$$

Applying the transfer theorem to this expansion yields

$$[z^n]\tilde{G}^{\bullet\bullet}(z,u) \underset{n\to\infty}{=} \frac{1-\rho}{\rho^2}\frac{C^{\frac{1}{\rho}+1}}{\Gamma\left(\frac{1}{\rho}+1\right)}n^{\frac{1}{\rho}}\rho^{-n} \cdot\left[1+\frac{\rho^{-2}}{n}\left(\ln n-\psi\left(\frac{1}{\rho}\right)+\mu+\frac{\rho+1}{2}+\frac{\rho(\rho^3-5\rho^2+8\rho-2)}{(1-\rho)(2-\rho)}\right)+O\left(\frac{(\ln n)^2}{n^2}\right)\right]. \tag{8}$$

B.1.3 First moments of the labeled-main distribution

The above estimates allow us to approximate the expected number of commits on the main branch of a git graph of size n sampled according to

$$\mathbb{P}(\gamma) = \frac{u^{k(\gamma)}}{k(\gamma)!\tilde{G}_n(u)}$$

where $k(\gamma)$ is the number of commits on the main branch of γ and where $\tilde{G}_n(u)$ denotes $[z^n]\tilde{G}(z,u)$.

Mean We have that, if γ is sampled according to this distribution, we can compute its main using from Equations (6) and (7):

$$\mathbb{E}(k(\gamma)) = \frac{u\partial_2\tilde{G}_n(u)}{\tilde{G}_n(u)} = \frac{[z^n]u\partial_2\tilde{G}(z,u)}{[z^n]\tilde{G}(z,u)} \underset{n\to\infty}{=} Cn\cdot\left[1+\rho^{-1}\frac{\ln n}{n}+\frac{1}{n}\left(\frac{\ln C-\psi\left(\frac{1-\rho}{\rho}\right)}{\rho}+\frac{1}{2-\rho}\right)+O\left(\frac{(\ln n)^2}{n^2}\right)\right].$$

Note that the ratio $C=\frac{1-\rho}{2-\rho}$ can take any value in the open interval $(0;\frac{1}{2})$ depending on the value of u. This implies that one can "tune" the value of u in order to target graphs with a given $k(\gamma)/n$ ratio.

Variance Similarly, we also have the variance of $k(\gamma)$ from Equations (6) and (8):

$$\begin{aligned}\mathbb{V}(k(\gamma)) = \mathbb{E}(k(\gamma)^2) - \mathbb{E}(k(\gamma))^2 &= \frac{[z^n]\tilde{G}^{\bullet\bullet}(z,u)}{[z^n]\tilde{G}(z,u)} - \left(\frac{[z^n]\tilde{G}^{\bullet}(z,u)}{[z^n]\tilde{G}(z,u)}\right)^2 \\ &= C^2 n \frac{\rho}{(1-\rho)(2-\rho)} = n \cdot \frac{\rho(1-\rho)}{(2-\rho)^3}.\end{aligned}$$

This means that $k(\gamma)$ is concentrated around its mean in a window of width about $\sqrt{n}$.