ANALYZING PMPQS
(INFORMAL NOTE)

FRANÇOIS MORAIN

Abstract. We use generating functions to analyze PMPQS.

1. Birthday paradox

Let $U = \{u_1, \ldots, u_m\}$ be an urn containing m balls. We suppose we draw n balls from it and that the probability we get u_i is π_i for some real π_i in $(0,1)$. Let $\mathcal{D}_n$ be the number of distinct balls and $\mathcal{M}_n$ be the number of matches. We have trivially

$$\mathcal{D}_n + \mathcal{M}_n = n.$$

We are looking for the average number of matches $\overline{\mathcal{M}}_n$. It is related to the average number of distinct balls, $\overline{\mathcal{D}}_n$, by

$$\overline{\mathcal{D}}_n + \overline{\mathcal{M}}_n = n.$$

Along the lines of [1], we can compute the exponential generating function of $\mathcal{D}_n$.

Theorem 1.1. *Let*

$$\Phi_m(z,u) = \prod_{j=1}^{m} (1 + u(e^{\pi_j z} - 1)).$$

Then

$$\overline{\mathcal{D}}_n = n![z^n] \left.\frac{\partial}{\partial u}\Phi_m(z,u)\right|_{u=1}.$$

We deduce

Corollary 1.1.

$$\overline{\mathcal{D}}_n = \sum_{j=1}^{m} (1 - (1-\pi_j)^n).$$

Proof: Differentiating $\Phi_m(z,u)$, one gets

$$\frac{\partial}{\partial u}\Phi_m(z,u) = \sum_{j=1}^{m} (e^{\pi_j z} - 1) \prod_{k \neq j} (1 + u(e^{\pi_k z} - 1))$$

and putting $u = 1$ in this formula, we obtain

$$\left.\frac{\partial}{\partial u}\Phi_m(z,u)\right|_{u=1} = \sum_{j=1}^{m} (e^{\pi_j z} - 1) \prod_{k \neq j} e^{\pi_k z} = \sum_{j=1}^{m} (e^z - e^{(1-\pi_j)z})$$

and the result follows. □

Date: January 1993.

On leave from the French Department of Defense, Délégation Générale pour l'Armement.

When n is small compared to m, we have

$$\sum_{j=1}^{m}(1-\pi_j)^n = m - n + \frac{n(n-1)}{2}S_2 + R$$

where

$$S_k = \sum_{j=1}^{m}\pi_j^k$$

for any k and the error term, R is negligible. We deduce from this:

Proposition 1.1.

$$\overline{\mathcal{M}}_n \approx \frac{n^2}{2}S_2.$$

2. Application to integer factorization

2.1. **PMPQS.** Suppose one wants for factor the integer N using the MPQS factorisation algorithm (cf. e.g., [2]). The heart of the algorithm is the collection of relations of the type:

$$(R_i) : u_i^2 \equiv q_i \prod_{r=1}^{k} p_r^{\alpha(r,i)} \bmod N$$

where u_i is an integer, p_j are the primes contained in the factor base $\mathcal{B} = \{p_1, p_2, \ldots, p_k\}$ and q_i belongs to $\mathcal{B}' = \{q_1, \ldots, q_m\}$, the so-called extra factor base. One can use this type of factorization if and only if one can find two relations (R_i) and (R_j) with $q_i = q_j$. In that case, one keeps the relation:

$$(u_i u_j / q_i)^2 \equiv \prod_{r=1}^{k} p_r^{\alpha(r,i)+\alpha_(r,j)} \bmod N.$$

2.1.1. *The model.* Suppose we have collected n relations of type (R_i). Our aim is to compute the estimated number of matches obtained from these relations. We use the following assumptions. We suppose that the prime q_i appears with probability π_i proportionnal to $1/q_i$. Letting

$$\Sigma_1 = \sum_{j=1}^{m}\frac{1}{q_j},$$

one gets $\pi_i = \Sigma_1/q_i$. Using the preceding section, we can now estimate the expected number of matches. It is

$$n^2\frac{1}{2}\sum_{j=1}^{m}\pi_j^2 = n^2\frac{\Sigma_1^2}{2}\sum_{j=1}^{m}\frac{1}{q_j^2}.$$

2.1.2. *Experimentations.* Let us consider the factorisation of the 66-digit cofactor of $z3c94 = \lceil 10^{94}\zeta(3)\rceil = 2 \times 59 \times 137 \times 241 \times 1913 \times 130393469 \times 25334347691 \times C_{66}$. We choose a multiplier of value 83 and take 4000 primes in the factor base. We let $P_M = 2000000$ (in the program `maxbig=2000000`). We get 70463 extra primes in the extra factor base. The exact value of S_2 is 0.0000312122.

(Note that

$$C_{66} = 140727812743537072348637234 3 \times 346932460355562829295892187566038362309.)$$

#par	observed	approx
241	0	0
525	2	4
836	11	10
1120	19	19
1433	28	32
1747	32	47
2043	43	65
2393	65	89
2678	82	111
3007	105	141
3316	127	171
3637	152	206
3969	181	245
4299	214	288
4631	250	334
4981	290	387
5337	336	444
5712	364	509
6049	409	571
6387	451	636
6730	503	706
7098	567	786
7460	631	868
7815	705	953
7815	705	953
8207	793	1051
8580	875	1148
8963	952	1253
9356	1042	1366
9777	1159	1491
10180	1227	1617
10503	1333	1721
10879	1449	1847
11259	1575	1978
11620	1678	2107
11978	1803	2239
12386	1941	2394
12744	2049	2534
13120	2183	2686
13561	2341	2869

2.1.3. *Heuristics.* We assume that q_1 (resp. q_k) is the K-th (resp. M-th) prime number, noted P_K (resp. P_M). We want to use the large prime variation so as to include all primes between P_K and P_M: This yields m prime numbers (do not forget that about half primes will be convenient for our purpose).

In that case, we suppose that we consider all primes between P_K and P_M and that each prime P_j occurs with probability $1/(2P_j)$. Next, we assume that $P_j \approx j \log j$. Then

$$\Sigma_1 \approx \sum_{j=K+1}^{M} \frac{1}{2j \log j} \approx 1/2 \int_{K+1}^{M} \frac{dx}{x \log x} = 1/2[\log \log x]_{K+1}^{M}.$$

We will show

$$S_2 \approx \frac{1}{\Sigma_1^2}[\frac{1}{x(\log x)^2}]_{K+1}^{M}.$$

References

[1] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms, and self–organizing search. *Discrete Applied Mathematics* (1991). To appear. Available as INRIA Research Report 720, August 1987, 17 pages.

[2] A. K. Lenstra and M. S. Manasse. Factoring by electronic mail. In *Advances in Cryptology* (1990), J.-J. Quisquater, Ed., vol. 434 of *Lect. Notes in Computer Science*, Springer-Verlag, pp. 355–371. Proc. Eurocrypt '89, Houthalen, April 10–13.

(F. Morain) Laboratoire d'Informatique de l'École polytechnique (LIX), F-91128 Palaiseau Cedex, France

E-mail address, F. Morain: `morain@polytechnique.fr`