Ann. Inst. Statist. Math.
Vol. 45, No. 3, 459–465 (1993)

A RANDOM CLUSTERING PROCESS

MASAAKI SIBUYA

Department of Mathematics, Keio University, Hiyoshi, Yokohama 223, Japan

(Received December 26, 1991; revised August 25, 1992)

Abstract. A clustering process which generates simple and uniform random partitions is studied. It has a single parameter and generates, for a special value of the parameter, the partition of a random permutation into its cycles. The limit distribution of the size index of the generated partition is the joint of the independent Poisson distributions with means determined by the size and the parameter.

Key words and phrases: Bell polynomial, cycles of random permutation, Poisson approximation, Pólya urn model, Stirling number of the first kind.

1. Random clustering process

Suppose the balls labeled $1, 2, \dots$ (one ball for each number) are thrown at random into an infinite number of indistinguishable urns as follows. Ball 1 is put into an urn. Ball 2 is put into an empty urn with probability p_1 and into the urn with Ball 1 with probability $1 - p_1$. Ball 3 is put into an empty urn with probability p_2 whichever Ball 2 is thrown in. If Balls 1 and 2 are in the same urn, Ball 3 is put into it with probability $1 - p_2$, otherwise Ball 3 is put into the urns with Ball 1 or 2 with equal probability $(1 - p_2)/2$.

Let the pattern $1|2|3$ denote that Balls 1, 2 and 3 are in separate urns, let $12|3$ denote that Balls 1 and 2 are in the same urn but Ball 3 is in a different one, and so on. If $p_2 = p_1/(2 - p_1)$, the patterns $12|3$ (or $23|1$) and $13|2$ are equally probable, and the probabilities of the patterns $1|2|3$, $12|3$ and 123 are

$$p_1^2/(2 - p_1), \qquad p_1(1 - p_1)/(2 - p_1) \qquad \text{and} \qquad 2(1 - p_1)^2/(2 - p_1),$$

respectively.

We generalize the steps for Balls 2 and 3 to define a random clustering process. A cluster means a set of balls thrown into the same urn, and the patterns of clusters are our concern. Mathematically, the clusters at the n-th step form a partition of the finite set $\mathcal{U}_n = \{1, 2, \dots, n\}$, i.e. the set of labeling numbers on balls up to Ball n. The family of all partitions of $\mathcal{U}_n$ is denoted by $\mathcal{A}_n$. If a partition $A \in \mathcal{A}_n$ has s_j subsets of cardinality j (i.e. s_j clusters of size j or s_j urns with j balls),

$j = 1, \dots, n$, this condition is denoted by

$$(1.1) \qquad S(A) = s \in \mathcal{S}_n, \qquad \mathcal{S}_n = \left\{ s = (s_1, \dots, s_n);\ s_j \geq 0, \sum_{j=1}^{n} j s_j = n \right\},$$

and s will be called the 'size index' of A. Further define

$$(1.2) \qquad p_k = \frac{\rho}{k - (k-1)\rho} = \frac{\alpha}{k + \alpha}, \qquad 0 < \rho < 1,\ \ 0 < \alpha < \infty,\ \ k = 1, 2, \dots,$$

where $\alpha = \rho/(1 - \rho)$.

Now a 'random clustering process' is defined as follows. If Ball 1, ..., Ball k are thrown to form a partition $A \in \mathcal{A}_k$, then Ball $k+1$ is put into an empty urn with probability p_k (1.2) and into an urn with j balls with probability $(1 - p_k) j/k$. This ball throwing is continued for $k = 2, 3, \dots$

PROPOSITION 1.1. *At the n-th step of the random clustering process mentioned above, the probability that Ball* $1, \dots,$ *Ball n form a partition* $A \in \mathcal{A}_n$ *is*

$$(1.3) \qquad P(A; \mathcal{A}_n) = f_n(s) = f_n(s; \rho) = \frac{\rho^{u-1}(1-\rho)^{n-u}}{\prod_{i=2}^{n-1}(i - (i-1)\rho)} \prod_{j=1}^{n} ((j-1)!)^{s_j}$$

$$= \frac{\alpha^u}{\alpha^{[n]}} \prod_{j=1}^{n} ((j-1)!)^{s_j}, \qquad 0 < \rho < 1,\ \ 0 < \alpha < \infty,$$

where $s = S(A) \in \mathcal{S}_n$, $u = \sum_{j=1}^{n} s_j$, *and* $\alpha^{[n]} = \alpha(\alpha+1)\cdots(\alpha+n-1)$.

PROOF. The probability (1.3) is shown by induction on n using the recurrence

$$(1.4) \quad f_{n+1}(s_1, \dots, s_{n+1}) = f_n(s_1 - 1, s_2, \dots, s_n) \cdot \frac{\alpha}{n+\alpha} \cdot 1[n+1 \in \{n+1\}]$$
$$+ \sum_{j=1}^{n} f_n(s_1, \dots, s_j + 1, s_{j+1} - 1, \dots, s_n)$$
$$\cdot \frac{j}{n+\alpha} \cdot 1[n+1 \in C \subset \mathcal{U}_n, |C| = j],$$

where $1[\cdot]$ is the indicator function of the bracketed event in $\mathcal{A}_{n+1}$. □

An important feature of (1.3) is that it is determined by the size index $S(A)$ and independent of the elements of the subsets of A. That is, $P(\cdot; \mathcal{A}_k)$ is independent of the order of n balls thrown in, and is invariant with respect to the permutation of the indices of the balls. Another feature is the Markovian property implied by (1.4) and typically shown by the following example (4). These facts are useful in calculating the probability of an event in the sample space $\mathcal{A}_n$ as the following examples.

(1) Any pair of elements of $\mathcal{U}_n$ is in different subsets or in the same subset with the probabilities $P(1|2;\mathcal{A}_2) = p_1$ and $P_2(12;\mathcal{A}_2) = 1 - p_1$, respectively.

(2) Let a subset $C \subset \mathcal{U}_n$ of cardinality $k > 0$ be given. The probability of the event that C is included in a subset of $A \in \mathcal{A}_n$ is equal to $P(1\cdots k;\mathcal{A}_k) = \prod_{j=1}^{k-1}(1-p_j)$.

(3) Let a subset $\{i, j_1, \ldots, j_k\} \subset \mathcal{U}_n$ be given. The probability of the event in $\mathcal{A}_n$ that i is not in a subset (of $A \in \mathcal{A}_n$) which includes any of $j_1, \ldots, j_k$ is equal to the probability that an element, say $k+1$, is a singleton in $A \in \mathcal{A}_{k+1}$, $\sum_{B \in \mathcal{A}_k} P(B \cup \{k+1\}; \mathcal{A}_{k+1}) = p_k$.

(4) Under the condition that $1 \in C \subset \mathcal{U}_n$, $|C| = k$, the random partition of $\mathcal{U}_n \backslash C$ follows $P(\cdot;\mathcal{A}_{n-k})$ provided that the elements of $\mathcal{U}_n \backslash C$ are relabeled.

(5) Let $A = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$ and $|A_1| = k$ $(0 < k < n)$. A random partition of A disregarding the elements of A_2, or a random subpartition of A_1 under the condition that a subpartition of A_2 is given, has the distribution $P(\cdot;\mathcal{A}_k)$ provided that the elements of A_1 are relabeled.

If $\rho = 1/2$, that is $\alpha = 1$ or $p_k = 1/(k+1)$, $f_n(s;1/2) = (n!)^{-1}\prod_{j=1}^{n}((j-1)!)^{s_j}$. This is the probability distribution of the random partition generated by cycles of a random permutation, and appears in many applications, Sibuya (1993). This fact suggests another way to generate (1.3), Yamato (private communication). Let $\mathcal{P}_n$ denote the symmetric group of all permutations of $\mathcal{U}_n$, and suppose that an element $\pi \in \mathcal{P}_n$ is chosen with the probability

$$P(\pi) = \alpha^u/\alpha^{[n]}, \qquad 0 < \alpha < \infty, \tag{1.5}$$

where u is the number of cycles of π. Then the partition $A(\pi) \in \mathcal{A}_n$ generated by cycles of π has the probability (1.3).

The random clustering process is almost the same to a Pólya urn model with balls of a continuum of colors, which was used by Blackwell and MacQueen (1973) to obtain Ferguson's Dirichlet process as $n \to \infty$. The distribution of colors at a finite step is studied by Yamato (1992) which is closely related to the present paper. In the population genetics, the model is known as Hoppe's urn model, see Hoppe (1984) and Ewens (1990).

2. Size index and number of clusters

If the balls are indistinguishable, only the size index of a random partition is observable. Or one may be interested just in the size index disregarding their contents.

PROPOSITION 2.1. *Let $S = (S_1, \ldots, S_n)$ be the size index of the partition at the n-th step of the random clustering process of Section* 1. *Then the probability that $S = s \in \mathcal{S}_n$ is*

$$\begin{aligned} g_n(s) = g_n(s;\rho) &= \frac{n!\rho^{u-1}(1-\rho)^{n-u}}{\prod_{j=1}^{n} j^{s_j} s_j! \prod_{i=2}^{n-1}(i-(i-1)\rho)} \\ &= \frac{n!\alpha^u}{\alpha^{[n]}\prod_{j=1}^{n} j^{s_j} s_j!}, \qquad 0 < \rho < 1, \; 0 < \alpha < \infty, \end{aligned} \tag{2.1}$$

where $u = \sum_{j=1}^{n} s_j$.

PROOF. The number of permutations of $\mathcal{U}_n$ having a size index s is $n!/\prod_{j=1}^{n}(j!)^{s_j}s_j!$. Multiplying this number and $f_n(s)$ given in (1.3) we obtain $g_n(s)$ because of the invariance of $f_n(s)$ with respect to the indexing of balls. □

The fuction $g_n(s)$ satisfies the recurrence relation,

$$\begin{aligned}(\alpha+n)g_{n+1}(s) &= \alpha g_n(s_1-1, s_2, \dots, s_n)\\ &\quad + \sum_{j=1}^{n} j(s_j+1)g_n(s_1, \dots, s_j+1, s_{j+1}-1, \dots, s_n),\end{aligned}$$

which is essentially the same as that in the proof of Proposition 1.1. In population genetics theory (2.1) is called the Ewens' sampling formula. See e.g. Hoppe (1984) and Ewens (1990).

The joint factorial moment of S is as follows.

$$M(r_1, \dots, r_n) = E\left(\prod_{j=1}^{n} S_j^{(r_j)}\right) = \frac{\alpha^r n^{(R)}}{(\prod_{j=1}^{n} j^{r_j})(\alpha+n-1)^{(R)}},$$
$$r_j = 0, 1, 2, \dots; \ j = 1, \dots, n,$$

where $n^{(R)} = n(n-1)\cdots(n-R+1)$, $r = \sum_{j=1}^{n} r_j$ and $R = \sum_{j=1}^{n} j^{r_j}$. Especially,

$$E(S_j) = \alpha n^{(j)}/j(\alpha+n-1)^{(j)}$$

and

$$\mathrm{Cov}(S_i, S_j) = \frac{\alpha^2}{ij}\left(\frac{n^{(i+j)}}{(\alpha+n-1)^{(i+j)}} - \frac{n^{(i)}n^{(j)}}{(\alpha+n-1)^{(i)}(\alpha+n-1)^{(j)}}\right), \quad i \leq j.$$

PROPOSITION 2.2. *Let m be a fixed positive integer, $1 \leq m < \infty$. The first m components $(S_1, \dots, S_m)$ of S in Proposition 2 converges as $n \to \infty$ to the joint distribution of independent Poisson with means $(\alpha, \alpha/2, \dots, \alpha/m)$.*

PROOF. The components have the joint factorial moment

$$M(r_1, \dots, r_m) = \left(\prod_{j=1}^{m}\left(\frac{\alpha}{j}\right)^{r_j}\right)\left(1 - \frac{R(\alpha-1)}{n} + O\left(\frac{1}{n^2}\right)\right),$$

where $R = \sum_{j=1}^{m} jr_j$. Since all the factorial moments converges to those of the independent Poisson distribution with means $(\alpha, \alpha/2, \dots, \alpha/m)$ the probability function (2.1) also converges to the Poisson. See, e.g., Bollobàs (1985) Theorem 21. □

The probability generating function of $S = (S_1, \dots, S_n)$ is

$$G_n(w_1, \dots, w_n) = \frac{1}{\alpha^{[n]}} Y_n(0!\alpha w_1, 1!\alpha w_2, \dots, (n-1)!\alpha w_n), \tag{2.2}$$

where Y_n is the exponential complete Bell polynomial of n variables defined by

$$Y_n(x_1,\ldots,x_n)=\sum_{s\in\mathcal{S}_n}\frac{n!}{\prod_{j=1}^n s_j!}\prod_{j=1}^n\left(\frac{x_j}{j!}\right)^{s_j}.$$

It has an exponential extended generating function

$$1+\sum_{n=1}^\infty Y_n(x_1,\ldots,x_n)\frac{t^n}{n!}=\exp\left(\sum_{k=1}^\infty\frac{x_k t^k}{k!}\right).$$

See Comtet (1974) for the properties of Y_n. The following alternative proof of the proposition is more complicated, but shows an interesting property of the Bell polynomial.

PROOF. (Mase, private communication) The probability generating function of the components $(S_1,\ldots,S_m)$, $m<n$, is

$$G_{mn}(w_1,\ldots,w_m)=G_n(w_1,\ldots,w_m,1,\ldots,1).$$

Define

$$H_n(w_1,\ldots,w_n)=\begin{cases}\alpha^{[n]}G_{mn}(w_1,\ldots,w_n)/n!, & \text{if } n>m,\\ Y_n(0!\alpha w_1,1!\alpha w_2,\ldots,(n-1)!\alpha w_n), & \text{if } n\le m,\end{cases}\tag{2.3}$$

and the exponential extended generating function of $(H_n)_1^\infty$ is

$$\begin{aligned}1+\sum_{n=1}^\infty H_n t^n&=\exp\left(\sum_{k=1}^m\frac{\alpha w_k}{k}t^k+\sum_{k=m+1}^\infty\frac{\alpha}{k}t^k\right)\\&=\exp\left(\sum_{k=1}^m\frac{\alpha(w_k-1)}{k}t^k\right)(1-t)^{-\alpha}.\end{aligned}\tag{2.4}$$

Expand the exponential function in the last expression as $1+\sum_{n=1}^\infty K_n(w_1,\ldots,w_m)t^n$ to obtain

$$H_n(w_1,\ldots,w_m)=\sum_{j=0}^n\frac{\alpha^{[n-j]}}{(n-j)!}K_j(w_1,\ldots,w_m).$$

Thus,

$$G_{mn}=1+\sum_{j=1}^n K_j\frac{n!\alpha^{[n-j]}}{(n-j)!\alpha^{[n]}}.\tag{2.5}$$

The coefficient $c_{n,j}$ of K_j in the series satisfies $c_{n,j}\le\max(2,1/\alpha)j$ and $c_{n,j}\to 1$ $(n\to\infty)$. Since K_j is the coefficient of the expansion of an analytic function, $\sum_j j|K_j|<\infty$. The absolute convergence shows that

$$\lim_{n\to\infty}G_{mn}=1+\sum_{j=1}^\infty K_j=\exp\left(\sum_{k=1}^m\frac{\alpha(w_k-1)}{k}\right).\tag{2.6}$$

□

The special case $g_n(s; 1/2)$ is as typical as $f_n(s; 1/2)$. An application of this case to the inelastic collisions of particles moving on a line was studied by Sibuya *et al.* (1990).

One of the interesting quantities in clustering is the number u of clusters. Propositions 1.1 and 2.2 show that u is a sufficient statistic of the distributions of $f_n(s; \rho)$ and $g_n(s; \rho)$. The number u of clusters in (1.3) or (2.1) has the following distribution function:

$$h_n(u; \rho) = h_n(u) = \begin{bmatrix} n \\ u \end{bmatrix} \frac{\rho^{u-1}(1-\rho)^{n-u}}{\prod_{i=2}^{n-1}(i-(i-1)\rho)} = \begin{bmatrix} n \\ u \end{bmatrix} \frac{\alpha^u}{\alpha^{[n]}}, \qquad u = 1, 2, \ldots, n, \quad 0 < \rho < 1, \tag{2.7}$$

where $\begin{bmatrix} n \\ u \end{bmatrix}$ is the (unsigned) Stirling number of the first kind, α and $\alpha^{[n]}$ are defined in (1.3). See, e.g., Graham *et al.* (1989) or Riordan (1968) for Stirling numbers. The probability function h_n is well known, see e.g. Johnson and Kotz (1977) and Ewens (1990) for a genetic application, and Bartholomew (1982) for a sociological application. It was discussed recently by Sibuya (1988) and Sibuya (1992).

Acknowledgements

The author thanks Prof. Hajime Yamato and Prof. Shigeru Mase for communicating the generation of (1.3) by the random permutation with (1.5) and a proof of Proposition 2.2, respectively. He also thanks the referees whose comments greatly improved the paper.

References

Bartholomew, D. J. (1982). *Stochastic Models for Social Processes*, 3rd ed., Wiley, New York.

Blackwell, D. and MacQueen, J. (1973). Ferguson distribution via Pólya urn schemes, *Ann. Statist.*, **1**, 353–355.

Bollobàs, B. (1985). *Random Graphs*, Academic Press, London.

Comtet, L. (1974). *Advanced Combinatorics*, Reidel, Dordrecht.

Ewens, W. J. (1990). Population genetics theory—the past and the future, *Mathematical and Statistical Developments of Evolutionary Theory* (ed. S. Lessard), NATO Adv. Sci. Inst. Ser. C-299, 177–227, Kluwer, Dordrecht.

Graham, R. L., Knuth, D. E. and Patashnik, O. (1989). *Concrete Mathematics*, Addison-Wesley, Reading, Massachusetts.

Hoppe, F. M. (1984). Pólya-like urns and the Ewens' sampling formula, *J. Math. Biol.*, **20**, 91–94.

Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Application*, Wiley, New York.

Riordan, J. (1968). *Combinatorial Identities*, Wiley, New York.

Sibuya, M. (1988). Log-concavity of Stirling numbers and unimodality of Stirling distributions, *Ann. Inst. Statist. Math.*, **40**, 693–714.

Sibuya, M. (1992). A cluster-number distribution and its application to the analysis of homonyms, *Japanese Journal of Applied Statistics*, **20**, 139–153 (in Japanese).

Sibuya, M. (1993). Random partition of a finite set by cycles of permutation, *Japan Journal of Industrial and Applied Mathematics*, **10**, 69–84.

Sibuya, M., Kawai, T. and Shida, K. (1990). Equipartition of particles forming clusters by inelastic collisions, *Physica A*, **167**, 676–689.

Yamato, H. (1992). A Pólya urn model with a continuum of colors, *Ann. Inst. Statist. Math.*, (to appear).

After the paper was accepted, another approach to Proposition 2.2 was published:

Arratia, R., Barbour, A. D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula, *Ann. Appl. Probab.*, **2**, 519–535.