Abstract
The Chinese restaurant process (Pitman, 1996) is a well-known sequential random construction which generates observations from the exchangeable partition of the positive integers induced by sampling from a Dirichlet process. A generalization is provided by Ishwaran and James (2003) for sampling from Pitman’s species sampling models. Here we derive a random graph representation of the exchangeable partition induced by sampling from the species sampling models. The growing random graph model is characterized in terms of attachment rules deduced by a variation of the generalized Chinese restaurant process, based on the associated sampling distribution.

1. Exchangeable random partitions
According to Kingman’s theory (1978) if \( X_1, \ldots, X_n \) is a sample from a random distribution \( F_0 \), the random partition \( \Pi_n = \{ A_1, \ldots, A_k \} \) of \([1, \ldots, n]\) induced by the exchangeable equivalence relation
\[
i \sim j \iff X_i(\omega) = X_j(\omega) \text{ for some } \omega.
\]
is an exchangeable partition of \([n]\). This means that for each \( n \), the distribution of \( \Pi_n \) is such that for each particular partition \( \{ A_1, \ldots, A_k \} \) of \([n]\) with \( |A_j| = n_j \), for \( 1 \leq j \leq k \), \n \geq 1 \text{ and } \sum n_j = n \),
\[
P(\Pi_n = \{ A_1, \ldots, A_k \}) = p(n_1, \ldots, n_k), \tag{1}
\]
is the exchangeable probability function of \([n]\), i.e., the random vector \( n \) of \([n]\) is exchangeable if its restriction \( \Pi_n \) to \([n]\) is exchangeable for every \( n \).

Different ways to encode the random sizes \( N_i = |A_i| \) of the blocks of an exchangeable partition as random compositions of \( n \) may be considered. One way is to consider the block counts vector, i.e., the random vector of non-negative integers
\[
M_i = \sum_{j \geq 1} 1(N_i = j),
\]
for \( i = 1, \ldots, n \) subject to \( \sum_{i=1}^n M_i = n \) and \( \sum_{i \neq j} M_i = k \), which counts how many blocks of size \( j \) there are in a given partition of \([n]\), for \( j = 1, \ldots, n \).
In Ishwaran and James (2003) the previous constraints are exploited to give a generalized version of the Chinese restaurant sequential construction, which provides samples from the partition structure induced by a species sampling model.

It may be shown (Cerqueti, 2005) that an alternative representation of the class of species sampling models may be obtained resorting to the following alternative partition rules, expressed in terms of the sampling formula (2):
\[
p(\Pi_n = \{ A_1, \ldots, A_k \}) = \frac{\prod_{i=1}^k (m_i + 1)}{\prod_{i=1}^k (n + 1)} \cdot \prod_{i=1}^k p(n_i), \tag{7}
\]
for \( i = 1, \ldots, n \),
\[
p(m_i, n_i) = \frac{p(n_i)}{p(n)} = \frac{p(n_i)}{p(n)} = \frac{p(n_i)}{p(n)} \tag{8}
\]
for \( 1 \leq j \leq k \), where \( p(n_i) = p(n_i, n_i + 1, \ldots, n_k) \), and
\[
p(n_i) = \frac{\prod_{i=1}^k (m_i + 1)}{\prod_{i=1}^k (n + 1)} \cdot \prod_{i=1}^k (n_i) \tag{9}
\]
for \( i = 1, \ldots, n \),
\[
p(m_i, n_i) = \frac{p(n_i)}{p(n)} = \frac{p(n_i)}{p(n)} \tag{10}
\]
for \( i = 1, \ldots, n \),
\[
p(m_i, n_i) = \frac{p(n_i)}{p(n)} = \frac{p(n_i)}{p(n)} \tag{11}
\]
for \( 0 \leq \alpha < 1 \) and \( \theta - \alpha > 0 \). For \( \alpha = 0, \theta = 0 \) this is the Ewens sampling formula.

2. Prediction rules for the species sampling models
By Theorem 2. in Hansen and Pitman (2000), the class of species sampling sequences, i.e. of all exchangeable sequences \( \{X_i\} \) admitting a prediction rule of the form:
\[
P(X_{i+1} \in [X_1, \ldots, X_i]) = \sum_{j=1}^i p_j p_{X_j}(\cdot) + p_{X_j}(\cdot), \tag{4}
\]
where \( [X_1, X_2] \), are the \( k \) distinct values in \( X_1, \ldots, X_i \), i.e., with non-atomic probability measure \( p \), is characterized by constraints on \( p_j \) and \( p_{X_j} \). It is shown that these quantities can be expressed in terms of EPPF associated with the random partition generated by \( \{X_1, \ldots, X_i\} \) as follows, provided \( p(n) > 0 \):
\[
p_j = p_j(n) = \frac{(n-1)!}{(n-j)!} p_j(n-j+1), \tag{5}
\]
for \( 1 \leq j \leq k \), where \( p_{X_j}(\cdot) = p(n_1, \ldots, n_j+1, \ldots, n_k) \), and
\[
p_j = p_j(n) = \frac{(n-1)!}{(n-j)!} p_j(n-j+1), \tag{6}
\]
for \( j = 1 \).
In Ishwaran and James (2003) the previous constraints are exploited to give a generalized version of the Chinese restaurant sequential construction, which provides samples from the partition structure induced by a species sampling model.

Remark 2. An alternative way to construct generalizations of the Ewens-Rényi model is to define a random graph by specifying its degree distribution. It easy to show that the model defined in section 3. is characterized by a random degree distribution governed by a variation of the sampling formula (2).

Acknowledgements
This poster was prepared while the author was visiting the research group in Probability and its Applications at the Boole Centre for Research in Informatics, University College Cork, Ireland. Thanks to Neil O’Connell for the kind hospitality and to SFI for supporting me.

References

Example 1. An explicit form for the sampling distribution (2) is known to be as follows for the two-parameter species sampling model discussed in Pitman (1996):
\[
p_j^*(m_j, n_j) = \frac{\prod_{i=1}^k (m_i + 1)}{\prod_{i=1}^k (n + 1)} \cdot \frac{1}{m_j}, \tag{11}
\]
for \( 0 \leq \alpha < 1 \) and \( \theta - \alpha > 0 \). For \( \alpha = 0, \theta = 0 \) this is the Ewens sampling formula.

The attachment rules defining the random evolution of the graph turn out to be as follows. Let a graph starts with a single vertex.

For \( n \geq 1 \), given \( (m_1, \ldots, m_n) \) the observed clique counts vector, the \( (n-1) \)-th adding vertex:

- joins one of the existing cliques of order \( i \), with probability:
\[
p^*_i = \frac{m_i(n_i - i)}{n - \theta}, \tag{12}
\]
for \( i = 1, \ldots, n \),

- starts a new clique, (i.e. remains isolated), with probability:
\[
p_0 = \frac{\theta}{n + \theta} \tag{13}
\]
At each time step, the observed clique counts vector is a sample from (11).

Remark 1. Consider a sequence of random graphs \( G_n = (V_n, E_n) \), whose vertices are labelled with an exchangeable sequence \( X_n \), and edges arise between nodes \( i \) and \( j \) such that \( X_i = X_j \), as in Cerqueti and Fortini (2003). If the vertices of the labels is a sample from a species sampling model, then, for each \( n \) (7) and (8) define a sequential construction of samples from \( G_n \). Moreover the random graph \( G_n \) decomposes almost surely into random cliques, and the clique counts vector distribution is given by (2).