Random Measures and Measure-Valued Processes

Jean Bertoin (Universität Zürich),
Shui Feng (McMaster University),
Paul Joyce (University of Idaho),
Ramsés H. Mena Chávez (Universidad Nacional Autónoma de México)

September 8 –13, 2013

1 Overview of the Field

Random measures and measure-valued processes are introduced to describe systems involving complex spatial and temporal structures. Over the past thirty years they have become the basic stochastic models in a wide range of subjects such as Bayesian non-parametrics, biology, communication networks, economics, financial analysis, machine learning, population genetics and statistical physics among others.

Consider a biological population of individuals that evolve under the influence of various factors such as mutation, genetic drift, natural selection and recombination. Each individual may reproduce, die or migrate according to the factors, and one is interested in asymptotic or macroscopic behaviours when the number of individuals is large. Under appropriate scaling, the system is approximated by a family of probability-valued processes including the Wright-Fisher diffusion in finite dimension and the Fleming-Viot processes in infinite dimensional space. When individual spatial displacement is involved, one is led to stepping-stone model, the interacting Fleming-Viot process, and the hierarchical process.

Sampling provides a natural connection between random measures and random partitions. This is especially important for measure-valued processes, as it enables the study of the genealogy of the population model in terms of certain coalescent processes, and also yields useful direct particle representations such as the look-down process of Donnelly and Kurtz [2]. In turn, sampling from certain random distributions (e.g. Poisson-Dirichlet) occurring in the setting of population genetics leads to remarkable combinatorial structures which have appeared independently in a variety of different fields.

Random measures arise naturally in Bayesian statistics as prior distributions. A basic requirement in Bayesian inference is the computability of the posterior distribution or the conjugacy property. Ferguson [1] introduced the Dirichlet process in the seventies and established the corresponding conjugate property. In recent years more random measures have appeared in Bayesian context. These include the Dirichlet process mixtures, the Chinese restaurant process, the Indian Buffet process, the Beta process, the Beta-Stacy process, the hierarchical Dirichlet process, etc.

A fundamental structure unifying different subjects is the underlying property of exchangeability. Many modern development is rooted in the relation between exchangeability, random partition, Kingman's coalescent, the GEM distribution (more generally the stick-breaking model), and the Poisson-Dirichlet distribution.

This workshop brings together researchers in several closely related areas including measure-valued processes, stochastic analysis, population genetics, and Bayesian non-parametric statistics, and provides a platform for inter-discipline dialog on strategies for future development.

2 Recent Developments and Open Problems

An important tool in the study of Fleming-Viot process is its dual process. The Fleming-Viot process describes the population evolution forward in time while the dual process offers a backward in time viewpoint. Various models of generalized Fleming-Viot processes have been introduced following the two different point of view.

Kingman's coalescent [3] is the first model that has been used for describing the genealogy of populations; it can be thought of as the dual to the neutral Fleming-Viot process. Kingman's coalescence has been extended to more general exchangeable coalescence processes (also known as coalescent with multiple and possibly simultaneous collisions) by Pitman [5], Sagitov [7], Möhle and Sagitov [4], and Schweinsberg [8] around 2000; these are coalescents known to describe the genealogy of generalized Fleming-Viot processes. Further more recent extensions incorporate immigration and neutral mutations; all these extensions exploit the fundamental analytic duality between the (forward in time) population model and the (backward in time) process of its genealogy. In this setting, the assumption of exchangeability of individuals is crucial; it implies that mutations must be neutral and that there should be no selection. An important challenge would thus be to develop a theory for the genealogy of population models with advantageous mutations and selection.

Ewens Sampling Formula (ESF) for Poisson-Dirichlet random measures is doubtless one of the most useful and best known mathematical results in population genetics, not to mention of course its applications to a variety of distinct areas in mathematics. It provides explicitly the distribution of the allelic partition (i.e. the partition of the current population into sub-populations with the same genetic type) for an n-sample of the idealized Wright-Fisher population model. ESF has been extended by Pitman to the so-called two-parameter model, and then further developments were obtained together with Gnedin. Recently, a different class of random partitions have appeared in connection to the allelic partitions of the total population generated by (sub)-critical Galton-Watson branching processes with neutral mutations. Although these allelic partitions possess simple characterizations, there are no known explicit sampling formulas. In a related direction, while the original models of coalescence did not take geometrical aspects into account, recent spatial coalescent models have been introduced by Limic and Sturm. This also yields new types of random partitions for which it would be interesting to obtain explicit sampling formulas.

While the Dirichlet and Poisson-Dirichlet distribution naturally arise in population genetics theory and the connections to various mathematical fields have been well established, one of the most recent genetics application of Dirichlet processes comes from phylogenetics. Phylogenetics explores evolution on long time scales and uses molecular genetics data to determine relationships between species. The most commonly used statistical approach to phylogenetics is called Mr. Bayes which was developed by Huelsenbeck. Dirichlet process priors are now central to drawing phylogenetic inferences. A stronger connection between the mathematics and phylogenetics world is currently lacking.

The theoretical advances in the field of measure-valued processes have permitted to incorporate a number of phenomena such as competition for resources, selection, immigration, mutations ..., which are intended to make stochastic models closer to reality. There is now a clear need of calibration methods for these more sophisticated models to fit data. This requires developing our knowledge about these stochastic population models. It is not sufficient to know that a certain model can be characterized e.g. as the unique solution to some martingale problem, one also needs more specific features concerning its distributions and its fine properties in order to develop efficient statistical methods.

Interestingly, the Bayesian non-parametric literatures provides with several models for random measures and also several ways of constructing them or representing them. Often these variety representations help to reveal distributional or other properties hard to get in other contexts. In particular, Ferguson's Dirichlet process can be represented in at least six different ways, e.g. as a normalized gamma process, as the limit of Pólya urns, as a species sampling model with stick-breaking weights, as a Pólya tree model, etc. Furthermore, these representations serve also as a way to construct more general models, other than Dirichlet process, for instance Lijoi and Prünster, and Lenk defined some general and important classes using respectively Kingman's completely random measures and Gaussian processes. Although the motivation to define random measures from this point of view is different from that used in other areas, these mathematical objects are widely used, applied and generalized within various statistical frameworks providing then with a wide variety of estimation and inference methods. Indeed, in a similar direction but with seemingly different purposes, Bayesian non-parametric methods have also evolved to allow for more complex dependence

structures, i.e. other than exchangeable samples. In particular, this has been done through what in this area are called dependent RPMs, namely sequences of RPMs indexed by some covariate or time index. These dependent processes directly connect to the concept of measure-valued processes, e.g. Fleming-Viot process with diploid fertility selection can be simply represented via a generalization of the Blackwell-MacQueen Pólya-urn scheme. Once more, within this viewpoint several distributional characterizations and estimation methods for measure-valued processes have been discovered. An important challenge here is to disentangle the connection of these constructions and properties with those used in areas such as super-processes or population genetics. From one side this would serve a starting point for implementations in those area but at the same time benefit from their canonical constructions and properties when defining new statistical methodologies.

3 Presentation Highlights

The workshop arranged twenty eight talks, and each talk is thirty five minutes long. These talks cover a wide range of topics and can be loosely categorized as follows:

- Measure-Valued Processes
- Random Measures and Bayesian Nonparametrics
- Coalescent
- Random Trees
- Stochastic Analysis

Measure-Valued Processes

The opening lecture by D. Dawson presented a dual approach to models with multilevel selection. The question of "levels of selection" has received a great deal of attention and controversy in the biological literature but relatively little attention in the mathematical literature.The proposed model is a multilevel interacting Fleming-Viot type process with selection at both individual level and the deem level. A related dual process is the main tool. J. Xiong discussed three nonlinear SPDEs arising from the study of continuous state branching processes in random environment. Several techniques were introduced to prove the uniqueness and other properties of these SPDEs. A link is drawn between SPDE and certain backward SDE. X. Zhou spoke about the modulus of continuity for Λ-Fleming-Viot processes with Brownian spatial motion. Properties such as uniform compactness, Hausdorff dimension and disconnectedness are also discussed for the support of the process. J. Blath gave a talk on the scaling limit of the interface of the symbiotic branching model. The limiting process of the interface of the continuous-space symbiotic branching model is investigated in the negative correlation regime under diffusive rescaling. Some properties of the shape of the limit is also discussed. A. Sturm considered long term behaviour of the law of a contact process started with a single infected site, distributed according to counting measure on the lattice. She showed that contact processes on general countable groups have in the subcritical regime a unique spatially homogeneous eigenmeasure.

A cluster of talks focuses on various generalizations of the Fleming-Viot process. R.C. Griffiths presented a generalized Wright-Fisher diffusion processes. The generalization is in the reproductive mechanism and the covariance structure of the offspring distributions. Connections with multi type branching diffusion processes are explored and stationary distribution in the two type case is identified explicitly. The process can be derived from the limit in a conditioned discrete multitype branching process in a constant-sized population of individuals of d-types. C. Foucart discussed the impact of selection on the disadvantage alleles in the Lambda-Wright-Fisher model. The resampling mechanism is governed by a finite measure Lambda on $[0, 1]$ and the selection by a parameter α. A critical value of α is identified and the disadvantaged allele is shown to vanish asymptotically when the selection is above the critical value. K. Handa discussed the ergodic properties for a class of generalized Fleming-Viot processes. The process is a natural generalization to the the Fleming-Viot process with parent-independent mutation. A generalized version of the Gamma-Dirichlet algebra is explored, the stationary distribution is identified, and the ergodic property is established. A. Etheridge gave an overview of the spatial Lambda-Fleming-Viot process. This is a class of processes modeling frequencies

of different genetic types in a population evolving in a spatial continuum. The biological motivation and mathmatical challenges were discussed. The talks also examined the relationship between the process and other familiar stochastic processes. In his concluding lecture, A. Wakolbinger presented an event-based construction and a look-down representation for a measure-valued equivalent to the spatial Lambda-Fleming-Viot process. The construction led to the derivation of several path properties of the measure-valued process as well as of the labeled trees describing the genealogical relations between a sample of individuals.

Two talks discussed stochastic modeling of some biological observation. S. Méléard reported some ongoing work about stochastic dynamics of adaptive trait and neutral marker driven by eco-evolutionary feedbacks. Each individual is characterized by trait and a genetic marker. The evolution of the population is driven by clonal reproduction, mutation and competition between individuals. The joint process of trait and marker dynamics is studied under various time scales. F. Yu considered the decoupling of linkage disequilibrium via recombination and its effect in increasing the fitness variance. These are demonstrated in two different settings.

Random Measures and Bayesian Nonparametrics

A major component of the workshop is random measures and Bayesian nonparametrics. Several presenters reported recent progresses and new challenges in the area.

Proposition 21 in [6] describes a representation of the two parameter Poisson-Dirichlet $PD(\alpha, \theta)$ using the jumps of a gamma subordinated generalized gamma subordinator for $\theta \geq 0, 0 < \alpha < 1$. L. F. James considered a class of random measures that generalize this proposition in various ways. This not only creates many possible new models for applications in Bayesian nonparametrics/machine learning but also has implications to certain types of fragmentation trees and coalescents. A. Lijoi discussed the recent Bayesian nonparametric literature and the issue on the proposal and the study of dependent random probability measures that are suited for the analysis of non-exchangeable data. He then presented classes of priors based on transformations of vectors of dependent completely random measures will be presented. Characterizations of the posterior distribution was displayed and used for estimating quantities of statistical interest in density estimation, clustering and survival analysis. P. Orbanz surveyed a class of random measures generating binary matrices and their application in Bayesian statistics. The matrices generalize exchangeable partitions to the case where blocks of the partition are not disjoint; each element of the underlying set can be contained in multiple, possibly overlapping blocks.

D.Spanò discussed the family of symmetric transition kernels with Gamma or Dirichlet stationary law. The focus is on a sub-family of such kernels that preserve the degree of polynomials. He also discussed open problems and potential applications in Bayesian inference. M. Ruggiero presented a link between optimal filtering for hidden Markov models and the notion of duality for Markov processes. He showed that when the signal is dual to a process that has two components, one deterministic and one a pure death process, and with respect to functions that define changes of measure conjugate to the emission density, the filtering distributions evolve in the family of finite mixtures of such measures and the filter can be computed at a cost that is polynomial in the number of observations. Reversibility in the signal plays a key role in the result.

Y. W. Teh proposed a novel Bayesian nonparametric model for genetic variations based on Markov processes over partitions, the fragmentation-coagulation processes. In comparison with the popular hidden Markov models, the new model is nonparametric and does not suffer from the label-switching issues. Statistical inference using an efficient Gibbs sampling algorithm reported encouraging result on genotype imputation.

Two talks discussed limit theorems such as law of large numbers (or consistency), central limit theorems,and large deviations. I. Pruenster considered the frequentist asymptotic behaviour of Gibbs-type priors. The focus was on posterior consistency and display conditions under which the posterior distribution accumulates in suitable neighbourhoods of what is assumed to be the 'true' data generating distribution. S. Favaro discussed Bayesian nonparametric inference for discovery probabilities. The basic setup is that given an initial observed sample of size n, an additional sample of size m is elected. One is interested in the asymptotic behaviour of certain conditional probabilities for large m.

Coalescent

Coalescent has been an active topic of research in recent years and several talks reported issues and progresses in this area. M. Möhle discussed conditions for an exchangeable coalescent to come down from infinity. A sufficient and necessary is found for a large subclass of coalescent and a conjecture is presented for the full class of exchangeable coalescents.

J. C. Pardo studied a weak law of large numbers for the total internal length of the Bolthausen-Szmitman coalescent. This is then applied to obtain the weak limit law of the centred and rescaled total external length. P. Pfaffelhuber presented recent work on some large deviation results in Kingman's coalescent. Let N_t be the number of blocks at small time t and $U_1, ..., U_{N_t}$ be the family sizes of the N_t families at time t. The large deviation results considered were associated with the laws of large number $tN_t \to 2$ and $N_t \sum_{i=1}^{N_t} U_i^2 \to 2$ as t converges to zero.

Random Trees

A major focus of the workshop is on random trees and applications. J. Delmas described the limits of all critical or subcritical Galton-Watson trees conditioned to have a large number of nodes or a large number of leaves or more generally a large number of nodes having a degree in a given set, and presented elementary proofs for these results. Key elements for the props are the Dwass formula and the strong ratio limit property for random walks. L. Popovic discussed the ancestral features such as MRCA of multitype branching trees. This is based on a multi-type coalescent point process description of the standing population of a general branching tree with finitely many different types. The dependence of these features on the offspring distribution is investigated. A. Winter presented recent work on the link between the discrete tree-valued pruning dynamics and its continuous counterpart. By introducing a new topology, the discrete and continuous dynamics become instances of the same Markov process with different initial conditions.

Stochastic Analysis

The Fleming-Viot process with parent-independent mutation is characterized by its reversibility among the family of Fleming-Viot processes. It is thus fit into the framework of symmetric diffusions in stochastic analysis. S. Fang presented a talk on Fokker-Planck equations on Lie groups. The talk started with a brief review of gradient flow on space of measures. The main result of construction of the equations on Lie groups was shown to follow from De Giorgie approximation. W. Sun discussed the relation between Markov processes and semi-Dirichlet forms as an extension to the relation between reversible Markov process and Dirichlet forms. Several recent results were presented including the representation of Markov processes, Hunts hypothesis (H) for Markov processes and Getoors conjecture for Lévy processes, Fukushimas decomposition for semi-Dirichlet forms.

4 Scientific Progress Made

In addition to formal presentations, time slots were allocated for group discussions and open problems. A real-dialogue took place between researchers from different areas. The development in random measures and measure-valued processes has been driven by progresses in biology and the demands in statistical inference. Different motivations led to different models. But a common theme is the development of models that are sophisticated enough to capture real world structures, and at the same are mathematical manageable so that explicit analysis can be performed. Several topics are identified for future development including dependent random measures, no-exchangeable random partitions, interacting Fleming-Viot process incorporating spatial structures and involving different level of selections, the application of non-reversible models in statistical inference, etc. J. Bertoin and L.F. James presented some open problems related respectively to the asymptotic behavior of partitions induced by large Galton-Waston trees, and to special properties of the two parameter Poisson-Dirichlet distribution.

Another progress made during the week associated with the model presented by C. Foucart. The process describes the evolution of the frequency of the disadvantaged allele in a two allele model. The random sampling is given by a measure Λ and the selection is the form $-\alpha x(1-x)$. The impact of selection is to reduce the frequency. A critical value α^* is identified. The behaviour of the process when t goes to infinity is studied for $\alpha < \alpha^*$ and $\alpha > \alpha^*$. R. Griffiths solved the case of $\alpha = \alpha^*$ during the workshop and gave a

brief presentation of his solution.

5 Outcome of the Meeting

The workshop provided an ideal venue for exchange of information between researchers in different areas. There is now a better understanding about the issues and challenges facing different research communities. A common goal is identified and the ground is laid for future development of the area and potential collaborations between participants.

The student participants were exposed to a wide range of topics and research areas. This will undoubtedly advance their research career in a very positive way.

Overall the workshop not only achieved but in some aspect exceeded the goal set in the original proposal.

References

[1] T.S. Ferguson (1973). A Baysian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230.

[2] P. Donnelly and T.G. Kurtz (1996). A countable representation of the Fleming–Viot measure-valued diffusion. *Ann. Prob.* **24**, No. 2, 698–742.

[3] J.C.F. Kingman (1982). The coalescent. *Stoch. Proc. Appl.* **13**, 235–248.

[4] M. Mölhe and S. Sagitov (2001). A classification of coalescent processes for haploid exchangeable partition models. *Ann. Probab.* **29**, 1547–1562.

[5] J. Pitman (1999). Coalescents with multiple collisions. *Ann. Probab.* **27**, 1870–1902.

[6] J. Pitman and M. Yor (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

[7] S. Sagitov (1999). The general coalescent with asynchronous mergers of ancestor lines. *J. Appl. Prob.* **36**, 1116–1125.

[8] J. Schweinsberg (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 1–50.