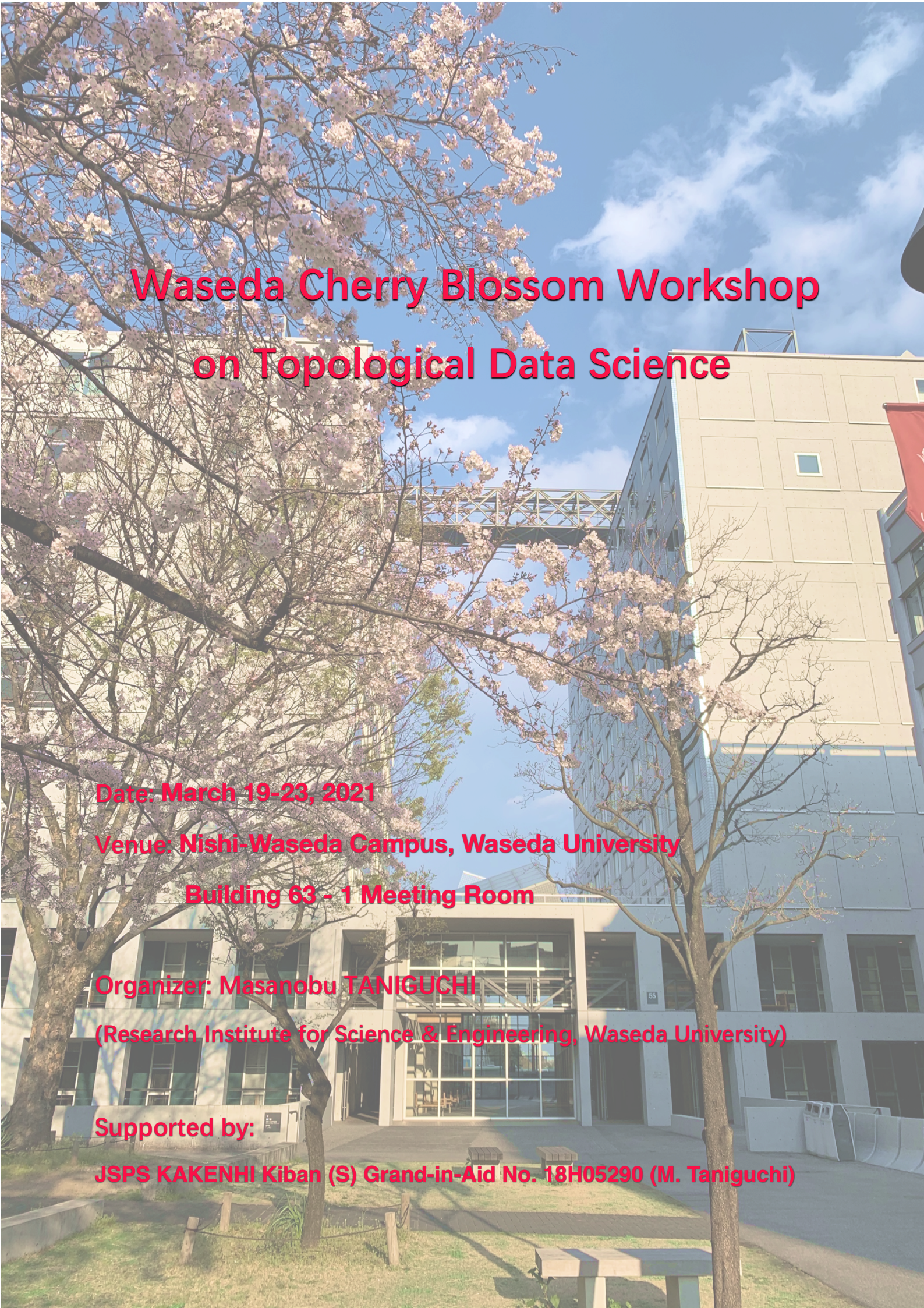

Waseda Cherry Blossom Workshop on Topological Data Science

Date: March 19-23, 2021

Venue: Nishi-Waseda Campus, Waseda University

Building 63 - 1 Meeting Room

Organizer: Masanobu TANIGUCHI

(Research Institute for Science & Engineering, Waseda University)

Supported by:

JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290 (M. Taniguchi)

Waseda Cherry Blossom Workshop on Topological Data Science

Date: March 19-23, 2021

Venue: Nishi-Waseda Campus, Waseda University

Building 63 - 1 Meeting Room

(Access map: https://www.waseda.jp/fsci/en/access/)

Organizer: Masanobu TANIGUCHI

(Research Institute for Science & Engineering, Waseda University)

This workshop is supported by:

JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290 (M. Taniguchi)

Program

March 19

09:50-10:00: Masanobu Taniguchi (Waseda Univ.)
Opening

■ Session I (10:00-12:00) chaired by Victor De Oliveira

10:00-11:00: Yan Liu (Waseda Univ.)
Statistical and Topological Inference of the Granger Causality

11:00-12:00: Takayuki Shiohama (Tokyo Univ. of Science)
Topological data analysis based classification and anomaly detection in time series

12:00-13:30: Lunch Time

■ Session II (13:30-17:00) chaired by Yan Liu

13:30-14:30: Yuichi Ike (Waseda Univ.) Zoom
Convergence result of stochastic subgradient descent for persistence-based functions

14:30-15:00: Coffee Break

15:00-16:00: Momoko Hayamizu (Waseda Univ.)
A structure theorem for tree-based phylogenetic networks: from theory to algorithms

16:10-17:00: Frederic Chazal (INRIA, France) Zoom
An Introduction to Topological Data Analysis, Part I

March 20

■ Session III (10:00-12:00) chaired by Takayuki Shiohama

10:00-12:00: Yusu Wang (UC San Diego) Zoom
Topological Data Analysis: How it can help in modern data analysis

Lunch & Cherry Blossom Festival

March 22

■ Session IV (9:00-11:50) chaired by Fumiya Akashi

9:00-9:50: Victor De Oliveira (Univ. of Texas) Zoom
An Introduction to Geostatistcs, Part I

10:00-10:50: Victor De Oliveira (Univ. of Texas) Zoom
An Introduction to Geostatistcs, Part II

11:00-11:50: Victor De Oliveira (Univ. of Texas) Zoom
Gaussian Copula Models for Geostatistical Count Data

11:50-13:30: Lunch Time

■ Session V (13:30-15:30) chaired by Xiaofei Xu

13:30-14:30: Yuichi Goto (Waseda Univ.)
Tests for a structural break and conditional variance of count time series

14:30-15:30: Fumiya Akashi (Univ. of Tokyo) Zoom
Robust regression methods in heavy-tailed processes and spherical predictors

15:30-16:00: Tea Time

■ Session VI (16:00-17:50) chaired by Masanobu Taniguchi

16:00-16:50: Frederic Chazal (INRIA, France) Zoom
An Introduction to Topological Data Analysis, Part II

17:00-17:50: Frederic Chazal (INRIA, France) Zoom
Linearization of persistence and the density of expected persistence diagrams

March 23

■ Session VII (10:00-12:00) chaired by Yuichi Goto

10:00-11:00: Xuze Zhang (Univ. of Maryland) Zoom
Estimation of residential radon concentration in Pennsylvania counties by data fusion

11:00-12:00: Xiaofei Xu (Waseda Univ.)
Adaptive log-linear zero-inflated generalized Poisson autoregressive model with applications to crime counts

12:00-13:30: Lunch Time

■ Session VIII (13:30-14:30) chaired by Masanobu Taniguchi

13:30-14:30: Tadashi Uratani (Hosei Univ.)
Pandemic, Insurance and Extreme Value Theory

Abstracts

March 19 (10:00–12:00)

Yan Liu

Title: Statistical and Topological Inference of the Granger Causality

Abstract: Granger causality has been employed to investigate causality relations between components of stationary multiple time series. Here, we generalize this concept by developing statistical inference for local Granger causality for multivariate locally stationary processes. Thus, our proposed local Granger causality approach captures time-evolving causality relationships in nonstationary processes. The proposed local Granger causality is well represented in the frequency domain and estimated based on the parametric time-varying spectral density matrix using the local Whittle likelihood. Under regularity conditions, we demonstrate that the estimators converge weakly to a Gaussian process. Additionally, the test statistic for the local Granger causality is shown to be asymptotically distributed as a quadratic form of a multivariate normal distribution. The finite sample performance is confirmed with several simulation studies for multivariate time-varying VAR models. For practical demonstration, the proposed local Granger causality method uncovered new functional connectivity relationships between channels in brain signals. Moreover, the method was able to identify

structural changes of Granger causality in financial data. (Joint work with Masanobu Taniguchi and Hernando Ombao)

Takayuki Shiohama

Title: Topological data analysis based classification and anomaly detection in time series

Abstract: Time series often contain outliers and level shifts or structural changes. These unexpected events are of the utmost importance in anomaly detection. The presence of such unusual events can easily mislead conventional time series analysis and yield erroneous conclusions. Anomaly detection methods for time series have been studied for decades and demonstrated to be useful in many applications. There exist many notable methods in machine learning, which include clustering analysis, isolation forests, and classifiers using artificial neural networks. Most of these techniques often are most effective when there are many additional features. In this study, we use topological data analysis (TDA) in order to provide more accurate classifier that can also detect unusual events in time series.

March 19 (13:30–17:00)

Yuichi Ike

Title: Convergence result of stochastic subgradient descent for persistence-based functions

Abstract: Optimization of functions and losses with topological flavor is an active and growing field of research in Topological Data Analysis, with plenty of applications to Machine Learning. In practice, one just applies stochastic subgradient descent to such a topological function, but the corresponding gradient and associated algorithm do not come with theoretical guarantees. In this talk, we will talk about a convergence result of stochastic subgradient descent for such a function, relying on the theory of o-minimal structures. This result includes all the constructions and applications for topological optimization in the literature. We show some experiments such as dimension reduction and filter selection to showcase the versatility of our approach. (Joint work with Mathieu Carrière, Frédéric Chazal, Marc Glisse, Hariprasad Kannan, and Yuhei Umeda)

Momoko Hayamizu

Title: A structure theorem for tree-based phylogenetic networks: from theory to algorithms

Abstract: While phylogenetic networks are useful to visualise non-treelike data or complex evolutionary histories, there are many

computationally hard problems regarding them. Therefore, it is important to define nice subclasses of phylogenetic networks that are mathematically tractable and biologically meaningful. In view of this, the concept of "tree-based" phylogenetic networks, which was originally introduced by Francis and Steel in 2015, has attracted great attention and given rise to various interesting research problems in combinatorial phylogenetics. In this talk, I provide the necessary background and explain how to solve those different problems in a unified manner. The talk is mainly based on arXiv:1811.05849 [math.CO]. I also mention more recent advancement that is joint work with Kazuhisa Makino (arXiv:1904.12432 [math.CO]).

Frederic Chazal

Title: An introduction to Topological Data Analysis
Part I: persistent homology theory

Abstract: Topological Data Analysis (TDA) is a recent and fast growing field providing a set of new topological and geometric tools to infer relevant features of possibly complex data. Among these tools, persistent homology plays a central role. It provides a mathematically well-founded basis to design efficient and robust methods to estimate, analyze and exploit the topological and geometric structure of data. This first talk will be dedicated to a brief introduction to persistent homology and its usage in TDA. We will introduce persistent homology for functions and point cloud data and study its stability properties. The talk does not require any specific background in

topology, the basic notions needed to introduce the persistent homology will be recalled or introduced during the talks.

March 20 (10:00–12:00)

Yusu Wang

Title: Topological Data Analysis: How it can help in modern data analysis

Abstract: In recent years, a new field for data, called Topological data analysis, has attracted much attention from researchers from diverse background, including computer science, applied mathematics and statistics. Leveraging various fundamental developments both in theoretical and algorithmic fronts in the past two decades, topological data analysis has been growing rapidly, and already applied in many applied domains, such as computational neuroscience, material science and bioinformatics.

In this time, I want to give some examples on where topological ideas could help with analyzing complex modern data. I will specifically focus on the following three aspects: (1) Topolgoical methods could provide flexibile yet generic framework for feature summarization / characterization. (2) Topological methods could help model, infer, and explore the hidden space behind data. (3) How to combine topological ideas with machine learning pipelines. I will use recent work from my research group to illustrate these points. Through the course, we will

touch upon multiple topological objects, including persistent homology, discrete Morse theory, and contour trees.

March 20 (9:00–11:50)

Victor De Oliveira

Title: An Introduction to Geostatistics, Part I

Abstract: In this talk I introduce some of the types of data and scientific problems for which geostatistics is used, the basic probabilistic tools needed to model geostatistical data, and the classical statistical methods of analysis. First, I describe the semivariogram function, the basic tool used in geostatistics to model the spatial association displayed by the quantity of interest, and then I describe the classical methods used for its estimation. These involve a two—step approach that is distribution-free as is based on moments and least squares. The pros and cons of these classical methods are discussed. Second, I describe several variants of the so--called ‘kriging’ prediction method, which are nothing other than applications of best linear unbiased prediction. I will review some of the properties of these predictors and their mean squared prediction errors, as well as the ability (or lack of) of the latter to properly account for the prediction uncertainty. The pros and cons of kriging predictors are discussed. The models and methods will be illustrated with several real data sets.

Victor De Oliveira

Title: An Introduction to Geostatistics, Part II

Abstract: In this talk I introduce models for geostatistical data based on Gaussian random fields. First, I describe the frequentist methods of maximum likelihood and restricted maximum likelihood to estimate the model parameters, as well as some of the properties of these. I also describe the optimal predictors and their relation to kriging predictors. The two main asymptotic frameworks for this type of data are reviewed, called increasing and fixed domain frameworks, and the dissimilar large-- sample properties of maximum likelihood estimators under these two frameworks are discussed. Second, Bayesian methods for estimation and prediction are described as well as some basic Markov chain Monte Carlo algorithms currently used to make inference about these models. The issue of how to select `good priors' for these model is also briefly discussed. Finally, two classes of non--Gaussian models are introduced to describe continuous data with skewed distributions and geostatistical count data that use Gaussian random fields as building blocks: transformed Gaussian random fields and Poisson hierarchical models. The models and methods will be illustrated with several real data sets.

Victor De Oliveira

Title: Gaussian Copula Models for Geostatistical Count Data

Abstract: In this talk I describe a class of random field models for geostatistical count data based on Gaussian copulas. Unlike hierarchical Poisson models often used to describe this type of data, Gaussian copula models allow a more direct modeling of the marginal distributions and association structure of the count data. I describe in detail the correlation structure of these random fields when the family of marginal distributions is either negative binomial or zero--inflated Poisson; these represent two types of overdispersion often encountered in geostatistical count data. I also contrast the correlation structure of one of these Gaussian copula models with that of a hierarchical Poisson model having the same family of marginal distributions. I also describe the computation of maximum likelihood estimators which are a computationally challenging task. Finally, a data analysis of Lansing Woods tree counts is used to illustrate the methods.

March 22 (13:30–17:50)

Yuichi Goto

Title: Tests for a structural break and conditional variance of count time series

Abstract: Count time series have been attracted attention and widely studied. We deal with count time series whose conditional expectation has dependence structure. This model is motivated by generalized

linear models. In this talk, we discuss two hypothesis testing problems for count time series. The first is a test for a structural break. We propose Wald type, score type, residual type of CUSUM test statistics, and show the asymptotic null distributions. This result enables us to construct distribution-free and asymptotic size alpha tests. Moreover, the tests based on a modified Wald statistic and a score type statistic are consistent. The second is a test for the conditional variance. We elucidate the asymptotic null distribution of a proposed test statistic and show the consistency of the proposed test. Moreover, the local alternative power is also clarified. This test can be applied to various testing problems such as a goodness of fit test, a specification test of intensity function, and a test for equidispersion. The simulation study illustrates the finite sample performance of the above methods. The number of patients with Escherichia coli in a state of Germany is also analyzed. (The test for a conditional variance of count time series is based on the joint work with K. Fujimori)

Fumiya Akashi

Title: Robust regression methods in heavy-tailed processes and spherical predictors

Abstract: Statistical treatment for non-stationarity, heteroscedasticity and heavy tails of the real data has attracted a lot of attention in these decades. The analysis for the locally stationary (LS) processes has been also developed under the finite variance assumptions. The former half of this talk extends the framework to the LS processes

with possibly infinite variance error terms and construct the L1-regression-based local linear estimator for the coefficients of the model. In addition, the self-weighting method is also employed to reduce the leverage effect brought by the past values of the observations. The proposed local-linear estimator is shown to have asymptotic normality regardless of whether the innovation process has finite variance or dependence structure. The latter half section of this talk considers a nonlinear regression model whose predictor is a random vector on a hyper-sphere. To construct a robust estimator for the nonlinear regression function, we consider a spherical kernel-type objective function, and elucidate robust properties of the estimator. Some simulation experiments illustrate desired finite sample properties of the proposed methods. (Joint works with Junichi Hirukawa, Konstantinos Fokianos and Holger Dette)

Frederic Chazal

Title: An introduction to Topological Data Analysis
Part II: statistical properties of persistent homology

Abstract: This second talk will be dedicated to the statistical study of persistent homology. We will show how the stability properties of persistence can be used to understand the behavior of persistence diagrams in various (selected) statistical settings. We will also illustrate how these statistical properties can be used to overcome some computational and noise issues encountered in practical TDA applications.

Frederic Chazal

Title: Linearization of persistence and the density of expected persistence diagrams

Abstract: Persistence diagrams play a fundamental role in Topological Data Analysis (TDA) where they are used as topological descriptors of data represented as point cloud. They consist in discrete multisets of points in the plane that can equivalently be seen as discrete measures. When they are built on top of random data sets, persistence diagrams become random measures. In this talk, we will show that, in many cases, the expectation of these random discrete measures has a density with respect to the Lebesgue measure in the plane. We will discuss its estimation and show that various classical representations of persistence diagrams (persistence images, Betti curves,...) can be seen as kernel- based estimates of quantities deduced from it. This is a joint work with Vincent Divol (ENS Paris / Inria DataShape team).

March 23 (10:00–12:00)

Xuze Zhang

Title: Estimation of residential radon concentration in Pennsylvania counties by data fusion

Abstract: Radon is a tasteless, colorless, and odorless radioactive gas that is considered as the leading cause of lung cancer among nonsmoker. Residential exposure to radon has been a serious public health problem in Pennsylvania (PA) in the past several years since record shows that a considerable proportion of PA houses have radon concentration beyond safety level 4 pCi/L. Thus, estimation of residential radon concentration, especially estimation of exceedance probability for a high threshold, becomes a prob- lem of interest. A multisample density ratio model (DRM) with variable tilts is proposed and applied to fused data from a reference county of interest and its neighboring counties to obtain the estimated distribution of radon concentration and confidence intervals that correspond to the estimates of exceedance probabilities of interest. (Joint work with Saumyadipta Pyne and Benjamin Kedem)

Xiaofei Xu

Title: Adaptive log-linear zero-inflated generalized Poisson autoregressive model with applications to crime counts

Abstract: This research proposes a comprehensive ALG model (Adaptive Log-linear zero-inflated Generalized Poisson integer-valued GARCH) to describe the dynamics of integer-valued time series of crime incidents with the features of autocorrelation, heteroscedasticity, over-dispersion, and excessive number of zero observations. The proposed ALG model captures time-varying nonlinear dependence and simultaneously incorporates the impact of

multiple exogenous variables in a unified modeling framework. We use an adaptive approach to automatically detect subsamples of local homogeneity at each time point of interest and estimate the time-dependent parameters through an adaptive Bayesian Markov Chain Monte Carlo (MCMC) sampling scheme. A simulation study shows stable and accurate finite sample performances of the ALG model under both homogeneous and heterogeneous scenarios. When implemented with data on crime incidents in Byron, Australia, the ALG model delivers a persuasive estimation of the stochastic intensity of criminals and provides insightful interpretations on both the dynamics of intensity and the impacts of temperature and demographic factors to different crime categories. (Joint work with Ying Chen, Cathy W. S. Chen and Xiancheng Lin)

March 23 (13:30–14:30)

Tadashi Uratani

Title: Pandemic, Insurance and Extreme Value Theory

Abstract: The pandemic of COVID-19 is the most devastating shocks experienced by the world in peacetime in mortality and economy.
“Excess deaths” is different in countries, more are Europe and America while less are Asians, but it affects uniformly national economy and government budget. Government spending for Covid-19 has increased sharply deficit finance. We discuss on the financing

to extreme event risk management by Catastrophe Bond in Extreme Value Theory.