ISI Informal Workshop on Learning, Algorithms and Networks

17th-18th February 2020
ISI Foundation, Via Chisola 5, Turin
Seminar Room (first floor)

Fondazione ISI
ISI Foundation

As we have a large number of friends and collaborators visiting in these days, instead of doing many seminars one at the time, we have decided to put together a 2-days informal workshop showcasing some of the on-going research in the wide area of AI at ISI Foundation. The workshop will feature 18 talks from senior researchers, post-docs, as well as PhD students, focussing on Machine Learning, Network Science, and Graph Algorithms.

The detailed program together with all the titles and abstracts is provided in the next pages.

As we have limited chairs, if you're external to ISI and would like to atterndd, please drop an email to: francesco.bonchi@isi.it

February 17, 2020		
Speaker	**Title**	**Time**
Marcello Restelli (Politecnico di Milano & ISI Foundation)	Safe Policy Search	11.00-11.45
Davide Mottin (Aarhus University, Denmark)	The user and the machine: Blending database and machine learning for smarter devices	11.45-12.30
Alan Perotti (ISI Foundation)	eXplainable Artificial Intelligence	12.30-13.00
Lunch break		13.00-14.00
Fabio Vitale (University of Lille, & INRIA France)	Compressing graph information for binary node classification	14.00-14.45
Tommaso Lanciano (La Sapienza University, Rome)	Extracting contrast subgraphs with applications to brain networks	14.45-15.15
António Leitão (NOVA IMS, Portugal)	Topology of spaces in learning systems	15.15-15.45
Coffee break		15.45-16.15
Corrado Monti (ISI Foundation)	Inference models for opinion formation in social media	16.15-16.45
Maddalena Torricelli (Univ. of Bologna & ISI Foundation)	weg2vec: Event embedding for temporal networks	16.45-17.15
Simone Piaggesi (Univ. of Bologna & ISI Foundation)	Temporal Network Embedding as Implicit Tensor Factorization	17.15-17.45

February 18, 2020		
Speaker	**Title**	**Time**
Giuseppe Manco (ICAR - CNR, Cosenza)	Adversarial Games for generative modeling of Temporally-Marked Event Sequences	11.00-11.45
David Garcia-Soriano (ISI Foundation)	Fair-by-design matching	11.45-12.30
Yllka Velaj (ISI Foundation)	Approximate Pricing in Networks: How to Boost the Betweenness and Revenue of a Node	12.30-13.00
Lunch break		13.00-14.00
Francesco Bonchi (ISI Foundation)	Distance-based community search	14.00-14.45
Cigdem Aslay (Aalto University, Finland)	Discovering Interesting Cycles in Directed Graphs	14.45-15.15
Gianmarco De Francisci Morales (ISI Foundation)	Quantifying Uncertainty in Online Regression Forests	15:15-15:45
Coffee break		15.45-16.15
Xiangyu Ke (Nanyang Technological University, Singapore)	Reliability Query on Uncertain Graphs: Algorithms and Applications	16.15-16.45
Martina Patone (University of Southampton)	Graph sampling and estimation	16.45-17.15
Alessio Angius (ISI Foundation)	Markov Decision Processes applied to Industry 4.0	17.15-17.45

Marcello Restelli (Politecnico di Milano & ISI Foundation):
“Safe Policy Search”

Abstract: Over the past decade, policy search methods have received a lot of attention within the reinforcement learning research community. They are commonly used in many real-world decision-making problems (e.g., robot control, autonomous driving, automatic trading, etc.) due to their ability to learn in continuous high-dimensional domains, the possibility to incorporate prior knowledge, and their robustness to partial observability. The application of learning systems that make decisions in the real-world asks for the development of algorithms able to interact safely with the environment.

This talk will focus on the policy search algorithms that achieve safety by guaranteeing (with high probability) a monotonically-improving learning process. Furthermore, it will be shown how the ideas behind these safe approaches have inspired the development of some (non-safe) policy search algorithms that belong to the state of the art of reinforcement learning.

Davide Mottin (Aarhus University, Denmark):
“The user and the machine: Blending database and machine learning for smarter devices”

Abstract: Can an algorithm discover what a user wants while accessing graph data? Can a system answer future queries without accessing the data again? In this talk, I will introduce two recent results that build upon machine learning and graph mining techniques to learn user preferences as fast as possible with the fewest number of interactions with the user. In particular, I will focus on techniques that exploit system approximations of user preferences to learn real user preferences more efficiently and effectively. Then, I will introduce a different perspective knowledge graph search which removes the need to access the data in a server by learning a small, informative summary on the user device.

Alan Perotti (ISI Foundation):
“eXplainable Artificial Intelligence”

Abstract: Several recent advancements in Machine Learning involve black-box models: algorithms that do not provide human-understandable explanations in support of their decisions. These models might be right for the wrong reasons, inadvertently implement racial or gender biases, not comply with GDPR, and so on. The field of eXplainable Artificial Intelligence (XAI) tries to solve this problem providing human-understandable explanations for black-box models - however, this research field is still developing and is far from matching the desiderata.

In this talk I will introduce some general lines of research on XAI we are pursuing at ISI. As a first line of research, we argue that in many real life application domains there is an untapped wealth of additional information, often represented in ontological form. For instance, all classes of the well known ImageNet dataset are in fact nodes in the Wordnet lexical database, and all medical codes that make up the MIMIC-III Critical Care dataset are nodes of the ICD-9 taxonomy. XAI techniques typically rely on the mere relationship between a data point and its black-box label, overlooking all domain information; we show how domain-agnostic XAI algorithms can be paired with said structured knowledge.

As a second line of research, we focus on explanation evaluation. The existing literature lists many desirable properties for explanations to be useful, but there is a scarce consensus on how to evaluate explanations in practice. Moreover, explanations are typically used only to inspect black-box models, and the proactive use of explanations as a decision support is generally overlooked. We introduce novel metrics and integrate them in a general framework for explanation evaluation. Other explainability tools that we are developing and that will be briefly introduced are: the topological inspection of the feature space, the relation between semantic losses and hierarchies in representation learning, embedding the dataset in (feature-level) Shapley values space, analysing the use of attention scores as feature importances, and exploring the concept of actionable explanations.

Fabio Vitale (University of Lille, France):
"Compressing graph information for binary node classification"

Abstract: We investigate the problem of sequentially predicting the binary labels on the nodes of an arbitrary weighted graph. We show that, under a suitable parametrization of the problem, the optimal number of prediction mistakes can be characterized (up to logarithmic factors) by the cut size of a random spanning tree of the graph. The cutsize is induced by the unknown adversarial labeling of the graph nodes. In deriving our characterization, we obtain a simple randomized algorithm achieving in expectation the optimal mistake bound on any polynomially connected weighted graph. Our algorithm draws a random spanning tree of the original graph and then predicts the nodes of this tree in constant expected amortized time and linear space. Experiments on real-world data sets show that our method compares well to both global (Perceptron) and local (label propagation) methods, while being generally faster in practice.

Tommaso Lanciano (La Sapienza University of Rome, Italy):
"Extracting contrast subgraphs with applications to brain networks"

Abstract: The task of classification in ML is well-known: we are given several instances of the same phenomena but that belongs to different classes, and we want to find a way, based on a set of features of such instances, to guess which class a possible new observation belongs. Supposing we are given these instances in the shape of networks, a good research question might be the following: are we able to characterize which is the portion of network that most differs between the 2 classes? In light of these, we introduce the problem of contrast subgraph: given 2 networks that shares the same set of nodes V, we want to find a subset of V that is highly connected in the first network, and extremely sparse in the other. Obviously, the final aim is to find such subset that maximizes this difference. We show that the problem is hard and we devise an approximated algorithm. Furthermore, we find an interesting application: given the brain networks of both control and ill patients, are we able to characterize the brain region that are involved in the disease? We performed several experiments that show how contrast subgraph is an important feature in order to give a correct diagnosis in patients affected by bipolar disorder.

António Leitão (NOVA IMS, Portugal)
"Topology of spaces in learning systems"

Abstract: Topology deals with spaces and maps between spaces, shapes and deformations. Basic equivalence relations in topology highlight robust qualitative features of a space. Crafting learning models that leverage topological invariants has yielded new insights and results, further elevating topological data analysis as an important step in machine learning.

By considering learning models as transformations acting on a given data space, topology becomes a natural lens with which to base the model's architecture on. In this talk I will expose how a neural network's architecture affects the topology of its parameter space and its representation of the data. Specifically, how one can adapt persistent homology to study the effects of a transformation on a given input space. I will also show how the topological characterization of the data might not reflect the topological characterization of a given classification problem, by focusing on the homology of the decision boundaries. I will give examples and results on how these two differ and why a clear distinction is important when tailoring effective and robust learning models.

Corrado Monti (ISI Foundation):
"Inference models for opinion formation in social media"

Abstract: Social media has significantly transformed how the general public consumes information. Despite the increasing attention received by the research community, modelling how opinions interact with social media remains an important open question. I am going to show how the design of such models can be tackled quantitatively thanks to maximum likelihood estimation on probabilistic graphical models.

The first question is about the space of opinions. We designed a model for how controversial content propagate on a social network, in which opinions live in a multidimensional, topic-based, ideological space. With a gradient-descent procedure, we can fit this model to data, turning a theoretical model for controversial content spreading into an algorithm for multi-dimensional opinion mining. In this framework we can infer people positions on different axes: economy, foreign policy, minorities, and study how different leanings relate to each other.

The second question is about how opinions evolve in time. Agent-based models are traditionally used for this aim, since they are easy to interpret, and encode sociological theories as causal mechanisms for opinion formation. However, these models do not exploit the widespread availability of social traces online. We show how by casting a traditional ABM as a problem of assigning signs on the edges of a temporal graph, we can perform inference and recover a likely latent opinion trajectory. This framework can then be used to analyze the evolution in time of internet communities, quantifying phenomena such as polarization and normalization.

Maddalena Torricelli (University of Bologna & ISI Foundation, Italy):
"weg2vec: Event embedding for temporal networks"

Abstract: Network embedding techniques are powerful to capture structural regularities in networks and to identify similarities between their local fabrics. However, conventional network embedding models are developed for static structures, commonly consider nodes only and they are seriously challenged when the network is varying in time. Temporal networks may provide an advantage in the description of real systems, but they code more complex information, which could be effectively represented only by a handful of methods so far. Here, we propose a new method of event embedding of temporal networks, called weg2vec,which builds on temporal and structural similarities of events to learn a low dimensional representation of a temporal network. This projection successfully captures latent structures and similarities between events involving different nodes at different times and provides ways to predict the final outcome of spreading processes unfolding on the temporal structure.

Simone Piaggesi (University of Bologna & ISI Foundation, Italy):
"Temporal Network Embedding as Implicit Tensor Factorization"

Abstract: Representation learning models for networks are proven to be a successful family of techniques that project nodes of static graphs into feature spaces that can be exploited by other machine learning algorithms. But many real-world networks are inherently dynamic, with interactions among nodes changing over time. Moreover, existing representation learning techniques that aggregate temporal networks into static representations and do not explicitly take time into account are not able to discern structure and dynamics.

We build upon the idea that the skip-gram embedding approach implicitly performs a matrix factorization, and we extend it to perform an implicit tensor factorization on different tensor representations of time-varying graphs. We show that our approach is able to disentangle the role of nodes and time, with a large reduction of the number of parameters needed by the model to represent the system.

Giuseppe Manco (ICAR - C.N.R., Cosenza):
"Adversarial Games for generative modeling of Temporally-Marked Event Sequences"

Abstract: Increasing amounts of data are becoming available in the form of "asynchronous" sequences of event records, associated each with a content and a temporal mark, e.g. sequences of activities on social media, clickstream data, user interaction logs, point-of-interest trajectories, business process logs, application logs and IoT logs, to name a few. Such a kind of data are more general than classic time series, as the lapse of time between consecutive events in a sequence may be an arbitrary continuous value. Usually, events in the same sequence exhibit hidden correlations (e.g., an event can cause or prevent the occurrence of certain kinds of events in the future). Generative models constitute a powerful and versatile means for analysing such data (e.g., by supporting variegate key tasks, such as data completion, data denoising, simulation analyses), as well as for enabling the very generation of new data instances (e.g., for preventing information leakage). In particular, if devised in a conditional fashion, these models can be exploited to predict which events will happen in the remainder of a given (unfinished) sequence and when, based on the sequence's history. Different kinds of neural generative models have been used in the last years for analysing sequence data, which range from Recurrent Neural Networks (RNNs), to Self-attention models, to more sophisticated frameworks like Variational Auto-encoders (VAEs) and Generative Adversarial Networks (GANs). In particular, basic GAN frameworks implement a sort min-max game, where a "discriminator" sub-net is trained to distinguish real data instances from those produced by a "generator" sub-net, which is trained instead to fool the former sub-net.

In principle, GANs could yield models that are more general (owing to the GANs' capability to learn implicit data distributions) and more robust to "exposure bias" issues (i.e., to the risk of accumulating errors at inference time). However, they are notoriously difficult to train and tune optimally —in particular, they may converge slowly, and eventually reach an equilibrium that does not ensure a good enough generator (e.g., the latter may suffer from severe mode collapse issues). Moreover, the discrete nature of temporally-marked event sequences calls for extending traditional GAN schemes, as to prevent the problem of breaking the differentiability of the discriminators output w.r.t. the generators parameters. The purpose of this talk is to review the recent contributions of the state of the art in the fields and to illustrate a proposal which combines generative modeling of events and discriminative capabilities in detecting event occurrences and yet it's capable of overcoming the current limitations.

David García-Soriano (ISI Foundation, Italy):
"Fair-by-design matching"

Abstract: Matching algorithms are used routinely to match donors to recipients for solid organs transplantation, for the assignment of medical residents to hospitals, record linkage in databases, scheduling jobs on machines, network switching, online advertising, and image recognition, among others. Although many optimal solutions may exist to a given matching problem, when the elements that shall or not be included in a solution correspond to individuals, it becomes of paramount importance that the solution be selected fairly. Given that many maximum matchings may exist, each one satisfying a different set of individuals, the only way to guarantee fairness is through randomization. Hence we introduce the distributional maxmin fairness framework which provides, for any given input instance, the strongest guarantee possible simultaneously for all individuals in terms of satisfaction probability (the probability of being matched in the solution). Specifically, a probability distribution over feasible solutions is maxmin-fair if it is not possible to improve the satisfaction probability of any individual without decreasing it for some other individual which is no better off. Our main contribution is a polynomial-time algorithm building on techniques from minimum cuts, and edge-coloring algorithms for regular bipartite graphs, and transversal theory. In the special case of bipartite matching, our algorithm runs in $O((|V|^2 + |E||V|^{2/3}) \cdot (\log |V|)^2)$ expected time. An experimental evaluation of our fair-matching algorithm shows its ability to scale to graphs with tens of millions of vertices and hundreds of millions of edges, taking only a few minutes on a simple architecture. Our analysis confirms that our method provides stronger satisfaction probability guarantees than non-trivial baselines. Joint work with Francesco Bonchi.

Yllka Velaj (ISI Foundation, Italy):
"Approximate Pricing in Networks: How to Boost the Betweenness and Revenue of a Node"

Abstract: We introduce and study two new pricing problems in networks: Suppose we are given a directed graph G= (V,E) with non-negative edge costs, k commodities (si, ti, wi) in [k] and a designated node u in V. Each commodity i is represented by a source-target pair (si,ti) and a demand wi>0, specifying that wi units of flow are sent from si to ti along shortest si,ti-paths (with respect to the cost of the edges). The demand of each commodity is split evenly over all shortest paths. Assume we can change the edge costs of some of the outgoing edges of u, while the costs of all other edges remain fixed; we also say that we price (or tax) the edges of u. We study the problem of pricing the edges of u with respect to the following two natural objectives: (i)max-flow: maximize the total flow passing through u, and (ii)max-revenue: maximize the total revenue (flow times tax) through u. Both variants have various applications in practice. For example, the max flow objective is equivalent to maximizing the betweenness centrality of u,which is one of the most popular measures for the influence of a node in a (social) network. We prove that (except for some special cases) both problems are NP-hard and inapproximable in general and therefore resort to approximation algorithms. We derive approximation algorithms for both variants and show that the derived approximation guarantees are best possible.

Francesco Bonchi (ISI Foundation, Italy):
"Distance-based community search"

Abstract: Suppose we have identified a set of subjects in a terrorist network suspected of organizing an attack. Which other subjects, likely to be involved, should we keep under control? Similarly, given a set of patients infected with a viral disease, which other people should we monitor? Given a set of proteins of interest, which other proteins participate in pathways with them? Each of these questions can be modeled as a graph-query problem: given a graph G = (V,E) and a set of query vertices Q, find a subgraph H of G which "explains" the connections existing among the nodes in Q, hat is to say that H must be connected and contain all query vertices in Q.

We start by providing a brief survey of various measures and methods defined for this network problem, then we turn our attention to the problem of finding a "minimum Wiener connector", i.e., the subgraph of G that connects all query vertices and that minimizes the sum of all pairwise shortest-path distances between its vertices (Wiener Index). We show that the minimum Wiener connector is smaller and denser than other methods in the literature, and it contains highly central nodes.

In the second part of the talk, we relax the constraint of connecting all the query vertices. Relaxing the connectedness requirement allows the connector to detect multiple communities and to be tolerant to outliers. We achieve this by introducing the new measure of network inefficiency and by instantiating our search for a selective connector as the problem of finding the minimum inefficiency subgraph. We show that our problem is hard and devise efficient algorithms to approximate it. By means of several case studies in a variety of application domains (such as human brain, cancer, and food networks), we show that our minimum inefficiency subgraph produces high-quality solutions, exhibiting all the desired behaviors of a selective connector. Finally, we extend the present notions to the case of temporal dynamic networks showing how our tools can be used to track a community of interest adaptively in time.

Cigdem Aslay (Aalto University, Finland):
"Discovering Interesting Cycles in Directed Graphs"

Abstract: Cycles in graphs often signify interesting processes. For example, cyclic trading patterns can indicate inefficiencies or economic dependencies in trade networks and cycles in food webs can identify fragile dependencies in ecosystems. Identifying such interesting cycles, which can also be constrained to contain a given set of query nodes, although not extensively studied, is a problem of considerable importance.

In this talk, I will introduce our recent result on the problem of discovering interesting cycles in graphs. We first address the problem of quantifying the extent to which a given cycle is interesting for a particular analyst. We then show that finding cycles according to this interestingness measure is related to the longest cycle and maximum mean-weight cycle problems (in the unconstrained setting) and to the maximum Steiner cycle and maximum mean Steiner cycle problems (in the constrained setting). We show that the problems of finding the most interesting cycle and Steiner cycle are both NP-hard, and are NP-hard to approximate within a constant factor in the unconstrained setting, and within a factor polynomial in the input size for the constrained setting. We also show that the latter inapproximability result implies a similar result for the maximum Steiner cycle and maximum mean Steiner cycle problems. Motivated by these hardness results, we propose a number of efficient heuristic algorithms. We verify the effectiveness of proposed methods and demonstrate their practical utility on two real-world use cases: a food web and an international trade-network dataset.

Gianmarco De Francisci Morales (ISI Foundation):
"Quantifying Uncertainty in Online Regression Forests"

Abstract: Accurately quantifying uncertainty in predictions is essential for the deployment of machine learning algorithms in critical applications where mistakes are costly. Most approaches to quantifying prediction uncertainty have focused on settings where the data is static, or bounded. Here, we investigate methods that quantify the prediction uncertainty in a streaming setting, where the data is potentially unbounded.

We propose two meta-algorithms that produce prediction intervals for online regression forests of arbitrary tree models; one based on conformal prediction theory, and the other based on quantile regression. We show that the approaches are able to maintain specified error rates, with constant computational cost per example and bounded memory usage.

Xiangyu KE (Nanyang Technological University, Singapore):
"Reliability Query on Uncertain Graphs: Algorithms and Applications"

Abstract: Uncertain, or probabilistic, graphs have been increasingly used to represent noisy linked data in many emerging applications, and have recently attracted the attention of the data management research community. A fundamental problem on uncertain graphs is the s-t reliability, which measures the probability that a target node t is reachable from a source node s in an uncertain graph. In this talk, I will first introduce the state-of-the-art algorithms for estimating s-t reliability on uncertain graphs. We conducted throughout experimental analyses for these algorithms, summarized their trade-offs, and provided guidance for future researchers and practitioners. In the second part of this talk, I will present two novel problems relevant to the reliability on uncertain graphs, including budgeted reliability maximization on uncertain graphs and jointly finding influential users and topics for social advertising. We designed efficient and scalable algorithms for each of them, and provided rich real-world case studies to demonstrate their effectiveness.

Martina Patone (University of Southampton, UK):
"Graph sampling and estimation"

Abstract: Finite graph sampling in an emerging topic area for design-based theory. This talk explores new developments on the areas of graph sampling and estimation. Two main advantages of using the graph representation of the data in the sampling process are shown: the possibility of enlarging the initial sample by means of different observation procedures and the variety of target parameters, which makes feasible to investigate the structural properties of the graph. As a framework of estimation for graph sampling, we define a bipartite incidence graph (BIG), which allows the construction of several linear design unbiased estimators and includes the HT estimators as a particular case. Finally, the incidence structure of the graph is explored as a form of auxiliary information, which can improve the estimation of the parameter of interest. This is a completely unexplored territory, which does not have a corresponding version in list sampling where a structure is clearly lacking.

Alessio Angius (ISI Foundation, Italy):
"Markov Decision Processes applied to Industry 4.0"

Abstract: After a brief introduction to the challenges of Industry 4.0, the talk will focus on three real use cases that have been tackled by using Markov Models. In particular, the talk will describe the following problems:

i) the minimization of the delivery span of batches in a flow shop i.e. a system where the jobs are executed in the same order on N machines. After being processed in the flow shop, the products must undergo an additional process step executed on a batch. The final goal is to minimize the time required to compose the batch.

ii) Minimizing the time to completion of lots in a multi-product system where machines are constrained shared resources with setup and processing times that depend on the product class.

iii) Minimizing the tardiness in a multi-server queue where the scheduling of ordinary production is altered by the stochastic arrival of products requiring maintenance.

The industrial application of the three use cases will be explained. Furthermore, the three mathematical models will be presented as Markov Decision Models in order to suggest future works in the context of reinforcement learning or other machine learning paradigms.