Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D10.6

SoBigData Interest groups report 2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (60 months)
ENDING DATE	31/12/2024
PROJECT WEBSITE	http://www.sobigdata.eu
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042

DELIVERABLE INFORMATION	
WORK PACKAGE	WP10 JRA3 - Exploratories
WORK PACKAGE LEADER	KTH and CNR
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETH Zürich, PSE, UNIROMA1, CNRS, CEU, URV, CSD, BSC, UPF, Eli, CRA, UvA
DELIVERABLE NUMBER	D10.6
DELIVERABLE TITLE	SoBigData Interest groups report 2
AUTHOR(S)	Luca Pappalardo (ISTI-CNR), Aris Gionis (KTH)
CONTRIBUTOR(S)	
EDITOR(S)	Valerio Grossi (CNR)
REVIEWER(S)	Michela Natilli (CNR), Ilaria Barsanti (CNR)
CONTRACTUAL DELIVERY DATE	31/12/2022
ACTUAL DELIVERY DATE	26/04/2023
VERSION	1.1
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	13
KEYWORDS	Exploratory, medicine, health

EXECUTIVE SUMMARY

The deliverable updates deliverable D10.5 “SoBigData Interest groups report 1” and contains the activities in the interest groups, reporting the creation of new ones and the status of the resources available.

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

This document contains information on SoBigData++ core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData++ Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The content of this publication is the sole responsibility of the SoBigData++ Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

GLOSSARY

AI	Artificial Intelligence
EC	European Commission
EU	European Union
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
MP	Micro-projects

TABLE OF CONTENTS

1 Relevance to SoBigData++

1.1 Purpose of this document

This document describes: (i) the activity carried out within the SoBigData++ interest groups during the second reporting period; and (ii) the report about the creation of a new exploratory, the "Network Medicine" exploratory. For each interest group, we report the results achieved and the activities carried out in terms of conferences/workshops, hackathons, data collection, and software development. Interest groups are possible future exploratories which will be investigated by the consortium to understand if there are interests and experiences which may be transformed in services. Those interest groups organise meetings with experts in the field, researchers and industries to eventually become exploratories in SoBigData++.

1.2 Relevance to project objectives

This document updates deliverable D10.5 "SoBigData Interest groups report 1"[1] and it is related to the activities for investigating and defining new exploratories inside SoBigData RI

1.3 Relation to other work packages

Since in the document we also describe some activities made or planned for the next period, this deliverable is also related to work packages WP3 - Dissemination, Impact, and Sustainability (because of workshops and conferences have been made or planned), WP4 - Training (because hackathons have been made or planned), and WP7 - Virtual Access (because data sets and software have been made available on the infrastructure or planned).

1.4 Structure of the document

Section 2 outlines all the activities performed for the definition and release of the new Network Medicine exploratory, while Section 3 reports some consideration about the creation of new exploratories.

[1] https://data.d4science.net/tnvY

2 Network Medicine Exploratory

2.1 Activities Report

Network and Sequence-Based Prediction of Protein-Protein Interactions
Partners Involved: UNIROMA1

Typically, proteins perform key biological functions by interacting with each other. As a consequence, predicting which protein pairs interact is a fundamental problem. Experimental methods are slow, expensive, and may be error prone. Many computational methods have been proposed to identify candidate interacting pairs. When accurate, they can serve as an inexpensive, preliminary filtering stage, to be followed by downstream experimental validation. Among such methods, sequence-based ones are very promising. In this activity, we designed a new algorithm that leverages both topological and biological information to predict protein-protein interactions. Preliminary results comparing our Framework with state-of-the-art approaches on reliable PPIs datasets, indicate that they have competitive or higher accuracy on biologically validated test sets. We claim that topological plus sequence-based computational methods can effectively predict the entire human interactome compared with methods that leverage only one source of biological information. Future work will evaluate the effectiveness of our hybrid approach in a comprehensive fashion.

Discovering epistatic interactions via neural network interpretability
Partners Involved: UNIROMA1

Epistatic interactions (EIs) of gene loci often determine complex trait phenotypes. EIs may indicate the underlying molecular mechanisms of multifactorial traits and diseases. Yet, the inference of EIs as well as EI-based gene–gene networks remain a great challenge. Neural networks have become very successful recently into classifying complex data, having revolutionised various fields; yet their complexity does not reveal how they combine the input features and this lack of explainability limits their use on genetic data. In this activity we designed EpiCID, a framework based on neural networks for discovering complicated interactions between input features (SNPs in our setting). In our future evaluation we plan to apply it on heart-related measurements, such as systolic and diastolic blood pressure.

Therapy evaluation of type-2 diabetes
Partners Involved: UNIROMA1

Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion/action. In 2019: 9.3% of adult population suffers from diabetes and diabetes was the 9th leading cause of death. By far, the most common diabetes is type-2 diabetes. It has multiple causes, and as a result there exist many potential therapies: more than 400 active ingredients exist in the market. Current medical approaches for addressing type-2 diabetes are based, to a large extent, on trial and error.

The goal of this project is to use historic data to recommend therapies, especially when a (potentially expensive therapy) has high chances to fail. To this end, we have prepared and preprocessed a large dataset of diabetes patients and, as a preliminary analysis, we have tested various graph neural network approaches for predicting therapies recommended by doctors.

Next steps include the design of GNN-based approaches for the prediction of a therapy failure.

Disease-gene identification with positive-unlabelled learning
Partners Involved: UNIROMA1

Gene-disease associations are fundamental for the understanding of disease mechanisms and for the development of effective interventions and treatments. Identifying genes not yet associated with a disease due to lack of studies is a challenging task in which prioritisation based on prior knowledge can be helpful. The computational search for new candidate disease genes may be eased by Positive-Unlabelled (PU) learning, the machine learning (ML) setting in which only a subset of instances are labelled as positive, while the rest of the data set is unlabelled. In this work, we propose a set of effective network-based features to be used in a novel Markov diffusion-based multi-class labelling strategy for putative disease gene discovery. The performances of the new labelling algorithm and the effectiveness of the proposed features have been tested on five different disease datasets using three ML algorithms. Such features have been compared against classical topological and functional/ontological features showing that they outperform the classical ones both in binary classification and in the multi-class labelling. Analogously, the predictive power of the integrated methodology in searching new disease genes has been found to be competitive against the state-of-the-art algorithms.

Prediction of thyroid-cancer recurrence
Partners Involved: UNIROMA1

The risk stratification of patients with differentiated thyroid cancer (DTC) is crucial in clinical decision making. The most used tool is included in the American Thyroid Association (ATA) Guidelines. Recent research focused on the inclusion of novel features or questioned the relevance of the current ones. This project aims to develop a comprehensive, data-driven prediction model, able to capture all available features, determining the weight of the predictors. To this end, we have collected a dataset of thyroid patients from 40 Italian centres, and we have used it to compare a variety of algorithms, such as decision and boosted trees. The next goal is to explore these experimental findings for identifying a method that is able to make accurate and explainable predictions.

Prioritisation of Gene-Disease Associations via Graph Neural Networks Explanations
Partners Involved: UNIROMA1

During the recent years, many powerful techniques have been developed to find Gene-Disease Associations (GDAs) using different approaches, such as text mining of a specific disease's literature or by employing Machine Learning (ML) algorithms, exploiting a wide variety of models, training datasets and feature extraction mechanisms. However, experimental validation of candidate genes is an expensive and resource-intensive task. Therefore, it is important to develop computational approaches able to reduce the search space by focusing expensive laboratory tests only on the most promising candidates. The task of discovering new GDAs is strictly under the responsibility of analytical laboratories that can validate the results through expensive experiments. Because it is too resource-consuming to analyse all the possible associations between all known genes with all known diseases, computational methods can help medical research by providing a ranking of the most likely associations, to help focusing limited resources on the most promising genes. The goal of this activity is to take advantage of the natural advantage of GNNs in capturing hidden information in biological networks and to apply it to protein-protein interaction networks to generate a ranking of probable new GDAs. Currently we have performed experiments to evaluate the performance of various GNN approaches. Future tasks involve the use of explainable AI approaches for prioritising the genes correlated with various diseases.

Partial Correlation for Functional COPD Subnetwork Genes Discovery
Partners Involved: UNIROMA1

Chronic obstructive pulmonary disease (COPD) is a complex disease influenced by environmental exposures (most notably, cigarette smoking) and genetic factors. Genome-wide association studies have identified thousands of genomic regions associated with complex diseases. The chromosome 4q region harbours multiple genetic risk loci for chronic obstructive pulmonary disease (COPD). To determine whether genes in this region are part of a gene expression network, we studied lung tissue RNA-Seq from COPD cases and controls. It is likely that the effects of genetic variants in complex diseases cannot be captured by a single type of omics data. However, including different biological data sources may facilitate the removal of indirect effects, allowing the identification of correlations that would not be detected otherwise. We leveraged protein-protein interaction information to build a partial correlation network, controlling for all of the genes in the genome while assessing for the correlation between pairs of genes.

2.2 Micro-projects

Discovering side effects of drugs through denoised GCN
Status: Suspended
Partner Involved: UNIROMA1, UT
Motivation of suspension: the MP was related to modelling polypharmacy side effects through graph convolutional networks. Its goal was to extend polypharmacy methodology by applying denoising techniques to GCN models to improve link prediction accuracy, aiming to discover unknown polypharmacy side effects. Unfortunately, UNIROMA1 had to abandon this MP (at least for now) due to insufficient data availability.

Predicting Tyroid Cancer Recurrence
Status: Ongoing
Partners involved: UNIROMA1
Outcomes: Blog post: not available yet
Dataset: not available yet

Gene expression Partial Correlation in chromosome 4 COPD patients
Status: Ongoing
Partners involved: UNIROMA1
Outcomes: Blog post: not available yet
Dataset: not available yet

Exploration of a Hybrid approach for Prediction of Protein-Protein Interactions
Status: Ongoing
Partners involved: UNIROMA1
Outcomes: Blog post: not available yet
Dataset: not available yet

Design of topological and biological approaches for Prediction of Protein-Protein Interactions
Status: Completed
Partners involved: UNIROMA1
Outcomes: Method
Blog post

Discovering epistatic interactions via neural network interpretability
Status: Completed
Partners involved: UNIROMA1
Outcomes: Method

Preparation of a large-scale dataset for therapy recommendation for type-2 diabetes
Status: Completed
Partners involved: UNIROMA1
Outcomes: Dataset

Comparison of approaches for type-2 diabetes patients
Status: Completed
Partners involved: UNIROMA1
Outcomes: Experiment

Disease-gene identification with positive-unlabeled learning
Status: Completed
Partners involved: UNIROMA1
Outcomes: Research paper

Comparison of approaches for predicting thyroid-cancer recurrence
Status: Completed
Partners involved: UNIROMA1
Outcomes: Experiment

Disease Genes Discovery with Explainable Graph Neural Networks
Status: Completed
Partners involved: UNIROMA1
Outcomes: Experiment

2.3 Publications

1. *Andrea Mastropietro, Giuseppe Pasculli, and Jürgen Bajorath*. "Protocol to explain graph neural network predictions using an edge-centric Shapley value-based approach." STAR protocols (2022). Accepted
2. *Riccardo De Luca, Marco Carfora, Gonzalo Blanco, Andrea Mastropietro, Manuela Petti, and Paolo Tieri*. "PROCONSUL: PRObabilistic exploration of CONnectivity Significance patterns for disease modULe discovery". IEEE BIBM 2022 Network Science and Artificial Intelligence for Biomedicine & Health informatics Workshop (2022). Accepted.

2.4 Planned Activities for next period

We have collected some datasets for medicine related purposes:

- type-2 diabetes
- thyroid cancer
- protein-protein interactions
- Genome-wide association studies

Because the data are sensitive, the data are available to SoBigData++ partners on site.

2.4.1 Data Collection activities

Partners involved: UNIROMA1

We have developed code for the “Disease-gene identification with positive-unlabeled learning,” which is available online.

2.4.2 Software Development Activities

Partners involved: UNIROMA1

We have developed code for the “Disease-gene identification with positive-unlabeled learning,” which is available online.

2.4.3 Scientific Activities

Partners involved: UNIROMA1

In the next period, we plan to perform experiments to evaluate our methods described in the Activities section and to prepare scientific publications. Please refer to the individual project for the future activities planned.

3 Conclusions - Next exploratories creation

This document reported the activities in the task T10.7 “SoBigData Interest groups interest groups”, reporting the creation and consolidation of the networking medicine exploratory. At the moment, the group did not plan the creation of new exploratory yet.