University of Padua

Doctorate Degree in Brain, Mind and Computer Science

Curriculum in Computer Science

KERNEL METHODS FOR LARGE-SCALE GRAPH-BASED HETEROGENEOUS BIOLOGICAL DATA INTEGRATION

Candidate
Dinh Tran Van

Supervisor
Prof. Alessandro Sperduti

Co-supervisor
Prof. Franca Stablum
University of Padova, Italy

October 30, 2017

Acknowledgments

First, I would like to express my sincere gratitude to my advisor Prof. Alessandro Sperduti for his invaluable insights, guidance and encouragement. I have been extremely lucky to have his continuous support despite his busy schedule.

Second, I would like to thank to my co-advisor Prof. Franca Stablum for her support, especially in scientific writing.

I also would like to thank Prof. Fabrizio Costa for his guidance that has been given to me since I visited the university of Freiburg.

I would like to thank all the members of BioInfogen project (Prof. Giorgio Valle, Prof. Fabio aiolli, Nicolo Navarin, Guido, etc) for the valuable discussions during our meetings.

I would like to thank the rest of my thesis committee for their insightful and valuable comments and suggestions.

I would like to thank all my friends, colleagues and officemates (Reza Khosh Kangini, Merylin Monaro, Mirko Polato, Li QianQian, Riccardo Spolaor, Kiyana Bahadori, Muhammad Hassan Raza Khan, Rabbani MD Masoom, Yan Hu, Giuseppe Cascavilla) here at the University of Padova for the stimulating discussions, and for all the fun we have had in the last three years.

Last but not least, a special thank goes to my family, that always supported me in these years.

Dinh Tran Van
Padova, October 30, 2017

Abstract

The last decade has been experiencing a rapid growth in volume and diversity of biological data, thanks to the development of high-throughput technologies related to web services and embeded systems. It is common that information related to a given biological phenomenon is encoded in multiple data sources. On the one hand, this provides a great opportunity for biologists and data scientists to have more unified views about phenomenon of interest. On the other hand, this presents challenges for scientists to find optimal ways in order to wisely extract knowledge from such huge amount of data which normally cannot be done without the help of automated learning systems. Therefore, there is a high need of developing smart learning systems, whose input as set of multiple sources, to support experts to form and assess hypotheses in biology and medicine. In these systems, the problem of combining multiple data sources or data integration needs to be efficiently solved to achieve high performances.

Biological data can naturally be represented as graphs. By taking graphs for data representation, we can take advantages from the access to a solid and principled mathematical framework for graphs, and the problem of data integration becomes graph-based integration. In recent years, the machine learning community has witnessed the tremendous growth in the development of kernel-based learning algorithms. Kernel methods whose kernel functions allow to separate between the representation of the data and the general learning algorithm. Interestingly, kernel representation can be applied to any type of data, including trees, graphs, vectors, etc. For this reason, kernel methods are a reasonable and logical choice for graph-based inference systems. However, there is a number of challenges for graph-based systems using kernel methods need to be effectively solved, including definition of node similarity measure, graph sparsity, scalability, efficiency, integration methods. The contributions of the thesis aim at investigating to propose solutions that overcome the challenges faced when constructing graph-based data integration learning systems.

The first contribution of the thesis is the definition of a novel decompositional graph node kernel, named conjunctive disjunctive node kernel (CDNK), to measure graph node similarities. Differently of existing graph node kernels that only exploit the topologies of graphs, the proposed kernel also utilizes the available information on the graph nodes. In CDNK, first the graph is transformed into a set of linked connected components in which we distinguish between "conjunctive" links whose endpoints are in the same connected components and "disjunctive" links that connect nodes located in different connected components. We then propose a graph node kernel that explicitly models the configuration of each nodes context. Empirical evaluation shows that the kernel presents better performance compared to state-of-the-art graph node kernels.

The second contribution aims at dealing with the graph sparsity problem. When working with sparse graphs, i.e graphs with a high number of missing links, the available information is not efficient to learn effectively. An idea is to overcome this problem is to enrich information for graphs by using link enrichment. Moreover, the performance of a link enrichment strongly depends on link prediction. Therefore, we propose an effective link prediction method (JNSL). In this method, first each link is represented as a joint neighborhood subgraphs. Then link prediction is considered as a binary classification. The proposed link prediction outperforms various other methods. Besides, we also present a method to boost the performance of diffusion-based kernels by coupling kernel methods with link enrichment. Evaluation on different datasets shows that performing link enrichment before applying diffusion-based kernels helps to considerably improve their performances.

The last contribution proposes a general kernel-based framework for graph integration that we name Graph-one. Graph-one is designed to overcome the challenges when handling with graph integration. In particular, it is a scalable and efficient framework. Besides, it is able to deal with unbanlanced settings where the number of positive and negative instances are much different. Different variations of Graph-one are evaluated in disease gene prioritization context. The results from experiments illustrate the power of the proposed framework. More precisely, Graph-one shows better performance than various methods. Moreover, Graph-one with data integration get higher results than it with any single data source. It presents the effectiveness of Graph-one in exploiting the complementary property of graph integration.

Contents

Chapter 1

Introduction

The release of advanced technologies is one of the main reasons for the revolution in various scientific research fields. In Biological and Medical domain, modern technologies are making it not only easier but also more economical than ever to undertake experiments and creating applications. As a consequence, a vast amount of biological data in terms of volume and type is generated through scientific experiments, published literatures, high-throughput experiment technologies, and computational analysis. This huge quantity of data are saved as biological datasets and made discoverable through web browsers, application programming interfaces, scalable search technology and extensive cross-referencing between databases. Biological databases normally contain information about gene function, structure, localization, clinical effects of mutations and similarities of biological sequences and structures.

The abundance of biological data, on the one hand, creates a golden chance for biologists to extract useful information. However, it, on the other hand, poses the challenge for scientists to wisely and effectively extract knowledge from such amount of data that normally cannot be done without the help of automated learning systems. Hence, the task of developing high performance learning systems, which help sciencists to form and assess hypotheses, plays an important role in the development of biology and medicine.

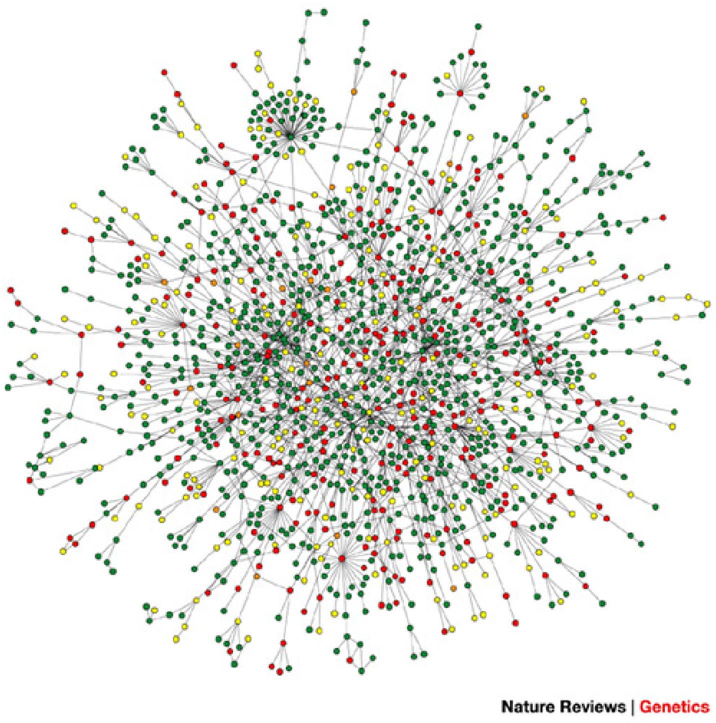

Figure 1.1: Yeast protein interaction network [2].

1.1 Why graph-based biological data integration?

Biological knowledge is distributed among general and specialized sources, such as gene expression, protein interaction, gene ontology, etc. It is common that information of a biological phenomenon is encoded over various heterogeneous sources. Each source captures different aspects of the phenomenon. The distribution of information over sources provides us an unprecedented opportunity to understand the phenomenon from multiple angles.

Therefore, the idea of data integration which allows multiple sources of information to be treated in a unified way can in priciple lead to an improvement of biological learning systems, i.e. systems that process with biological data. Despite the fact that data integration is a promising solution, it poses a challenge for machine learning experts and data scientists to find out optimal solutions for combining multiple sources in a big space of solutions.

Relations between entities encoded in biological sources can be naturally represented in form of graphs (networks) whose vertices describe for biological entities and links characterize the relations between entities. An example is the protein-protein interaction network (Figure 1.1) where each vertex represents a protein and a link connecting two vertices if they interact. Graph theory provides a mathematical abstraction for the description of such relationships. Thus, using graphs to represent for biological data allows us to i) access to a principled and solid mathematical framework built for graphs that most scientists are familiar with, ii) develop concepts and tools which are independent of the concrete applications. By using a graph presentation, the problem of biological data integration now can be converted into graph-based data integration. The final representation of data obtained by integration is used as input for the construction of inference systems (graph-based learning systems).

1.2 Kernel methods for graph-based data integration and challenges

Kernel methods whose best known member is support vector machine (SVM) [28], has emerged as one of the most famous and powerful frameworks in machine learning. Kernel methods with the use of kernels allow to decouple the representation of the data (via kernel function) from the specific learning algorithm. Kernel representation is flexible and efficient and it provides a principled framework that allows universal type of data to be represented, including images, graphs, vectors, strings, etc. As a consequence, kernel methods are the state-of-the-art learning technique for graph-based inference systems. However, there are a number of challenges that need to be efficiently solved, if we desire to have high performance graph-based data integration learning systems. Following are the main challenges: definition of node similarity measure, graph sparsity, data integration method.

1.2.1 Definition of node similarity measure

In machine learning, one of the main factors which impacts on the performance of learning systems is the definition of example similarity measure. In our context, large-scale graph-based inference systems, examples are nodes of graphs. Hence, it is necessary to have a good definition of node similarity measure. Node similarity is normally measured by *graph node kernels*. However, there is not a clear way to define a graph node kernel which can be efficiently applied to a wide range of graphs.

1.2.2 Graph sparsity

The input of a graph-based data integration system is a set of graphs which often contain sparse graphs whose number of links is much less than the number of possible links. This is typically due to the lack of information. For instance, in the disease gene network, links connecting genes are formed when genes are involved in the same diseases. However, new genes associated to a certain disease could be discovered over time. This means that new links could be added into the networks over time. At a given time, a number of discovered links can be very limited, so discovered links cause the sparsity problem. When working with sparse graphs, systems encounter difficulties in performing an effective training since not enough information is available to correctly learn the target function. As a consequence, an effective solution helping to overcome the sparsity problem is crucial and needs to be proposed.

1.2.3 Scalability and efficiency

Given an adopted learning algorithm, the complexity of a large-scale graph-based data integration learning system incurs with the growth of the input graph set. In other words, the complexity of a graph-based learning system strongly depends on the size and the number of graphs used as its input. Therefore, scalability is an important property that a graph-based learning system is supposed to possess. It allows systems to run in reasonable time and with a reasonable memory consumption.

1.2.4 Data integration methods

Information encoded in multiple sources (graphs) provide complementary views of the phenomenon of interest. Combining information from collective sources helps to form a complete picture of the phenomenon or problem at hand. Nonetheless, the search for efficient and scalable integration methods that allow to improve the performance of the learning system with respect to the same system where a single source of information is used, is normally expensive.

1.3 Contributions

The contributions of the thesis focus on solutions to overcome the challenges faced when working with large-scale graph-based biological data integration.

The first contribution considers the problem of defining an effective node similarity measure by introducing a novel graph node kernel, named conjunctive disjunctive node kernel (CDNK). Most existing graph node kernels are based on a notion of information diffusion which can be applied to dense networks with high values of average node degree. However, a drawback of these approaches is their relatively low discriminative capacity. This is in part due to the fact that information is processed in an additive and independent fashion which prevents them from accurately modeling the configuration of each node context. To address this issue, we propose to employ a decompositional graph kernel technique in which the similarity function between graphs can be formed by decomposing each graph into subgraphs and by devising a valid local kernel between the subgraphs. In CDNK, to exploit its higher discriminative capacity, first the network is decomposed into a collection of connected sparse graphs and then a suitable kernel is developed.

The second contribution is the introduction of a link prediction method, which is adopted later on for link enrichment with the aim of solving the problem of graph sparsity. We get the motivation from the current link prediction methods that do not effectively exploit the contextual information available in the neighborhood of each edge. In our method, we propose to cast the problem as a binary classification task over the union of the pairs of subgraphs located at the endpoints of each edge. We model the classification task using a support vector machine endowed with an efficient graph kernel and achieve state-of-the-art results on several benchmark datasets.

The third contribution proposes a method that boosts performance of diffusion-based kernels when working with sparse graphs by tackling them with link enrichment methods. In particular, given a sparse graph, our proposed method consists of two phases. In the first phase, a link prediction method is employed to rank unobserved links based on their probabilities to be related to missing links. The top links in the ranking are then added into the graph. In the second phase, diffusion-based graph node kernels are applied to the graph obtained from the first phase to compute the kernel matrix.

The last contribution presents a general framework, named Graph-one, for graph integration. Graph-one is an efficient and scalable kernel-based framework which is able to deal with the unbalanced settings. We evaluate

Graph-one by introducing its different variations (Scuba, PLC, DIGI) for disease gene prioritization. Experimental results illustrate that *i*) Graph-one outperforms various methods, and *ii*) Graph-one with data integration shows better performance than it with any single data source.

1.4 Thesis roadmap

The thesis is organized as follows:

Chapter 2 presents preliminary concepts, notations and comprehensive review of the state-of-the-art in the field.

Chapter 3 proposes an effective convolutional graph node kernel, Conjunctive Disjunctive Graph Node Kernel (CDNK).

Chapter 4 introduces solutions to solve the graph sparsity problem: a novel link prediction method (JNSL) and a method to boost the performance of diffusion-based kernels when working with sparse graphs.

Chapter 5 describes a general kernel-based framework for graph integration.

Chapter 6 summarizes the contributions of the thesis and discusses the directions for future work.

1.5 List of publications

- Dinh Tran Van, Alessandro Sperduti and Fabrizio Costa, Conjunctive Disjunctive Node Kernel, the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 26-28 April, 2017, ISBN 978-287587039-1

- Dinh Tran Van, Alessandro Sperduti and Fabrizio Costa, Link Enrichment for Diffusion-based Graph Node Kernels, the 26th International Conference on Artificial Neural Networks, Alghero, Italy, 11-15, Steptember, 2017

- Dinh Tran Van, Alessandro Sperduti and Fabrizio Costa, Joint Neighborhood Subgraphs Link Prediction, the 24th International Conference on Neural Information Processing, Guangzhou, China, November 1418, 2017

- Guido Zampieri, Dinh Van Tran, Michele Donini, Nicol Navarin, Fabio Aiolli, Alessandro Sperduti and Giorgio Valle, Scuba: scalable kernel-based gene prioritization, BMC Bioinformatics Journal (under revision)

- Dinh Tran Van, Alessandro Sperduti and Fabrizio Costa , Conjunctive Disjunctive Node Kernel, Neurocomputing Journal (Selected paper from the ESANN 2017 for an extension that will be published in a special issue of the Neurocomputing journal) (under revision)

Chapter 2

Background

In this chapter, we describe preliminary knowledge and notaions used for the remaining parts of the thesis to make it easy for readers to follow the exposition of its original contributions.

2.1 Machine learning

Recently, *machine learning* has become a must-know term not only in academia but also in daily life due to the popularity of it's applications in various fields. Machine learning can be considered as a branch of Artificial Intelligence which aims at providing systems the ability to automatically adapt to their environment and learn from experience without being explicitly programmed. According to [59], machine learning is formally defined as:

Definition 2.1.1. *A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E.*

We denote $\mathbb{D}$ as a set of training examples which come from some generally unknown probability distribution. $\mathbb{D}$ is resulted from any observation, measurement or recording apparatus for a certain domain A machine learning technique aims at exploiting $\mathbb{D}$ to build a general model about the example space. This model is then used to produce sufficiently accurate predictions in unseen examples.

Machine learning algorithms can be classified into three paradigms: supervised learning, unsupervised learning and reinforcement learning. Supervised learning is the machine learning task of inferring a function from labeled training examples in which each example is a pair consisting of an entity and a desired output value (label). A supervised learning algorithm analyzes the training data and forms an inferred function, which is used for mapping unseen examples. Unsupervised learning is a machine learning task that models a set of inputs where labeled examples are not available. Reinforcement Learning aims at designing machines and software agents that can automatically determine the ideal behaviour within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behaviour; this is known as the reinforcement signal. In this thesis, we focus on supervised learning scenario.

We consider a training set $\mathbb{D}$ generated by an unknown probability distribution $\mathcal{P}$, $\mathbb{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{X}$ are instances and $y_i \in \mathbb{Y}$ are labels. The relations between x_i and y_i are defined by a true function (target function) $f : \mathbb{X} \longmapsto \mathbb{Y}$. What we desire to do is to learn the function f. However, the only information we can access is from the training set. Therefore, a supervised learning method aims at estimating a function h based on $\mathbb{D}$ to be as close to f as possible. Depending on the domain of $\mathbb{Y}$, we can further group supervised learning into the following sub-groups:

- if $\mathbb{Y} \subseteq \mathbb{R}$, the problem is called regression;
- if $|\mathbb{Y}| = 2$, we have a binary classification problem;
- if $|\mathbb{Y}| = n$ with $n > 2$, we have multi-class classification problem.

Besides, a multi-class learning is called multi-label if examples have more than one label associated with. It is worth highlighting that there is normally more than one possible choice for h. We refer each choice of h as a hypothesis or model and the set of all possible h as hypothesis space, $\mathbb{H}$. The goal of the learning process is to find the final hypothesis that best approximates the unknown target function. In order to measure the difference between a hypothesis and the target function, the risk function is used:

$$\mathcal{R}(h) = \int_{\mathbb{X}\times\mathbb{Y}} \mathcal{L}(h(x), y) dP(x, y), \tag{2.1}$$

where $\mathcal{L}$ is a loss function that measures the classification error of h. An example of the loss function for classification can be defined as:

$$\mathcal{L}(h(x), y) = (h(x) - y)^2.$$

The optimal hypothesis is the one which minimizes the risk and it is the solution of the following optimization problem:

$$h^* = \arg\min_{h \in \mathbb{H}} \mathcal{R}(h). \tag{2.2}$$

Unfortunately, it is impossible to directly solve the optimization problem described in eq. 2.1 since the probability distribution $\mathcal{P}$ in the true loss function presented in eq. 2.2 is an unknown function and we only have access to a finite training set $\mathbb{D}$. In this case, an alternative approach is to use the empirical loss instead of the true loss function. The empirical loss function is defined over the training set as follows:

$$\mathcal{R}_{emp}(h) = \frac{1}{n}\sum_{i=1}^{n} |h(x_i) - y_i|, \tag{2.3}$$

where n is the number of traning examples. However, in order to use $\mathcal{R}_{emp}(h)$, we need to guarantee that the value of $\mathcal{R}_{emp}(h)$ converges to the value of $\mathcal{R}(h)$.

In the typical statistical learning framework, every data instance is embedded in a suitable space. However, most real world data has no natural representation as vectorial forms. Kernel methods have been successful in various learning tasks on data represented by vectors, but structured forms, including graphs. In this thesis, we are interested in investigating graph-based integration methods. Therefore, in the next section we describe Kernel methods.

2.2 Kernel methods

In classical machine learning techniques, each data instance is mapped to a point in the feature space, $x \in \mathbb{X} \longrightarrow \phi(x) \in \mathbb{F}$. Then a model is constructed from the training set and is used to predict for unseen data. Although these approaches have sucessfully applied in some cases, they share two common limitations: i) If the dimension of the feature space is high, it leads to high complexity algorithms. ii) It is difficult or even impossible in some cases to find the mapping.

Kernel method has been proposed and shown the state-of-the-art results in many cases of various fields. SVM [28] is a typical example of Kernel Methods. Unlike the presentation of data in traditional machine learning, in Kernel methods, data instances are not individually represented in the feature space, instead they are represented by similarity measures between pairs of instance images, which are computed by using dot product. The dot product of instance image pairs can be computed through input instances only by kernel functions. Kernel functions enable Kernel methods to enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data instances in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors. A kernel methods can be modularized into two components: the design of a specific kernel function and the design of a general learning algorithm (kernel machine).

2.3 Kernel functions

In kernel methods, the definition of kernel functions is independent from the definition of general learning algorithms. Therefore, a given generl learning algorithm can go with any kinds of kernel functions. A number of kernel functions have been proposed for different types of data. In this section, we first formally define what is a kernel function. We then introduce some kernels defined on graphs that later on are used in our experiments.

Definition 2.3.1. *Given a set of entities $\mathbb{X}$, a function $k : \mathbb{X} \times \mathbb{X} \longmapsto \mathbb{R}$ is called a kernel on $\mathbb{X} \times \mathbb{X}$ iff k is*

- *symmetric: it means $k(x_1, x_2) = k(x_2, x_1)$, where $x_1, x_2 \in \mathbb{X}$.*
- *positive semi-definite: that is $\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(x_i, x_j) \geq 0$ for any $N > 0$, $c_i, c_j \in \mathbb{R}$, and $x_i, x_j \in \mathbb{X}$.*

The similarity measures computed by a kernel over a set of input instances can be represented in a matrix, called Gram matrix, K. K is symmetric and positive semi-definite, i.e. its eigenvalues are non-negative.

$$K = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_3) & \dots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & k(x_2,x_3) & \dots & k(x_2,x_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & k(x_n,x_3) & \dots & k(x_n,x_n) \end{bmatrix}.$$

The simplist kernel is Linear kernel which is defined on vectors, $\mathbb{X} \subseteq \mathbb{R}^n$:

$$k_L(x_1, x_2) = x_1^\intercal x_2, \tag{2.4}$$

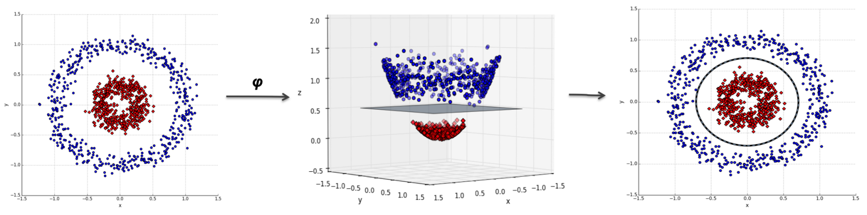

Figure 2.1: The kernel trick transforms the data in a feature space where the instances from the two classes may be linearly separable.

where $x_1, x_2 \in \mathbb{X}$. This kernel suggests a systematic way to define kernels. Given a general set of representations of entities $\mathbb{X}$, we first project each element in $\mathbb{X}$ into a vector space, called feature space, $x \in \mathbb{X} \longrightarrow \phi(x) \in \mathbb{R}^m$ such that $m \gg n$. This is due to that fact that it is easier to find a linear decision line in higer dimensional space (m). Next, we define a kernel as:

$$k(x_1, x_2) = \phi(x_1)^\intercal \phi(x_2) \tag{2.5}$$

Interestingly, any kernel defined on $\mathbb{X}$, there exists a Hilbert space, $\mathbb{F}$, and a mapping $\phi : \mathbb{X} \longrightarrow \mathbb{F}$ such that $k(x_1, x_2) = \phi(x_1)^\intercal \phi(x_2)$, where $x_1, x_2 \in \mathbb{X}$.

There are two problems we might face with if we would like to explicitly embed objects into a vector space. *i*) if m is too big, we face with the high computation. *ii*) if data instances are in the structured forms (strings, graphs, trees, etc), we need to transform them into vectorial forms. One way is to decompose each instance into a set of sub-structures which are considered as elements of vectors. However, in general, there is not a clear way to projecct instacnes in structured forms into feature space without loosing much information. These limitations are effectively solved by using the so called *Kernel trick*. The kernel trick (see Figure 2.1 for an illustration) avoids the explicit mapping. Instead, it allows the operations (dot product) between vectors in the feature space to be done by computing in the input space. It is worth to notice that a kernel is considered as a similarity (proximity) measure since its value computed for two objects is proportional to their similarity.

Most kernels are defined on vectorial form of data among which Basis Function kernel (RBF) [81] is the most used one. However, real-world data often cannot be represented in the vectorial form without loosing important information. Therefore, a high number of kernels are proposed to deal with structured data, including trees, graphs, etc.

Convolution kernels

On the development of kernels for structured data, R-convolution kernels originally proposed in [40] and a generalization of the framework is proposed in [74] can be considered as one of the most important frameworks. The basic idea of convolution kernels is that the semantics of composite entities can often be captured by a relation R between the entity and its parts. The kernel between entities is then made up from kernels defined on different parts.

Let x be a composite structure whose $x_1, x_2, \ldots, x_N = \hat{x}$ are parts of x, such that $x \in \mathbb{X},\ x_i \in \mathbb{X}_i,\ i = \overline{1, N}$ and $\mathbb{X}, \mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_N$ are non-empty and separable metric spaces. We define a relation $R(\hat{x}, x)$ on $\mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_N \times \mathbb{X}$ is true iff $x_1, x_2, \ldots, x_N$ are the parts of x. We denote with R^{-1} the inverse relation of R and it is defined as $R^{-1}(x) = \{\hat{x} | R(\hat{x}, x)\}$.

If there exists kernel k_i defined on X_i, the similarity between $x, y \in \mathbb{X}$ is defined on $\mathbb{S} \times \mathbb{S}$, where $\mathbb{S} = \{x | R^{-1}(x) \neq \emptyset\}$, as:

$$K(x, y) = \sum_{\hat{x} \in R^1(x),\ \hat{y} \in R^1(y)} \prod_{i=1}^{N} k_i(x_i, y_i) \tag{2.6}$$

K is referred as finite convolution, if R is finite. The zero expansion of K to $\mathbb{X} \times \mathbb{X}$ is called R-convolution and it is denoted as $K_1 \star K_2 \star \ldots K_N(x, y)$.

Theorem 1. *If $K_1, K_2, \ldots, K_N$ are kernels on $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_N$, respectively, and R is a finite relation on $\mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_N$, then $K_1 \star K_2 \star \ldots K_N(x, y)$ is a kernel on $\mathbb{X} \times \mathbb{X}$.*

Constructing kernels

Kernels can be constructed from predefined kernels. Let k_1, k_2 be kernels over $\mathbb{X} \times \mathbb{X}$, $\mathbb{X} \subseteq \mathbb{R}^n$, $\alpha_1, \alpha_2 \in \mathbb{R}^+$, $f(.)$ a real valued function on $\mathbb{X}$, ϕ a mapping $\mathbb{X} \longmapsto \mathbb{R}^N$, k_3 a kernel over $\mathbb{R}^N \times \mathbb{R}^N$, and $\mathbf{B}_{n \times n}$ is a symmetric, positive demi-definite. The following functions are kernels:

- $k(x, y) = \alpha_1 k_1(x, y) + \alpha_2 k_2(x, y)$
- $k(x, y) = k_1(x, y) k_2(x, y)$
- $k(x, y) = f(x) f(y)$
- $k(x, y) = k_3(\phi(x), \phi(y))$
- $k(x, y) = x' B y$

The proof of above kernels and other ways to form kernels from predefined kernels are presented in [71].

2.4 Kernel machines

In machine learning, there is a high number of techniques which aim at finding linear relations in datasets which are represented in vectorical forms. However, in many cases, the expected linear relations do not exist. A solution to overcome these situations is to first explicitly perform a non-linear transformation of input instances into a higher dimensional space and then search for linear relations in that space. Unlike traditional machine methods, kernel methods with the use of kernel functions are able to operate in a high-dimensional space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates.

A number of traditional machine learning methods exploiting dot products for computing can be turned into versions that exploit kernels, i.e. replacing the dot products with kernel computations. Therefore, their expressive power are increased. Examples are Kernel Perceptron, Support Vector Machines (SVM), Kernel Gaussian Processes, Kernel Principal Components Analysis (PCA), etc. In the next sections, we will introduce in detail two famous algorithms: Kernel Perceptron and SVM. For the simplicity, we describe these algorithms in the context of binary classification.

2.4.1 Kernel Perceptron algorithm

Perceptron [13] is an old, online leanring algorithm which is based on error-driven learning. It desires to learn a hyper-plane, $w^\intercal x+b=0$ or $w^\intercal x=0$ for simplicity, to separate positive instances from negative ones in the training set. It is then used to predict a label for each unseen instance, x, through the *sgn* function. If $w^\intercal x \geq 0$, the output of the Perceptron is $\hat{y} = sgn(w^\intercal x) = 1$, otherwise $\hat{y} = -1$.

Perceptron works by first initilizing values for weight vector, w. It then iteratively improves the performance by updating the weight vector whenever a misclassification is found in the training set. Consider y_i and $\hat{y_i}$ as the true label and the predicted label for x_i, respectively, if $y_i \neq \hat{y_i}$, w is updated as follows:

$$w \longleftarrow w + \alpha y_i x_i,$$

where $\alpha \in (0, 1]$ is the learning rate. Suppose that n misclassified examples are observed, the weight vector w can be expressed as:

$$w = \sum_{i=1}^{n} \alpha y_i x_i. \tag{2.7}$$

The update of the weights is done if the current input is not misclassified. This algorithm guarantees that a linear separation is found if it exists. When the linear separation does not exist, a possible solution is to embed input data into a higher dimensional space. By virtue of doing so, there is a higher chance to have linear separation. However, when the algorithm operates in a high dimensional space, it faces with the high complexity. As a consequence, Kernel Perceptron method [14], an extension of original Perceptron, is proposed to cope with high dimensional spaces.

Suppose that $\phi(x)$ is the image of x in feature space. We rewrite eq. 2.7 to compute the weight vector in the feature space as:

$$w = \sum_{i=1}^{n} \alpha y_i \phi(x_i).$$

Then we get

$$sgn(w^\intercal \phi(x)) = \sum_{i=1}^{n} \alpha y_i \phi(x_i)^\intercal \phi(x_j).$$

One limitation of both Perceptron and Kernel Perceptron method is that they are not able to find the optimal linear separation. Normally, among all possible linear separations, there might exist some which show bad predicting ability for unseen data. In the next section, we describe SVM, a kernel method, which aims at finding an optimal hyperplane to separate positive instances from negative ones.

2.4.2 Support Vector Machine

The original Support Vector Machine is a linear classifier and it was invented by Vladimir N. Vapnik [79]. SVM became popular when Vladimir et al introduced in [18] a way to create nonlinear classifiers by employing the notion of kernel trick. In particular, a SVM searches for an optimal hyperplane in the feature space, $\mathcal{H}$, through operations in input space only.

Given a set of training examples $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ in which $x_i \in \mathbb{R}^n$ and $y_i \in \{\pm 1\}$, we first assume that the examples in the training set are linearly separable. SVM tries to learn a decision function

$$f_{w,b}(x) = w^\intercal x + b, \tag{2.8}$$

such that $f_{w,b}(x_i)y_i \geq 0,\ b \in \mathbb{R}$ is the bias, and $w \in \mathbb{R}^n$ is the norm vector.

This function forms two half-spaces of instances: $h^+ = \{x : f(x) \geq 1\}$ and $h^- = \{x : f(x) \leq -1\}$. The distance between these two half-spaces is referred as *margin* and equal to $\frac{2}{\|w\|}$.

The optimal hyperplane is the solution of the below quadratic optimization problem (primal form):

$$\begin{aligned} \underset{w,b}{\text{maximize}} \quad & \frac{1}{\|w\|} \\ \text{subject to} \quad & y_i(w^\intercal x_i + b) \geq 1. \end{aligned} \tag{2.9}$$

It is equivalent to

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i(w^\intercal x_i + 1) \geq 1. \end{aligned} \tag{2.10}$$

It is easier to use the dual form that can be obtained by the introduction of Lagrangian multipliers. Since the the optimization is convex, the solution of dual form is the same as primal form. The resulting Lagrange multiplier equation we desire to optimize is

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(w^\intercal x_i + b) - 1 \right], \tag{2.11}$$

where $\alpha_i \geq 0$ are Lagrange multipliers. Solving Lagrangian optimization eq. 2.11, we obtain values for w, b and α which determin a unique hyperplane. The $x_i s$ corresponding to $\alpha_i s$ which differ from 0 are called support vectors.

The formula of the hyperplane decision function eq. 2.8 can be rewritten as:

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i x^\intercal x_i + b. \tag{2.12}$$

In the case that the training set is not linearly separable, we can apply the kernel trick to let SVM to operate in possibly higher dimensional Hilbert space. By doing so, we hope in a higher dimensional space, there exist a hyperplane that separates images of positive instances from the negative ones.

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i \phi(x)^\intercal \phi(x_i) + b. \tag{2.13}$$

There are usually very few $\alpha_i s$ which are equal to 0. Therefore, it requires a low computation to predict for unseen examples.

In practice, there are two problems that we need to take into account. First, in many cases, the separating hyperplane does not exist in the feature space due to the high level of noise in data. Second, the learning function is so complex that it not only fits instances, but it also fits the noise. Therefore, the function is able to classify the training set, but it fails to generalize for unseen data. The latter problem is called overfitting. In order to solve such problems, a solution one may think is to allow examples to violate eq. **??**.

A soft margin SVM is introduced in which a trade-off between the mistakes on the training set and the complexity of the hypothesis is defined. The optimization eq. 2.10 is modified by introducing slack variables ξ_i:

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & y_i(w^\intercal x_i + 1) \geq 1 - \xi_i, \end{aligned} \tag{2.14}$$

where $\xi_i \geq 0$ and C is a constant which determines the trade-off between margin maximazation and the training error minimization.

2.5 Kernels on graphs

Canonical machine learning methods take vectorial data, data that are represented by vectors of features, as their input. However, there are many fields where data are not naturally represented by vectors, but by structured forms in which graph is one of the most popular representation. Therefore, the task of developing methods which are able to learn from structured data in general or graphs in particular is very important. In this thesis, we focus on graphs, a special type of structured data representation. An example of data that can be represented by graph is the genetic network, where each node represents a gene and each link is formed between two genes, if they encode common protein(s). Another example is the social network whose nodes are users and links depict friendship between users. Systems that deal with problems where data are naturally represented as graphs are called graph-based systems.

One of the key points that determines the performance of a learning system is the similarity measure definition. In our context is the similarity measure definition between graphs. An idea is to find ways that map graphs into vectorial forms: $\mathbb{X} \longmapsto \mathbb{R}^n$, and then employ similarity functions defined on vectors. However, the task of designing these mappings, which are able to encapsulate all information in a vectorial form, is a difficult task since they need to:

- map isomorphic graphs into the same vector;
- non-isomorphic graphs into different vectors;
- be efficient in terms of time computation and memory consumption.

Recently, kernel methods with the use of kernel functions have emerged as one of the most powerful frameworks in machine learning. Kernel functions are considered as similarity functions which can be defined on any type of data representation. Therefore, the similarity measures defined on graphs mostly are kernels. There are two groups of kernels defined on graphs. The first group consists of kernels that aim at measuring the similalrities between graphs and they are referred as graph kernels. To have an overview of graph kernels, we recommend to readers a survey on graph kernels presented in [82]. The second one includes kernels which intend to measure the similarities between nodes inside graphs and are called graph node kernels or node kernels in short. For analysis of different graph node kernels, we suggest to read the work proposed in [31].

In the following, we first give formal definitions and notations related to a graph. We then give an overview of graph kernels followed graph node kernels.

Definition 2.5.1. A graph is a structure $G = (\mathbb{V}, \mathbb{E}, \mathcal{L}_n, \mathcal{L}_e)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the node (vertex) set, $\mathbb{E} = \{(v_i, v_j)\} \subseteq (\mathbb{V} \times \mathbb{V})$ is the link (edge) set and $\mathcal{L}_n, \mathcal{L}_e$ are the node and edge function, respectively.

Definition 2.5.2. *An undirected graph is a graph in which edges have no orientation. The edge (u, v) is identical to the edge (v, u), i.e. they are not ordered pairs, but sets $\{u, v\}$ (or 2-multisets) of vertices. The maximum number of edges in an undirected graph without a loop is $n \times (n - 1)/2$.*

Definition 2.5.3. *An adjacency matrix $\boldsymbol{A}$ is a symmetric matrix used to characterize the direct links between vertices v_i and v_j in the graph. Any entry A_{ij} is equal to w_{ij} when there exists a link connecting v_i and v_j, and is 0 otherwise.*

Definition 2.5.4. *The Laplacian matrix* $\boldsymbol{L}$ *is defined as* $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, *where* $\boldsymbol{D}$ *is the diagonal matrix with non-null entries equal to the summation over the corresponding row of the adjacency matrix, i.e.* $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{A}_{ij}$.

Definition 2.5.5. *The transition matrix of a graph* G, *denoted as* P, *is a matrix in which each element* $P_{ij} = A_{ij}/\sum_i A_{ij}$ *is the probability of stepping on node* j *from node* i.

We define the *distance* $\mathcal{D}(u, v)$ between two nodes u and v, as the number of edges on shortest path between them. The *neighborhood* of a node u with radius r, $N_r(u) = \{v \mid \mathcal{D}(u, v) \leq r\}$, is the set of nodes at distance no greater than r from u. The corresponding *neighborhood subgraph* $\mathcal{N}_r^u$ is the subgraph induced by the neighborhood (i.e. considering all the edges with endpoints in $N_r(u)$). The *degree* of a node u, $deg(u) = |\mathcal{N}_1^u|$, is the cardinality of its neighborhood. The maximum node degree in the graph G is denoted by $deg(G)$.

2.5.1 Graph kernels

The task of designing effcient and expressive graph kernels play an important role in the development of graph-based predictive systems. One of the first systematic works that strongly impacts on the development of research on graph kernels is convolution or decomposition kernel [40] (see Section 2.3). Existing graph kernels are decompositional kernels and can be classified into two categories: sequence-based graph kernels and subgraph-based graph kernels. The sequence-based graph kernels decompose graphs into "parts" in sequence-based forms, such as paths and walks; meanwhile, the subgraph-based graph kernels dissolve graphs into subgraphs. Following we first introduce some consequence-based graph kernels: product graph kernel, shorstest path kernels; we then describe several subgraph-based graph kernels: Weisfeiler-Lehman kernels, neighborhood subgraph pairwise distance kernel.

2.5.1.1 Product graph kernel

Product graph kernel was originaly proposed in [33] with the aim to measure the similarity between two labeled graphs by counting their common walks. To compute the similarity between two graphs (factor graphs), first a graph, called direct product graph, is constructed from two factor graphs. Then the similarity is computed based on the obtained graph.

Formally, we consider two factor graphs G_1 and G_2 and $\mathcal{L}_{n1}$, $\mathcal{L}_{n2}$, $\mathcal{L}_{e1}$, $\mathcal{L}_{e2}$ as the node and edge labeling functions of G_1 and G_2, respectively. We define the direct product graph of G_1, G_2 as a graph $G_\times = (V_\times, E_\times)$ where

- $V_\times = \{(u, v) : u \in V(G_1) \wedge v \in V(G_2) \wedge \mathcal{L}_{n1}(u) = \mathcal{L}_{n2}(v)\}$;
- $E_\times = \{((u, v), (u', v')) \in V_\times \times V_\times : (u, v) \in E(G_1) \wedge (u', v') \in E(G_2)$ $\wedge\, \mathcal{L}_{e1}((u, v)) = \mathcal{L}_{e2}((u', v'))\}$.

Given $\lambda_1, \lambda_2, \dots (\lambda_i \in \mathbb{R}, \lambda_i \geq 0,\ \forall i \in \mathbb{N})$, the direct product kernel is defined as follows

$$k(G_1, G_2) = \sum_{i,j=1}^{|V_\times|} \left[\sum_{n=0}^{\infty} \lambda_n A_\times^n \right]_{ij}, \tag{2.15}$$

if the limit exists, in which $A_\times$ is the adjacency matrix of the direct product graph $G_\times$. The computation of this limit is high, $O(n^6)$. There are different modifications of this kernel that can be more efficient to compute, such as the method proposed in [83] that has complexity of $O(n^3)$.

2.5.1.2 Shortest path kernels

The simple idea behind shortest path kernels is that they consider the common shortest paths between two graphs to measure their similarity. Although the computation of shortest paths in a graph can be computed in polinomial time, taking into account shortest paths leads to some problems. First, the shortest paths between two nodes normally are not unique and there has not been a way to deterministically choose one shortest path among others. Second, if we keep all shortest paths into account, it would lead to a NP-hard kernel. Although the shortest paths are not unique, the length of them is unique. As a consequence, a shortest path graph kernel is proposed in [15] with total runtime of $O(n^4)$.

Given two graphs G_1, G_2, we first construct their two corresponding shortest graphs G_{s1}, G_{s2}, respectively. The shortest graph of a graph, G, is a labeled graph defined as $G_s = (V_s, E_s)$ where

- $V_s = V(G)$;
- $E_s = \{(u, v) : u, v \in V(G) \wedge |p(u, v)| \geq 1\}$ where $p(u, v)$ is the number of shortest paths connecting u and v;
- $\mathcal{L}_e((u, v)) = \mathcal{D}(u, v)$.

We then define the shortest graph kernel for G_1 and G_2 as follows:

$$k_s(G_1, G_2) = \sum_{(u_1,v_1)\in E(G_{s1})} \sum_{(u_2,v_2)\in E(G_{s2})} k^1_{walk}((u_1, v_1), (u_2, v_2)), \qquad (2.16)$$

where k^1_{walk} is a positive semi-definite kernel on 1-length walks.

2.5.1.3 Weisfeiler-Lehman kernels

The Weisfeiler-Lehman kernel framework is proposed in [73]. It is derived from the kernel proposed in [72]. The idea is to first decompose each graph into a sequence of graphs. Then the similarity between two graphs G and G' is the summation of values computed by a given kernel defined on the two corresponding graph sequences, derived from G and G' by applying different labeling functions.

Formally, the Weisfeiler-Lehman graph sequence of a given graph $G = (V, E, \mathcal{L})$ up to the height h is defined as:

$$\{G_0, G_1, \dots, G_h\} = \{(V, E, \mathcal{L}_0), (V, E, \mathcal{L}_1), \dots, (V, E, \mathcal{L}_h)\},$$

where $G_0 = G$, $\mathcal{L}_0 = \mathcal{L}$, $\mathcal{L}_i$s ($i = \overline{1, h}$) are different labeling function on G.

Given two graphs G, G' and a graph kernel k, the Weisfeiler-Lehman kernel, k_{WL} is defined as:

$$k_{WL}(G, G') = k(G_0, G_0^{'}) + k(G_1, G_1^{'}) + \ldots + k(G_h, G_h^{'}). \qquad (2.17)$$

2.5.1.4 The neighborhood subgraph pairwise distance kernel

The NSPDK is an instance of convolution kernel [40] where given a graph $G \in \mathcal{G}$ and two rooted graphs A_u, B_v, the relation $R_{r,d}(A_u, B_v, G)$ is true *iff* $A_u \cong \mathcal{N}^u_r$ is (up to isomorphism $\cong$) a neighborhood subgraph of radius r of G and so is $B_v \cong \mathcal{N}^v_r$, with roots at distance $\mathcal{D}(u, v) = d$ (see Figure 2.2). We denote with R^{-1} the inverse relation that returns all pairs of neighborhoods of radius r at distance d in G, $R^{-1}_{r,d}(G) = \{A_u, B_v | R_{r,d}(A_u, B_v, G) = true\}$. The kernel $\kappa_{r,d}$ over $\mathcal{G} \times \mathcal{G}$, counts the number of such fragments in common in two input graphs:

$$\kappa_{r,d}(G, G^{'}) = \sum_{\substack{A_u, B_v \,\in\, R^{-1}_{r,d}(G) \\ A'_{u'}, B'_{v'} \,\in\, R^{-1}_{r,d}(G')}} \mathbf{1}_{A_u \cong A'_{u'}} \cdot \mathbf{1}_{B_v \cong B'_{v'}} \qquad (2.18)$$

where $\mathbf{1}_{A \cong B}$ is the *exact matching function* that returns 1 if A is isomorphic to B and 0 otherwise. Finally, the NSPDK is defined as

$$K(G, G') = \sum_r \sum_d \kappa_{r,d}(G, G'), \tag{2.19}$$

where for efficiency reasons, the values of r and d are upper bounded to a given maximal r^* and d^*, respectively.

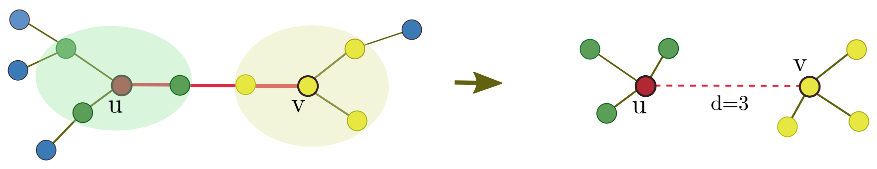

Figure 2.2: Example of a pairwise neighborhood subgraphs rooted at u with radius $r = 1$ and distance $d = 3$.

2.5.2 Graph node kernels

Different from graph kernels which aim at measuring similarities between graphs, graph node kernels intend to measure similarities between nodes in graphs. The ideas behind graph node kernels are similar to graph kernels. Therefore, they can be devided into sequence-based graph node kernels and subgraph-based graph node kernels. However, graph node kernels attempt to exploit the configuration concerning two nodes in the graph in order to define their similarity. Most available graph node kernels are sequence-based graph node kernels and they are based on the diffusion phenomenon. In other words, they consider paths connecting two given nodes in order to form their similarity. In the following, we introduce some of the most used graph node kernels.

2.5.2.1 Laplacian exponential diffusion kernel

One of the most well-known kernels for graphs is the Laplacian exponential diffusion kernel **LEDK**, as it is widely used for exploiting discrete structures in general and graphs in particular. On the basis of the heat diffusion dynamics, Kondor and Lafferty proposed **LEDK** in [48]: imagine to initialize each vertex with a given amount of heat and let it flow through the edges until an arbitrary instant of time. The similarity between any vertex couple v_i, v_j is the amount of heat starting from v_i and reaching v_j within the given time. Therefore, **LEDK** can capture the long range relationship between vertices of a graph to define the global similarities. Below is the formula to compute **LEDK** values:

$$K = e^{-\beta \mathbf{L}} = \mathbf{I} - \beta \mathbf{L} + \frac{\beta \mathbf{L}^2}{2!} - \ldots, \tag{2.20}$$

where β is the diffusion parameter and is used to control the rate of diffusion and $\mathbf{I}$ is the identity matrix. Choosing a consistent value for β is very important: on the one side, if β is too small, the local information cannot be diffused effectively and, on the other side, if it is too large, the local information will be lost. **LEDK** is positive semi-definite as proved in [48].

2.5.2.2 Exponential diffusion kernel

In **LEDK**, the similarity values between high degree vertices are generally higher compared to those between low degree ones. Intuitively, the more paths connect two vertices, the more heat can flow between them. This could be problematic since peripheral nodes have unbalanced similarities with respect to central nodes. In order to make the strength of individual vertices comparable, a modified version of **LEDK** introduced by Chen et al in [24] is called Markov exponential diffusion kernel **MEDK** and given by the following formula:

$$\mathbf{K} = e^{-\beta \mathbf{M}}. \tag{2.21}$$

The difference with respect to the Laplacian diffusion kernel is the replacement of $\mathbf{L}$ by the matrix $\mathbf{M} = (\mathbf{D} - \mathbf{A} - n\mathbf{I})/n$ where n is the total number of vertices in graph. The role of β is the same as for **LEDK**.

2.5.2.3 Markov diffusion kernel

The original Markov diffusion kernel **MDK** introduced by Fouss et al. [32] exploits the idea of diffusion distance, which is a measure of how similar the pattern of heat diffusion is among a pair of initialized nodes. In other words, it expresses how much nodes ”influence” each other in a similar fashion. If their diffusion ways are alike, the similarity will be high and, vice-versa, it will be low if they diffuse differently. This kernel is computed starting from the transition matrix $\mathbf{P}$ and by defining $\mathbf{Z}(t) = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{P}^{\tau}$, as follows:

$$\mathbf{K} = \mathbf{Z}(t)\mathbf{Z}^{\top}(t). \tag{2.22}$$

2.5.2.4 Regularized Laplacian kernel

Another popular graph node kernel function used in graph mining is the regularized Laplacian kernel **RLK**. This kernel function was introduced by

Chebotarev and Shamis in [22] and represents a normalized version of the random walk with a restart model. It is defined as follows:

$$\mathbf{K} = \sum_{n=0}^{\infty} \beta^n (-\mathbf{L})^n = (\mathbf{I} + \beta \mathbf{L})^{-1}, \tag{2.23}$$

where the parameter β is again the diffusion parameter. **RLK** counts the paths connecting two nodes on the graph induced by taking **-L** as the adjacency matrix, regardless of the path length. Thus, a non-zero value is assigned to any couple of nodes as long as they are connected by any indirect path. **RLK** remains a relatedness measure even when the diffusion factor is large, by virtue of the negative weights assigned to self-loops.

2.6 Disease gene prioritization

The identification of the genes underlying human diseases is a major goal in current molecular genetics research. Dramatic progresses have been made since the 1980s, when only a few DNA loci were known to be related to disease phenotypes. Nowadays opportunities for the diagnosis and the design of new therapies are progressively growing, thanks to several technological advances and the application of statistical or mathematical techniques. For instance, positional cloning has allowed to map a vast portion of known Mendelian diseases to their causative genes [75, 19]. Despite the huge advances, much remains to be discovered. On December 21st 2016, the Online Mendelian Inheritance in Man database (OMIM) registered 4,908 Mendelian phenotypes of known molecular basis and 1,483 Mendelian phenotypes of unknown molecular origin [4]. Moreover, 1,677 more phenotypes were suspected to be Mendelian. But it is among oligogenic and poligenic (and multifactorial) pathologies that the most remains to be elucidated: for the majority of them, only a few genetic loci are known [75, 19].

Independently of the type of disease, the search of causative genes usually concerns a large number of suspects. It is therefore necessary to recognise the most promising candidates to submit to additional investigations, as experimental procedures are often expensive and time consuming. Gene prioritization is the task of ordering genes from the most promising to the least. In traditional genotype-phenotype mapping approaches - as well as in genome-wide association studies - the first step is the identification of the genomic region(s) wherein the genes of interest lie. Once the candidate region is identified, the genes there residing are prioritized and finally analysed for the presence of possible causative mutations [75]. More recently, in new generation sequencing studies this process is inverted as the first step is the identification of mutations, followed by prioritization and final validation [68]. Prioritization criteria are usually based on functional relationships, co-expression and other clues linking genes together. In general, all of them follow the "guilt-by-association" principle, i.e. disease genes are sought by looking for similarities to genes already associated to the pathology of interest [75].

In the last few years, computational techniques have been developed to aid researchers in this task, applying both statistics and machine learning [61]. Thanks to the advent of high-throughput technologies and new generation sequencing, a huge amount of data is in fact available for this kind of investigations. In particular, computational methods are essential for multi-*omics* data integration, that has been recognised as a valuable strategy for understanding genotype-phenotype relationships [67]. In fact, clues are often embedded in different data sources and only their combination leads to the emergence of informative patterns. Furthermore, incompleteness and noise of the single sources can be overcome by inference across multiple levels of knowledge.

Several popular algorithms for pattern analysis are based on *kernels*, which are mathematical transformations that permit to estimate the similarity among items (in our case genes) taking into account complex data relations [71]. Importantly, kernels provide a universal encoding for any kind of knowledge representation, e.g. vectors, trees or graphs. When data integration is required, a multiple kernel learning (MKL) strategy allows a data-driven weighting/selection of meaningful information [38]. The goal of MKL is indeed to learn optimal kernel combinations starting from a set of predefined kernels obtained by various data sources. Through MKL the issue of combining different data types is then solved by converting each dataset in a kernel matrix.

Numerous MKL approaches have been proposed for the integration of genomic data [85, 16] and some of them have been applied to gene prioritization [30, 91, 60, 92]. De Bie *et al* formulated the problem as a one-class support vector machine (SVM) optimization task [30], while Mordelet and Vert tackled it through a biased SVM in a *positive-unlabelled* framework [60, 20]. Recently, Zakeri *et al* proposed an approach for learning non-linear log-euclidean kernel combinations, showing that it can more effectively detect complementary biological information compared to linear combinations-based approaches [92]. However, as highlighted in a recent work by Wang *et al* [85], current methods share two limitations: high computational costs - given by a (at least) quadratic complexity in the number of training examples - and the difficulty to predefine optimal kernel functions to be fed to the MKL machine.

Let us formally define the problem of disease gene prioritization which is later on employed in our empirical experiments to evaluate of different methods. We consider a list of genes $\mathcal{G} = \{g_1, g_2, ..., g_n\}$ that could either be the full list of human genes or a subset of it. Considering a specific disease, there exists a set $P_i \subseteq \mathcal{G}$ of genes known to be associated with it. Its complementary set $U_i = \mathcal{G} - P_i$ contains genes that are not a priori related to the disease, but we assume that inside U_i some positive genes, i.e. causing the disease, are hidden. Gene priorization is a task that ranks the genes in U_i based on their likelihood to be related to P_i.

2.7 Biological datasets

The development of computational biology makes a high number of biological datasets available. Many biological datasets can naturally be represented as networks which are later on used as the input of graph-based biological systems. In biological networks, vertices are biological entities (genes and proteins, etc) and links describe the relation between entities. The relations can be discovered by either physical experiments and results from inferring methods (systems). We describe how information from some biological datasets are extracted and transformed into undirected networks, represented by adjacency matricies. These networks will be employed in our experiments for evaluating the performance of proposed algorithms.

Human Protein Reference Database (**HPRD**): database of curated proteomic information pertaining to human proteins. It is derived from [45] with 9,465 vertices and 37,039 edges. We employ the HPRD version used in [21] that forms a graph which contains 7,311 vertices (genes) and 30503 links. In the graph, two vertices are linked if proteins encoded by their corresponding genes interact.

BioGPS [88]: contains expression profiles for 79 human tissues, which are measured by using the Affymetrix U133A array. Gene co-expression, defined by pairwise Pearson correlation coefficients (PCC), is used to build an unweighted graph. A pair of genes are linked by an edge if the PCC value is larger than 0.5.

Pathways: Pathway datasets are obtained from the database of KEGG [64], Reactome [80], PharmGKB [87] and PID [69], which contain 280, 1469, 99 and 2,679 pathways, respectively. A pathway co-participation network is constructed by connecting genes that co-participate in any pathway.

String [43]: the String database gathers protein information covering seven levels of evidence: genomic proximity in procaryotes, fused genes, co-occurrence in organisms, co-expression, experimentally validated physical interactions, external databases and text mining. Overall, these aspects focus on functional relationships that can be seen as edges of a weighted graph, where the weight is given by the reliability of that relationship. To perform unbiased evaluation we employed the version 8.2 of String from which we extracted functional links among 17,078 human genes.

Phenotype similarity: we use the OMIM [56] dataset and the phenotype similarity notion introduced by Van Driel et al. [77] based on the relevance and the frequency of the Medical Subject Headings (MeSH) vocabulary terms in OMIM documents. We built the graph linking those genes whose associated phenotypes have a maximal phenotypic similarity greater than a fixed cut-off value. Following [77], we set the similarity cut-off to 0.3. The resulting graph has 3,393 nodes and 144,739 edges.

Biogridphys: this dataset encodes known physical interactions among proteins. The idea is that mutations can affect physical interactions by changing the shape of proteins and their effect can propagate through protein graphs. We introduce a link between two genes if their products interact. The resulting graph has 15,389 nodes and 155,333 edges.

Biogridgen: Genetic interaction is the phenomenon through which the effects of a gene are modified by one or several other genes. This occurs in indirect way by means of knock-on effects of multiple physical interactions. In practice, this is observed when the effects of two mutations in distinct genes is not equal to the sum of the effects of the mutations alone. This kind of interaction is complementary in respect to the physical one and is important especially for complex diseases involving a large number of genes. In the adjacency matrix, the entries of coordinates (i, j) and (j, i) are equal to 1 if gene i and j interact. Otherwise, they are equal to 0.

Omim: OMIM is a public database of disease-gene association. Genes implicated in the same disease are more likely to be involved in other similar diseases as well. Therefore, Omim network is formed by connecting genes which are involved in common disease(s).

2.8 Link prediction

We are witnessing a constant increase of the rate at which data is being produced and made available in machine readable formats. Interestingly it is not only the quantity of data that is increasing, but also its complexity, i.e. not only we are measuring a number of attributes or features for each data point, but we are also capturing their mutual relationships, that is, we are considering non independent and identically distributed (non i.i.d.) data. This yields collections that are best represented as graphs or relational data bases and requires a more complex form of analysis. As cursory examples of application domains are social networks, where nodes are people and edges encode a type of association such as friendship or co-authorship; bioinformatics, where nodes are proteins and metabolites, and edges represent a type of chemical interaction such as catalysis or signaling; and e-commerce, where nodes are people and goods, and edges encode a "buy" or "like" relationship. A key characteristic of this type of data collection is the sparseness and dynamic nature, i.e. the fact that the number of recorded relations is significantly smaller than the number of all possible pairwise relations, and the fact that these relations evolve in time. A crucial computational task is then the "link prediction problem" which allows to suggest friends, or possible collaborators for scientists in social networks, or to discover unknown interactions between proteins to explain the mechanism of a disease in biological networks, or to suggest novel products to be bought to a customer in an e-commerce recommendation system. Many approaches to link prediction that exist in literature can be partitioned according to *i*) whether additional or "side" information is available for nodes and edges or rather only the network topology is considered and *ii*) whether the approach is unsupervised or supervised.

Unsupervised methods are, in this setting, non-adaptive, i.e. they do not have parameters that are tuned on the specific problem instance, and can therefore be computationally efficient. In general they define a score for any node pair that is proportional to the existence likelihood of an edge between the two nodes. *Adamic-Adar* [6, 53] computes the weighted sum over the common neighbors where the weight is inversely proportional to the (log of) each neighbor node degree. The *preferential attachment* [12] method computes a score simply as the product of the node degrees in an attempt to exploit the "rich get richer" property of certain network dynamics. *Katz* [44] takes into account the number of common paths with different lengths between two nodes, assigning more weight to shorter paths. The *Leicht-Holme-Newman* method [50] computes the number of intermediate nodes. In [57] the score is derived from the singular value decomposition of the adjacency matrix. Two methods, named Local Random Walk (LRW) and Superposed Random Walk (SRW), are proposed in [52]. These methods work based on random walk. Besides, graph node kernels, including [48, 24, 32, 22], can be applied to measure the similarities between nodes and link prediction is made based on these similarities. For more information of link prediction methods, we refer readers to [54, 57].

Supervised link prediction methods convert the problem into a binary classification task where links present in the network (at a given time) are considered as positive instances and a subset of all the non links are considered as negative instances. Following [57], we can further group these methods into four classes: feature-based models, graph regularization models, latent class models, and latent feature models. A Bayesian non-parametric approach is used in [58] to compute a non-parametric latent feature model that does not need a user defined number of latent features but rather induces it as part of the training phase. In [57] a matrix factorization approach is used to extract latent features that can take into consideration the output of an arbitrary unsupervised method. The authors show a significant increase in predictive performance when considering a ranking loss function suitable for the imbalance problem, i.e. when the number of negative instances is much larger than the number of positive instances.

In general, supervised methods exhibit better accuracies compared to unsupervised methods although incurring in much higher computational and memory complexity costs.

2.9 Biological data integration

Data integration has attracted many researchers because of its important role in building high performance learning systems for biological data (as discussed in the Chapter 1). As a consequence, there is a high number of data integration methods which have been proposed in the last decades. According to [36], existing methods can be divided into three classes: early data integration, late data integration, and intermediate data integration.

Early data integration first combines different data sources into a single one. It then builds a model for inference. Typical approaches are proposed in [49, 94, 60, 26, 46]. A common requirement for the methods in this class is that data sources need to be transformed into a common representation. This might lead to the problem of information loss.

Late data integration builds models for each data source separately. It then combines different obtained models to have a unified one. A common technique for combining is to use the majority voting policy. Late data integration methods often show relatively low performance since models are built from each dataset in isolation from others. Examples in this class are [34, 84, 55, 62].

Intermediate data integration combines data through inference of a joint model. An advantage of this strategy is that it does not require any data transformation. Therefore, it does not lead to the problem of information loss. Usually, it shows high performance in many applications, including [49, 34, 78, 94, 65].

2.10 Multiple kernel learning

A common way to represent data in data integration is using kernels since kernels are well-known as universal methods for data representation (see Section 2.3). First kernels are defined on each data source. Then the obtained kernels are combined into a single higher abstract level of data representation. The combination is often performed by using multiple kernel learning (MKL) algorithms [38, 85] (see [38] for a recent and quite exhaustive survey).

The task of Multiple Kernel Learning is to combine kernels derived from multiple sources in a data-driven way with the aim of improving the accuracy of a target kernel machine. MKL algorithms are normally in linear forms because of the two following reasons. Firstly, the time required to solve the associated optimization problem grows, normally more than linearly, w.r.t the number of pre-defined kernels. Secondly, employing sophisticated algorithms often do not significantly outperform the simple average of kernels. However, most of them still require a long computation time and a high memory consumption, especially when the number of pre-defined kernels is high. To tackle these limitations, a scalable multiple kernel learning named EasyMKL has been previously proposed [10]. This method focuses on learning a linear combination of the input kernels with positive linear coefficients, namely

$$\mathbf{K} = \sum_{r=1}^{R} \eta_r \mathbf{K}_r, \;\; \eta_r \geq 0\,, \tag{2.24}$$

where $(\eta_1, \ldots, \eta_R)$ is the coefficient vector and R is the number of kernels to combine. For a fully supervised binary task, EasyMKL computes the optimal kernel by maximizing the distance between positive and negative examples. The base learner is a kernel-based approach for the optimization of the margin distribution in binary classification or ranking [9].

In order to present its formulation, let us first define the probability distribution $\gamma \in \mathbb{R}_+^N$ representing weights assigned to training examples and living in the domain $\Gamma = \{\gamma \in \mathbb{R}_+^N | \sum_{i \in \mathcal{P}} \gamma_i = 1, \sum_{i \in \mathcal{N}} \gamma_i = 1\}$, where $\mathcal{N}$ is the set of negative examples. From this definition, it follows that any element $\gamma \in \Gamma$ represents a pair of points in the input space: the first one is constrained to the convex hull of positive training examples and the second one to the convex hull of negative training examples. As stated above, EasyMKL maximizes the distance between positive and negative examples, optimizing the margin distribution at the same time. Under this notation, the task can be posed as a min-max problem over variables γ and η as follows:

$$\max_{\eta:\|\eta\|_2\leq 1} \min_{\gamma\in\Gamma} (1-\lambda)\gamma^\top \mathbf{Y}(\sum_r \eta_r \mathbf{K}_r)\mathbf{Y}\gamma + \lambda\,\gamma^\top\gamma\,. \tag{2.25}$$

Here $\mathbf{Y}$ is a diagonal matrix containing the vector of example labels, +1 for the positive and -1 for the negative. Optimization of the first term alone leads to an optimal γ^* representing the two nearest points in the convex hulls of positive and negative examples, equivalently to a hard SVM task using a kernel $\mathbf{K}$ [9]. The second term represents a quadratic regularization over γ whose solution is the squared distance between positive and negative centroids in the feature space. The regularization parameter $\lambda \in [0, 1]$ permits to tune the objective to optimize, by balancing between the two critical values $\lambda = 0$ and $\lambda = 1$. When $\lambda = 0$ we obtain a pure hard SVM objective, while when $\lambda = 1$ we get a centroid-based solution.

It can be shown that this problem has analytical solution in η, so that the previous expression can be reshaped into:

$$\min_{\gamma\in\Gamma} (1-\lambda)\gamma^\top \mathbf{Y}\mathbf{K}^s\mathbf{Y}\gamma + \lambda\,\gamma^\top\gamma\,, \tag{2.26}$$

where $\mathbf{K}^s = \sum_r^R \mathbf{K}_r$ is the sum of the pre-defined kernels. This minimization can be efficiently solved and only requires the sum of the kernels. The computation of the kernel summation can be easily implemented incrementally and only two matrices need to be stored in memory at a time. As shown in [10], EasyMKL can deal with an arbitrary number of kernels using a fixed amount of memory and a linearly increasing computation time.

Once the problem in Eq. 2.26 is solved, we have an optimal γ^* and we are able to obtain the optimal kernel weights η_r^* by using the formula:

$$\eta_r^* = \frac{\gamma^*\mathbf{Y}\mathbf{K}_r\mathbf{Y}\gamma^*}{\sum_{r=1}^R \gamma^*\mathbf{Y}\mathbf{K}_r\mathbf{Y}\gamma^*}\,. \tag{2.27}$$

The optimal kernel is thus evaluated as $\mathbf{K}^* = \sum_r^R \eta_r^*\mathbf{K}_r$. Finally, by replacing $\mathbf{K}^s$ with $\mathbf{K}^*$ in Eq. 2.26, we can get the final probability distribution γ^*.

Chapter 3

Conjunctive Disjunctive Graph Node Kernel

In this chapter, we propose a graph node kernel, named conjunctive disjunctive node kernel (CDNK), which is an instance of convolutional kernels. To measure similarities between nodes in large-scale graphs, our graph node kernel not only effectively exploits graph structures, but also takes the side information (auxiliary information) associated to graph nodes into account. The empirical evaluation results on several datasets shows that CDNK significantly outperforms various famous graph node kernels.

3.1 Motivation

As discussed before, node similarity measure definition is the key that determines the performance of graph-based learning systems. The state of the art graph node kernels used to measure node similarity, are based on the notion of information diffusion, including LEDK [48], MEDK [24], MDK [32], RLK [22], etc. These graph node kernels often show relatively low discriminative capacity, especially when working with sparse graphs, i.e. graphs whose high numbers of missing links. This is due to they share the following limitations. First, they process information in an additive and independent fashion which prevents them from accurately modeling the configuration of each node's context. Second, they do not take into account auxiliary information associated to nodes of graphs when they are available. Additional information normally provides a complement to graph topology. Therefore, using additional information would help to improve the expressiveness of graph node kernels.

We propose convolutional graph node kernel, *the conjunctive disjunctive node kernel*, which is able to i) effectively exploit the nodes' context, ii) utilize the side information available on the graph nodes.

3.2 Conjunctive disjunctive graph node kernel

We start from the type of similarity notion computed by a neighborhood based decomposition kernel between graph instances, NSPDK [29], and adapt it to form a convolutional graph node kernel which aims at expressing the similarity between nodes in a single graph. Our intended graph node kernel takes an undirected, labeled graph as its input.

Given an input labeled graph $G = (\mathbb{V}, \mathbb{E}, \mathcal{L}_1, \mathcal{L}_2)$, our kernel consists of two phases. In the first phase, a network decomposition procedure is applied to transform the graph into a set of linked sparse connected components. In this procedure, we define two different kinds of link: *conjunctive* and *disjunctive* in which we treat in distinct manners.

In the second phase, the similarity between any node couple $(u,\ v)$, $u, v \in \mathbb{V}$ is computed by adopting NSPDK on two neighborhood subgraphs rooted at u and v. Notice that these neighborhood subgraphs are extracted the decomposed graph. In the following, we describe each phase in detail.

3.2.1 Network decomposition

In genetic networks, it is not uncommon to find nodes with high degrees. Unfortunately these cases cannot be effectively processed by a neighborhood based decomposition kernel (see Section 2.5.1.4) since they are based on the notion of exact matches. Neighborhood subgraphs rooted at high degree nodes are relatively big due to a high number of neighbors. It leads to a low likelihood of finding neighborhood subgraphs rooted at high degree nodes which are isomorphic. Thus, the probability of having identical neighborhoods decreases exponentially as the degree increases. This means that in a finite network it quickly becomes impossible to find any match and hence learn or generalize at all.

We propose a procedure to "sparsify" the network that is observed by the neighborhood kernel. In practice, we mostly keep the cardinality of the edge set unchanged. However we mark the edges with special attributes so that the kernel is able to treat them differently when computing the output value. The result is a procedure that decomposes the network in a linked collection

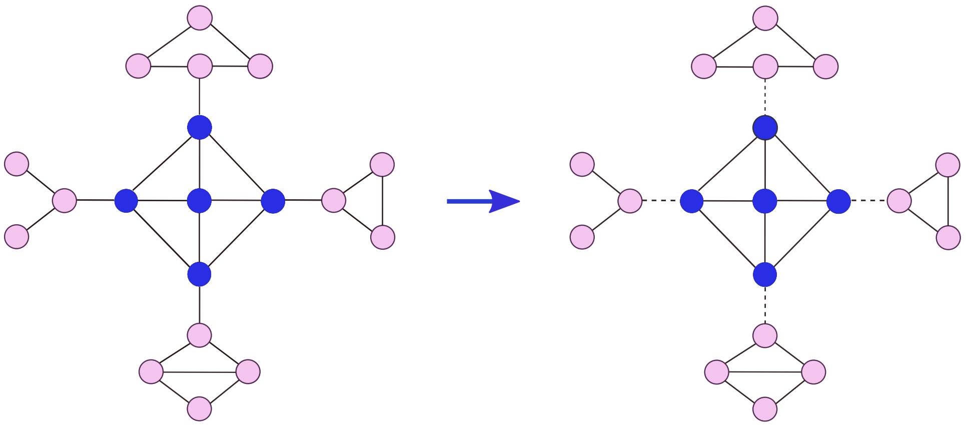

Figure 3.1: K-core decomposition with degree threshold $D = 4$. Links in solid style are conjuctive, while links in dashed style are disjuctive.

of sparse sub-networks where each node has a reduced connectivity when considering the edges of a specific type. However the other edges are still available to connect the various sub-networks. We distinguish two types of edges: *conjunctive* and *disjunctive* edges. Nodes linked by conjunctive edges are going to be used jointly to define the notion of context and will be visible to the neighborhood graph kernel. Nodes linked by disjunctive edges are instead used to define features based only on the pairwise co-occurrence of the nodes at the endpoints and are processed by our novel kernel.

The network decomposition works by first applying the iterative k-core decomposition and then the clique decomposition. Next, we describe both the k-core and the clique decomposition procedure.

Iterative k-core decomposition [11]: the node set is partitioned in two groups on the basis of the degree of each node w.r.t. a threshold degree D, the first part contains all nodes with degree smaller than D and the second part the remaining ones. The node partition is used to induce the "conjunctive" vs "disjunctive" notion for the edge partition: edges that have both endpoints in the same part are marked as conjunctive, otherwise they are marked as disjunctive. Figure 3.1 shows an illustration of k-core with the threshold $D = 4$. We apply the k-core decomposition iteratively, where at each iteration we consider only the graph induced by the conjunctive edges. We stop iterating the decomposition after a user defined number of steps. Note that this decomposition does not alter the cardinality of the edge set, it is simply a procedure to mark each edge with the attribute conjunctive or disjunctive.

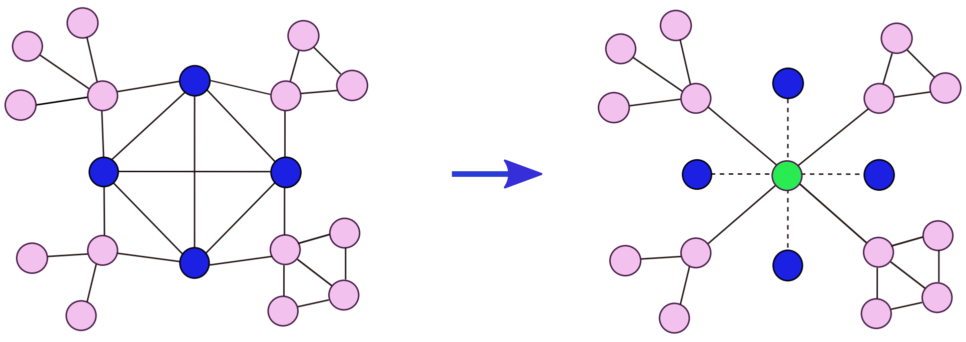
Figure 3.2: Clique decomposition with threshold $C = 4$. Links in solid style are conjuctive, while links in dashed style are disjuctive.

Clique decomposition [76]: to model the notion that nodes in a clique are tightly related, we summarize the whole clique with a new "representative" node. All the cliques (completely connected subgraphs) with a number of nodes greater than or equal to a given threshold size C are identified. The endpoints of all edges incident on the clique's nodes are moved to the representative node. Disjunctive edges are introduced to connect each node in the clique to the representative. Finally all edges with both endpoints in the clique are removed. An example of clique decomposition is presented in Figure 3.2.

3.2.2 Kernel definition

We define a node kernel $K(G_u, G_{u'})$ between two copies of the same network G where we distinguish the nodes u and u' respectively. The idea is to define the features of a node u as the subset of NSPDK features (pairwise neighborhood subgraphs) that always have u as one of the roots. When computing distances to induce neighborhood subgraphs, only conjunctive edges are considered. When choosing the pair of neighborhoods to form a single feature, we additionally consider roots u and v that are not at distance d but such that u is connected to w via a disjunctive edge and such that w is at distance d from v (Figure 3.3 is an illustration of extracted pairwise neighborhood subgraphs). In this way disjunctive edges can still allow an *information flow* even if their endpoints are only considered in a pairwise fashion and not jointly.

Formally, we define two relations: (i) the *conjunctive relation* $R^{\wedge}_{r,d}(A_u, B_v, G_u)$ identical to the NSPDK relation $R_{r,d}(A_u, B_v, G)$, and (ii) $\mathcal{D}(u, v) = d$; the *disjunctive relation* $R^{\vee}_{r,d}(A_u, B_v, G_u)$ is true *iff* (a) $A_u \cong \mathcal{N}^u_r$

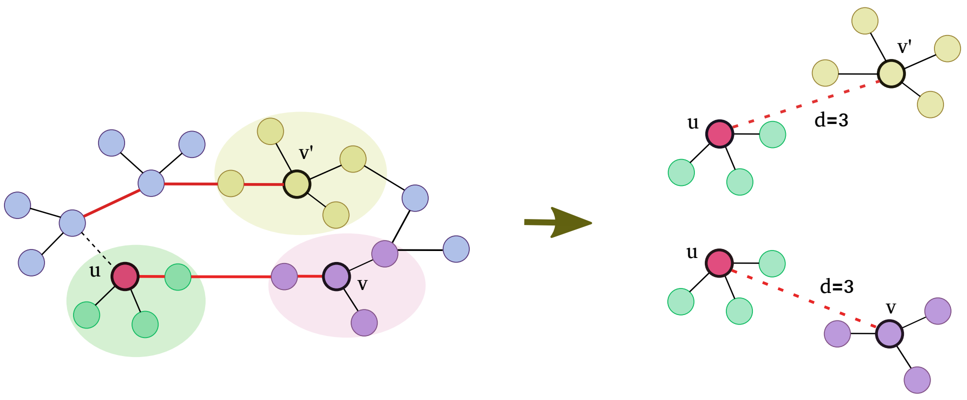

Figure 3.3: Pairwise neighborhood subgraphs for the "red" node with $r = 1$ and $d = 3$ using "*conjuntive*" and "*disjuctive*" edges.

and $B_v \cong \mathcal{N}_r^v$ are true, (b) $\exists w$ s.t. $\mathcal{D}(w, v) = d$, and (c) (u, w) is a disjunctive edge.

We define $\kappa_{r,d}$ on the inverse relations ${R^{\wedge}_{r,d}}^{-1}$ and ${R^{\vee}_{r,d}}^{-1}$:

$$\kappa_{r,d}(G_u, G_{u'}) = \sum_{\substack{A_u, B_v \in {R^{\wedge}_{r,d}}^{-1}(G_u) \\ A'_{u'}, B'_{v'} \in {R^{\wedge}_{r,d}}^{-1}(G_{u'})}} \mathbf{1}_{A_u \cong A'_{u'}} \cdot \mathbf{1}_{B_v \cong B'_{v'}} + \sum_{\substack{A_u, B_v \in {R^{\vee}_{r,d}}^{-1}(G_u) \\ A'_{u'}, B'_{v'} \in\ {R^{\vee}_{r,d}}^{-1}(G_{u'})}} \mathbf{1}_{A_u \cong A'_{u'}} \cdot \mathbf{1}_{B_v \cong B'_{v'}},$$

where $\mathbf{1}_{A_u \cong B_v}$ is the matching function which returns 1 if A_u and B_v are isomorphic and 0 otherwise (see [29] for an efficient approximate function for graph isomorphism).

The CDNK is finally defined as $K(u, u') = \sum_r \sum_d \kappa_{r,d}(G_u, G_{u'})$, where for efficiency reasons, the values of r and d are upper bounded to a given maximal r^* and d^*.

In order to integrate the real valued vector labels, we proceed as follows. We compute a sparse vector representation for the neighborhood graph rooted in node v following [29]: for each neighborhood subgraph we calculate the quasi-isomorphism certificate hash code; we then combine the hashes for the pair of neighborhoods and use the resulting integer as a feature indicator. This yields a direct sparse vector representation (associated to node u in graph G) $f : G_u \longmapsto I\!R^N$ where $N \approx 10K - 1M$. Given the real valued vector information (associated to node u in graph G) $g : G_u \longmapsto I\!R^P$ computed as the multi-class similarity to the P clusters (c.f.r. Section 3.4.2), we update the computation of CDNK considering the discrete convolution of the discrete information with the real valued information:

$$K(u, u') = \left\langle f(G_u) \bigotimes g(G_u), f(G_{u'}) \bigotimes g(G_{u'}) \right\rangle, \tag{3.1}$$

where the discrete convolution is defined as:

$$(f \bigotimes g)[n] = \sum_{m=0}^{K-1} f[n-m]g[m]. \tag{3.2}$$

In words, we are starting a scaled copy of the real valued vector at the position indicated by each feature computed on the basis of the discrete information. Intuitively, when both the real valued and the discrete information match, the kernel computes a large similarity, but if there is a discrepancy in either one of the sources of information, the similarity will be penalized.

3.3 Parameter space

Our kernel consists of five parameters: the threshold degree D, the clique size threshold C, the maximal radius r^*, the maximal distance d^* and the number of clusters P. In order to choose the optimal tuple of values for kernel parameters in a specific setting, a model selection procedure is normally adopted. The parameter space of our kernel seems large due to the relatively high number of parameters. However, most parameter values are supposed to be in limited natural ranges. Therefore, it is actually not too big. In particular, the clique size threshold C is in $\{4, 5\}$. The maximal radius r^* is set with a value smaller than or equal to 3, otherwise it will lead to big neighborhood subgraphs. The maximal distance d^* is usually assigned with values less than or equal to 4 because the value of the maximal shortest distance between nodes in a connected component is often not too high. The values of the threshold degree D should neither be too high nor too low. If it is high, we have to face with the high degree node problem, and if it is too low, the obtained graph containing too sparse connected components. Therefore, we suggest value for D in the range $[6, 20]$.

3.4 Empirical evaluation

In this section, we evaluate the performance of CDNK and other graph node kernels to answer the two following questions:

- Does CDNK show better performance comparing to other graph node kernels?
- Does the use of side information (real vector labels) help to improve the performance of CDNK?

3.4.1 Experimental settings and evalution method

We carry out experiments in the context of disease gene prioritization (see Section 2.6).

The experiments are performed on two separate networks derived from BioGPS and Pathways datasets, described in Section 2.7. We follow the experiment procedure in [24] where 12 diseases [37] are used in which each disease is associated to at least 30 positive (known) genes (see Table A.1. For each disease, we construct a positive set $\mathcal{P}$ with all positive (confirmed) disease genes, and a negative set $\mathcal{N}$ which contains random genes associated at least to one disease class which is not related to the class that is defining the positive set. In [24] the ratio between the dataset sizes is chosen as $|\mathcal{N}| = \frac{1}{2}|\mathcal{P}|$. This is due to the fact that genes known to be related to at least a genetic disease, but not to the considered one, are well studied, so they have a low probability to be associated to the current disease.

To compare the performance of graph node kernels, we fix the learning algorithm and use diffrent graph node kernels: LEDK, MEDK, MDK, RLK and CDNK.

The predictive performance of the system using each kernel is evaluated via a leave-one-out cross validation: one gene is kept out in turn and the rest are used to train an SVM model. We compute a decision score q_i for the test gene g_i as the top percentage value of score s_i among all candidate gene scores. We collect all decision scores for every gene in the test set to form a global decision score list on which we compute the AUC-ROC.

3.4.2 Node labeling

Our graph node kernel define the node similarities in labled graphs in which labels can be discrete and/or real valued vectors. Therefore, to be able to carry out experiments on graphs, which are not labeled, derived from BioGPS and Pathways, we need ways to label for these genetic graphs. This is due to In the following, we describe our proposed node labeling methods for genetic networks.

Discrete labels: we introduce two different approaches to associate genes with dicrete labels.

In the first approach, we use same label for every node of the graphs. In this case, the node contexts expressed by the pairwise neighborhood subgraphs in CDNK now become pairwise neighborhoods.

The second approach aims to use nodes labels to encode abstract information about genes. In this way, it allows downstream machine learning algorithms to generalize from similar examples and the identification of overlooked but related genes. We employ a gene ontology (GO) [27] to construct binary vectors representing a bag-of-words encoding for each gene, i.e. each element of a binary vector is equal to 1 if its corresponding GO-term is associated to the gene, and is equal to 0 otherwise. The resulting vectors are then clustered using the k-means algorithm into a user defined number of clusters, P, so that genes with similar description profiles receive the same class identifier as label.

Real vector labels: in addition to encoding the functional information as a discrete label we add a richer description by computing the similarity vector w.r.t. to each cluster. In this way, we can fully exploit the latent description of the genes in terms of the different functional groups captured by the clustering procedure. Formally, given a vector $v \in I\!R^{26501}$ $(26,501$ is the number of GO-terms), we compute a similarity vector $S(v) = s_1, s_2, \ldots s_P$ with entries $s_i = \frac{1}{1+\ell(v,c_i)}$ where $\ell(v, c_i)$ is the Euclidean distance of v from the center of the i^{th} cluster $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ computed as the geometric mean of the elements in the cluster C_i.

3.4.3 Model Selection

The hyper-parameters of the various methods are tuned using a k-fold cross validation. However due to the non i.i.d. nature of the problem, we employ a stronger setup to ensure no information leakage. The dataset on which we are validating the performance is never subsequently used in the predictive performance estimation. The values for diffusion parameter β in LEDK and MEDK are sampled in $\{10^{-3}, 10^{-3}, 10^{-2}, 10^{-1}\}$, time steps t in MDK in $\{1, 10, 100\}$ and RLK parameter β in $\{1, 4, 7\}$. For CDNK, the degree threshold values, D, are sampled in $\{10,\ 15,\ 20\}$, clique size threshold, C, in $\{4,\ 5\}$, maximum radius, r^*, in $\{1,\ 2\}$, maximum distance, d^*, in $\{2,\ 3,\ 4\}$, number of clusters P in $\{5,\ 7\}$. Finally, the regularization trade off parameter of the SVM is sampled in $\{10^{-5},\ 10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1}, 1,\ 10,\ 10^{2}\}$.

3.5 Results and discussion

Figures 3.4 and 3.5 show the AUC-ROC performance of the different models using state-of-the-art graph node kernels and different variations of CDNK on BioGPS and Pathways networks, respectively. In what follows, we analyze the results to answer the two questions in Section 3.4.

Concerning the performance of different graph node kernels, CDNK variations outperform diffusion-based graph node kernels in all cases on BioGPS dataset and 8 out of 11 cases on Pathways. CDNK is ranked first when considering both the average AUC-ROC and the average rank (see Tables A.2 and A.3) with a difference, w.r.t. the best diffusion-based kernel, ranging from 5.4% to 10% and from 1.2% to 3.4% on BioGPS and Pathways, respectively. As a consequence, we can conclude that CDNK show state-of-the-art performance in graph node kernels.

Regarding the variations of CDNK, the integration of real valued vectors improves the performance in most cases. In particular, considering the use of discrete labels based on ontology, using side information helps to increase the performance in all diseases on BioGPS and in 9 out of 11 diseases on Pathways. Similarly, by employing auxiliary information associated to graph nodes, the performance of CNDK in case of using uniform discrete labels is also improved in 9 and 8 out of 11 on BioGPS and Pathways, respectively.

3.6 Summary

We have shown how decomposing a network in a set of connected sparse graphs allows us to take advantage of the discriminative power of CDNK, a novel decomposition kernel, to achieve state-of-the-art results. Moreover, we have also introduced the way to integrate “side” information in form of real valued vectors when it is available on graph to get even better performance of CDNK. In future work we will investigate how to *i*) decompose networks in a data driven way and *ii*) extend the CDNK approach to gene-disease association problems exploiting multiple heterogeneous information sources in a joint way.

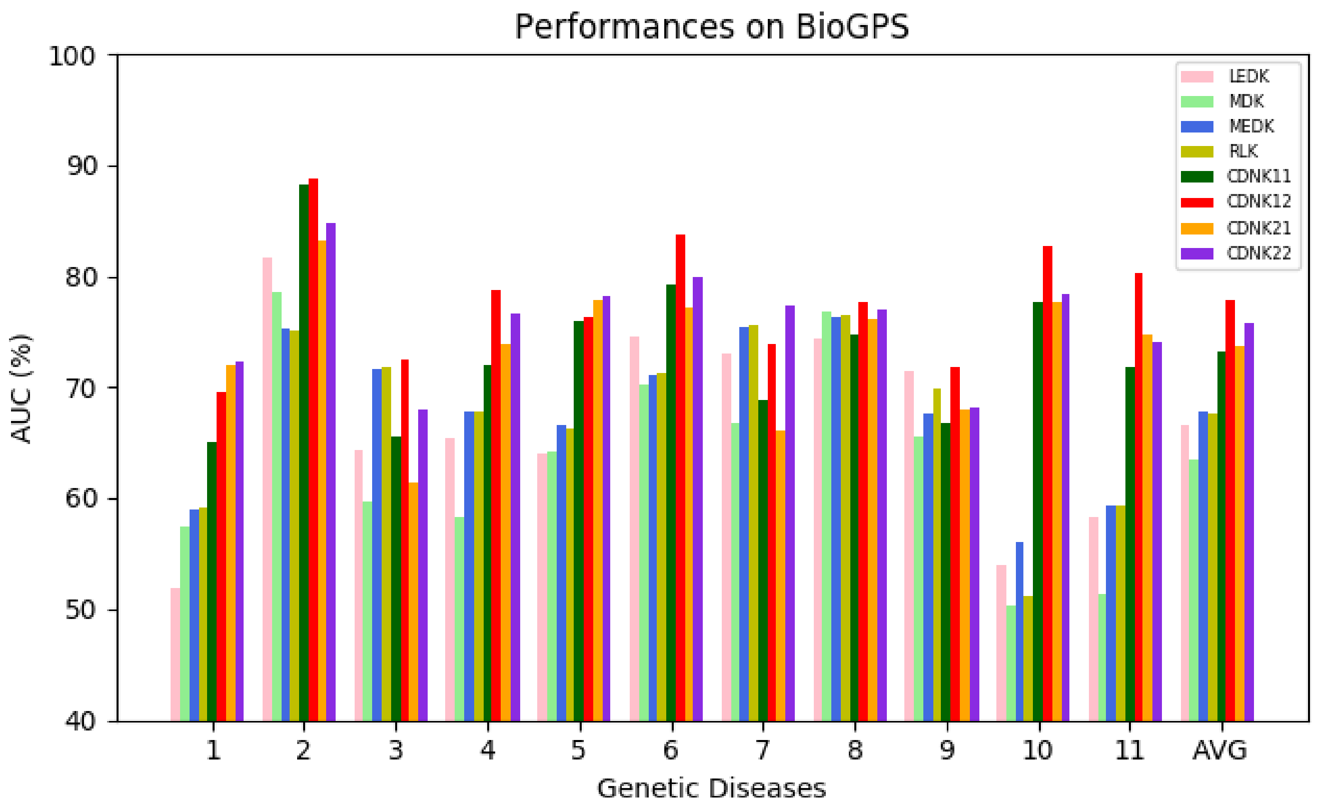

Figure 3.4: *AUC-ROC performance on 11 genetic diseases using network induced by the BioGPS. CDNK11: ontology for discrete labels, CDNK12: ontology for both discrete and vector labels, CDNK21: node degree for discrete labels and CDNK22: node degree for discrete labels and ontology for vector label.*

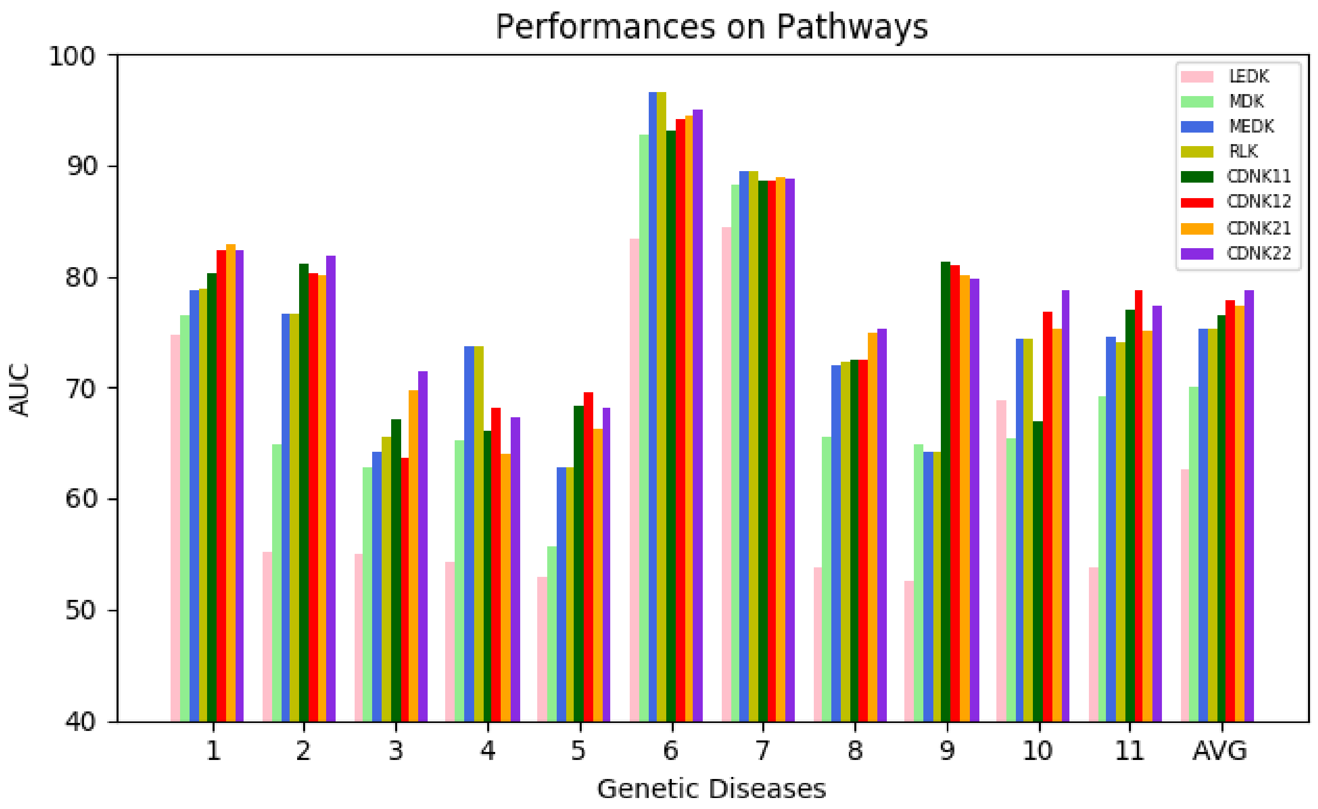

Figure 3.5: *AUC-ROC performance on 11 genetic diseases using network induced by the BioGPS. CDNK11: ontology for discrete labels, CDNK12: ontology for both discrete and vector labels, CDNK21: node degree for discrete labels and CDNK22: node degree for discrete labels and ontology for vector label.*

Chapter 4

Solutions for Graph Sparsity

One of the problems that we often face with when constructing graph-based learning systems is the graph sparsity problem, i.e. input graphs have high a number of missing links. In this case, the systems can only access to limited information, so they are not able to learn effectively. This leads to low performance of the learning systems. In this chapter, we aim at investigating and proposing solutions to deal with the graph sparsity problem.

4.1 Motivation

Despite the increase of relational data, which are best presented by graphs, in terms of diversity and amount, the number of relations between entities encoded in data is much smaller than the number of possible relations. This yields the problem of graph sparsity when we employ graphs to represent the entity relations. The graph sparsity problem often leads to low performance of graph-based learning systems since the information, which available, for learning is restricted. For this reason, we need effective solutions to overcome the graph sparsity issue.

An effective solution for the graph sparsity problem is to use link enrichment methods. Given a graph, a link prediction aims at finding the top un-observed links, i.e. links which have the highest possibilities of being missing links, to add into the graph. However, the performance of a link enrichment directly depends on the employed link prediction method. Therefore, in the following sections, we first propose a novel, effective link prediction method, named Joint Neighborhood Subgraphs Link Prediction.

We then propose a paradigm, called Link Enrichment for Diffusion-based Graph Node Kernels, to boost the performance of graph-based kernels that usually show low performance on sparse graphs.

4.2 Joint neighborhood subgraphs link prediction method

Due to the importance of link prediction, there is a high number of link prediction approaches which have been proposed and applied in a wide range of fields (see Section 2.8). They can be classified into supervised and unsupervised learning. Supervised methods normally show better accuracies compared to unsupervised methods. However, they face with much higher computational and memory complexity costs. Most link prediction methods in both categories implicitly represent the link prediction problem and the inference used to tackle it as a disjunction over the edges, that is, information on edges is propagated in such a fashion so that for a node to have k neighbors or $k+1$ does not make a drastic difference. We claim that this hypothesis is likely putting a cap on the discriminative power of classifiers and therefore we propose a novel supervised method, JNSL, that employs a conjunctive representation. We call the method "joint neighborhood subgraphs link prediction" (JNSL). The key idea here is to transform the link prediction task into a binary classification, on suitable small subgraphs, which we then solve using an efficient convolutional graph kernel method.

Given a labeled, unweighted graph, $G = (\mathbb{V}, \mathbb{E}, \mathcal{L}_1, \mathcal{L}_2)$, the set $\mathbb{E}$ is partitioned into the subset of observed links ($\mathcal{O}$) and the subset of unobserved links (U). The set U is further devided into two subsets: missing links (U_m), i.e. links that have not discovered, and non-links (U_n). We define the link prediction problem as the task of ranking candidate links, i.e. links in U, from the most to the least probable of being missing links. In the next sections, we first describe how links are represented in our method, and we then show how we cast link prediction as a binary classication problem.

4.2.1 Link encoding as subgraphs union

Most methods for link prediction compute pairwise nodes similarities treating the nodes defining the candidate edge independently. Instead we propose to jointly consider both candidate endpoint nodes together with their extended "context". To do so we build a graph starting from the two nodes and the underlying network. Given nodes u and v, we first extract the two neighborhood sets with a user defined radius r rooted at u and v to obtain

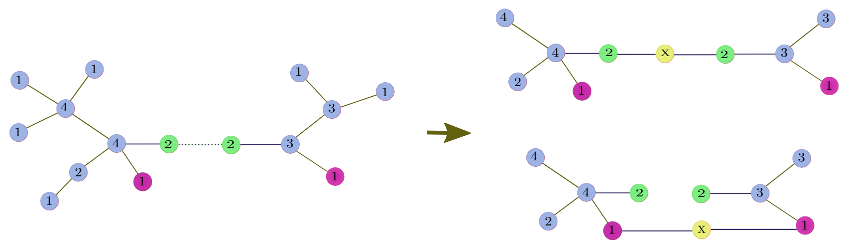

Figure 4.1: Left) We represent with solid lines edges belonging to the training material and with a dotted line edges belonging to the test material. Right) joint neighborhood subgraphs for an existing (green endpoints) (top) and a non existing (red endpoints) link (bottom) ($r = 2$). These graphs will receive a positive and a negative target, respectively.

$N_r(u)$ and $N_r(v)$, respectively. We then consider the graph $\mathcal{J}$ induced by the set union $N_r(u) \cup N_r(u)$. Finally we add an auxiliary node w and the necessary edges to connect it to u and v.

4.2.2 Joint neighborhood subgraphs link prediction

In the link prediction problem we are given a graph $G = (\mathbb{V}, \mathbb{E}, \mathcal{L}_1, \mathcal{L}_2)$ and a binary target vector $Y = \{y_{(0,0)}, y_{(0,1)}, \cdots, y_{(|V|,|V|)}\}$ where $y_{(u,v)} = 1$ if $(u, v) \in E$ and 0 otherwise. The training data is obtained considering a random subset of edges in $E^{tr} \in E$ and inducing a training graph $G^{tr} = (V, E^{tr})$. Note that the graph used for training does not contain any of the edges that will be queried in the test phase. The remaining edges $E^{ts} = E \setminus E^{tr}$ are used to partition the target vectors: $Y^{tr} = \{y_{(u,v)} | (u, v) \in E^{tr}\}$, $Y^{ts} = \{y_{(u,v)} | (u, v) \in E^{ts}\}$. We can now cast the problem as a standard classification problem in the domain of graphs. Given G^{tr} we build a train and test set as the corresponding joint neighborhood subgraphs as detailed in Section 4.2.1 and Figure 4.1. We can now compute a Gram matrix of the instances and solve the classification task using for example the efficient LinearSVC library [3].

4.2.3 Empirical evaluation

In this section, we carry out two separate experimental settings to evaluate the performance of JNSL and compare it with various link prediction methods which only exploit topological features.

In the first setting, we follow the experimental setting and procedure performed in [57] in which 6 datasets belonging to different domains are employed.

- *Protein* [84]: nodes are proteins and edges encodes a thresholded interaction confidence between proteins. It has 2,617 nodes and 11,855 links with an average degree of 9.1.

- *Metabolic* [89]: nodes are enzyme and metabolites, edges are present if the enzyme catalyzes for a reaction that include those chemical compounds. It has 668 nodes and 2,782 links with an average degree of 8.3.

- *Nips* [5]: nodes are authors at the NIPS conference from the first to the 12^{th} edition. Links encode the co-authorhip relation, i.e. if two authors have published a paper together. This network contains 2,865 nodes and 4,733 links with an average degree of 3.3.

- *Condmat* [51]: nodes are scientists working in condensed matter physics, edges encode co-authorship. This network has 14,230 nodes and 1,196 links with an average degree of 0.17.

- *Conflict* [35, 86]: nodes are countries and edges encode a conflict or a dispute. We have 130 nodes and 180 links in total in this network with an average degree of 2.5.

- *Powergrid* [42]: a network of electric powergrid in US. It has 4,941 nodes and 6,594 links. The average degree is 2.7.

We evaluate the performance of employed methods by splitting 10 times the data in a train and a test part. For *Protein*, *Metabolic*, *Nips* and *Conflict* networks, we use 10% of the edges to induce the training set while for *Condmat* and *Powergrid* we use 90% of the links. On *Condmat* and *Powergrid*, if we set a 10% for training, it would cause most nodes to have no observed links in the training set, so, it makes difficult to predictions based solely on the topological structure. The performance of each method is computed as the average of the AUC-ROC over the 10 rounds.

In the second setting, we aims at applying diffusion-based graph node kernels LEDK, MEDK, MDK, RLK (see Section 2.5.2) for link prediction and comparing their performances with JNSL. We test on three datasets BioGridgen, Omim and HPRD presented in 2.7. Similar to the first setting, we evaluate the performance of employed methods by splitting 10 times the data in a training (90%) and a testing part (10%). The performance of each method is computed as the average of the AUC-ROC over the 10 rounds.

Model Selection: The values of different parameters are chosen through the model selection. Different hyper-parameters are set by using a modified 3-fold cross-validation procedure on the training set, that is, always considering only the training network, we use one fold for fitting the parameters and the rest two folds for validating the effect of the hyper-parameter choice. We tune the values of radius for extracting subgraphs r in $\{1, 2\}$, λ in node label function in $\{10, 15\}$, for r^* and d^* parameters of NSPDK in $\{1, 2\}$ and $\{1, 2, 3\}$, respectively. The values for diffusion parameter β in LEDK and MEDK are sampled in $\{10^{-3}, 10^{-3}, 10^{-2}, 10^{-1}\}$, time steps t in MDK in $\{1, 10, 100\}$, and RLK parameter β in $\{1, 4, 7\}$. Finally, the regularization tradeoff of the SVM is picked up in $\{10^{-4}, 10^{-3}, 10^{-2},\ 10^{-1}, 1,\ 10,\ 10^2, 10^3,\ 10^4\}$.

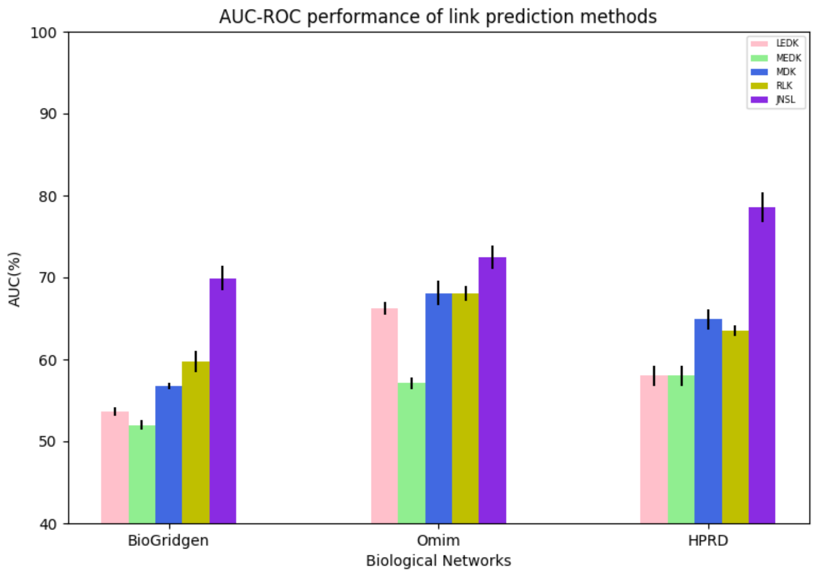

Figure 4.2: Performance of diffusion-based methods and JNSL on BioGridgen, Omim and HPRD.

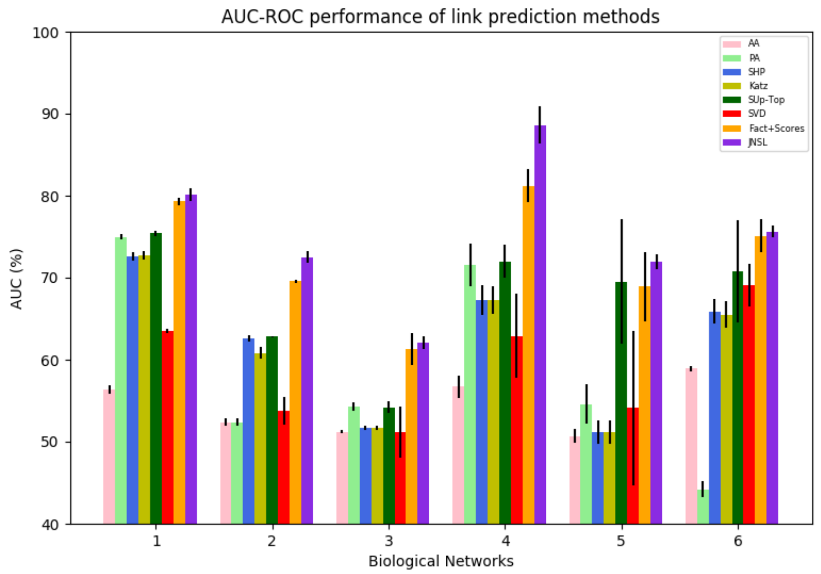

Figure 4.3: Performance of link prediction methods on 6 datasets: Protein: 1; Metabolic: 2; Nips: 3, Condmat: 4, Conflict: 5, PowerGrid: 6. Legend: AA: Adamic-Adar, PA [12] preferential Attachment, SHP: Shortest Path, Sup-Top [57]: Linear regression running on unsupervised scores, SVD [57]: Singular value decomposition, Fact+Scores [57]: Factorization with unsupervised scores.

4.2.4 Results and discussion

Figures 4.3 and 4.2 show the performance in AUC-ROC of different link prediction in the first and second experimental setting, respectively (see Tables B.2 and B.1 for more detail).

Concerning Figure 4.3, we can group methods into two groups based on their performances: supervised methods, AA, PA, SHP, Sup-Top SVD, and unsupervised methods, Fact+Scores, JNSL. The performance of supervised methods are considerably higher than unsuperivsed ones in most cases, except in the Conflict dataset where Sup-Top outperforms Fact+Scores, but with a very small difference. Concerning supervised methods, JNSL outperforms Fact-Scores in all cases. The difference between their performance is small in PowerGrid and Protein datasets with 0.5% and 0.8%, respectively. And the big gap is in the Condmat dataset with 7.4%.

In Figure 4.2, it can be clearly seen that JNSL outperforms the compared methods that use diffusion-based kernels which are unsupervised methods. In particular, JNSL gets from 10.2%, 4.4% and 13.7% higher w.r.t the rest methods on BioGridgen, Omim and HPRD, respectively. Related to diffusion-based prediction methods, MDK and RLK present better results, in all cases, than LEDK and MEDK.

4.2.5 Summary

We have presented a novel approach to link prediction in absence of side information that can effectively exploit the topological contextual information available in the neighborhood of each edge. We have empirically shown that this approach achieves very competitive results compared to other state-of-the-art methods.

For the comming work, we will first investigate how to make use of multiple and heterogeneous information sources when these are available for nodes and edges. We then apply link prediction methods in general and JNSL in particular to link enrichment for improving performance of learning systems when dealing with sparse graphs (see Section 4.3).

4.3 Link enrichment for diffusion-based graph node kernels

Diffusion-based graph node kernels, such as LEDK, MEDK, MDK and RLK (see Section 2.5.2), are popularly used to measure graph node similarities. Nevertheless, they show low discriminative capacity on sparse graphs due to two main reasons: *i*) the lower the average node degree is, the smaller the number of paths through which information can travel, and *ii*) missing links can end up separating a graph into multiple disconnected components. In this case, since information cannot travel across disconnected components, the similarity of nodes belonging to different components is null. To address these problems we propose to solve a link prediction task prior to the node similarity computation and start studying the question: *how can we improve the discriminative capacity of diffusion-based node kernels by using link prediction?* In this section, we take into account both the link prediction literature (presented in 2.8) and the diffusion-based graph node kernel literature (described in 2.5.2), select a subset of approaches in both categories that seem well suited, focus on a set of node predicting problems in the bioinformatics domain and empirically investigate the effectiveness of the combination of these approaches on the given predictive tasks. The encouraging result that we find is that all the strategies for link prediction we examined consistently enhance the performance on downstream predictive tasks, often significantly improving state of the art results.

4.3.1 Method

Often the relational information that defines the graph structure is incomplete because certain relations are not known at a given time or have not been yet investigated. When this happens the resulting graphs tend to become sparse and composed of several disconnected components. Diffusion-based kernels are not suited in these cases and so they show a degraded predictive capacity. Our key idea is to introduce a *link enrichment phase* that can address both issues and enhance the performance of diffusion-based kernels.

Given a link prediction algorithm M, a diffusion-based graph node kernel K and a sparse graph $G = (\mathbb{V}, \mathbb{E}, \mathcal{L}_1, \mathcal{L}_2)$ where $|\mathbb{V}| = n$ and $|\mathbb{E}| = m$, the link enrichment method consists of two phases:

1. enrichment: the link prediction algorithm M is used to score all unobserved links $\frac{n(n-1)}{2} - m$. A score associated to a link represents for its likelihood of being a missing link. The top scoring t links are added to E to obtain E' that defines the new graph $G' = (V, E')$;

2. kernel computation: the diffusion-based graph node kernel k is applied to graph $G^{'}$ to compute the kernel matrix K which captures the similarities between any couple of nodes, possibly belonging to different components in the graph G.

The kernel matrix K can be used directly by a kernelized learning algorithm, such as a support vector machine, to make predictive inferences.

4.3.2 Empirical evaluation

To empirically study the answer to the question: *how can we improve node similarity using link prediction?* we would need to define a taxonomy of prediction problems on graphs that make use of the notion of node similarity and analyze which link prediction strategies can be effectively coupled with specific node similarity computation techniques for each given class of problems. In addition we should also study the quantitative relation between the degree of missingness and the size of the improvement offered by prepending the link prediction to the node similarity assessment. In this section, we start such endeavor restricting the type of predictive problems to that of node ranking in the sub-domain of gene-disease association studies with a fixed but unknown degree of missingness given by the current medical knowledge. More in detail, the task, known as *gene prioritization*, consists of ranking candidate genes based on their probabilities to be related to a disease on the basis of a given set of genes experimentally known to be associated to the disease of interest. We have studied the proposed approach on the following 4 datasets: BioGPS, HPRD, Phenotype similarity and Biogridphys which we described in Section 2.7.

Evaluation method: to evaluate the performance of the diffusion kernels, we follow the experimetal setting from [24] and used in Section 3.4. However, we employ 14 disease-gene associations with at least 30 confirmed genes (see Table B.3). For each disease, we construct a positive set $\mathcal{P}$ with all confirmed disease genes. To build the negative set $\mathcal{N}$, we randomly sample a set of genes that are associated at least to one disease class, but not related to the class which defines the positive set such that $|\mathcal{N}| = \frac{1}{2}|\mathcal{P}|$. We replicate this procedure 5 times such that the positive set is held constant, while the negative set varies. We assess the performance of kernels via a modified 3-fold cross-validation, where, after partitioning the dataset $\mathcal{P} \cup \mathcal{U}$ in 3 folds, we use one fold for training a model using SVM, and the two remaining folds for testing. For each test gene g_i, the model returns a score s_i proportional to the likelihood of being associated to the disease. Next a decision score q_i is computed as the top percentage value of s_i among all candidate gene scores. We collect all decision scores for every test genes to compute the area under the curve for the receiver operating characteristic (AUC-ROC). The final performance on the disease class is obtained by taking the average over 5 trials and 3 folds for each trial.

Model selection: the hyper-parameters of the various methods are set using a 3-fold on training set as described previously. We try the values for LEDK and MEDK in $\{0.01, 0.05, 0.1\}$, time steps in MDK in $\{3, 5, 10\}$ and RLK parameter in $\{0.01, 0.1, 1\}$. For CDNK, we set the degree threshold value D in $\{10, 15, 20\}$, the clique size threshold C in $\{4, 5\}$, the maximum radius r^* in $\{1, 2\}$, the maximum distance d^* in $\{2, 3, 4\}$. The number of links used for the enrichment are chosen in $\{40\%, 50\%, 60\%, 70\%\}$ of the number of existing links. Finally, the regularization tradeoff in the SVM is chosen in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3,\ 10^4\}$.

4.3.3 Results and discussion

Figures 4.4, 4.5, 4.6 and 4.7 show the average performance on BioGPS, Biogridphys, HPRD and Omim, respectively. Note that each resource yields a graph with different characteristic sparsity and number of components. We compare the average AUC-ROC scores in two scenarios: plain diffusion kernel (in green color) and diffusion kernel on a modified graph G' (in red color) which includes the novel edges identified by a link prediction. Here we report the aggregated results (a detailed breakdown is available in Tables B.4 - B.12).

The noteworthy result is how consistent the result is: each link prediction method improves each diffusion kernel algorithm, and on average using link prediction yields a 15% to 20% relative error reduction for diffusion-based kernels. What varies is the amount of improvement, which depends on the coupling between the four elements: the disease, the information source, the link prediction method and the diffusion kernel algorithm. In particular, we obtain that the largest improvement is obtained for disease 3 (Connective) where we have a maximum improvement of 20% ROC points, while the minimum improvement is for disease 8 (Immunological) with a minimal improvement of 0% ROC points. On average, the largest improvement is of 13% ROC points, while the smallest improvement is on average of 1% ROC point. Such stable results are of interest since diffusion-based kernels are currently state-of-the-art for gene-disease prioritization tasks, and hence a technique that can offer a consistent and relatively large improvement can have important practical consequences in the understanding of disease mechanisms.

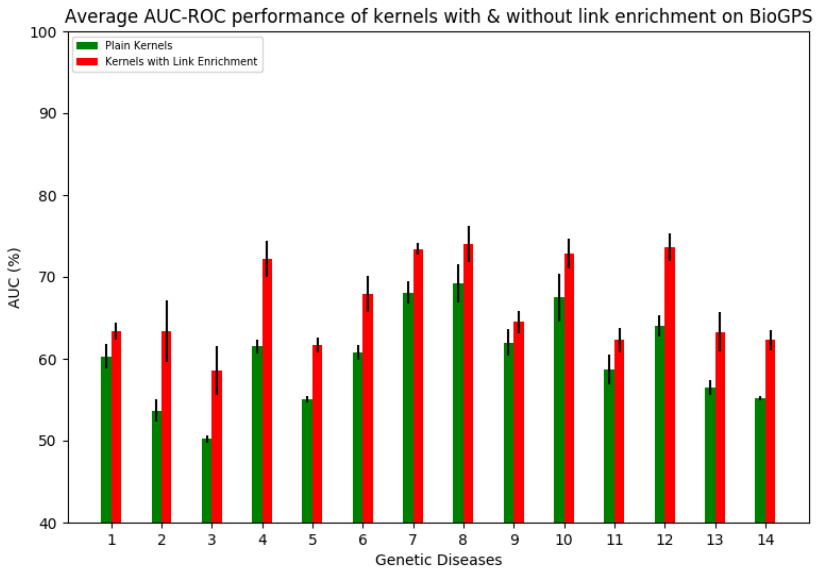

Figure 4.4: The AUC-ROC performance using kernels with and without using link enrichment on BioGPS. See Table B.3 for ID-disease list.

4.3.4 Summary

In this section, we have proposed the notion of *link enrichment* for diffusion-based graph node kernels, that is, the idea of carrying out the computation of information diffusion on a graph that contains edges identified by link prediction approaches. We have discovered a surprisingly robust signal that indicates that diffusion-based node kernels consistently benefit from the coupling with similarity-based link prediction techniques on large scale datasets in biological domains.

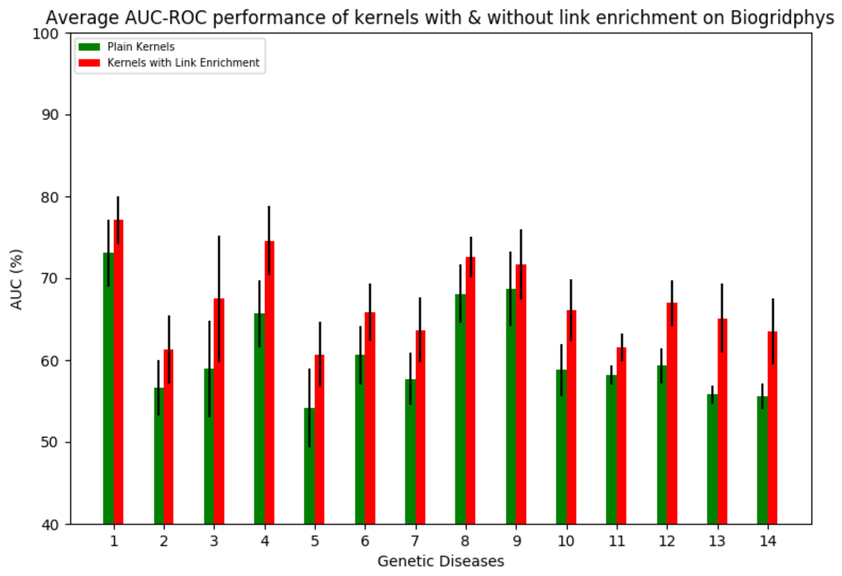

Figure 4.5: The AUC-ROC performance using kernels with and without using link enrichment on Biogridphys. See Table B.3 for ID-disease list.

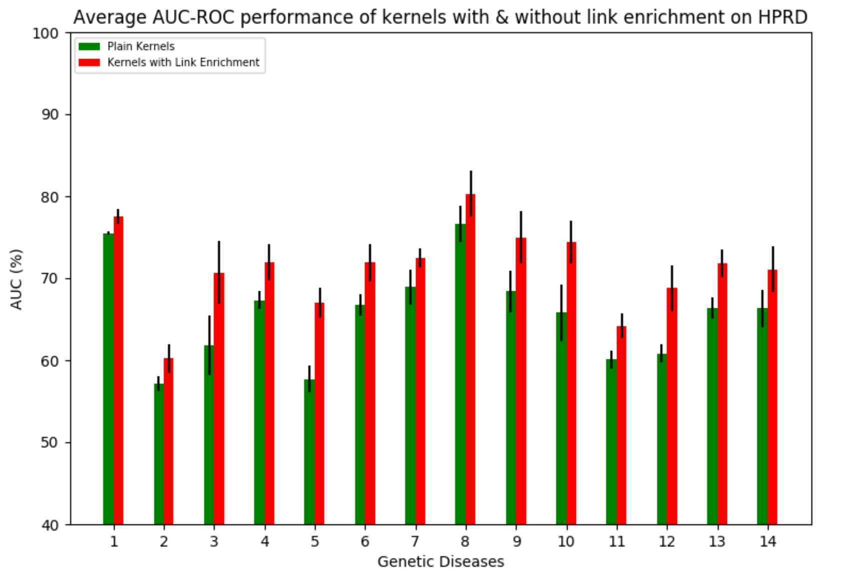

Figure 4.6: The AUC-ROC performance using kernels with and without using link enrichment on HPRD. See Table B.3 for ID-disease list.

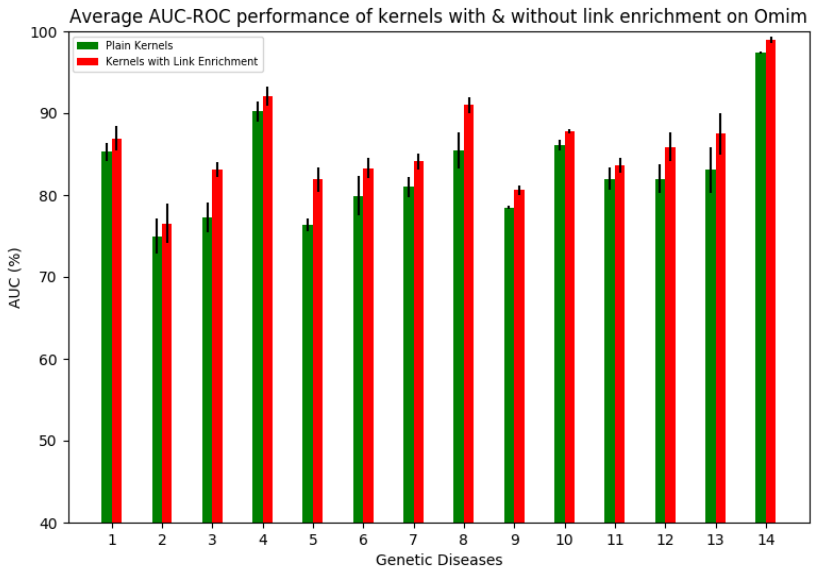

Figure 4.7: The AUC-ROC performance using kernels with and without using link enrichment on Omim. See Table B.3 for ID-disease list.

Chapter 5

Graph-based Data Integration Approaches

As discussed in the Chapter 1, there is a high need of proposing graph-based integration methods which allow to build high performance graph-based predictive systems. The task of designing such methods is not trivial due to its requirements. In particular, a good graph-based integration should possess the following features. First, it needs to effectively exploit the complementary property of graph combination. Indeed, each graph encodes an aspect of a considered phenomenon. Therefore, combining graphs together would bring a more unified view of that phenomenon which is useful for leanring systems. Second, it is required to be scalable to deal with the constant increase of data sources resulting from different researches in which each research views the phenomenon from a certain angle. Third, it should be efficient, i.e. it is required to execute in a reasonable time and with limited memory consumption. Last, the performance of a graph-based system using graph integration shows better performance than it using any single graph.

In this chapter, we propose a general kernel-based framework for graph-based biological data integration, named Graph-one, that meets the above requirements. In what follows, we first show the flow of Graph-one, followed by the description for each step of the flow. We then evaluate Graph-one by introducing different variations of Graph-one created under the general framework for disease gene prioritization.

5.1 Graph-one: a general kernel-based framework for graph-based data integration

In this section, we describe Graph-one, a kernel-based general framework for graph-based data integration.

5.1.1 Graph-one flow

Figure 5.1.2 shows the flow of Graph-one which aims at finding an effective and efficient kernel that represents a set of data sources. Precisely, our framework consists of four main steps: graph construction, graph layer combination, kernel definition and kernel learning. Given a set of data sources that encode information of a certain phenomenon, $S = \{S_1, S_2, \dots, S_p\}$. In the following, we explain each step in detail.

- Graph construction: for each data source $S_i \in S$, we build an undirected, unweighted, labeled graph G_i. As a consequence, we obtain a set of graphs $\{G_1, G_2, \dots, G_p\}$. We consider this graph set as a single graph G where each graph of the set is one layer, $G = \{L_1, L_2, \dots, L_n\}$.

- Graph layer combination: we allow different graph layers to be *combined* in specific ways. By doing so, a combined layer would have richer information w.r.t every single layer used to combine. Therefore, it can be useful for learning process. For example, sparse graph layers can be combined to get a denser layer which allows diffusion-based kernels to perform better. As a consequence of the combining process, we obtain a graph $G' = \{L_1, L_2, \dots, L_q\}$ such that $q \leq p$.

- Kernel definition: we apply suitable graph node kernels to each graph layer $L_i \in G'$ (see Section 2.5.2) with various selected parameter values. Given a graph node kernel applying a layer, different tuple values of its parameters define different kernel matrices. Thus, we obtain a set of kernel matrices in which each matrix encodes a different node similarity measure in a layer.

- Kernel learning: we employ UEasyMKL (see Section 5.1.2), a scalable MKL algorithm that is able to effficiently deal with unbalanced settings, to get an optimal positive linear combination of predefined computed kernels in a data driven way. As a result, we achieve a kernel which is likely to be better than every base kernel.

The final kernel can be fed into any kernel machine to build a model for different tasks of interest: classification, regression, etc.

Graph-one meets the requirements for graph-based data integration described above. First, it allows to define various kernels from each graph layer in which each kernel would capture the node similarity in a specific way. The obtained kernels defined on all graph layers are combined into a single one. Therefore, the final kernel encodes a unified view of information from all layers. Second, Graph-one is a scalable and efficient framework due to the properties of the adopted multiple kernel learning, UEasyMKL. UEasyMKL can deal with an arbitrary number of kernels using a fixed amount of memory and a linearly increasing computation time w.r.t the number of defined kernels. Interestingly, Graph-one is desinged to handle with unbalanced settings, i.e. high difference between number of positive and negative examples. Moreover, the graph layer combination step in the framework lets users to decide which graph layers should be combined. The combination decision are based on propertices of graph layers. The performance of Graph-one is evaluated in disease gene prioritization context (see in Section 5.2).

5.1.2 Scalable unbalanced multiple kernel learning: UEasyMKL

In Section 2.10, we introduce EasyMKL, a scalable, efficient kernel integration approach. However, in many applications, we often observe strong imbalance between the number of positive and negative examples, making it more difficult to build high performance learning systems. Disease gene prioritization is a typical example where given a genetic disease, the number of positive genes known to be associated to the disease is much smaller than the number of the unknown genes. For this reason, inspired by a previous work proposed in [66], we propose a novel MKL algorithm, based on EasyMKL, that we name UEasyMKL. UEasyMKL is not only scalable and efficient, as it inherits from EasyMKL, but also able to deal with unbalanced settings.

In order to clearly present our method, we first need to highlight the different contributions given by positive and unlabelled examples. In many cases, a portion of unlabeled instances are selected and considered as the negative ones. We define K^+, K^- and K^{+-} the sub-matrices of K^s, kernel concerning the positive set only, pertaining to positive-positive, unlabelled-unlabelled and positive-unlabelled instance pairs, respectively. Schematically, we have:

$$K^s = \begin{pmatrix} K^+ & K^{+-} \\ K^{-+} & K^- \end{pmatrix},$$

where K^{-+} is the transpose of K^{+-}. In other words, K^{+} contains the degree of similarities between positive instances, K^{-} contains the degree of similarities between unlabelled instances and K^{+-} includes the degree of similarities between positive-unlabelled example pairs. In the same way, we define γ_+ and γ_- as the probability vectors associated to positive and unlabelled examples, respectively.

Under this change of variables, we reformulate eq. 2.26 as:

$$\min_{\gamma\in\Gamma} \gamma_+^\top K^+\gamma_+ \;+\; 2\,\gamma_+^\top K^{+-}\gamma_- \;+\; \gamma_-^\top K^-\gamma_- \;+\; \lambda_+\gamma_+^\top\gamma_+ \;+\; \lambda_-\gamma_-^\top\gamma_- \,.$$

In this new formulation, the original EasyMKL problem is obtained by setting $\lambda_+ = \lambda_- = \frac{\lambda}{1-\lambda}$. However, due to the unbalanced PU nature of the problem, i.e. the number of positive instances is much smaller than the number of unlabeled ones, we are interested in using two different regularizationss. In our case, we decide to fix *a priori* the regularization parameter $\lambda_- = +\infty$, corresponding to fixing $\lambda = 1$ over unlabelled examples only. Then, the solution of part of the objective function is defined by the uniform distribution $\gamma_- = (\frac{1}{n}, \frac{1}{n}, \dots \frac{1}{n}) \equiv u$, where n is the number of unlabelled examples.

We inject this analytic solution of part of the problem in our objective function as

$$\min_{\gamma\in\Gamma^+} \gamma_+^\top K^+\gamma_+ \;-\; 2\,\gamma_+^\top K^{+-}u \;+\; u^\top K^-u \;+\; \lambda_+\gamma_+^\top\gamma_+ \;+\; \lambda_-u^\top u,$$

where $\Gamma^+ = \{\gamma \in \mathbb{R}_+^m | \sum_{i:y_i=1}\gamma_i = 1, \gamma_j = 1/n\;\forall j : y_j = -1\}$ is the probability distribution domain where the distributions over the unlabelled examples correspond to the uniform distribution. It is trivial that $u^\top K^- u$ and $\lambda_- u^\top u$ are independent from the γ_+ variable. Then, they can be removed from the objective function obtaining

$$\min_{\gamma\in\Gamma^+} \gamma_+^\top K^+\gamma_+ - 2\,\gamma_+^\top K^{+-}u + \lambda_+\gamma_+^\top\gamma_+. \tag{5.1}$$

In this expression, we only need to consider the entries of the kernel K^s concerning the positive set, avoiding all the entries with indices in the unlabelled set. The complexity becomes quadratic in the number of positive examples m, which is always much smaller than the number of examples to prioritize. Moreover, this algorithm still depends linearly on the number of kernels R and the overall time complexity is then $\mathcal{O}(m^2 \cdot R)$. In this way, we greatly simplify the optimization problem, while being able to take into account the diverse amount of noise present in positive and unlabelled example sets.

Like in Section 2.10, after solving the problem of eq. 5.1 we use eq. 2.27 to compute the optimal kernel weights. Next, we solve again the Scuba optimization problem to get the final optimal probability distribution γ^*. The likelihood to be positive of every test example is given by the vector of scores s defined as

$$s = K^* Y \gamma^*, \tag{5.2}$$

where K^* is the final kernel matrix, computed by $K^* = \sum_r^R \eta_r^* K_r$. We apply the formula 5.3 to compute scores for the test examples as:

$$s = K^{t*} Y \gamma^*\,, \tag{5.3}$$

where K^{t*} is the final test kernel matrix, obtained by taking the weighted sum over all test kernel matrices:

$$K^{t*} = \textstyle\sum_r \eta_r^* K_r^t,$$

K^{t*}, K_r^ts are matrices of size (p, q) with p, q are the number of test examples and training examples, respectively. More particularly, a score for s_i of the test example g_i is computed as:

$$s_i = \textstyle\sum_j y_j \gamma_j^* K_{ij}^{t*}.$$

In words, s_i is the weighted sum over the weighted similarities between the test example g_i and all examples in the training set g_js. Once we get the scores for test examples, candidate examples are then prioritized based on the score values.

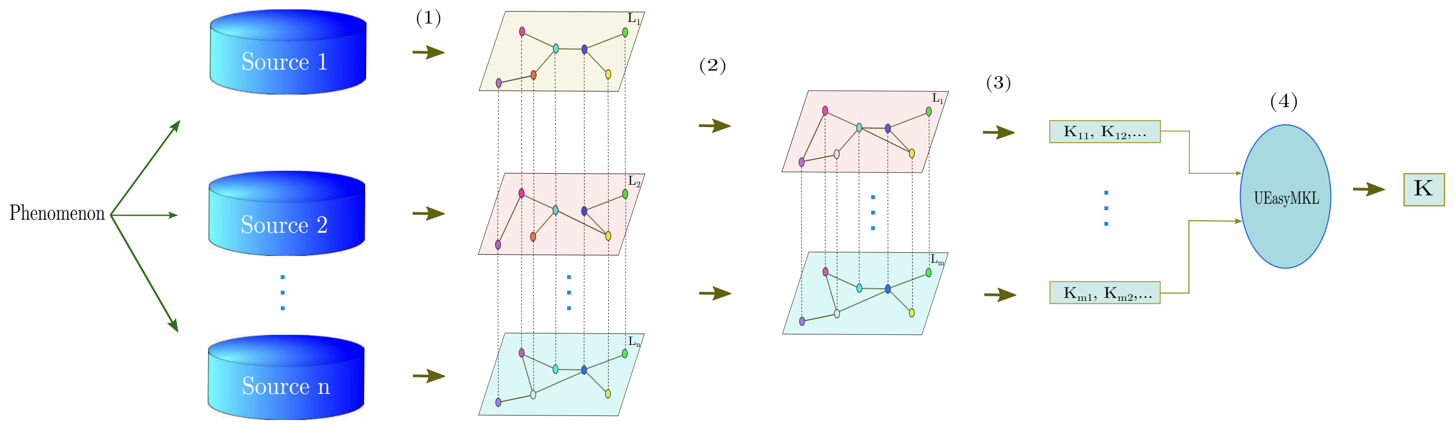

Figure 5.1: Graph-one general framework flow for graph-based data integration where (1) Graph construction, (2) Layer combination , (3) Kernel definition, (4) Kernel learning

5.2 Graph-one for disease gene prioritization

In this section, we evaluate the performance of Graph-one through the disease gene prioritization, one of the main tasks in Biomedicine as discussed in Section 2.6. Since Graph-one is a general framework which allow the user to define their preference in some steps. Different choices leads to different variations of Graph-one. Here, we propose three variations for disease gene prioritization: Scuba, PLC and DIGI. These variations are evaluated and compared with various methods by using two separate experimental settings: Cross-validation and Unbiased evaluation. Through performing these experiments, we would like to have answers for following questions.

- Does Graph-one show better performance comparing to other methods for disease gene prioritization?
- Does Graph-one with graph integration show better results in comparison with the use of single graph?
- Is Graph-one stable in term of performance between its different variations?

5.2.1 Graph-one variations for disease gene prioritization

We propose different variations created under Graph-one general framework.

Graph-one: Scuba

We introduce Scuba as an variation of Graph-one. Consider a n-layer graph $G = G_1, G_2, \dots, G_n$ where $G_i \in G$ is derived from the source $S_i \in S$. In this method, kernels are defined on every single graph layer, i.e. different kernels with different parameter values are applied on each $G_i \in \mathcal{G}$. As a consequence, we get a set of kernel matrices $\mathcal{K}_i = \{K_{i1}, K_{i2}, \dots, K_{iH}\}$ for each layer G_i. By collecting all kernels from all $\mathcal{K}_i$, we achieve a final kernel matrix set $\mathcal{K} = \bigcup_{i \in I}^{n} \mathcal{K}_i$ comprising $L \cdot H$ matrices. Next, all matrices in $\mathcal{K}$ and the training are fed into UEasyMKL to obtain the final kernel K. In this way, we directly use UEasyMKL to perform an automatic selection of optimal kernel parameters. As the method is an instance of Graph-one, it inherits the strength of the general graph integration framework Graph-one.

We apply Scuba for experiments in 5.2.2 with three diffusion-based kernels. Each kernel is used to define three kernel matrices from each graph layer. The value for λ_+ is chosen by employing k-fold cross validation on the training set, using the the grid of values $\{0,\ 0.1,\ 0.2,\ \ldots,1\}$. Kernel parameter values are set as follows: $\{0.01,\ 0.04,\ 0.07\}$ for MEDK, as suggested in [26] and used in [24], $\{2,\ 4,\ 6\}$ for MDK and $\{1,\ 10,\ 100\}$ for RLK, as suggested in [32]. Besides, we use Scuba in two cases: using plain diffusion-based kernels (Scuba1), using diffusion-based kernels with association with link enrichment (Scuba2).

Graph-one: disjuctive interconnection graph integration (DIGI)

We consider a graph including n layers $G = \{G_1, G_2, \ldots, G_n\}$ where $G_i \in G$ is derived from $S_i \in S$. In this method, we first process and combine all graph layers in a specific way. We then apply CDNK, which is defined in Chapter 3, with different parameter value tuples on the obtained graph to get a set of kernel matrices. These kernel matrices are then fed into UEasyMKL to get a final kernel matrix. More precisely, the method consists of the following steps:

- Graph decomposition: for each graph layer $G_i \in G$, we apply the network decomposition procedure proposed in 3.2.1 to have a graph composed of linked sparse connected components.

- Graph combination: consider a layer tuple G_i, G_j $(G_i, G_j \in G)$ and two nodes $u \in G_i,\ v \in G_j$ such that u and v are both represent for a same entity, we connect u and v by a "disjuctive" link. As the consequence, we achieve a graph $G^{'}$ whose nodes represeting the same entities are linked by disjuctive links. Figure 5.2 is a visualization for this graph integration idea.

- Similarity definition: we employ CDNK to define similarity for the similarity for any couple of genes. The similarity between two genes g_i and g_j is the summation of similarity of between their corresponding nodes on all layers.

For experiments in 5.2.2, we use values for parameters of graph decomposition procedure and CDNK the same as experiments in 3.4

Graph-one: partial layer combination (PLC)

Here we propose another variation of Graph-one, named PLC. Given a graph with a set of layers, we first combine different graph layers together by using the same graph combination method proposed in DIGI. We then define kernel matrices on each layer of the obtained graph where only CDNK is

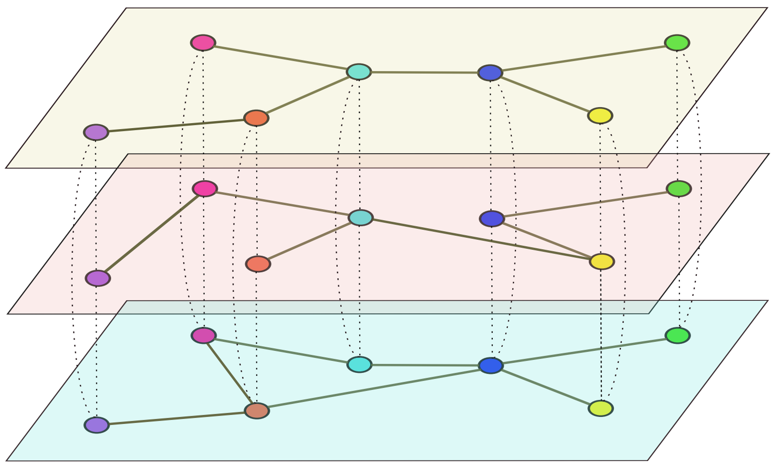
Figure 5.2: Graph Disjunctive Interconnection Illustration

applied on the combined layers, while various node kernels are applied each of other layers.

In Cross-validation setting presented in the next section, PLC combine BioGPS and HPRD layers since they are sparser comparing to Pathways layer.

5.2.2 Experiments and results

Cross-validation

As a first evaluation of Graph-one, we followed the experimental protocol used by Chen *et al* to test predictive performance of other prioritization methods [23]. In this setting, we employed three data sets: BioGPS, HPRD and Pathways (see the section 2.7), which we borrowed from the authors of the work. To perform the experiments, we employed known gene-disease associations from OMIM, grouped into 20 classes on the basis of disease relatedness by Goh *et al* [37]. Among those classes we selected the 12 with at least 30 confirmed genes. We then built a training set consisting of a positive set $\mathcal{P}$ and an unlabelled set $\mathcal{U}$ for each of them. $\mathcal{P}$ contains all its disease gene members. $\mathcal{U}$ is constructed by randomly picking genes from known disease genes such that $|\mathcal{U}| = \frac{1}{2}|\mathcal{P}|$. The unlabelled genes relate to at least one disease class, but do not relate to the current class. We chose the genes in $\mathcal{U}$ from the other disease genes because we assume that they were less likely to be associated to the considered class. In fact, disease genes are generally more studied and a potential association has more chances to have already been identified.

After that, leave-one-out cross validation was used to evaluate the performance of the algorithm. Iteratively, every gene in the training set was selected to be the test gene and the remaining genes in $\mathcal{P}$ and $\mathcal{U}$ were used to train the model. Once the model was trained, a score list for the test gene and all genes associated to no disease was computed. Then, we computed a decision score for each test gene representing the percentage of candidate genes ranked lower than it. We collected all decision scores for every gene in all disease classes to form a global decision score list. The performance of Scuba was measured by calculating the area under the curve (AUC) in the receiver-operating-characteristic plot obtained from the decision score list. The AUC expresses the probability that a randomly chosen disease gene is ranked above a randomly picked non-disease gene for any disease class.

In this setting, we compare not only the performance between Graph-one variations, but also the performance between Graph-one with other methods for disease gene prioritization, including F3C3 [23], MRF [24], DIR [26], GeneWanderer [46] and Avg, an implementation that takes an average of kernels defined on graph layers as the final representation for datasets).

The Figure 5.3 shows the performance of Graph-one variations (Scuba1, Scuba2, PLC and DIGI) along with other methods, while the Figure 5.4 presents the performance of Scuba in two cases: using single graph and graph integration. It can be seen from the Table 5.3 that all variations of Graph-one perform significantly better than the other methods (also see the Table C.1), getting an AUC from 3.8% and 4.9% greater than the F3PC and Avg, respectively. Among Graph-one variations, DIGI and Scuba2 show slightly better than Scuba1 and PLC. This first proves again the benifit of stacking link enrichment with diffusion-based kernels. Second, this illustrates the stability of performance between Graph-one variations. Besides, the effectiveness of Graph-one for data integration is shown in the Table 5.4. Graph-one with data integration presents higher performance in most of diseases (11 out of 12) w.r.t its that uses single data source.

Unbiased evaluation

Although the previous evaluation is useful to compare Graph-one with other methods, predictive performance in cross-validation experiments may be inflated compared to real applications. Indeed, the retrieval of known disease genes can be facilitated by various means. One mean is the crosstalk between data repositories: for example, KEGG [64] draws its information also from medical literature. Moreover, often the discovery of the link between a gene and a disease coincides with the discovery of a functional annotation or of a molecular interaction. In practice, instead, researchers are interested in novel associations, which in most cases are harder to find due to a lack of information around them.

To achieve a thorough evaluation of Graph-one, we test it in a more realistic setting, following the work of Börnigen *et al* [17]. In this study, eight gene prioritization web tools were benchmarked as follows. Newly discovered gene-disease associations were collected over a timespan of six months, gathering 42 test genes associated to a range of disorders. As soon as a new association was discovered, each web tool was queried with a disorder-specific set of positive genes $\mathcal{P}$ to prioritize a set of candidates $\mathcal{U}$ containing the corresponding test gene (or to prioritize the whole genome where possible). In other words, the test gene was treated as unlabelled to simulate the re-discovery of its association with the disease. Rank positions of the 42 test genes were ultimately used to assess the ability of the tools to successfully prioritize disease genes. The idea behind this procedure is to anticipate the integration of the associations in the data sources and so avoid biased predictions.

In order to test Graph-one in this setting, we backdate our data to a time prior to May 15, 2010 by employing String v8.2 data [43] and Omim (see the section 2.7). After that, we recovere positive sets and test genes from the original publication and we followe its experimental protocol as follows [17]. We performe prioritizations for each test gene in two distinct cases: genome-wide and candidate set-based prioritizations. In any genome-wide prioritization all genes in the String dataset - except those in $\mathcal{P}$ - belong to $\mathcal{U}$ and were prioritized. In any candidate set-based prioritization, the set of candidates $\mathcal{U}$ was constructed by considering all genes with Ensembl [1] gene identifier within the chromosomal regions around the test gene, in order to get on average 100 candidates. In both cases, we normalize the ranking positions over the total number of considered genes in order to get median, mean and standard deviation of the normalize ranks for test genes. We also compute the true positive rate (TPR) relatively to some representative thresholds (5%, 10% and 30% of the ranking) and the AUC obtained by averaging over the 42 prioritizations.

Along with Graph-one, we evaluated in this setting also MKL1class [91] and ProDiGe [60], two state-of-the-art kernel based gene prioritization methods. In Table 5.1, it is possible to see performances for all three methods. The significance of rank median differences between Scuba and competing methods was assessed by Wilcoxon signed rank tests, one for each comparison. At a significance threshold of 0.05, Scuba achieves significantly higher performances in genome-wide tasks compared to both baselines. In the candidate set-based setting, it performs significantly better than ProDiGe and better, although not significantly, than MKL1class. These differences can be visually appreciated in Figure 5.5, where we compare the rank distributions of test genes obtained by the three methods. Scuba and MKL1class present moderate rank differences, particularly in the central region of the ranks. On the other hand, differences between Scuba and ProDiGe are smaller (Pearson $r = 0.98$ in both cases) and almost all in favour of Scuba.

In Table 5.2 we show results for Graph-one compared to the results obtained in the work of Börnigen *et al*, pertaining to eight prioritization systems [17]. In genome-wide predictions, Scuba dominates over the other tools. On predictions over smaller candidate sets, it is still competitive although best results are achieved by GeneDistiller [70], Endeavour [8] and ToppGene [25]. It is important to underline that in this case considered tools rely on different data sources, so we are comparing different prioritization systems rather different algorithms. Furthermore, tools are in some cases unable to provide an answer to a given task, depending on the underlying data sources (for more details see the original work [17]). We report the fraction of prioritizations on which tools are actually evaluated as response rate. This table has the purpose of showing the potentiality of Scuba relatively to what is easily accessible by non-bioinformaticians. However, since we used the String data for instance Graph-one:Scuba1 is directly comparable with Pinta [17, 63].

Next, we expanded this validation by employing gene-disease annotations derived from the Human Phenotype Ontology (HPO) [47]. This resource gathers information from several databases and makes available its monthly updates, permitting to trace the annotations history. We downloaded the HPO build 29 - dating March 2013 - and build 117 of February 2017. We compared the two annotations corresponding to these versions of HPO and extracted the gene-disease associations that were added in this time gap. We concentrated on the multifactorial diseases covered in the previous analysis, that could possibly have some previously undiscovered associations. We thus analyzed how the obtained genes are ranked in genome-wide prioritizations of the previous analysis, applying the same performance measures as before. The outcome is an analogous evaluation, but this time target genes are those extracted from HPO.

In Table 5.3 results for Scuba, MKL1class and ProDiGe are shown. We can observe a slightly different trend compared to previous results, with Scuba and ProDiGe having very close performance and MKL1class being significantly worse than Scuba. As a confirmation, in Figure 5.5 we can see that there is no clear difference between test genes rank distributions for Scuba and ProDiGe. Instead, MKL1class ranks several test genes neatly lower compared to Scuba, with the associated Pearson correlation coefficient dropping to $r = 0.85$.

5.2.3 Discussion

Gene prioritization is progressively becoming essential in molecular biology studies. In fact, we are assisting to a continuous proliferation of a variety of *omic* data brought by technological advances. In the near future it is then likely that more heterogeneous knowledge will have to be combined. Moreover, the classes of biological agents to be prioritized are going to enlarge. For instance, we are only beginning to understand the complex regulation machinery involving non-coding RNA and epigenetic agents. It is estimated that around 90.000 human long non coding genes exist, whose functional implications are progressively emerging [93]. Facing this challenge, the development of novel methods is still strongly needed in order to enhance predictive power and efficiency.

Compared to the considered benchmark kernel methods - MKL1class and ProDiGe - Scuba has some important advantages. ProDiGe is one of the first proposed kernel-based PU learning method for gene prioritization [60]. It implements a PU learning strategy based on a biased SVM, which over-weights positive examples during training. In order to reach scalability to large datasets, it leverages a bagging procedure. Like ProDiGe, Scuba implements a learning strategy based on a binary classification set up, but from a different perspective. In a PU problem, the information on positive labels is assumed secure, while the information on negative labels is not. In terms of margin optimization, this translates in unbalanced entropy on the probability distributions associated to the two sets of training examples. It is then required to regularize more on the unlabelled class - having higher entropy - and in the limit of maximum uncertainty we get the uniform distribution.

MKL1class implements another effective approach for data integration, namely single class learning. This means that the model is obtained solely based on the distribution of known disease genes, disregarding unlabelled ones. Scuba has enhanced scalability compared to MKL1class, as it involves the optimization of the 1-norm of the margin vector from the different kernels. In contrast, MKL1class optimizes its 2-norm, which is more computationally demanding. Importantly, another distinctive feature of Scuba is a time complexity dependent on the number of positive examples and not on the number of total examples. As a consequence, Scuba can exploit the

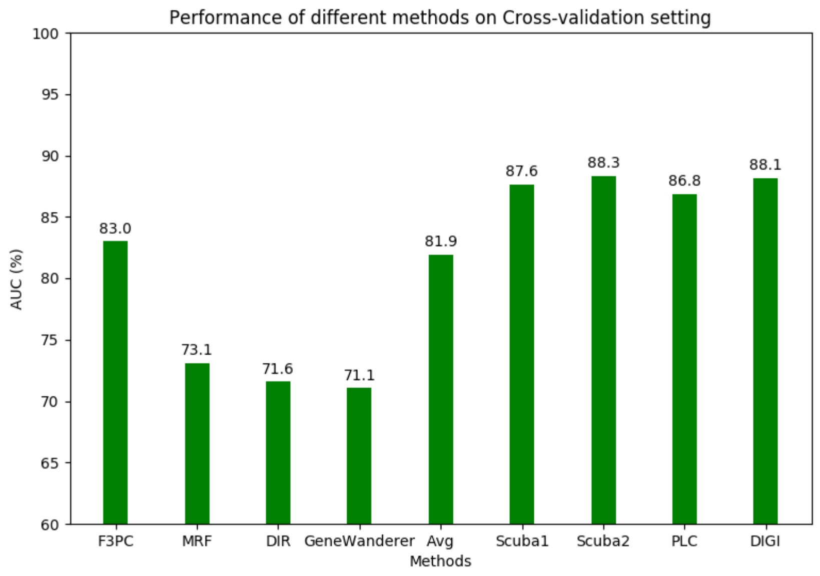

Figure 5.3: The performance of different techniques in the experimental setting of Chen *et al* [23] expressed in terms of AUC. Except for our Graph-one's variations, and the average kernels (Avg), these results were taken from that work.

information on the whole data distribution and at the same time scale to large datasets without the need of sub-sampling the examples. This may be of great advantage as typically disease genes are orders of magnitude less numerous than the candidates.

Results from two different evaluation settings show that our proposed method Scuba outperforms many existing methods, particularly in genome-wide analyses. Compared to the two considered existing kernel-based methods, Scuba performances (considering AUC) are always higher, and often significantly higher. Moreover, Scuba has two main levels of scalability that make it particularly suitable for gene prioritization

Altogether, our results show that Scuba is a valuable tool to achieve efficient prioritizations, especially in large-scale investigations.

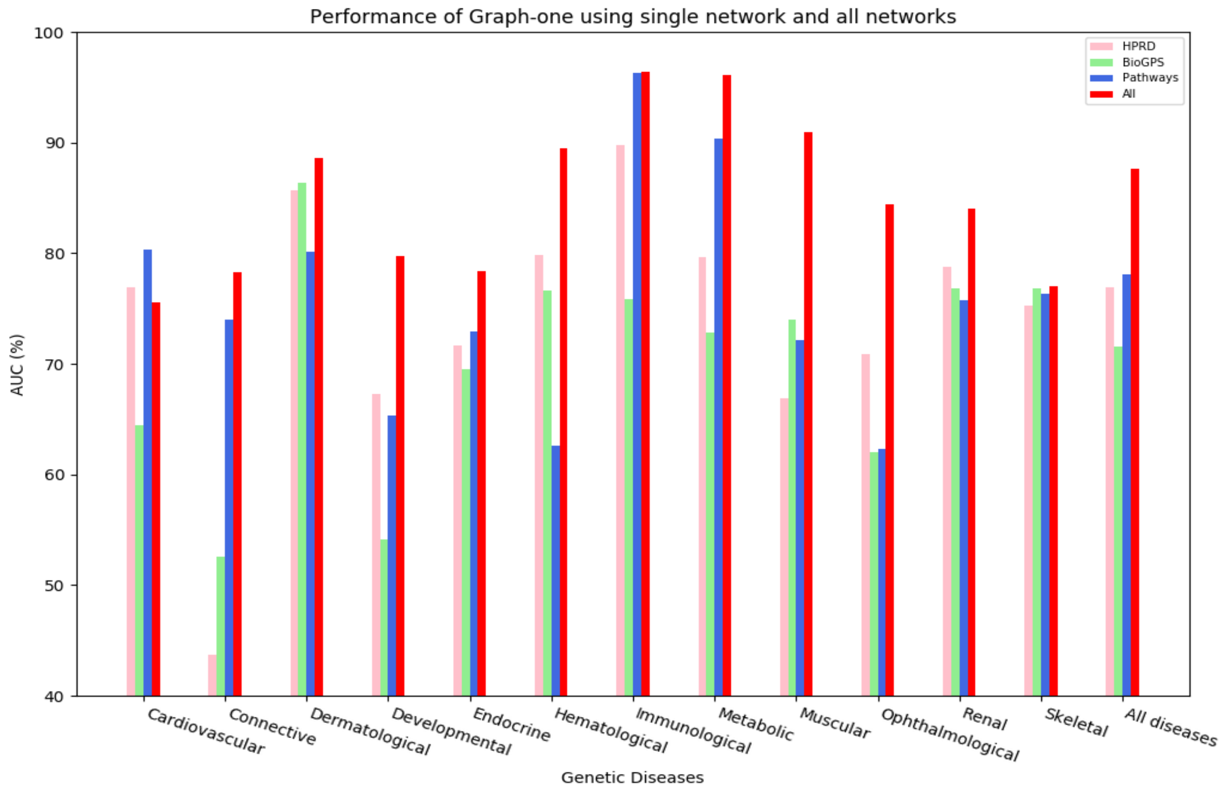

Figure 5.4: The performance of Graph-one (Scuba1): using single layer and combine all layers

Table 5.1: Performances of Graph-one, MKL1class and ProDiGe in the unbiased setting of Börnigen *et al* [17]. Values refer to predictions on all the 42 gene-disease associations. Rank difference p-values were obtained using Wilcoxon signed rank tests comparing separately Scuba1/MKL1class and Scuba1/ProDiGe ranks differences. Asterisks indicate significance of the tests at a threshold of 0.05.

Tool/Method	Rank median	Rank average	TPR in top 5% (%)	TPR in top 10% (%)	TPR in top 30% (%)	AUC	Rank difference p-value
Genome-wide prioritization methods							
Scuba1	**10.55**	**20.48**±23.53	**33.3**	**47.6**	**78.6**	**0.80**	-
MKL1class [91]	13.30	23.42±23.23	21.4	**47.6**	69.0	0.77	$2.5 \cdot 10^{-2}$ *
ProDiGe [60]	11.73	24.45±27.33	31.0	45.2	71.4	0.76	$3.0 \cdot 10^{-7}$ *
Candidate set-based prioritization methods							
Scuba1	**12.95**	**23.32**±25.46	**28.6**	**45.2**	**73.8**	**0.78**	-
MKL1class [91]	15.07	25.63±24.73	23.8	40.5	61.9	0.76	$9.7 \cdot 10^{-2}$
ProDiGe [60]	14.41	26.39±29.09	26.2	40.5	71.4	0.75	$2.7 \cdot 10^{-3}$ *

Table 5.2: Performances of Graph-one variations and of some gene prioritization web tools in the unbiased setting of Börnigen *et al* [17]. Response rate is the percentage of gene-disease associations considered by each tool. Values for Suspects were computed on the first 27 associations only (highlighted by ^{a}).

Tool/Method	Response rate (%)	Rank median	Rank average	TPR in top 5(%)	TPR in top 10(%)	TPR in top 30(%)	AUC
Genome-wide prioritization methods							
Scuba1	100	10.55	20.48±23.53	33.3	47.6	78.6	0.80
DIGI	100	**4.73**	**12.72±18.20**	**52.4**	**59.5**	**85.7**	**0.87**
Candid [41]	100	18.10	27.35±24.62	21.4	33.3	64.3	0.73
Endeavour [8]	100	15.49	21.47±22.37	28.6	38.1	71.4	0.79
Pinta [63]	100	19.03	23.52±23.58	26.2	31.0	71.4	0.77
Candidate set-based prioritization methods							
Scuba1	100	12.95	23.32±25.46	28.6	45.2	73.8	0.78
DIGI	100	**6.10**	**13.70±17.98**	**50.0**	**59.5**	88.1	**0.86**
Suspects [7]	88.9^a	12.77^a	24.64 ± 26.42^a	33.3^a	33.3^a	63.0^a	0.76^a
ToppGene [25]	97.6	16.80	34.53±35.31	35.7	42.9	52.4	0.66
GeneWanderer-RW [46]	88.1	22.10	29.55±26.28	16.7	26.2	61.9	0.71
Posmed-KS [90]	47.6	31.44	42.07±30.98	4.7	7.1	23.8	0.58
GeneDistiller [70]	97.6	11.11	15.37±13.77	26.2	47.6	78.6	0.85
Endeavour [8]	100	11.16	18.41±21.39	26.2	42.9	**90.5**	0.82
Pinta [63]	100	18.87	25.23±24.72	28.6	31.0	71.4	0.75

Table 5.3: Performances of Graph-one, MKL1class and ProDiGe in the expanded unbiased setting involving seven multi-factorial diseases. Values refer to predictions on 48 gene-disease associations. Rank difference p-values were obtained using Wilcoxon signed rank tests comparing separately Scuba/MKL1class and Scuba/ProDiGe ranks differences. Asterisks indicate significance of the tests at a threshold of 0.05.

Method	Rank median	Rank average	TPR in top 1% (%)	TPR in top 5% (%)	TPR in top 10% (%)	TPR in top 30% (%)	AUC	Rank difference p-value
Genome-wide prioritizations								
Scuba1	8.13	**17.45**±22.33	**10.4**	41.7	**58.3**	**79.2**	**0.83**	-
MKL1class [91]	14.28	25.79±26.96	2.1	27.1	45.8	66.7	0.74	$1.2 \cdot 10^{-5}$ *
ProDiGe [60]	**7.89**	18.40±23.77	**10.4**	**43.8**	54.2	**79.2**	0.82	$9.5 \cdot 10^{-2}$

5.3 Summary

In this chapter, we propose a general kernel-based framework for large-scale graph-based integration, named Graph-one. We theoretically and practically show that Graph-one meets the necessary requirements for graph-based integration methods: effective exploitating the graph integration complementary property, scalability, efficiency. Besides, it is able to handle with unbalance settings. Interestingly, Graph-one allows to create different variations under its general framework.

We evaluate Graph-one performance and compare it with various methods through disease gene prioritization context by using introducing its four different variations. The results in many cases illustrate that Graph-one shows state-of-the-art performance in disease gene prioritization and it is potential to apply for graph-based integration systems.

For future work, we plan to apply Graph-one other tasks where data can be represented in form of graphs.

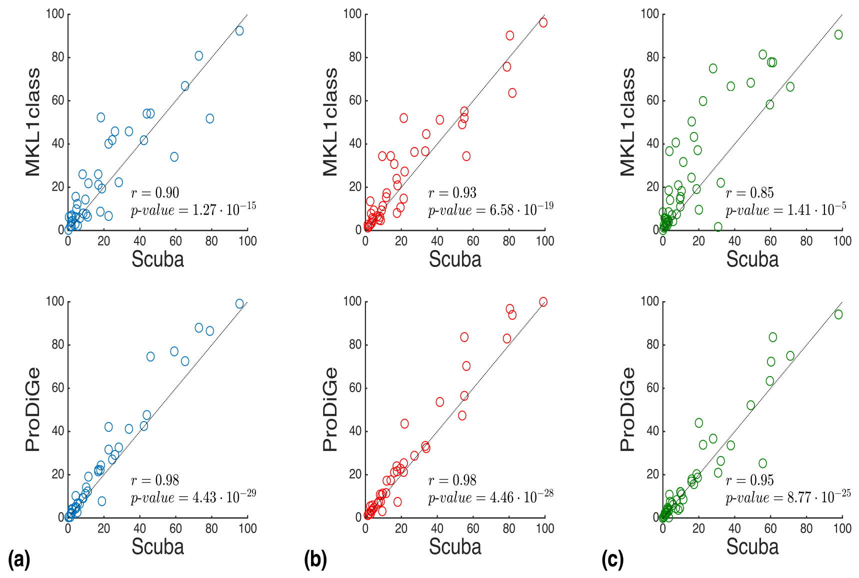

Figure 5.5: Comparison of normalized ranks predicted by Scuba and competing kernel methods. Normalized test genes rank distributions predicted by Scuba, MKL1class and ProDiGe for test genes in **(a)** genome-wide prioritizations in the unbiased evaluation of Table 5.1 - **(b)** candidate set-based prioritizations in the unbiased evaluation of the Table 5.1 - **(c)** genome-wide prioritizations in the expanded unbiased evaluation of the Table 5.3. In all cases, each point represents a test gene and lower values on the axes indicate better predictions. Genes lying on a diagonal have the same rank according to both methods considered on a plot. The further a gene lies above (below) a diagonal and the better it was ranked by Scuba (MKL1class/ProDiGe) compared to MKL1class/ProDiGe (Scuba). In each plot we show the Pearson correlation coefficient r between the test genes rank distributions and its associated p-value.

Chapter 6

Conclusion and Future Work

The abundance of biological data have been made available thanks to the rapid technological advances. In order to effectively extract useful knowledge from such huge data, automated systems are needed to support biologists in forming and assessing hypotheses. Biological data are distributed in different sources. Each source encrypts a piece of information related to a certain biological phenomenon and normally contains interactions between biological entities that can be naturally represented by graphs. Besides, it has been well known that a comprehensive understanding of a biological phenomenon can only come from a joint analysis of all data sources associated with that phenomenon. For these reasons, a high number of graph-based data integration learning systems have been proposed and applied to solve different tasks. Despite a high number of proposed methods for graph-based integration, the performance is still far from actual expectation. This is due to the challenges that graph-based integration methods need to effectively solve to have high performance: *node similarity measure definition*, *graph sparsity*, *scalability*, *efficiency* and *complemetary propery exploitation.* In this thesis, we investigated solutions to overcome those challenges with the target of having high performance graph-based integration learning systems.

6.1 Graph node measure definition

Graph node measure definition is one of the main factors which is responsible for the performance of any graph-based integration learning system. Node similarities are normally measured by graph node kernels. However, most graph node kernels are based on diffusion phenomenon. They do not often show high discriminative capacity and also do not allow to process with auxiliary information available on graph nodes. In Chapter 3, we propose a high discriminative convolutional graph node kernel, CDNK. We start from NSPDK, a kernel for measuring graph similarities, and adapt it to measure graph node similarities. CDNK takes an undirected, labeled graph for the input. First, it transforms the graph into a collection of linked connected components in which conjunctive and disjunctive links are introduced. Nodes linked by conjunctive links are jointly used to define the notion of context, while nodes linked by disjunctive edges are employed to define features based only on the pairwise co-occurrence of nodes at the endpoints. Second, the similarity between any node couple is measured by adopting NSPDK on two neighborhood subgraphs rooted as each node, extracted from the obtained graph. Last, CDNK integrates the "side" information available in form of real valued vectors by updating the computation of CDNK considering the discrete convolution of the discrete information with the real valued vectors. The empirical experiments on several biological networks on the context of disease gene prioritization show that our kernel outperforms the compared diffusion-based graph node kernels.

6.2 Graph sparsity

Graph-based integration learning systems take a set of graphs as the input. Despite the fact that huge amount of data have been made available, there exist a high number of graphs which are sparse, i.e. the number of links on a graph is much less than the possible links. This creates a challenge for graph-based learning systems, since the available information is not efficient to learn. A reasonable solution is to recover missing links and add into graphs. This task is known as link enrichment. However, the performance of link enrichment methods strictly depend on the employed link prediction methods which intend to predict the probability of being missing links for each unobserved one. However, most existing link prediction methods do not show high discriminative power since they implicitly represent the link prediction problem and the inference used to tackle it as a disjunction over the edges. In the Chapter 4, we first propose an effective link prediction method and we then show the way to boost the performance of diffusion-based kernels when dealing with sparse graphs.

We propose a link prediction method, named joint neighborhood subgraphs link prediction (JNSL). In this method, we first use a joint neighborhood subgraphs rooted at each node of a node couple as to represent the "link" between them. We then cast the problem of link prediction as a standard classification in the domain of graphs in which NSPDK is adopted for example similarity measure. Experiments on various networks illustrate JNSL outperforms many state-of-the-art topology-based link prediction methods.

Diffusion-based node kernels are popularly used for dense graphs. In case of sparse graphs, they relatively show low performance. We introduce an approach to boost the performance diffusion-based kernels by coupling link enrichment with diffusion kernels. More precisely, we perform link enrichment on the input graphs applying kernels. This is due to proper link enrichment helps to enrich the information of the graph, so it allows kernels to capture more information. We evaluate the effectiveness of this approach by comparing plain diffusion kernels and diffusion kernels with link enrichment. We carry out experiments on four different graphs in the context of disease gene prioritization on four graphs. The results show that associating diffusion kernels with link enrichment improve their average peroformance in all cases.

6.3 Graph-one: an efficient, scablable framework for graph-based integration

Graph-based integration methods can be considered as the backbone for the graph integration learning systems. Proposing such methods are challenging since they need to meet the above requirements. In Chapter 5, we propose a general kernel-based framework for graph-based integration, Graph-one. Given a set of data sources, it first constructs a graph whose each layer is derived from a data source. It then allows different graph layers to combine by a user defined combination function, so the obtained graph has less than or equal to the number of graph layers compared to the previous one. Next, various graph node kernels are applied on every layer to achieve a set of kernel matrices. These kernel matices are fed into an efficient, scalable MKL, UEasyMKL, to get a final kernel matrix. Finally, the final kernel matrix is used as the input for a kernel machine to build models for prediction.

Graph-one is a framework that meets all the requirements for graph integration. First, link enrichment can be applied on sparse graph layers to enrich the performance of kernels. Second, it is efficient and scalable because of the adoption of UEasyMKL that allow to deal with high number of kernels in a constant memory consumption and linear computation time w.r.t the number of defined kernels. Next, it effectively exploits the complementary property of data integration since various kernels are defined on each graph layer and these kernels are combined in a data driven way to get a final representation. Besides, using Graph-one is able to deal with unbalanced setting as shown in 5.

We practically evaluate the performance of Graph-one by introducing different variations created under the general framework and applying them to disease gene prioritization: Scuba, PLC and DIGI. The exmperimental results show that Graph-one variations outperform many compared methods, including some kernel-based data integration ones.

6.4 Future work

For future work, we keep investigating on graph-based integration because of its important role in the development of graph-based learning systems. In particular, we plan to focus on the following topics.

6.4.1 Graph side information

We often observe the availability of side information associated to nodes and/or links of graphs. These side information provide a complement for graph topological information. Therefore, methods defined on graphs are supposed to effectively not only exploit topological information, but also side information. In this thesis, our two proposed methods, JNSL and CDNK, do not fully exploit all information sources. Therefore, we will investigate to propose methods which are able to wisely combine all information available on graphs so that the performance of JNSL and CDNK will be improved.

6.4.2 Graph combination

Graph-one general framework allows graph layers to be combined in the user defined ways. Although we have proposed a graph combination method in DIGI, it is not clear in general that which layers should be combined together and how to combine them. We also put this problem in our future work. We aim at introducing suggestions for users to decide not only which layers should be combined, but by which ways to combine.

6.4.3 Applications

We theoretically show the strength of Graph-one for graph integration. However, we only practically show its performance on the disease gene prioritization context. Therefore, we plan to apply Graph-one in further contexts with two purposes. First, we would like to check the stability. Second, we investigate solutions to overcome problems arisen when applying to other contexts so that we can improve the strength of Graph-one in general.

Bibliography

[1] Ensembl, http://www.ensembl.org/.

[2] https://www.researchgate.net.

[3] http://www.scikit-learn.org/.

[4] Online mendelian inheritance in man, http://omim.org/.

[5] Rowies, s.: Nips dataset, http://www.cs.nyu.edu/rwoeis/data.html.

[6] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[7] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, 2006.

[8] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

[9] Fabio Aiolli, Giovanni Da San Martino, and Alessandro Sperduti. A kernel method for the optimization of the margin distribution. *Artificial Neural Networks-ICANN 2008*, pages 305–314, 2008.

[10] Fabio Aiolli and Michele Donini. Easymkl: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215–224, 2015.

[11] José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: A tool for the visualization of large scale networks. *arXiv preprint cs/0504107*, 2005.

[12] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[13] Hans-Dieter Block. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1):123, 1962.

[14] Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6(Sep):1579–1619, 2005.

[15] Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[16] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1):i47–i56, 2005.

[17] Daniela Börnigen, Léon-Charles Tranchevent, Francisco Bonachela-Capdevila, Koenraad Devriendt, Bart De Moor, Patrick De Causmaecker, and Yves Moreau. An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088, 2012.

[18] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[19] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics*, 33(3s):228, 2003.

[20] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[21] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'donnell, et al. The biogrid interaction database: 2015 update. *Nucleic acids research*, 43(D1):D470–D478, 2014.

[22] Pavel Chebotarev and Elena Shamis. The matrix-forest theorem and measuring relations in small social groups. *arXiv preprint math/0602070*, 2006.

[23] Bolin Chen, Min Li, Jianxin Wang, Xuequn Shang, and Fang-Xiang Wu. A fast and high performance multiple data integration algorithm for identifying human disease genes. *BMC medical genomics*, 8(3):S2, 2015.

[24] BoLin Chen, Min Li, JianXin Wang, and Fang-Xiang Wu. Disease gene identification by using graph kernels and markov random fields. *Science China. Life Sciences*, 57(11):1054, 2014.

[25] Jing Chen, Huan Xu, Bruce J Aronow, and Anil G Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC bioinformatics*, 8(1):392, 2007.

[26] Yixuan Chen, Wenhui Wang, Yingyao Zhou, Robert Shields, Sumit K Chanda, Robert C Elston, and Jing Li. In silico gene prioritization by integrating multiple data sources. *PloS one*, 6(6):e21137, 2011.

[27] Gene Ontology Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

[28] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[29] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262. Omnipress, 2010.

[30] Tijl De Bie, Léon-Charles Tranchevent, Liesbeth MM Van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics.*

[31] François Fouss, Kevin Francoisse, Luh Yen, Alain Pirotte, and Marco Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural networks*, 31:53–72, 2012.

[32] Francois Fouss, Luh Yen, Alain Pirotte, and Marco Saerens. An experimental investigation of graph kernels on a collaborative recommendation task. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 863–868. IEEE, 2006.

[33] Thomas Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1):49–58, 2003.

[34] Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[35] Faten Ghosn, Glenn Palmer, and Stuart A Bremer. The mid3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 21(2):133–154, 2004.

[36] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

[37] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[38] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268, 2011.

[39] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[40] David Haussler. Convolution kernels on discrete structures. Technical report, Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[41] Janna E Hutz, Aldi T Kraja, Howard L McLeod, and Michael A Province. Candid: a flexible method for prioritizing candidate genes for complex human traits. *Genetic epidemiology*, 32(8):779–790, 2008.

[42] Koki Ichinose, Jihoon Park, Yuji Kawai, Junichi Suzuki, Minoru Asada, and Hiroki Mori. Local over-connectivity reduces the complexity of neural activity: Toward a constructive understanding of brain networks in patients with autism spectrum disorder.

[43] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, et al. String 8a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(suppl_1):D412–D416, 2008.

[44] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[45] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772, 2008.

[46] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[47] Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, et al. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876, 2017.

[48] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322, 2002.

[49] Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[50] Elizabeth A Leicht, Petter Holme, and Mark EJ Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[51] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[52] Weiping Liu and Linyuan Lü. Link prediction based on local random walk. *EPL (Europhysics Letters)*, 89(5):58007, 2010.

[53] Linyuan Lü, Ci-Hang Jin, and Tao Zhou. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4):046122, 2009.

[54] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[55] Edward M Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W Rice, Todd O Yeates, and David Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[56] Victor A McKusick. Mendelian inheritance in man and its online version, omim. *The American Journal of Human Genetics*, 80(4):588–604, 2007.

[57] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer, 2011.

[58] Kurt Miller, Michael I Jordan, and Thomas L Griffiths. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284, 2009.

[59] Tom MITCHELL and McGraw HILL. *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.

[60] Fantine Mordelet and Jean-Philippe Vert. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC bioinformatics*, 12(1):389, 2011.

[61] Yves Moreau and Léon-Charles Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature reviews. Genetics*, 13(8):523, 2012.

[62] Ralf Mrowka, Wolfram Liebermeister, and Dirk Holste. Does mapping reveal correlation between gene expression and protein–protein interaction? *Nature genetics*, 33(1):15–16, 2003.

[63] Daniela Nitsch, Joana P Gonçalves, Fabian Ojeda, Bart De Moor, and Yves Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC bioinformatics*, 11(1):460, 2010.

[64] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34, 1999.

[65] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Stafford Noble. Learning gene functional classifications from multiple data types. *Journal of computational biology*, 9(2):401–411, 2002.

[66] Mirko Polato and Fabio Aiolli. Kernel based collaborative filtering for very large scale top-n item recommendation. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 11–16, 2016.

[67] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics*, 16(2):85, 2015.

[68] David Salgado, Matthew I Bellgard, Jean-Pierre Desvignes, and Christophe Béroud. How to identify pathogenic mutations among all those variations: variant annotation and filtration in the genome sequencing era. *Human mutation*, 2016.

[69] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1):D674–D679, 2008.

[70] Dominik Seelow, Jana Marie Schwarz, and Markus Schuelke. Genedistillerdistilling candidate genes from linkage intervals. *PLoS One*, 3(12):e3874, 2008.

[71] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis.* Cambridge university press, 2004.

[72] Nino Shervashidze and Karsten M Borgwardt. Fast subtree kernels on graphs. In *Advances in neural information processing systems*, pages 1660–1668, 2009.

[73] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011.

[74] Kilho Shin and Tetsuji Kuboyama. A generalization of haussler's convolution kernel: mapping kernel. In *Proceedings of the 25th international conference on Machine learning*, pages 944–951. ACM, 2008.

[75] T Strachan and AP Read. Human molecular genetics (new york: Garland science, taylor & francis group). 2011.

[76] Robert E Tarjan. Decomposition by clique separators. *Discrete mathematics*, 55(2):221–232, 1985.

[77] Marc A Van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack AM Leunissen. A text-mining analysis of the human phenome. *European journal of human genetics: EJHG*, 14(5):535, 2006.

[78] Martin H Van Vliet, Hugo M Horlings, Marc J Van De Vijver, Marcel JT Reinders, and Lodewyk FA Wessels. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PloS one*, 7(7):e40358, 2012.

[79] Vladimir Vapnik. Pattern recognition using generalized portrait method. *Automation and remote control*, 24:774–780, 1963.

[80] Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Geeta Joshi-Tope, Gopal Gopinath, David Croft, Bernard de Bono, Marc Gillespie, Bijay Jassal, Suzanna Lewis, et al. Reactome: a knowledge base of biologic pathways and processes. *Genome biology*, 8(3):R39, 2007.

[81] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. *Kernel Methods in Computational Biology*, pages 35–70, 2004.

[82] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.

[83] SVN Vishwanathan, Karsten M Borgwardt, and Nicol N Schraudolph. Fast computation of graph kernels. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 1449–1456. MIT Press, 2006.

[84] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[85] Xuefeng Wang, Eric P Xing, and Daniel J Schaid. Kernel methods for large-scale genomic data analysis. *Briefings in bioinformatics*, 16(2):183–192, 2014.

[86] Michael D Ward, Randolph M Siverson, and Xun Cao. Disputes, democracies, and dependencies: A reexamination of the kantian peace. *American Journal of Political Science*, 51(3):583–601, 2007.

[87] Michelle Whirl-Carrillo, EM McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, RB Altman, and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

[88] Chunlei Wu, Camilo Orozco, Jason Boyer, Marc Leglise, James Goodale, Serge Batalov, Christopher L Hodge, James Haase, Jeff Janes, Jon W Huss, et al. Biogps: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology*, 10(11):R130, 2009.

[89] Yoshihiro Yamanishi, Jean-Philippe Vert, and Minoru Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(suppl_1):i468–i477, 2005.

[90] Yuko Yoshida, Yuko Makita, Naohiko Heida, Satomi Asano, Akihiro Matsushima, Manabu Ishii, Yoshiki Mochizuki, Hiroshi Masuya, Shigeharu Wakana, Norio Kobayashi, et al. Posmed (positional medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic acids research*, 37(suppl_2):W147–W152, 2009.

[91] Shi Yu, Tillmann Falck, Anneleen Daemen, Leon-Charles Tranchevent, Johan AK Suykens, Bart De Moor, and Yves Moreau. L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC bioinformatics*, 11(1):309, 2010.

[92] Pooya Zakeri, Sarah Elshal, and Yves Moreau. Gene prioritization through geometric-inspired kernel data fusion. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1559–1565. IEEE, 2015.

[93] Yi Zhao, Hui Li, Shuangsang Fang, Yue Kang, Yajing Hao, Ziyang Li, Dechao Bu, Ninghui Sun, Michael Q Zhang, Runsheng Chen, et al. Noncode 2016: an informative and valuable data source of long non-coding rnas. *Nucleic acids research*, 44(D1):D203–D208, 2016.

[94] Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53, 2015.

Appendices

Chapter A

Conjunctive Disjunctive Graph Node Kernel

Table A.1: Indices and gene number of genetic disease classes

Index	Disease	Number
0	Connective	35
1	Cardiovascular	75
2	Dermatological	54
3	Developmental	37
4	Endocrine	62
5	Hematological	106
6	Immunological	94
7	Metabolic	159
8	Muscular	55
9	Ophthalmological	61
10	Renal	42
11	Skeletal	35

Table A.2: Predictive performance on 11 gene-disease associations in percentage using network induced by the BioGPS. Best results in bold. We report the AUC ROC and the rank for each kernel method. CDNK11: ontology for discrete labels, CDNK12: ontology for both discrete and vector labels, CDNK21: uniform discrete labels and CDNK22: uniform discrete labels and ontology for real vector labels.

				BioGPS				
Disease	LEDK	MDK	MEDK	RLK	CDNK11	CDNK12	CDNK21	CDNK22
1	51.9/8	57.4/7	59.0/6	59.2/5	65.1/4	69.5/3	72.0/2	**72.3/1**
2	81.7/5	78.5/6	75.2/7	75.0/8	88.3/1	**88.8/1**	83.2/4	84.8/3
3	64.3/6	59.6/8	71.6/3	71.8/2	65.5/5	**72.5/1**	61.4/7	67.9/4
4	65.3/7	58.2/8	67.8/6	67.8/5	71.9/4	**78.7/1**	73.8/3	76.7/2
5	64.0/8	64.1/7	66.5/5	66.2/6	75.9/4	76.2/3	77.8/2	**78.2/1**
6	74.6/5	70.2/8	71.0/7	71.2/6	79.3/3	**83.7/1**	77.2/4	80.0/2
7	73.0/5	66.7/7	75.4/3	75.6/2	68.8/6	73.9/4	66.1/8	**77.4/1**
8	74.4/8	76.8/3	76.2/5	76.4/4	74.7/7	77.7/1	76.1/6	76.9/2
9	71.5/2	65.6/8	67.7/6	69.9/3	66.8/7	**71.7/1**	67.9/5	68.1/4
10	54.0/6	50.3/8	56.1/5	51.1/7	77.6/4	**82.7/1**	77.7/3	78.3/2
11	58.2/7	51.3/8	59.3/6	59.3/5	71.8/4	**80.2/1**	74.8/2	74.0/3
$\overline{AUC}$	66.6	63.5	67.8	67.6	73.2	**77.8**	73.5	75.9
$\overline{Rank}$	6.09	7.09	5.36	4.82	4.45	**1.64**	4.18	2.27

Table A.3: Predictive performance on 11 gene-disease associations in percentage using network induced by the Pathways. Best results in bold. We report the AUC ROC and the rank for each kernel method. CDNK11: ontology for discrete labels, CDNK12: ontology for both discrete and vector labels, CDNK21: uniform discrete labels and CDNK22: uniform discrete labels and ontology for real vector labels.

				Pathways				
Disease	LEDK	MDK	MEDK	RLK	CDNK11	CDNK12	CDNK21	CDNK22
1	74.7/8	76.4/7	78.7/6	78.8/5	80.2/4	82.4/2	**82.8/1**	82.3/3
2	55.1/8	64.9/7	76.6/6	76.6/5	81.1/2	80.3/4	80.1/3	**81.9/1**
3	55.0/8	62.7/7	64.1/5	65.6/4	67.1/3	63.6/6	69.7/2	**71.4/1**
4	54.3/8	65.2/6	**73.7/1**	73.7/2	66.1/5	68.1/3	64.0/7	67.3/4
5	52.9/8	55.7/7	62.7/5	62.7/6	68.3/2	**69.6/1**	66.2/4	68.1/3
6	83.4/8	92.7/7	96.5/2	**96.5/1**	93.0/6	94.1/5	94.5/4	94.9/3
7	84.5/7	88.3/8	89.4/2	**89.5/1**	88.5/6	88.5/5	89.0/3	88.7/4
8	53.7/8	65.6/7	72.0/6	72.3/5	72.5/4	72.5/3	74.9/2	**75.3/1**
9	52.5/8	64.9/5	64.2/7	64.2/6	**81.3/1**	81.0/3	80.1/2	79.8/4
10	68.8/6	65.4/8	74.4/5	74.4/4	66.9/7	76.8/2	75.2/3	**78.7/1**
11	53.7/8	69.2/7	74.6/5	74.1/6	77.0/3	**78.7/1**	75.1/4	77.3/2
$\overline{AUC}$	62.6	70.1	75.2	75.3	76.5	77.8	77.4	**78.7**
$\overline{Rank}$	7.73	6.82	4.55	4.09	3.91	**2.27**	3.18	2.45

Chapter B

Solutions for Graph Sparsity

B.1 Joint neighborhood subgraphs link prediction

Table B.1: The AUC-ROC performance of link prediction methods on 3 datasets. In bold the highest score.

	Methods				
Datasets	LEDK(%)	MEDK(%)	MDK(%)	RLK(%)	JNSL(%)
BioGridgen	53.6±0.5	52.0±0.6	56.7±0.4	59.7±1.3	**69.9**±**1.5**
Omim	66.2±0.8	57.1±0.7	68.1±1.5	68.1±0.9	**72.5**±**1.4**
HPRD	58.0±1.2	58.0±1.2	64.9±1.2	63.5±0.7	**78.6**±**1.8**

Table B.2: The AUC-ROC performance of link prediction methods on 6 datasets. Legend: AA: Adamic-Adar, PA [12] preferential Attachment, SHP: Shortest Path, Sup-Top [57]: Linear regression running on unsupervised scores, SVD [57]: Singular value decomposition, Fact+Scores [57]: Factorization with unsupervised scores.

	Datasets					
Methods	Protein(%)	Metabolic(%)	Nips(%)	Condmat(%)	Conflict(%)	PowerGrid(%)
AA	56.4±0.5	52.4±0.5	51.2±0.2	56.7±1.4	50.7±0.8	58.9±0.3
PA	75.0±0.3	52.4±0.5	54.3±0.5	71.6±2.6	54.6±2.4	44.2±01.0
SHP	72.6±0.5	62.6±0.4	51.7±0.3	67.3±1.8	51.2±1.4	65.9±1.5
Katz	72.7±0.5	60.8±0.7	51.7±0.3	67.3±1.7	51.2±1.4	65.5±1.6
Sup-Top	75.4±0.3	62.8±0.1	54.2±0.7	72.0±2.0	69.5±7.6	70.8±6.2
SVD	63.5±0.3	53.8±1.7	51.2±3.1	62.9±5.1	54.1±9.4	69.1±2.6
Fact+Scores	79.3±0.5	69.6±0.2	61.3±1.9	81.2±2.0	68.9±4.2	75.1±2.0
JNSL	**80.1±0.8**	**72.5±0.7**	**62.1±0.8**	**88.6±2.3**	**72.0±0.9**	**75.6±0.7**

B.2 Link enrichment for diffusion-based graph node kernels

In this section, we present in detail the results obtained from the experiment described in 4.3.2. We report the performance on 14 disease gene classes and four different biological graphs induced by BioGPS, Biogridphys, HPRD and Omim. For each disease class and graph, we show the average AUC(%) for each diffusion-based graph node kernel in two cases: plain diffusion kernel (denoted by "-" and followed by kernel notation) and diffusion kernel on a modified graph obtained by link enrichment process (denoted by + followed by notation of kernel which is used for link enrichment). We notate each kernel as follow: A: LEDK, B: MEDK, C: MDK, D: RLK, E: CDNK.

Table B.3: Indices of genetic disease classes

Index	Disease
1	Cancer
2	Cardiovascular
3	Connective
4	Dermatological
5	Developmental
6	Endocrine
7	Hematological
8	Immunological
9	Metabolic
10	Muscular
11	Neurological
12	Ophthalmological
13	Renal
14	Skeletal

Table B.4: Average predictive performance on 14 gene-disease associations using four different graphs induced by the BioGPS, Biogridphys, Hprd and Omim. We report the average AUC-ROC (%) and standard deviations for all difussion-based kernels with (+) and without (-) link enrichment.

	BioGPS		**Biogridphys**		**Hprd**		**Omim**	
Disease	-	+	-	+	-	+	-	+
1	60.3±1.5	63.4±1.0	73.1±4.1	77.1±2.9	75.5±0.2	77.5±0.9	85.3±1.1	86.9±1.5
2	53.7±1.4	63.4±3.8	56.6±3.4	61.3±4.1	57.1±0.9	60.2±1.8	75.0±2.2	76.5±2.4
3	50.2±0.4	58.6±3.0	58.9±5.9	67.5±7.7	61.8±3.6	70.7±3.8	77.3±1.8	83.1±0.9
4	61.5±0.9	72.2±2.2	65.7±4.1	74.6±4.2	67.3±1.1	71.9±2.2	90.2±1.2	92.1±1.2
5	55.1±0.4	61.7±0.9	54.2±4.8	60.7±4.0	57.7±1.6	67.0±1.8	76.4±0.8	81.9±1.5
6	60.8±0.9	67.9±2.2	60.6±3.6	65.9±3.5	66.8±1.3	71.9±2.3	79.9±2.4	83.3±1.2
7	68.1±1.4	73.4±0.7	57.7±3.2	63.7±4.0	68.9±2.1	72.5±1.2	81.0±1.2	84.1±1.0
8	69.2±2.3	74.0±2.2	68.1±3.6	72.6±2.5	76.6±2.2	80.3±2.8	85.4±2.2	91.0±1.0
9	62.0±1.6	64.5±1.4	68.7±4.6	71.7±4.3	68.4±2.5	75.0±3.2	78.5±0.2	80.6±0.6
10	67.5±2.9	72.9±1.8	58.8±3.2	66.1±3.8	65.8±3.4	74.4±2.6	86.1±0.6	87.8±0.3
11	58.7±1.8	62.3±1.5	58.2±1.2	61.6±1.7	60.1±1.1	64.2±1.5	82.0±1.4	83.6±0.9
12	64.0±1.3	73.6±1.7	59.3±2.1	67.0±2.8	60.8±1.1	68.8±2.8	82.0±1.8	85.9±1.7
13	56.5±0.9	63.3±2.4	55.8±1.1	65.1±4.2	66.4±1.3	71.8±1.7	83.1±2.8	87.5±2.5
14	55.2±0.3	62.3±1.2	55.6±1.6	63.5±4.0	66.3±2.3	71.1±2.8	97.4±0.1	99.0±0.4
$\overline{AUC}$	60.2±0.3	66.7±1.2	60.8±1.6	67.0±4.0	65.7±2.3	71.2±2.8	82.8±0.1	86.0±0.4

Table B.5: Predictive performance on 14 gene-disease associations using network induced by the BIOGPS

Disease	-A	+A	+B	+C	+D	+E	-B	+A	+B	+C	+D	+E
1	61.8	63.8	62.5	63.7	63.9	63.4	57.9	62.8	62.8	62.2	62.6	62.2
2	55.6	67.5	62.3	70.4	69.5	63.5	52.0	62.8	57.3	57.9	60.1	58.3
3	50.3	62.3	61.2	63.6	61.9	60.0	49.5	57.0	55.1	56.1	57.2	55.2
4	61.3	73.9	73.1	73.3	73.7	73.9	62.7	72.7	67.9	72.5	74.7	68.5
5	55.7	62.8	60.6	63.2	62.9	62.9	55.1	61.7	61.0	62.9	62.6	60.5
6	60.1	68.8	66.8	70.4	69.4	67.4	59.7	66.1	62.2	68.2	67.0	63.4
7	68.8	73.6	73.0	73.6	74.2	73.4	67.7	73.5	73.1	74.2	74.8	72.1
8	71.1	77.9	72.3	76.5	77.8	73.2	65.6	71.7	70.8	70.4	71.8	70.9
9	62.4	63.8	63.1	64.4	64.3	63.8	63.7	66.9	65.7	67.2	66.7	66.5
10	69.4	72.9	71.2	74.5	72.6	71.8	62.5	73.6	70.4	72.7	71.9	71.1
11	60.0	63.9	63.1	64.6	63.6	63.3	58.4	61.4	61.9	61.4	61.8	61.7
12	64.4	76.2	73.5	76.1	75.2	73.2	61.9	71.2	70.3	71.7	72.3	70.5
13	57.2	66.1	62.4	65.8	63.0	59.3	55.1	63.7	63.5	62.4	61.8	60.6
14	55.4	61.7	61.7	61.6	63.0	60.1	55.0	63.7	64.8	62.1	63.1	62.5
$\overline{AUC}$	61.0	68.2	66.2	68.7	68.2	66.4	59.1	66.3	64.8	65.8	66.3	64.6

Table B.6: Predictive performance on 14 gene-disease associations using network induced by the BIOGPS

Disease	-C	+A	+B	+C	+D	+E	-D	+A	+B	+C	+D	+E
1	60.4	65.2	63.6	64.7	65.3	64.6	61.1	63.0	62.2	63.0	63.3	62.5
2	53.0	67.5	64.2	66.8	66.9	65.1	54.4	63.7	59.7	59.8	63.8	61.0
3	50.4	62.7	52.8	61.8	62.3	57.1	50.4	57.4	56.7	56.9	58.6	56.1
4	60.2	74.0	72.4	72.8	74.3	74.9	61.9	72.0	68.2	69.6	71.7	69.9
5	54.8	60.8	61.1	61.4	61.5	61.9	54.8	61.2	60.2	61.9	61.5	61.3
6	62.0	70.6	67.6	70.8	69.5	66.1	61.3	70.0	67.0	69.3	69.5	67.6
7	66.0	73.4	71.7	72.5	72.9	73.2	69.8	73.7	73.3	73.7	74.1	73.4
8	71.4	76.9	73.9	75.3	76.0	74.0	68.7	75.0	73.3	74.4	74.2	72.8
9	59.3	64.2	62.1	63.6	63.3	62.4	62.4	64.4	63.7	65.0	64.7	64.3
10	69.6	75.9	74.4	76.1	76.2	73.9	68.4	72.8	70.2	73.5	71.8	70.5
11	56.0	61.4	58.1	61.7	61.9	60.0	60.6	63.5	63.1	63.5	63.6	62.8
12	65.4	74.2	72.7	73.6	74.2	74.0	64.5	76.2	73.4	74.3	74.6	74.3
13	57.3	67.2	60.5	67.3	67.7	63.5	56.4	63.3	61.7	61.7	63.3	60.5
14	55.5	63.1	60.0	62.3	63.8	63.2	54.9	62.0	61.8	62.0	62.4	61.1
$\overline{AUC}$	60.1	68.4	65.4	67.9	68.3	66.7	60.7	67.0	65.3	66.3	66.9	65.6

Table B.7: Predictive performance on 14 gene-disease associations using network induced by the BIOGRIDphys

Disease	-A	+A	+B	+C	+D	+E	-B	+A	+B	+C	+D	+E
1	76.4	79.2	77.8	77.8	79.6	78.3	73.2	79.7	75.0	73.9	79.5	74.9
2	60.5	66.4	61.6	60.8	67.3	63.4	52.3	63.3	55.5	53.0	61.3	53.8
3	62.2	77.2	67.7	67.4	77.2	69.2	49.1	65.2	53.7	51.8	70.0	54.2
4	69.2	79.7	75.4	75.0	79.1	76.3	63.5	77.5	73.9	71.9	77.9	72.7
5	59.2	66.6	63.3	62.1	67.5	62.7	50.2	59.5	56.2	53.4	58.7	55.0
6	63.1	69.7	66.7	66.7	70.1	67.1	54.4	62.9	60.3	58.7	63.0	59.0
7	60.1	67.2	62.3	61.8	72.2	61.8	53.2	64.1	60.3	59.1	67.8	59.1
8	70.6	75.1	71.0	71.1	75.7	71.8	62.9	73.6	68.0	67.4	74.5	67.6
9	69.1	72.1	70.5	70.5	72.3	73.7	61.3	69.2	63.6	62.1	69.0	64.9
10	61.3	70.4	65.6	66.7	70.4	68.1	57.3	67.4	60.8	60.3	68.6	60.3
11	58.9	64.0	61.4	60.6	64.1	62.1	56.2	61.5	59.3	58.4	61.4	59.6
12	61.0	68.3	70.8	66.3	69.5	67.4	59.6	68.6	67.6	67.1	69.4	66.7
13	55.4	66.3	64.7	57.3	69.1	61.7	57.6	71.5	67.2	66.2	71.9	68.3
14	57.4	68.7	61.4	60.9	72.4	64.3	53.2	64.7	60.0	57.8	64.5	58.9
$\overline{AUC}$	63.2	70.8	67.2	66.1	71.9	67.7	57.4	67.8	63.0	61.5	68.4	62.5

Table B.8: Predictive performance on 14 gene-disease associations using network induced by the BIOGRIDphys

Disease	-C	+A	+B	+C	+D	+E	-D	+A	+B	+C	+D	+E
1	66.4	75.6	69.9	72.1	79.2	73.5	76.6	80.0	78.2	78.1	80.2	78.7
2	54.4	63.6	57.5	57.8	65.9	58.7	59.3	66.5	61.8	59.5	65.9	62.8
3	60.0	74.0	60.6	60.3	74.0	65.5	64.4	76.3	68.7	68.4	76.3	71.4
4	60.2	73.3	65.1	67.3	71.8	67.3	70.1	79.5	76.9	75.8	79.7	76.7
5	48.6	61.9	56.5	55.4	62.7	57.4	58.7	66.0	62.6	60.9	64.6	61.1
6	61.7	68.8	65.2	64.5	68.8	65.2	63.2	69.5	67.1	67.1	70.1	67.3
7	56.4	65.5	58.5	58.8	68.7	60.3	61.3	66.8	63.1	62.3	71.1	62.5
8	66.8	74.0	73.1	74.0	74.6	72.5	72.3	75.3	72.3	72.3	75.4	72.7
9	74.0	78.5	75.1	75.6	79.1	75.1	70.4	73.0	71.6	71.7	72.6	74.8
10	54.4	66.9	64.1	61.3	68.2	58.7	62.3	71.0	66.5	67.3	70.8	68.3
11	58.4	64.5	60.0	60.5	63.9	61.4	59.1	62.9	61.0	60.3	63.2	61.8
12	55.9	66.9	56.9	65.4	66.7	64.4	60.8	67.2	70.1	66.1	68.5	67.1
13	54.8	66.2	64.4	58.7	69.4	63.5	55.2	65.5	64.1	57.0	68.1	60.9
14	55.1	66.6	59.7	57.5	67.6	62.6	56.4	67.5	61.7	60.7	68.7	64.1
$\overline{AUC}$	59.1	69.0	63.3	63.5	70.0	64.7	63.6	70.5	67.6	66.3	71.1	67.9

Table B.9: Predictive performance on 14 gene-disease associations using network induced by the HPRD

Disease	-A	+A	+B	+C	+D	+E	-B	+A	+B	+C	+D	+E
1	75.5	77.7	77.8	76.7	77.9	76.1	75.1	77.2	77.0	76.5	76.0	75.9
2	56.6	61.1	58.6	59.6	63.5	58.6	55.9	59.5	56.6	58.0	60.2	58.0
3	61.2	73.0	69.0	71.3	76.1	67.9	56.1	70.0	61.1	67.7	71.1	63.3
4	67.3	74.9	71.0	70.9	74.1	69.7	65.5	72.7	68.6	68.0	71.8	67.5
5	57.6	66.9	65.6	68.6	68.4	68.2	55.1	67.3	62.9	64.5	66.5	62.6
6	67.1	73.7	72.0	72.1	73.5	69.3	64.8	71.8	69.8	69.1	68.6	66.4
7	68.8	73.6	72.4	72.7	73.7	71.9	65.5	71.9	69.4	71.6	72.4	69.8
8	76.2	82.4	76.7	81.9	83.3	77.1	73.4	80.7	73.6	79.5	82.3	74.8
9	68.2	76.2	75.6	71.2	74.5	77.3	64.4	72.2	70.9	68.6	69.4	76.0
10	66.0	76.2	75.9	74.6	76.6	73.8	60.4	74.9	71.8	69.2	71.0	69.4
11	60.5	65.2	63.2	64.4	65.0	63.8	58.3	64.1	60.7	62.7	63.2	61.5
12	61.9	69.8	69.3	68.5	72.1	64.3	60.5	68.8	68.8	67.6	68.0	63.0
13	67.6	72.2	71.9	73.3	72.6	71.4	65.1	71.5	69.4	69.4	69.7	67.1
14	68.4	74.1	72.7	71.8	74.1	71.2	64.6	70.5	66.7	67.8	66.9	67.5
$\overline{AUC}$	65.9	72.6	70.8	71.3	73.2	70.0	63.2	70.9	67.7	68.6	69.8	67.3

Table B.10: Predictive performance on 14 gene-disease associations using network induced by the HPRD

Disease	-C	+A	+B	+C	+D	+E	-D	+A	+B	+C	+D	+E
1	75.6	78.2	78.0	78.8	78.6	77.1	75.6	78.2	78.1	77.7	78.8	77.7
2	58.0	62.5	60.6	60.7	63.0	60.7	58.0	60.4	59.2	60.0	63.5	59.2
3	65.5	73.4	70.7	73.5	76.6	69.4	64.4	72.0	70.1	73.7	74.6	70.2
4	68.1	73.9	73.2	72.0	74.3	71.3	68.3	75.1	72.1	71.2	74.0	71.5
5	59.0	68.1	67.1	67.2	68.3	66.9	59.0	68.5	66.7	69.0	68.2	67.8
6	68.3	75.5	72.2	75.3	73.8	71.9	67.1	73.4	72.3	72.1	73.6	70.7
7	71.1	73.5	72.7	72.8	74.2	73.1	70.1	73.3	72.3	72.9	73.5	72.4
8	77.6	82.3	78.8	81.4	83.3	79.2	79.3	82.7	80.0	82.9	83.2	80.3
9	71.0	77.5	77.1	77.2	77.5	81.0	69.8	76.0	75.6	72.5	75.3	78.8
10	67.1	77.0	69.9	73.8	77.5	74.8	69.6	76.3	75.2	76.8	77.1	75.2
11	60.3	65.2	63.4	67.2	65.6	64.4	61.3	65.8	63.5	65.1	65.1	64.5
12	59.2	71.7	70.9	68.6	73.0	64.2	61.7	70.9	69.7	69.7	72.0	64.3
13	65.3	73.3	71.1	73.3	73.3	71.8	67.7	74.1	72.5	74.0	73.2	71.7
14	63.4	72.4	67.1	67.3	73.7	72.1	68.7	74.3	72.5	71.8	74.5	73.9
$\overline{AUC}$	66.4	73.2	70.9	72.1	73.8	71.3	67.2	72.9	71.4	72.1	73.3	71.3

Table B.11: Predictive performance on 14 gene-disease associations using network induced by the OMIM

Disease	-A	+A	+B	+C	+D	+E	-B	+A	+B	+C	+D	+E
1	84.9	86.9	85.6	87.4	86.7	85.8	84.4	86.3	85.3	85.9	85.9	85.4
2	76.6	78.4	78.4	78.8	78.4	78.4	75.7	77.1	77.4	77.3	77.1	77.1
3	78.4	83.3	83.3	83.8	83.2	84.6	74.4	82.1	80.8	82.5	82.4	82.2
4	91.3	93.0	93.0	93.0	93.0	93.1	89.3	93.0	91.5	92.6	93.0	92.1
5	76.1	82.3	79.7	84.1	80.4	80.7	75.6	83.9	79.6	84.3	82.4	79.9
6	81.9	84.7	83.7	83.9	84.6	83.8	79.1	82.7	80.8	83.6	84.1	80.8
7	81.2	85.2	84.5	84.1	84.9	83.2	79.5	84.3	82.8	82.2	84.3	81.9
8	84.3	90.5	91.4	91.0	90.5	92.2	83.9	90.1	89.8	89.7	90.1	90.9
9	78.8	80.6	80.4	80.8	80.4	80.4	78.2	79.9	79.9	80.1	80.3	80.4
10	86.3	87.6	87.6	87.6	87.6	87.6	86.7	88.1	88.1	88.1	88.1	88.1
11	83.3	84.6	84.6	84.6	84.6	84.6	79.8	83.0	81.7	83.1	82.9	82.2
12	82.0	87.3	85.2	87.0	86.8	85.3	79.8	85.3	81.9	87.2	85.4	81.8
13	85.0	88.6	88.9	88.6	89.4	89.2	84.1	88.9	88.6	89.4	89.6	89.1
14	97.4	99.2	98.6	99.2	99.2	99.5	97.2	99.3	98.2	99.1	99.1	99.6
$\overline{AUC}$	83.4	86.6	86.1	86.7	86.4	86.3	82.0	86.0	84.8	86.1	86.0	85.1

Table B.12: Predictive performance on 14 gene-disease associations using network induced by the OMIM

Disease	-C	+A	+B	+C	+D	+E	-D	+A	+B	+C	+D	+E
1	87.3	89.7	88.7	90.2	89.5	88.4	84.7	86.8	85.5	86.5	86.5	85.7
2	71.2	72.9	72.1	72.1	72.8	72.1	76.4	78.0	78.0	78.2	78.0	78.0
3	78.9	84.2	83.3	84.2	84.0	84.2	77.5	82.7	82.6	83.2	82.9	83.6
4	88.8	90.2	90.0	90.2	90.0	90.0	91.4	92.9	92.8	92.8	92.9	93.1
5	77.7	82.2	80.6	83.5	83.1	82.5	76.3	82.3	80.1	83.8	81.3	80.9
6	76.3	82.4	81.9	82.0	83.7	82.2	82.2	84.9	83.8	83.8	84.6	84.5
7	82.7	85.4	84.1	85.4	85.8	84.1	80.6	84.6	83.9	84.3	83.9	83.3
8	89.3	91.6	91.3	91.8	91.7	92.4	84.3	90.5	90.1	90.5	90.3	93.5
9	78.5	81.5	81.5	81.5	81.5	81.5	78.6	80.4	80.4	80.8	80.4	80.4
10	85.1	88.0	88.0	88.1	88.3	88.0	86.0	87.5	87.5	87.6	87.5	87.5
11	82.0	82.9	82.8	82.7	82.8	82.6	83.0	84.4	84.4	84.4	84.4	84.4
12	84.7	87.8	86.5	88.2	87.7	85.9	81.3	86.5	83.8	86.4	86.3	85.0
13	78.2	83.1	83.1	83.1	83.8	83.1	85.0	88.5	88.5	88.5	89.0	88.5
14	97.5	98.7	98.2	98.4	98.6	98.7	97.4	99.2	98.3	99.2	99.2	99.5
$\overline{AUC}$	82.7	85.7	85.2	85.8	86.0	85.4	83.2	86.4	85.7	86.4	86.2	86.3

Chapter C

Graph-based Data Integration Approaches

Table C.1: The performance of different techniques in the experimental setting of Chen *et al* [23] expressed in terms of AUC. Except for our Graph-one's variations, these results were taken from that work. The p-values indicate significance of the pairwise AUC differences with respect to Scuba AUC [39]. Asterisks indicate significance of the test (p-value < 0.05) of Scuba1 with compared methods.

Method	AUC(%)	p-value
F3PC [23]	0.830	$1.39 \cdot 10^{-4}$ *
MRF [24]	0.731	$<10^{-6}$ *
DIR [26]	0.716	$<10^{-6}$ *
GeneWanderer [46]	0.711	$<10^{-6}$ *
Avg	81.9	$<10^{-6}$ *
Scuba1	87.6	-
Scuba2	88.3	-
PLC	86.8	-
DIGI	88.1	-

Table C.2: Scuba1 AUC performance for each disorder in the evaluation of gene prioritization tools [17]

Disease	**Associated genes**	**genome-wide**	**candidate set**
Abdominal aortic aneurysm	ENSG00000136848	0.77	0.84
Alcohol dependence	ENSG00000148680	0.98	0.98
Arthrogryposis	ENSG00000152818	0.98	1.0
Asthma	ENSG00000182578	0.93	0.94
Autosomal recessive primary microcephaly	ENSG00000075702	0.41	0.44
Behcet's disease	ENSG00000136634	0.98	0.97
Bipolar schizoaffective disorder	ENSG00000146276 ENSG00000139618	0.97	0.98
Complex heart defect	ENSG00000121068	0.98	1.0
Congenital anomalies of the kidney and the urinary tract	ENSG00000164736 ENSG00000178188	0.97	0.96
Congenital diaphragmatic hernia	ENSG00000004961 ENSG00000154309	0.86	0.87
Crohn's disease	ENSG00000176920 ENSG00000185651 ENSG00000069399	0.89	0.89

Disease	**Associated genes**	**genome-wide**	**candidate set**
Dursun syndrome	ENSG00000141349	0.58	0.46
Ehlers-Danlos syndrome	ENSG00000169105	0.99	1.0
Esophageal squamous cell carcinoma	ENSG00000138193 ENSG00000101276	0.3	0.23
Leprosy	ENSG00000111537	0.96	0.9
Lung adenocarcinoma	ENSG00000073282	0.89	0.84
Methylmalonic aciduria	ENSG00000167775	0.9	0.93
Metopic craniosynostosis	ENSG00000106571	0.98	0.98
Mitochondrial complex I deficiency	ENSG00000177646	0.95	0.96
Multiple sclerosis	ENSG00000120088	0.83	0.84
Myelodysplastic syndromes	ENSG00000106462	0.81	0.83
Nasopharyngeal carcinoma	ENSG00000085276 ENSG00000127863	0.81	0.8
Nonsyndromic cleft lip/palate	ENSG00000148175	0.82	0.8
Parkinson's disease	ENSG00000175104	0.82	0.8
Periventricular heterotopia	ENSG00000102103	0.54	0.45
Primary biliary cirrhosis	ENSG00000142606 ENSG00000142539	0.82	0.77

Disease	**Associated genes**	**genome-wide**	**candidate set**
Psoriasis	ENSG00000056972	0.96	1.0
Retinal-renal ciliopathy	ENSG00000054282	1.0	1.0
Single-suture craniosynostosis	ENSG00000124813	0.98	0.98
Smooth pursuit eye movement abnormality	ENSG00000099901	0.27	0.2
Testicular germ cell tumor	ENSG00000137090 ENSG00000171681	0.5	0.41
Tetralogy of Fallot	ENSG00000145012	0.74	0.67
Type 2 diabetes	ENSG00000182247	0.21	0.19

Table C.3: Scuba1 AUC performance for single disorders considered in 5.3 in the main text. These are the multifactorial diseases employed by Börnigen *et al* [17] with at least a new gene annotation between March 2013 and February 2017 as reported by the Human Phenotype Ontology [47].

Disease	**Associated genes**	**genome-wide**
Behcet's disease	ENSG00000162594	0.87
	ENSG00000168811	
	ENSG00000138378	
	ENSG00000163823	
	ENSG00000136869	
	ENSG00000183542	
	ENSG00000164307	
	ENSG00000026103	
	ENSG00000206340	
	ENSG00000206450	
	ENSG00000134882	
Bipolar schizoaffective disorder	ENSG00000175344	0.68
	ENSG00000138592	
	ENSG00000151702	
	ENSG00000124782	
	ENSG00000171988	
	ENSG00000176986	
Crohn's disease	ENSG00000140368	0.90
Parkinson's disease	ENSG00000064692	0.89
	ENSG00000153234	
	ENSG00000116675	
	ENSG00000159082	
	ENSG00000184381	
	ENSG00000138246	
Primary biliary cirrhosis	ENSG00000128604	0.76
	ENSG00000181634	
	ENSG00000105329	
	ENSG00000110777	
	ENSG00000064419	
	ENSG00000016602	
	ENSG00000141076	
	ENSG00000106089	
	ENSG00000132912	

Disease	**Associated genes**	**genome-wide**
Psoriasis	ENSG00000206237	0.94
	ENSG00000196126	
	ENSG00000179344	
	ENSG00000206306	
	ENSG00000163599	
	ENSG00000206240	
	ENSG00000077150	
	ENSG00000141527	
	ENSG00000198246	
Smooth pursuit eye movement abnormality	ENSG00000104133	0.73
	ENSG00000171385	
	ENSG00000020922	
	ENSG00000070610	
	ENSG00000013503	
	ENSG00000167658	