

# Generalized Optimization Framework for Graph-based Semi-supervised Learning

Konstantin Avrachenkov  
INRIA Sophia Antipolis, France  
K.Avrachenkov@sophia.inria.fr

Alexey Mishenin  
St. Petersburg State University, Russia  
Alexey.Mishenin@gmail.com

Paulo Gonçalves  
INRIA Rhone-Alpes, France  
paulo.goncalves@inria.fr

Marina Sokol  
INRIA Sophia Antipolis, France  
marina.sokol@inria.sophia.fr

## Abstract

We develop a generalized optimization framework for graph-based semi-supervised learning. The framework gives as particular cases the Standard Laplacian, Normalized Laplacian and PageRank based methods. We have also provided new probabilistic interpretation based on random walks and characterized the limiting behaviour of the methods. The random walk based interpretation allows us to explain differences between the performances of methods with different smoothing kernels. It appears that the PageRank based method is robust with respect to the choice of the regularization parameter and the labelled data. We illustrate our theoretical results with two realistic datasets, characterizing different challenges: Les Miserables characters social network and Wikipedia hyper-link graph. The graph-based semi-supervised learning classifies the Wikipedia articles with very good precision and perfect recall employing only the information about the hyper-text links.

**Keywords:** Graph-Based Semi-Supervised Learning, Optimization Framework, PageRank, Wikipedia Article Classification.

## 1 Introduction

Semi-supervised classification is a special form of classification. Traditional classifiers use only labeled data to train. Semi-supervised learning use large amount of unlabeled data, together with labeled data, to build better classifiers. Semi-supervised learning requires less human effort and gives high accuracy. Graph-based semi-supervised methods define a graph where the nodes are labeled and unlabeled instances in the dataset, and edges (may be weighted) reflect the similarity of instances. These methods usually assume label smoothness over the graph (see the excellent book on the graph-based semi-supervised learning [16]). In this work we often omit “graph-based” term as it is clear that we only consider graph-based semi-supervised learning methods.

Up to the present, most literature on the graph-based semi-supervised learning studied the following two methods: the Standard Laplacian based method

(see e.g., [15]) and the Normalized Laplacian based method (see e.g., [14]). Here we propose a generalized optimization framework which implies the above two methods as particular cases. Moreover, our generalized optimization framework gives PageRank based method as another particular case. The PageRank based methods have been proposed in [3] as a classification stage in a clustering method for large hyper-text document collections. In [3] only a linear algebraic formulation was proposed but not the optimization formulation. A great advantage of the PageRank based method is that it has a quasi-linear complexity. In [5] a method also based on PageRank has been proposed. However, the method of [5] cannot be scaled to large datasets as it is based on the K-means method. The generalized optimization framework allows us to provide intuitive interpretation of the differences between particular cases. Using the terminology of random walks on graphs we also provide new probabilistic interpretation for the Standard Laplacian based method and the PageRank based method. With the help of the random walk terminology we are able to explain differences in classifications provided by the Standard Laplacian based method and the PageRank based method. The generalized optimization framework has only two free parameters to tune. By choosing the first parameter, we vary the level of credit that we give to nodes with large degree. By choosing the second parameter, the regularization parameter, we choose a trade-off between the closeness of the classification function to the labeling function and the smoothness of the classification function over the graph. We study sensitivity of the methods with respect to the value of the regularization parameter. We conclude that only the PageRank based method shows robustness with respect to the choice of the value of the regularization parameter. We illustrate our theoretical results and obtain further insights from two datasets. The first dataset is

a graph of co-appearance of the characters in the novel *Les Misérables*. The second data set is a collection of articles from Wikipedia for which we have expert classification. We have compared the quality of classification of the graph-based semi-supervised learning methods with the quality of classification based on Wikipedia categories. It is remarkable to observe that with just few labeled points and only using the hyper-text links, the graph-based semi-supervised methods perform nearly as good as Wikipedia categories in terms of precision and even better in terms of recall. With the help of the two datasets we confirm that the PageRank based method is more robust than the other two methods with respect to the value of the regularization parameter and with respect to the choice of labeled points.

The rest of the paper is organized as follows: In Section 2 we describe a generalized optimization framework for the graph-based semi-supervised learning and discuss differences among particular cases. In Section 3 we demonstrate our theoretical results by numerical examples. We conclude the paper in Section 4 with directions for future research.

## 2 Generalized Optimization Framework

The input to a semi-supervised classification consists of a set of data instances  $X = \{X_1, \dots, X_P, X_{P+1}, \dots, X_N\}$ . An instance could be described by a fixed collection of attributes. For example, all attributes can take real numbers as values and these numbers can be normalized. Suppose we have  $K$  classes and the first  $P$  instances in our dataset are labeled as  $k(i) \in 1, \dots, K$ ,  $i = 1, \dots, P$ . Let matrix  $W$  represent degrees of similarity between instances in  $X$ . The construction of  $W$  can be done by various method. If we continue with the example where attributes are given by normalized real numbers, the Radial Basis Function (RBF)

$$W_{ij} = \exp(-\|X_i - X_j\|^2/\gamma)$$

or k-Nearest Neighbors (kNN) method

$$W_{ij} = \begin{cases} 1, & \text{if } X_j \text{ is one of the } k \text{ nearest neighbors of } X_i, \\ 0, & \text{otherwise} \end{cases}$$

can be chosen to construct the similarity matrix. In the datasets of this article we assume that the matrix  $W$  is symmetric. The RBF method in fact gives a symmetric similarity matrix. In general, the kNN method can give non-symmetric matrix, but it could be easily transformed to the symmetric one by  $W' = (W + W^T)/2$ . Denote by  $D$  a diagonal matrix with its  $(i, i)$ -element equals to the sum of the  $i$ -th row of matrix

$W$ :

$$d_{i,i} = \sum_{j=1}^N w_{i,j}.$$

In some applications, which is also the case for our datasets, the similarity graph is available as a part of the data.

Define  $N \times K$  matrix  $Y$  as

$$Y_{ik} = \begin{cases} 1, & \text{if } X_i \text{ is labeled as } k(i) = k, \\ 0, & \text{otherwise.} \end{cases}$$

We refer to each column  $Y_{\cdot k}$  of matrix  $Y$  as labeling function. Also define  $N \times K$  matrix  $F$  and call its columns  $F_{\cdot k}$  classification functions. A general idea of the graph-based semi-supervised learning is to find classification functions so that on the one hand they will be close to the corresponding labeling function and on the other hand they will change smoothly over the graph associated with the similarity matrix. This general idea can be expressed with the help of optimization formulation. In particular, there are two widely used optimization frameworks. The first formulation, the Standard Laplacian based formulation [15], is as follows:

$$(2.1) \min_F \left\{ \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|F_i - F_j\|^2 + \mu \sum_{i=1}^N d_i \|F_i - Y_i\|^2 \right\}$$

and the second, the Normalized Laplacian based formulation [14], is as follows:

$$(2.2) \min_F \left\{ \sum_{i=1}^N \sum_{j=1}^N w_{ij} \left\| \frac{F_i}{\sqrt{d_{ii}}} - \frac{F_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i=1}^N \|F_i - Y_i\|^2 \right\}$$

where  $\mu$  is a regularization parameter. In fact, the parameter  $\mu$  represents a trade-off between the closeness of the classification function to the labeling function and its smoothness.

Here we propose a generalized optimization framework, which has as particular cases the two above mentioned formulations. Namely, we suggest the following optimization formulation

$$(2.3) \min_F \left\{ \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|d_{ii}^{\sigma-1} F_i - d_{jj}^{\sigma-1} F_j\|^2 + \mu \sum_{i=1}^N d_{ii}^{2\sigma-1} \|F_i - Y_i\|^2 \right\}$$

In addition to the Standard Laplacian formulation ( $\sigma = 1$ ) and the Normalized Laplacian formulation ( $\sigma = 1/2$ ), we obtain the third very interesting case when  $\sigma = 0$ . We show below that this particular case corresponds to

PageRank based clustering [3], for which (2.3) can be rewritten as:

$$\min_F \sum_{i=1}^N \sum_{j=1}^N w_{ij} \left\| \frac{F_i}{d_{ii}} - \frac{F_j}{d_{jj}} \right\|^2 + \mu \sum_{i=1}^N \frac{1}{d_{ii}} \|F_i - Y_i\|^2$$

Since the objective function of the generalized optimization framework is a sum of a positive semi-definite quadratic form and a positive quadratic form, we can state the following proposition.

**PROPOSITION 2.1.** *The objective of the generalized optimization framework for semi-supervised learning is a convex function.*

One way to find  $F$  is to apply one of many efficient optimization methods for convex optimization. Another way to find  $F$  is to find it as a solution of the first order optimality condition. Fortunately, we can even find  $F$  in explicit form.

**PROPOSITION 2.2.** *The classification functions for the generalized semi-supervised learning are given by*

$$(2.4) \quad F_{.k} = \frac{\mu}{2+\mu} \left( I - \frac{2}{2+\mu} D^{-\sigma} W D^{\sigma-1} \right)^{-1} Y_{.k},$$

for  $k = 1, \dots, K$ .

**Proof:** The objective function of the generalized semi-supervised learning framework can be rewritten in the following matrix form

$$Q(F) = 2 \sum_{k=1}^K F_{.k}^T D^{\sigma-1} L D^{\sigma-1} F_{.k} + \mu \sum_{k=1}^K (F_{.k} - Y_{.k})^T D^{2\sigma-1} (F_{.k} - Y_{.k}),$$

where  $L = D - W$  is the Standard Laplacian. The first order optimality condition  $D_{F_{.k}} Q(F) = 0$  gives

$$2F_{.k}^T (D^{\sigma-1} L D^{\sigma-1} + D^{\sigma-1} L^T D^{\sigma-1}) + 2\mu (F_{.k} - Y_{.k})^T D^{2\sigma-1} = 0.$$

Multiplying the above expression from the right hand side by  $D^{-2\sigma+1}$ , we obtain

$$2F_{.k}^T (D^{\sigma-1} (L + L^T) D^{-\sigma}) + 2\mu (F_{.k} - Y_{.k})^T = 0.$$

Then, substituting  $L = D - W$  and rearranging the terms yields

$$F_{.k}^T (2I - D^{\sigma-1} (W + W^T) D^{-\sigma} + \mu I) - \mu Y_{.k}^T = 0.$$

Since  $W$  is a symmetric matrix, we obtain

$$F_{.k}^T (2I - 2D^{\sigma-1} W D^{-\sigma} + \mu I) - \mu Y_{.k}^T = 0.$$

Thus, we have

$$F_{.k}^T = \mu Y_{.k}^T (2I - 2D^{\sigma-1} W D^{-\sigma} + \mu I)^{-1},$$

which proves the proposition.

As a corollary, we have explicit expressions for the classification functions for the three mentioned above particular semi-supervised learning methods. Namely, from expression (2.4) we derive

- if  $\sigma = 1$ , the Standard Laplacian method:  
 $F_{.k} = \frac{\mu}{2+\mu} (I - \frac{2}{2+\mu} D^{-1} W)^{-1} Y_{.k},$
- if  $\sigma = 1/2$ , the Normalized Laplacian method:  
 $F_{.k} = \frac{\mu}{2+\mu} (I - \frac{2}{2+\mu} D^{-\frac{1}{2}} W D^{\frac{1}{2}})^{-1} Y_{.k},$
- if  $\sigma = 0$ , PageRank based method:  
 $F_{.k} = \frac{\mu}{2+\mu} (I - \frac{2}{2+\mu} W D^{-1})^{-1} Y_{.k}.$

Let us now explain why the case  $\sigma = 0$  corresponds to the PageRank based clustering method. Denote  $\alpha = 2/(2 + \mu)$  and write  $F_{.k}$  in a transposed form

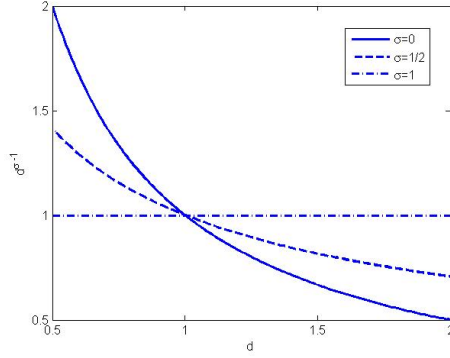
$$F_{.k}^T = (1 - \alpha) Y_{.k}^T (I - \alpha D^{-1} W)^{-1}.$$

If the labeling functions are normalized, this is exactly an explicit expression for PageRank [10, 9]. This expression was used in [3] but no optimization framework was provided.

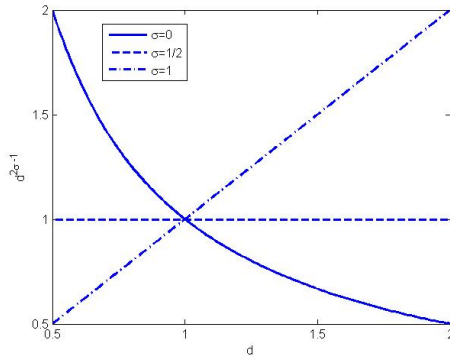
Note that  $D^{-1}W$  represents the transition probability matrix for the random walk on the similarity graph. Then, the  $(i, j)$ -th element of the matrix  $(I - \alpha D^{-1}W)^{-1}$  gives the expected number of visits to node  $j$  starting from node  $i$  until the random walk restarts with probability  $1 - \alpha$ . This observation provides the following probabilistic interpretation for the Standard Laplacian and PageRank based methods. In the Standard Laplacian method,  $F_{ik}$  gives up to a multiplicative constant the expected number of visits before restart to the labeled nodes of class  $k$  if the random walk starts from node  $i$ . In the PageRank based method with normalized labeling functions,  $F_{ik}$  gives up to a multiplicative constant the expected number of visits to node  $i$ , if the random walk starts from a uniform distribution over the labeled nodes of class  $k$ .

The random walk approach can explain why in some cases Standard Laplacian and PageRank based methods provide different classifications. For instance, consider a case when a node  $v$  is directly connected to the labeled nodes  $k_1$  and  $k_2$  belonging to different classes. Furthermore, let the labeled node  $k_1$  have a higher degree than the node  $k_2$  and let the node  $k_1$  belong

to a denser cluster than node  $k_2$ . From [4] we know that the expected number of visits to node  $j$  starting from node  $i$  until the restart is equal to the product of the probability to visit node  $j$  before the absorption and the expected number of returns to node  $j$  starting from node  $j$ . Then, the PageRank based method will classify the node  $v$  into the class of the labeled node  $k_2$  as it is more likely that the random walk misses the node  $v$  starting from node  $k_1$ . In other words, when the random walk starts from  $k_2$ , there are less options how to choose a next node and it is more likely to choose node  $v$  as a next node. In the Standard Laplacian method we need to compare the average number of visits to the labeled nodes starting from the node  $v$ . Since the random walk can reach either node  $k_1$  or node  $k_2$  in one step the probabilities of hitting these nodes before absorption are similar and what matters is how dense are the classes. If the class associated with the labeled node  $k_1$  is more dense than the class associated with the labeled node  $k_2$ , the node  $v$  will be classified to the class associated with  $k_1$ . We shall illustrate the above reasoning by a specific example in the next section.



(a) Smoothness term



(b) Fitting term

Figure 1: Fitting and smoothness terms

Based on the formulation (2.3), we could give some further intuitive interpretation for various cases of the generalized semi-supervised learning. Let us consider the first term in the r.h.s. sum of (2.3), which corresponds to the smoothness component. Figure 1(a) shows that if  $\sigma < 1$  we do not give much credit to the connections between points with large degrees. Let us now consider the second term which corresponds to the fitting function. Figure 1(b) shows that  $\sigma < 1/2$  does not give much credit to samples that pertain to a dense cluster of points (i.e.  $d_{ii}$  is large), whereas samples that are relatively isolated in the feature space (corresponding to small value of  $d_{ii}$ ), are given higher confidence. If  $\sigma = 1$ , the node degree does not have any influence. And if  $\sigma > 1/2$ , we consider that the nodes with higher weighted degree are more important than the nodes with smaller degree.

Next we analyze the limiting behavior of the semi-supervised learning methods when  $\mu \rightarrow 0$  ( $\alpha \rightarrow 1$ ). We shall use the following Laurent series expansion (2.5)

$$(1 - \alpha)[I - \alpha D^{-1}W]^{-1} = \alpha[\underline{1}\pi + (1 - \alpha)H + o(1 - \alpha)],$$

where  $\pi$  is the stationary distribution of the random walk ( $\pi D^{-1}W = \pi$ ),  $\underline{1}$  is a vector of ones of appropriate dimension and  $H = (I - D^{-1}W + \underline{1}\pi)^{-1} - \underline{1}\pi$  is the deviation matrix [12]. Let us note that if the similarity matrix  $W$  is symmetric, the random walk governed by the transition matrix  $D^{-1}W$  is time-reversible and its stationary distribution is given by

$$(2.6) \quad \pi = (\underline{1}^T D \underline{1})^{-1} \underline{1}^T D.$$

Let us insert the Laurent series expansion (2.5) into the expression for the general classification function (2.4):

$$(2.7) \quad \begin{aligned} F_{.k} &= (1 - \alpha)(I - \alpha D^{-\sigma} W D^{\sigma-1})^{-1} Y_{.k} \\ &= (1 - \alpha)[D^{-\sigma+1}(I - \alpha D^{-1}W)D^{\sigma-1}]^{-1} Y_{.k} \\ &= (1 - \alpha)D^{-\sigma+1}[I - \alpha D^{-1}W]^{-1} D^{\sigma-1} Y_{.k} \\ &= \alpha D^{-\sigma+1}[\underline{1}\pi + (1 - \alpha)H + o(1 - \alpha)]D^{\sigma-1} Y_{.k} \\ &= \alpha[D^{-\sigma+1}\underline{1} \sum_{i:k(i)=k} \pi_i Y_{ik} d_i^{\sigma-1} \\ &\quad + (1 - \alpha)D^{-\sigma+1} H D^{\sigma-1} Y_{.k} + o(1 - \alpha)]. \end{aligned}$$

Next, using the expression for the stationary distribution (2.6), we can specify (2.7) as follows:

$$(2.8) \quad \begin{aligned} F_{.k} &= \alpha[D^{-\sigma+1}\underline{1}(\underline{1}^T D \underline{1})^{-1} \sum_{i:k(i)=k} Y_{ik} d_i^{\sigma} \\ &\quad + (1 - \alpha)D^{-\sigma+1} H D^{\sigma-1} Y_{.k} + o(1 - \alpha)]. \end{aligned}$$

Hence, in the Semi-supervised methods when  $\alpha$  is sufficiently close to one, a class with the largest  $\sum_{i:k(i)=k} Y_{ik} d_i^\sigma$  attracts all instances. This implies that the limiting behavior of the PageRank based method ( $\sigma = 0$ ) is quite different from the limiting behaviors of the other methods. In particular, if the number of labelled points in each class is same or if the labeling functions are normalized, then there is no dominating class which attracts all instances and the classification results most likely make sense even for  $\alpha$  very close to one. The conclusion is that the PageRank based method is more robust to the choice of the regularization parameter than the other graph-based semi-supervised learning methods.

**Illustrating example:** to illustrate the limiting behaviour of the methods we generated an artificial example of the planted partition random graph model [6] with two classes with 100 nodes in each class. The probability of link creation inside the first class is 0.3 and the probability of link creation inside the second class is 0.1. So the first class is three times denser than the second class. The probability of link creation between two classes is 0.05. We have generated a sample of this random graph model. In each class we have chosen just one labelled point. In the first class we have chosen as the labelled point the point with the smallest degree (degree=28, 24 edges inside the class and 4 edges leading outside). In the second class we have chosen as the labelled point the point with the largest degree (degree=31, 27 edges inside the class and 4 edges leading outside). We have indeed observed that the second class attracts all points when  $\alpha$  is close to one for all semi-supervised methods except for the PageRank based method. This is in accordance with theoretical conclusions as the labelled point in the second class has a larger weight than the labelled point in the first class. It is interesting to observe that in this example the first class loses all points when  $\alpha$  is close to one even though the first class is denser than the second one.

### 3 Experiments

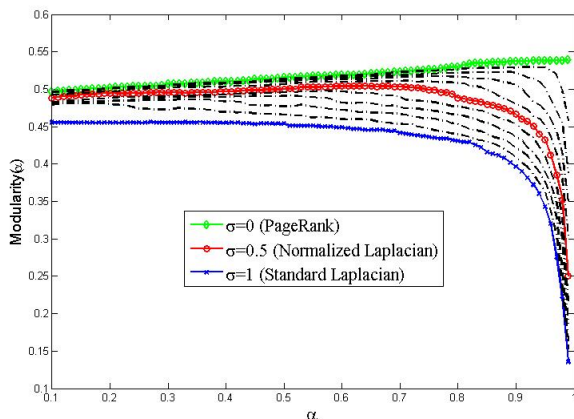
In this section we apply the developed theory to two datasets. The first dataset is the network of interactions between major characters in the novel *Les Misérables*. If two characters participate in one or more scenes, there is a link between these two characters. The second dataset is a subset of Wikipedia pages. Wikipedia articles correspond to the data points and hyper-text links correspond to the edges of the similarity graph. We disregard the direction of the hyper-text links.

**3.1 Les Misérables example** The graph of the interactions of *Les Misérables* characters has been com-

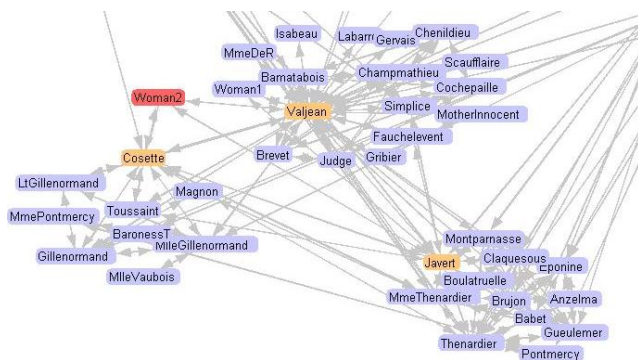
puted by Knuth [8]. There are 77 nodes and 508 edges in the graph. Using the betweenness based algorithm of Newman [11] we obtain 6 clusters which can be identified with the main characters: Valjean (17), Myriel (10), Gavroche (18), Cosette (10), Thenardier (12), Fantine (10), where in brackets we give the number of nodes in the respective cluster. We have generated randomly 100 times labeled points (one labeled point per cluster). In Figure 2(a) we plot the modularity measure averaged over 100 experiments as a function of  $\alpha$  for methods with different values of  $\sigma$  ranging from 0 to 1 with granularity 0.1. The modularity measure is based on the inter-cluster link density and the average link density and reflects the quality of clustering [11]. From Figure 2(a) we conclude that on average the PageRank based method performs best in terms of modularity and it is robust with respect to the choice of the regularization parameter. In particular, we observe that as was predicted by the theory the Standard Laplacian method and Normalized Laplacian method perform badly when  $\alpha$  is close to 1 (one class attracts all instances). The PageRank based method is robust even for the values of  $\alpha$  which are very close to one.

Next let us use the random walk based interpretation to explain differences between the Standard Laplacian based method and the PageRank based method. Let us consider the node Woman 2 (see Figure 2(b)). The node Woman 2 is connected with three other nodes: Valjean, Cosette and Javert. Suppose we have chosen labeled points so that only the nodes Valjean and Cosette are labeled but not Javert. Since the node Valjean has many more links than the node Cosette, the random walk starting from the node Valjean will less likely hit the node Woman 2 than the random walk starting from the node Cosette in some given time. Thus, the PageRank based method classifies the node Woman 2 into the class corresponding to Cosette. Since the node Woman 2 is just one link away from both Valjean and Cosette, the probability to hit these nodes before absorption is approximately equal. Thus, if we apply the Standard Laplacian method the classification will be determined by the expected number of returns to the labeled nodes before absorption. Since the labeled node Valjean lies in the larger and denser class, the Standard Laplacian method classifies the node Woman 2 into the class corresponding to Valjean.

**3.2 Wikipedia-math example** The second dataset is derived from the English language Wikipedia. In this case, the similarity graph is constructed by a slight modification of the hyper-text graph. Each Wikipedia article typically contains links to other Wikipedia articles which are used to explain specific terms and concepts.



(a) Modularity as a function of  $\alpha$ .



(b) Difference in classifications.

Figure 2: Les Misérables example.

Thus, Wikipedia forms a graph whose nodes represent articles and whose edges represent hyper-text inter-article links. For our experiments we took a snapshot (dump) of Wikipedia from January 30, 2010<sup>1</sup>. Based on this dump we have extracted outgoing links for other articles. The links to special pages (categories, portals, etc.) have been ignored. In the present experiment we did not use the information about the direction of links, so the graph in our experiments is undirected. Then we have built a subgraph with mathematics related articles, a list of which was obtained from “List of mathematics articles” page from the same dump. In the present experiments we have chosen the following three mathematical topics: “Discrete mathematics” (DM), “Mathematical analysis” (MA), “Applied mathematics” (AM). With the help of AMS MSC Classification<sup>2</sup> and experts we have classified related Wikipedia mathematical articles into the three above mentioned topics. According

to the expert annotation we have built a subgraph of the Wikipedia mathematical articles providing imbalanced classes DM (106), MA (368) and AM (435). The subgraph induced by these three topics is connected and contains 909 articles. Then, the similarity matrix  $W$  is just the adjacency matrix of this subgraph. Thus,  $w_{ij} = 1$  means that Wikipedia article  $i$  is connected with Wikipedia article  $j$ . Then, we have chosen uniformly at random 100 times 5 labeled nodes for each class. In Figure 3(a) we plot the modularity averaged over 100 experiments as a function of  $\alpha$  for methods with different values of  $\sigma$  ranging from 0 to 1 with granularity 0.1. Figure 3(a) confirms the observations obtained from Les Misérables dataset that the PageRank based method ( $\sigma = 0$ ) has the best performance in terms of the modularity measure. Next, in Figure 3(b) we plot the precision as a function of the regularization parameter for each of the three methods with respect to the expert classification. For the most values of  $\alpha$  the PageRank based method performs better than all the other methods and shows robust behaviour when the regularization parameter approaches one. This is in agreement with the theoretical conclusions at the end of Section 2. Both Figure 3(a) and Figure 3(b) demonstrate that the PageRank based method is also more robust than the other two methods with respect to the choice of labeled points.

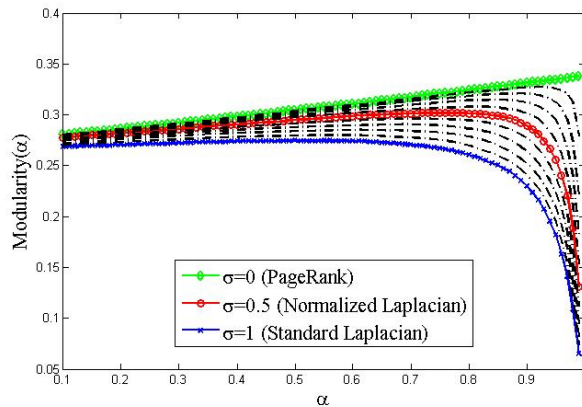
Figure 3(a) and Figure 3(b) also suggest that we can use the modularity measure as a good criterion for the choice of the regularization parameter for the Standard Laplacian and Normalized Laplacian methods. Now, let us investigate the effect of the quantity of the labelled data on the quality of the classification. Figures 4(a) 4(b) and 4(c) show that on average the modularity of the classification increases when we increase the quantity of the labelled data. Moreover, the quality of classification improves significantly when we increase the quantity of labelled data for each class from few points to about 50 points. The further increase of the quantity of the labelled data does not result in significant improvement in classification quality. The same behaviour manifests itself with respect to the precision measure (see Figures 5(a) 5(b) and 5(c)).

Both Les Misérables and Wikipedia-math datasets indicate that for the PageRank based method it is better to choose the value of the regularization parameter as close to one as possible but at the same time keeping the system numerically stable and efficient. This is an example of the singular perturbation phenomena [2, 13]

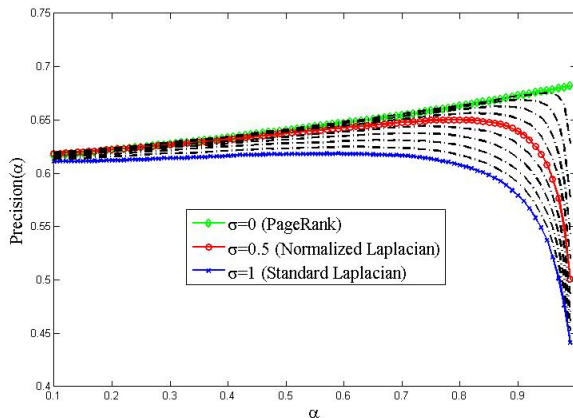
We have also compared the results obtained by the semi-supervised learning methods with the classification provided by Wikipedia Categories. As Wikipedia categories we have chosen: `Applied mathematics`,

<sup>1</sup><http://download.wikimedia.org/enwiki/20100130>

<sup>2</sup><http://www.ams.org/mathscinet/msc/msc2010.html>



(a) Modularity as a function of  $\alpha$ .



(b) Precision as a function of  $\alpha$ .

Figure 3: Wikipedia-math example

`Mathematical.analysis` and `Discrete.mathematics`. It turns out that the precision of the Wikipedia categories with respect to the expert classification is 78% (with 5 random labelled points the PageRank based method can achieve about 68%). However, the recall of the Wikipedia categorization is 72%. With the help of the semi-supervised learning approach we have classified all articles. It is quite interesting to observe that just using the link information the semi-supervised learning can achieve precision nearly as good as the Wikipedia categorization produced by hard work of many experts and the semi-supervised learning can do even better in terms of recall.

#### 4 Conclusion and future research

We have developed a generalized optimization approach for the graph-based semi-supervised learning which im-

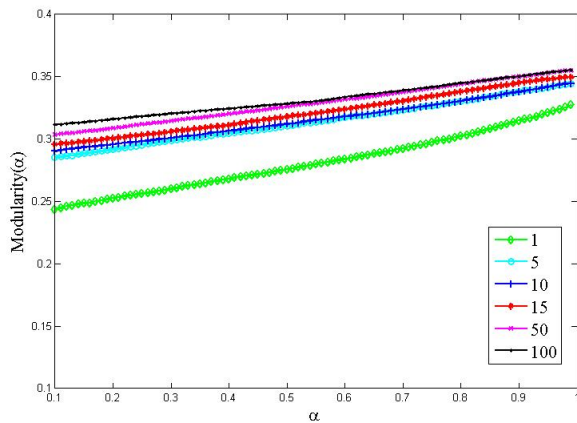
plies as particular cases the Standard Laplacian, Normalized Laplacian and PageRank based methods and provides the new ones based on parameter  $\sigma$ . We have provided new probabilistic interpretation based on random walks. This interpretation allows us to explain differences in the performances of the methods. We have also characterized the limiting behaviour of the methods as  $\alpha \rightarrow 1$  which based on the weight of the labelled points. We have illustrated theoretical results with the help of Les Miserables example and Wikipedia-math example. Also, we show how the number of labeled points has an influence on the quality of the classification. Both theoretical and experimental results demonstrate that the PageRank based method outperforms the other methods in terms of clustering modularity and robustness with respect to the choice of labelled points and regularization parameter. We propose to use the modularity measure for the choice of the regularization parameter in the cases of the Standard Laplacian method and the Normalized Laplacian method. In the case of the PageRank based method we suggest to choose the value of the regularization parameter as close to one as possible but at the same time keeping the system numerically stable and efficient. It appears that remarkably we can classify the Wikipedia articles with very good precision and perfect recall employing only the information about the hyper-text links. As future research we plan to apply the cross-validation technique to the choice of the kernel and to apply our approach to inductive semi-supervised learning [1, 7], which will help us to work with out-of-sample data.

#### Acknowledgement

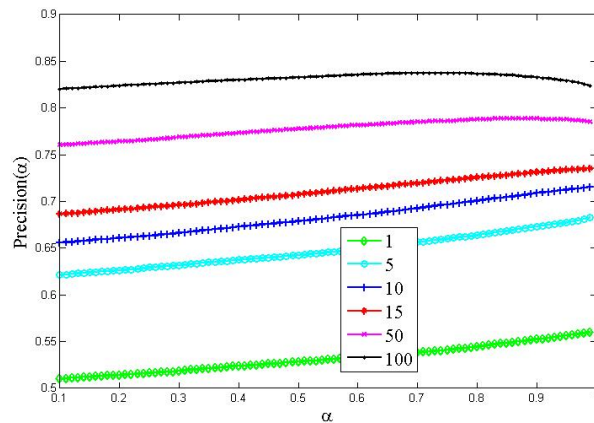
The work has been supported by the joint INRIA Alcatel-Lucent Laboratory.

#### References

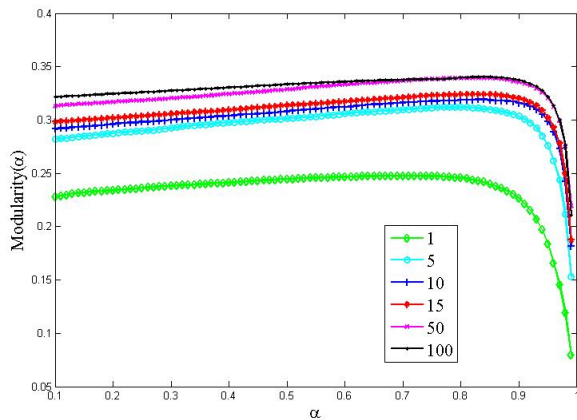
- [1] Yasemin Altun, David McAllester, and Mikhail Belkin. Maximum margin semi-supervised learning for structured variables. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 33–40. MIT Press, Cambridge, MA, 2005.
- [2] Konstantin Avrachenkov. *Analytic Perturbation Theory and its Applications*. University of South Australia, 1999.
- [3] Konstantin Avrachenkov, Vladimir Dobrynin, Danil Nemirovsky, Son Kim Pham, and Elena Smirnova. Pagerank based clustering of hypertext document collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 873–874. ACM, 2008.



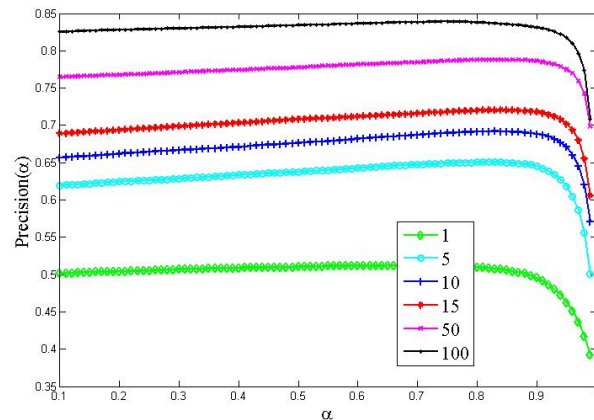
(a) Modularity as a function of  $\alpha$  for PageRank.



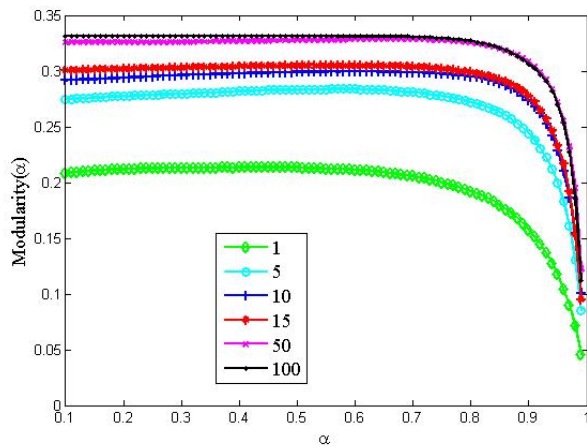
(a) Precision as a function of  $\alpha$  for PageRank.



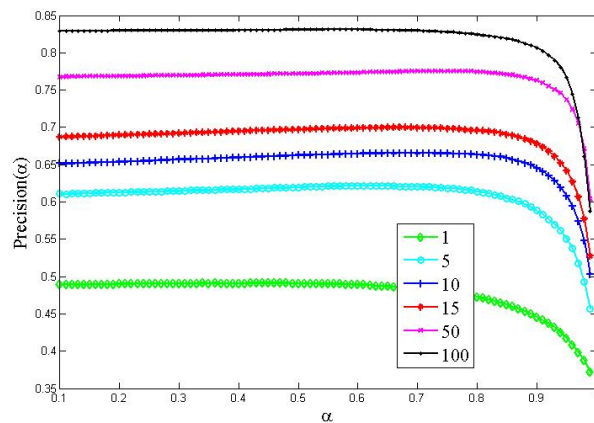
(b) Modularity as a function of  $\alpha$  for Normalized Laplacian.



(b) Precision as a function of  $\alpha$  for Normalized Laplacian.



(c) Modularity as a function of  $\alpha$  for Standard Laplacian.



(c) Precision as a function of  $\alpha$  for Standard Laplacian.

Figure 4: Wikipedia-math example

Figure 5: Wikipedia-math example



- [4] Konstantin Avrachenkov and Nelly Litvak. The effect of new links on google pagerank. *Stochastic Models*, 22(2):319–332, 2006.
- [5] Fan Chung and Alexander Tsias. Finding and visualizing graph clusters using pagerank optimization. In Ravi Kumar and Dandapani Sivakumar, editors, *Algorithms and Models for the Web-Graph*, volume 6516 of *LNCS*, pages 86–97. Springer Berlin / Heidelberg, 2010.
- [6] Anne Condon and Richard M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms*, 18(2):116–140, 2001.
- [7] Zhen Guo, Zhongfei (Mark) Zhang, Eric P. Xing, and Christos Faloutsos. Semi-supervised learning based on semiparametric regularization. In *SDM*, pages 132–142, 2008.
- [8] Donald E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*. ACM, New York, NY, USA, 1993.
- [9] Amy N. Langville and Carl D. Meyer. *Google page rank and beyond*. Princeton University Press, 2006.
- [10] Cleve B. Moler. *Numerical Computing with MATLAB*. 2004.
- [11] Mark E.J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.
- [12] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.
- [13] George Yin and Qing Zhang. *Continuous-time Markov chains and applications: a singular perturbation approach*. Applications of mathematics. Springer, 1998.
- [14] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004.
- [15] Dengyong Zhou and Christopher J. C. Burges. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th international conference on Machine learning, ICML '07*, pages 1159–1166. ACM, 2007.
- [16] Xiaojin Zhu and Andrew B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.